

Bioinformatic Methods for Allergenicity Assessment Using a Comprehensive Allergen Database

Ronald E. Hileman^a Andre Silvanovich^a Richard E. Goodman^a
Elena A. Rice^a Gyula Holleschak^a James D. Astwood^a Susan L. Hefle^b

^aMonsanto Company, Product Safety Center, St. Louis, Mo., and ^bUniversity of Nebraska, Food Allergy Research and Resource Program, Lincoln, Nebr., USA

Key Words

Allergen · Allergy · *Bacillus thuringiensis* ·
Bioinformatics · Biotechnology · Database · Food
allergy · IgE epitope · Genetically modified crops

Abstract

Background: A principal aim of the safety assessment of genetically modified crops is to prevent the introduction of known or clinically cross-reactive allergens. Current bioinformatic tools and a database of allergens and gliadins were tested for the ability to identify potential allergens by analyzing 6 *Bacillus thuringiensis* insecticidal proteins, 3 common non-allergenic food proteins and 50 randomly selected corn (*Zea mays*) proteins. **Methods:** Protein sequences were compared to allergens using the FASTA algorithm and by searching for matches of 6, 7 or 8 contiguous identical amino acids. **Results:** No significant sequence similarities or matches of 8 contiguous amino acids were found with the *B. thuringiensis* or food proteins. Surprisingly, 41 of 50 corn proteins matched at least one allergen with 6 contiguous identical amino acids. Only 7 of 50 corn proteins matched an allergen with 8 contiguous identical amino acids. When assessed for overall structural similarity to allergens,

these 7 plus 2 additional corn proteins shared $\geq 35\%$ identity in an overlap of ≥ 80 amino acids, but only 6 of the 7 were similar across the length of the protein, or shared $>50\%$ identity to an allergen. **Conclusions:** An evaluation of a protein by the FASTA algorithm is the most predictive of a clinically relevant cross-reactive allergen. An additional search for matches of 8 amino acids may provide an added margin of safety when assessing the potential allergenicity of a protein, but a search with a 6-amino-acid window produces many random, irrelevant matches.

Copyright © 2002 S. Karger AG, Basel

Introduction

Agricultural biotechnology has been used to engineer a variety of agronomically important crops, such as corn, soybean, potato and cotton. Some are genetically modified to be resistant to insect damage. Insect-resistant varieties are generally desirable because of enhanced yield, reduced need for chemical insecticide and the high degree of pest control. Some crops that are resistant to insects have been produced by genetic modification with genes obtained from the bacterium *Bacillus thuringiensis*.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2002 S. Karger AG, Basel
1018–2438/02/1284–0280\$18.50/0

Accessible online at:
www.karger.com/journals/iaa

Correspondence to: Dr. Susan L. Hefle
University of Nebraska, Food Allergy Research and Resource Program
255 Food Industry Building
Lincoln, NE 68583-0919 (USA)
Tel. +1 402 472 4430, Fax +1 402 472 1693, E-Mail shefle1@unl.edu

B. thuringiensis is a gram-positive bacterium commonly present in soil. Many different strains of *B. thuringiensis* have been shown to produce crystal proteins or inclusion bodies that are specifically effective in controlling certain orders and species of insect pests [1]. *B. thuringiensis* insecticidal proteins have been used safely for more than 40 years as biopesticides [2]. During this time, no reports of adverse human health or environmental effects have been documented [2–4]. They control larval pests, yet are safe to mammals, birds, fish and beneficial insects [3, 5]. *B. thuringiensis* crystal proteins have been generally classified based on their insecticidal activity. For example, Cry1, Cry2, Cry3 and Cry4 proteins are toxic to lepidopteran, lepidopteran/dipteran, coleopteran and dipteran pests, respectively [6, 7]. The crops that have been genetically modified to produce the *B. thuringiensis* insect control proteins Cry1Ab, Cry1Ac, Cry2Aa, Cry2Ab, Cry3Aa and Cry3Bb, are thus resistant to predation by specific insects.

Commercialization of agricultural biotechnology products requires premarket authorization. In the United States, the US Department of Agriculture, the Food and Drug Administration and the Environmental Protection Agency regulate genetically modified crops. Other countries typically have one or two agencies with similar responsibilities. Extensive food, feed and environmental safety assessments are conducted for each product prior to commercialization; this includes an allergy assessment for each introduced protein (e.g. a Cry protein) in the genetically modified crop. A principal aim is to prevent the introduction of known allergens or clinically cross-reactive incomplete allergens. An incomplete allergen is one that may elicit allergic symptoms, but does not sensitize individuals *de novo* [8]. The use of bioinformatic algorithms and specialized databases containing known and suspected allergens [9–13] provides important and effective tools to identify known and potentially cross-reactive allergens that may pose a risk to those with preexisting allergies. If a significant match is identified, the introduced protein must be evaluated further using sera from individuals who are allergic to the similar allergenic protein to determine if it is likely to cause cross-reactions.

Exposure to allergens in foods poses a risk to sensitized, allergic individuals. Type I (immediate hypersensitivity) reactions occur when a sufficient quantity of individual allergenic proteins, each bound by two IgE antibodies, cross-links IgE receptors on the surface of mast cells and basophils, thereby stimulating these cells to release histamine and other allergenic mediators [14]. Sensitization to produce the IgE antibodies requires a

third allergen-specific recognition sequence, a CD4+ T cell epitope to provide appropriate (Th2) costimulation and cytokines to antigen specific B cells to induce IgE production [15]. Cross-reactions may be induced in sensitized individuals by a different protein if it contains IgE-binding epitopes that are identical or nearly identical to those of the sensitizing protein, as is the case for incomplete allergens, which do not include T cell epitopes for the induction of the immune response [8]. While cross-reactive structures frequently have lower affinity than the immunogen for the antibody and are therefore typically less effective at inducing an immune response [16], clinical evidence has demonstrated that foods containing proteins that share greater than 40% identity and appear to share cross-reactive IgE epitopes can induce clinical responses in some individuals [17]. However, it has not been conclusively proven that individuals with apparent cross-reactivity were not sensitized to both allergens or that those proteins are responsible for the clinical reactivity [18].

It is important to recognize that there is insufficient information to exclude the possibility that a protein that does not share sequence similarity to any known allergen will ever cause an allergic reaction. At this time there is no well-established evidence that specific allergenic motifs or structures exist that are allergenic in all contexts. Therefore, the most appropriate bioinformatic approach at this time seems to be reliance on a comparison to known allergens. Further, the additional considerations of stability to pepsin digestion and the abundance of the protein in the food may provide useful information in that regard [9, 13].

