

Bioinformatic Methods for Identifying Known or Potential Allergens in the Safety Assessment of Genetically Modified Crops

**Richard E. Goodman, Andre Silvanovich,
Ronald E. Hileman, Gary A. Bannon,
Elena A. Rice, and James D. Astwood**

Product Safety Center, Monsanto Company,
St. Louis, Missouri, USA

Keywords: allergen, allergy, bioinformatic, biotechnology, food safety, genetically modified, sequence comparison

INTRODUCTION

Agricultural crops have been genetically improved through centuries of breeding to select phenotypes that are controlled by combinations of genes, typically with undefined mutations, which produce the desired traits. Changes in the plant characteristics (e.g., disease resistance, insect resistance, food quality) typically have been slow, except when partial or full genomes have been combined. This combination as is hypothesized to have occurred thousands of years ago in the generation of modern hexaploid wheat as a hybrid cross of tetraploid and diploid progenitors (1). Such crosses result in the combination of hundreds to thousands of different proteins in a single food. In contrast, modern tools of biotechnology have allowed the introduction of one or a few new genes and resulting proteins into crop varieties that are carefully selected and studied for safety and performance before being allowed into commercial production (2, 3). As methods to introduce desired traits in plants have improved, so too has the public

We thank Dr. Sue Hefle, co-director of the Food Allergy Research and Resource Program at the University of Nebraska, for her support in producing and maintaining the online database and FASTA3 algorithm at FARRP and Gyula Holleschak from Monsanto for his technical expertise in designing and maintaining the web site. We also thank Dr. W.R. Pearson for posting his FASTA3 program as a publicly available algorithm on the University of Virginia web site.

Address correspondence to Richard E. Goodman, Monsanto Company, Mail Stop O3F, 800 North Lindbergh Blvd. St. Louis, MO 63167, USA. E-mail: richard.e.goodman@monsanto.com

expectation that modern safety tests must be improved to reduce the chance of inadvertently introducing a new allergen into a food.

The U.S. Food and Drug Administration recognized in 1992 the need for evaluating the potential allergenicity of the newly expressed proteins in genetically engineered crops, although the difficulties of trying to predict whether a protein would be an allergen were well recognized (4). For example, Pioneer Hi-Bred International developed a nutritionally improved soybean variety by inserting the 2S albumin gene from the Brazil nut to increase methionine levels. Food allergy experts evaluated the safety of the protein and discovered that it was a previously unknown cause of allergic reactions to Brazil nuts (5). As a result, that product was stopped during development. This experience helped guide the development of an allergenicity assessment scheme that has been widely used to evaluate genetically engineered food crops (6). Major agricultural biotechnology companies generally follow these recommendations. However, some investigators have either unwittingly, or intentionally, introduced known allergens into different crop species, (7, 8) but at this time those products have not been registered for commercial production. It is therefore essential that the methods for evaluating the potential allergenicity of the introduced proteins be improved, and communicated widely to those who are attempting to improve crops through biotechnology.

A key step in the assessment process is evaluating the introduced protein to determine if it is known to be an allergen or is similar enough to a known allergen to suspect that it would be cross-reactive, i.e., capable of eliciting an allergic response in those with existing allergies. Cross-reactivity occurs when allergen-specific IgE reacts with proteins that are sufficiently similar in amino acid sequence and structure to mimic the specific allergen. Good clinical examples of allergies caused by cross-reactive allergens are found in individuals with allergies to shrimp. These patients typically have allergic responses to other crustacea, sometimes to mollusks, and often have IgE to housedust mites and cockroach antigens, especially tropomyosins (9). In some cases, these represent co-sensitization, but many individuals react to the first exposure of the new food indicating probable cross-reactivity. Another example is that of individuals with allergies to European birch pollen that frequently experience oral allergy syndrome (OAS), but rarely systemic reactions, when eating specific fresh fruits or vegetables (10). The OAS symptoms usually do not occur if the food is cooked. Sensitization to two major birch pollen allergens (Bet v 1 and Bet v 2), or homologues of these allergens that are sometimes referred to as pathogenesis-related proteins, has been implicated through in vitro IgE binding studies with sera from many patients (10, 11). However, some IgE binding results may represent weak or nonspecific interactions unlikely to be clinically relevant. Therefore, definitive patient histories and food challenges are needed to demonstrate cause and effect before allergenicity is truly established. Even skin prick tests can yield false-positive results (12). Proteins that are truly

cross-reactive are typically greater than 70% identical in primary amino acid sequence and have similar secondary structure (13). Therefore, comparing the translated sequences of genes transferred into crops with those of known allergens is used to predict whether or not a new protein is similar enough to a known allergen to suspect that it might cause cross-reactions. If the sequence is highly similar to an allergen, direct IgE binding, or more appropriately IgE inhibition, binding studies should be performed using serum from individuals with allergies to the identified similar protein (10, 13, 14). Additional confirmation may require double-blind placebo-controlled food challenge (6).

Existing computer algorithms such as FASTA (15) or simple scanning window comparative algorithms may be used to compare each new protein with sequences in a database of known allergens to identify effectively known allergens and proteins that are likely to be cross-reactive. Here we describe the compilation of an allergen database and provide examples of two different algorithms to compare hypothetical candidate protein products to provide potential criteria for determining significant amino acid sequence matches to allergens.

