

Proteomic and bioinformatic analyses for the identification of proteins with low allergenic potential for hazard assessment

Nora L. Krutz^{*1}, Jason Winget², Cindy A. Ryan², Rohan Wimalasena², Sebastian Maurer-Stroh³, Rebecca J. Dearman⁴, Ian Kimber⁴, G. Frank Gerberick^{2,5}

¹NV Procter & Gamble Services Company SA, Strombeek-Bever, Belgium.

²The Procter & Gamble Company, Mason, OH, United States.

³Bioinformatics Institute, Agency for Science, Technology and Research, Singapore and Department of Biological Sciences, National University of Singapore, Singapore.

⁴University of Manchester, the Faculty of Biology, Medicine and Health, Manchester, United Kingdom.

⁵Present address: GF3 Consultancy LLC, West Chester, OH, United States.

***corresponding author:** Dr. Nora L. Krutz

NV Procter & Gamble Services Company SA

Temselaan 100, 1853 Strombeek-Bever, Belgium

Phone: +32 245 66040; Email: krutz.n@pg.com

Keywords: Type I Allergy, Risk Assessment, label-free proteomic analysis, *in silico* prediction model, AllerCatPro

Abstract

Use of botanicals and natural substances in consumer products has increased in recent years. Such extracts can contain protein that may theoretically represent a potential risk of IgE-mediated allergy. No method has yet been generally accepted or validated for assessment of the allergenic potential of proteins. For development of suitable methods datasets of allergenic and non-allergenic (or low allergenic) proteins are required that can serve, respectively, as positive and negative controls. However, data are unavailable on proteins that lack or have low allergenic potential. Here, low allergenic potential proteins are identified based upon the assumption that proteins with established human exposure, but with a lack of an association with allergy, possess low allergenic potential. Proteins were extracted from sources considered to have less allergenic potential (corn, potato, spinach, rice and tomato) as well as higher allergenic potential (wheat) regarding common allergenic foods. Proteins were identified and semi-quantified by label-free proteomic analysis conducted using mass spectrometry. Predicted allergenicity was determined using AllerCatPro (<https://allercatpro.bii.a-star.edu.sg/>). In summary, 9077 proteins were identified and semi-quantified from six protein sources. Within the top 10% of the most abundant proteins identified, 178 characterized proteins were found to have no evidence for allergenicity predicted by AllerCatPro and were considered to have low allergenic potential. This panel of low allergenic potential proteins provides a pragmatic approach to aid the development of alternative methods for robust testing strategies to distinguish between proteins of high and low allergenic potential to assess the risk of proteins from natural or botanical sources.

Introduction

Botanicals and natural substances are of increasing interest for consumer products since they are naturally occurring and can confer desirable properties. Such extracts contain many proteins and some of these may have a potential to cause, IgE antibody-mediated (Type I) allergy, which can affect the respiratory tract (Blackburn et al. 2015). Methods and tools have been developed to determine the allergenic potential of proteins, with the focus recently on foods containing transgene products (Remington et al. 2018). However, none of these approaches has been widely accepted or validated (Verhoeckx et al. 2016).

One of the challenges in developing new approaches for the identification of allergenic hazards of proteins is the lack of a dataset that can serve as a ‘gold standard’ with positive and negative controls. Without such controls it is difficult to calibrate any proposed system to determine the sensitivity, selectivity and overall accuracy of the method. While the identification of allergenic proteins is based on well-defined criteria (e.g. Pomes et al. 2018), there are no commonly accepted criteria or strategies for identifying non-allergenic proteins that lack inherent allergenic potential. Although it is understood that some proteins have a stronger allergenic potential than others, available experimental data on proteins that are considered to lack allergenic potential, or at least to have only low allergenic activity, are very limited in the literature and comprehensive databases that are in general focused on listing proteins associated with allergenicity, but not the lack of allergenic potential (e.g. Allergome, <http://www.allergome.org/>). Moreover, it is inappropriate to include as negative controls for allergenicity proteins of human origin, which would preclude the complexity of interindividual tolerizing mechanisms and autoreactive IgE antibodies (Aalberse 2000). Generating non-allergenic protein datasets purely based on protein sequences in UniProt (<http://www.uniprot.org/>) that lack the annotation as allergen (e.g.

Dimitrov et al. 2014; Lu et al. 2018), is considered as less persuasive as well, due to the lack of evidence for human exposure.

To generate an initial robust set of proteins of low allergenic potential, we here generated evidence for abundant proteins for which there is no evidence for allergenic activity, despite opportunities for relatively high levels of human exposure. Abundant proteins are considered as those that are within the top 10% of all identified proteins extracted and semi-quantified from one protein source and each of them represent a minimum of 0.1% relative abundance of the total protein level. We selected protein sources that are steadily consumed in high amounts, but are reported less frequently with regard to allergic reactions than the ‘big 8’ major foods or food groups that account for 90% of all food allergies (Food And Drug Administration 2004). Corn (or Maize, *Zea mays*), potato (*Solanum tuberosum*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), and spinach (*Spinacia oleracea*) were chosen for analyses as protein sources that are considered to have less allergenic potential as well as one of the ‘big 8’ common food allergens, wheat (*Triticum aestivum*).

To identify the most abundant proteins within these prevalent protein sources, we applied label-free proteomic analyses. Abundant proteins were considered to have low allergenic potential, if their protein sequence and their predicted 3D structure showed no similarity to any known protein allergen, including respiratory allergens. The criteria for ‘no evidence for allergenic potential’ can be predicted by the recently developed AllerCatPro model (<https://allercatpro.bii.a-star.edu.sg/>; Maurer-Stroh et al. 2019).

The purpose of this article is to provide a list of proteins with low allergenic potential and describe the approach used to identify them. Based on our observations and investigations by proteomic and bioinformatic (AllerCatPro) analyses, we believe that it is now possible to provide

legitimate proteins with evidence of low allergenic potential as controls that can be used with confidence for the calibration of protein safety assessment methods for IgE-mediated allergenic potential.

Material and Methods

Selection of protein sources for the identification of low allergenic proteins

Based on literature research, we selected protein sources with rather rare cases of allergic reactions, but with reported high consumption (Knudsen et al. 2008; Lomnitski et al. 2003) including corn (*Zea mays*), potato (*Solanum tuberosum*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), and spinach (*Spinacia oleracea*). Wheat (*Triticum aestivum*), one of the ‘big 8’ foods, was included as a source of positive allergens. Organic corn meal, medium grind (made from 100% whole grain with germ and bran), organic potato flour (made from whole peeled dehydrated russet potatoes) and organic whole grain brown rice flour (contains bran) were obtained from Bob’s Red Mill and purchased via Amazon (Received June 2016). Freeze-dried tomato chunks (with peel) was obtained from Emergency Essentials (Received June 2016) and freeze-dried spinach from Nuts.com (Received May 2016). Wheat was purchased as organic stone ground whole wheat pastry flour from Bob’s Red Mill.

Label-free proteomic analysis

Samples of spinach and tomato (pulverized under liquid nitrogen to yield powder) and corn, potato, rice, and wheat flours were analyzed to determine the protein content via Amino Acid Analysis at Molecular Structure Facility, University of California, Davis using Hitachi analyzers and ion-exchange chromatography (detection limit: ~100 pM). Raw materials (each nominally

500 µg total protein) were suspended in 10:1 v/w quantity of SDT buffer (4% Sodium dodecyl sulfate (SDS), 100 mM dithiothreitol (DTT) and 100 mM Tris, pH 7.6). A standard protein mixture containing beta-lactoglobulin A, myoglobin, bovine serum albumin, and lysozyme (obtained from Sigma as powders) was initially used to ensure that the quantification variance in different matrices was stable (data not shown). Samples were sonicated (5 min), heated at 95°C (5 min), and sonicated (5 min). Digestion enzymes Trypsin and LysC (obtained from Promega) were added during the Filter-Aided Sample Preparation (Wisniewski et al. 2009). Peptides were desalted on MacroSpin columns (NEST group). Dried peptides were reconstituted to a nominal concentration of 1 µg/µl in 0.1% Formic acid in water and 5 µl were injected onto a 75 µm x 20 cm EasySpray column packed with 3 µm C18 (Thermo) driven by an Acquity M-class LC (Waters) at 300 nl/min. The gradient proceeded from 97% Buffer A (0.1% Formic acid in water) to 30% Buffer B (0.1% Formic acid in acetonitrile) over 240 min. Eluted peptides were introduced to a QExactive HF mass spectrometer (Thermo) operating in data-dependent mode with settings as indicated in Kelstrup et al. (2012).