Food produced from genetically engineered crops should be as safe as that produced from traditional crops. Thus, the protection of food-allergic individuals from unexpected and unwanted exposures to offending foods or allergens is a primary health concern. Bioinformatic tools can be used to address this problem by asking two questions:

(1) *Is the structure of an introduced protein sufficiently similar to a known allergen to suspect that it is an allergen or would be cross-reactive?*

Current bioinformatic algorithms compare linear (local) sequences and global (three-dimensional) predictions of structure to estimate similarity. Programs such as FASTA [19] (an algorithm used to find local high-scoring alignments between a pair of protein or nucleotide sequences) and BLASTP [20] (basic local alignment search tool for proteins) can be effectively used to predict functional similarity. The basic principle of these bioinformatic algorithms is that if two proteins share sufficient linear

sequence similarity, they will also share three-dimensional structure and therefore functional homology. By definition, homologous proteins share secondary structure and common three-dimensional folds [19]. Homologous proteins are more likely to share allergenic cross-reactive conformational and linear epitopes than unrelated proteins due to similarity in sequence and structure. Because the degree of similarity between homologues varies widely, data derived from these algorithms need to be carefully evaluated in order to predict potential cross-reactivity. While related (homologous) proteins may share only 25% amino acid identity in a 200-amino-acid overlap [19], this is not generally sufficient to indicate IgE-mediated cross-reactivity [18]. Indeed, allergenic cross-reactivity caused by proteins sharing conformational or linear epitopes is rare at 50% identity and typically requires >70% amino acid identity across the full length of the protein sequences [8]. Such high levels of identity are readily detected using either FASTA or BLASTP. Additionally, proteins closely related to gliadins or glutenins, the proteins that trigger celiac disease, a non-IgE-mediated allergic disorder, can be easily identified using FASTA or BLASTP algorithms.

(2) Does the introduced protein share a short linear sequence(s) identity with an allergen?

Sequence alignment tools such as FASTA are valuable for addressing overall similarity, but proteins structurally unrelated to allergens may still contain small immunologically significant epitopes that may elicit allergic symptoms in sensitized individuals. Thus a second comparison methodology accounting for linear epitope structure has been proposed [10, 12, 13]. The presence of small (local) regions of identity could indicate the presence of common IgE-binding epitopes shared between a query sequence and a known allergen. For practical (optimal signal to noise ratio) and theoretical (approximate average linear epitope length) reasons, the current approach recommends searching for 8 or more contiguous amino acid identities [13]. In developing this approach, it was recognized that while IgE epitope maps are available for some allergens, relatively few food allergens have been fully mapped using in vitro IgE-binding studies with sera from allergic individuals. Although some linear IgE epitopes may be as short as 5 amino acids [21, 22], the majority of characterized IgE linear epitopes are 8 amino acids or longer [23]. However, only a few reports have demonstrated the avidity (of affinity) of these antibodies, or their in vivo allergic significance for these small epitopes. It is clear from some reports that high avidity binding requires 8 or more amino acids [22, 24].

All allergens do not share common structural characteristics and there is no indication of a common structure within B cell or T cell epitopes [25]. Consequently, the predictive value of a bioinformatic assessment depends on the completeness of the allergen database [11]. The identification and molecular characterization of food and other allergens have resulted in the availability of DNA and protein sequences of many important allergens in public databases and scientific literature, albeit the data are often difficult to find [11, 12]. The construction of an inclusive allergen and gliadin protein sequence database (ALLERGEN3) compiled from publicly available sequence databases and current literature is described. Further, the efficiency of searching for similarities, either by FASTA or the sliding window of 8 amino acids is greatly enhanced by using an allergen database as compared to all publicly available protein sequences.

While the focus of the bioinformatic search is on proteins that are bound by IgE, gluten-associated proteins cause celiac disease (gluten-sensitive enteropathy), a non-IgE-mediated allergic reaction, in genetically predisposed individuals. These proteins are gliadins and glutenins in wheat, triticale, rye and barley [26]. Gliadin- and glutenin-like proteins that may cause celiac disease can be identified using the same bioinformatic tools as those used to identify type I allergens. Thus clinically relevant gliadins and glutenins have been included in the allergen database.

The most recent FAO/WHO (2001) recommendations for allergenicity assessment include the use of a 6 amino acid window in a pair-wise fashion with known allergens to identify potential cross-reactive proteins in genetically modified crops used for food production [27]. In an attempt to determine if this is a justifiable alternative to the currently used 8-amino-acid window, 3 commonly consumed nonallergenic proteins and 6 proteins derived from the bacterium *B. thuringiensis* were compared to sequences in the ALLERGEN3 database using window sizes of 5–8 as well as a FASTA search. Additionally, 50 randomly selected proteins from corn (*Zea mays*) were compared to sequences in the assembled ALLERGEN3 database by scanning with a 6, 7, or 8 amino acid window and by a FASTA comparison. Corn was chosen because it is one of the least allergenic of the major grain crops and there are only one or two proteins in corn that have been identified as clinically relevant allergens [28].

Materials and Methods

All analyses were performed using the UNIX-based Genetics Computer Group (GCG) software package (version 10.0, Madison, WI) on a personal computer supported with Reflection X Client Manager network software (version 7.20, WRQ, Inc. Seattle, Wash., USA; the UNIX computer-operating system was originally termed 'UNiplexed Information and Computing Service'). Database construction was performed in the SeqLab environment of GCG. Searches were performed from an X terminal window.

Sequence Database Preparation

The allergen and gliadin sequence database (ALLERGEN3) was assembled from public domain databases including Genbank and EMBL [29], PIR [30], the NRL3D version of RCSB PDB [31, 32] and SwissProt [33]. 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 [13], allergens recognized by the IUIS (<ftp://biobase.dk/resourcespub/who-iuis/allergen.list>) and from searches of current literature available on PubMed and Entrez (<http://www.ncbi.nlm.nih.gov/>) [34]. For publications reporting only an N-terminal sequence, the sequence was added if 8 or more residues were provided. Duplicate sequences were identified using a FASTA search with each individual sequence. Identical entries were removed, but unique isoforms were retained (e.g. there are 35 entries of Bet v 1). The selected allergen, gliadin and glutenin sequences were compiled in the searchable database, called ALLERGEN3, containing 635 protein sequences. ALLERGEN3 is available and may be used in similarity searches at the website <http://www.allergenonline.com/>.

Selection of Query Protein Sequences

Six publicly available insecticidal *B. thuringiensis* protein sequences, Cry1Ab10, Cry1Ac9, Cry2Aa1, Cry2Ab2, Cry3Aa4 and Cry3Bb1 (accession No. A29125, U89872, M31738, X55416, M30503 and M89794, respectively) are representative of the *B. thuringiensis* proteins expressed in some approved genetically modified insect-resistant crops. Three proteins that are commonly consumed and not known to be allergenic were also selected for comparison (spinach rubisco large and small subunits, and bovine myoglobin). In addition, 4,116 corn protein sequences were compiled and randomly ordered from an Entrez [34] protein search using a keyword search for 'zea or maize'. The first 50 sequences found to be of 95 amino acids or greater in length were evaluated in these analyses.

