THE USE OF AMINO ACID SEQUENCE ALIGNMENTS TO ASSESS POTENTIAL ALLERGENICITY OF PROTEINS USED IN GENETICALLY MODIFIED FOODS

STEVEN M. GENDEL

Biotechnology Studies Branch
Food and Drug Administration
National Center for Food Safety and Technology
Summit-Argo, Illinois 60501

I. INTRODUCTION

Food allergies occur in approximately 5% of children and 2% of adults (Sampson, 1992; Hefle 1996). Although allergic reactions to foods can range from mild to life threatening, sensitive individuals experience extreme reactions when exposed to small amounts of allergen (Sampson, 1992). The only reliable way to deal with food allergy is to avoid the offending food; therefore, it is important that allergic individuals be aware of the content of all foods consumed. An allergic individual must avoid both whole foods that cause reactions (for example milk) and mixtures that contain components of the allergenic food (such as casein).

The production of transgenic foods raises two major concerns regarding allergenicity: the transfer of allergenic proteins to new hosts, and the poten-

tial for proteins from organisms that have not previously been part of the food supply to become allergens (FDA, 1992; Fuchs and Astwood, 1996). If a transferred protein is from a "commonly allergenic" donor, it may be possible to obtain some measure of the allergenicity of the protein in a new host by testing with sera from allergic individuals. This has been done in one case in which a Brazil nut protein was transferred to soybean and retained allergenicity (Nordlee *et al.,* 1996). However, if a transferred protein is derived from a "less commonly allergenic" source, this approach to safety assessment is not practical because of the difficulty in obtaining a sufficient number of allergic sera. Similarly, there are no direct biochemical tests for potential allergenicity that can be used to assess new proteins in the food supply.

Food allergens, and allergens in general, are a diverse group of proteins. Food allergen proteins have been described as generally being between 10 and 70 kDa, highly expressed, possibly glycosolated, and resistant to degradation (Hefle, 1996). However, there are no data to show that any of these properties are necessary or sufficient to cause sensitization or an allergic reaction in a previously sensitized individual.

Several publications, including a recent report from the International Food Biotechnology Council, have suggested that the potential allergenicity of a transferred protein can be assessed by examining a set of physiochemical properties (including stability to digestion, prevalence, and stability to processing) and by comparing the sequence of the protein to those of known allergens. (Astwood and Fuchs, 1996; Fuchs and Astwood, 1996; Metcalfe *et al.,* 1996). The sequence-based component of an allergenicity assessment can be carried out by aligning a query sequence with each member of a database of allergen sequences. A negative result, the failure to find significant sequence similarity between the query sequence and any known allergen, can be considered an indication of low probability of potential allergenicity. Such comparisons have, in fact, been used in the safety assessment process for several transgenic foods, although little specific information has been published on how these comparisons were performed (Astwood and Fuchs, 1996; Fuchs and Astwood, 1996).

Sequence alignments of this type can be carried out using programs that implement a number of different algorithms (Gribskov and Devereux, 1991). Most of these algorithms were developed primarily to detect evolutionary or functional relationships. However, allergenicity assessment involves the detection of short regions of structural similarity that are not evolutionarily or functionally related. Therefore, some assumptions that are built into the alignment programs may not be relevant in this context. In addition, these programs often have multiple user-definable input parameters that affect their functioning. The data sets used for alignment also

affect the results obtained and the ease with which significant results can be recognized.

To determine the best method for utilizing sequence information in assessing the potential allergenicity of proteins used in new food varieties, I compared the results obtained by using different sequence alignment strategies with several test sequences and two allergen sequence databases. The test sequences included both synthetic control sequences and sequences for proteins that are currently being used in transgenic foods. The results of these tests showed that local alignment algorithms are more appropriate for use in this context than global allignment algorithms, use of the proper scoring matrix is necessary to reliably locate significant matches, and the lack of reliable criteria for defining an allergenic epitope makes it difficult to assess the biological significance of the matches that are identified.

II. METHODS

All of the sequence analysis programs used were part of Version 8 of the GCG package (Genetics Computer Group, Inc., Madison, WI) running on an AXP 2100 computer (Digital Equipment Corp., Maynard, MA) under a VMS 6.1 operating system. The individual programs and parameters used are described in detail under Results.

Construction of the two allergen sequence databases has been described (Gendel, chapter 3 of this volume). Briefly, accessions for food and nonfood allergen sequences were identified in three large reference databases. These sequences were compared both within and between the reference databases to identify a complete set of accessions that includes all available allergen sequence variants. Because it is not known whether common sequence properties are involved in the allergenicity of food and nonfood allergens, each group of sequences was treated as a separate database. The overall composition of the allergen databases is described in Table I. The food allergen sequence database does not include wheat gluten proteins because it is not clear whether food allergies and gluten-associated enteropathies share a common etiology (O'Mahony and Ferguson, 1991; Metcalfe, 1992).