WHOLE PROTEIN AND SHORT AMINO ACID SEQUENCE HOMOLOGY SEARCHES

The proteins introduced into all genetically engineered plants on the market in the United States have been screened by comparing their amino acid sequence with those of known allergens as one of many assessments performed to evaluate product safety (6, 16). Additionally, the amino acid sequence of the introduced protein is screened against known proteins in publicly available sequence databases to identify proteins that could have other potential safety concerns based on activity or known toxicity. The extent of sequence similarities between the introduced protein and database sequences of allergens, gliadins, and other proteins can be effectively assessed using the FASTA sequence alignment tool (15). Although the FASTA algorithm directly compares amino acid sequences and provides the best alignment between the query sequence and similar database sequences (*i.e.*, primary protein structure), the alignment data may be used to infer higher order structure (*i.e.*, secondary and tertiary protein structures). Proteins that share a high degree of similarity throughout the entire length are considered homologous and share secondary structure and common three-dimensional folds (17). Homologous proteins are more likely to share allergenic cross-reactive conformational and linear epitopes than unrelated proteins; however, the degree of similarity between homologues varies widely.

The minimum requirement for allergenic cross-reactivity of conformational epitopes has been estimated as 50% identity across the full length of

the protein sequences, and they may not be common unless there is at least 70% identity (18). The Codex Alimentarius Task Force on Foods Derived from Biotechnology recently accepted 35% identity to an allergen, in a segment of 80 or more amino acids, as a reasonable criteria for considering a protein likely to be allergenic and therefore requiring further testing (19). However, there is relatively little direct evidence to support this low specific cut-off as many studies claiming cross-reactivity are based on single IgE binding studies or skin tests that may be the result of either multiple sensitivities or low affinity IgE.

Questions concern the ability of the FASTA search to identify short regions of matching amino acid sequence that might be bound by allergen specific IgE and represent a cross-reactive epitope. Because IgE-binding epitopes have only been identified for a few allergens, it is not possible at this time to construct a comprehensive database for a more accurate search. Further, most of the identified "epitopes" are known through *in vitro* mapping studies without regard to antibody affinity and have used sera from only a few allergic individuals. In the absence of a complete description of IgE epitopes for all known allergens, a theoretical IgE epitope search may be performed by scanning all possible overlapping 8 amino acid peptides from each sequence in the allergen database with all possible overlapping peptides of 8 amino acids from the test protein, in a pair-wise fashion to identify identical matches (6). Identification of these short matches is a highly conservative approach, as few of the theoretical peptides compared with the query sequence represent *bona fide* IgE epitopes. This approach also is all-inclusive because it includes the possibility of matching regions to the allergen, most of which would not bind IgE.

The use of an 8 amino acid window represents a compromise to identify most potentially cross-reactive linear epitopes, while reducing the probability of identifying a large number of irrelevant similarities that would be identified using a window smaller than 8 amino acids and unlikely to present a risk of cross-reactivity. Also, two IgE-binding epitopes on the same molecule are required to cross-link high-affinity IgE receptors on mast cells to induce an intracellular signal. However, because not all potential epitopes could be identified by this search, a single match (that may well not represent an epitope) would be important to note. While some IgE epitopes may be as short as 5 amino acids (20, 21), the majority of characterized IgE-linear epitopes are 8 amino acids or longer (22–24). Although many of these reports have demonstrated IgE binding, few have tested the affinity of the binding, or the allergic significance of the *in vitro* binding, and it is clear from some reports that high-affinity binding requires 8 or more amino acids (20, 25). Whereas researchers have attempted to identify structures or sequences that differentiate allergens from nonallergenic proteins, common features have not been identified that would allow accurate predictions based on structure alone (18). Therefore, the safety assessment must focus on evaluating whether the primary amino acid

sequence of the introduced protein is a known allergen or is similar enough to suspect possible cross-reactivity.

Three essential components are required for an effective bioinformatic approach to evaluating the potential allergenicity of the protein to be introduced into a food crop: 1. A useful and broadly representative allergen database. 2. A simple computer search method or algorithm. 3. Predictive criteria for the identification of potentially cross-reactive proteins (26).

ALLERGEN DATABASES

While public databases including GenBank, EMBL, and SwisProt contain DNA and/or amino acid sequences of most of the allergens that have been identified, these sequences are not annotated fully regarding their allergenicity. Additionally, a search of these databases with existing algorithms primarily identifies proteins that are similar to the query sequence and does not differentiate between allergenic and nonallergenic proteins. A more productive search can be accomplished if an allergen-specific database is used. Such a database was constructed by collaborators at Monsanto Company and the Food Allergy Research and Resource Program at the University of Nebraska (FARRP) and is now publicly accessible on the Internet at <http://www.allergenonline.com> (27). The Allergenonline database currently contains 635 sequences of known or putative allergens and is searchable with an interactive online FASTA3 algorithm (17). Other databases, or lists of allergens are available through Protall (<http://www.ifr.bbsrc.ac.uk/protall>) at the Institute of Food Research in the United Kingdom; the International Union of Immunological Society (IUIS), Allergen Nomenclature Subcommittee list of allergens (<http://www.allergen.org/List.htm>); a database of food and nonfood allergens (<http://www.iit.edu/~sgendel/>), at the National Center for Food Safety and Technology in Chicago, IL, USA; (28) and a list of food and aeroallergens by Allergy Resources International in South Africa (<http://www.allallergy.net/>). The protein structural database at Protein Data Bank (Rutgers University, NJ) also could be searched based on three-dimensional structure, but the number of resolved structures of allergens is quite low and therefore at present a search would not have much predictive power.