Sequence Database Searches for Structural Similarity

Structural similarities shared between query sequences and the ALLERGEN3 sequence database were assessed using the FASTA algorithm [35, 36]. FASTA comparisons are initiated by aligning the first match of a specific word size (*k*-tuple parameter), in this case two. The alignment is then extended based on the chosen scoring matrix. Specific parameters used for these analyses included an expectation threshold of 10, a gap creation penalty of 12 and gap extension penalty of two. FASTA comparisons were performed using the BLOSUM50 scoring matrix [37]. Multiple alignments are made between the query sequence and each sequence in the database with a score calculated for each alignment. Only the best-scoring alignment is reported for each database sequence. The BLOSUM matrix series

[37] was derived from a set of aligned, ungapped regions from protein families, called the BLOCKS database. Sequences from each block were clustered based on the percent of identical residues in the alignments [38]. The BLOSUM50 matrix will identify blocks of conserved residues that are at least 50% identical. BLOSUM50 works well for identifying sequence similarities that include gaps, and thus recognizes distant evolutionary relationships [19].

The extent of similarity was evaluated by visual inspection of the aligned sequences, the calculated percent identity and *E* score (expectation score). The *E* score reflects the degree of similarity between a pair of sequences based on matches of identical or functionally similar amino acids. Similar amino acids are structurally related and share polar, hydrophobic, or charged states. Such substitutions are referred to as 'conservative' since they are unlikely to change the structure of the protein, and by inference the function, of homologous proteins. The calculated *E* score depends on the overall length of joined (gapped) local sequence alignments, the quality (percent identity, similarity) of the overlap and the size of the database [35]. A large *E* score (e.g. > 1) indicates a low degree of similarity between the query sequence and the sequence from the database. For polypeptides, a sequence similarity may indicate homology (i.e., sequences derived from a common ancestral gene with homologous function). Sequences that share extensive amino acid sequence identity and/or similarity throughout the entire alignment or identified domains will have very small *E* scores and will often share similar overall structures. Proteins that are nearly identical are very likely to share cross-reactive IgE (or IgG) and T cell epitopes.

Sequence Database Searches for Short Peptide Matches

In addition to the FASTA comparisons used to assess overall structural similarity, a pairwise comparison algorithm referred to as ALLERGENSEARCH was used to find matches of linearly contiguous amino acid identities between the query sequence and sequences within the ALLERGEN3 database. The algorithm was run from a UNIX terminal window in GCG. This program compares the query sequence to each protein sequence in a database using a sliding window of *n* amino acids, with an *n* - 1 amino acid overlap relative to the preceding window. Values of '*n*' equal to 5, 6, 7 and 8 amino acids were used in this analysis, where specified.

Results

Construction of the ALLERGEN3 (Allergen and Gliadin) Database

Keyword searches of publicly available genetic databases, using the term 'allergen', were used to retrieve allergen sequences. This approach was insufficient for retrieving allergen sequences. For example, many of the entries retrieved using this method were duplicates due to the redundancy of the public databases and those were removed. Entries determined to be irrelevant based on published clinical, IgE binding or significant structural similarity data to support the declaration of allergenicity were also removed. In addition, allergen sequences were identified through a literature review and Internet searches and

Table 1. Summary results using FASTA for nonallergenic dietary proteins and selected Cry proteins

Query protein	Accession No.	Query sequence length, aa	Best-matched allergen accession No.	Allergen sequence length, aa	E score	Identity %	Length of overlap, aa
Bovine myoglobin	MYBO	154	Chi t 1.01 (P02229)	151	0.0028	24.8	137
Spinach rubisco SSU	P00870	123	Garden pea (Q41043)	258	0.64	42.9	21
Spinach rubisco LSU	RKSPL	475	Hev b 6 (P02877)	204	1.2	37.0	46
Cry1Ab10	A29125	1,156	Cor a 1.0102 (X71000)	160	1.0	27.6	98
Cry1Ac9	U89872	1,178	Cor a 1.0102 (X71000)	160	1.2	28.6	98
Cry2Aa1	M31738	633	Cop c 1 (CAB39376)	81	1.5	30.8	52
Cry2Ab2	X55416	633	Cop c 1 (CAB39376)	81	0.44	32.7	52
Cry3Aa4	M30503	644	Asp f 4 (CAA04959)	286	0.68	29.4	102
Cry3Bb1	M89794	652	Asp f 9 (CAA11266)	302	3.4	33.8	65

The query protein name, accession number and length is shown. Each protein was compared to the ALLERGEN3 database. The identity of the allergen found to produce the best alignment, the sequence length of that allergen (aa), expectation (*E*) score, percent identity and length of the overlapped (aligned) regions are listed.

added to the database. From this analysis, 635 protein sequence entries were deemed to be allergens. The compiled allergen sequence database is available at the web site <http://www.allergenonline.com/>, and includes links to the corresponding public database accessions. This database is considered to be conservative relative to the identification of potential allergens because putative (i.e. not clinically proven) allergens were included along with well-characterized, proven allergens. It will be important to update this database periodically as additional knowledge is gained regarding the allergenicity of the proteins currently included and as new allergens are characterized and their sequences published.

Assessment of the ALLERGEN3 Database to Identify a Known Allergen

Searching the ALLERGEN3 Database with Known Nonallergenic Sequences. Three polypeptide sequences that represent commonly consumed, nonallergenic food proteins were compared to the allergen database: the small and large subunit of spinach rubisco and bovine myoglobin. The highest-scoring similarity identified for

each of these proteins is shown in table 1. The *E* score, percent identity and percent similarity values for spinach rubisco subunits did not suggest structural homology to any alleged allergen or gliadin [39]. Bovine myoglobin received a modest *E* score (0.0028) in the alignment with the *Chironomus thummi* (chironomid midge) erythrocrucorin III allergen Chi t 1.01 (accession No. P02229) but shared only 24.8% amino acid identity over the 137 amino acid overlap (of 154 amino acids). These scores suggest that myoglobin and Chi t 1.01 proteins may be homologous [40]. Indeed, Chi t 1.01 is a monomeric hemoglobin and has previously been shown to share similarity to other globin proteins [41]. While this limited similarity indicates that these proteins are distantly related, the low observed percent identity does not indicate that these proteins share cross-reactive IgE-binding epitopes [18].

The sequence of each nonallergenic food protein was also compared to the ALLERGEN3 database to search for linearly contiguous amino acid identities using the ALLERGENSEARCH algorithm. None of the protein sequences contained 8 amino acid identities to any of the sequences in the ALLERGEN3 database (table 2).