All known food allergens are proteins (Taylor, 1992; Hefle, 1996). Therefore, amino acid sequence comparisons should be used for assessing potential allergenicity. Direct comparison of amino acid sequences avoids three problems with nucleic sequence comparison that could obscure significant matches. First, because the genetic code is degenerate, proteins with identical amino acid sequences can have significantly different coding sequences. Second, because all known food allergen sequences originate from eukaryotes, the genomic sequences may contain introns. Although it may be

TABLE I

SUMMARY OF THE CONTENTS OF THE ALLERGEN DATABASES (GENDEL, CHAPTER 3 OF THIS VOLUME)

Food allergen database	
Unique sequences	138
GenPept accessions	89
SwissProt accessions	53
PIR accessions	90
Species	15
Proteins	44
Nonfood allergen database	
Unique sequences	218
GenPept accessions	118
SwissProt accessions	105
PIR accessions	162
Species	65
Proteins	142

possible in many cases to identify and use only the coding regions of the nucleic acid sequences, this can be much more complex than simply using the translated amino acid sequence. Third, some allergen sequences have been obtained from cDNA while others represent genomic clones. Again, this means that the possible presence of introns needs to be considered when making comparisons at the nucleic acid level.

Construction of three positive control sequences was carried out as follows. The sequence of a known food allergen, the 113-amino-acid cod parvalbumin protein known as allergen M or Gad cl (SwissProt accession A94236), was randomized by using the program SHUFFLE. This produces a random sequence with the same amino acid composition as the original sequence. The 10 amino acids numbered 51–60 in the original sequence were used to replace amino acids 11–20, 51–60, or 101–110 in the shuffled sequence. This produced sequences with regions located near the N-terminus (control sequence C1), the middle (control sequence C2), or the C-terminus (control sequence C3) that are identical to part of a known food allergen.

A set of transgene test sequences, proteins currently being used in transgenic plants, were identified from a variety of sources, including direct searching of database annotation, regulatory documents from both the Food and Drug Administration and the U.S. Department of Agriculture, and literature sources (Table II). Sequence testing was carried out using the accessions listed in Table II; the actual transgenic plants may express

TABLE II

ACCESSIONS OF GENES USED TO CONSTRUCT TRANSGENIC FOOD PLANTS

Gene	Original source organism	Target organism(s)	GenPept accession	SwissProt accession	PIR accession	References[a]
ACC deaminase	*Pseudomonas* 6G5	Tomato	M80882	P30297	JQ1330	1, 3
Bacillus toxin CrylA(b)	*Bacillus thuringiensis kurstaki/ berliner*	Tomato	A09398[b]	P06578	JD0002	1, 4
Bacillus toxin CrylA(c)	*Bacillus thuringiensis kurstaki*	Cotton	X54159		S11445	1
Bacillus toxin Cry3A	*Bacillus thuringiensis kurstaki*	Potato	X70979			8
Bacillus toxin	*Bacillus thuringiensis tenebrionis*	Patato	M30503	P07130	A29987	1, 5
Neomycin phosphotransferase II (NPT)	*Escherichia coli*	Tomato, cotton, potato, etc.	V00618	P00552	A00663	1, 2
Nitrilase (BXN)	*Klebsiella pneumoniae*	Cotton	J03196	P10045	A28658	6
Phosphinothricin acetyltransferase 1	*Streptomyces viridochromogenes*	Corn	M22827		JT0409	1, 7
Phosphinothricin acetyltransferase 2	*Streptomyces hygroscopicus*	Corn	X17220	P16426	S08615	1, 10
Thioesterase	*Umbellularia californica* (California bay)	*Brassica*	M94159		A40229	6, 9
ZYMV coat protein	Zucchini yellow mosaic virus	Squash	M35095[c]	P18479	JH0103	11, 12
MWV coat protein	Watermelon mosaic virus	Squash	D00535[d]	P20235	PS0084	11, 12

[a] 1. Fuchs and Astwood (1996); 2. FDA (1994); 3. Klee *et al.* (1991); 4. Noteborn and Kuiper (1995); 5. Fuchs *et al.* (1995); 6. Redenbaugh *et al.* (1995); 7. USDA Petition 94-357-01; 8. USDA Petition 94-257-01; 9. Voelker *et al.* (1992); 10. USDA Petition 94-319-01; 11. Quemada (1995); 12. USDA Petition 92-204-01.

[b] A09398, from a from Ciba-Ceigy patent, is 100% identical at the nucleotide level to M15271 and (the slightly shorter) X54939, which were the accessions translated to create P06527 in SwissProt and JD0002 in PIR. The amino acid sequences in the GP, SWP, and PIR accessions listed are 100% identical.

[c] The coat protein is made from a polyprotein that is processed proteolytically. The actual coat protein starts at amino acid 109.

[d] The coat protein is made from a polyprotein that is processed proteolytically. The actual coat protein starts at amino acid 22.

proteins with slightly different amino acid sequences if the gene used originated from a different strain or if it was modified during construction of the transgenic plant.