Protein databases for the various materials were downloaded from UniProt (<http://www.uniprot.org/>). When available, a defined ‘proteome’ (according to UniProt curation) was used. Otherwise, all sequences matching a search for the genus/species were included. Protein databases were each then clustered at 90% identity using the UniRef approach (Suzek et al. 2015) to reduce redundant entries. A false positive rate of protein sequences was estimated with false (decoy) sequences using pseudo-randomization with the decoyFastaGenerator functionality of the Trans-Proteomic Pipeline (TPP), an open-source (<https://sourceforge.net/projects/sashimi/>) proteomic MS/MS analysis platform utilizing XML file formats (v. 5.1.0 “SYZYGY”, Build 201711031215-7670). Spectra were searched against

the appropriate database using the Comet search engine. Resulting peptide-spectrum matches were refined and scored using the TPP to yield peptide and inferred protein identifications. Quantification was performed on proteins identified at <1% false positive rate using the StPeter module of the TPP (Hoopmann et al. 2018) and is given as the sum of fragment ion intensities (normalized spectral index, SI_N) for all spectra counted for a protein and normalized by the protein length and the total signal for a given LC-MS/MS run (Sardiu and Washburn 2010).

Bioinformatic analysis of protein allergenicity by AllerCatPro

Identified proteins were analyzed for their predicted allergenicity potential by AllerCatPro (<https://allercatpro.bii.a-star.edu.sg/>), which is described in detail in Maurer-Stroh et al. (2019). This *in silico* model screens protein sequences of interest by their protein sequence and 3D structure level and performs with 84% accuracy at 100% sensitivity to predict allergens versus non-allergens with the same structural fold. Proteins in FASTA format are first screened for their similarity to Gluten-like proteins, then for their similarity on a 3D structure level against a 3D structure database of 713 templates covering the 3D structure of 74% of known allergens and finally for similarity on the protein sequence level towards the most complete dataset of 4180 protein sequences derived from the union of the major databases FARRP, COMPARE, WHO/IUIS, UniProtKB and Allergome.

A protein is predicted with ‘strong evidence’ for allergenicity, if it is either similar to gluten (based on a Gluten-like repeat pattern detection) or its most similar 3D structure surface epitope is more than 93% identical to the corresponding 3D region of a known allergen or its protein sequence is either more than 35% identical to known allergens in a linear window of 80 amino acids or a sequence identity of short three or more hexamer hits [3 x 6-mer sequence identity

rule, adapted from (FAO/WHO) and described in Maurer-Stroh et al. (2019)] to known allergens was found. A protein is predicted with ‘weak evidence’ for allergenicity if it shares the same fold with sequence similarity up to 93% in 3D structure epitopes of protein allergens. Proteins are predicted with ‘no evidence’ for allergenic potential if the protein sequence does not trigger a hit for similarity to Gluten-like proteins, similarity to known allergens on a 3D structure level and on a sequence level above 35% sequence identity to known allergens within the 80-amino acid window as well as a sequence identity of at least three short hexamers with known allergens (3 x 6-mer sequence identity rule).

Results

1) Low allergenic proteins retrieved from literature and Allergome

A protein can be considered with low allergenic potential if there is evidence for the lack of inherent allergenic potential. Since there are neither validated methods nor commonly accepted criteria for identifying non-allergenic proteins, we searched in the literature and Allergome for proteins that were considered to lack or have low allergenic potential and were used *in vitro* and/or *in vivo* experiments.

Based on the literature research, only 11 proteins were found that were used as either negative controls or were considered to have low allergenic potential based on different experiments focusing on various endpoints in (**Table 1**). In the following, the reasons for including each of the proteins are described in more detail.

RuBisCO (Ribulose-1,5 biphosphate carboxylase/oxygenase) is one of the most abundant proteins on earth (Raven 2013) and consumed in high amounts by oral uptake or inhalation due to smoking of tobacco (Geada et al. 2007). It may cross-react nonspecifically with IgE in humans

(Foti et al. 2012; Hoff et al. 2007; Nayak et al. 2013), but does not induce specific IgG1 nor IgE in mice (Smit et al. 2016). Potato agglutinin is considered to lack allergenic potential (Jeannet-Peter et al. 1999) since it fails to stimulate detectable IgE response in mice after 14 days (Dearman and Kimber 2005). Potato acid phosphatase was tested negative to induce IgE in mice after 28 days (Dearman and Kimber 2001) and 42 days in all (Jia et al. 2005) and in 4/6 (Dearman and Kimber 2001) tested mice. Lectin in Kidney beans is considered as a potent mitogen (Ceuppens et al. 1988) reacting rarely with IgE from bean-sensitized individuals (Kasera et al. 2011). The heat stable (Samudzi et al. 1997) Trypsin inhibitor from the Lima bean is considered unlikely to induce allergic reactions in bean-sensitized individuals (Bernhisel-Broadbent and Sampson 1989; Kumar et al. 2006). The often favored negative control Concanavalin A from Jack bean (e.g., Dearman et al. 2003; Foster et al. 2013) is a potent mitogen (Mendez et al. 1998), which may elicit nonspecific histamine release in basophils (Magro and Bennich 1977; Pramod et al. 2007). Lipxygenase from soybean was tested negative for both, the induction of specific IgG1 and IgE in mice after 16 days (Smit et al. 2016) and the production of pro-inflammatory cytokines in a DC-T cell assay (Van Bilsen et al. 2015). The glyphosate-resistant enzyme CP4-EPSPS in soybean was found non-allergenic based on the lack of positive responses in soy-sensitized individuals (Hoff et al. 2007). The mannose-binding tuber agglutinin from wild taro was considered to lack allergenic potential based on results derived by bioinformatic analysis, pepsin digestibility, thermal stability assay, immuno-screening and allergenicity assessment in mice (Das et al. 2018). The only available experiments conducted to assess the allergenic potential of lectin from garden pea and peroxidase from horseradish are based on their instability in simulated gastric fluid (Herman et al. 2007; Thomas et al. 2004).

In summary, too few proteins with low allergenic potential were identified in the published literature. The number of proteins is insufficient to conduct further analyses such as a bioinformatic analyses based on their protein sequence.

Besides the literature search, Allergome (<http://www.allergome.org/>; Mari and Scala 2006), a major database of allergens was searched for proteins annotated as ‘No IgE-binding Antigens’. These proteins are considered having no evidence for allergenicity since they were found either with a negative test result (indicated with ‘-1’) or with no data available (indicated with ‘0’) in allergenicity tests such as skin tests, functional tests (e.g. basophil or mast cell activation tests), non-functional tests (e.g. IgE immunoblotting) as well as epidemiology data from literature. From a total of 31 ‘No IgE-binding Antigens’ in Allergome (last accessed September 2018) only 16 proteins are characterized and linked to a protein sequence (**Table 2**). Most of the proteins are derived from bacteria and, except CP4 EPSPS, none is similar to any of those found in literature listed in **Table 1**. These proteins may represent low allergenic potential, but the confidence is low, since the experiments are based on different approaches and direct human exposure information has not yet been evaluated explicitly.

