NIH Public Access
Author Manuscript
Immunol Allergy Clin North Am. Author manuscript; available in PMC 2007 August 8.

Published in final edited form as:
Immunol Allergy Clin North Am. 2007 February ; 27(1): 1–27.

Bioinformatics Approaches to Classifying Allergens and Predicting Cross-Reactivity

Catherine H. Schein[1,2], **Ovidiu Ivanciuc**[1], and **Werner Braun**[1,*]

1 *Sealy Center for Structural Biology and Molecular Biophysics, Departments of Biochemistry and Molecular Biology, University of Texas Medical Branch, 301 University Blvd., Galveston TX 77555-0857*

2 *Sealy Center for Structural Biology and Molecular Biophysics, Departments of Microbiology and Immunology, University of Texas Medical Branch, 301 University Blvd., Galveston TX 77555-0857*

Abstract

The major advances in understanding why patients respond to several seemingly different stimuli have been through the isolation, sequencing and structural analysis of proteins that induce an IgE response. The most significant finding is that allergenic proteins from very different sources can have nearly identical sequences and structures, and that this similarity can account for clinically observed cross-reactivity. The increasing amount of information on the sequence, structure and IgE epitopes of allergens is now available in several databases and powerful bioinformatics search tools allow user access to relevant information. Here, we provide an overview of these databases and describe state-of-the art bioinformatics tools to identify the common proteins that may be at the root of multiple allergy syndromes. Progress has also been made in quantitatively defining characteristics that discriminate allergens from non-allergens. Search and software tools for this purpose have been developed and implemented in the Structural Database of Allergenic Proteins (SDAP, http://fermi.utmb.edu/SDAP/). SDAP contains information for over 800 allergens and extensive bibliographic references in a relational database with links to other publicly available databases. SDAP is freely available on the Web to clinicians and patients, and can be used to find structural and functional relations among known allergens and to identify potentially cross-reacting antigens. Here we illustrate how these bioinformatics tools can be used to group allergens, and to detect areas that may account for common patterns of IgE binding and cross-reactivity. Such results can be used to guide treatment regimens for allergy sufferers.

Keywords

Structural Database of Allergenic proteins (SDAP); allergen databases; predicting IgE binding epitopes

Introduction

Allergy is a steadily increasing health problem for all age groups in our society. Food allergies, mostly against milk, eggs, peanuts, soy or wheat affect up to 8% of infants and young children

*To whom correspondence should be addressed: Telephone: 409-747 6810; Telefax: 409-747 6000; webraun@utmb.edu

Grant support: This work was supported by grants from the Food and Drug Administration (FD-U-002249), the Texas Higher Education Coordinating Board (ATP grant 004952-0036-2003) and the National Institute of Health (R01 AI 064913).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

[1–3]. In addition, many allergies to food (including shellfish, nuts, and fruits) and air-borne particulate matter (such as insect residue, tree and grass pollens) can develop later in life. One hypothesis is that this late onset may be the result of the individual being sensitized by long term exposure to environmental factors that contain proteins similar to those in the known triggers of allergenic response [1,4–6]. Recent studies have identified common molecular features of proteins from quite different sources, which could account for clinically important cross reactivity [7,8] and sensitivity [9,10]. For example, major allergenic proteins in peanut have been isolated, and peptides from their sequences that react with IgE from patient sera identified [6,11,12]. Proteins similar to these allergens were subsequently found in other foods that are known to elicit clinically significant responses in peanut allergic individuals [13], such as tree nuts [14], soy [15], and legumes [16,17].

To show how valuable comparing the proteins that are known to cause allergy can be, allergens classified as pathogenesis response proteins have extremely conserved sequences in many different plants [18–21]. Some of the sources of allergens belonging to pathogenesis related proteins of group 5 (PR5) are shown pictorially in Figure 1. The first allergen in this group was isolated from cedar pollen [22,23], but related proteins that bound to IgE from patients allergic to cherries, bell pepper, apple and tomato were subsequently identified [24]. As the sources of such closely related allergenic proteins are so different, sensitive individuals may indeed feel they are suffering from “multiple allergy syndrome”, when in reality they may have a strong reaction to one protein type that is found in many sources. This means that by identifying the common protein, using bioinformatics tools described in this article, more specific treatments can be defined [25].

Allergenic proteins from pollens, particularly from cedar [22,26–28], birch [29–31], and grass [32–35] are quite similar in overall sequence and structure. Other families of allergenic proteins have been isolated from primary food sources such as milk [36–38] (casein [37,39,40] and lactoglobulin [41–43]), egg [44] (ovomucoid [45–47] and lysozyme [48]), shrimp and related species (tropomyosins [49–51]), fish (parvalbumin [52–54]) legumes (albumins [55–58] and glycinins [59–61]).

Many of the allergens in pollen and fruits have similar sequences and structures, as Figure 1 makes clear. Identifying the protein group can be used to guide specific immunotherapy. For example, clinically important cross reactivities between birch pollen and several different types of fruits could be accounted for by demonstrating that they contained similar proteins [62–66]. In a group of patients with oral allergy syndrome to apples, as well as birch pollen sensitivity, the root cause was hypothesized to be the similarity between the allergen Bet v 1 and the apple protein Mal d 1. Specific immunotherapy with a Bet v 1 containing extracts was able to mitigate sensitivity to this fruit [67].

Successes of this sort explain the current interest in developing bioinformatics approaches to interpret the accumulated body of knowledge about the proteins that elicit severe IgE mediated reactions [8,68–70]. Until recently, much of the information about allergens was distributed in many different literature sources, making oversight and direct sequence comparisons difficult, especially for clinicians dealing directly with patients [71]. However, there are now several databases were that contain sequences and information about allergenic proteins (Table 1). This chapter will discuss the specific features in each of these databases and highlight how our Structural Database of Allergenic Proteins (SDAP, http://fermi.utmb.edu/SDAP/) [72,73] can be used for comparing the sequence, structure, and epitopes of allergens. SDAP has been designed to be user friendly, and to be of maximum use to clinicians as well as scientists interested in determining the molecular basis of allergen cross reactivity [13,24,74].

Part 1: Database approaches to classifying allergenic proteins

Overview of Allergen Databases

In Table 1 we list public databases dedicated to allergens, together with their URL addresses and a short list of features. Most of these databases are simply lists of allergenic proteins or sources, with limited cross-indexing. For example, the IUIS (International Union of Immunological Societies) site, available at http://www.allergen.org, simply lists alphabetically the official names of all the proteins this organization recognizes as allergens. The tables contain brief information for each allergen, such as source, references, and Genbank accession numbers for the sequence. However, there is no cross-indexing, making it difficult to identify relationships between proteins. Another database, AllAllergy (http://allallergy.net/), is a collection of links for allergy related Internet sites, which can be useful for obtaining clinical information, organizations, publications, events, and databases. Other recent reviews provide more details on general allergy databases [25,71].

Other databases are cross-indexed and much more useful to those seeking to identify cross-reacting allergens and their natural sources. Allergome (http://www.allergome.org) is a comprehensive database of clinical, biological and structural information about IUIS and non-IUIS recognized allergens. Allergome records the name, source, biochemical and immunochemical features for each allergen. While no computational tools are integrated in Allergome, there is a list of allergen MEME "sequence motifs" introduced by Stadler and Stadler [75]. Unlike the IgE based sequences we will present later, these motifs are quite long, and could be more properly called conserved domains in the sequences of closely related allergens.

Another database, maintained by the CSL (Central Science Laboratory, UK, http://www.csl.gov.uk/allergen/index.htm), lists official names of allergens with sequence links to Genbank. This is similar to the Biotechnology Information for Food Safety Database (National Center for Food Safety and Technology, http://www.iit.edu/~sgendel/fa.htm), which provides comprehensive lists of allergens with links to sequences from PIR, SwissProt, and Genbank. The Protall project (http://www.ifrn.bbsrc.ac.uk/protall/) developed a database of plant food allergens that contains links to detailed biochemical, structural, and clinical data. Since 2001 the development of the Protall database is part of the InformAll project (http://foodallergens.ifr.ac.uk/).

A few databases allow direct comparison of allergen sequences, using conventional search tools, and permit the use of the WHO guidelines for predicting potential allergenicity [76]. These guidelines specify that a protein might cross-react with an allergen if it is at least 35% identical over a frame of 80 amino acids or contains an exact match of any peptide of 6–8 amino acids with the allergen [77]. The FARRP database (http://allergenonline.com/asp/public/login.asp) contains a searchable list of allergens, sequence links to Genbank, and a FASTA search for related sequences. ADFS (Allergen Database for Food Safety, http://allergen.nihs.go.jp/ADFS/) contains allergen sequences and epitope information, and, implements the WHO allergenicity rules using FASTA. ALLERDB (http://sdmc.i2r.a-star.edu.sg/Templar/DB/Allergen/) lists official names of allergens, has a BLAST search, and a sequence comparison tool that implements the WHO guidelines.

Overview of the SDAP Database

Methods to check for how well a test protein matches any allergen, according to the WHO guidelines, are also implemented in our Structural Database of Allergenic Proteins (SDAP, http://fermi.utmb.edu/SDAP/) [72,73]. Unlike the publicly available allergen databases discussed above, SDAP has integrated search tools to allow a user to rapidly compare the molecular properties of allergenic proteins and their epitopes. In addition, SDAP contains special tools that were developed to compare short sequences, and these tools permit rapid identification of allergens that contain sequences statistically similar to known linear IgE epitopes [13,24]. SDAP was developed for basic research on the nature of allergenic proteins, and to allow regulatory agencies, food scientists and engineers a way to determine if a novel protein has allergenic potential. No special training is needed to access the data, and the tools are implemented in a user-friendly fashion. Searches are direct and rapid, so they can be done from a computer in a clinic with Internet access.

SDAP contains information on sequence, 3D structures and epitopes of known allergens from published literature and compiled databases on the Web [72,73]. The major uses of SDAP by clinicians are to determine food sources that could induce cross-reactions in sensitive individuals, and help in preparing dietary recommendations for allergic patients with a known sensitivity. For example, patients with a food allergy can be told to avoid other foods contain similar proteins to the ones in that know to trigger an allergic reaction. Alternatively, a recombinant protein can be used to determine the scope of the allergic response, and suggest candidates for specific immunotherapy.

The Basic Structure of SDAP

The best way to learn using SDAP is to go to the website (http://fermi.utmb.edu/SDAP/) and open the main search page, which has a link to a list of all the allergenic proteins in SDAP. Select an allergen of interest (for example, Asp f 1, a major fungal allergen) which opens a new descriptive page for that allergen. This page contains a summary of all the data archived in SDAP for the allergen, including the official name (according to the IUIS website listing, http://www.allergen.org/), P scientific and common name for the species, general source of the allergens, allergen type; species; systematic name; brief description; sequence accession numbers from SwissProt, PIR, NCBI and, where available, the PDB file name for a structure. All of this information is cross-referenced to the original data sources, which can be directly accessed by clicking on the links from each allergen page.