Note. To avoid any implication of evolutionary or functional relationship between the sequences involved in this work, all sequence comparisons are referred to in terms of sequence identity or similarity, rather than homology. Identical sequences have the same amino acid sequence over the region involved, while similar sequences have some amino acid mismatch. Sequence homology is a consequence of evolutionary divergence from a common ancestor.

III. RESULTS

A. GLOBAL ALIGNMENT—GAP WITH CONTROL SEQUENCES

In general, sequence alignment algorithms can be divided into global algorithms that optimize alignments across the entire length of the sequences involved and local algorithms that attempt to optimize alignments only with regions of high similarity (Gribskov and Devereux, 1991). Global alignment algorithms are of greatest utility when the sequences involved are related. Allergenicity assessment involves sequence alignments between proteins that are not evolutionarily related. Therefore, it is likely that local alignment will be more useful. To confirm this, the three control sequences were each aligned with the unmodified Gad c1 sequence by using the GCG implementation of the Needleman and Wunsch algorithm in the program GAP (Needleman and Wunsch, 1970). The program was able to find the correct 10-amino-acid match only for the control sequence that contained the sequence identity in the middle, that is, in the original position (Fig. 1). Because there is no reason to expect regions of potential allergenic cross-reactivity to be located in the same region of different allergens, global alignment algorithms appear unsuitable for assessing potential allergenicity.

B. LOCAL ALIGNMENT—FASTA WITH CONTROL SEQUENCES

The most widely used local alignment program is FASTA, developed by Pearson and Lipman (Pearson and Lipman, 1988; Pearson, 1990). FASTA has several user-definable parameters that can be altered depending on the nature of the search being conducted and that significantly affect the matches obtained. A series of FASTA searches was carried out using the

A. Global Alignment of Control Sequence C1 with Gad c1.

```
  1 ...AFLDIKERKADED...KEGFIEEKAKEGGEKWSFKGFGADFDAGGAE 44
        |     || |    |||   |           || |
  1 AFKGILSNADIKAAEAACFKEGSFDE..........DGFYAKVGLDAFS 39
            .         .         .         .         .
 45 ADSEDDDLFKDGADKDKLCAAEALEEFL.ALDIGLFTFSVYKEDTKDGFI 93
    ||           ||        |  || |    |         ||
 40 ADELKKLFKIADEDKEGFIEEDELKLFLIAFAADLRALT..DAETKAFLK 87
            .         .
 94 DDLAKALGKVFASIAELEAI........ 113
          ||      |  |
 88 AGDSDGDGKI..GVDEFGALVDKWGAKG 113
```

B. Global Alignment of Control Sequence C2 with Gad c1.

```
  1 AFLDIKERKAVIFLAANGGKKAKEGGEKWSFKGFGADFDAGGAEADSEDD 50
    ||  |          |   |           |     | |
  1 AFKGILSNADIKAAEAACFKEGSFDEDGFYAKVGLDAFSADELKKLFKIA 50
            .         .         .         .         .
 51 DEDKEGFIEEKLCAAEALEEFL.ALDIGLFTFSVYKEDTKDGFIDDLAKA 99
    |||||||||         |  || |    |         ||
 51 DEDKEGFIEE.....DELKLFLIAFAADLRALT..DAETKAFLKAGDSDG 93
            .         .
100 LGKVFASIAELEAI........ 113
     ||      |  |
 94 DGKI..GVDEFGALVDKWGAKG 113
```

C. Global Alignment of Control Sequence C3 with Gad c1.

```
  1 AFLDIKERKAVIFLAANGGKKAKEGGEKWSFKGFGADFDAGGAEADSEDD 50
            |      |   || |      ||     ||  |  |    |
  1 .........AFKGILSNADIKAAEAA...CFK..EGSFDEDGFYAKVGLD 36
            .         .         .         .         .
 51 DLFKDGADKDKLCAAEALEEFL...ALDIGLFTFSVYKEDTKDGFIDDLA 97
       |   |    | |  | |     |   |  |        |
 37 AFSADELKKLFKIADEDKEGFIEEDELKLFLIAFAADLRALTDAETKAFL 86
            .         .
 98 KALDEDKEGFIEEEAI........... 113
    || |  |  | |
 87 KAGDSDGDGKIGVDEFGALVDKWGAKG 113
```

FIG. 1. Alignment of the three control sequences with the original Gad c1 sequence using the Needleman and Wunsch (1970) global alignment algorithm. The 10-amino-acid match present in each control sequence is underlined. The control sequence is the top sequence in each case; the Gad c1 sequence is on the bottom.

control sequences and allergen databases to determine the optimal parameters for allergenicity assessment.

To confirm that FASTA could locate short regions of sequence identity regardless of location, each of the control sequences described above was aligned with the unmodified Gad c1 sequence by using FASTA (Fig. 2). FASTA located the correct alignment in all three cases.

