

RESEARCH ARTICLE

AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity

Richard E. Goodman¹, Motohiro Ebisawa², Fatima Ferreira³, Hugh A. Sampson⁴, Ronald van Ree⁵, Stefan Vieths⁶, Joseph L. Baumert¹, Barbara Bohle⁷, Sreedevi Lalithambika¹, John Wise¹ and Steve L. Taylor¹

¹ Food Allergy Research and Resource Program, Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln, NE, USA

² Department of Allergy, Sagamihara National Hospital, Sagamihara, Japan

³ Department of Molecular Biology, University of Salzburg, Salzburg, Austria

⁴ Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA

⁵ Departments of Experimental Immunology and of Otorhinolaryngology, Academic Medical Center, Amsterdam, The Netherlands

⁶ Department of Allergology, Paul-Ehrlich-Institut, Langen, Germany

⁷ Department of Pathophysiology and Allergy Research, Medical University of Vienna, Vienna, Austria

Scope: Increasingly regulators are demanding evaluation of potential allergenicity of foods prior to marketing. Primary risks are the transfer of allergens or potentially cross-reactive proteins into new foods. AllergenOnline was developed in 2005 as a peer-reviewed bioinformatics platform to evaluate risks of new dietary proteins in genetically modified organisms (GMO) and novel foods.

Methods and results: The process used to identify suspected allergens and evaluate the evidence of allergenicity was refined between 2010 and 2015. Candidate proteins are identified from the NCBI database using keyword searches, the WHO/IUIS nomenclature database and peer reviewed publications. Criteria to classify proteins as allergens are described. Characteristics of the protein, the source and human subjects, test methods and results are evaluated by our expert panel and archived. Food, inhalant, salivary, venom, and contact allergens are included. Users access allergen sequences through links to the NCBI database and relevant references are listed online. Version 16 includes 1956 sequences from 778 taxonomic-protein groups that are accepted with evidence of allergic serum IgE-binding and/or biological activity.

Conclusion: AllergenOnline provides a useful peer-reviewed tool for identifying the primary potential risks of allergy for GMOs and novel foods based on criteria described by the Codex Alimentarius Commission (2003).

Keywords:

Allergens / Bioinformatics / Food allergy / Genetically modified / Risk assessment

Received: October 1, 2016

Revised: February 4, 2016

Accepted: February 9, 2016

1 Introduction

Food allergy has become an important food safety issue that was hardly recognized 50 years ago. In the past 20 years

the prevalence of IgE-mediated food allergy has grown from 1–2% overall to 5–8% in the United States, with similar numbers in the European Union [1, 2]. Diverse environmental and behavioral factors likely contribute to the increased prevalence of sensitization and allergy, however, objectively defined causal factors to explain the increases are still lacking [3]. Control of manufacturing, sourcing, and labeling of ingredients in processed foods has become a major focus of regulators in many countries to help protect allergic consumers. However, this does not address risks associated with restaurant or home-cooked meals, which more frequently cause severe reactions [4]. While only a few foods and proteins are

Correspondence: Prof. Richard E. Goodman

E-mail: rgoodman2@unl.edu

Abbreviations: AOL, AllergenOnline.org; FARRP, Food Allergy Research and Resource Program; FDA, US Food and Drug Administration; GM, genetically modified; GMO, genetically modified organisms; IgE, Immunoglobulin E; NCBI, National Center for Biotechnology Information; SPT, Skin Prick Test

responsible for the majority of severe clinical reactions and the focus of food safety initiatives, many additional foods and proteins are “falsely” labeled as allergens in public sequence databases and scientific literature confounding regulatory dossiers.

The possibility that new food crop varieties developed through genetic engineering, often referred to as genetically modified (GM) crops, might cause an increased risk of food allergy was recognized in 1992 by the US Food and Drug Administration (FDA) (www.aphis.usda.gov/brs/fedregister/coordinated_framework). As described previously, development of a protective allergenicity assessment process based on scientific knowledge of allergy and allergens has been successful in protecting consumers [5–8]. To date there is no proof that an introduced protein in an approved GM crop has caused food allergy or risks for consumers. Arguably the most important assessment tool is the use of bioinformatics (sequence comparisons) to evaluate candidate GM proteins to prevent the transfer of an allergen or likely cross-reactive protein from one source into a different species used for food production [8].

2 Allergenonline.org database

The AllergenOnline.org (AOL) is maintained by the Food Allergy Research and Resource Program (FARRP) at the University of Nebraska-Lincoln. The database was developed in 2005 and provides a risk assessment tool for evaluating the potential allergenicity of new food proteins produced by genetically modified organisms (GMO) and novel protein ingredients in processed foods as well as a research aid. An overview of the management, expert panel, criteria for including proteins and financial support are described for transparency. The rationale for the processes of construction and updating the database are included. The protein risk assessment sequence search evaluation for GM products and novel proteins is described briefly but this manuscript focuses on the identification of allergenic proteins from sources causing allergic reactions upon ingestion, inhalation, contact, sting, and bite, respectively. Importantly there are no uniform objective characteristics that determine whether a protein will become an allergen. Investigators have tried to develop computer programs to predict the allergenic potential of proteins based on amino acid sequences or structural predictions, but so far none have proven predictive over diverse allergen types [9]. The most predictive comparison for risk evaluation is whether a new protein is at least 50% identical to a known allergenic protein based on local sequence alignment (FASTA or BLASTP) indicating that it may be cross-reactive in IgE binding or may be a *bonafide* allergen as observed by Aalberse [10]. Post-translational modifications including disulfide bond linkages, proteolysis, glycosylation, myristoylation, hydroxylation, or other alterations can change physical properties that might enhance or reduce potential allergenicity. The abundance of the protein, matrix composition, and the

route of exposure may also impact potential sensitization or elicitation. The comparison may demonstrate whether a new protein is already an allergen or is sufficiently identical to an allergen to suspect likely cross-reactivity. It should be recognized that all allergens are not equal. Some are extremely potent while other proteins bind IgE in a high percentage of people, yet rarely, if ever cause reactions [11].

2.1 Defining allergens

The task of defining allergens is complex. Type I allergens are typically proteins that sensitize (induce both protein-specific Th2 helper cells and protein-specific IgE producing B cells) and elicit an IgE response by cross-linking IgE on surface receptors of mast cells and basophils, leading to mediator release and symptoms. However, nearly identical proteins may elicit a cross-reactive allergic response due to IgE binding to a structurally related protein. Allergic symptoms vary remarkably between subjects, and include urticaria, angioedema, asthma, hypotension, vomiting, diarrhea, or multisystem involvement defining anaphylaxis that can lead to death. Objective clinical reactions to individual proteins are rarely determined. Normally only the allergenic source is identified by testing with a complex mixture of proteins extracted from the source. While some publications suggest that allergens are typically glycoproteins of 10–70 kDa, the amino acid sequence, structure, size, abundance, stability in heat and to proteases and the sources of allergens are highly variable and few allergens fit all of those characteristics [5]. Demonstration of specific IgE binding to a protein *in vitro* does not prove that the protein will cause allergy *in vivo* as two or more IgE-binding sites are necessary to cross-link receptors on effector cells. Other factors including affinity of binding, abundance, and exposure will determine reactivity and relative risks. Identification of the protein(s) causing allergic reactions is complicated as most allergenic sources contain a number of proteins that bind IgE of some allergic subjects.

Thus deciding which proteins to include in a searchable database for risk assessment requires a rigorous review by allergy experts. We have set criteria to define *bonafide* allergens, putative allergens, or proteins with unproven allergenicity (Table 1). The latter comprise proteins considered as allergens based only on sequence homology, IgE binding without clinical disease, immune responses to parasites where IgE aids in protection, or sequences that are too short to cross-link IgE bound to mast cell or basophil FcεRI (Table 2).

2.2 Food allergy assessment of novel proteins for GM crops: USA and EU

In order to reduce health risks of GM crops, a consensus evaluation process was developed through consultations with a variety of scientists prior to 1990 in the United States [12]. The primary concerns for food safety of GM crops include

Table 1. Evaluating the evidence of allergenicity for new proteins

Evidence	Allergen	Putative allergen	Unproven
Evidence the protein is expressed in an allergenic source including the correct tissues	Direct protein by LC-MS/MS, N-terminal sequencing, specific antibody binding or as a minimum mRNA	Direct protein by LC-MS/MS, N-terminal sequencing, specific antibody binding or as a minimum mRNA	Negative
Description of source of protein as recombinant or natural source and characterization	Clear as to genus and species, appropriate organ-specific material (fruit, pollen, etc.), or cDNA with expression of recombinant protein	Recombinant protein from mRNA without proof the protein is in material humans are exposed to.	In adequate description or genomic sequence only
Allergic serum donors	Description of clinical symptoms for allergic subjects. Clinical challenge to the source and specific IgE binding to protein (e.g. blinded food allergen challenge, inhalation challenge, or skin prick tests)	Appropriate clinical description for allergic subjects. Specific IgE binding to an extract of the source, but no challenge with extract or natural source	Vague clinical symptom description (e.g. asthma, atopy) without reference to the source. Or serum IgE binding to protein without symptoms
In vitro IgE-binding data	Immunoblot or ELISA with pure protein and extract of source, with appropriate negative controls. Method described with appropriate controls. Inhibition of binding to protein using extract or potentially cross-reactive protein	In vitro IgE binding with recombinant protein only, without an extract or purified natural protein from relevant raw material	IgE binding to extract not pure protein, or IgE binding to protein without a benchmark of binding to extract or other control proteins, threshold set based on a very low threshold such as 2 Std. Dev. Above negative control
Protein includes potential cross-reactive carbohydrate determinants (CCD)	Demonstration that CCD is not responsible for binding, based on expression in <i>E. coli</i> , or proven deglycosylation	Presence of CCD not tested, probable glycoprotein from plant or arthropod, with IgE binding possibly due to CCD	
Proven biological activity of IgE binding	Skin prick test or basophil activation with pure protein with appropriate controls	Skin prick test or basophil activation without a benchmark of activation.	No biological assay and no clear specific IgE binding

The primary information used to categorize a protein as an allergen, or putative allergen is defined in peer-reviewed publications. Since few proteins meet all criteria, therefore a weight of evidence is used to judge the outcome. Proteins that do not meet minimum criteria are categorized as unproven but will be reevaluated if relevant new publications are identified.

the potential risks of allergenicity or toxicity from novel proteins and nutritional equivalence of the GM plant. Nearly 150 academic, regulatory and industrial scientists from various disciplines were involved in developing the FDA recommendations [8]. An important conclusion was that the safety of GMOs should be evaluated using existing U.S. food safety laws, administered by the FDA, Department of Agriculture (USDA), and the Environmental Protection Agency (EPA) [13].