As noted in the Introduction, allergenic proteins belong to discrete groups, or families of structures. SDAP has several different methods for comparing the sequences of allergens within the database. The in-house bioinformatics methods permit almost instantaneous sequence similarity searching, with direct connections to much larger databases. To determine what other sequences are similar to the amino acid sequence of an allergen, one can do a full-length sequences for similarity to the known allergens, by selecting the BLAST and FASTA [78] searches. These searches can be activated directly from the main descriptive page for any of the allergens in SDAP of known sequence.

The Pfam grouping, discussed below in more detail, is available for most of the allergens in SDAP and rapidly indicates which other allergens a given protein resembles. Links to structural models for the various allergens allows one to visualize common areas of the proteins in 3-D. Yet another box on the allergen page indicates what IgE epitopes are known, and each sequence is linked to tools that can be used to automatically identify similar or identical sequences in all the other sequences in SDAP. Other tools allow the user to map the IgE containing peptides onto the experimental modeled structures of allergens. Figure 2 shows two examples of this tool, where the IgE epitopes of two fungal allergens Asp f 1 and Asp f 3 have been mapped on an experimentally determined structure (PDB 1AQZ) and on our MPACK model, respectively.

SDAP is also integrated with other bioinformatics servers, allowing the user to investigate structural similarity and neighbors using SCOP (Structural Classification Of Proteins) [79], TOPS (TOpological representation of Protein Structure) [80], CATH (Class, Architecture, Topology and Homologous superfamily) [81], CE (Combinatorial Extension of the optimal path) [82], FSSP (Fold Classification based on Structure-Structure alignment of Proteins) [83], and VAST (Vector Alignment Search Tool) [84].

Epitope Lists For Allergenic Proteins

SDAP is a unique allergen data source because it contains lists of IgE binding epitopes of allergenic proteins, assembled from the primary literature. Most of these sequence segments have been identified by in vitro binding to short peptides on solid phases that are assumed to represent epitopes that may be involved in eliciting allergic reactions. In a few cases the biological importance of the identified epitopes has been tested, for example by mutating these areas and showing that the IgE binding capacity was thereby diminished [85–87] or that the isolated peptides can interfere with antibody binding to the whole protein [27,28]. Currently, IgE epitope information is available for 27 allergens in SDAP: fungi (Alt a 1, Asp f 1, Asp f 2, Asp f 3, Asp f 13, Pen ch 18), pollen from Texas mountain cedar (Jun a 1, Jun a 3) and related species (Cha o 1, Cry j 1, Cry j 2), weed pollen (Par j 1, Par j 2), rubber (Hev b 1, Hev b 3, Hev b 5), yellowjacket venom (Ves v 5), peanut (Ara h 1, Ara h 2, Ara h 3), buckwheat (Fag e 1), hen egg (Gal d 1), soybean (Gly m glycinin G1, Gly m glycinin G2), English walnut (Jug r 1), shrimp (Pen a 1, Pen i 1).

We have developed special methodology for finding homologues of known epitopes in the sequences of other SDAP entries, the *PD* search, which will de discussed in some detail in Part III. The *PD* distance method was developed in our group specifically to detect meaningful similarities when comparing short sequences [72,73]. PD searches can reveal similarities in IgE epitope sequences, even from different allergenic proteins [13,24].

FASTA for Comparing the Overall Sequences of Allergenic Proteins

The first step in determining whether proteins are potentially cross-reactive is to determine their overall sequence similarity to other allergens by FASTA [78]. FASTA can be run automatically from any sequence file in SDAP by a mouse-click, and outputs a table that lists all similar allergens in SDAP with their “E-value”. Table 2 is an example of a FASTA search result in SDAP for the cedar pollen allergen Jun a 3. The last column of Table 2 lists the E-value which indicates the statistical significance of the hit. The E-value, or expectation value, is a measure of how many matches with the same sequence similarity one would expect to occur randomly in a database of a given size. Thus a low E-value (*e.g*., less than 10^{-6}) indicates a high significance of the sequence match.

Note that the most similar entry in SDAP for Jun a 3 (i.e, that with the lowest E-value) is another pathogenesis related protein (group 5) is from a related cypress tree, Cup a 3. However, similar allergenic proteins have been isolated from various vegetables and fruits, which are shown pictorially in Figure 1. Based on these FASTA alignments, we can anticipate that someone with severe cedar pollen allergy might also develop oral allergy symptoms [88,89] when eating, for example, apples or cherries. Other clinically relevant cross-reactivities, in addition to those mentioned in the introduction, have been shown, such as that linking dust mite sensitization and development of food allergies to shrimp and other crustaceans. This is postulated based on the similarities of the tropomyosins in these organisms [72,73,88].

Part II: Nomenclature and Classes of Allergenic Proteins

SDAP Can Aid in Determining Names for Newly Identified Allergens

Names of allergens are only official after approval by the IUIS (International Union of Immunological Societies, http://allergen.org/). Submitted allergenic proteins are named by a thorough process, agreed to by the member societies. Allergens are named by abbreviating the Latin name of the species from which they were isolated (*e.g*., *Cryptomeria japonica* becomes

Cry j) followed by a number that indicates the order in which they were identified (Cry j 1, a vicilin related to Jun a 1 from *Juniperus ashei* and similar allergens from other Taxaceae). After the original rounds of naming, the committee has tried, when possible, to maintain a structural or functional relationship across related taxa in the allergen numbering system. In the ideal case, the number would also be consistent with the protein class of the allergen. Thus, the PR5 related allergens in cypress pollens, regardless of species, would be number 3 (Cry j 3, Jun a 3). However, the numbering is not always consistent, and the PR5 allergens from apple, cherry, and bell pepper are Mal d 2, Pru av 2, and Cap a 1, respectively. Used at an early stage in the nomenclature process, SDAP can provide enough information to determine rapidly what other related allergens have similar sequences, and thus should have the same number,

Routine use of SDAP can prevent certain problems, such as those that may crop up when discoverers of a new allergen give it a name based on their understanding of how many other allergens have been previously isolated from the given biological source. Although this name may be changed by the IUIS, once a protein has been named in the literature, it is often difficult to obtain wide acceptance of a different designation [90]. For example, an allergen identified in *Juniper oxycedrus* was originally called Jun o 2 by its discoverer. However, when the IUIS examined this allergen, they realized that this was a different protein from the other cypress allergens named type 2, such as Jun a 2 and Cyp a 2 (see Table 3). Thus they assigned the protein an official name of Jun o 4 [91], with relevance to its similarity to the Bet v 4 protein of birch pollen (lower part of Table 3). SDAP names this protein Jun o 4, but alerts the user who types in Jun o 2 of the name change. This is because other databases, including PIR, GenBank and Swissprot, continue to identify Jun o 4 as Jun o 2. If SDAP had been used by initially to name this allergen, this confusing situation (which will get worse if a real Jun o 2 is isolated) would not have arisen.

The other feature of allergen nomenclature illustrated by Table 3 is that allergens with very closely related sequences, such as those in the bottom section of the table, can have widely differing numbers. Further, the names of closely related proteins can differ. For example, food allergens are generally categorized “seed storage proteins or “albumins”. Depending on their degree of identity to other proteins they may be additionally categorized as vicilins, proglycinins, 7S albumin, etc., which in turn can determine the common name of the allergen. As it is unlikely that a radically different nomenclature scheme will be introduced anytime soon, database searches like this one will provide the best way to truly indicate which allergens are most related to one another.

Grouping Allergenic Proteins According to Major Pfam Families

Classification of allergens into functional groups of proteins can indicate important relationships and has the additional advantage that structural and sequence groupings allow one to identify significant similarities in proteins with different names. These structural similarities may also underly functional similarities that are probably not related to the allergenic potential of the protein [74]. The most common protein groups for plant allergens are cupin, prolamin, plant defense system. Representative allergens from the cupin superfamily are vicilin and legumin from tree nuts, peanuts, and soybean. Important allergens from the prolamin superfamily are amylase and protease inhibitors, nonspecific lipid transfer proteins, and 2S albumin seed storage proteins. Plant defense proteins comprise allergens from several classes, such as pathogenesis-related proteins, proteases, and protease inhibitors [89,92,93].

The user of SDAP can identify allergens that are significantly similar to one another according to their Pfam or enzyme classification. Pfam (http://www.sanger.ac.uk/Software/Pfam/) is a list of multiple sequence alignments of related protein domains, classified in two ways. The Pfam-A database lists protein families that are grouped by their common function as well as sequence, using expert knowledge and experimental data. Pfam-B is computer-generated and contains alignments of proteins sequences selected based on a minimum level of sequence identity, regardless of their protein function. Most SDAP entries have now been classified to one of these groupings. Easy access to this Pfam classification for any allergen can be accessed from the “List SDAP” menu item.

The most common Pfam families for allergens are listed in Table 4, where we show 18 Pfam families that have between 34 and 7 allergens each. The most common Pfam families for allergens are PF00234 (protease inhibitor/seed storage/LTP family, 34 allergens), PF00235 (profilins, 27 allergens), PF00036 (EF hand, 23 allergens), and PF01357 (pollen allergen, 20 allergens). Table 5 lists allergens from these four Pfam families, with representative structures shown in Figure 3 (Pru p 3 from PF00234, Hev b 8 from PF00235, Bet v 4 from PF00036, and Phl p 2 from PF01357).

Part III: Computational Methods for Prediction of Cross-reactivity

Cross-reactive allergenic proteins are usually very similar in sequence and structure, at the molecular level, regardless of their source. As noted in the previous section, most allergens can be grouped into discrete sequence families, according to their Pfam classification. However, a typical Pfam will contain allergenic and non-allergenic proteins. Quantitatively discriminating the allergenic members in a group of similar proteins is a difficult task, and, as we will see below, one that eludes the programs currently implemented in popular databases. Experimentally, cross-reactive allergens typically have high sequence identity, which can drop to as low as 35% (hence the WHO guidelines). Still, point mutations are known to eliminate IgE binding [5,14,30,85] and one need only point to the example of isoforms of Bet v 1, which are 98% identical, that are not cross reactive [94–96]. For this reason, the reader should treat the results of current methods for predicting the allergenicity of a given protein with caution. The general consensus of a recent International Bioinformatics Workshop Meeting [97] on this problem concluded that more detailed statistical analysis of the properties of allergenic proteins versus non-allergens are needed and that numerical benchmarks for prediction methods should be developed. Below, we describe the current methodology, and outline other methods, based on identifying motifs of allergenic protein groups, that may have more success in defining cross-reactive allergens and areas for potential IgE recognition. The use of some of the more advanced methods implemented in SDAP, for comparison of IgE epitopes are emphasized, as this is one of the key features of our database that can be used in future to both design allergen vaccines and proteins with reduced allergenic potential.