FASTA can compare a query sequence to all members of a sequence database (Pearson, 1990). Therefore, there are two possible approaches to using FASTA to assess potential allergenicity. First, the query sequence could be aligned with all sequences in a large reference database, such as GenPept, and the resulting "hits" examined to determine if any allergen sequences are present. Alternatively, the query sequence could be aligned with the allergen sequence database, and those sequences with the most extensive similarities could be examined in detail. The first approach seems inherently undesirable because no single large database contains all the relevant allergen sequences. Therefore, it would be necessary to carry out multiple searches and analyses to ensure complete testing. Also, in most cases, searches of large reference databases produce a number of high scoring alignments with sequences that are evolutionarily related to the

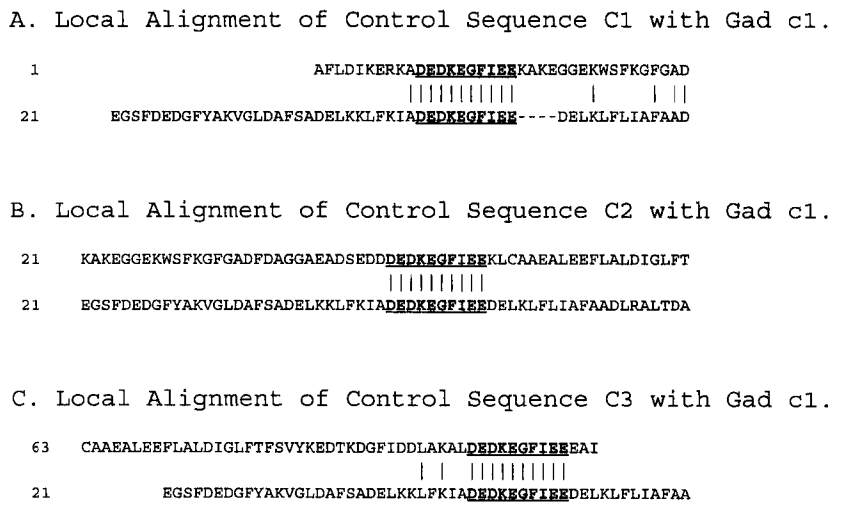

FIG. 2. Alignment of the three control sequences with the original Gad c1 sequence using the FASTA local alignment algorithm (Pearson, 1990). The 10-amino-acid match present in each control sequence is underlined. The control sequence is the top sequence in each case; the Gad c1 sequence is on the bottom. Compare these alignments to those shown in Fig. 1.

query sequence. Sequences containing regions of short, but immunologically significant, sequence similarity produce alignments with low scores. Depending on the gene involved and the parameters used to determine how many matches are displayed, immunologically significant alignments may be lost. Direct searching of allergen databases ensures complete testing and, because the databases involved are relatively small, all alignments can be examined.

Each of the three control sequences was aligned with the complete PIR database, using FASTA with the default parameters to demonstrate the difficulty in using large reference databases for allergenicity assessment. Because the control sequences were derived from a randomized sequence, matches to evolutionarily related sequences were eliminated. Despite this simplification, the correct match was not found in the top 50 scores for two of the control sequences and was the 30th highest scoring sequence for the third. In contrast, when the control sequences were aligned with the food allergen database using FASTA and the default parameters, the correct sequence produced the highest score in one case and was among the top 5 scores for the other two control sequences.

FASTA, as well as all other commonly used sequence alignment programs, uses a scoring system that allows for evolutionarily conservative substitutions. The score for each alignment is determined by using a scoring matrix that assigns a numerical value to each possible pair of amino acid matches. Although several different scoring matrices exist, they all assign positive scores to some amino acid substitutions on the basis of observed evolutionary patterns (Gribskov and Devereux, 1991). As a result, an alignment of two proteins that contains a number of "evolutionarily conservative" substitutions may produce a higher score than an alignment of the same proteins that contains a shorter region of sequence identity. The alignment programs only reports the highest scoring alignment between the two sequences. Because allergenic cross-reactivity likely involves short regions of high sequence similarity, the use of an evolutionary scoring matrix may produce alignments that miss significant matches. Therefore, an alternative scoring matrix (an identity matrix) that assigned a positive score only to exact amino acid matches was constructed. When FASTA was run using this identity matrix instead of the default evolutionary matrix with the three control sequences and the food allergen database, the correct match produced the highest score in all three cases.