1) Abundances of proteins in selected protein sources

Due to the limited available data on proteins with evidence to support low allergenic potential, we developed a strategy to identify and semi-quantify proteins from selected protein sources with the opportunity of relatively high human exposure by label-free proteomic analysis and to further analyze the proteins for their allergenic potential by bioinformatic analysis.

In summary, 9077 proteins were identified and semi-quantified in the six analyzed protein sources with a range from 1205 proteins in wheat (average of 1110 proteins within 3 independent

extracts of the same raw material) up to 2009 proteins in corn, and furthermore 1224 in spinach, 1519 in potato, 1245 in rice, and 1875 in tomato. With regard to each protein source, the number of different proteins present at a relative abundance over 0.1% of the measured total protein level reflect less than 10% of all identified proteins but covers at least 76% of the total protein content. In wheat, 77 proteins (and thus 6.4% of all identified proteins) are above 0.1% relative abundance and cover 93% of the total protein content. With regard to the less allergenic protein sources, the number of all identified proteins above 0.1% relative abundance are similar: in corn, 114 proteins (5.7%) cover 86% of the total protein content, in spinach 84 proteins (6.9%) cover 88% of the total protein content, in potato 114 proteins (7.5%) cover 83%, in rice 79 proteins (6.3%) cover 91% of the total protein content, and in tomato, 162 proteins (8.6%) cover 76% of the total protein content. Besides the proteins with significant relative abundance (above 0.1%), a part of the identified protein content was neglected for the further identification of proteins with low allergenic potential due to their low protein abundance. The number of neglected low abundant proteins range between 1140 proteins (spinach) and 1895 proteins (corn). The percentage of low abundant proteins in relation to the total protein content is 7% and 9% in wheat and rice, respectively, 12% to 17% in spinach, corn and potato, and 24% in tomato. Based on the distribution of abundances and the total number of identified and semi-quantified proteins in each protein source, the selection of only those proteins with a relative abundance above 0.1% is considered to reflect the major protein content of each protein source.

2) Protein allergenicity prediction of abundant proteins by bioinformatic results

Protein allergenicity can be predicted based on the similarity approach with already known allergens (4180 unique allergenic protein sequences derived from the union of major databases) by AllerCatPro (Maurer-Stroh et al. 2019). Proteins that are not similar to any known allergens despite opportunities for high levels of human exposure, can be considered as proteins with low allergenic potential. In order to identify low allergenic potential proteins, the sequences (given as cluster ID and as listed in UniProt) of each identified and semi-quantified protein from wheat, corn, spinach, potato, rice, and tomato were predicted for their strong, weak or no evidence for allergenic potential by AllerCatPro (**Suppl. Table 1**).

Considering only the most abundant proteins (above 0.1% in each protein source) reveals different distribution profiles of their predicted allergenic potential (

Figure 1). In detail, spinach provides the least number (30) of proteins with strong or weak evidence for allergenic potential, followed by rice (52), corn and potato (each 59), wheat (64), and tomato (82). In spinach, the most abundant proteins identified belong to RuBisCO (small and large chain), which has been found low allergenic based on different studies (**Table 1**). However, it is annotated as an allergen in databases, such as Allergome and thus, is predicted with allergenic potential by AllerCatPro. With regard to the percentage of the total protein content, proteins with evidence for allergenicity comprise 49% in tomato, followed by corn (58%), potato (60%), spinach (62% or 18% when considering RuBisCO with no evidence for allergenicity), rice (84%), and wheat (88%). the highest number of significantly abundant proteins with no evidence for allergenicity was found in tomato (80), followed by potato, corn (each 55), spinach (54 or 56 with RuBisCO), rice (27), and wheat (13). Similarly, the percentage of abundant

proteins related to the total protein content with no evidence of allergenicity can be summed up to 29% in corn followed by 27% in tomato, 26% in spinach (69%, when considering RuBisCO with no evidence for allergenicity), 23% in potato and only 7% in rice and 4% in wheat. Thus, wheat and rice reveal the lowest number and percentage of abundant proteins with no evidence for allergenicity and a quite high number and the highest percentage of proteins with evidence for allergenicity.

Conclusively, proteins with significant relative abundance and with predicted no evidence of allergenicity can be obtained from all selected protein sources, whereas the least number of those low allergenic potential proteins can be obtained by wheat, a ‘big 8’ allergenic food, followed by rice.

3) *Relative abundances of identified WHO/IUIS allergens*

The results of the distribution of the allergenic potential of the six analyzed protein sources revealed surprisingly numerous proteins predicted with evidence for allergenicity based on either the identification as a known allergen or due to their similarity to known allergens by AllerCatPro. Proteins with a relative abundance above 0.1% that are recognized in detail by WHO/IUIS (<http://www.allergen.org/>) as well-known allergens were identified within the analyzed protein sources based on their UniProtID and are listed in

Table 3. Only WHO/IUIS allergens were listed, since they are recognized and characterized as allergens according to well-defined criteria (e.g. Pomes et al. 2018). Well-known allergens were found with relative abundances below 0.1% and up to 11% (Tri a 30 in wheat). Some protein sequences showed high similarity towards sequences recognized by WHO/IUIS and in line with this, some WHO/IUIS allergens were found in other protein sources, e.g. wheat allergens were found in tomato, rice and potato. Allergens such as, Sola t 2, a known potato allergen, was identified 10 times in potato and 2 times in tomato based on the similarity of its 3d structure with the predicted 3D structure of the identified protein sequences by AllerCatPro. The similarity is given, but the protein sequences are not completely identical: 4 hits have a percent 3D identity below 93% and 1 hit has a percent linear identity within a window of 80 amino acids below 90%. Additionally, 4 more hits were excluded due to their low relative abundance and thus, only 3 Sola t 2 hits are included in

Table 3. Many more sequences with high similarity to allergens as well as those annotated as allergens in Allergome (Data available in supplementary material). Thus, proteomic and bioinformatic analyses can also be used to identify and semi-quantify allergens as well as proteins with great similarity to allergens, which is a powerful feature to understand the allergenic components in complex mixtures.

4) *Proteins considered with low allergenic potential*

Abundant proteins (above 0.1%) that are predicted with no evidence for allergenic potential in AllerCatPro, and thus no 3D structure similarity nor similarity on the amino acid sequence level, are considered as candidates for proteins with low allergenic potential. These candidates (in total 284 proteins) are further adjusted by those proteins that were listed as ‘uncharacterized’ or ‘putative uncharacterized’ in the UniProt database.

lists the in total 178 proteins selected from the analyzed protein sources. In corn, spinach and potato, 54 to 55 significantly abundant proteins are predicted with no evidence of allergenic potential and 31, 36 and 33 proteins were selected based on their characterized status in UniProt. In tomato, of 80 abundant low allergenic potential proteins only 44 were characterized in UniProt. The lowest overall number of low allergenic potential proteins were found in rice with 22 selected from 27 and in wheat with 12 selected from 13 low allergenic potential proteins. The characterized low allergenic proteins comprise 22% in corn, 20% in spinach (excluding the large and small chain of RuBisCO), 18% in potato and tomato and only 7% and 4% in rice and wheat, respectively.

Based on the proposed approach to generate evidence for the absence of allergenic potential and accompanied by confidence for high human exposure, a total of 178 proteins from the six analyzed protein sources were considered as proteins with low allergenic potential.