Allergenonline, Sequence Database Construction

The allergen and gliadin sequence database was assembled from public domain databases including Genbank and EMBL version 119, PIR version 56, the NRL3D version 56 of RCSB PDB and SwissProt version 38 (29–34). A preliminary list of sequences was compiled from the public databases using STRINGSEARCH (keyword = allergen). Nonallergen entries were identified by reviewing published information for each entry

and removed. Additional sequences were added from a previously published allergen list (6), allergens recognized by the IUIS and from searches of current literature available on PubMed and Entrez (35). For some allergens, sequences were added from publications reporting only an N-terminal sequence if the sequence included at least 8 residues. Duplicate sequences were removed, but those of unique isoforms were retained (e.g., multiple Bet v 1 isoforms). There is clear clinical evidence in the literature for the designation of many of these proteins as allergens. However, many have been labeled allergens in the literature based solely on *in vitro* IgE binding using sera from individuals without documented allergies to the protein source (36). Proteins have been called allergens merely because of a high degree of sequence similarity to known allergens without any *in vivo* or *in vitro* tests with appropriately allergic subjects (37). When *in vitro* IgE binding has been used, the specificity of binding or avidity frequently has not been assessed, and few studies have evaluated the potential clinical relevance of IgE binding through *in vivo* challenges. However, we have constructed this database to be widely inclusive as a precautionary measure and have included many of these putative allergens. Therefore, a sequence match to a protein listed in the database warrants further evaluation of the authenticity of the claim of allergenicity once a sequence match is found.

Gliadins, proteins associated with celiac disease, were included in the database even though there is no evidence that IgE is involved in the etiology of this disease. While the mechanism that leads to celiac disease is thought to be a combination of gliadin-specific IgA antibodies and cytolytic T cells, disease management is through avoidance as it is for food allergies (38). Therefore, individuals who developed the original safety assessment also felt that transfer of proteins that are likely to trigger the clinical symptoms of celiac disease also should be avoided (6).

Searching the Allergen Database

A variety of algorithms could be used to compare a query sequence with those in the allergen database. As described above, FASTA is readily available and has been widely used to identify potential homologous sequences. It is flexible in allowing choices for specific search parameters and the use of scoring matrices to find sequences that are nearly identical or are evolutionarily divergent. BLASTP, available through the National Center for Biotechnology Information web site (www.ncbi.nlm.nih.gov) is another viable choice. Both report the maximum (best overall) sequence alignment and are capable of identifying proteins that are likely to share both primary and conformational antibody-binding epitopes due to significant primary amino acid sequence similarities. The similarity of the test sequence relative to any aligned protein can be gauged by the *E* score value, a statistical

estimate of the probability that any one alignment could have happened by chance alone (17). The magnitude of the *E* score is used to infer homology of the aligned sequence pair with very small values ($<10^{-40}$) typically found for homologues from closely related species. The *E* score is affected by the size and complexity of the database, but more so by the size and complexity of the test sequence. The *E* score is not directly related to the percentage of identical amino acid matches shared by the two proteins over the length of the best alignment. Criteria for accepting either an *E* score or a percent identity as significant depends on the intent. When the intent is to characterize a protein as a possible evolutionary homologue, with likely similar function, an *E* score <0.001 may be significant. However, when the intent is to evaluate the potential cross-reactivity of antibody binding, our experience with known clinically cross-reactive allergens compared through the Allergenonline database indicates that a more relevant criteria for predicting cross-reactivity is likely to be much smaller than 10^{-7} and that the value should be used to identify proteins that should be more closely evaluated (27). The percent identity, also calculated by the FASTA program, provides a more informative gauge for the likelihood that two proteins may be cross-reactive in allergic individuals, but further evaluation is necessary to determine a reasonable minimum cut-off (18).

There also may be value in identifying short identical sequences that might be missed by FASTA or BLASTP. We have developed a proprietary UNIX-based program (ALLERGENSEARCH) that runs from a terminal window in GCG (27). This program searches for short segments of identical amino acid matches in the query protein compared with any segment of any protein in the allergen database. Similar programs could be written in any computer language. Alternatively, the sequence could be parsed into overlapping 8-mers for use as query sequences by FASTA against the Allergenonline database.

The sequences of all proteins evaluated in this article were evaluated using both a FASTA3 (version 3 of FASTA) (17) and the ALLERGENSEARCH program. Default parameters were used with the FASTA3 algorithm including an expectation threshold of 10, a wordsize (*k-tuple*) of 2, a gap-creation penalty of -12 , and a gap-extension penalty of -2 . The BLOSUM50 scoring matrix (39) was used to identify matches and estimate the percent identity of two primary sequences. This scoring matrix works well for sequence alignments with gaps and has become a standard for recognizing distant evolutionary relationships (17). The ALLERGENSEARCH program tested sequentially for matches of 6, 7, or 8 amino acid matches.

Interpretation of Results

Reliable use of bioinformatic searches for safety assessment requires a sound strategy and guidelines for interpretation. Suggestions of criteria are

offered in this article, but there are insufficient clinical data on which to establish absolute rules to determine the level of match required to elicit an allergic response. Neither a high percent identity from the FASTA search nor a match of 8 or more identical contiguous amino acids means that a protein is an allergen, or that allergic epitopes have been identified. Rather, such matches indicate that there is an increased chance that a protein may be cross-reactive. After identification of the matching protein and determination of whether published data support the conclusion that it is an allergen, a determination is made regarding the significance of the sequence match. If deemed significant, an evaluation is performed to identify an appropriate allergic population that would be contacted to provide serum to perform in vitro IgE-binding studies and, if appropriate, skin prick tests or food challenges. Clinical testing may be carried out to evaluate objectively the potential cross-reactivity, or the protein may be dropped from consideration as a product.

A match of approximately 100% identifies the protein as a probable allergen in which case the product should not be marketed unless it can be accurately labeled to alert consumers. At some future date, there may be sufficient data to establish the minimum concentration for a threshold of elicitation, in which case allergens or putative allergens may be introduced into crops without labeling if the concentration is significantly below this threshold (40).