Regulations and guidance documents in most other countries were developed following the 1992 US FDA recommendations. The utility of the assessment was demonstrated by the example of a potential new soybean product developed by Pioneer Hi-Bred, which incorporated a gene for the 2S

albumin from Brazil nut to improve the nutritional quality of soybeans by increasing the cysteine and methionine content [6]. At that time no allergens from Brazil nut were identified as being responsible for eliciting allergic responses. However, because the gene was derived from a commonly allergenic source (tree nuts), in vitro serum IgE-binding tests were performed using sera from Brazil nut allergic volunteers. Additionally, skin prick tests (SPT) with extracts of Brazil nut, non-GM soybean, and the GM soybean were performed. Results demonstrated that the transferred gene encoded a previously unrecognized major allergen that would have put Brazil nut allergic consumers at risk if they consumed the GM soybean [6]. Therefore developers stopped development of that product, which was never used in food production. The

Table 2. Defining evidence for exclusion from the annual NCBI download for the expert panel review

Evidence	Explanation
Sequence Homology Human proteins	NCBI entries labeled as “allergen” or “allergen-like” simply due to low-sequence identity matches. Although there are published reports of IgE binding to some human proteins, those are involved in autoimmune responses either due to high sequence identity to exogenous allergens (e.g. profilins) or to marked inflammation and auto-immune responses in those with lupus or arthritis.
Classification of protein type	Some entries are classified as allergens simply based on the class, or type of protein (e.g. trypsin inhibitor, SCP, MD2), without evidence of IgE binding or allergenicity.
Patented Sequences	Individuals and companies are entering sequences in NCBI that are patented, including proteins that are defined as allergens without established proof of IgE binding or allergic elicitation; or sequences that might be useful for immunotherapy. These are not identifiable by the NCBI BLAST alignment tool. There are not links to information that would allow independent verification of allergenic activity.
Genome model organism sequences with “allergen-like designations”	Proteins from <i>Arabidopsis thaliana</i> , <i>Danio rerio</i> , <i>Drosophila melanogaster</i> , and other sequences from genome projects and without evidence of expression (e.g. mRNA, cDNA, or protein) are excluded unless there is a publication clearly linking the sequence to an allergy study.
Parasite proteins	Proteins from parasitic worms and arthropods are excluded unless there are clear published data demonstrating the source causes allergic disease in humans (not just in domestic animals), since the natural immune response to most parasites is generation of IgE antibodies and development of basophil and eosinophil responses. True allergy is elicited by only a few parasitic organisms, for example, the fish parasite <i>Anisakis simplex</i> , causes sensitization during first infection in humans with live worms, but can elicit allergic responses in later exposure to heat-killed worms in infected fish.
Sequences less than 8 amino acids long	The shortest useful sequences for bioinformatics inquiry regarding potential cross-reactivity are 8 or more amino acids long. Although unlikely to be unique in NCBI at 8 amino acids, there is no rationale for including sequences of less than eight.
No public reference of allergenicity beyond brief comments in NCBI or UNIPROT	Sequences that do not have associated publication references in the NCBI or UNIPROT databases and no identified reference in PubMed or Web of Science are not included in for panel review.

The following guideline was developed to eliminate thousands of protein sequences that have no evidence of allergenicity or are unlikely to have published evidence.

experience provided an important example of the predictive value of methods detailed by Metcalfe et al. (1996) and adopted by the CODEX Alimentarius Commission (2003), that serves as an international guideline for the safety assessment of genetically engineered crops [5, 7, 14]. Subsequently, the European Food Safety Authority (EFSA) used an extensive review process from 2007 until 2010 that produced recommendations for ways the assessment might be improved. The primary finding was the stated need for maintaining a curated, inclusive allergen database that is updated regularly [15].

2.3 Allergen protein databases for risk assessment

Individuals may be allergic to airway, contact or injected allergens and transfer of those allergens into a food source may cause food allergy. The risk assessment includes evaluate of the similarity of introduced proteins to all types of allergens. Researchers are identifying proteins as allergens from a variety of source organisms, but often the evidence of allergenicity is weak. In 1996, Astwood compiled a set of 219 allergens and gliadins (potential celiac eliciting proteins) from primary public sequence databases for risk assessment of Monsanto's GM products [16]. Prior to 2004 each

biotechnology development company generated their own allergen databases for the evaluation of GM crops. Then seven companies agreed to fund an independent academic group, the FARRP at the University of Nebraska, to develop a common, publically available allergen database designed for use in bioinformatics searches to screen candidate GM proteins for potential risks of cross-reactivity [17]. The AllergenOnline.org database, (also referred to as the FARRP database), originally included protein sequences from the NCBI protein database that were labeled as “allergen” or “allergenic,” as versions 5 (1189 total sequences) and version 6 (1537 total sequences) if expressed from a known allergenic source. However, a search for publications associated with many of the proteins in version 6 of AOL (unpublished, 2006) failed to identify scientific studies demonstrating human serum IgE binding or allergy to the source of the proteins. Therefore we developed and applied criteria for use in updates of AOL, version 7 that reduced the list of allergenic sequences to 1251 sequences. Since 2007 we have made minor changes in criteria yet man the number of sequences accepted as allergens has grown during annual updates as shown in the version history of AOL (Table 3).

While a few other allergen databases exist that might be used for risk assessment, they do not have a clearly

Table 3. Version history AOL sequence download and decisions

NCBI Searches for candidate allergens				AllergenOnline decisions (combined putative and bonafide allergens)			
Date range	Allerg*	Allergy	Allergen	Version #	Total # seq	Taxon protein groups	Species
1960/1/1–2005/06/01	33 992	16 934	4 376	6	1537	672	255
2005/06/01 – 2006/06/01	34 470	14 442	962	7	1251	451	221
2006/06/01 – 2007/06/01	8523	5211	2174	8	1313	483	229
2007/06/01 – 2008/06/01	27 558	24 754	2377	9	1386	502	236
2008/06/01 – 2009/06/01	126 837	123 259	3141	10	1471	529	254
2009/06/01 – 2010/06/01	138 033	133 234	3176	11	1491	553	265
2010/06/01 – 2011/06/01	227 716	223 351	3609	12	1603	603	273
2011/06/01 – 2012/06/01	308 529	299 853	3332	13	1630	612	275
2012/06/01 – 2013/06/01	935 229	921 032	8905	14	1706	645	290
2013/06/01 – 2014/06/01	932 310	911 991	19 793	15	1897	744	335
2014/06/01 – 2015/06/01	166 730	147 289	18 082	16	1956	778	345
NO date limit	2 953 644	2 843 504	70 705	na	na	na	na

Potential allergen sequences identified by keyword searches of the NCBI protein database using date limitations are listed as candidates along with the final determinations (cumulative) for versions of AOL. Note: for version 7, more stringent criteria were implemented reducing the number of allergens. The version number, total number of sequences, taxonomic protein groups and number of species accepted in AllergenOnline.org are shown (right side).

defined review process, or do not use the same risk assessment search algorithms. The WHO/IUIS Allergen Nomenclature Subcommittee maintains a database that provides recognizable names for specific allergens (e.g. Ara h 2 for the 2S albumin from *Arachis hypogaea*) based on taxonomy and protein types (www.allergen.org), prior to publication. The SDAP database has fewer sequences (1526) compared to AOL. Importantly, the review process has not been described and SDAP was last updated in February, 2013 (http://fermi.utmb.edu/SDAP/sdap_man.html). Proteins in the AllFam database (1091 allergens in 186 allergen families) are organized into structural protein families to provide information and associations to help identify potentially shared conformational epitopes. AllFam was last updated in 2011 (www.meduniwien.ac.at/allergens/allfam). The Allergome database (www.allergome.org) is a very good source of references for scientific information regarding allergy to sources of allergens and to many allergens. However, many of the entries in Allergome do not include sequences. The general sequence protein database, NCBI Protein (www.ncbi.nlm.nih.gov/protein) is a repository of more than 70 million protein (amino acid) sequences including those translated from GenBank, RefSeq, and TPA as well as sequences from SwissProt, PIR, PRF, and PDB as described on their website databases. The NCBI protein database is searchable by key word queries and provides a BLAST alignment tool for local sequence alignments to all entries, or to subsets based on user queries. However, keywords related to allergy are often entered based on auto-annotation from very low-level sequence identities to an allergen. Sequence searches of NCBI can be made using “allergen,” but there are 69 000 plus “allergen” entries and most are not associated

with published data demonstrating allergenicity. Thus the focus of the AOL is to provide a comprehensive, sequence searchable database of proteins based on publication of at least protein-specific IgE binding, and preferably biological activity using sera from subjects allergic to the source.