Testing Automatic Computational Procedures for Allergenicity Prediction

One of the most difficult task in allergen recognition is to distinguish features of proteins that are allergenic from closely related proteins that are not [98]. The tropomyosin family is a particularly difficult problem, as the allergenic members of the family, such as Der p10 from dust mite and Met e 1 from shrimp, are highly identical to mammalian tropomyosins that are not allergenic. We tested the ability of three allergenicity prediction servers (WebAllergen, Allermatch, and AlgPred) to discriminate between four non-allergenic tropomyosins from animal sources and four allergenic tropomyosins from insects and shellfish (Table 6). The first server, WebAllergen [99], found that all the tropomyosins have 5 wavelet allergenic motifs [100], in common, while the allergenic tropomyosins have a few additional wavelet motifs that distinguish them. This is a promising start, if one could demonstrate that these allergenic motifs also contain IgE binding areas.

On the other hand, other rapid methods were unable to discriminate the two groups at all. Allermatch, which applies the FAO/WHO allergenicity guidelines, predicts that all eight tropomyosins are allergens. We also tested the MEME classifier based on motifs identified in groups of allergenic proteins [75]. This method, as implemented in the AlgPred server, predicted that all eight tropomyosins are non-allergens. Besides the MEME classification approach, the AlgPred server evaluates the allergenicity by scanning for known IgE epitopes, by a BLAST search, by a support vector machines (SVM) prediction based on amino acid composition or on dipeptide composition, and with a hybrid approach that combines the above five procedures [101]. The SVM dipeptide composition classifier predicts that all eight tropomyosins are allergens. In conclusion, it seems that, consistent with earlier results [102, 103] the WHO guidelines current allergenicity prediction servers cannot discriminate between closely related proteins, such as the non-allergenic and allergenic tropomyosins, as their overall sequences are too similar. However, other studies have shown that these guidelines are useful for detecting proteins that are sufficiently similar to known allergens that they might cross-react [104]. Several reports demonstrate that the succession of bioinformatics and experimental procedures from the FAO/WHO decision tree may be valuable in investigating the protein allergenicity [105–107].

Efforts to improve correlations based on the whole protein sequence are ongoing. Methods based on analyzing the statistics of FASTA alignments with machine learning procedures have been tested [108,109]. In one case, such an algorithm was able to classify 81% of 91 food allergens and 98% of 367 non-allergens correctly [110]. This level of accuracy could make the method useful for clinically discriminating protein groups to be avoided by patients with a known sensitivity to a related protein.

Another alternative is to detect discrete areas of a protein sequence similar to known IgE epitopes. The SDAP list of IgE epitopes, most which were identified by IgE recognition of linear arrays of synthetic peptides, is unique, as are the tools incorporated at the site for detecting identical and similar sequences (the PD tool) in other known allergens.

Motif-based Methods for Allergenicity Prediction

Alternatively, one can define discrete areas of residue conservation, “motifs”, in related allergenic proteins of known clinical cross-reactivity, as possible areas for IgE binding. Several groups have defined conserved sequences in groups of allergens [75,100,103,111]. For the purpose of this discussion, a motif is defined as an area of sequence that is extremely conserved in a group of related proteins. While motifs can be quite long, for the practical purpose of defining areas likely to be IgE epitopes, a normal length is between 6 and 15 amino acids. In our work, we look for areas where the side chains show conserved physical-chemical properties (PCPs), such as hydrophobicity, size or alpha-helical propensity, rather than strict identify. The underlying assumption is that for a group of cross-reactive allergenic proteins, the IgE epitopes areas will have common physical-chemical properties. Our method begins by aligning the sequences of known allergens that are related to one another, such as those in the tropomyosin or vicilin family. The PCPMer suite (available at http://landau.utmb.edu:8080/WebPCPMer/HomePage/index.html) finds sequence motifs in protein families by identifying regions with highly conserved physical-chemical properties. These “PCP-motifs” are determined by conservation of the five quantitative property vectors E_1–E_5 which summarize many different physicochemical properties of the side chains of the amino acids, including size, hydrophobicity, and tendency to form helical or strand secondary structures [112,113]. The descriptors E_1 to E_5 were determined by multidimensional scaling of 237 such physical-chemical properties, and encode numerically the maximum distance between the various side chains in a five dimensional space. These descriptors are also the basis of our PD scale for classifying sequences similar to know IgE epitopes, as described below.

Other efforts to predict allergenicity were directed towards identification of linear epitopes [114]. Saha and Raghava used a recurrent neural network for the prediction of continuous B-cell epitopes [115]. While short amino acid sequences matches seem to have little value for allergenicity prediction [116], peptide motifs common to groups of allergens may be a better way to distinguish allergens. An efficient machine learning classification scheme, based on identifying a set of allergen-representative peptides that appear in allergens but have a low or no occurrence in non-allergens, outperformed the FAO/WHO allergenicity rules [117]. Further, as Table 6 shows, the motif-based allergenicity prediction scheme based on wavelet transform did find areas of the allergenic tropomyosins that were not present in the non-allergenic. Further development of this method, taking into account the physicochemical properties of the amino acids, and solvent accessibility, was able to correctly classify 93.0% of 229 allergens tested, and 99.95 of non-allergenic proteins [118].

Sequence Similarity Ranking in SDAP: The Property Distance Scale *PD*

The FASTA search in SDAP is a rapid way to determine the overall similarity of large proteins. However, this program was not designed to compare short sequences, such as the linear IgE epitopes that have been identified by peptide mapping for many allergens [13,39,42,119]. Two different tools were incorporated in SDAP to look for short sequences in other known allergens, an “exact search”, that finds short sequences identical to that of a known epitope, and a second tool, to determine sequences that are close to the IgE epitope in “property space”. The *PD* tool determines similar sequences in other allergen entries in SDAP that have similar overall physical-chemical properties [72,73]. Peptides with identical sequences have a *PD* value of 0, and peptides with conservative substitutions of a few amino acids have a small *PD* value, typically in the range of 0 to 3. Peptides with a recognizable similarity in their physical chemical properties tend to have *PD* values lower than 10, while unrelated peptides have *PD* values that are much higher [120]. Table 7 shows two typical *PD* searches, done with the automatic tools in SDAP, starting with the sequence of two IgE epitopes of the Jun a 3 protein.

Additional data is needed to determine the statistical significance of the identified regions in the other allergens, particularly in a database as large as SDAP. The *PD* search is designed to compare protein regions with lengths comparable with those of published linear IgE epitopes. Significance levels for the sequence-similarity index *PD* are set high enough to detect all peptides that are similar to an IgE epitope, but low enough to discriminate them from other regions in the ensemble of allergenic proteins that would match randomly. For the search, each area of all the sequences is individually matched, with a window for the sequence segment that moves progressively by one position. Thus a 200 amino acid protein would have 194 different “sequence windows” of 7 amino acids, and 191 for a 10 mer. All *PD* searches in SDAP are followed by two histograms. The “lowest scoring window” is determined for the “best matching” peptide in the ~850 protein entries in SDAP. As the data in the heading of Table 7 indicate, for a *PD* search starting from a known epitope of Jun a 3, the average *PD* value for the best scoring sequence window in each of the 854 full length entries in SDAP was 12.15 (*SD* = 1.29) According to this test, values below about 9 (mean value – 2× the standard deviation) would be significantly similar to the test peptide. However, there are many similar sequences and isoforms in SDAP, which tends to skew the statistics for peptides. As a better estimate of what a random match would be, a second histogram summarizes the scores for all ~190,000 windows of a given size in all the SDAP allergen sequences. The average values in this histogram range from 17–26, depending on peptides. For the second example of Table 7, the average *PD* value for all 190530 possible windows was 17.24 (SD=1.78). According to these statistics for a random match, peptides with PD values below 10 would be clear outliers.

To give the reader a sense of the significance of the match, Z scores (which indicate the quality of the match relative to the database random distribution), are calculated automatically along with the PD value. The *lower* the *PD* score, the more closely related the peptide sequences are, but *high* Z-scores indicate better significance for the match.

The *PD* searches from two Jun a 3 IgE epitopes (Table 7) illustrate the usefulness of using the *PD* value to identify potential epitopes and potentially cross reactive allergenic proteins. For both epitopes, areas of thaumatin proteins from other pollens and fruits have the lowest *PD* value, consistent with their overall similarity according to our FASTA search (Table 2). Note that the order of the sequences in the two tables is a bit different. Although the IgE epitopes have not been identified for the fruit allergens, the Jun a 3 epitope 4 has a low *PD* value to a known IgE epitope of the latex allergen Hev b 3. The significance of this finding for cross-reactivity has not been determined. Other results with the peanut allergen epitopes [13] have identified epitopes with similar IgE reactivity and predicted structure for *PD* scores as high as 9.5–10.

We should at this point emphasize that the *PD* search is a computational way to define the sequence relationship between known IgE epitopes and other sequences in allergenic proteins. The correlation of *PD* values to meaningful IgE cross-reactivity, and eventually, to clinically relevant ones, is ongoing. However, initial tests indicate that this is a rapid way to quantify local similarities in known allergens. The structure of epitopes and their location on the protein surface (“solvent exposure”) are other possible factors determining whether a given sequence will bind IgE or not [13]. We believe the most promising methodology is to compare not just the sequences, but also the structures of areas before suggesting possible cross reactivity, as described in the next section.

Combining Sequence and Structural Information to Improve Prediction

Many questions about the nature of the IgE epitopes of allergens remain to be answered. Why, for example, do some individuals show cross-reactivity to homologous proteins in peanuts and tree nuts, while others with strong allergies only react to one or another of the homologous proteins [121]? While single amino acid differences may be quite important in individual reactivity, a 3D view of the identified IgE binding sites can provide missing information about the possible relationships between structure and sequence. If IgE binding sequences of related proteins have similar properties, the proposed methods that combine PD values with structural clues will have predictive ability, if properly calibrated.

Once similar sequences have been identified by PD values, the structural information in SDAP can be used to understand which parts of an allergen sequence are likely to be surface exposed, and thus likely to form an IgE binding surface [13,24]. Determining which residues are on the surface of an allergen (and thus most likely to react with an antibody) can also be determined rapidly and automatically using a program developed in this group, GETAREA (http://www.scsb.utmb.edu/cgi-bin/get_a_form.tcl). This program has also been incorporated at the site, where the data can be quickly accessed for SDAP allergens of known or modeled structure. SDAP allows direct access to the experimental structures (out of 586 SDAP allergens, 45 have known PDB structures). We estimate that for more than 90% of allergens where the experimental structure is unknown, we can make reliable models based on results from fold recognition servers such as 3DPSSM, http://www.sbg.bio.ic.ac.uk/~3dpssm/.