Table 2. Effect of the scanning window size on the number of matches identified to allergens, using ALLERGEN-SEARCH

Query protein	Accession No.	Length, aa	Search window size, aa			
			5	6	7	8
Bovine myoglobin	MYBO	154	11	1 ^a	0	0
Spinach rubisco SSU	P00870	123	5 ^b	0	0	0
Spinach rubisco LSU	RKSPL	475	58	1 ^c	0	0
Cry1Ab10	A29125	1,156	103	13	2 ^d	0
Cry1Ac9	U89872	1,178	86	22	6 ^e	0
Cry2Aa1	M31738	633	46	2 ^f	0	0
Cry2Ab2	X55416	633	61	2 ^f	0	0
Cry3Aa4	M30503	644	30	1 ^g	0	0
Cry3Bb1	M89794	652	38	5 ^h	0	0

The query protein names and accession numbers identify the common food protein, or bacterial insecticidal protein that were compared to the allergens in ALLERGEN3. The length of the query protein is shown in amino acids (aa). Values reported for the number of amino acids under each search window size correspond to the number of allergens that had an identical match to some segment of the query protein at the corresponding search window size.

^a Matched allergen: bovine β -lactoglobulin (CAA32835).

^b Matched allergens: bovine serum albumin (AAA51411), dust mite Der f 14 (BAA04558), storage mite Eur m 14 (AAF14270), midge Chi t 5 (P02224), and avocado endochitinase Per a 1 (CAB01591).

^c Matched allergen: peanut lectin hemagglutinin (S14765).

^d Matched allergens: garden pea pollen-like allergen (CAA59470) and cucumber expansin (AAB37746).

^e Matched allergens: mountain cedar isoforms Jun a 1-2 (AAF80164), Jun a 1-1 (AAD03608) and Jun v 1-1 (AAF80166), garden pea pollen-like allergen (CAA59470), cucumber expansin (AAB37746) and Cup a 1 (CAB62551).

^f Matched allergens: leucine zipper protein, mushroom Cop c 1 (CAB39376) and mustard 2S albumin Bra j 1 (P80207).

^g Matched allergen: dust mite protein Eur m 14 (AAF14270).

^h Matched allergens: dust mite protein Eur m 14 (AAF14270), *Aspergillus niger* xylosidase (AAD13106, CAB06417), *Aspergillus fumigatus* heat shock protein (P40292), and *Cladosporium herbarum* aldehyde dehydrogenase (S43114).

The effect of decreasing the size of the window searched below 8 contiguous amino acids on the number of hits obtained was investigated using a window size of 5, 6, 7 or 8 to search for exact matches between any portion of these query sequences and any sequence in the ALLERGEN3 database (table 2). All proteins searched using a window size of 5 amino acids matched between 5 and 58 allergens, suggesting that these hits represented noise. The number of random matches decreased dramatically when larger window sizes were used. No matches were observed for either rubisco subunit or myoglobin when a window size of 8 amino acids was used.

Assessment of B. thuringiensis Protein Structural Similarity to Allergens and Gliadins. Six publicly available *B. thuringiensis* protein sequences similar to those used in approved genetically modified crops were compared to the ALLERGEN3 database (table 1). Cry1Ab10 and

Cry1Ac9 proteins are nearly identical (89%) to each other and both identified the *Corylus avellana* (European hazelnut) tree pollen Cor a 1.0102 (Accession No. X71000) as having the highest score. The optimized alignments for both Cry1A proteins were to the same 98-amino-acid region of Cor a 1.0102. The alignments contained multiple gaps (3 and 5, respectively), shared only 28–29% amino acid identity to Cor a 1.0102 and were short relative to the length of the Cry1A proteins (>1,150 amino acids). The low percent identity for both of these Cry1A proteins did not indicate structural homology to Cor a 1.0102 [39] as supported by the relatively large observed *E* scores (1.0 and 1.2).

The highest-scoring similarity observed to both the Cry2Aa1 and Cry2Ab2 proteins, which share 88% identity to each other, was with the *Coprinus comatus* (shaggy cap mushroom) aeroallergen Cop c 1 (accession No.

CAB39376, table 1). While approximately 31–33% amino acid identity was observed between the two Cry2 proteins and Cop c 1, the aligned region was only 52 amino acids long. These results do not indicate significant structural homology [39], again supported by the relatively large observed *E* scores (1.5 and 0.44).

The highest-scoring similarity observed to the Cry3Aa4 protein was with the *Aspergillus fumigatus* minor allergen, Asp f 4 (accession No. CAA04959, table 1). This protein has been reported to cause an allergic response only in patients with allergic bronchopulmonary aspergillosis [42]. Approximately 29% amino acid identity was shared between the Cry3Aa4 and Asp f 4 proteins in the 102-amino-acid-aligned region, with an *E* score of 0.68. The Cry3Aa4 and Cry3Bb1 protein sequences share 69% amino acid identity with each other and unlike the Cry1 and Cry2 pairs of proteins, different best-scoring allergens were identified. The highest-scoring similarity observed to Cry3Bb1 protein was with the Asp f 9 aeroallergen (accession No. CAA11266, table 1). These proteins shared approximately 34% amino acid identity with an *E* score of 3.4 for the 65 amino acid-aligned region. Neither Cry3Aa4 nor Cry3Bb1 protein was considered to be homologous to an allergenic protein due to the observed short alignments, low identity values and the relatively large observed *E* scores.

The sequence of each *B. thuringiensis* protein was also compared to the ALLERGEN3 database to search for linearly contiguous amino acid identities using the ALLERGENSEARCH algorithm (table 2). None of the 6 *B. thuringiensis* protein sequences contained 8 amino acid identities to any of the sequences in the ALLERGEN3 database. The effect of decreasing the window size below 8 contiguous amino acids on the number of hits obtained was investigated. All *B. thuringiensis* proteins searched using a window size of 5 amino acids were matched to several allergens, suggesting that these matches again represented noise. All *B. thuringiensis* proteins matched at least one allergen when a window size of 6 was used. Importantly, the number of matches with a window size of 5 increased markedly as the length of the query protein increased, suggestive of a random match phenomenon. This result was consistent with a previous report that demonstrated that matches between allergens and Cry proteins could be found at 5 and 6 contiguous identical amino acids [10]. While the Cry1A proteins matched two allergen sequences using a window of 7 amino acids, neither of these allergens was identified as having significant similarity by FASTA as the sequences of the query protein and the allergens are very

dissimilar (data not shown), indicating a low probability of allergenic cross-reactivity.