2.4 Allergenic protein sequence comparisons for food risk evaluation

Sequence searching algorithms that have been recommended for evaluating food allergy risks of GM proteins use a full-length FASTA (or BLASP), a sliding 80mer FASTA and/or exact 8 amino acid matching algorithms to compare the newly introduced protein with sequences of allergens to identify proteins that allergens or are likely to pose a risk of cross-reactivity [9, 18, 19]. The predictive value of the computer search depends on the quality of the database (inclusion of allergens, exclusion of proteins without proof of IgE binding or elicitation of allergy), the specificity of search algorithms and the criteria of significance. Although some researchers claim to have developed computer programs that predict B-cell epitopes, potential IgE binding is predicted by estimating structural motifs based on amino acid properties including charge hydrophobicity and nearest neighbor analysis, 3D surface plots or similarities to other known allergenic epitopes [20–22], but without broad validation. Development and maintenance of allergy to a given protein requires establishment of both T-helper cell epitopes and B-cell epitopes. However, specific proteins may elicit reactions through cross-reactive recognition of IgE epitopes without T-cell help [10]. The great diversity in individual allergic patient responses

to different proteins in complex food and environmental allergenic sources and of epitope recognition to the same protein makes validation of predictions complex as demonstrated by recent studies of allergen microarrays for allergy diagnosis [23, 24]. Based on current studies it seems that the best method to predict potential risks of allergy or cross-reactivity is still to use a local alignment method (BLASTP or FASTA) with identity scores of greater than 50% to identify proteins that are known to be allergens or so similar in sequence to indicate a need to perform serum testing using samples from those with allergy to the matched protein as recommended by the CODEX Alimentarius Commission in 2003 [9, 14]. The CODEX Alimentarius Commission guideline is not a regulation, however, CODEX member countries including members of the European Union are expected to follow CODEX [7, 25, 26]. The guidance suggests “significant sequence identity matches” are those exceeding 35% identity in alignments over segments of 80 amino acids or longer. In addition, short segments (six or eight) amino acids with 100% identity match were suggested as possibly being meaningful although the predictive value of such matches without overall long identity alignments has been largely discredited [9, 17]. As scientific knowledge is gained about allergy and allergens, new methods and criteria may eventually be accepted if proven predictive [27]. Two studies suggest using an alternative to the sliding window of 80 amino acid approach, by using the statistical *E* score from FASTA or BLASTP to judge the relevance of overall identity matches might help in judging potential cross-reactivity when used to evaluate the significance of percent identity scores [28, 29]. However, *E* scores vary markedly with the size of the database and the length of aligned proteins. The *E* scores should be more useful if utilized as an additional criterion for judging marginal identity alignments (e.g. 35–50%) rather than being used as the primary criteria for potential cross-reactivity [17]. The results of the bioinformatics search inform developers and regulators regarding whether human serum IgE studies are warranted. However, appropriate serum testing requires well-characterized allergic and control sera, specific control and test substance characterization, validation of IgE detection reagents and inhibition to verify specificity of results [6, 9, 30, 31].

3 Origin of allergenonline database and selection of candidate allergens

The first public version of the AOL was compiled in 2005 (called version 5). It included 1189 protein amino acid (aa) sequences that represented 608 protein-taxonomic groups including isoallergens and variants (isoforms) as defined by King et al. [32]; from a total of 208 species of plants, animals, and fungi [16]. That dataset was seeded from Monsanto's 2004 database, which is an expanded version of the ALLERGEN3 dataset described by Hileman et al. [33]. Version 5 of AOL also included the entries from four other biotechnology developers (Bayer CropScience, Dow AgroSciences,

DuPont-Pioneer, and Syngenta AG). Duplicates and substrings of longer sequences were removed. Additional sequences were added through keyword searches (allerg*) of the NCBI Protein database.

Sequences in the AOL include environmental allergens representing proteins from pollen of flowering plants, common molds, house dust mites (HDM), and insects; contact allergens (e.g. latex) and food allergens. The complete dataset of unique sequences (>2100 proteins) was compared to NCBI sequence entries (February, 2005) to verify annotation and the appropriate GenInfo Identifier (GI) number. During annual AOL updates (2006–2015) large increases in the number of keyword identified sequences were identified without proof of allergy as shown in Table 3. Table 3 also shows the number of sequences, groups, and species that were accepted as allergens based on criteria described below. The sequences are organized in taxonomic-protein groups based on unique species (or occasionally two species) and protein type (at least 65% identical, similar functional protein).

Because literature searches and reviews are quite time consuming, we developed a screening system for removing a large number of proteins that are highly unlikely to cause IgE-mediated allergies (described in Section 2.2).

3.1 Identifying new candidate allergenic proteins

Many well-characterized allergens have been identified from clearly defined allergenic sources over the past few decades. Understanding the source and exposure are important considerations for inclusion. Food allergens are often well characterized as exposure is usually identifiable and controllable, allowing clear definition of study subjects. While diverse foods may occasionally cause allergy, eight foods, or food groups (peanuts, soybeans, tree nuts, milk, eggs, fish, crustaceans, and wheat) are thought to account for nearly 90% of reactions in the United States [5, 34]. Recent studies demonstrate that even among the common allergens, peanut, and a few tree nut species (e.g. almond, walnut, pecan, and hazelnut), shrimp and milk cause most of the fatal and life-threatening reactions [35]. Only a few proteins in each of these complex food sources bind the majority of IgE and are likely responsible for eliciting serious reactions [36, 37]. But identification of the biologically important allergens is difficult. While a large percent of people believe they have food allergy, careful diagnosis using IgE binding, skin prick tests, and food challenges demonstrate no food allergy, even if individuals have IgE that binds to a protein of the source [38]. A recent study of 4000 cases of severe anaphylaxis in central Europe implicated many foods commonly causing anaphylaxis in the United States (peanuts, tree nuts, milk, and crustaceans, fish, and hens egg) although a surprising number of wheat and celeriac allergic subjects were also identified [39]. Population studies using the same reagents and methods in China, India, and Russia demonstrated much higher rates of sensitization (IgE binding) than allergy as judged by lack of SPT reactions

and histories [40]. However, allergy in young children to commonly consumed foods, e.g. milk and egg, were similar to those in the United States and EU [40]. A new study in the EU demonstrated that self-reported allergy and specific IgE binding using an allergen-protein microarray overestimated clinical reactivity [41].

The dominant allergenic proteins in peanuts in various populations are 2S albumins (Ara h 2 and Ara h 6), 7S vicilin (Ara h 1), and 11S glycinin (Ara h 3) according to Palmer et al. [42] and Vickery et al. [43]. Homologous proteins in walnut have been identified as the major allergens in walnut [44]. Yet in spite of moderately high sequence identities (33–53%) between peanut and walnut homologous allergens, there is little evidence that they are cross-reactive [44]. A non-specific lipid transfer protein (nsLTP) is the major allergen identified in peach, which can cause severe food-allergic reactions in some subjects from Italy and Spain [45, 46], but similar allergies are rarely reported elsewhere. While nsLTP sequences and structures are highly conserved in fruits of *Rosaceae* and across much broader taxonomic groups and in vitro IgE binding is often shared, allergy to the sources is often much more restricted [47, 48]. Many other proteins from these foods have been identified as allergens. While careful studies have demonstrated IgE binding to a number of other proteins from these sources, the potency, and relevance of many for allergy is hard to demonstrate.

Inhalation allergy and allergens are often difficult to diagnose and proof of the identity of the proteins responsible for eliciting allergic reactions is often lacking. The abundance and time of exposure is important. There is little doubt that the pathogenesis related proteins (e.g. PR-10) of birch pollen (Bet v 1) and related trees are prominent allergens for many subjects experiencing rhinitis, conjunctivitis, and asthma in the spring in northern Europe [49, 50]. However, there is less proof that the profilin of birch pollen (Bet v 2) or cyclophilin (Bet v 7) are important allergens. Conservation of the amino acid sequence and three dimensional structures of proteins across diverse taxonomic organisms (birch, apple, carrot, and hazelnut) means that sera from some donors allergic to birch pollen may bind to homologues in apple, carrot, and hazelnut by in vitro IgE assays. More than 25 proteins from two species of *Dermatophagoides* sp. are listed as allergens in the WHO/IUIS allergen database, however, few of them have been proven to be significant allergens [51]. Further complexity is added due to the existence of other diverse species of mites and arthropods that are from other environments, with different possibilities of exposure for the individual.

Studies reporting proteins as allergens often rely on in vitro protein-specific serum IgE-binding data. The formats of the tests vary from simple immunoblotting of extracts followed by purification or identification of proteins by MS or N-terminal sequence to IgE binding to highly purified natural or recombinant proteins purified and presented in native or denatured forms. Typically inhibition assays with homologous proteins from other sources are required to demonstrate specificity of binding. Appropriate negative control sera

and nonallergenic proteins are required to demonstrate specificity. The specificity of secondary reagents, blocking methods should be well described as reviewed and demonstrated in a number of studies [30, 52]. Test proteins must be well characterized, appropriately folded, and presented to provide accurate IgE binding compared to the multiple proteins present in the source materials in order to appropriately diagnose allergy and identify allergens [53–55]. However, IgE binding to common pan-allergens may be misleading without inhibition data [47, 52].