To evaluate the potential significance of the percent identity and *E* score values from the FASTA3 search, and to test the utility of a smaller window of identical contiguous amino acid matches, three separate comparisons were performed. The first evaluated 20 arbitrarily chosen corn proteins. Corn proteins were chosen because corn is thought to be one of the least allergenic foods; therefore, very few proteins should be allergenic. In addition, corn pollen also is thought to be relatively less allergenic than that of most grasses. The second comparison evaluated the similarity of a range of tropomyosins from allergenic and nonallergenic sources. The third evaluated similarities between the most and least similar of cross-reactive proteins from two commonly recognized pollen-food allergen families.

Analysis of Corn Sequences

Twenty corn proteins of greater than 100 amino acids in length were arbitrarily chosen from a list of corn peptides compiled from GenBank (National Center for Biotechnology Information web site, <http://www.ncbi.nlm.nih.gov/>). To our knowledge, there are only two documented reports of corn allergic reactions in the past five years including one case of a single anaphylactic reaction (41). Another recent study identified maize lipid transfer protein as the major IgE-binding protein in a number of subjects who claimed to have experienced systemic allergic

reactions upon eating products made from maize, but those individuals were not clinically challenged (42).

Results of the FASTA3 comparison (Table 1) indicate that 4 of the 20 corn proteins are homologues of allergens or suspected allergens based on significant *E* score values (17). The corn proteins matched with a latex allergen manganese superoxide dismutase (83% identity to the corn superoxide dismutase), *Cladosporium* heat shock protein (64% identity to the corn luminal binding protein), *Cladosporium* aldehyde dehydrogenase (53% identity to the corn male sterility restoration protein), and a wheat pollen allergen (62% identity to the corn beta expansin protein). These percent identity and *E* score values indicate a moderate to high likelihood of cross-reactivity (18) for those having allergies to the indicated allergens. Two other corn proteins, dehydrin and alcohol dehydrogenase, had relatively small *E* score values indicating they may be homologues of allergens. However, they are at best evolutionarily distant homologues based on their low percent identity to the allergens in question indicating a low likelihood of cross-reactivity. The other 14 corn proteins do not appear to be homologous to any allergens in the database and are unlikely to cross-react.

Evaluation of the short peptide scans of the 20 proteins for matches with all sequences in the Allergenonline database clearly illustrate the difficulty of using an overlapping amino acid window of less than 8 contiguous residues to identify potentially cross-reactive allergens. A scientific advisory panel of the FAO/WHO recommended using an identity match of 6 amino acids (43). However, the results (Table 2) indicate that 14 of the 20 corn proteins would be considered likely cross-reactive or allergenic, if a 6 amino acid search window is used. While all 4 corn proteins that were identified by the FASTA3 search to have significant sequence homology to allergens also were identified using a scanning window of 6 amino acids, so too were the two evolutionary distant homologues as well as 8 proteins that have less than 35% identity over 80 amino acids to any known allergen and *E* scores that would not suggest homology to known allergens. The results indicate that proteins identified by a 6 amino acid scanning window with low percent identities (<35%) over significant sequence stretches (>80 aa) and insignificant *E* score values (greater than 10^{-2}) are most likely false-positives and should not be considered allergenic.

Scanning window criteria of 7 identical amino acid matches also overestimates potentially cross-reactive allergenic proteins. Using this window size, in addition to the 4 proteins identified by the FASTA3 search, an additional 4 proteins were identified. Similar to the previous search, these 4 proteins appeared unlikely to be allergens based on their low percent identity to known allergens and their insignificant *E* score values.

Finally, a scanning window of 8 amino acids identified 4 likely cross-reactive corn proteins and 2 additional corn proteins. The 2 additional corn proteins, dehydrin and gibberellin responsive modulator, were identified as

TABLE 1 FASTA3 Search of the FARRP Allergen Database with 20 Arbitrarily Chosen Corn Proteins. The protein sequences of 20 corn proteins of greater than 100 amino acids in length were arbitrarily chosen from publicly available databases to use as query sequences against the FARRP protein allergen database (<http://www.allergenonline.com/>)

Corn protein accession #	Length # aa	E score ^a	Identity percent/ Alignment length (aa)	Best allergen match access #	Length # aa
zLikely to cross-react^b					
Superoxide dismutase (SOD) P41980	233	1.1e-73	83%/197	Latex Mn SOD CAB53458	205
Luminal-binding protein P24067	663	3.2e-124	64%/621	Cladosporium heat shock Cla h 4 P40918	643
Male sterility restoration AAK58370	549	3.6e-100	53%/495	Cladosporium aldehyde dehydrogenase S43114	496
Beta expansin AAK56131	286	9e-55	62%/246	Wheat pollen allergen homologues AAD10496	271
zUnlikely to cross-react^c					
Dehydrin AAA33480	289	1.2e-02	31%/96	Latex Hev b 5 AAC49447	151
Alcohol dehydrogenase AAK52444	195	2.6e-05	26%/172	Candida alcohol dehydr P43067	350
Catalase isozyme P18122	492	0.57	41%/29	Turnip chitin-Bra r 2 P81729	91
Malate dehydrogenase AAK58078	110	0.5	36%/45	Kiwi fruit protease Act c 1 P00785	380
Gluc. 6-p- adenylyltransferase P55241	516	0.4	31%/88	Honeybee PLA2 Api m 1 P00630	162
Rubisco activator Q9ZT00	433	1.2	24%/140	Soy glycinin CAA26723	495
Gibberellin response AAK59925	266	0.39	28%/123	Velvet grass pollen grp V CAB10766	240