Discussion

Little information on the identification of proteins with low allergenic potential is available in the published literature but needed to generate a robust set of negative controls (in addition to known allergens) for the development and validation of methods to assess proteins for IgE-mediated (Type I) allergy potential. For IgE-mediated respiratory allergy to proteins, the mechanism is not fully understood and no ‘Gold Standard’ of positive and negative controls can be retrieved from generally accepted *in vivo* studies (Verhoeckx et al. 2016). In order to validate test methods, the accuracy and specificity of the method is crucial, which in turn, requires a robust dataset as ‘Gold Standard’ or at least reliable positive and negative controls. A limited number of positive and negative controls (e.g. skin sensitizers and irritants) has been proposed for testing chemicals for respiratory sensitization potential (Chary et al. 2018), but none have been proposed as controls for protein allergenicity testing.

Based on the literature, a few proteins were used as negative controls in experimental studies or were considered to lack allergenic potential depending on the analyzed endpoint and the applicability domain of the method. However, some of these proteins may need more supporting data for their classification as low allergenic potential proteins. For example, lectin from Garden pea and peroxidase from Horseradish were considered as low allergenic solely based on the criterion of instability towards enzymatic degradation (Herman et al. 2007; Thomas et al. 2004), which is a questionable parameter for allergenicity (Herman and Ladics 2018; Schnell and Herman 2009). Therefore, this criterion alone may not be a strong argument to classify these proteins as low allergenic without any further evidence and moreover, the overall number of proteins found in literature and the Allergome database is insufficient for a robust dataset of ‘negative’ controls.

To generate further evidence, analyses by *in silico* tools can be useful to provide more information on allergenic potential. For *in silico* analysis, the required protein sequence can be obtained by UniProt via the Accession ID. However, detailed information on the protein material used for experiments is frequently not given or easy to reconstruct. To specifically identify the protein material, label-free proteomic analysis helps to identify the protein sequences and aids understanding the exposure to proteins. Databases with protein abundance information are already available, such as PaxDb (pax-db.org). Nevertheless, the number of protein sources with abundance data is limited and focused on plant physiology and thus the whole organism, while the protein sources selected for the work presented here are parts of plants with likely opportunities for high levels of human exposure via consumption and production processes.

Most of the selected protein sources for this work are commonly considered as less allergenic compared to the ‘big 8’; corn, spinach, tomato, potato and rice. Wheat was selected as a highly consumed protein source, which is recognized as one of the ‘big 8’ allergenic food groups (Food And Drug Administration 2004).

The abundance of these protein sources in the environment is reflected in their high production rates within the range of million tons worldwide per year (<http://www.fao.org/faostat/en/#data>; <https://www.statista.com/>, last accessed July 2018). It is arguable that proteins derived from the ‘big 8’, such as wheat, can be considered as low allergenic. Wheat is one of the best characterized protein sources and its allergens have been studied rigorously (Pahr et al. 2012). Moreover, wheat and corn proteins have relevant human respiratory exposure. Rankin (1986) reported that symptoms of cough and/or expectoration are elicited at low levels from grain dust such as wheat and least likely from corn, which is in line with the recommended occupational

exposure limits of 2 µg/m³ for wheat grain dust and 7.5 µg/m³ for the less allergenic corn grain dust (Blackburn et al. 2015).

The difference in allergenic potential of these two protein sources is also reflected in the results generated by proteomic and bioinformatic (AllerCatPro) analyses: Wheat was found to contain more abundant proteins that are predicted with strong evidence for allergenicity compared to corn. Rice had a distribution profile of protein abundances and predicted allergenic potential more similar to wheat than to any other analyzed protein source. The higher number of proteins predicted with strong or weak evidence of allergenicity in rice is in accordance with the numerous allergens described in more detail recently (Satoh et al. 2016; Satoh et al. 2019) and the substantial number of cases of allergic reactions in patients in Japan (Ikezawa et al. 1992) and Madagascar (Ramavovololona et al. 2014).

Proteomic analyses open new paths to determine the relative abundance of major allergens in relation to minor allergens as well as putative non-allergens and unknown proteins, which is useful for risk assessment. The relative abundances of the identified proteins with strong evidence for allergenicity including allergens specifically recognized by WHO/IUIS (

Table 3) or by Allergome (Data available in supplementary material), is presented here for the first time for the protein sources selected for this work. While this data is not sufficient to assess allergenic potency, it may provide confidence to evaluate allergenic potential and determine exposure. In this context it should be noted that the here described proteomic analysis has its typical limitations with regard to correctly identifying and semi-quantifying e.g. highly modified proteins. Furthermore, the results of the bioinformatics analysis are dependent on the quality of the defined ‘proteome’ in UniProt. Besides these limitations, this investigation demonstrates that proteomic analysis can identify and semi-quantify with precision single proteins with their protein sequence in a complex mixture, which in turn enables bioinformatic analysis (e.g. AllerCatPro) to predict the allergenic potential of the individual proteins as well as their protein sources.

AllerCatPro, novel computational workflow, analyzes the sequence and predicted 3D structure of proteins for a comprehensive prediction of allergenic potential which can be used to improve the risk assessment process for Type I allergy (Maurer-Stroh et al. 2019). We used the UniRef cluster ID that represents the most reliable identified protein sequence for a certain identified protein within a sequence identity of 90%. The so called UniRef90 sequence may be represented by another species than the analyzed source, which in turn makes it difficult to find e.g. WHO/IUIS allergens by their UniProt ID. Here, AllerCatPro matches automatically the input, an UniRef90 sequence, with the most similar protein sequence within the internal database of 4180 protein allergens.

By using AllerCatPro in combination with label-free proteomic analysis on the selected protein sources with opportunities for high levels of human exposure, we are able to generate evidence for proteins and their allergenic potential and confidence to assess proteins with having low

allergenic potential. Of the resulting 284 proteins with significant abundance and ‘no evidence’ for allergenicity, 178 proteins were selected, because these were not annotated as ‘uncharacterized protein’ in the UniProt database. Some of the 178 proteins have not been reviewed yet and some UniProt IDs may be considered as obsolete, since UniProt is curating its database constantly (The UniProt Consortium 2017). The reviewed proteins by UniProt can be consolidated to be part of the stress response and plant defense in corn and wheat, biosynthesis, transport and transfer in tomato, transport, protein storage and stress response in rice, transport and production of energy in spinach and a variety of binding-proteins, enzymes and enzyme inhibitors as well as transfer proteins in potato.

The results generated here on the allergenic potential of proteins with opportunities for significant human exposure open new fields of research for elucidating factors that make a protein less allergenic than major well-known allergens. It is not well understood if low allergenic potential proteins also lack immunogenic potential and if they are less similar to human proteins or share specific characteristics with regard to bioavailability and stability compared to known allergens. However, if these low allergenic potential proteins are used as ‘negative’ controls in experiments other than *in silico* then it is recommended that the individual proteins be isolated from protein sources that are less allergenic than the ‘big 8’ to mitigate the risk of allergenic impurities.

We here provide a pragmatic approach to identify proteins with low allergenic potential by proteomic and bioinformatic analyses. Based on our observations and investigations, we believe that it is appropriate to use the 178 selected low allergenic potential proteins as negative controls to aid the development and validation of alternative methods for more robust testing strategies to

assess the risk of novel proteins from natural or botanical sources for IgE-mediated respiratory allergy.

Declaration of Interest

The author's affiliation is as shown on the cover page. The authors have sole responsibility for the writing and content of the paper.

Supplement

Supplement: List of all identified proteins, their relative abundances in the analyzed proteins sources corn, potato, rice, tomato, spinach, and wheat as well as their predicted strong, weak or no evidence for allergenic potential based on proteomic and bioinformatic analyses (AllerCatPro).