C. LOCAL ALIGNMENT—FASTA WITH TRANSGENE PROTEINS

The efficacy of a FASTA database alignment for assessing potential allergenicity was further tested with the 12 transgene sequences listed in

Table II. These proteins were chosen because they (or closely related proteins) have been used to produce transgenic crop plants that are being used in commerce (Astwood and Fuchs, 1996; Fuchs and Astwood, 1996). Each test sequence was aligned with each entry in both the food and nonfood allergen databases by using both the evolutionary and identity scoring matrices. In each case, the entire set of matches was scanned to locate the longest contiguous region of sequence identity (Table III). Longer stretches of sequence identity were found by using the identity matrix than by using the evolutionary matrix in more than half of the database alignments. Figure 3 is an example of the different alignments produced between the same two proteins by using the different matrices. In this case, the alignment produced using the evolutionary matrix missed a possibly

TABLE III

THE LONGEST CONTIGUOUS REGION OF SEQUENCE IDENTITY FOUND BY FASTA ALIGNMENT OF EACH TRANSGENE SEQUENCE; WORDSIZE = 2

Gene	Database	Scoring matrix	
		Ident.	Evol.
ACC deaminase	Food	5	5
	Nonfood	5	5
Bacillus toxin CrylA(b)	Food	6	6
	Nonfood	7	7
Bacillus toxin CrylA(c)	Food	6	6
	Nonfood	5	4
Bacillus toxin Cry3A	Food	5	3
	Nonfood	5	4
Bacillus toxin	Food	5	5
	Nonfood	5	4
NPT	Food	7	4
	Nonfood	5	5
Nitralase	Food	5	4
	Nonfood	7	5
Phosphinothricin acetyltransferase 1	Food	5	4
	Nonfood	5	4
Phosphinothricin acetyltransferase 2	Food	5	4
	Nonfood	5	5
Thioesterase	Food	5	4
	Nonfood	5	5
WMV coat protein	Food	4	4
	Nonfood	5	4
ZYMV coat protein	Food	6	5
	Nonfood	5	5

A. FASTA Alignment using the Evolutionary Matrix

```
      210        220        230        240        250        260
NPT    CGRLGVADRYQDIALATRDIAEELGGEWADRF-LVLYGIAAPDSQRIAFYRLLDEFF
        :: :|:   :::::  :: :::  ::::::| :|: : :: ::: |||:
Lectin PAKSTVGRVLHSTQVRLWEKSTNRLTNFQAQFSFVIKSPNDIGADGIAFFIAAPDSQIPK
            50         60         70         80         90        100
```

B. FASTA Alignment using the Identity Matrix

```
         220        230        240        250        260
NPT    DRYQDIALATRDIAEELGGEWADRFLVLYGIAAPDSQRIAFYRLLDEFF
                                        |||||||
Lectin STNRLTNFQAQFSFVIKSPNDIGADGIAFFIAAPDSQIPKNSAGGTLGLFDPQTAQNPSA
            70         80         90        100        110        120
```

FIG. 3. An example of different FASTA alignments produced using the evolutionary and identity scoring matrices. In both cases, the top sequence is part of NPT, the bottom sequence is part of a peanut lectin protein. Lines indicate exact amino acid matches; dots indicate conservative substitutions, as defined by the evolutionary scoring matrix. The seven amino acids that are aligned using the identity matrix are underlined.

significant exact match. This confirms the previous conclusion that use of the identity matrix is more appropriate for allergenicity assessment.

The distribution of lengths of sequence identity shown in Table III suggests that exact matches of five or fewer amino acids are likely to occur frequently by chance. This was confirmed independently by using the GCG WORDSEARCH implementation of the Wilbur and Lipman algorithm (data not shown) (Wilbur and Lipman, 1983). Matches of seven amino acids or greater were sufficiently rare (approximately the top 10% of scores) to suggest that they should be evaluated for biological relevance (see below).

One critical user-definable parameter in FASTA is wordsize. The wordsize defines the window size used in the initial steps of the sequence alignment. The default value for proteins, which was used above, is 2, meaning that only those sites containing at least two adjacent matching amino acids are used for initiating possible alignments. The complete set of database alignments was also run using a wordsize of 1 with the identity scoring matrix (data not shown). Use of the smaller wordsize did not locate any longer regions of sequence identity, and in some cases decreased the size of the longest region found.

Although it is reasonable to assume that immunologically cross-reactive sequences will have a high degree of sequence identity, it is possible that

some amino acid mismatches might be tolerated. Therefore, the FASTA results used to construct Table III were also examined to identify regions containing ≥70% or ≥80% sequence identity over ≥10 amino acids (Table IV). Eight of the test sequences matched at least one allergen at the 70% identity level, while two sequences produced matches at the 80% level. In most cases, as expected, the regions of sequence similarity identified by these criteria were the same as, or only slightly larger than, the longest regions of contiguous identity located previously. It should be noted that the two proteins that had matches at the 80% level are related, and the high similarity occurs in a region that has the same sequence in both.