Diphosnucl. Phosphatase AAK57915	355	1.5	26%/179	Alternaria minor allerg CAA55069	204
Ferr. glut. synthase P23225	1616	0.91	26%/125	Alternaria aldehyde dehydrogenase P42041	495
Cytochrome f CAA60298	320	1.1	28%/81	Bean PvPR1 pathogen CAA43637	156
Large SU of Rubisco NP_043033	476	2.1	35%/46	Latex Hev b 6 P02877	204
Chloroplast envelope P46641	230	1.5	33%/46	Birch isoflavr P46641	300
Tubulin beta-5 chain Q43697	445	1.2	22%/109	Walnut 2S alb AAB41308	139
Transl. initiation factor P55876	451	1.3	22%/106	Cow kappa-casein AAA30433	190
Tryptophan synthase Q59992	412	3.1	22%/107	Honeybee hylaromidase Q08169	382
40S Ribosomal Q08069	221	2.2	42%/36	Latex Hev b 5 AAC49447	151

Searches were performed with the FASTA3 algorithm of Pearson (17). Only the allergen with the best scoring match is presented. The smaller the expectation or *E* score is, the more significant the similarity, with a greater likelihood that the proteins are cross-reactive. The program also calculates the percent identity over the best-aligned region.

^a*E* score values smaller than 0.1 are shown as exponential notation where $e = 10$, e.g., $3.6e-100$ is 3.6×10^{-100} .

^bClear homology to an allergen; may be cross-reactive; should be tested.

^cUnlikely to cross-react, low degree of identity and similarity.

TABLE 2 Sliding Window Search for 6, 7, and 8 Identical Contiguous Amino Acid Matches between 20 Arbitrarily Chosen Corn Proteins and the FARRP Allergen Database. The protein sequences of 20 corn proteins of greater than 100 amino acids in length were arbitrarily chosen from publicly available databases to use as query sequences against the FARRP protein allergen database (<http://www.allergenonline.com/>)

Corn protein accession #	Length in #aa	Hits w/6aa (#ptn)	Hits w/7aa (#ptn)	Hits w/8aa (#ptn)	Best allergen match access	Longest matching segment
zLikely to cross-react^a						
Superoxide dismutase (SOD) P41980	233	3	2	2	Latex Mn SOD CAB53458	> 8
Luminal binding protein P24067	663	20	1	1	Cladosporium heat shock Cla h 4 P40918	> 8
Male sterility restoration AAK58370	549	5	2	2	Cladosporium aldehyde dehydrogenase S43114	> 8
Beta expansin AAK56131	286	21	18	8	Wheat pollen allergen homologues AAD10496	> 8
zUnlikely to cross-react^b						
Dehydrin AAA33480	289	18	4	1	Vitellogenin II precursor VJCH2	8 ^c
Alcohol dehydrogenase AAK52444	195	1	0	0	Chironomus globulin P02228	6
Catalase isozyme P18122	492	0	0	0		
Malate dehydrogenase AAK58078	110	0	0	0		
Gluc. 6-p- adenylyltrans P55241	516	4	1	0	Aspergillus asp f 9 CAA11266	7
Rubisco activator Q9ZT00	433	8	0	0	Ryegrass Lol p 1 A37881	6
Gibberellin response AAK59925	266	5	3	1	Vitellogenin II precursor VJCH2	8 ^c

Diphosnucl. phosphatase AAK57915	355	9	0	0	Chironomid globin and others P02224	6
Ferr. glut. synthase P23225	1616	6	0	0	Walnut Jug r 2 AAF18269	6
Cytochrome f CAA60298	320	2	0	0	Timothy phl p 5b AAC25998	6
Large SU of Rubisco NP_043033	476	1	0	0	Peanut lectin S14765	6
Chloroplast envelope P46641	230	0	0	0		
Tubulin beta-5 chain Q43697	445	0	0	0		
Transl. initiation factor P55876	451	8	4	0	House dust mite tropomyosin CAA75141	7
Tryptophan synthase Q59992	412	0	0	0		
40S Ribosomal Q08069	221	0	0	0		

Searches were performed with a proprietary search program that scans each database sequence with a sliding window of 6, 7, or 8 amino acids of the query sequence, advancing 1 amino acid per cycle. The number of matches to allergens is listed for each query, along with the best (or representative) allergen that was matched and the longest identical contiguous segment. Two corn proteins contain 8-contiguous serine residues and therefore match the chicken vitellogenin II precursor protein that contains a stretch of 80 residues that are primarily serine. That protein of vitellogenin is not known to be part of the allergenic portion of this protein and polyserine is unlikely to be immunogenic.

^aClear homology to allergens: these corn proteins matched the same allergens by FASTA and by ALLERGENSEARCH with a window of 8 and should be tested for cross-reactivity with allergic donor sera.

^bUnlikely to cross-react; "poor matches identified by ALLERGENSEARCH and not identified by FASTA.

^c8-amino acid segment matches are due to 8 and 9 contiguous serine residues, proteins unlikely to cross-react.

having relatively low identities to the latex Hev b 5 and the velvet grass group V allergens, respectively. Interestingly, both also matched the chicken vitellogenin II precursor using ALLERGENSEARCH as both contain a single segment of 8-serine residues, which matched the polyserine region in the chicken protein. Neither the FASTA3 results nor the match of 8-contiguous serine residues found for the dehydrin and gibberellin proteins is thought to be significant.