References

- Aalberse RC (2000) Structural biology of allergens. *The Journal of allergy and clinical immunology* 106(2):228-38 doi:10.1067/mai.2000.108434
- Bernhisel-Broadbent J, Sampson HA (1989) Cross-allergenicity in the legume botanical family in children with food hypersensitivity. *The Journal of allergy and clinical immunology* 83(2 Pt 1):435-40
- Blackburn K, N'Jai A U, Dearman RJ, Kimber I, Gerberick GF (2015) Respiratory allergenic potential of plant-derived proteins: Understanding the relationship between exposure and potency for risk assessments. *Critical reviews in toxicology* 45(9):799-811 doi:10.3109/10408444.2015.1067876
- Ceuppens JL, Baroja ML, Lorre K, Van Damme J, Billiau A (1988) Human T cell activation with phytohemagglutinin. The function of IL-6 as an accessory signal. *Journal of immunology* 141(11):3868-74
- Chary A, Hennen J, Klein SG, Serchi T, Gutleb AC, Blomeke B (2018) Respiratory sensitization: toxicological point of view on the available assays. *Archives of toxicology* 92(2):803-822 doi:10.1007/s00204-017-2088-5
- Das A, Ghosh P, Das S (2018) Expression of *Colocasia esculenta* tuber agglutinin in Indian mustard provides resistance against *Lipaphis erysimi* and the expressed protein is non-allergenic. *Plant cell reports* 37(6):849-863 doi:10.1007/s00299-018-2273-x
- Dearman RJ, Kimber I (2001) Determination of protein allergenicity: studies in mice. *Toxicology letters* 120(1-3):181-6
- Dearman RJ, Kimber I (2005) Characterisation of immune responses to food allergens in mice. *The Proceedings of the Nutrition Society* 64(4):426-33
- Dearman RJ, Skinner RA, Herouet C, Labay K, Debruyne E, Kimber I (2003) Induction of IgE antibody responses by protein allergens: inter-laboratory comparisons. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association* 41(11):1509-16
- Dimitrov I, Naneva L, Bangov I, Doytchinova I (2014) Allergenicity prediction by artificial neural networks. *Journal of Chemometrics* 28(4):282-286 doi:doi:10.1002/cem.2597
- FAO/WHO (2001) Evaluation of Allergenicity of Genetically Modified Foods: Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, 22-25 January 2001, Rome, Italy. http://www.who.int/foodsafety/publications/biotech/en/ec_jan2001.pdf. Food and Agriculture Organization of the United Nations (FAO) and World Health Organization (WHO)
- Food And Drug Administration, FDA (2004) Food Allergen Labeling and Consumer Protection (FALCP) Act of 2004. PL 108-282. <https://www.fda.gov/downloads/Food/GuidanceRegulation/UCM179394.pdf>.
- Foster ES, Kimber I, Dearman RJ (2013) Relationship between protein digestibility and allergenicity: comparisons of pepsin and cathepsin. *Toxicology* 309:30-8 doi:10.1016/j.tox.2013.04.011
- Foti C, Damiani E, Zamboni CG, et al. (2012) Urticaria and angioedema to rubisco allergen in spinach and tomato. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology* 108(1):60-1 doi:10.1016/j.anai.2011.09.011

- Geadia D, Valdes R, Escobar A, et al. (2007) Detection of Rubisco and mycotoxins as potential contaminants of a plantibody against the hepatitis B surface antigen purified from tobacco. *Biologicals : journal of the International Association of Biological Standardization* 35(4):309-15 doi:10.1016/j.biologicals.2007.02.007
- Herman RA, Ladics GS (2018) Allergenic sensitization versus elicitation risk criteria for novel food proteins. *Regulatory toxicology and pharmacology : RTP* 94:283-285 doi:10.1016/j.yrtph.2018.02.016
- Herman RA, Woolhiser MM, Ladics GS, et al. (2007) Stability of a set of allergens and non-allergens in simulated gastric fluid. *International journal of food sciences and nutrition* 58(2):125-41
- Hoff M, Son DY, Gubesch M, et al. (2007) Serum testing of genetically modified soybeans with special emphasis on potential allergenicity of the heterologous protein CP4 EPSPS. *Molecular nutrition & food research* 51(8):946-55 doi:10.1002/mnfr.200600285
- Hoopmann MR, Winget JM, Mendoza L, Moritz RL (2018) StPeter: Seamless Label-Free Quantification with the Trans-Proteomic Pipeline. *Journal of proteome research* 17(3):1314-1320 doi:10.1021/acs.jproteome.7b00786
- Ikezawa Z, Miyakawa K, Komatsu H, et al. (1992) A probable involvement of rice allergy in severe type of atopic dermatitis in Japan. *Acta dermato-venereologica Supplementum* 176:103-7
- Jeannet-Peter N, Piletta-Zanin PA, Hauser C (1999) Facial dermatitis, contact urticaria, rhinoconjunctivitis, and asthma induced by potato. *American journal of contact dermatitis : official journal of the American Contact Dermatitis Society* 10(1):40-2
- Jia XD, Li N, Wu YN, Yang XG (2005) Studies on BN rats model to determine the potential allergenicity of proteins from genetically modified foods. *World journal of gastroenterology* 11(34):5381-4
- Kasera R, Singh BP, Lavasa S, Prasad KN, Sahoo RC, Singh AB (2011) Kidney bean: a major sensitizer among legumes in asthma and rhinitis patients from India. *PloS one* 6(11):e27193 doi:10.1371/journal.pone.0027193
- Kelstrup CD, Young C, Lavallee R, Nielsen ML, Olsen JV (2012) Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. *Journal of proteome research* 11(6):3487-97 doi:10.1021/pr3000249
- Knudsen I, Soborg I, Eriksen F, Pilegaard K, Pedersen J (2008) Risk management and risk assessment of novel plant foods: concepts and principles. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association* 46(5):1681-705 doi:10.1016/j.fct.2008.01.022
- Kumar R, Singh BP, Srivastava P, Sridhara S, Arora N, Gaur SN (2006) Relevance of serum IgE estimation in allergic bronchial asthma with special reference to food allergy. *Asian Pacific journal of allergy and immunology* 24(4):191-9
- Lomnitski L, Bergman M, Nyska A, Ben-Shaul V, Grossman S (2003) Composition, efficacy, and safety of spinach extracts. *Nutr Cancer* 46(2):222-31 doi:10.1207/S15327914NC4602_16
- Lu W, Negi SS, Schein CH, Maleki SJ, Hurlburt BK, Braun W (2018) Distinguishing allergens from non-allergenic homologues using Physical-Chemical Property (PCP) motifs. *Molecular immunology* 99:1-8 doi:10.1016/j.molimm.2018.03.022
- Magro AM, Bennich H (1977) Concanavalin A induced histamine release from human basophils in vitro. *Immunology* 33(1):51-8