To this point, we have only talked about linear epitopes, and occur next to one another in the sequence of a protein. A folded protein in 3D may have epitopes formed from several areas of the sequence. The ConSurf [122] method has been used to detect patches of residues common to many allergens on the surface of allergen structures for Arat 8, Act c 1, Bet v 1, and Ves v 5 [123]. The findings have not yet been tested experimentally. A combined method that uses sequence similarity and comparison of 3D models was used to identify potentially cross-reactive peanut-lupine proteins [124] and to search for potential new latex allergens [125]. Our PCPMer program contains methodology to map conserved residues on the surface of proteins, to detect such common areas. These “stereophysicochemical variability plots” are useful for distinguishing functional areas of viruses [126], and can also be used to identify regions that might constitute IgE epitopes. Alternatively, we are also in the process of developing methods to map peptides that bind IgE to surface areas of allergens of known structure.

Conclusions and Future Developments

As discussed, similarities in sequence and structure of allergenic proteins can account for cross-reactivities between allergen sources [5,15,29,127–130] that may complicate the management of severely allergic patients [76,128,131].

Proteins are classified as allergens based on their ability to trigger responses in patients. Allergens may just be more potent forms of other proteins with similar surface areas that may have been the true sensitizing antigens during development of the disease. Thus it is not clear how similar proteins must be to known triggers in order to represent significant risk for cross-reactivity. The problem is made more difficult by the fact that some potent allergens can be rendered non-allergenic by selected point mutations and highly similar proteins, such as the glycinins of soybean and peanut (62% identity), provoke quite different responses [132].

We have outlined here computational methodology to identify cross-reacting proteins at the molecular level, using the databases of allergenic proteins and their structures. Recent identification of the sequence and structure of allergenic proteins from pollen and foods has revealed how similarities which might offer a structural explanation for their allergenicity and cross-reactivity [7,18,23,29,31,127,133]. Some of the recently developed search software tools, such as those implemented in SDAP can help clinicians and patients to find structural and functional relations among known allergens and to identify potentially cross-reacting antigens.

It is clear that available methods cannot, with 100% accuracy, discriminate between closely related proteins according to their allergenicity. Instead, they provide an indication that certain proteins may be cross-reactive. These predictions can certainly be useful in developing dietary guidelines for individual patients, and in designing specific immunotherapy.

Acknowledgements

Grant support: This work was supported by grants from the Food and Drug Administration (FD-U-002249), the Texas Higher Education Coordinating Board (ATP grant 004952-0036-2003) and the National Institute of Health (R01 AI 064913).

References

1. Sampson HA. Food allergy. Part 2: Diagnosis and management. J Allergy Clin Immunol 1999;103(6):981–989. [PubMed: 10359874]
2. Sampson HA. Food allergy. Part 1: Immunopathogenesis and clinical disorders. J Allergy Clin Immunol 1999;103(5):717–728. [PubMed: 10329801]
3. Sampson H. Food allergy: When mucosal immunity goes wrong. J Allergy Clin Immunol 2005;115:139–141. [PubMed: 15637560]
4. Vanek-Krebitz M, Hoffmannsommergruber K, Machado MLD, et al. Cloning and Sequencing of Mal-D-1, the Major Allergen from Apple (Malus-Domestica), and Its Immunological Relationship to Bet-V-1, the Major Birch Pollen Allergen. Biochem Biophys Res Commun 1995;214(2):538–551. [PubMed: 7677763]
5. Scheurer S, Son DY, Boehm M, et al. Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen. Mol Immunol 1999;36(3):155–167. [PubMed: 10403481]
6. Rabjohn P, Helm EM, Stanley JS, et al. Molecular cloning and epitope analysis of the peanut allergen Ara h 3. J Clin Invest 1999;103(4):535–542. [PubMed: 10021462]
7. Breiteneder H, Ebner C. Molecular and biochemical classification of plant-derived food allergens. J Allergy Clin Immunol 2000;106(1):27–36. [PubMed: 10887301]
8. Jenkins J, Griffiths-Jones S, Shewry P, et al. Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: an *in silico* analysis. J Allergy Clin Immunol 2005;115:163–170. [PubMed: 15637564]

9. Ferreira F, Hawranek T, Gruber P, et al. Allergic cross-reactivity: from gene to the clinic. Allergy 2004;59(3):243–267. [PubMed: 14982506]
10. Mari A. Multiple pollen sensitization: A molecular approach to the diagnosis. Int Arch Allergy Immunol 2001;125(1):57–65. [PubMed: 11385289]
11. Burks AW, Shin D, Cockrell G, et al. Mapping and mutational analysis of the IgE-binding epitopes on Ara h 1, a legume vicilin protein and a major allergen in peanut hypersensitivity. Eur J Biochem 1997;245(2):334–339. [PubMed: 9151961]
12. Shin DS, Compadre CM, Maleki SJ, et al. Biochemical and structural analysis of the IgE binding sites on Ara h 1, an abundant and highly allergenic peanut protein. J Biol Chem 1998;273(22):13753–13759. [PubMed: 9593717]
13. Schein CH, Ivanciuc O, Braun W. Common physical-chemical properties correlate with similar structure of the IgE epitopes of peanut allergens. J Agric Food Chem 2005;53(22):8752–8759. [PubMed: 16248581]
14. de Leon MP, Glaspole IN, Drew AC, et al. Immunological analysis of allergenic cross-reactivity between peanut and tree nuts. Clin Exp Allergy 2003;33(9):1273–1280. [PubMed: 12956750]
15. Eigenmann PA, Burks AW, Bannon GA, et al. Identification of unique peanut and soy allergens in sera adsorbed with cross-reacting antibodies. J Allergy Clin Immunol 1996;98(5):969–978. [PubMed: 8939161]
16. Lopez-Torrejon G, Salcedo G, Martin-Esteban M, et al. Len c 1, a major allergen and vicilin from lentil seeds: protein isolation and cDNA cloning. J Allergy Clin Immunol 2003;112(6):1208–1215. [PubMed: 14657885]
17. Wensing M, Knulst AC, Piersma S, et al. Patients with anaphylaxis to pea can have peanut allergy caused by cross-reactive IgE to vicilin (Ara h 1). J Allergy Clin Immunol 2003;111(2):420–424. [PubMed: 12589366]
18. Midoro-Horiuti T, Brooks EG, Goldblum RM. Pathogenesis-related proteins of plants as allergens. Ann Allergy Asthma Immunol 2001;87(4):261–271. [PubMed: 11686417]
19. Asensio T, Crespo JF, Sanchez-Monge R, et al. Novel plant pathogenesis-related protein family involved in food allergy. J Allergy Clin Immunol 2004;114(4):896–899. [PubMed: 15480331]
20. Hoffmann-Sommergruber K. Pathogenesis-related (PR)-proteins identified as allergens. Biochem Soc Trans 2002;30:930–935. [PubMed: 12440949]
21. Elbez M, Kevers C, Hamdi S, et al. The plant pathogenesis-related PR-10 proteins. Acta Bot Gall 2002;149(4):415–444.
22. Midoro-Horiuti T, Goldblum R, Kurosky A, et al. Isolation and characterization of the mountain cedar (Juniperus ashei) pollen major allergen, Jun a 1. J Allergy Clin Immunol 1999;104:608–612. [PubMed: 10482835]
23. Soman KV, Midoro-Horiuti T, Ferreon JC, et al. Homology modeling and characterization of IgE epitopes of mountain cedar allergen Jun a 3. Biophys J 2000;79(3):1601–1609. [PubMed: 10969020]
24. Ivanciuc O, Mathura V, Midoro-Horiuti T, et al. Detecting potential IgE-reactive sites on food proteins using a sequence and structure database, SDAP-Food. J Agric Food Chem 2003;51(16):4830–4837. [PubMed: 14705920]
25. Mari A. Importance of databases in experimental and clinical allergology. Int Arch Allergy Immunol 2005;138(1):88–96. [PubMed: 16127277]
26. Midoro-Horiuti T, Goldblum RN, Kurosky A, et al. Molecular cloning of the mountain cedar (Juniperus ashei) pollen major allergen, Jun a 1. J Allergy Clin Immunol 1999;104(3):613–617. [PubMed: 10482836]
27. Midoro-Horiuti T, Mathura VS, Schein CH, et al. Major linear IgE epitopes of mountain cedar pollen allergen Jun a 1 map to the pectate lyase catalytic site. Mol Immunol 2003;40(8):555–562. [PubMed: 14563374]
28. Midoro-Horiuti T, Schein CH, Mathura V, et al. Structural basis for epitope sharing between group 1 allergens of cedar pollen. Mol Immunol 2006;43(6):509–518. [PubMed: 15975657]
29. Fedorov AA, Ball T, Mahoney NM, et al. The molecular basis for allergen cross-reactivity: crystal structure and IgE epitope mapping of birch pollen profilin. Structure 1997;5(1):33–45. [PubMed: 9016715]

30. Ferreira F, Ebner C, Kramer B, et al. Modulation of IgE reactivity of allergens by site-directed mutagenesis: potential use of hypoallergenic variants for immunotherapy. FASEB J 1998;12(2):231–242. [PubMed: 9472988]

31. Spangfort MD, Mirza O, Holm J, et al. The structure of major birch pollen allergens: Epitopes, reactivity and cross-reactivity. Allergy 1999;50:23–26. [PubMed: 10466032]

32. Petersen A, Schramm G, Schlaak M, et al. Post-translational modifications influence IgE reactivity. Clin Exp Allergy 1998;28(3):315–321. [PubMed: 9543081]

33. Schramm G, Bufe A, Petersen A, et al. Mapping of IgE-binding epitopes on the recombinant major group I allergen of velvet grass pollen, rHol 1 1. J Allergy Clin Immunol 1997;99(6):781–787. [PubMed: 9215246]

34. Lalla C, Tamborini E, Longhi R, et al. Human recombinant antibody fragments specific for a rye-grass pollen allergen: Characterization and potential applications. Mol Immunol 1996;33:1049–1058. [PubMed: 9010244]

35. Flicker S, Vrtala S, Steinberger P, et al. A human monoclonal IgE antibody defines a highly allergenic fragment of the major timothy grass pollen allergen, Phl p 5: molecular, immunological, and structural characterization of the epitope-containing domain. J Immunol 2000;165:3849–3859. [PubMed: 11034391]

36. Wal JM. Bovine milk allergenicity. Ann Allergy Asthma Immunol 2004;93(5):S2–S11. [PubMed: 15562868]

37. Natale M, Bisson C, Monti G, et al. Cow's milk allergens identification by two-dimensional immunoblotting and mass spectrometry. Mol Nutr Food Res 2004;48(5):363–369. [PubMed: 15672476]

38. Pourpak Z, Mostafaie A, Hasan Z, et al. A laboratory method for purification of major cow's milk allergens. J Immunoass Immunoch 2004;25(4):385–397.