Analysis of tropomyosins

A FASTA3 algorithm search was performed using a major allergen, shrimp tropomyosin (Met e 1) to evaluate the *E* scores and percent identities for a group of proteins with a large degree of amino acid sequence similarity across a large evolutionary distance. Tropomyosins are important proteins involved in muscle contraction in vertebrates and invertebrates, and because they all perform a similar function, they also share significant amino acid sequence homology. The tropomyosins of shrimp, crab, and lobster are known to be potent, cross-reactive food allergens that are highly stable to denaturation by heat (44), and relatively stable to digestion by pepsin (unpublished, Monsanto). Tropomyosins of mollusks including snails, oysters, and clams also are potent food allergens that are only partly cross-reactive with tropomyosins of crustaceans. The homologous tropomyosins of house dustmites and cockroaches are potent airway sensitizers, cause significant numbers of cases of rhinitis and asthma, and bind IgE from sera of many shrimp allergic individuals (45). Tropomyosins of chickens, rabbits, and pigs are more similar to human tropomyosins than to invertebrate tropomyosins and are not known to be allergenic. Figure 1 depicts a phylogenetic dendrogram of the predicted evolutionary distances between tropomyosins of the various species that was calculated by a clustal alignment using the MegAlign program (DNASTAR, Inc.). The percent identity, compared with shrimp tropomyosin (Met e 1), is shown in brackets along with the *E* score values. The percent identity of the typical cross-reactive tropomyosins is in the range of >55% identity relative to shrimp tropomyosin. Surprisingly, nonallergenic rabbit, pig, and chicken tropomyosins are between 53% and 57% identical to shrimp tropomyosin and share segments of 8 or more contiguous amino acid matches with the allergens. *E* score values also were surprisingly small for the nonallergenic tropomyosins (10^{-33} to 10^{-49}). The reason that humans do not have allergic responses to vertebrate tropomyosins may be the development of thymic tolerance of T cells to proteins that are highly similar to the human tropomyosin protein. Whereas vertebrate tropomyosins share between 49–55% homology with the allergenic shrimp protein, the human tropomyosin is >70% identical to that of other vertebrate tropomyosins. This example illustrates the difficulty in assigning an absolute cut-off value for shared identity meaningful for predicting all

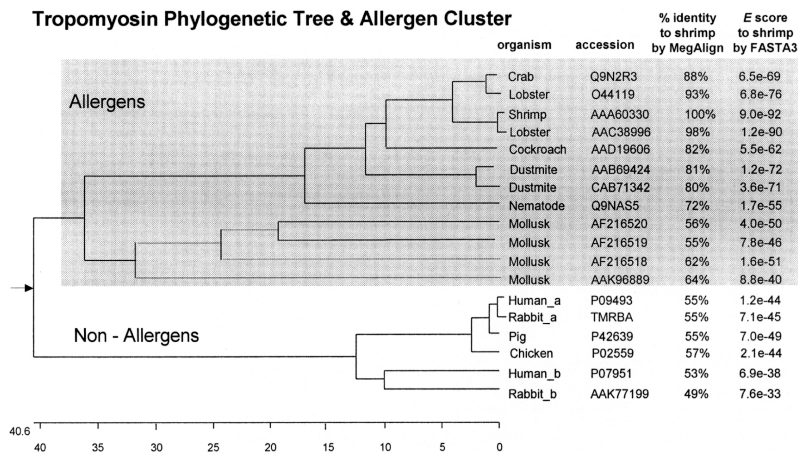


FIGURE 1 Phylogentic dendrogram of tropomyosins. A phylogenetic dendrogram of the predicted evolutionary distances between tropomyosins of the various species was calculated by a clustal alignment using the MegAlign program (DNASTAR, Inc.). The percent identity (calculated by MegAlign), compared with shrimp tropomyosin (Met e 1), is shown along with the *E* score values (calculated by FASTA3). *E* score values are shown in common exponential notation with *e* = 10, e.g., 6.5e-69 is 6.5×10^{-69} . Allergic responses have been documented to tropomyosins of each of the invertebrate species shown (gray box), but none has been reported against the vertebrate proteins.

possible cross-reactive matches, and there are special cases of highly similar proteins that do not cause allergic cross-reactions.

Analysis of Birch Pollen Homologues

An evaluation of sequence alignment search results for cross-reactive birch pollen and food proteins with shared conformational epitopes was performed using both the FASTA algorithm and scanning window searches of 6, 7, or 8 identical amino acids. Examples of closely and more distantly related food-plant homologues of Bet v 1 (apricot allergen Pru ar 1, accessionU93165; celery allergen Api g 1.02, accession Z75662) and Bet v 2 (Lychee fruit profilin, accession AY049013; apple profilin Mal d 4, accession AF129427) proteins were evaluated by FASTA with ALLERGENSEARCH. We assume that many individuals become sensitized by airway exposure to birch pollen that leads to allergic reactions to homologues in foods, though the reverse could be true. An estimated 35% to 53% of birch pollen allergic individuals may have

IgE-mediated reactions when consuming hazelnuts, 25% to 47% with apples, and lower percentages to peaches, cherries, almonds, celery, carrots and potatoes (10).

The Bet v 1 homologues are members of the pathogenesis-related protein family and occasionally induce systemic allergic reactions. Pru ar 1 (apricot) was found to be approximately 70% identical to Mal d 1 (apple), Pru av 1 (cherry), Pyr c 1 (pear), and Cor a 1 (hazelnut), and roughly 60–65% identical to pollen allergens Bet v 1 (birch), Aln g 1 (alder), and Car b 1 (hornbeam). All these share short contiguous amino acid matches of 6–8 identical residues. Celery contains two Bet v 1 homologues, Api g 1.01 which does not share contiguous amino acid matches with Bet v 1, and Api g 1.02 which does. Both celery proteins are approximately 50% identical in overall sequence to Bet v 1. Evaluation of the E scores from these comparisons indicated that values greater than 2×10^{-17} are unlikely to identify a cross-reactive protein.

The heat-labile profilin family of proteins includes the prototypical Bet v 2 pollen allergen and a taxonomically diverse group of homologues that share high percentages of identity. Many pollen, fruit, and nut profilins (e.g., Mal d 4 [apple], Lychee fruit, and Ara h 5 [peanut] proteins) share >75% identity with E scores smaller than 1×10^{-35} and commonly displayed matches of 8 or more contiguous amino acids. Based on various reported clinical responses to various foods, the level of significance for potentially cross-reactive matches is probably greater than 35% identity over the entire protein, or at least 80 amino acids and an E score less than 1×10^{-14} .