Allergenicity Assessment of 50 Randomly Selected Corn (Z. mays) Protein Sequences

The FASTA sequence alignment tool was used to compare 50 randomly selected corn protein sequences to the ALLERGEN3 sequence database as was previously described for the *B. thuringiensis* insect control proteins. Data presented in table 2 indicated that a window size of 5 amino acids was not discriminatory. Therefore the ALLERGENSEARCH algorithm was performed using only window sizes of 6, 7 and 8 amino acids. A summary of the search results is shown in table 3.

Of the 50 sequences used to search the ALLERGEN3 database, 10 sequences (shaded in table 3) yielded FASTA *E* scores that were likely to reflect some conservation of structure [40]. These *E* scores ranged from of 2.3×10^{-5} to 1.4×10^{-76} . Furthermore, one of these proteins displayed 100% identity along its entire length with a known corn pollen allergen [43] and could therefore be positively identified as an allergen. While the *E* score of 2.3×10^{-5} observed between zein and a wheat gliadin may suggest that these proteins share structural similarity, both are unusually rich in the amino acids glutamine and proline. This alignment was biased and included multiple inserted gaps to create the long overlap. Corn, and therefore zein, is known not to cause celiac disease. The other 8 proteins would require further analysis and testing with sera from appropriate patients to determine whether they would likely constitute a risk for individuals with specific allergies. The results of searches for identical contiguous matches of 6, 7 and 8 amino acids are also presented (table 3) and showed that 41 of the 50 query proteins contained at least 1 identical 6-amino-acid sequence match with an allergen; 15 of the 50 contained at least 1 identical 7 amino acid sequence match with an allergen. All 7 of the 50 sequences that had short linear matches of 8 amino acids with an allergen were also identified by FASTA as having significant *E* scores ($<10^{-7}$). Not all of these, however, had $\geq 35\%$ amino acid identity over ≥ 80 amino acids [27] or greater than 50% amino acids identity over the entire length of the protein [18], which have been suggested as probable lower limits for predicting cross-reactivity.

Table 3. Bioinformatic analyses of 50 randomly chosen corn proteins from public databases, comparing the results of searches for matches of 6, 7, and 8 contiguous identical amino acids (aa) with the results of a FASTA alignment search using the BLOSUM50 scoring matrix against the database of known or suspected (putative) allergens and gliadin sequences

Corn protein (Accession No.)	Corn protein (name)	Length, aa	Matches			'Best' aligned allergen using FASTA		
			6 aa	7 aa	8 aa	E score	% ID (aa overlap)	Accession No.
A55092 ^a	catalase	493	1	0	0	0.0069	22.3 (301)	Am cockroach Per a 1, Cr-PII (U69260)
AAA33523 ^b	16 kDa zein	183	11	1	0	2.3×10^{-5}	35.4 (181)	wheat gamma-gliadin B-I (M11077)
AAA68209 ^a	Sus1 gene product	816	4	0	0	2.5	30.0 (60)	cherry thaumatin-like (U32440)
AAA73960 ^a	kaurene synthase	823	3	0	0	2.2	30.4 (56)	hornbeam major pollen allergen (P38949)
AAA87580 ^a	GAPDH	337	1	0	0	0.27	26.3 (152)	mite allergen Eur m 3 (AF047615)
AAB71078 ^c	ribosomal protein P3a	120	15	7	6	5.4×10^{-8}	69.4 (49)	cladosporium 60S acidic ribosomal P2, minor allergen (P42039)
AAB86960 ^d	profilin	131	20	19	19	3.6×10^{-48}	83.2 (131)	wheat profilin (X89826)
AAC72193 ^a	pyruvate dehydrogenase	374	3	2	0	6.0	31.3 (48)	beta expansin
AAC78468 ^a	RNA polymerase	349	2	0	0	1.2	19.8 (96)	wheat gamma-gliadin B-III (M11335)
AAC79953 ^a	sigma factor							
AAC79953 ^a	anionic peroxidase H	253	2	1	0	0.096	26.7 (101)	mouse urinary protein MUP V (M16360)
AAK26754 ^a	plasma membrane	292	0	0	0	0.16	26.7 (90)	apple lipid transfer protein (AAF26450)
AAK30114 ^a	integral protein ZmPIP-3							
AAK30114 ^a	teosinte branched protein 1	132	2	0	0	4.3	29.2 (72)	aspergillus allergen Asp f 4 (AAF26450)
AAK51777 ^a	MURB-like	207	0	0	0	1.5	28.4 (81)	mite allergen Der f 7 (Q26456)
AAK51786 ^a	transposable element							
AAK51786 ^a	MURB-like protein	207	1	0	0	2.1	28.8 (81)	mite allergen Der f 7 (Q26456)
AAK56122 ^d	hMURB12							
AAK56122 ^d	alpha-expansin 4	197	3	3	3	7.8×10^{-67}	81.2 (197)	pea petal protein similar to pollen allergen (X85187)
AAK56130 ^d	beta-expansin 7	268	25	18	8	2.2×10^{-55}	57.0 (268)	wheat pollen allergen (U91981)
AAK59898 ^a	kaurene synthase A	202	1	0	0	2.2	26.2 (65)	birch pollen allergen Bet v 1 (Z72435)
AAK60245 ^a	teosinte branched	317	10	6	0	1.8	27.7 (83)	salmon parvalbumin (Q91483)
AAK60502 ^a	protein							
AAK60502 ^a	sucrose export defective 1	474	2	0	0	0.14	32.8 (61)	pollen allergen 3-2 <i>Juniperus virginiana</i> (Q91483)
BAA05550 ^a	Lea group 3 protein	221	1	0	0	0.11	24.9 (177)	mollusk allergen (AAB69424)
BAA22410 ^a	Ca+ protein kinase-kinase	452	2	0	0	0.048	33.3 (57)	bermuda grass pollen allergen (A28050)
CAA28734 ^a	40.1 kDa A1 protein	357	3	0	0	0.46	25.3 (170)	pear isoflavone reductase-related protein Pyr c 2 (AF071477)
CAA31221 ^a	cytochrome oxidase su III	265	2	1	0	9.1	38.5 (26)	mold allergen Cop c 1 (AJ132235)
CAA35589 ^a	pyruvate decarboxylase	610	16	4	0	0.22	22.5 (182)	cockroach Per a 1 allergen (U69957)
CAA37038 ^a	dihydrodipicolinate	380	0	0	0	0.42	42.4 (33)	perennial ryegrass pollen Lol p 5 (Q40237)
CAA39438 ^a	ribosomal protein S11	159	14	8	0	7.2	25.2 (111)	soy trypsin inhibitor (CAA45778)
CAA60366 ^a	hypothetical plastid protein	139	0	0	0	3.7	40 (25)	para rubber tree beta glucanase (S65077)
CAA72196 ^a	cytochrome p450	538	4	0	0	0.49	35.3 (51)	storage mite group 4 homologue of D pter. (AF144061)
CAA87634 ^a	unknown function	485	5	0	0	6.4	35.1 (37)	aspergillus allergen (CAB64688)
	w/APELTALA2-like binding domain							