39. Elsayed S, Hill DJ, Do TV. Evaluation of the allergenicity and antigenicity of bovine-milk alpha s1-casein using extensively purified synthetic peptides. Scand J Immunol 2004;60(5):486–493. [PubMed: 15541041]

40. Cocco RR, Jarvinen KM, Sampson HA, et al. Mutational analysis of major, sequential IgE-binding epitopes in alpha(s1)-casein, a major cow's milk allergen. J Allergy Clin Immunol 2003;112(2):433–437. [PubMed: 12897753]

41. Ehn BM, Ekstrand B, Bengtsson U, et al. Modification of IgE binding during heat processing of the cow's milk allergen beta-lactoglobulin. J Agric Food Chem 2004;52(5):1398–1403. [PubMed: 14995152]

42. Jarvinen KM, Chatchatee P, Bardina L, et al. IgE and IgG binding epitopes on alpha-lactalbumin and beta-lactoglobulin in cow's milk allergy. Int Arch Allergy Immunol 2001;126(2):111–118. [PubMed: 11729348]

43. Adel-Patient K, Creminon C, Boquet D, et al. Genetic immunisation with bovine beta-lactoglobulin cDNA induces a preventive and persistent inhibition of specific anti-BLG IgE response in mice. Int Arch Allergy Immunol 2001;126(1):59–67. [PubMed: 11641607]

44. Mine Y, Rupa P. Immunological and biochemical properties of egg allergens. Worlds Poult Sci J 2004;60(3):321–330.

45. Mizumachi K, Kurisaki J. Localization of T cell epitope regions of chicken ovomucoid recognized by mice. Biosci Biotechnol Biochem 2003;67(4):712–719. [PubMed: 12784609]

46. Mine Y, Sasaki E, Zhang JW. Reduction of antigenicity and allergenicity of genetically modified egg white allergen, ovomucoid third domain. Biochem Biophys Res Commun 2003;302(1):133–137. [PubMed: 12593859]

47. Mine Y, Zhang JW. Identification and fine mapping of IgG and IgE epitopes in ovomucoid. Biochem Biophys Res Commun 2002;292(4):1070–1074. [PubMed: 11944924]

48. Fremont S, Kanny G, Nicolas JP, et al. Prevalence of lysozyme sensitization in an egg-allergic population. Allergy 1997;52(2):224–228. [PubMed: 9105530]

49. Ayuso R, Lehrer SB, Reese G. Identification of continuous, allergenic regions of the major shrimp allergen Pen a 1 (tropomyosin). Int Arch Allergy Immunol 2002;127(1):27–37. [PubMed: 11893851]

50. Reese G, Ayuso R, Leong-Kee SM, et al. Epitope mapping and mutational substitution analysis of the major shrimp allergen Pen a 1 (tropomyosin). J Allergy Clin Immunol 2002;109(1):S307–S307.

51. Samson KTR, Chen FH, Miura K, et al. IgE binding to raw and boiled shrimp proteins in atopic and nonatopic patients with adverse reactions to shrimp. Int Arch Allergy Immunol 2004;133(3):225–232. [PubMed: 14976390]

52. Van Do T, Hordvik I, Endresen C, et al. Characterization of parvalbumin, the major allergen in Alaska pollack, and comparison with codfish Allergen M. Mol Immunol 2005;42(3):345–353. [PubMed: 15589323]

53. Swoboda I, Bugajska-Schretter A, Verdino P, et al. Recombinant carp parvalbumin, the major cross-reactive fish allergen: a tool for diagnosis and therapy of fish allergy. Allergy 2002;57:80–80.

54. Swoboda I, Bugajska-Schretter A, Valenta R, et al. Recombinant fish parvalbumins: Candidates for diagnosis and treatment of fish allergy. Allergy 2002;57:94–96. [PubMed: 12144564]

55. Moreno FJ, Maldonado BM, Wellner N, et al. Thermostability and in vitro digestibility of a purified major allergen 2S albumin (Ses i 1) from white sesame seeds (Sesamum indicum L.). Biochim Biophys Acta 2005;1752(2):142–153. [PubMed: 16140598]

56. Robotham JM, Wang F, Seamon V, et al. Ana o 3, an important cashew nut (Anacardium occidentale L.) allergen of the 2S albumin family. J Allergy Clin Immunol 2005;115(6):1284–1290. [PubMed: 15940148]

57. Palomares O, Cuesta-Herranz J, Rodriiguez R, et al. A recombinant precursor of the mustard allergen Sin a 1 retains the biochemical and immunological features of the heterodimeric native protein. Int Arch Allergy Immunol 2005;137(1):18–26. [PubMed: 15785078]

58. Moreno FJ, Mellon FA, Wickham MSJ, et al. Stability of the major allergen Brazil nut 2S albumin (Ber e 1) to physiologically relevant in vitro gastrointestinal digestion. FEBS J 2005;272(2):341–352. [PubMed: 15654873]

59. Beardslee TA, Zeece MG, Sarath G, et al. Soybean glycinin G1 acidic chain shares IgE epitopes with peanut allergen Ara h 3. Int Arch Allergy Immunol 2000;123(4):299–307. [PubMed: 11146387]

60. Helm RM, Cockrell G, Connaughton C, et al. A soybean G2 glycinin allergen - 1. Identification and characterization. Int Arch Allergy Immunol 2000;123(3):205–212. [PubMed: 11112856]

61. Rabjohn P, Burks AW, Sampson HA, et al. Mutational analysis of the IgE-binding epitopes of the peanut allergen, Ara h 3: a member of the glycinin family of seed-storage proteins. J Allergy Clin Immunol 1999;103(1):S101–S101.

62. Bolhaar S, van Ree R, Bruijnzeel-Koomen C, et al. Allergy to jackfruit: a novel example of Bet v 1-related food allergy. Allergy 2004;59(11):1187–1192. [PubMed: 15461600]

63. Bolhaar S, van Ree R, Ma Y, et al. Severe allergy to sharon fruit caused by birch pollen. Int Arch Allergy Immunol 2005;136(1):45–52. [PubMed: 15591813]

64. Bolhaar S, Zuidmeer L, Ma Y, et al. 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 2005;35(12):1638–1644. [PubMed: 16393331]

65. Bucher X, Pichler WJ, Dahinden CA, et al. Effect of tree pollen specific, subcutaneous immunotherapy on the oral allergy syndrome to apple and hazelnut. Allergy 2004;59(12):1272–1276. [PubMed: 15507095]

66. Mari A, Ballmer-Weber BK, Vieths S. The oral allergy syndrome: improved diagnostic and treatment methods. Curr Opin Allergy Clin Immunol 2005;5(3):267–273. [PubMed: 15864087]

67. Bolhaar S, Tiemessen MM, Zuidmeer L, et al. Efficacy of birch-pollen immunotherapy on cross-reactive food allergy confirmed by skin tests and double-blind food challenges. Clin Exp Allergy 2004;34(5):761–769. [PubMed: 15144469]

68. Gendel SM. Bioinformatics and food allergens. J AOAC Int 2004;87:1417–1422. [PubMed: 15675454]

69. Glaspole IN, de Leon MP, Rolland JM, et al. Characterization of the T-cell epitopes of a major peanut allergen, Ara h 2. Allergy 2005;60:35–40. [PubMed: 15575928]

70. Breiteneder H, Mills ENC. Molecular properties of food allergens. J Allergy Clin Immunol 2005;115:14–23. [PubMed: 15637541]

71. Brusic V, Millot M, Petrovsky N, et al. Allergen databases. Allergy 2003;58(11):1093–1100. [PubMed: 14616118]

72. Ivanciuc O, Schein CH, Braun W. SDAP: Database and computational tools for allergenic proteins. Nucleic Acids Res 2003;31(1):359–362. [PubMed: 12520022]

73. Ivanciuc O, Schein CH, Braun W. Data mining of sequences and 3D structures of allergenic proteins. Bioinformatics 2002;18(10):1358–1364. [PubMed: 12376380]

74. Schein, CH.; Ivanciuc, O.; Braun, W. Structural Database of Allergenic Proteins (SDAP). In: Maleki, SJ.; Burks, AW.; Helm, RM., editors. Food Allergy. Washington, D.C: ASM Press; 2006. p. 257-283.

75. Stadler MB, Stadler BM. Allergenicity prediction by protein sequence 2003;17(6):1141–1143.

76. WHO. Report of a joint FAO/WHO expert consultation. Geneva: World Health Organization; 2001. Evaluation of allergenicity of genetically modified foods.

77. WHO. Codex Ad Hoc Intergovernmental Task Force on Foods Derived from Biotechnology. Yokohama: World Health Organization; 2003. Joint FAO/WHO Food Standards Programme. http://www.codexalimentarius.net/

78. Pearson W. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 1990;183:63–98. [PubMed: 2156132]

79. Conte LL, Ailey B, Hubbard TJP, et al. SCOP: A Structural Classification of Proteins Database. Nucleic Acids Res 2000;28(1):257–259. [PubMed: 10592240]

80. Gilbert D, Westhead D, Nagano N, et al. Motif-based searching in TOPS protein topology databases. Bioinformatics 1999;15(4):317–326. [PubMed: 10320400]

81. Pearl FMG, Martin N, Bray JE, et al. A rapid classification protocol for the CATH Domain Database to support structural genomics. Nucleic Acids Res 2001;29(1):223–227. [PubMed: 11125098]

82. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11(9):739–747. [PubMed: 9796821]

83. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–602. [PubMed: 8662544]

84. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6(3):377–385. [PubMed: 8804824]

85. Rabjohn P, West C, Connaughton C, et al. Modification of peanut allergen Ara h 3: effects on IgE binding and T cell stimulation. Int Arch Allergy Immunol 2002;128:15–23. [PubMed: 12037397]

86. Bannon G, Cockrell G, Connaughton C, West CM, Helm R, Stanley JS, King N, Rabjohn P, Sampson HA, Burks AW. Engineering, characterization and in vitro efficacy of the major peanut allergens for use in immunotherapy. Int Arch Allergy Immunol 2001;124:70–72. [PubMed: 11306930]

87. Li XM, Srivastava K, Huleatt JW, et al. Engineered recombinant peanut protein and heat-killed Listeria monocytogenes coadministration protects against peanut-induced anaphylaxis in a murine model. J Immunol 2003;170:3289–3295. [PubMed: 12626588]

88. vanRee R, Antonicelli L, Akkerdaas JH, et al. Possible induction of food allergy during mite immunotherapy. Allergy 1996;51(2):108–113. [PubMed: 8738516]

89. Breiteneder H, Mills ENC. Plant food allergens - structural and functional aspects of allergenicity. Biotechnol Adv 2005;23(6):395–399. [PubMed: 15985358]

90. Schein CH. The shape of the messenger: using protein structural information to design novel cytokine-based therapeutics. Curr Pharm Des 2002;8(24):213–230.