3.2 Excluding proteins that are unlikely to represent risks of allergy

In order to reduce the number of likely irrelevant proteins that are not from allergenic sources and may be not be expressed, we have developed a set of criteria to exclude sequences from NCBI that are highly unlikely to be *bonafide* allergens (Table 2). However, annotation of NCBI sequences changes over time. Thus each year we challenge the screening process to reduce the likelihood of missing potential allergens. The rapid increase in the number and diversity of genomes that have been sequenced recently and annotated with predicted (hypothetical) proteins labeled as allergens or allergen-like proteins is surprising. Data fields of NCBI entries have to be screened to exclude entries where the only association with allergy is that the authors are from an allergy institute (e.g. National Institute of Allergy and Infectious Diseases) or the organism is known to cause allergies (e.g. *Aspergillus*). Criteria defining major groups for exclusion are listed in Table 2. Candidate allergen entries and information from publications describing studies of allergenicity related to the taxonomic source or the specific proteins are entered in a private MySQL database in FARRP to provide a mechanism for reviewing and archiving for each annual review. Detailed descriptions, expert opinions, referenced publications, and decisions are recorded and maintained by FARRP each year. The public database (<http://www.allergenonline.org/>) only shows sequences of proteins with sufficient published evidence of allergy, at a minimum-specific IgE binding from sera of subjects allergic to the source, to suggest a role in elicitation of symptomatic allergy. The public AOL database includes a description of the database (Home, About, and Contact) with a link to the version's historical record. The Browse function shows all accepted allergens, with species name, IUIS allergen name (if available), GI: number links to protein sequences and a links on the Group column listing relevant references. The Sequence Search link goes to a sequence entry box (FASTA format or plain text sequence, with or without numbers) and has a link to a Support page that describes the full-length and 80mer FASTA searches as well as an exact eight amino acid identity match section with suggestions for interpretation.

The download search criteria and resulting total candidate sequence numbers from the NCBI search for the current

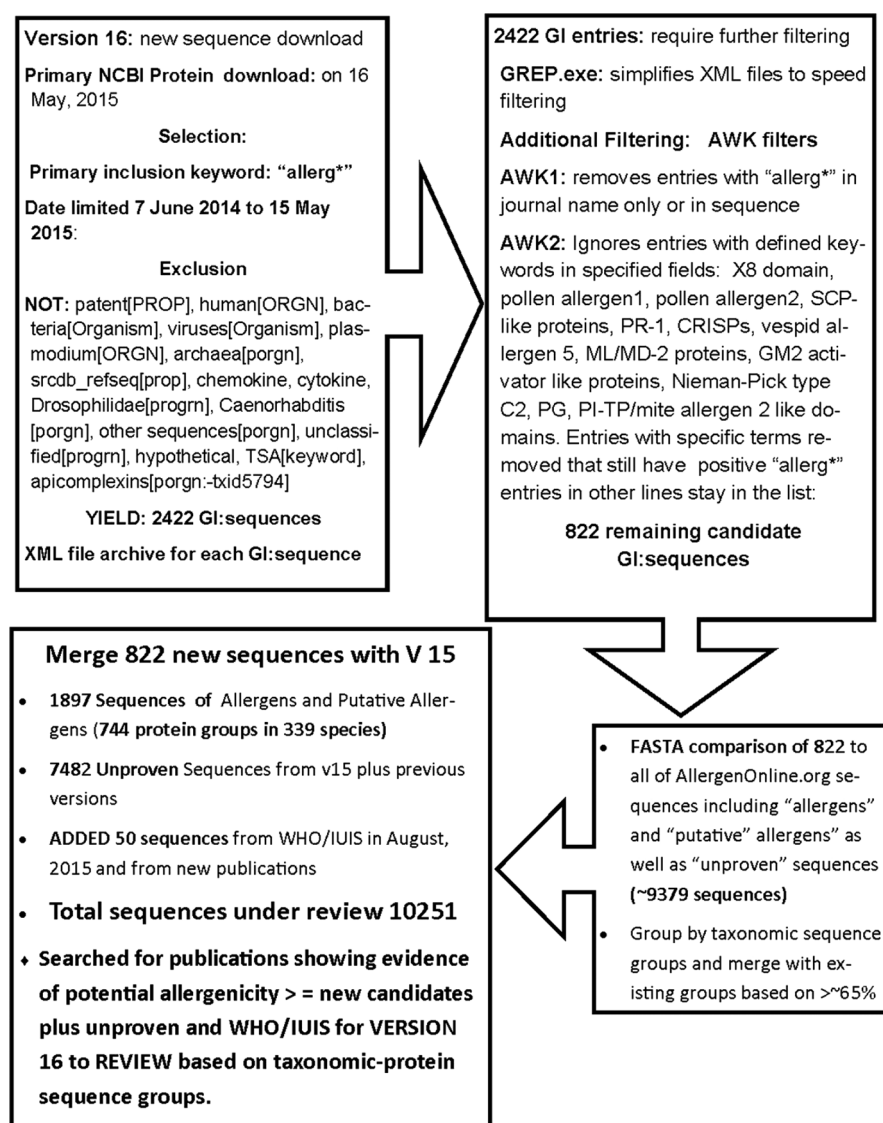


Figure 1. Download process of new candidate protein sequences for version 15. The upper-left box lists the query terms for NCBI to accumulate candidates. A total of 1511 sequences were gathered and subjected to further filtering with AWK filters to remove irrelevant entries (upper-right box), bringing the list to 558 candidates that were compared to all preexisting entries (>9100 sequences) in the database archive (lower right box) using the FASTA3 comparison. All new entries (in existing groups or new groups) were reviewed by the expert panel. In addition, the "unproven" groups were reviewed for new published information and reviewed by the panel if there was new information.

version (v16) are shown in the upper left box of Fig. 1. The search was downloaded from the NCBI Protein database on 15 May, 2015 as the end date using "allerg*" as the inclusive primary keyword in the Entrez Query, followed by terms to limit the entries based on date (beginning June 7, 2014) and other group identifiers as explained here. Exclusion of major groups of sequences that are highly unlikely to include allergenic proteins was accomplished using the Boolean logic terms "NOT" and "AND" in the search query string with keywords and categories (Fig. 1). Patented sequences are excluded because of limited information to judge their involvement in allergy. A number of patented protein sequences are modified for reduced risk of IgE binding for therapeutic use. Human sequences were excluded as we consider human proteins not to be allergens, although some display high sequence identity to allergens of molds, and some act as targets of auto-immune reactions [55, 56]. However, there is no evidence that human proteins are the primary sensitizers.

Sequences from common genomic model organisms (e.g. *Arabidopsis thaliana*, *Danio rerio*, *Drosophila melanogaster*) are excluded as these organisms have not been demonstrated to cause allergies in humans. However, they contain many sequences labeled as allergenic or allergen-like, based on sequence homology. Viral sequences are excluded as there is no evidence that they elicit allergies. Specific viruses certainly skew immune responses, e.g. increase or exacerbate atopic asthma by enhancing the expression of signals including thymic stromal lymphopoietin (TSLP) or interleukins-25 and -33 [57, 58]. Respiratory syncytial viral (RSV) has been shown to skew the immune response away from a Th2- allergic response even though RSV and human rhinovirus are known to lead to development of asthma [59, 60]. While a number of mechanistic studies point to various immunomodulatory signals or outcomes induced by viruses, the evidence is primarily from studies in animal models. Evidence suggesting IgE binding to some viral proteins implicates cross-reactivity

that is unlikely to elicit reactions as demonstrated for IgE binding to VP4 of rotavirus [61]. Another investigation claims that a yellow fever virus vaccine caused anaphylaxis in 40 individuals (out five million recipients) to the first injection. However, it has been suggested that these reactions were not due to viral proteins, but rather egg proteins or gelatin used as carrier proteins [63]. Similarly, sequences from plasmodia, bacteria, and whole organism genome sequencing projects and immunological marker proteins (cytokines and chemokines) are excluded from the broad NCBI download using the search scheme shown in Fig. 1. This screening and filtering process allows us to avoid having to perform PubMed searches on thousands of proteins that are from organisms without known histories of allergy. Proteins from parasites are complicated to evaluate for potential allergenicity because in a number of cases, IgE and Th2 response provide immune-mediated protection against a number of parasites due to IgE triggering mast cells and helping develop eosinophil responses [64, 65]. However, there are a few parasites that clearly cause allergic responses. Humans can be infected with *Anisakis simplex* if ingesting raw, parasite infected fish. After the parasite dies in humans, some experience parasite protein-specific allergy when consuming well-cooked fish with dead *Anisakis simplex* worms [66]. *Ascaris sp.* has been suggested as another source of allergens. Therefore we search for publications of specific parasite proteins found in the WHO/IUIS database (www.allergen.org) and referenced in PubMed.

3.3 Integrating new entries with existing entries

The version 16 download identified 2422 potential new allergen sequences from NCBI (Fig. 1). A second set of screening filters (AWK filters) was used to filter out additional sequences that were annotated simply based on similarities to a few protein types (e.g. GM2, pollen allergen 2) without evidence the proteins have been detected or studied, leaving 822 sequences to integrate into existing entries (1897 allergens and 7482 unproven sequences). An additional 50 candidate entries were later identified (August, 2015) by a comparison to the WHO/IUIS database (www.allergen.org) and from our review of publications. The entries were compared to previous entries in AOL using FASTA to organize the entries in previously defined or new taxonomic-protein groups. The groups are defined by having 65% identity or more to the previous clearly defined allergen or within a new group. Specific PubMed literature searches were used to identify bacterial and viral proteins that might have been identified as allergens based on direct studies, but were missed due to our exclusion criteria since a few sequences from bacterial or viral sources may stimulate an allergic response [67–69].