➤

Table 3 (continued)

Corn protein (Accession No.)	Corn protein (name)	Length, aa	Matches			'Best' aligned allergen using FASTA		
			6 aa	7 aa	8 aa	<i>E</i> score	% ID (aa overlap)	Accession No.
CAB56627 ^a	SBP domain protein 1	440	5	0	0	0.92	44.1 (34)	mite hsp-70 (P39674)
CAC18100 ^a	putative legumain	485	1	0	0	0.045	25.9 (189)	mite allergen (P39673)
CAC35069 ^a	VIP3 protein	126	0	0	0	0.05	25.9 (112)	shrimp topomyosin (U08008)
JC1524 ^c	major allergen m1, pollen allergen	191	49	16	15	1.4×10^{-76}	100 (191)	corn pollen allergen (Q07154)
O24578 ^a	adenylosuccinate synthetase precursor	484	2	0	0	1.6	33.3 (45)	mite group 2 (Y12690)
P04705 ^a	zein alpha-precursor	186	8	0	0	0.003	28.2 (174)	wheat alpha-type gliadin (K02068)
P15590 ^d	embryo globulin-1	573	7	1	0	1.1×10^{-41}	36.7 (499)	English walnut vicilin like protein (AAF18269)
P33489 ^a	auxin-binding protein-5	150	11	0	0	1.4	25.8 (66)	white mustard Sin a 1 (X91799)
P40280 ^a	histone H2a	159	0	0	0	0.83	30.2 (96)	fungal 60S acidic ribosomal protein (P42039)
P46251 ^a	actin depolymerizing factor	139	0	0	0	0.45	20.5 (78)	canary grass pollen allergen (P56167)
P51059 ^a	phosphoenolpyruvate carboxylase 2	967	0	0	0	0.52	20.9 (201)	mollusk tropomyosin
PWZMAM ^a	F1 ATPase mitochondria	508	2	0	0	1.0	25.6 (117)	fungi manganese superoxide dismutase (U53561)
S12043 ^a	leucoanthocyanidin dioxygenase	395	6	0	0	0.52	26.5 (136)	canary grass pollen Pha a 5.1 (P56164)
S30062 ^d	polygalacturonase	95	7	7	7	5.1×10^{-31}	70.6 (95)	polygalacturonase (CAB42866)
S37379 ^a	catalase 3	496	2	0	0	0.33	20.4 (113)	wheat gamma gliadin BI (M11336)
S58532 ^a	chloroplast matK protein	544	3	0	0	0.27	24 (101)	wheat alpha/beta gliadin class AII (M10092)
T02242 ^a	GTP-binding protein rab2	209	2	0	0	4.2	31.6 (57)	parasitic amoeba profilin (1PRQ)
T02763 ^a	probable malate dehydrogenase	652	1	0	0	0.62	23.9 (71)	canary grass pollen PHA A 5.1 (P56164)
T02990 ^d	cinnamyl alcohol dehydrogenase	367	5	1	1	1.5×10^{-25}	30.4 (342)	fungi alcohol dehydrogenase (P43067)
T02993 ^c	Ca ⁺ protein kinase	531	3	0	0	4.0×10^{-11}	37.9 (145)	juniper pollen Ca ⁺ binding protein (AF031471)
T03397 ^a	hypothetical protein	132	0	0	0	0.83	23.9 (71)	mouse urinary protein V (M16360)
Totals	50		41	15	7	$10 (<10^{-4})$	9 ($\geq 35\%$ and ≥ 80 aa overlap)	

Shaded entries correspond to those that had either $\geq 35\%$ amino acid identities or an overlap of ≥ 80 amino acid. The matches corresponds to the number of allergens or putative allergens that matched within the specified search window. The 'Best' *E* score, % ID and allergen indicate the allergen that produced the best scoring alignment identified using the FASTA algorithm.

^a Insignificant similarity, not a likely homologue based on the *E* score and little chance for cross-reactivity.

^b Possible homology with the indicated allergen, low degree of identity, many gaps in alignment, little chance of cross-reactivity.

^c Possible homology with the indicated allergen, low to moderate degree of identity, little chance of cross-reactivity.

^d Clear homology with the indicated allergen, high degree of identity, moderate chance of cross-reactivity.

^e Previously identified as an aeroallergen.

Discussion

Methods of biotechnology allow the insertion of genes from any source into crops that will be utilized for foods or feeds. Commercialization of these products is regulated to ensure that the foods produced from these products are as safe as foods derived from conventional crops. One important consideration is the possibility that the newly introduced gene may encode an allergen. The most significant potential risk of allergy associated with crops developed through biotechnology is the potential transfer of a known allergen from one source to another. Since proteins that are structurally very similar may be immunologically cross-reactive, it is also important to determine whether the newly introduced protein is significantly similar to any known allergen.

Comparison of the primary amino acid sequence and overall structure of the introduced protein with that of known allergens is an important step in the safety assessment process. The strategy for assessing the potential allergenicity of each protein should begin early in the developmental stages of the genetically modified plant. A decision tree describing this process has been previously reported [13]. Proteins obtained from sources with no known history of allergenicity should be evaluated using bioinformatic tools with a comprehensive allergen sequence database. Proteins that are found to be highly similar in sequence or predicted conformational structure to known allergens are then further tested with sera from individuals having allergies to the identified allergen to evaluate possible IgE binding as an indication of their allergenicity. Those proteins with specific IgE binding would not be considered further as potential products without clinical testing to prove that they are not cross-reactive.

Various computer algorithms could be used for comparing query sequences with those of known allergens. If these tools are to be predictive of allergenicity, the criteria used in the bioinformatic search must be carefully considered. The criteria for assessment of potential allergenicity of proteins produced in genetically modified crops using bioinformatic tools are currently being debated. For example, a review panel convened jointly by the FAO and WHO recommended using either a match of 6 contiguous identical amino acids, or a match of 35% identity over more than 80 amino acids to indicate that a protein is likely to be allergenic [27]. While the FASTA pairwise comparison algorithm provides a number of useful calculated values, these values are interpreted to infer homology and potential cross-reactivity. Homologous sequences

share a common evolutionary history [39, 44, 45]. The percent amino acid identity is a measure of similarity for a pair of proteins. Although sequences that share a high degree of similarity (e.g. measured by percent identity) may be related, they are not necessarily homologous. In addition, a query protein that is homologous to an allergen is not itself necessarily an allergen or cross-reactive [8]. For example, bovine myoglobin (table 1), while homologous to the midge hemoglobin allergen [41], is not known to cause allergic reactions. Also, vertebrate tropomyosins, often sharing ~50% amino acid identity to known allergenic invertebrate tropomyosins (e.g. shrimp, cockroach or mite) [46], are not known to induce clinical allergic reactions.