- Mari A, Scala E (2006) Allergome: a unifying platform. *Arb Paul Ehrlich Inst Bundesamt Sera Impfstoffe Frankf A M*(95):29-39; discussion 39-40
- Maurer-Stroh S, Krutz NL, Kern PS, et al. (2019) AllerCatPro - Prediction of protein allergenicity potential from the protein sequence. *Bioinformatics* doi:10.1093/bioinformatics/btz029
- Mendez A, Vargas RE, Michelangeli C (1998) Effects of Concanavalin A, fed as a constituent of Jack bean (*Canavalia ensiformis* L.) seeds, on the humoral immune response and performance of broiler chickens. *Poultry science* 77(2):282-9 doi:10.1093/ps/77.2.282
- Nayak AP, Green BJ, Sussman G, et al. (2013) Characterization of *Cannabis sativa* allergens. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology* 111(1):32-7 doi:10.1016/j.anai.2013.04.018
- Pahr S, Constantin C, Mari A, et al. (2012) Molecular characterization of wheat allergens specifically recognized by patients suffering from wheat-induced respiratory allergy. *Clin Exp Allergy* 42(4):597-609 doi:10.1111/j.1365-2222.2012.03961.x
- Pomes A, Davies JM, Gadermaier G, et al. (2018) WHO/IUIS Allergen Nomenclature: Providing a common language. *Molecular immunology* doi:10.1016/j.molimm.2018.03.003
- Pramod SN, Venkatesh YP, Mahesh PA (2007) Potato lectin activates basophils and mast cells of atopic subjects by its interaction with core chitobiose of cell-bound non-specific immunoglobulin E. *Clinical and experimental immunology* 148(3):391-401 doi:10.1111/j.1365-2249.2007.03368.x
- Ramavovololona, Senechal H, Andrianarisoa A, et al. (2014) High IgE sensitization to maize and rice pollen in the highlands of Madagascar. *The Pan African medical journal* 19:284 doi:10.11604/pamj.2014.19.284.4654
- Raven JA (2013) Rubisco: still the most abundant protein of Earth? *The New phytologist* 198(1):1-3 doi:10.1111/nph.12197
- Remington B, Broekman HCH, Blom WM, et al. (2018) Approaches to assess IgE mediated allergy risks (sensitization and cross-reactivity) from new or modified dietary proteins. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association* 112:97-107 doi:10.1016/j.fct.2017.12.025
- Samudzi C, Schroeder S, Griffith S, Chen X, Quinn TP (1997) Crystallization and preliminary studies of lima bean trypsin inhibitor. *Proteins* 27(2):311-4
- Sardiu ME, Washburn MP (2010) Enriching quantitative proteomics with SI(N). *Nature biotechnology* 28(1):40-2 doi:10.1038/nbt0110-40
- Satoh R, Teshima R, Kitta K, et al. (2016) Inter-laboratory optimization of protein extraction, separation, and fluorescent detection of endogenous rice allergens. *Bioscience, biotechnology, and biochemistry*:1-10 doi:10.1080/09168451.2016.1206810
- Satoh R, Tsuge I, Tokuda R, Teshima R (2019) Analysis of the distribution of rice allergens in brown rice grains and of the allergenicity of products containing rice bran. *Food chemistry* 276:761-767 doi:10.1016/j.foodchem.2018.10.080
- Schnell S, Herman RA (2009) Should digestion assays be used to estimate persistence of potential allergens in tests for safety of novel food proteins? *Clinical and molecular allergy : CMA* 7:1 doi:10.1186/1476-7961-7-1
- Smit J, Zeeuw-Brouwer ML, van Roest M, de Jong G, van Bilsen J (2016) Evaluation of the sensitizing potential of food proteins using two mouse models. *Toxicology letters* 262:62-69 doi:10.1016/j.toxlet.2016.09.005

- Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6):926-32 doi:10.1093/bioinformatics/btu739
- The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic acids research* 45(D1):D158-D169 doi:10.1093/nar/gkw1099
- Thomas K, Aalbers M, Bannon GA, et al. (2004) A multi-laboratory evaluation of a common in vitro pepsin digestion assay protocol used in assessing the safety of novel proteins. *Regulatory toxicology and pharmacology : RTP* 39(2):87-98 doi:10.1016/j.yrtph.2003.11.003
- Van Bilsen J, De Zeeuw-Brouwer M, Van Roest M, et al. (2015) The prediction of the allergenicity food proteins by combining an in vitro and an in vivo model. *Clinical and translational allergy* 5(3):P153 doi:10.1186/2045-7022-5-s3-p153
- van Bilsen JHM, Sienkiewicz-Szlapka E, Lozano-Ojalvo D, et al. (2017) Application of the adverse outcome pathway (AOP) concept to structure the available in vivo and in vitro mechanistic data for allergic sensitization to food proteins. *Clinical and translational allergy* 7:13 doi:10.1186/s13601-017-0152-0
- Verhoeckx K, Broekman H, Knulst A, Houben G (2016) Allergenicity assessment strategy for novel food proteins and protein sources. *Regulatory toxicology and pharmacology : RTP* 79:118-124 doi:10.1016/j.yrtph.2016.03.016
- Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. *Nature methods* 6(5):359-62 doi:10.1038/nmeth.1322

Table 1: List of proteins considered as low allergenic based on expert judgement and/or based on results derived by *in vitro* and/or *in vivo* experiments.

Protein name	Organism	References
RuBisCO	<i>Spinacia oleracea</i> (spinach)	(Herman et al. 2007; Hoff et al. 2007; Smit et al. 2016; Thomas et al. 2004; Van Bilsen et al. 2015)
Agglutinin, Mannose-specific lectin	<i>Solanum tuberosum</i> (potato)	(Dearman and Kimber 2005; Jeannet-Peter et al. 1999)
Acid phosphatase	<i>Solanum tuberosum</i> (potato)	(Dearman and Kimber 2001; Jia et al. 2005)
Lectin, Phytohemagglutinin	<i>Phaseolus vulgaris</i> (kidney bean)	(Ceuppens et al. 1988; Kasera et al. 2011)
Trypsin inhibitor	<i>Phaseolus lunatus</i> (lima bean)	(Bernhisel-Broadbent and Sampson 1989; Herman et al. 2007; Kumar et al. 2006; Samudzi et al. 1997)
Concanavalin A	<i>Canavalia ensiformis</i> (jack bean)	(Herman et al. 2007; Magro and Bennich 1977; Mendez et al. 1998; Pramod et al. 2007; Thomas et al. 2004)
Lectin	<i>Pisum sativum</i> (garden pea)	(Herman et al. 2007)
Peroxidase	<i>Armoracia rusticana</i> (horseradish)	(Herman et al. 2007; Thomas et al. 2004)
Lipoxygenase	<i>Glycine max</i> (soybean)	(Smit et al. 2016; van Bilsen et al. 2017)
CP4-EPSPS	<i>Glycine max</i> (soybean)	(Hoff et al. 2007)
Mannose-binding tuber agglutinin	<i>Colocasia esculenta</i> (wild taro)	(Das et al. 2018)

Table 2: List of proteins linked to a protein sequence and annotated with no evidence for allergenicity ('No IgE-binding Antigens') in Allergome. Proteins were found in the following tests either with a negative test result (-1) or with no data available (0): skin test, functional tests (e.g. basophil or mast cell activation tests), non-functional tests (e.g. IgE immunoblotting), epidemiology data.

Protein name	Organism	Allergome ID	UniProt ID	Skin Test	Functional Test	Non-Functional Test	Epidemiology
Agr sp CP4 EPSPS	<i>Agrobacterium sp.</i>	2952	Q9R4E4	-1	0	-1	-1
Art gl CO	<i>Arthrobacter globiformis</i>	5976	Q5911, Q7X2H8	0	0	-1	-1
Bac t Cry1Fa	<i>Bacillus thuringiensis</i>	2833	Q45749	0	0	-1	-1
Bac t Cry3Bb1	<i>Bacillus thuringiensis</i>	8256	Q06117	0	0	-1	-1
Bac t Cry1Ab	<i>Bacillus thuringiensis</i>	3852	A0MLX0, Q7BE98 C3FWH8, C3DAS4, P0A371, P0A370, P0A372	0	0	-1	-1
Bac t Cry9c	<i>Bacillus thuringiensis</i>	2953	Q45733	-1	0	-1	-1
Bet v BB18	<i>Betula pendula</i>	9881	O49813	0	0	-1	-1
Cat r 17kD	<i>Catharanthus roseus</i>	2901	P93105	-1	-1	-1	0
Der p 30	<i>Dermatophagoides pteronyssinus</i>	10870	Q96217	0	0	-1	-1
Bac st amyM	<i>Geobacillus/ Bacillus stearothermophilus</i>	3854	P19531	0	0	-1	-1
Sac c Profilin	<i>Saccharomyces cerevisiae</i>	8836	P07274, C8ZI84, B3LJG4, A6ZNW9, D6W2I0	0	0	-1	0
Sta a SEH	<i>Staphylococcus aureus</i>	3907	Q334J6, P0A0M0, Q6GD45, P0A0L9, D6UER2, E1E4S4, Q6VAM8	0	0	-1	-1
Sta a SEJ	<i>Staphylococcus aureus</i>	3909	Q76LS7	0	0	-1	-1
Sta a SEK	<i>Staphylococcus aureus</i>	3910	O54476, D1GR64, D9RN08, E0P9Y9, Q5HHK0, Q93CC5, C8KPR5, D1QCY2	0	0	-1	-1
Str v PAT	<i>Streptomyces viridochromogenes</i>	2954	Q5IW47, D9XF32, Q285M3, Q57146	0	0	-1	-1
Zoa a ISP III	<i>Zoarcis americanus</i>	3847	P19614	-1	-1	-1	0

Table 3: List of proteins recognized as allergens according to WHO/IUIS with their according relative abundances in corn, potato, tomato, rice, spinach, and wheat obtained from proteomic analyses. Proteins were identified based on their sequence or UniProtID via AllerCatPro. Included are only those proteins with a relative abundance above 0.1%.