Specific literature searches were performed for groups that were previously identified as “unproven” as well as new candidates. Many references were identified from the sequence entries in NCBI although those often describe cloning and

characterization of the protein and may not have relevant information regarding proof of allergy. Keyword searches of PubMed looks for references to the organism, the authors of sequences and the protein. Complex searches are needed for some searches to uncover relevant papers or to limit publications using Boolean terms. Publications were also identified in the WHO/IUIS database and occasionally Allergome. Literature searches were performed by two people (SL and RG). Copies of the articles deemed relevant are loaded in the review database for RG and other members of the expert panel.

4 Criteria for allergens, allergic subjects, and test interpretation

From 2005 to 2007 the expert panel developed criteria for selection of allergens, taking into consideration the uncertainties of accurate diagnosis of allergy, methods for characterizing proteins, serum testing methods and biological assays that have been used to identify clearly allergenic proteins (e.g. 2S albumins, lipid transfer proteins, vicilins, tropomyosins, proteases of house dust mites) for classifying proteins as proven allergens, putative allergens, or unproven (incomplete or absence of evidence of allergenicity).

4.1 Source of the protein

The protein must be isolated from natural sources or synthesized from cDNA of a source of proven allergy. Characterization must include confirmation of at least a partial protein sequence that matches a translated sequence of cDNA or genomic clone, or a nearly full-length amino acid sequence from purified protein with a description of methods of purification and characterization. Immunoblots of extracts associated with human exposure are useful in addition to IgE binding to purified protein. Recombinant vectors and hosts must be appropriate for producing recombinant proteins depending on the characteristics of the protein; considering glycosylation, disulfide bond formation, or other probable posttranslational processing. If the complete amino acid sequence of the protein tested for IgE binding is listed in the publication, the GI: number and characterization (including partial amino acid sequence) must be described.

4.2 Serum IgE binding

Natural, purified proteins, or recombinant proteins may be used as the target for serum IgE binding. Test methods must be described and may include immunoblotting from SDS-PAGE under reducing or nonreducing conditions; ELISA, radioallergosorbent assay (RAST), dot-blotting with purified protein, or microarray assays with a description of test and control materials. Irrelevant or nonallergenic control proteins should be used in the same assay along with nonallergic sera

to demonstrate that any secondary detection system (e.g. anti-IgE) is specific. The source of specific critical reagents used in the assays should be clearly described (e.g. monoclonal or polyclonal anti-IgE; company and product number).

Description of serum donors used for tests must be included. Symptoms should be described and associated with the source of the protein. The use of “atopic” sera without an association of symptoms to the source by the subject is not acceptable. Allergies to foods are often clear if the source is consumed intact (tree nuts or fruits). However, many sources of food allergy (e.g. corn, soybean, wheat) are typically found in processed foods that contain multiple ingredients. Identification of the causative agent often requires confirmation by serum IgE tests, skin prick tests, elimination diets, or even food challenges [70]. Symptoms for food allergy usually begin within 10–20 min following ingestion [70,71]. However, some subjects experience delayed reactions four to eight hours after consumption, as recently described for allergy to red meat in individuals sensitized to the galactose- α -1,3-galactose glycan on nonhuman mammalian muscle proteins, mimicking epitopes of tick salivary proteins [72]. Airway allergy including rhinosinusitis or allergic asthma may be seasonal due to pollen exposure or mold exposure during peak seasons, or perennial due to pet dander, house dust mites, molds or cockroach in indoor environments. Diagnosis is typically based on history and skin prick tests with appropriate extracts or in vitro IgE-binding assays. Venoms from stinging and biting insects, spiders, or reptiles are often easily diagnosed, but are generally not of common occurrence. Self-reported symptoms by the study subjects without specific IgE binding or other antigen-specific tests are questioned and in most cases not accepted. Studies reporting results with multiple subjects having allergy to the same source are much more convincing than single case reports, but the expert panel recognizes that the prevalence of allergy to many allergenic sources is low. Well-conducted single case reports can be more convincing than studies using a number of poorly characterized subjects. Scoring information (e.g. SPT, wheal diameter; ImmunoCAP reagents and scores) and positive and negative control reagents should be described. A clear demonstration that a well-characterized protein specifically binds IgE from appropriately allergic human subjects is taken as proof that a protein is at least a putative allergen.

Studies reporting allergic reactions in nonhuman animals are not accepted as demonstration that a given protein is an allergen for humans. Therefore proteins that have only been tested in rodents or domestic animals are excluded since the intended purpose of AllergenOnline.org is evaluating food safety for humans.

4.3 Biological activity

IgE antibody binding alone is not sufficient to merit classification as a *bonafide* allergen [73]. Allergic responses require effective cross-linking of IgE bound to Fc ϵ RI on the surface

of mast cells or basophils to cause release of histamine and other mediators. Therefore, the ultimate proof that a protein is an allergen is through biological activity, demonstrated by skin prick tests with pure protein, specific challenges (airway challenge for inhalation allergens or food challenge for dietary allergens) or by in vitro basophil activation or histamine release [73,74]. Our consensus opinion is that demonstration of basophil activation or skin prick test activity to the purified protein demonstrates sufficient evidence to consider a protein as a *bonafide* allergen.

Proteins that are highly similar in sequence and from the same taxonomic group to an accepted allergen, according to our criteria, should not require the same level of testing and evaluation as a new, unique candidate. Therefore proteins from a single species sharing more than 65% identity across most of the sequence are accepted as equivalent.

Since version 7 there have been only a few minor changes in our criteria, which are reviewed every year by the expert panel. Since few publications match our ideal data recommendations, some flexibility is incorporated in the system. For instance, we prefer studies testing at least ten allergic subjects and two to ten negative controls (those not allergic to the material under investigation). However, since specific allergies to some sources are relatively rare, exceptions using one to five allergic subjects may be accepted. Negative controls should include subjects with allergies to unrelated sources who are exposed to the allergenic source material without reacting.

4.4 Special exclusions

A few glycoproteins have been identified to bind IgE from atopic subjects selectively to asparagine linked, complex carbohydrate determinants. Some research indicates that such glycoproteins are unlikely to elicit allergic responses following their ingestion [75]. For example, the nonglycosylated recombinant tomato protein Lyc e 2 (also called Sola l 2) produced in *E. coli* did not bind IgE from tomato-allergic subjects. The native glycosylated protein caused basophil histamine release whereas the recombinant protein without glycan did not [76]. The expert panel will reevaluate the rationale for keeping this protein in AllergenOnline during the current update for version 17. A similar challenge is presented by the celery allergen Api g 5 [77]. Those two proteins are in the current version of AOL.

5 Review process

After new candidate proteins are sorted into taxonomic protein groups and references are loaded in the review database, if we did not find relevant publications, the database manager (RG) searches again during his review of all entries. Text boxes are used to describe the review and findings. For groups voted as unproven in one year, the group is re-reviewed again

for over the next two annual updates. Each group is then assigned to two additional experts for review and recording their comments. Grading decisions are recorded by selection in a drop-down menu as “allergen,” “putative,” or “unproven.” The reviews are then locked down and cannot be changed except by the database manager. When groups are fully scored, a final voting system is prepared and the reviewers have to revote, with the previous scores visible. They can check the detailed records prior to voting. The database manager then reviews all votes. If there is more than one dissenting vote or the arguments are compelling for a different outcome than voted by the majority, the database manager summarizes the dispute and returns the entry to the other experts for a rereview. Approximately one month after completion of final votes, the sequences and GI: numbers are checked again to ensure accuracy compared to the NCBI entries. In version 16 there are three entries that are not in the NCBI Protein database, but have been manually entered into AOL. Taxonomic-protein groups judged as “allergen” or “putative allergen” are referred to as allergens in the public database. Groups judged as “unproven” are held in the review system for later reevaluation in the next review cycles.

An archive of the new version is submitted to the database sponsors approximately one month prior to public release so they can prepare the databases (behind their firewall) that they use for regulatory submissions. A PDF of the complete set of allergens is included generated and posted on the AllergenOnline public website for convenience of those preparing regulatory dossiers.

5.1 AllergenOnline example decisions on candidate proteins

A sample of entries, references, and decisions are shown in Table 4. Five example proteins from diverse source organisms (cashew, *Anisakis* sp. parasite, honey bee venom, and house dust mite) judged as having clear evidence of allergenicity are described. Three examples judged as putative allergens (kiwi protein; *Alternaria* sp. mold protein and dog lipocalin) are described. And two sequences judged as unproven (a very short peptide from a protein of black gram, *Vigna mungo*; and a protein from Chinese jujube) are described.

5.2 Bioinformatics methods for safety evaluation of a novel protein

A hypothetical example of a protein that might be used in a GM crop {2S albumin (193 AA) in *Eucalyptus grandis* (GI:702384944)} was compared to AOL version 16. The full-length FASTA identified the best overall alignments with the 2S albumin of black walnut (*Juglans regia*) based on the smallest *E* score (8.9e-006) and 37.6% identity over a 181 AA alignment, which indicates probable homology. By a sliding 80 amino acid window, FASTA identified highly significant

alignments to 2S albumins of hazelnut (48.7%), pecan (45%), English walnut and black walnut (45%), sesame (42.5%), and castor bean (41.2%). Only a single 8 AA identity match was found using the 8mer identity match, to the castor bean 2S albumin.