91. Weber RW. Patterns of pollen cross-reactivity. Curr Rev Allergy Clin Immunol 2003;112:229–239.

92. Breiteneder H, Radauer C. A classification of plant food allergens. J Allergy Clin Immunol 2004;113(5):821–830. [PubMed: 15131562]

93. Jenkins JA, Griffiths-Jones S, Shewry PR, et al. Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: An in silico analysis. J Allergy Clin Immunol 2005;115(1):163–170. [PubMed: 15637564]

94. Hartl A, Kiesslich J, Weiss R, et al. Isoforms of the major allergen of birch pollen induce different immune responses after genetic immunization. Int Arch Allergy Immunol 1999;120(1):17–29. [PubMed: 10529585]

95. Ferreira F, Hirthenlehner K, Briza P, et al. Isoforms of atopic allergens with reduced allergenicity but conserved T cell antigenicity: Possible use for specific immunotherapy. Int Arch Allergy Immunol 1997;113(1–3):125–127. [PubMed: 9130500]

96. Ferreira F, Hirtenlehner K, Jilek A, et al. Dissection of immunoglobulin E and T lymphocyte reactivity of isoforms of the major birch pollen allergen Bet v 1: Potential use of hypoallergenic isoforms for immunotherapy. J Exp Med 1996;183(2):599–609. [PubMed: 8627171]

97. Thomas K, Bannon G, Hefle S, et al. In silico methods for evaluating human allergenicity to novel proteins: International Bioinformatics Workshop Meeting Report, 23–24 February 2005. Toxicol Sci 2005;88(2):307–310. [PubMed: 16107555]

98. Aalberse RC, Stadler BM. In silico predictability of allergenicity: From amino acid sequence via 3-D structure to allergenicity. Mol Nutr Food Res 2006;50(7):625–627. [PubMed: 16764015]

99. Riaz T, Hor HL, Krishnan A, et al. WebAllergen: a web server for predicting allergenic proteins. Bioinformatics 2005;21(10):2570–2571. [PubMed: 15746289]

100. Li KB, Issac P, Krishnan A. Predicting allergenic proteins using wavelet transform. Bioinformatics 2004;20(16):2572–2578. [PubMed: 15117757]

101. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res 2006;34:W202–W209. [PubMed: 16844994]

102. Fiers M, Kleter GA, Nijland H, et al. Allermatch (TM), a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. BMC Bioinformatics 2004:5. [PubMed: 14718068]

103. Brusic V, Petrovsky N. Bioinformatics for characterisation of allergens, allergenicity and allergic crossreactivity. Trends Immunol 2003;24(5):225–228. [PubMed: 12738409]

104. Hileman RE, Silvanovich A, Goodman RE, et al. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. Int Arch Allergy Immunol 2002;128(4):280–291. [PubMed: 12218366]

105. Bindsley-Jensen C, Sten E, Earl LK, et al. Assessment of the potential allergenicity of ice structuring protein type III HPLC 12 using the FAO/WHO 2001 decision tree for novel foods. Food Chem Toxicol 2003;41(1):81–87. [PubMed: 12453731]

106. Baderschneider B, Crevel RWR, Earl LK, et al. Sequence analysis and resistance to pepsin hydrolysis as part of an assessment of the potential allergenicity of ice structuring protein type III HPLC 12. Food Chem Toxicol 2002;40(7):965–978. [PubMed: 12065219]

107. Singh AK, Mehta AK, Sridhara S, et al. Allergenicity assessment of transgenic mustard (Brassica juncea) expressing bacterial codA gene. Allergy 2006;61(4):491–497. [PubMed: 16512812]

108. Soeria-Atmadja D, Zorzet A, Gustafsson MG, et al. Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. Int Archiv Allergy Immunol 2004;133(2):101–112.

109. Soeria-Atmadja D, Wallman M, Björklund ÅK, et al. External cross-validation for unbiased evaluation of protein family detectors: Application to allergens. Proteins 2005;61(4):918–925. [PubMed: 16231294]

110. Zorzet A, Gustafsson M, Hammerling U. Prediction of food protein allergenicity: A bioinformatic learning systems approach. In Silico Biol 2004;2:525–534. [PubMed: 12611632]

111. Mills EN, Jenkins J, Marigheto N, et al. Allergens of the cupin superfamily. Biochem Soc Trans 2002;30(6):925–929. [PubMed: 12440948]

112. Venkatarajan MS, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. J Mol Model 2001;7(12):445–453.

113. Mathura VS, Schein CH, Braun W. Identifying property based sequence motifs in protein families and superfamilies: Application to DNase I related endonucleases. Bioinformatics 2003;19(11): 1381–1390. [PubMed: 12874050]

114. Kleter GA, Peijnenburg A. Presence of potential allergy-related linear epitopes in novel proteins from conventional crops and the implication for the safety assessment of these crops with respect to the current testing of genetically modified crops. Plant Biotechnol J 2003;1(5):371–380. [PubMed: 17166136]

115. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 2006;65:40–48. [PubMed: 16894596]

116. Silvanovich A, Nemeth MA, Song P, et al. The value of short amino acid sequence matches for prediction of protein allergenicity. Toxicol Sci 2006;90(1):252–258. [PubMed: 16338955]

117. Björklund ÅK, Soeria-Atmadja D, Zorzet A, et al. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. Bioinformatics 2005;21(1):39–50. [PubMed: 15319257]

118. Cui J, Han LY, Li H, et al. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. Mol Immunol 2007;44(4):514–520. [PubMed: 16563508]

119. Shreffler WG, Beyer K, Chu TH, et al. Microarray immunoassay: association of clinical history, in vitro IgE function, and heterogeneity of allergenic peanut epitopes. J Allergy Clin Immunol 2004;113(4):776–782. [PubMed: 15100687]

120. Ivanciuc O, Oezguen N, Mathura V, et al. Using property based sequence motifs and 3D modeling to determine structure and functional regions in CASP5 targets. Curr Med Chem 2004;11(5):583–593. [PubMed: 15032606]

121. Teuber SS, Beyer K. Peanut, tree nut and seed allergies. Curr Opin Allergy Clin Immunol 2004;4 (3):201–203. [PubMed: 15126942]

122. Glaser F, Pupko T, Paz I, et al. ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. Bioinformatics 2003;19(1):163–164. [PubMed: 12499312]

123. Furmonaviciene R, Sutton BJ, Glaser F, et al. An attempt to define allergen-specific molecular surface features: a bioinformatic approach. Bioinformatics 2005;21(23):4201–4204. [PubMed: 16204345]

124. Guarneri F, Guarneri C, Benvenga S. Identification of potentially cross-reactive peanut-lupine proteins by computer-assisted search for amino acid sequence homology. Int Arch Allergy Immunol 2005;138(4):273–277. [PubMed: 16220003]

125. Guarneri F, Guarneri C, Guarneri B, et al. In silico identification of potential new latex allergens. Clin Exp Allergy 2006;36(7):916–919. [PubMed: 16839407]

126. Schein CH, Zhou B, Braun W. Stereophysicochemical variability plots highlight conserved antigenic areas in Flaviviruses. Virology J 2005;2:40. [PubMed: 15845145]

127. Ipsen H, Lowenstein H. Basic features of crossreactivity in tree and grass pollen allergy. Clin Rev Allergy Immunol 1997;15(4):389–396. [PubMed: 9484576]

128. Lehrer, SI.; Ayuso, R.; Reese, G. Current understanding of food allergens. In: Fu, TJ.; Gendel, SM., editors. Genetically engineered foods: assessing potential allergenicity. 964. New York, NY: The New York Academy of Sciences; 2002. p. 69-85.

129. Leung PSC, Chow WK, Duffey S, et al. IgE reactivity against a cross-reactive allergen in crustacea and mollusca: Evidence fev tropomyosin as the common allergen. J Allergy Clin Immunol 1996;98 (5):954–961. [PubMed: 8939159]

130. Sparholt SH, Larsen JN, Ipsen H, et al. Crossreactivity and T-cell epitope specificity of Bet v 1-specific T cells suggest the involvement of multiple isoallergens in sensitization to birch pollen. Clin Exp Allergy 1997;27(8):932–941. [PubMed: 9291292]

131. Bousquet J, Knani J, Hejjaoui A, et al. Heterogeneity of atopy. 1. Clinical and immunological characteristics of patients allergic to cypress pollen. Allergy 1993;48(3):183–188. [PubMed: 8506986]

132. Beardslee TA, Zeece MG, Sarath G, et al. Soybean glycinin G1 acidic chain shares IgE epitopes with peanut allergen Ara h 3. Int Arch Allergy Immunol 2000;123:299–307. [PubMed: 11146387]

133. Aalberse RC. Structural biology of allergens. J Allergy Clin Immunol 2000;106(2):228–238. [PubMed: 10932064]

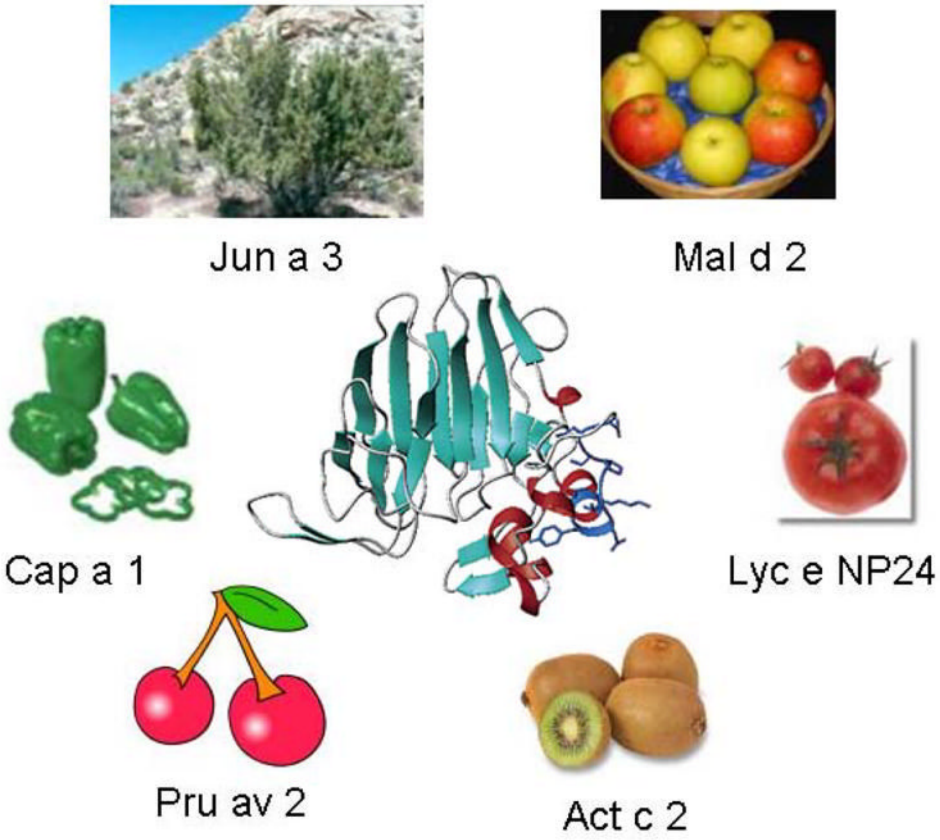

Figure 1.