Concerns have been raised that proteins structurally unrelated to allergens may still share an immunologically relevant epitope (a short local overlap) and may cause cross-reactions. While the number of identified allergenic epitopes is increasing, relatively few are known. A database of authenticated epitopes could be used to screen any protein for allergenic activity. However, in the absence of an extensive, empirically determined epitope database for known allergens, comparisons between overlapping peptides using algorithms such as ALLERGENSEARCH are beneficial. The ALLERGENSEARCH may be more appropriate than FASTA for identification of short potentially allergenic epitopes because it only requires that the specified window size (e.g., 8-amino-acid residues) match. While recognizing that an IgE-mediated allergic response requires at least two IgE-binding epitopes, the use of the single match of 8 identical amino acids represents a precautionary approach to identifying proteins that might cross-react and should be tested further. This search technique may be used to complement the FASTA search. FASTA comparisons are initiated by aligning a matched word size (*k-tuple* parameter) and the alignment extended based on the chosen scoring matrix. The final computed alignment is by default the highest-scoring alignment. Thus, a short local overlap may be ignored for a higher-scoring alignment containing a longer overlap but relatively few linearly contiguous amino acid identities. The use of window sizes smaller than 8 contiguous amino acids appeared to only increase the frequency of questionable positives. Together these two approaches identify similarities that may be immunologically important and relevant to the safety assessment. The use of FASTA with an inclusive allergen database will most efficiently identify known allergens and also identify proteins that are likely to share sufficient structural similarity to suspect possible cross-reactivity.

The results of the ALLERGENSEARCH (table 3) indicated that any match of 8 contiguous identical amino acids is a conservative yet appropriate criterion for such searches. The criterion of a 6 amino acid match has previously not been rigorously assessed for its predictive value and does not appear to be scientifically justified due to the high rate of matches to allergens that appear to be random. Based on these results, the FAO/WHO recommendation [27] to use a window size of 6 contiguous amino acids did not discriminate between commonly consumed nonallergenic food proteins (myoglobin and rubisco, table 2) and allergens. The use of extremely conservative criteria for the bioinformatic searches could raise unnecessary public concerns and confusion relative to the safety of our current food supply. This would also lead to significant overestimation of potential allergenicity with rejection of beneficial nonallergenic proteins or possibly necessitate considerable clinical testing to demonstrate safety. The FASTA search and additional information regarding the extent of similarity (*E* score) or the >35% identity in any segment of 80 or more amino acids [27] may provide another method that might prove more useful than strict adherence to 8 amino acid matches in the search. However, this may be overly conservative, as evidence suggests that typically greater than 50% identity is required for clinical cross-reactions [8].

None of the *B. thuringiensis* protein sequences met criteria that would indicate a need for further testing. For the 50 randomly selected corn proteins, it is interesting that the 8-amino-acid match criterion missed two allergen homologues (embryo globulin-1, P15590 and Ca⁺ protein kinase, T02993) and one gliadin homologue (zein, AAA33523). While each of these displayed $\geq 35\%$ identity over an ≥ 80 -amino-acid stretch, they were well below the 50% identity level, suggesting that they are unlikely to cause cross-reactions [8]. Five proteins (pyruvate dehydrogenase, AAC72193; anionic peroxidase, AAC79953; cytochrome oxidase subunit III, CAA31221; pyruvate decarboxylase, CAA35589 and ribosomal protein S11)

matched 7 amino acids of at least one allergen that are clearly not homologues of allergens and are therefore false positives. The results of these analyses suggest that using a FASTA (or BLASTP) search, either with or without an additional 8-amino-acid match search will provide the most effective means to identify candidate product proteins that are sufficiently similar to an allergen to require further in vitro testing such as serum IgE-binding studies using sera from appropriately allergic subjects, in order to prevent the unintended introduction of allergens or cross-reactive proteins into the food supply.

As bioinformatic tools are invaluable for discovery of protein function, they are equally well suited for use in the assessment of potential allergenicity, especially in combination with other techniques [12, 13]. A comprehensive allergen database was assembled for this purpose. Overall structural similarity to known and suspected allergens should be evaluated using a method such as FASTA or BLASTP. Another sensitive screening criterion that retained selectivity for allergens or incomplete allergens versus nonallergens was a match of 8 or more contiguous identical amino acids to any known allergen. Finally, the results of these analyses demonstrate that additional work is necessary to evaluate specific threshold criteria (e.g. *E* score or between 35 and 50% identity over a significant length) for bioinformatic analysis for the prediction of allergenicity, specifically when addressing overall structural similarity and the potential for allergenic cross-reactivity.

Acknowledgments

The authors would like to thank Gary A. Bannon for critically reviewing the manuscript and Conrad H. Halling for providing the code for the algorithm used in this report and referred to as ALLERGENSEARCH. Published as Journal Series No. 13568, Agricultural Research Division, University of Nebraska-Lincoln, Lincoln, Nebr., USA.

References

- 1 Aronson AI, Shai Y: Why *Bacillus thuringiensis* insecticidal toxins are so effective: Unique features of their mode of action. *FEMS Microbiol Lett* 2001;195:1–8.
- 2 McClintock JT, Schaffer CR, Sjoblad RD: A comparative review of the mammalian toxicity of *Bacillus thuringiensis*-based pesticides. *Pestic Sci* 1995;45:95–105.
- 3 Siegel JP: The mammalian safety of *Bacillus thuringiensis*-based insecticides. *J Invertebr Pathol* 2001;77:13–21.
- 4 EPA: EPA registration eligibility decision (RED) *Bacillus thuringiensis*. EPA 738-R-98–004 1998; March 1998.
- 5 Betz FS, Hammond BG, Fuchs RL: Safety and advantages of *Bacillus thuringiensis*-protected plants to control insect pests. *Regulatory Toxicol Pharmacol* 2000;32:156–173.
- 6 de Maagd RA, Bravo A, Crickmore N: How *Bacillus thuringiensis* has evolved specific toxins to colonize the insect world. *Trends Genet* 2001;17:193–199.