6 AllergenOnline expert panel

Individual experts who have participated on the expert review panel are listed on the website and are included as coauthors. All have significant extensive records of publishing on allergy and allergens. We maintain records of decisions by the panel and provide the sponsors with an archival copy of each new version of the database (www.AllergenOnline.org) that they subscribe to.

7 Public access to the AllergenOnline database

The FARRP at the University of Nebraska maintains the database as a free public service database. We do not track users or record or view any sequences entered by users for searches.

8 Discussion

The AOL was established in 2005 to provide a tool for evaluating the safety of dietary proteins that might be introduced into foods through genetic modification of plants, animal, or microbes and used in food production, or proteins that might be introduced as novel ingredients through food processing. It is intended to provide a sequence and information dataset representing known and putative allergenic proteins of foods and environmental sources (inhalation, contact, or injected). Selected sequences represent known proteins from allergenic sources, which specifically bind IgE from some individuals allergic to the source. The database is not intended to include every minor variant of sequences of every allergen, as publications and available comprehensive databases include many redundant and theoretical sequences. References used to judge taxonomic-protein groups are listed, but cannot be provided to the public due to copyright laws. Novel proteins that match proteins from AllergenOnline at a level above the conservative threshold of identity recommended by regulatory agencies should be considered for testing to evaluate specific IgE binding and allergy if the novel protein will be present in food.

It must be recognized that the level of proof and the potency of allergenicity of the proteins in sequences selected in the database are not equal. Some are clearly proven, abundant, and potent allergens from allergenic sources demonstrated to cause systemic anaphylaxis in sensitized individuals. There is evidence that many others are likely to contribute to the aller-

Table 4. Examples of allergens, putative allergens, and unproven, version 16

Genus/species	Candidate proteins	GI# protein	Publications	Current status
Anacardium occidentale Cashew	Legumin seed storage protein WHO/IUIS Ana o 2.0101	25991543	[79] Wang et al., 2003	Allergen
Anisakis simplex Fish parasite	Cysteine-protease inhibitor WHO/IUIS Ani s 4.0101	47605398 110346534	[80] Moneo et al., 2005, [81] Rodriguez-Mahillo et al., 2007	Allergen
Apis mellifera Honey bee	Dipeptidyl peptidase venom protein WHO/IUIS Api m 5.0101	187281543	[82] Blank et al., 2010	Allergen
Betula verrucosa Betula pendula	Pathogenesis related protein 10 (PR10) WHO/IUIS Bet v 1.0101	17938 Many more	[49] Swoboda et al., 1995 [50] Erdmann et al., 2005	Allergen
Birch tree pollen Dermatophagoides pteronyssinus European house dust mite	ML domain lipid binding protein WHO/IUIS Der p 2.0101	9280543 Many more	[83] Chua et al., 1990 [84] An et al., 2013	Allergen
Alternaria alternata Plant pathogenic fungus	Manganese superoxide dismutase WHO/IUIS Alt a 14.0101	529279957	[85] Postigo et al., 2011 [86] Gabriel et al., 2015	Putative allergen
Canis lupus familiaris Dog dander lipocalin	Lipocalin like protein WHO/IUIS Can f 4.0101	262232390	[87] Mattsson et al., 2010 [88] Niemi et al., 2014	Putative allergen
Actinidia deliciosa Kiwi fruit	Pectinmethylesterase inhibitor WHO/IUIS Act d 6.0101	27544452	[89] Irifune et al., 2004 [75] Bublin et al., 2011	Unproven
Vigna mungo Black gram	28 kDa protein purified by only 12 amino acids determined	410591582	[90] Kumari et al., 2012	Unproven
Ziziphus jujube Chinese jujube tree	Bet v 1 homologue Erroneous IUIS-like number	631901704	No publication	Unproven

gic response. Yet for many sequences included in AOL, there is only evidence of weak IgE binding or the proteins may represent minor components in suspected allergenic sources. Interpretation of results obtained from the use of the AllergenOnline database is the responsibility of the user. High percent identity matches to sequences with many clear references presenting objective symptoms associated with allergic disease indicate probable risk for some allergic consumers. Modest or marginal identity matches probably represent low risk. High identity matches to proteins from sources rarely reported as causing allergy or only subjective symptoms probably pose little risk. However, the context of future exposure and processing of foods that might include those proteins should be taken into consideration.

During the past 20 years, the number of putative allergens has increased rapidly. Yet less than 2000 proteins have been demonstrated to specifically bind IgE from subjects experiencing allergic symptoms following exposure to the respective proteins or protein source. Even less of those proteins have been demonstrated to have specific biological allergenic activity defined by SPT or basophil activation. The

majority of data is from studies in Europe or North America. However, allergy studies are now being reported from diverse geographies where different source organisms of food and environmental organisms exist and the list of diverse allergens is expanding. The risk assessment process has become more challenging during the past 10 years as the genomes of many organisms are being added to the database accompanied by auto-annotation of proteins as “allergens” or “allergen-like” based on low level sequence or structural similarity.

The AOL is intended to help protect consumers around the world by providing a tool that is updated regularly based on global scientific information to help developers and regulators protect allergic consumers from unexpected and accidental exposure to their allergens in new food sources. The AOL has been used by the sponsoring companies to evaluate potential risks of allergenicity of their GM proteins. It has also been used in studies to evaluate potential products developed by nonprofit organizations and academic developers for example, Indian mustard plants designed to enhance hybrid seed production [78].

Financial support for the AOL was provided (2004–2015) by contractual agreements between the University of Nebraska-Lincoln (UNL) and the following companies: BASF Plant Science, Bayer CropScience, Dow AgroSciences, DuPont-Pioneer, KWS SAAT, Limagrain (CIE & Vilmorin), Monsanto Company, Proctor and Gamble (one of the seven original sponsors); as well as the Food Allergy Research and Resource Program at UNL. BASF, DuPont-Pioneer, and FARRP faculty are continuing to support AOL in 2016 and beyond. We gratefully acknowledge Sue Hefle (deceased) who was a Co-Director of FARRP and played an important part in development of the concept of the collaborative database.

R.G. drafted the manuscript, but all authors reviewed and contributed. R.G. and S.T. are the project leaders. J.W. was responsible for developing the web-based review system and the public database website, which is now maintained by S.L. The expert panel (S.T., J.B., H.S., M.E., R.vR., B.B., F.F., R.G., and previously S.V.) develop the guidelines for criteria used to classify proteins as allergens, putative allergens, or unproven proteins and provide the technical allergy review. R.G. is the database manager with overall responsibility for ensuring the reviews and records are complete.

Contract support for the allergen database sponsors paid partial salaries for R.G., J.W., and S.L. The external panelists (B.B., F.F., H.S., M.E., S.V., R.vR.) receive an honorarium and expenses for attending the annual expert panel meeting. S.V. has received grant funding previously for scientific studies from Monsanto and Pioneer. R.G. has received grant funding for previous scientific studies from BASF, Bayer, Monsanto, and Pioneer. Expert panel members are not assigned as primary or secondary reviewers of publications or entries they coauthored. Authors are screened by R.G. during triage. Reviews are recorded for each expert and archived. All experts perform final votes on all entries, which are recorded.

The authors have declared no conflict of interest.

9 References

- [1] Sicherer, S. H., Epidemiology of food allergy. *J. Allergy Clin. Immunol.* 2011, 127, 594–602.
- [2] Rona, R. J., Keil, T., Summers, C., Gislason, D. et al., The prevalence of food allergy: a meta-analysis. *J. Allergy Clin. Immunol.* 2007, 120, 638–646.
- [3] Cochrane, S., Beyer, K., Clausen, M., Wjst, M. et al., Factors influencing the incidence and prevalence of food allergy. *Allergy* 2009, 64, 1246–1255.
- [4] Roses, J. B., Food allergen law and the Food Allergen Labeling and Consumer Protection Act of 2004: falling short of true protection for food allergy sufferers. *Food Drug Law J.* 2011, 66, 225–242.
- [5] Metcalfe, D. D., Astwood, J. D., Townsend, R., Sampson, H. A. et al., Assessment of the allergenic potential of foods derived from genetically engineered crop plants. *Crit. Rev. Food Sci. Nutr.* 1996, 36S, S165–S186.
- [6] Nordlee, J. A., Taylor, S. L., Townsend, J. A., Thomas, L. A. et al., Identification of a Brazil-nut allergen in transgenic soybeans. *N. Engl. J. Med.* 1996, 334, 688–692.
- [7] Goodman, R. E., Vieths, S., Sampson, H. A., Hill, D. et al., Allergenicity assessment of genetically modified crops—what makes sense? *Nat. Biotechnol.* 2008, 26, 73–81.
- [8] Goodman, R. E., Biosafety: evaluation and regulation of genetically modified (GM) crops in the United States. *J. Huazhong Agric. Univ.* 2014, 33, 85–114.
- [9] Goodman, R. E., Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. *Food Chem. Toxicol.* 2008, 46, S24–S34.
- [10] Aalberse, R. C., Structural biology of allergens. *J. Allergy Clin. Immunol.* 2000, 106, 228–238.
- [11] Fernandez-Rivas, M., Fruit and vegetable allergy. *Chem. Immunol. Allergy* 2015, 101, 162–170.
- [12] ILSI-IFBC, Supplemental issue on the safety of foods and food ingredients developed through the application of genetic modification. *Reg. Toxicol. Pharma.* 1990, 12, 1–196.
- [13] FDA., Department of Health and Human Services: Food and Drug Administration (FDA): Statement of Policy: Foods Derived from New Plant Varieties. Federal Register Vol. 57, No. 104. Rockville, FDA, 1992.
- [14] CODEX Alimentarius Commission., Proposed draft annex on the assessment of possible allergenicity of the draft guideline for the conduct of food safety assessment of foods derived from recombinant DNA plants. Joint FAO/WHO Food Standard Program, CAC/GL 45–2003, Appendix IV; 2003, pp. 57–60.
- [15] EFSA Panel on Genetically Modified Organisms (GMO), Draft scientific opinion on the assessment of allergenicity of GM plants and microorganisms and derived food and feed. *EFSA J.* 2010, 8, 1700, Appendix 3.
- [16] Astwood, J. D., Fuchs, R. L., Lavric, P. B., in: Metcalfe, D. D., Sampson, H. A., Simmon, R., *Food Allergy*, Chapter 4, 2nd edn. Blackwell Sciences Inc., Cambridge, MA, USA 1997, pp. 65–92.
- [17] Goodman, R. E., Hefle, S. L., Taylor, S. L., van Ree, R., Assessing genetically modified crops to minimize the risk of increased food allergy: a review. *Int. Arch. Allergy Immunol.* 2005, 137, 153–166.
- [18] Goodman, R. E., Practical and predictive bioinformatics methods for the identification of potentially cross-reactive protein matches. *Mol. Nutr. Food Res.* 2006, 50, 655–660.
- [19] van Ree, R., Vieths, S., Poulsen, L. K., Allergen-specific IgE testing in the diagnosis of food allergy and the event of a positive match in the bioinformatics search. *Mol. Nutr. Food Res.* 2006, 50, 645–654.
- [20] Barrio, A. M., Soeria-Atmadja, D., Nister, A., Gustafsson, M. G. et al., EVALLER: a web server for in silico assessment of potential protein allergenicity. *Nuc. Acids Res.* 2007, 35, 694–700.
- [21] Oezguen, N., Zhou, B., Negi, S. S., Ivanciuc, O. et al., Comprehensive 3D-modeling of allergenic proteins and amino acid composition of potential conformational IgE epitopes. *Mol. Immunol.* 2008, 45, 3740–3747.