Different products can contain similar allergenic proteins. In this case, an allergen, Jun a 3, originally isolated from the pollen of mountain cedar was found to be a member of the PR5 family of proteins, and was modeled based on its similarity to a protein of known structure, thaumatin [23]. Subsequently, similar allergenic proteins were isolated from many food plants [24], including bell pepper (Cap a 1), cherry (Pru av 2), kiwi (Act c 2), tomato (Lyc e NP24), and apple (Mal d 2).

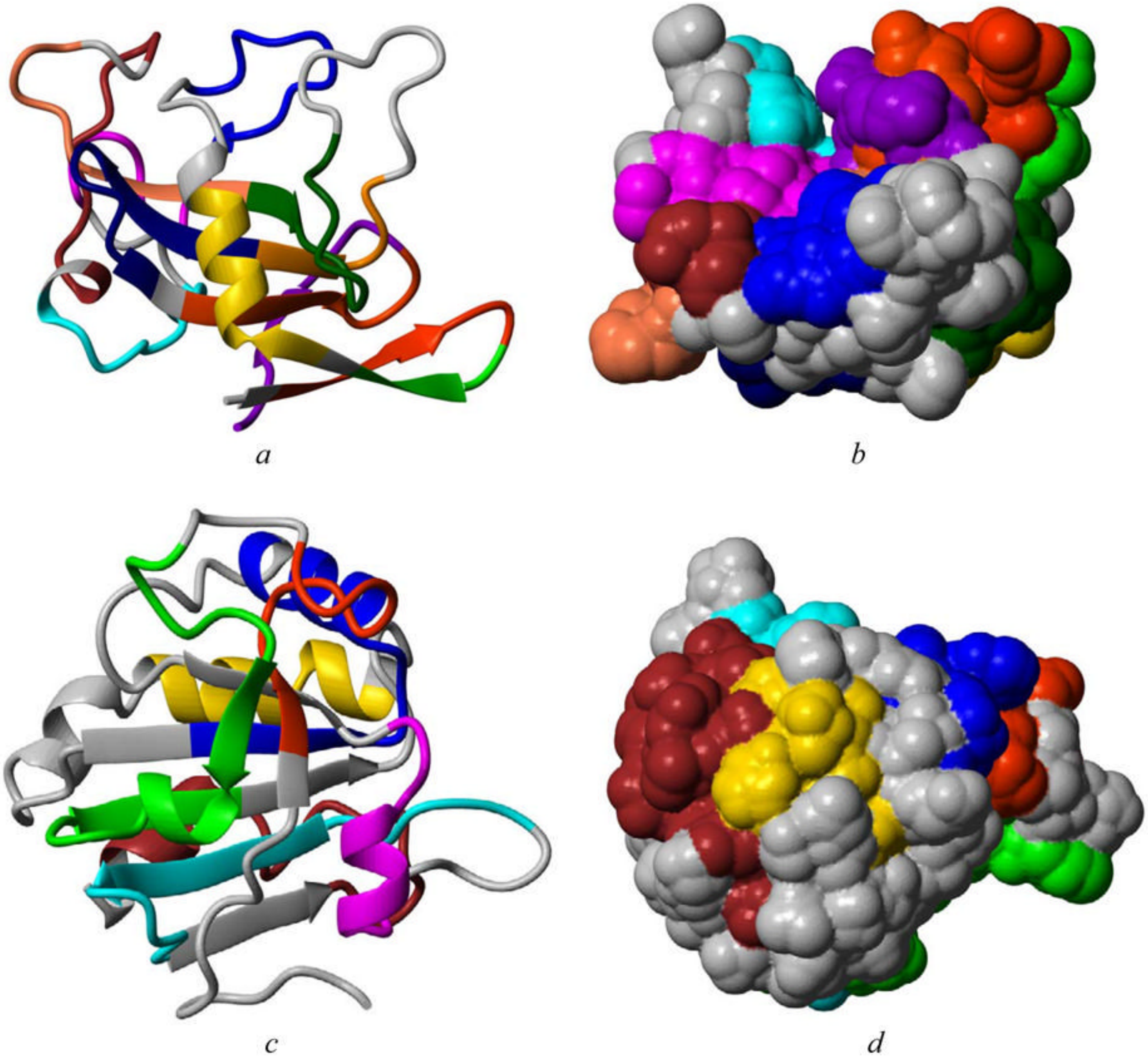

Figure 2.
IgE epitopes mapped on the experimental structure (PDB 1AQZ) of Asp f 1 (*a* and *b*) and out MPACK model of Asp f 3 (*c* and *d*), two allergens from the fungus causing aspergillosis.

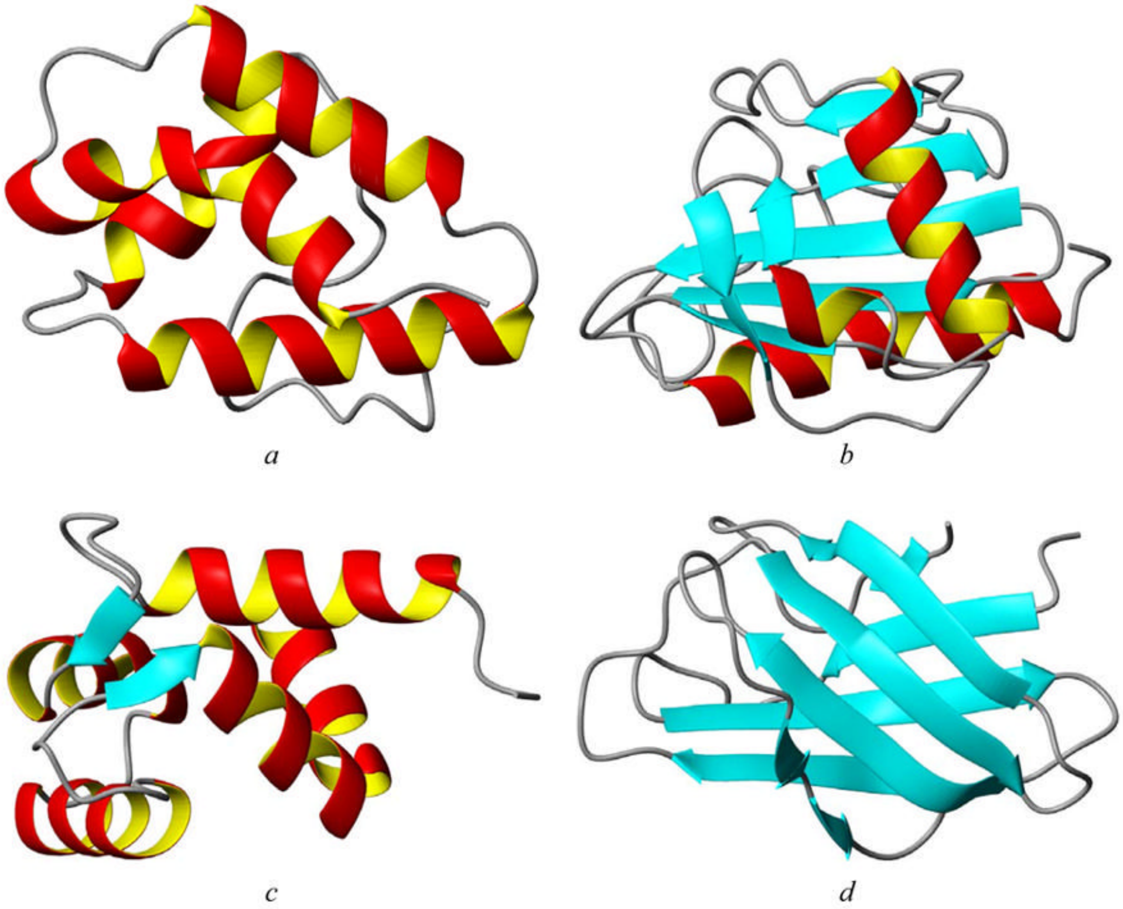

Figure 3.

PDB structures for allergens from the four most abundant Pfam families: (a) Pru p 3 (PF00234, protease inhibitor/seed storage/LTP family); (b) Hev b 8, (PF00235, profilin); (c) Bet v 4, (PF00036, EF hand); (d) Phl p 2, (PF01357, pollen allergen).