D. LOCAL ALIGNMENT—BLAST WITH TRANSGENE PROTEINS

The other major local alignment tool that is used for database searching is the BLAST implementation of the Altschul *et al.* (1990) algorithm. Although BLAST is usually used with large databases, it is possible to construct local BLAST databases. Both the food allergen and nonfood allergen databases were converted to the BLAST format, and the test sequences in Table II were aligned with each using two scoring matrices, as was done with FASTA. The results of these alignments (Table V) were similar to

TABLE IV

PRESENCE OF AT LEAST ONE HIGH HOMOLOGY MATCH AT THE INDICATED LEVEL OF IDENTITY OVER ≥10 AMINO ACIDS AFTER FASTA ALIGNMENT OF THE TRANSGENE SEQUENCES WITH EACH ALLERGEN DATABASE; WORDSIZE = 2

	≥70% identity		≥80% identity	
Gene	Food	Nonfood	Food	Nonfood
ACC deaminase	+	−	−	−
Bacillus toxin Cry1A(b)	+	+	+	−
Bacillus toxin Cry1A(c)	+	+	+	−
Bacillus toxin Cry3A	+	+	−	−
Bacillus toxin	+	+	−	−
NPT	−	−	−	−
Nitralase	+	+	−	−
Phosphinothricin acetyltransferase 1	−	+	−	−
Phosphinothricin acetyltransferase 2	+	−	−	−
Thioesterase	−	−	−	−
WMV coat protein	−	−	−	−
ZYMV coat protein	−	−	−	−

TABLE V

THE LONGEST CONTIGUOUS REGION OF SEQUENCE IDENTITY FOUND BY BLAST ALIGNMENT OF EACH TRANSGENE, SEQUENCE[a]

Gene	Database	Scoring matrix	
		Ident.	Evol.
ACC deaminase	Food	5	5
	Nonfood	5	5
Bacillus toxin CrylA(b)	Food	6	6
	Nonfood	7	7
Bacillus toxin CrylA(c)	Food	6	6
	Nonfood	5	4
Bacillus toxin Cry3A	Food	5	4
	Nonfood	5	5
Bacillus toxin	Food	5	5
	Nonfood	5	5
NPT	Food	7	7
	Nonfood	5	5
Nitralase	Food	5	5
	Nonfood	7	7
Phosphinothricin acetyltransferase 1	Food	5	4
	Nonfood	5	5
Phosphinothricin acetyltransferase 2	Food	5	4
	Nonfood	5	4
Thioesterase	Food	5	5
	Nonfood	5	5
WMV coat protein	Food	5	5
	Nonfood	5	4
ZYMV coat protein	Food	6	5
	Nonfood	5	5

[a] Compare with Table III.

those obtained using FASTA. BLAST did not locate any regions of sequence identity ≥6 amino acids in length that were not found by FASTA.

E. EXTENDING LOCAL ALIGNMENTS

Although it is likely that immunological cross-reactivity requires extensive sequence similarity, absolute identity may not be necessary (for example, see Elsayed *et al.,* 1982). Therefore, the regions of the alignments described above with highest degree of sequence identity were examined to determine if additional sequence similarity was present. This was done using FASTA to realign these regions containing the sequence identities employing either the evolutionary scoring matrix or a biochemical scoring

matrix. The biochemical scoring matrix divides the amino acids into six classes based on biochemical characteristics (i.e., hydrophilic acid amino acids, hydrophilic basic amino acids, etc). (GCG Program Manual, Version 8). Alignment of members of the same class is scored as self-match; alignment of members of different classes is scored as a mismatch. The realignment was confined to a region of 15 to 20 amino acids in each case to preserve the previously located identities.

The relevant regions of each of the alignments containing 7 amino acid exact matches (Table III) and all of the alignments containing ≥70% identity over ≥10 amino acids (Table IV) were realigned and examined for similarity. For two of the three sequence pairs that included 7 amino acid exact matches, the realignment did not indicate that significant additional similarity was present in these regions. In the third case, use of the evolutionary matrix indicated that the region contained a stretch of 17 amino acids with 82% similarity, but the matches involved were not similar biochemically.

Realignment did not indicate that additional similarity was present in 9 of the 11 regions with ≥70% identity over ≥10 amino acids. However, in two cases, significant additional similarity was present (Fig. 4). First, the

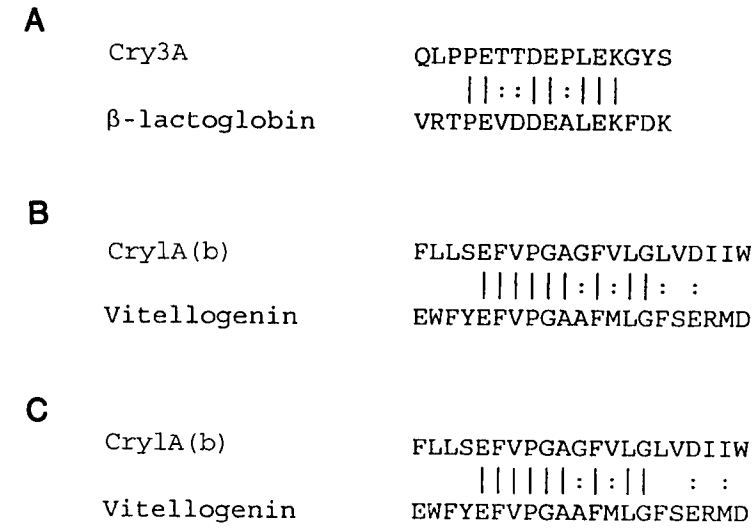

FIG. 4. Realignment of sequences with regions of high similarity using the evolutionary and biochemical similarity scoring matrices. (A) *Bacillus* toxin gene Cry3A aligned with β-lactoglobin using the similarity matrix. (B) *Bacillus* toxin gene CrylA(b) aligned with vitellogenin using the evolutionary matrix. (C) *Bacillus* toxin gene CrylA(b) aligned with vitellogenin using the similarity matrix. The lines indicate exact matches; the dots indicate similar amino acids.