DISCUSSION

Since the mid-1990s, Monsanto has used the FASTA algorithm and the 8-contiguous amino acid scan methods to evaluate potential products against a proprietary allergen database (16). These searches have proven useful in identifying potential allergens that were being considered for development (unpublished, Monsanto). The third and most controversial component of a successful searching strategy is the interpretation of a significant match (26). The standing paradigm is that a match of 8 or more contiguous amino acids with a known allergen would indicate that a protein is sufficiently similar to consider it may cause cross-reactions, and it should be tested further with allergic patient sera and if warranted, in vitro challenges (6). However, Aalberse has noted that proteins sharing less than 50% identity are unlikely to be cross-reactive, and only when proteins share more than 70% identity does cross-reactivity become common (18). Yet the FAO/WHO scientific advisory panel suggested using 35% identity over 80 or more amino acids and/or 6 identical, contiguous amino acids as indicators of potential cross-reactivity (43). We

recently tested selected proteins using FASTA3 and searches for 6 or 7 contiguous identical amino acid matches and found an unacceptably high rate of false-positive matches (27). Therefore, we performed the current analysis to broaden the evaluation. Clearly more work is needed to provide additional refinement of the criteria that would lead to the decision to perform difficult clinical testing of a protein if appropriate allergic subjects can be located, or reject a potentially beneficial protein-product.

This article presents evidence that even 35% identity is very conservative for identifying potentially cross-reactive proteins. Additionally, these data confirm that matches of 6 contiguous amino acids are relatively common and do not provide reasonable predictions of potential allergenicity. These data also indicate that a FASTA search alone may be sufficient to identify potentially cross-reactive proteins when a reasonably low percent identity (35% to 50%) over the length of a protein, or an *E* score of less than 1×10^{-14} , is used as the minimum cut-off.

CONCLUSIONS

The identification of the potentially allergenic proteins that could be transferred into food crops through biotechnology is aided by computer algorithms to assess whether the protein is known to be an allergen or is sufficiently similar to known allergens to be considered potentially cross-reactive. The FASTA3 algorithm provides a fast and thorough comparison of the query sequence to those in an allergen database. FASTA3 calculates two numbers that are good indicators of potentially significant similarity/identity. One is an *E* score that is commonly used to identify probable evolutionary homologues. The second is the calculated percent identity over the best-aligned segment of two similar sequences. While more evaluation is necessary to set scientifically justified cut-off scores, it appears from our comparison of a few well-known allergens and proteins unlikely to be allergens, that an *E* score less than $\sim 1 \times 10^{-14}$ and/or a percent identity of greater than 35% to 50% may be reasonable values to indicate possible cross-reactivity. Search results producing lower *E* scores, or higher percent identities, indicate a probable need for further testing. However, we recommend that the data leading to the classification of a matched "allergen" from the database be examined to make sure that inclusion of the "allergen" sequence in the database not be based on inference alone, but be supported by clinical data and tests of avidity, as there is little evidence that some proteins in the database cause clinical disease. We also recommend that any search for isolated short regions of identity be restricted to matches of 8 or more contiguous identical amino acids.

REFERENCES

1. Poehlman, J.M. 1979. *Breeding Field Crops*, 2nd ed. Westport, CT: Avi Publishing, 173–175.
2. Harrison, L.A., Bailey, M.R., Naylor, M.W., Ream, J.E., Hammond, B.G., Nida, D.L., Burnette, B.L., Nickson, T.E., Mitsky, T.A., Taylor, M.L., Fuchs, R.L., and Padgett, S.R. 1996. *J Nutr* 126(3):728.
3. Perlak, F.J., Oppenhuizen, M., Gustafson, K., Voth, R., Sivasupramaniam, S., Heering, D., Carey, B., Ihrig, R.A., and Roberts, J.K. 2001. *Plant J* 27(6):489.
4. U.S. Food and Drug Administration. 1992. Statement of policy: Foods derived from new plant varieties. *Federal Register* 57:22984.
5. Nordlee, J.A., Taylor, S.L., Townsend, J.A., Thomas, L.A., and Bush, R.K. 1996. *New England Journal of Medicine* 334:688.
6. Metcalfe, D.D., Astwood, J.D., Townsend, R., Sampson, H.A., Taylor, S.L., and Fuchs, R.L. 1996. *Crit Rev Food Sci Nutr* 36:S165.
7. Molvig, L., Tabe, L.M., Eggum, B.O., Moore, A.E., Craig, S., Spencer, D., and Higgins, T.J.V. 1997. *Proc Natl Acad Sci USA* 94:8393.
8. Momma, K., Hashimoto, W., Ozawa, S., Kawai, S., Katsube, T., Takaiwa, F., Kito, M., Utsumi, S., and Murata, K. 1999. *Biosci Biotechnol Biochem* 63(2):314.
9. Ayuso, R., Lehrer, S.B., and Reese, G. 2001. *Int Arch Allergy Immunol* 127:27.
10. Breiteneder, H., and Ebner, C. 2000. *J Allergy Clin Immunol* 106:27.
11. Vieths, S., Scheurer, S., and Ballmer-Weber, B. 2002. *Ann N Y Acad Sci* 964:47.
12. Sicherer, S.H. 2001. *J Allergy Clin Immunol* 108:881.
13. Ganglberger, E., Radauer, C., Wagner, S., Riordain, G.O., Beezhold, D.H., Brehler, R., Niggemann, B., Scheiner, O., Jensen-Jarolim, E., and Breiteneder, H. 2001. *Int Arch Allergy Immunol* 125:216.
14. Kazemi-Shirazi, L., Pauli, G., Purohit, A., Spitzauer, S., Froschl, R., Hoffman-Sommergruber, K., Breiteneder, H., Scheiner, O., Kraft, D., and Valenta, R. 1999. *J Allergy Clin Immunol* 105:116.
15. Pearson, W.R., and Lipman, D.J. 1998. *Proc Natl Acad Sci USA* 85:2440.
16. Astwood, J.D., and Fuchs, R.L. 1996. *Mongr Allergy* 32:105.
17. Pearson, W.R. 2000. *Methods Mol Biol* 132:185.
18. Aalberse, R.C. 2000. *J Allergy Clin Immunol* 106:228.
19. Codex Alimentarius Commission. 2002. Report of the third session of the Codex ad hoc intergovernmental task force on foods derived from biotechnology (ALINORM 03/34). Yokohama, Japan: 4–8 March, 2002.
20. Banerjee, B., Greenberger, P.A., Fink, J.N., and Kurup, V.P. 1999. *Infect Immun* 67:2284.
21. Beezhold, D.H., Hickey, V.L., Slater, J.E., and Sussman, G.L. 1999. *J Allergy Clin Immunol* 103:1166.
22. Chatchatee, P., Jarvinen, K.M., Bardina, L., Beyer, K., and Sampson, H.A. 2001. *J Allergy Clin Immunol* 107:379.
23. Reese, G., Ayuso, R., and Lehrer, S.B. 1999. *Int Arch Allergy Immunol* 119:247.
24. Shin, D.S., Compadre, C.M., Maleki, S.J., Kopper, R.A., Sampson, H.A., Huang, S.K., Burks, A.W., and Bannon, G.A. 1998. *J Biol Chem* 273:12753.
25. Rabjohn, P., Helm, E.M., Stanley, J.S., West, C.M., Sampson, H.A., Burks, A.W., and Bannon, G.A. 1999. *J Clin Invest* 103:535.
26. Gendel, S.M. 1998. *Adv Food Nutr Res* 42:45.
27. Hileman, R., Silvanovich, A., Goodman, R.E., Rice, E.A., Holleschak, G., Astwood, J.D., and Hefle, S.L. 2002. *Int Arch Allergy Immunol* 128:280.