Table 1

Websites with information about allergens, allergen databases, and allergenicity prediction servers

Websitename	URL	Information available
Nomenclature and general information databases:		
All Allergy	http://allallergy.net/	a portal to allergy information, useful for the general public
IUIS (International Union of Immunological Societies)	http://www.allergen.org	lists official names, grouped by source, and Genbank accession numbers of allergens
Allergome	http://www.allergome.org	lists the official names of allergens, and links to PubMed & sequence databases
CSL (Central Science Laboratory, UK)	http://www.csl.gov.uk/allergen/index.htm	lists official names of allergens with sequence links to Genbank
National Center for Food Safety and Technology	http://www.iit.edu/~sgendel/fa.htm	lists official names of food allergens with links to Genbank
Protall	http://www.ifrn.bbsrc.ac.uk/protall/	allergen names, plus links to detailed biochemical, structural, and clinical data
InformAll	http://foodallergens.ifr.ac.uk/	biochemical information, mainly for food allergens, epitopes, sequences, links to literature
Cross-referenced databases with tools to compare sequences:		
FARRP	http://allergenonline.com/asp/public/login.asp	lists official names of allergens, sequence links to Genbank, and a FASTA search for related sequences
ADFS – Allergen Database for Food Safety	http://allergen.nihs.go.jp/ADFS/	allergen sequences, implements the WHO allergenicity rules using FASTA
Cross-referenced databases with tools for prediction of allergenicity:		
SDAP (Structural database of Allergenic proteins) and SDAP-FOOD	http://fermi.utmb.edu/SDAP	allergens sequences, on-site and cross-referenced by source and protein type, with links to all major sequence and structural databases, IgE epitopes collection, tools for sequence and epitope comparison, on site information about experimental structures of allergens, and high-quality protein models
ALLERDB	http://sdmc.i2r.a-star.edu.sg/Templar/DB/Allergen/	lists official names of allergens, a BLAST search, and implements the WHO allergenicity rules
Web servers for allergenicity prediction		
WebAllergen	http://weballergen.bii.a-star.edu.sg/	predicts the potential allergenicity of proteins using motifs found by a wavelet algorithm
Allermatch	www.allermatch.org/	implements the WHO allergenicity rules using FASTA
AlgPred	http://www.imtech.res.in/raghava/algpred/	Predicts allergenicity with MEME/MAST motifs

Table 2

Output of an automatic FASTA search in SDAP starting from the file for the allergen Jun a 3, an allergen isolated from cedar pollen (see Figure 1). This PR5 protein is related (i.e., the sequence match has an E-score <0.01) to 7 other allergenic proteins in SDAP.

No	Allergen	Sequence	Source	Sequence Length	bit score	E score
1	Jun a 3	P81295	cedar pollen	225	311.0	1.0e-86
2	Cup a 3	CAC05258	cypress	199	272.9	2.7e-75
3	Cap a 1w	CAC34055	bell pepper	246	167.5	1.7e-43
4	Lyc e NP24	P12670	tomato	247	161.2	1.4e-41
5	Cap a 1	AAG34078	bll pepper	180	136.2	3.4e-34
6	Mal d 2	CAC10270	apple	246	77.0	3.0e-16
7	Pru av 2	P50694	cherry	245	75.1	1.2e-15
8	Act c 2	P81370	kiwi	29	36.0	7.9e-05

Table 3

Results of two automatic FASTA searches in SDAP that illustrate how to use this site to correctly name a new allergen. While the cedar pollen allergens Jun a 2 and Cry j 2 are very close in sequence (*i.e*, have a very low expectation (E) value; top table), they are not related to the allergen originally called Jun o 2. This protein, officially named Jun o 4 by the IUIS, is similar to Bet v 4 and other allergens of that sequence family (bottom table).

No	**Allergen**	**Sequence Link in SwissProt/NCBI/PIR**	**Sequence Length**	**bit score**	**E score**
1	Jun a 2	CAC05582	507	794.0	0.0e+00
2	Cry j 2	P43212	514	579.2	9.5e-167
4	Phl p 13	CAB42886	394	198.8	2.4e-52

No	Allergen	Sequence Link in SwissProt/NCBI/PIR	Sequence Length	bit score	E score
1	Jun o 4	O64943	165	229.0	2.7e-62
2	Ole e 8	Q9M7R0	171	89.8	2.2e-20
4	Syr v 3	P58171	81	62.5	1.7e-12
5	Bra n 2	BAA09633	82	61.3	4.0e-12
7	Bra r 2	Q39406	83	61.3	4.0e-12
8	Aln g 4	O81701	85	61.3	4.1e-12
9	Bet v 4	Q39419	85	60.9	5.4e-12
11	Ole e 3	O81092	84	59.6	1.3e-11
12	Phl p 7	O82040	78	58.8	2.2e-11
15	Bra n 1	Q42470	79	52.8	1.4e-09
16	Bra r 1	Q42470	79	52.8	1.4e-09
17	Bet v 3	P43187	205	35.8	4.9e-04
18	Gad c 1	P02622	113	34.9	4.9e-04
19	Sal s 1	Q91482	109	33.6	1.1e-03
20	Sco j 1	P59747	109	31.3	5.5e-03

Table 4

The most abundant Pfam A allergen families from SDAP

Pfam code	Family name	No Allergens
PF00234	Protease inhibitor/seed storage/LTP family	34
PF00235	Profilin	27
PF00036	EF hand	23
PF01357	Pollen allergen	20
PF00188	SCP-like extracellular protein	19
PF00407	Pathogenesis-related protein Bet v I family	16
PF00190	Cupin	15
PF00261	Tropomyosin	15
PF00061	Lipocalin /cytosolic fatty-acid binding protein family	12
PF03330	Rare lipoprotein A (RlpA)-like double-psi beta-barrel	12
PF00042	Globin	9
PF00544	Pectate lyase	9
PF00112	Papain family cysteine protease	8
PF00428	60s Acidic ribosomal protein	8
PF00082	Subtilase family	7
PF00314	Thaumatin family	7
PF01190	Pollen proteins Ole e I family	7
PF01620	Ribonuclease (pollen allergen)	7

Table 5

Allergens from the four most populated with allergens Pfam families

PF00234: Protease inhibitor/seed storage/LTP family					
Amb a 6	Ana o 3	Ara h 2	Ara h 6	Ber e 1	Bra j 1
Bra n 1	Cor a 8	Fag e 8kD	Gly m 1	Hev b 12	Hor v 1
Hor v 21	Jug n 1	Jug r 1	Lyc e 3	Mal d 3	Ory s TAI
Par j 1	Par j 2	Pru ar 3	Pru av 3	Pru d 3	Pru p 3
Pyr c 3	Ric c 1	Ses i 1	Ses i 2	Sin a 1	Tri a gliadin
Tri a glutenin	Tri a TAI	Vit v 1	Zea m 14		
PF00235: Profilin					
Ana c 1	Api g 4	Ara h 5	Ara t 8	Bet v 2	Cap a 2
Che a 2	Cor a 2	Cuc m 2	Cyn d 12	Dau c 4	Gly m 3
Hel a 2	Hev b 8	Lit c 1	Lyc e 1	Mal d 4	Mer a 1
Mus xp 1	Ole e 2	Par j 3	Phl p 11	Phl p 12	Pru av 4
Pru p 4	Pyr c 4	Tri a profilin			
PF00036: EF hand					
Aln g 4	Bet v 3	Bet v 4	Bos d 3	Bra n 1	Bra n 2
Bra r 1	Che a 3	Cyn d 7	Cyp c 1	Gad c 1	Gad m 1
Hom s 4	Jun o 4	Ole e 3	Ole e 8	Phl p 7	Ran e 1
Ran e 2	Sal s 1	Sco j 1	Syr v 3	The c 1	
PF01357: Pollen allergen					
Ara t expansin	Cyn d 1	Cyn d 15	Cyn d 2	Dac g 2	Dac g 3
Gly m 2	Hol l 1	Lol p 1	Lol p 2	Lol p 3	Ory s 1
Pan s 1	Pha a 1	Phl p 1	Phl p 2	Poa p a	Tri a 3
Tri a ps93	Zea m 1				

Table 6

Test of three allergenicity prediction servers. Three database prediction methods are unable to discriminate the allergenic sequences from insects shellfish (bottom 4) from the non-allergenic tropomyosins from animals (top sequences). The former proteins do, however, contain motifs that are not present in the latter.

Tropomyosin	SwissProt	WebAllergen	Allermatch FAO/ WHO	AlgPred MEME	AlgPred SVM dipeptide
Human	TPM1_HUMAN	6 motifs: 13, 15, 25, 26, 30, 31	Y	N	Y
Bovine	Q5KR49_BOVIN	6 motifs: 13, 15, 25, 26, 30, 31	Y	N	Y
Pig	TPM1_PIG	6 motifs: 13, 14, 15, 26, 30, 31	Y	N	Y
Chicken	TPM1_CHICK	7 motifs: 13, 14, 15, 25, 26, 30, 31	Y	N	Y
Der p 10, house dust mite	TPM_DERPT	10 motifs: 13, 14, 15, 16, 25, 26, 29, 30, 31, 55	Y	N	Y
Per a 7 American cockroach	TPM_PERAM	10 motifs: 13, 14, 15, 25, 26, 29, 30, 31, 55, 7	Y	N	Y
Met e 1 sand shrimp	TPM_METEN	9 motifs: 13, 14, 15, 25, 26, 30, 31, 55, 7	Y	N	Y
Ani s 3 herring worm	TPM_ANISI	9 motifs: 13, 14, 15, 25, 26, 29, 30, 31, 55	Y	N	Y

Table 7

Sequences most closely related to epitope 3 of Jun a 3 from mountain cedar pollen by the PD search reveal other PR-5 proteins from fruits (top); that of epitope 4 reveals another known IgE epitope from latex (bottom; bold sequence). The column lists the order of sequences found.starting from each epitope sequence, the allergen name and source, the PD index (the lower the number, the more significant the sequence match), the Z score* (the higher the number, the more significant the match) and finally the matching sequence. Note that no epitopes have been published for any of the fruit proteins, and that sequence of the kiwi thaumatin like protein Act c 2 is only of the first 29 amino acids. The average PD value for the best scoring sequence in the 854 full length entries in SDAP was 12.15 (SD=1.29); the average PD value for all 190530 possible windows was 17.24 (SD=1.78).

No	Allergen	Source	*PD* Sequence Similarity Index	z (PD,all)	Start Residue	Matching region	End Residue
SDAP Search starting from Epitope 3 of Jun a 3:							
1	Jun a 3	Juniper	0.00	8.3364	146	ADINAVCPSELK	157
2	Cup a 3	Cedar	0.00	8.3364	120	ADINAVCPSELK	131
3	Pru av 2	Cherry	1.66	7.5620	157	ANVNAVCPSELQ	168
4	Mal d 2	Apple	6.63	5.2499	158	ANVNKVCPAPLQ	169
5	Lyc e NP24	Tomato	7.10	5.0309	148	ANINGECPRALK	159
6	Cap a 1	Pepper	7.31	4.9363	121	ANINGECPGSLR	132
SDAP Search starting from Epitope 4 of Jun a 3:							
1	Jun a 3	Juniper	0.00	9.0000	158	VDGGCNSACNVFKT	171
2	Cup a 3	Cedar	1.29	8.3504	132	VDGGCNSACNVLQT	145
3	Lyc e NP24 Tomato 8.20	4.8843	160			VPGGCNNPCTTFGG	173
4	Cap a 1	Pepper	8.20	4.8843	133	VPGGCNNPCTTFGG	146
5	Hev b 3	Latex	8.55	4.7090	41	**LKPGVDTIEN**VVKT	54

*Using the *PD* distribution, the score *z*(*PD*) for a given match is calculated as: $z(PD) = \frac{\left| PD_{\mathrm{min}} - PD_{\mathrm{ave}} \right|}{\mathbf{SD}(PD)}$