Allergen	Protein name	UniProt ID	Relative abundance [%]
Corn			
Zea m 14	Non-specific lipid-transfer protein	P19656	8.59
Zea m 8	Endochitinase A	P29022	0.52
Potato			
Sola t 3	Putative cysteine proteinase inhibitor P7B11 (Fragment)	Q8S382	1.79
Sola t 2	KTI-A protein (Fragment)	A0A097H108	1.78
Sola t 2	Aspartic protease inhibitor 7	Q41448	1.15
Sola t 4	KTI-B protein	A0A097H180	1.07
Sola t 4	Serine protease inhibitor 1	P58514	0.97
Sola t 4	Serine protease inhibitor 4 (Fragment)	P58517	0.57
Sola t 2	Aspartic protease inhibitor 2	Q43646	0.18
Sola t 4	Putative kunitz-type proteinase inhibitor P1D4	Q8S381	0.15
Tomato			
Sola l 2	Acid beta-fructofuranosidase	P29000	4.04
Sola l 3	Non-specific lipid-transfer protein 2	P93224	1.55
Sola l 4	PR10 protein	K4CWC4	0.63
Sola l 5	Peptidyl-prolyl cis-trans isomerase	P21568	0.21
Sola l 2	Acid beta-fructofuranosidase	P29000	4.04
Rice	None		
Spinach	None		
Wheat			
Tri a 30	Alpha-amylase/trypsin inhibitor CM3	P17314	11.00
Tri a 26	Glutenin, high molecular weight subunit DX5	P10388	0.96
Tri a 31	Triosephosphate isomerase, cytosolic	P12863, P46226	0.24, 0.12
Wheat allergens found in tomato, rice, potato			
Tri a 32	l-Cys peroxiredoxin A (found in rice)	P0C5C9	1.65
Tri a 31	Triosephosphate isomerase, cytosolic (found in rice)	P48494	0.43
Tri a 31	Uncharacterized protein (found in tomato)	K4BPK4	0.36
Tri a 39	Type I serine protease inhibitor (found in potato)	E0WCF2	0.24

Table 4: List of proteins considered with low allergenic potential based on their relatively high abundance in selected protein sources measured by proteomic analyses accompanied by no evidence for allergenic potential based on bioinformatic analysis. Included are only proteins with a relative abundance above 0.1% and a characterized status according to UniProt.

Protein Source	Protein name	UniRef90 Cluster ID	Relative abundance [%]
Corn	Antimicrobial peptide MBP-1	P28794	5.28
	Zein-beta	P06673	4.27
	Defensin-like protein 2	P81009	2.26
	Late embryogenesis abundant protein EMB564	P46517	1.68
	Nucleoside diphosphate kinase 1	P93554	0.85
	Sucrose synthase 2	P30298	0.57
	Embryonic abundant protein 1	B6T8E4	0.50
	Glyoxalase family protein superfamily	B6SGF3	0.49
	Stress-inducible membrane pore protein	B6TTP4	0.46
	10 kD zein protein	Q41881	0.45
	Late embryogenesis abundant protein Lea14-A	B6UH99	0.42
	Eukaryotic translation initiation factor 5A	P80639	0.39
	Late embryogenesis abundant protein D-34	B6UH67	0.37
	Oil body-associated protein 1A	B4FFZ9	0.34
	Oil body-associated protein 2A	B4FFK9	0.33
	Aldose reductase	E9JVD4	0.33
	Adenine nucleotide transporter BT1	P29518	0.27
	Defensin-like protein 1	B6SL97	0.26
	14-3-3-like protein GF14-12	Q01526	0.25
	Aspartate aminotransferase	B4FUH2	0.23
	Bowman-Birk type wound-induced proteinase inhibitor WIP1	B4FPL1	0.23
	Cysteine proteinase inhibitor	K7UX83	0.22
	Seed maturation protein	B6UGJ4	0.21
	Glucose-1-phosphate adenylyltransferase small subunit (Fragment)	P55240	0.19
	Glucose-1-phosphate adenylyltransferase	Q84J79	0.16
	Embryonic protein DC-8	K7UNW7	0.12
	40S ribosomal protein S9	B6T7B2	0.11
	Elongation factor 2	Q9ASR1	0.11
	Phosphoglucomutase, cytoplasmic 1	P93804	0.10
	Jasmonate-induced protein	B4F7S2	0.10
	Seed maturation protein	K7TQC8	0.10
Spinach	Cytochrome b559 subunit alpha	P56779	4.67
	ATP synthase subunit alpha, mitochondrial (Fragment)	P80082	2.23
	Ribulose biphosphate carboxylase/oxygenase activase, chloroplastic	P10871	1.89
	Carbonic anhydrase	A0A0K9QE98	1.12
	Oxygen-evolving enhancer protein 3, chloroplastic	P12301	1.00
	Photosystem II CP43 reaction center protein	P06003	0.88
	ATP synthase epsilon chain, chloroplastic	P00833	0.88
	Photosystem II CP47 reaction center protein	P04160	0.81
	Cytochrome b559 subunit beta	P60128	0.76
	Histone H4	P59259	0.75
	Photosystem I iron-sulfur center	P42046	0.48
	Photosystem II protein D1	P83755	0.48
	Photosystem I reaction center subunit III, chloroplastic	P12355	0.33
	Photosystem II D2 protein	Q8CM25	0.33
	Histone H2A	A0A0K9QP00	0.30
	Oxygen-evolving enhancer protein 1, chloroplastic	P12359	0.24
	Ferredoxin--NADP reductase, chloroplastic	P00455	0.24
	Phosphoglycerate kinase	A0A0J8CZ08	0.22
	Photosystem I reaction center subunit II, chloroplastic	P12353	0.20
	Photosystem I reaction center subunit XI, chloroplastic	Q41385	0.19
	ATP synthase delta chain, chloroplastic	P11402	0.17
	30S ribosomal protein S4, chloroplastic	P13788	0.15
	Sedoheptulose-1,7-bisphosphatase, chloroplastic	O20252	0.15