- [22] Dimitrov, I., Bangov, I., Flower, D. R., Doytchinova, I. Allertop v 2.—a server for in silico prediction of allergens. *J. Mol. Model.* 2014, 20, 2278–6 DOI 10.1007/s00894-014-2278-5.
- [23] Prosperi, M. C., Belgrave, D., Buchan, I., Simpson, A. et al., Challenges in interpreting allergen microarrays in relation to clinical symptoms: a machine learning approach. *Pediatr. Allergy Immunol.* 2014, 25, 71–79.
- [24] Pascal, M., Grishina, G., Yang, A. C., Sanchez-Garcia, S. et al., Molecular diagnosis of shrimp allergy: efficiency of several allergens to predict clinical reactivity. *J. Allergy Clin. Immunol. Pract.* 2015, 3, 521–529.
- [25] CODEX Alimentarius Commission. *Foods derived from modern biotechnology*, 2nd ed. FAO and WHO of the United Nations. CAC/GL 55–2003, CAC/GL 45–2003, CAC/GL 46–2003 and CAC/GL 68–2008. FAO Viale delle Terme di Caracalla, Rome, Italy 2009.
- [26] EFSA. European Food Safety Authority Draft Scientific opinion on the assessment of allergenicity of GM plants and microorganisms and derived food and feed 2009.
- [27] Goodman, R. E., Tetteh, A. O., Suggested improvements for the allergenicity assessment of genetically modified plants used in foods. *Curr. Allergy Asthma Rep.* 2011, 11, 317–324.
- [28] Cressman, R. F., Ladics, G., Further evaluation of the utility of “sliding window” FASTA in predicting cross-reactivity with allergenic proteins. *Regul. Toxicol. Pharmacol.* 2008, 54, S20–S25.
- [29] Silvanovich, A., Bannon, G., McClain, S., The use of E-scores to determine the quality of protein alignments. *Regul. Toxicol. Pharmacol.* 2009, 54, S26–S31.
- [30] Hoff, M., Son, D. Y., Gubesch, M., Ahn, K. et al., Serum testing of genetically modified soybeans with special emphasis on potential allergenicity of the heterologous protein CP4 EPSPS. *Mol. Nutr. Food Res.* 2007, 51, 946–955.
- [31] Holzhauser, T., van Ree, R., Poulsen, L. K., Bannon, G. A., Analytical criteria for performance characteristics of IgE binding methods for evaluating safety of biotech food products. *Food Chem. Toxicol.* 2008, 46, S15–S19.
- [32] King, T. P., Hoffman, D., Lowenstein, H., Marsh, D. G. et al., Allergen nomenclature. *Allergy* 1995, 50, 765–774.
- [33] Hileman, R. E., Silvanovich, A., Goodman, R. E., Rice, E. A. et al., Bioinformatics methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.* 2002, 128, 280–291.
- [34] Sampson, H. A., Food allergy—accurately identifying clinical reactivity. *Allergy* 2005, 60, 19–24.
- [35] Bock, S. A., Munoz-Furlong, A., Sampson, H. A., Further fatalities caused by anaphylaxis to foods (Letter to the editor). *J. Allergy Clin. Immunol.* 2007, 119, 1016–1018.
- [36] Palmer, G. W., Dibbern, D. A., Burks, A. W., Bannon, G. A. et al., Comparative potency of Ara h 1 and Ara h 2 in immunochemical and functional assays of allergenicity. *Clin. Immunol.* 2005, 115, 302–312.
- [37] Pascal, M., Munoz-Cano, R., Reina, Z., Palacin, A. et al., Lipid transfer protein syndrome: clinical pattern, cofactor effect and profile of molecular sensitization to plant-foods and pollens. *Clin. Exp. Allergy* 2013, 42, 1529–1539.
- [38] Le, T.-M., van Hoffen, E., Kummeling, I., Potts, J. et al., Food allergy in the Netherlands: differences in clinical severity, causative foods, sensitization and DBPCFC between community and outpatients. *Clin. Transl Allergy* 2015, 5, 8, DOI: 10.1186/s.
- [39] Worm, M., Eckermann, O., Dolle, S., Aberer, W. et al., Triggers and treatment of anaphylaxis: an analysis of 4,000 cases from Germany, Austria and Switzerland. *Dtsch. Arztebl. Int.* 2014, 111, 367–375.
- [40] Wong, G. W., Mahesh, P. A., Ogorodova, L., Leung, T. F. et al., The EuroPrevall-INCO surveys on the prevalence of food allergies in children from China, India and Russia: the study of methodology. *Allergy* 2010, 65, 385–390.
- [41] Incorvaia, C., Mauro, M., Ridolo, E., Makri, E. et al., A pitfall to avoid when using an allergen microarray: the incidental detection of IgE to unexpected allergens. *J. Allergy Clin. Immunol. Pract.* 2015, 3, 879–882.
- [42] Palmer, G. W., Dibbern, D. A. Jr., Burks, A. W., Bannon, G. A. et al., Comparative potency of Ara h 1 and Ara h 2 in immunochemical and functional assays of allergenicity. *Clin. Immunol.* 2005, 115, 302, 312.
- [43] Vickery, B. P., Scurlock, A. M., Jones, S. M., Burks, A. W., Mechanisms of immune tolerance relevant to food allergy. *J. Allergy Clin. Immunol.* 2011, 127, 576–584.
- [44] Rosenfeld, L., Shreffler, W., Bardina, L., Niggemann, B. et al., Walnut allergy in peanut-allergic patients: significance of sequential epitopes of walnut homologous to linear epitopes of Ara h 1, 2 and 3 in relation to clinical reactivity. *Int. Arch. Allergy Immunol.* 2012, 157, 238–245.
- [45] Pastorello, E. A., Ortolani, C., Baroglio, C., Pravettoni, V. et al., Complete amino acid sequence determination of the major allergen of peach (*Prunus persica*) Pru p 1. *Biol. Chem.* 1999, 380, 1315–1320.
- [46] Fernandez-Rivas, M., Benito, C., Gonzalez-Mancebo, E., de Durana, D. A., Allergies to fruits and vegetables. *Pediatr. Allergy Immunol.* 2008, 19, 675–681.
- [47] Pascal, M., Munoz-Cano, R., Reina, Z., Palacin, A. et al., Lipid transfer protein syndrome: clinical, cofactor effect and profile of molecular sensitization of plant-foods and pollens. *Clin. Exp. Allergy* 2012, 42, 1529–1539.
- [48] Asero, R., Lipid transfer protein cross-reactivity assessment in vivo and in vitro in the office: pros and cons. *J. Investig. Allergol. Clin. Immunol.* 2011, 21, 129–136.
- [49] Swoboda, I., Jilek, A., Ferreira, F., Engel, E. et al., Isoforms of Bet v 1, the major birch pollen allergen, analyzed by liquid chromatography, mass spectrometry, and cDNA cloning. *J. Biol. Chem.* 1995, 270, 2607–2613.
- [50] Erdmann, S. M., Sachs, B., Schmidt, A., Merk, H. F. et al., In vitro analysis of birch-pollen-associated food allergy by use of recombinant allergens in the basophil activation test. *Int. Arch. Allergy Immunol.* 2005, 136, 230–238.
- [51] Thomas, W. R., Hierarchy and molecular properties of house dust mite allergens. *Allergol. Intl.* 2015, 64, 304–311.
- [52] Goodman, R. E., Leach, J. N., Assessing the allergenicity of proteins introduced into genetically modified crops