initial alignment between Cry 3A and β-lactoglobin located subsequences in which 7 of 10 amino acids matched exactly. Realignment with both the evolutionary and biochemical matrices indicated that the intercalary amino acids were similar, meaning that the alignment was 100% similar over 10 amino acids (Fig. 4). Second, the initial alignment between CrylA(b) and vitellogenin located subsequences in which 9 to 11 amino acids were identical (82% similarity). Realignment indicated that these regions contained stretches of 11 biochemically similar and 12 evolutionarily similar amino acids (100% similarity over 11 or 12 amino acids) (Fig. 4). (CrylA(c) has the same sequence as CrylA(b) in the region involved, and therefore produced the same alignments, but this was not considered an independent alignment because the proteins are closely related).

IV. DISCUSSION

In the absence of reliable biochemical indicators, assessing the potential allergenicity of new proteins in the food supply is difficult. Sequence analysis may contribute to this analysis by determining whether the test protein shares any significant sequence features with known allergens. Three conditions must be met for this analysis to be meaningful. First, the database of allergen sequences used must be extensive enough to encompass all relevant sequence features. Second, the sequence analysis must be carried out with the proper algorithms and scoring parameters. Third, the criteria that define significant matches must be clear and technically valid.

The allergen sequence databases used here have beed described (Gendel, chapter 3 of this volume). Although they are the most extensive allergen databases available, it is clear that a large number of allergens, particularly food allergens, have not been identified or sequenced. Interestingly, a number of important sequence motifs (such as nucleic acid or cofactor binding regions) have been identified from much smaller data sets (Bairoch *et al.*, 1996). The absence of a common motif or sequence pattern for food allergens suggests that there is no single common pattern or that secondary structural patterns that are not easily recognized from the primary structure are critical. In the absence of an appropriate motif or pattern, allergenicity assessment should be carried out by aligning each test protein to all members of the allergen databases.

These results show that allergenicity assessment should be carried out using local alignment algorithms such as those implemented in FASTA or BLAST. Because BLAST requires compilation of special format databases and index files, FASTA is easier to use with small data sets, particularly if those data sets are subject to change. The optimal strategy for locating

sequence matches with a high degree of sequence identity is to carry out an initial alignment between a test sequence and each database with FASTA using a wordsize of 2 and an identity scoring matrix. All regions showing a relatively high degree of sequence identity should then be realigned using either an evolutionary or biochemical similarity scoring matrix. The results of the second set of alignments can then be evaluated for biological relevance.

Because little is known about the etiology of food allergy, the criteria that define significant matches for potential allergens are not clear. It is difficult to determine how much sequence similarity between distantly related proteins is significant in this context when some closely related proteins that share extensive homology are not allergenicity cross-reactive. The only published recommendations assert that an exact match of at least eight amino acids is necessary to raise concerns about potential allergenicity (Astwood and Fuchs, 1996; Metcalfe *et al.*, 1996). These results show that exact matches of eight or more amino acids occur infrequently by chance, and therefore any such occurrence may be of biological significance. However, it is not clear whether this is the minimum level of significant sequence similarity for potential allergenicity.

The work of Elsayed *et al.* (1982, 1991), Miller *et al.* (1996), Walsh and Howden (1991) suggests that shorter sequences may determine recognition specificity as long as they occur within peptides that provide the proper (nonsequence specific) structural context. This is consistent with studies showing that the antigen binding site may be able to accommodate only six to eight amino acids (Berzofsky and Berkower, 1993). Further, although it is clear that some amino acid residues are critical for specific binding, some conservative substitutions may not affect allergenicity. Therefore, it may be prudent to treat sequence matches with a high degree of identity that occur within regions of similarity as significant even if the identity does not extend for eight or more amino acids. For example, the similarity between CrylA(b) and vitellogenin (Fig. 4) might be sufficient to warrant additional evaluation.

These results emphasize the need for further research to define minimal food allergen epitopes, determine the effects of amino acid substitutions on allergenicity, and clarify the possible role of so-called discontinuous epitopes in food allergy. The development of techniques for the rapid production of large numbers of synthetic peptides should greatly facilitate this research, especially in systems for which well-characterized human sera can be readily obtained. These allergen databases are currently being used to determine if potential sequence motifs can be identified for food allergens.

ACKNOWLEDGMENT

This work was partially supported by Cooperative Agreement No. FD000431 between the FDA and the National Center for Food Safety and Technology.

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Astwood, J., and Fuchs, R. 1996. Allergenicity of foods derived from transgenic plants. *Monogr. Allergy* **32,** 105–120.