28. Gendel, S.M. 1998. *Adv Food Nutr Res* 42:63.
29. Benson, D.A., Boguski, M.S., Lipman, D.J., and Ostell, J. 1997. *Nucl Acids Res* 25:1.
30. Stoesser, G., Sterk, P., Tuli, M.A., Stoehr, P.J., and Cameron, G.N. 1997. *Nucleic Acids Res* 25:7.
31. George, D.G., Dodson, R.J., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Sidman, K.E., Srinivasarao, G.Y., Yeh, L.S.L., Arminski, L.M., Ledley, R.S., Tsugita, A., and Barker, W.C. 1997. *Nuc Acids Res* 25:24.
32. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1999. *Eur J Biochem* 80:319.
33. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. *Nuc Acids Res* 28:235.
34. Bairoch, A., and Apweiler, R. 1997. *Nuc Acids Res* 25:31.
35. Schuler, G.D., Epstein, J.A., Ohkawa, H., and Kans, J.A. 1996. *Methods Enzymol* 266:141.
36. Fujita, C., Moriyama, T., and Ogawa, T., 2001. *Int. Arch. Allergy Immunol.* 125:44.
37. Gardner, R.C., Richards, K.D., Carrucan, G., Bradbury, L.E., and Delhaize, E. 1999. Unpublished mRNA (accession # U91981) and protein (accession # AAD10496) sequences in NCBI.
38. Ferguson, A. 1996. Gluten-sensitive enteropathy (celiac disease). In *Food allergy: Adverse reactions to foods and food additives*, eds. D.D. Metcalfe, H.A. Sampson, and R.A. Simon. Cambridge, MA: Blackwell Science, pg. 281.
39. Henikoff, S., and Henikoff, J.G. 1992. *Proc Natl Acad Sci USA* 89:10915.
40. Taylor, S.L., Hefle, S.L., Bindslev-Jensen, C., Bock, S.A., Burks, A.W., Christie, L., Hill, D.J., Host, A., Hourihane, J. O'B., Lack, G., Metcalfe, D.D., Moneret-Vautrin, D.A., Vadas, P.A., Rance, F., Skrypec, D.J., Trautman, T.A., Yman, I.M., and Zieger, R.S. 2002. *J Allergy Clin Immunol* 109:24.
41. Tanaka, L.G., El-Dahr, J.M., and Lehrer, S.B. 1999. *J Allergy Clin Immunol* 107(4):744.
42. Pastorello, E.A., Farioli, L., Pravettoni, V., Spano, M., Scibola, E., Trambaioli, C., Giuffrida, M.G., Ansaloni, M.G., Godovac-Zimmermann, J., Conti, A., Fortunato, D., and Ortolani, C. 2000. *J Allergy Clin Immunol* 106(4):744.
43. FAO/WHO expert consultation. 2001. Evaluation of allergenicity of genetically modified foods, http://www.who.int/fsf/Gmfood/Consultation_Jan2001/report20.pdf. Rome, Italy: 22–25 January, 2001.
44. Leung, P.S.C., Chen, Y.-C., Gershwin, M.E., Wong, S.H., Kwan, H.S., and Chu, K.H., 1998. *Journal of Allergy Clin Immunol* 102:847.
45. Arruda, L.K., Vailes, L.D., Ferriani, V.P.L., Santos, A.B.R., Pomes, A., and Chapman, M.D. 2001. *J Allergy Clin Immunol* 107:419.