Protein Source	Protein name	UniRef90 Cluster ID	Relative abundance [%]
	50S ribosomal protein L16, chloroplastic	P17353	0.15
	ATP synthase subunit alpha	A0A0J0Y4H9	0.15
	Fructose-1,6-bisphosphatase, chloroplastic	P22418	0.14
	ATP synthase subunit alpha (Fragment)	A1XIR6	0.14
	Elongation factor 1-alpha (Fragment)	A8ILY0	0.14
	Chloroplast ribose-5-phosphate isomerase	Q8RU73	0.12
	Elongation factor Tu	A0A0J8CAD8	0.12
	Phosphoribulokinase, chloroplastic	P09559	0.12
	ATP synthase gamma chain, chloroplastic	P05435	0.12
	30S ribosomal protein S19 alpha, chloroplastic	P06508	0.11
	40S ribosomal protein S14-1	Q9SIH0	0.11
	Nucleoside diphosphate kinase 1	Q02254	0.11
	Photosystem II 22 kDa protein, chloroplastic	Q02060	0.10
Potato	Proteinase inhibitor II type D	J7ENS8	2.88
	Lipoxygenase (Fragment)	O24378	2.23
	Proteinase inhibitor type-2 CM7	Q43652	2.04
	Annexin	Q9M3H3	1.97
	UTP--glucose-1-phosphate uridylyltransferase	P19595	1.07
	Linoleate 9S-lipoxygenase 2	O24379	1.01
	Nucleoside diphosphate kinase (Fragment)	P47921	0.76
	Alpha-1,4 glucan phosphorylase L-1 isozyme, chloroplastic/amyloplastic	P04045	0.70
	14-3-3 protein 10	P93207	0.45
	Malic enzyme	M1D7J7	0.40
	S-formylglutathione hydrolase	M1AYK4	0.39
	Snakin-2	Q93X17	0.38
	Annexin	M0ZNV9	0.30
	Proteinase inhibitor type-2	Q00782	0.29
	40S ribosomal protein S24	M1BPE5	0.23
	Alpha-glucan water dikinase, chloroplastic	Q9AWA5	0.21
	60S ribosomal protein L13	Q3HRY7	0.18
	Elongation factor 1-alpha	P43643	0.18
	60S ribosomal protein L12	O50003	0.17
	Eukaryotic initiation factor 4A-2	P41377	0.17
	Fructokinase	P37829	0.17
	Phosphoglucomutase, chloroplastic	Q9M4G5	0.16
	Phosphoglucomutase, cytoplasmic	Q9M4G4	0.16
	Protease inhibitor-related protein	C6F3B7	0.15
	Induced stolen tip protein TUB8 (Fragment)	P33191	0.15
	Glucose-6-phosphate 1-dehydrogenase, cytoplasmic isoform	P37830	0.15
	Leucine aminopeptidase 2, chloroplastic	Q42876	0.13
	Fructokinase	K7WJT8	0.12
	Metalloprotease inhibitor	Q948Z8	0.11
	Glucose-1-phosphate adenylyltransferase small subunit	P23509	0.11
	60S ribosomal protein L34-2	Q9FE65	0.11
	Ran binding protein-1	Q94K24	0.11
	Pyridoxal 5'-phosphate synthase subunit PDX1.3	Q8L940	0.10
Rice	Os02g0202400 protein	Q6Z782	0.91
	10 kDa prolamin	A2XMB2	0.80
	Aspartate aminotransferase, cytoplasmic	P37833	0.56
	Embryonic abundant protein 1	P46520	0.30
	Nucleoside diphosphate kinase 1	Q07661	0.29
	Glutaredoxin-C6	P55142	0.28
	Late embryogenesis abundant protein, group 3	P0C5A4	0.26
	60S ribosomal protein L12	O50003	0.24
	Pyruvate, phosphate dikinase 1, chloroplastic	Q6AVA8	0.21
	40S ribosomal protein S13-1	Q69UI2	0.18
	H0502B11.4 protein	Q01L47	0.15
	Os01g0705200 protein	Q94JF2	0.15
	Sucrose synthase 4	Q10LP5	0.14
	Alpha-1,4 glucan phosphorylase	Q9AUV8	0.14
	Os01g0159600 protein	Q5ZEL0	0.14
	Os04g0510900 protein (Fragment)	C7J0T2	0.13

Protein Source	Protein name	UniRef90 Cluster ID	Relative abundance [%]
	Os02g0715400 protein	Q6ZHP6	0.13
	Sucrose synthase 2	P30298	0.12
	Phosphoglycerate kinase, cytosolic	P12783	0.11
	Os03g0159600 protein	Q8H8B0	0.11
	AWPM-19-like family protein, expressed	Q8H920	0.10
	Late embryogenesis abundant protein 1	A3AHG5	0.10
Tomato	Fruit-specific protein	P14903	4.41
	Linoleate 9S-lipoxygenase B	P38416	3.45
	Abscisic stress-ripening protein 1	Q08655	1.14
	Adenosylhomocysteinase	Q9SWF5	0.73
	1-aminocyclopropane-1-carboxylate oxidase homolog	P10967	0.71
	Nucleoside diphosphate kinase (Fragment)	P47921	0.58
	Alcohol acyl transferase	Q6QLX4	0.50
	Annexin	Q9M3H3	0.48
	V-type proton ATPase catalytic subunit A	O23654	0.40
	Flower-specific gamma-thionin-like protein/acidic protein	Q40128	0.36
	V-type proton ATPase subunit B1	P11574	0.33
	Probable aquaporin PIP-type pTOM75	Q08451	0.25
	Monodehydroascorbate reductase	Q43497	0.24
	14-3-3 protein 10	P93207	0.23
	Harpin binding protein 1	Q5QJA7	0.22
	1-aminocyclopropane-1-carboxylate oxidase 1	P05116	0.22
	40S ribosomal protein S25	P46301	0.20
	ATP synthase subunit alpha, chloroplastic	P00823	0.19
	40S ribosomal protein SA	K4C8T4	0.18
	60S ribosomal protein L13	Q3HRY7	0.18
	Cytochrome b-c1 complex subunit 9	P46270	0.17
	60S ribosomal protein L34-2	Q9FE65	0.16
	UTP--glucose-1-phosphate uridylyltransferase	P19595	0.15
	Aconitate hydratase	M1AGW8	0.14
	Carbonic anhydrase	Q5NE21	0.14
	S-formylglutathione hydrolase	M1AYK4	0.14
	Response to dessication 2	K4B3I4	0.14
	40S ribosomal protein S6	Q9M3V8	0.14
	Alpha-1,4-glucan-protein synthase [UDP-forming] 2	Q8RU27	0.14
	Aspartate aminotransferase	K4CK47	0.13
	40S ribosomal protein S14-1	Q9SIH0	0.13
	60S ribosomal protein L1	K7XKT4	0.13
	ADP,ATP carrier protein, mitochondrial	P25083	0.13
	Glutaredoxin	Q9ZR41	0.12
	Cytochrome b-c1 complex subunit 7	P48502	0.12
	Obg-like ATPase 1	G7IKV9	0.12
	Plastid-lipid-Associated Protein (Fragment)	O24024	0.12
	Phosphoglucomutase, cytoplasmic	Q9M4G4	0.11
	Elongation factor 1-alpha	P43643	0.11
	Leucine aminopeptidase 2, chloroplastic	Q42876	0.11
	Small nuclear ribonucleoprotein E	A9PB49	0.11
	GTP-binding nuclear protein Ran1	P38546	0.11
	Protein phosphatase 2C	Q6QLU0	0.10
	Pyruvate kinase, cytosolic isozyme	P22200	0.10
Wheat	Puroindoline-A	P33432	1.38
	Histone H4	A0A096USG9	0.83
	Em protein H2	Q08000	0.36
	Defensin-like protein 1	P20158	0.27
	Elongation factor 1-alpha	M8AVR5	0.21
	Late embryogenesis abundant protein, group 3	Q03968	0.17
	Caleosin	W5BUF4	0.17
	Glucose-1-phosphate adenylyltransferase large subunit	P12299	0.16
	Histone H2A.2.1	P02276	0.14
	Glucose-1-phosphate adenylyltransferase small subunit	P30523	0.14
	Phosphoglycerate kinase, cytosolic	P12783	0.11
	Nucleoside diphosphate kinase	W4ZQ38	0.11

Figure 1: Distribution of relative abundances and predicted allergenic potential of all identified proteins above 0.1% extracted from wheat, corn, spinach, potato, rice, and tomato with strong (red), weak (light green) or no evidence (dark green) of allergenic potential based on proteomic and bioinformatic analyses (AllerCatPro).