- 7 Bravo A: Phylogenetic relationships of *Bacillus thuringiensis* delta-endotoxin family proteins and their functional domains. *J Bacteriol* 1997; 179:2793–2801.
- 8 Aalberse RC: Structural biology of allergens. *J Allergy Clin Immunol* 2000;106:228–238.
- 9 Taylor SL, Hefle SL: Will genetically modified foods be allergenic? *J Allergy Clin Immunol* 2001;107:765–771.
- 10 Gendel SM: The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods. *Adv Food Nutr Res* 1998;42:45–62.
- 11 Gendel SM: Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv Food Nutr Res* 1998;42: 63–92.
- 12 Astwood JD, Fuchs RL: Allergenicity of foods derived from transgenic plants. *Monogr Allergy* 1996;32:105–120.
- 13 Metcalfe DD, Astwood JD, Townsend R, Sampson HA, Taylor SL, Fuchs RL: Assessment of the allergenic potential of foods derived from genetically engineered crop plants. *Crit Rev Food Sci Nutr* 1996;36:S165–186.
- 14 Taylor SL, Hefle SL, Gauger BJ: Food allergies and sensitivities; in Helferich W, Winter CK (eds): *Food Toxicology*. New York, CRC Press, 2000, pp 1–36.
- 15 Plant M, Zimmerman EM: Allergy and mechanisms of hypersensitivity; in Paul WE (ed): *Fundamental Immunology*. New York, Raven Press, 1993, pp 1399–1425.
- 16 Berzofsky JA, Berkower IJ, Epstein SL: Antigen-antibody interactions and monoclonal antibodies; in Paul WE (ed): *Fundamental Immunology*. New York, Raven Press, 1993, pp 421–465.
- 17 Scheurer S, Wangorsch A, Nerkamp J, Skov PS, Ballmer-Weber B, Wuthrich B, Hausteiner D, Vieths S: Cross-reactivity within the profiling panallergen family investigated by comparison of recombinant profilins from pear (Pyr c 4), cherry (Pru av 4) and celery (Api g 4) with birch pollen profiling Bet v 2. *J Chromatogr B* 2001;756:315–325.
- 18 Aalberse RC, Akkerdaas J, van Ree R: Cross-reactivity of IgE antibodies to allergens. *Allergy* 2001;56:478–490.
- 19 Pearson WR: Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* 2000;132:185–219.
- 20 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J. Mol Biol* 1990;215:403–410.
- 21 Beezhold DH, Hickey VL, Slater JE, Sussman GL: Human IgE-binding epitopes of the latex allergen Hev b 5. *J Allergy Clin Immunol* 1999; 103:1166–1172.
- 22 Banerjee B, Greenberger PA, Fink JN, Kurup VP: Conformational and linear B-cell epitopes of Asp f 2, a major allergen of *Aspergillus fumigatus*, bind differently to immunoglobulin E antibody in the sera of allergic bronchopulmonary aspergillosis patients. *Infect Immun* 1999;67:2284–2291.
- 23 Chatchatee P, Jarvinen KM, Bardina L, Beyer K, Sampson HA: Identification of IgE- and IgG-binding epitopes on alpha(s1)-casein: Differences in patients with persistent and transient cow's milk allergy. *J. Allergy Clin Immunol* 2001;107:379–383.
- 24 Rabjohn P, Helm EM, Stanley JS, West CM, Sampson HA, Burks AW, Bannon GA: Molecular cloning and epitope analysis of the peanut allergen Ara h 3. *J Clin Invest* 1999;103:535–542.
- 25 Bredehorst R, David K: What establishes a protein as an allergen? *J Chromatogr B Biomed Sci Appl* 2001;756:33–40.
- 26 Ferguson A: Gluten-sensitive enteropathy (celiac disease); in Metcalfe DD, Sampson HA, Simon RA (eds): *Food Allergy: Adverse Reactions to Foods and Food Additives*. Cambridge, Blackwell Science, 1996, pp 287–301.
- 27 FAO/WHO expert consultation: Evaluation of allergenicity of genetically modified foods, Rome, Jan, 2001. <http://www.who.int/fsf/Gmfood/Consultation-Jan2001/report20.pdf>.
- 28 Pastorello EA, Farioli L, Pravettoni V, Ispano M, Scibola E, Trambaioli C, Giuffrida MG, Ansaloni R, Godovac-Zimmermann J, Conti A, Fortunato D, Ortolani C: The maize major allergen, which is responsible for food-induced allergic reactions, is a lipid transfer protein. *J Allergy Clin Immunol* 2000;106:744–751.
- 29 Benson DA, Boguski MS, Lipman DJ, Ostell J: GenBank. *Nucl Acids Res* 1997;25:1–6.
- 30 George DG, Dodson RJ, Garavelli JS, Haft DH, Hunt LT, Marzec CR, Orcutt BC, Sidman KE, Srinivasarao GY, Yeh LSL, Arminski LM, Ledley RS, Tsugita A, Barker WC: The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database. *Nucleic Acids Res* 1997;25:24–28.
- 31 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
- 32 Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* 1977; 80:319–324.
- 33 Bairoch A, Apweiler R: The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res* 1997;25:31–36.
- 34 Schuler GD, Epstein JA, Ohkawa H, Kans JA: Entrez: Molecular biology database and retrieval system. *Methods Enzymol* 1996;266:141–162.
- 35 Pearson WR, Lipman DJ: Improved tools for biological comparison. *Proc Natl Acad Sci USA* 1988;85:2440–2448.
- 36 Lipman DJ, Pearson WR: Rapid and sensitive protein similarity searches. *Science* 1985;227: 1435–1441.
- 37 Henikoff S, Henikoff JG: Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:10915–10919.
- 38 Henikoff JG, Henikoff S: Blocks database and its applications. *Methods Enzymol* 1996;266: 88–105.
- 39 Doolittle RF: Searching through sequence databases. *Methods Enzymol* 1990;183:99–110.
- 40 Pearson WR: Effective protein sequence comparison. *Methods Enzymol* 1996;266:227–258.
- 41 Buse G, Steffens GJ, Braunitzer G, Steer W: Hemoglobins. XXV. Hemoglobin (erythrocrutorin) CTT III from *Chironomus thummi thummi* (Diptera). Primary structure and relationship to other heme proteins (author's transl). *Hoppe Seylers Z Physiol Chem* 1979; 360:89–97.
- 42 Hemmann S, Menz G, Ismail C, Blaser K, Cramer R: Skin test reactivity to 2 recombinant *Aspergillus fumigatus* allergens in *A. fumigatus*-sensitized asthmatic subjects allows diagnostic separation of allergic bronchopulmonary aspergillosis from fungal sensitization. *J Allergy Clin Immunol* 1999;104:601–607.
- 43 Broadwater AH, Rubinstein AL, Chay CH, Klapper DG, Bedinger PA: Zea m1, the maize homolog of the allergen-encoding Lol p1 gene of rye grass. *Gene* 1993;131:227–230.
- 44 Reece GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, Zuckerkandl E: 'Homology' in proteins and nucleic acids: A terminology muddle and a way out of it. *Cell* 1987;50:667.
- 45 Baxevanis AD, Ouellette BFF (eds): *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Methods of Biochemical Analysis.. New York, Wiley, 1998, vol 39.
- 46 Reese G, Ayuso R, Lehrer SB: Tropomyosin: An invertebrate pan-allergen. *Int Arch Allergy Immunol* 1999;119:247–258.