- using human IgE assays. *J. AOAC Int.* 2004, *87*, 1423–1432.
- [53] Orovitg, A., Guardia, P., Barber, D., de la Torre, F. et al., Enhanced diagnosis of pollen allergy using specific immunoglobulin E determination to detect major allergens and pan allergens. *J. Investig. Allergol. Clin. Immunol.* 2011, *21*, 253–259.
- [54] Ferreira, F., Wolf, M., Wallner, M., Molecular approach to allergy diagnosis and therapy. *Yonsei Med. J.* 2014, *55*, 839–852.
- [55] Chapman, M. D., Wuenschmann, S., King, E., Pomes, A., Technological innovations for high-throughput approaches to in vitro allergy diagnosis. *Curr. Allergy Asthma Rep.* 2015, *15*, 36.
- [56] Cramer, R., Immunoglobulin E-binding autoantigens: biochemical characterization and clinical relevance. *Clin. Exp. Allergy* 2012, *42*, 343–351.
- [57] Dema, B., Pellefigues, C., Hasni, S., Gault, N. et al., Autoreactive IgE is prevalent in systematic lupus erythematosus and is associated with increased disease activity and nephritis. *PLoS One* 2014, *9*: e90424, doi: 10.1371/journal.pone.0090424.
- [58] Gavalá, M. L., Bashir, H., Gern, J. E., Virus/allergen interactions in asthma. *Curr. Allergy Asthma Rep.* 2013, *13*, 298–307.
- [59] Agrawal, R., Wisniewski, J., Yu, M. D., Kennedy, J. L. et al., Infection with human rhinovirus 16 promotes enhanced IgE responsiveness in basophils of atopic asthmatics. *Clin. Exp. Allergy* 2014, *44*, 1266–1273.
- [60] Zhang, L., Liu, J., Wang, E., Zeng, S. et al., Respiratory syncytial virus protects against the subsequent development of ovalbumin-induced allergic responses by inhibiting Th2-type $\gamma\delta$ T cells. *J. Med. Virol.* 2013, *85*, 149–156.
- [61] Rossi, G. A., Colin, A. A., Infantile respiratory syncytial virus and human rhinovirus infections: respective role in inception and persistence of wheezing. *Eur. Respir. J.* 2015, *45*, 774–789.
- [62] di Somma, C., Fiore, L., Di Lonardo, A., Ridolfi, B. et al., Cross-reactivity between the major *Parietaria* allergen and rotavirus VP4 protein. *Allergy* 2003, *58*, 503–510.
- [63] Kelso, J. M., Mootrey, G. T., Tsai, T. F., Anaphylaxis from yellow fever vaccine. *J. Allergy Clin. Immunol.* 1999, *103*, 698–701.
- [64] Farouk, S. E., Dolo, A., Berczy, S., Kouniba, B. et al., Different antibody and cytokine-mediated responses to *Plasmodium falciparum* parasite in two sympatric ethnic tribes living in Mali. *Microbes Infect.* 2005, *7*, 110–117.
- [65] Halstensen, T. S., Why are we not all allergic: basic mechanisms for tolerance development. *Environ. Toxicol. Pharmacol.* 1997, *4*, 25–31.
- [66] Arlian, L. G., Morgan, M. S., Quirce, S., Maranon, F. et al., Characterization of allergens in *Anisakis simplex*. *Allergy* 2003, *58*, 1299–1303.
- [67] Hofer, M. F., Harbeck, R. J., Schlievert, P. M., Leung, D. Y., Staphylococcal toxins augment specific IgE responses by atopic patients exposed to allergen. *J. Invest. Dermatol.* 1999, *112*, 171–176.
- [68] Wehner, J., Neuber, K., Staphylococcus aureus enterotoxins induce histamine and leukotriene release in patients with atopic eczema. *Br. J. Dermatol.* 2001, *145*, 302–305.
- [69] Becker, Y., Respiratory syncytial virus (RSV) evades the human adaptive immune system by skewing the Th1/Th2 cytokine balance toward increased levels of Th2 cytokines and IgE, markers of allergy—a review. *Virus Genes* 2006, *33*, 235–252.
- [70] Sampson, H. A., Update on food allergy. *J. Allergy Clin. Immunol.* 2004, *113*, 805–819.
- [71] Sicherer, S. H., Sampson, H. A., Food allergy: epidemiology, pathogenesis, diagnosis and treatment. *J. Allergy Clin. Immunol.* 2014, *133*, 291–307.
- [72] Commins, S. P., Satinover, S. M., Hosen, J., Mozena, J. et al., Delayed anaphylaxis, angioedema, or urticaria after consumption of red meat in patients with IgE antibodies specific for galactose- α -1,3-galactose. *J. Allergy Clin. Immunol.* 2009, *123*, 426–433.
- [73] Sampson, H. A., Historical background, definitions and differential diagnosis. *Chem. Immunol. Allergy* 2015, *101*, 1–7.
- [74] Santos, A. F., Du Toit, G., Douiri, A., Badulovic, S. et al., Distinct parameters of the basophil activation test reflect the severity and threshold of allergic reactions to peanut. *J. Allergy Clin. Immunol.* 2015, *135*, 179–186.
- [75] Mari, A., Ooievaar-de Heer, P., Scala, E., Giani, M. et al., Evaluation by double-blind placebo-controlled oral challenge of the clinical relevance of IgE antibodies against plant glycans. *Allergy* 2008, *63*, 891–896.
- [76] Westphal, S., Kolarich, D., Foetisch, K., Lauer, I. et al., Molecular characterization and allergenic activity of Lyc e 2 (β -fructofuranosidase), a glycosylated allergen of tomato. *Eur. J. Biochem.* 2003, *270*, 1327–1337.
- [77] Bublin, M., Dennstedt, S., Buchegger, M., Antonietta Ciardiello, M. et al., The performance of a component-based allergen microarray for the diagnosis of kiwifruit allergy. *Clin. Exp. Allergy* 2011, *41*, 129–136.
- [78] Siruguri, V., Bharatraj, D. K., Vankudavath, R. N., Rao Mendu, V. V. et al., Evaluation of Bar, Barnase and Barstar recombinant proteins expressed in genetically engineered *Brassica juncea* (Indian mustard) for potential risks of food allergy using bioinformatics and literature searches. *Food Chem. Toxicol.* 2015, *83*, 93–102.
- [79] Wang, F., Robotham, J. M., Teuber, S. S., Sathe, S. K. et al., Ana o 2, a major cashew (*Anacardium occidentale* L.) nut allergen of the legumin family. *Int. Arch. Allergy Immunol.* 2003, *132*, 27–39.
- [80] Moneo, I., Caballero, M. L., Gonzalez-Munoz, M., Rodriguez-Mahillo, A. I. et al., Isolation of a heat-resistant allergen from fish parasite *Anisakis simplex*. *Parasitol. Res.* 2005, *96*, 285–289.
- [81] Rodriguez-Mahillo, A. I., Gonzalez-Munoz, M., Gomez-Aguado, F., Rodriguez-Perez, R. et al., Cloning and characterization of the *Anisakis simplex* allergen Ani s 4 as a cysteine-protease inhibitor. *Int. J. Parasitol.* 2007, *37*, 907–917.

- [82] Blank, S., Seismann, H., Bockisch, B., Braren, I. et al., Identification, recombinant expression, and characterization of the 100 kDa high molecular weight hymenoptera venom allergens Api m 5 and Ves v 3. *J. Immunol.* 2010, *184*, 5403–5413.
- [83] Chua, K. Y., Doyle, C. R., Simpson, R. J., Turner, K. J. et al., Isolation of cDNA coding for the major mite allergen Der p II by IgE plaque immunoassay. *Int. Arch. Allergy Appl. Immunol.* 1990, *91*, 118–123.
- [84] An, S., Chen, L., Long, C., Liu, X. et al., Dermatorphagoides farina allergens diversity identification by proteomics. *Mol. Cell Proteomics* 2013, *12*, 1818–1828.
- [85] Postigo, I., Gutierrez-Rodriguez, A., Fernandez, J., Guisantes, J. A. et al., Diagnostic value of Alt a 1, fungal enolase and manganese-dependent superoxide dismutase in the component-resolved diagnosis of allergy to Pleosporaceae. *Clin. Exp. Allergy* 2011, *41*, 443–451.
- [86] Gabriel, M. F., Postigo, I., Gutierrez-Rodriguez, A., Sunen, E. et al., Characterisation of *Alternaria alternata* manganese-dependent superoxide dismutase, a cross-reactive allergen homologue of Asp f 6. *Immunobiology* 2015, *220*, 851–858.
- [87] Mattsson, L., Lundgren, T., Olsson, P., Sundberg, M. et al., Molecular and immunological characterization of Can f 4: a dog dander allergen cross-reactive with a 23 kDa odorant-binding protein in cow dander. *Clin. Exp. Allergy* 2010, *40*, 1276–1287.
- [88] Niemi, M. H., Rytönen-Nissinen, M., Janis, J., Virtanen, T. et al., Structural aspects of dog allergies: the crystal structure of a dog dander allergen Can f 4. *Mol. Immunol.* 2014, *61*, 7–15.
- [89] Irifune, K., Nishida, T., Egawa, H., Nagatani, A., Pectin methylesterase inhibitor cDNA from kiwi fruit. *Plant Cell Rep.* 2004, *23*, 333–338.
- [90] Kumari, D., Arora, N., Kasera, R., Sridhara, S. et al., Isolation and characterization of a 28 kDa major allergen from blackgram (*Phaseolus mungo*). *Immunobiology* 2012, *217*, 895–904.