Bairoch, A., Bucher, P., and Hofman, K. 1996. The PROSITE database: Its status in 1995. *Nucl. Acids Res.* **24,** 189–196.

Berzofsky, J., and Berkower, I. 1993. Immunogenicity and antigen structure. *In* "Fundamental Immunology" (E. Paul, ed.) 3rd ed. Raven Press, New York.

Elsayed, S., Apold, J., Holen, E., Vik, H., Florvagg, E., and Dybendal, T. 1991. The structural requirements of epitopes with IgE binding capacity demonstrated by three major allergens from fish, egg, and tree pollen. *Scand. J. Clin. Lab. Invest.* **51** (Suppl. 204), 17–31.

Elsayed, S., Sornes, S., Aplod, J., Vik, H., and Florvagg, E. 1982. The immunological reactivity of the three homologous repetitive tetrapeptides in the region 41–64 of allergen M from cod. *Scand. J. Immunol.* **16,** 77–82.

FDA. 1992. Statement of policy: Food derived from new plant varities. *Fed. Reg.* **57,** 22,984–23,005.

FDA. 1994. Secondary direct food additives permitted in food for human consumption: Food additives permitted in feed and drinking water of animals; aminoglycoside 3′-phosphotransferase II. *Fed. Reg.* **59,** 26,700–26,711.

Fuchs, R., and Astwood, J. 1996. Allergenicity assessment of food derived from genetically modified plants. *Food Technol.* **50**(2), 83–88.

Fuchs, R., Rogan, G., Kech, P., Love, S., and Lavrik, P. 1995. Safety evaluation of Colorado potato beetle-protected potatoes. *In* "Application of the Principles of Substantial Equivalence to the Safety Evaluation of Foods or Food Components Derived by Modern Biotechnology." World Health Organization, Geneva.

Gendel, S. 1998. Sequence Databases for Assessing the Potential Allergenicity of Proteins Used in Transgenic Foods. *In* "Advances in Food and Nutrition Research" (Steve L. Taylor, ed.), vol. 42, pp. 63–92. Academic Press, San Diego.

Gribskov, M., and Devereux, J. 1991. "Sequence Analysis Primer." Stockton Press, New York.

Hefle, S. 1996. The chemistry and biology of food allergens. *Food Technol.* **50**(3), 86–92.

Klee, H., Hayford, M., Kretamek, K., Barry, G., and Kishore, G. 1991. Control of ethylene synthesis by expression of a bacterial enzyme in transgenic tomato plants. *Plant Cell* **3,** 1187–1193.

Metcalfe, D. 1992. The nature and mechanisms of food allergies and related diseases. *Food Technol.* **46**(5), 136–139.

Metcalfe, D., Astwood, J., Townsend, R., Sampson, H., Taylor, S., and Fuch, R. 1996. Assessment of the allergenic potential of foods derived from genetically engineered corp plants. *Crit. Rev. Food Sci. Nutr.* **36**(S), S165–S186.

Miller, K., Bradley, A., Fitzgerald, R., Maggi E., Morgan, M., and Wal, J. 1996. Chemical composition and structure of food constituents: Defining allergenic potential. *Monogr. Allergy* **32,** 100–104.

Needleman, S., and Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453.

Nordlee, J., Taylor, S., Townsend, J., Thomas, L., and Bush, R. 1996. Identification of a Brazil-nut allergen in transgenic soybeans. *N. Engl. J. Med.* **334,** 688–692.

Noteborn, H., and Kuiper, H. 1995. Safety evaluation of transgenic tomatoes expressing *Bt* Endotoxin. *In* "Application of the Principles of Substantial Equivalence to the Safety Evaluation of Foods or Food Components Derived by Modern Biotechnology." World Health Organization, Geneva.

O'Mahony, S., and Ferguson, A. 1991. Gluten-sensitive enteropathy (celiac disease). *In* "Food Allergy: Adverse Reactions to Food and Food Additives" (D. Metcalfe, H. Sampson, and R. Simon, eds.), pp. 186–198. Blackwell Publisher, Boston.

Pearson, W. 1990. Rapid and sensitive sequence comparision with FASTF and FASTA. *Meth. Enzymol.* **83,** 63–98.

Pearson, W., and Lipman, D. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.

Quemada, H. 1995. Food safety evaluation of a transgenic squash. *In* "Proceeding of the OECD Workshop on Food Safety Evaluation." Organisation for Economic Cooperation and Development, Paris.

Redenbaugh, K., Lindemann, J., and Malyj, L. 1995. Application of the principles of substantial equivalence in the safety evaluation of FLAVR SAVR tomato, BXN cotton and oil-modified rapeseed. *In* "Application of the Principles of Substantial Equivalence to the Safety Evaluation of Foods or Food Components Derived by Modern Biotechnology." World Health Organization, Geneva.

Sampson, H. 1992. Food hypersensitivity: Manifestations, diagnosis, and natural history. *Food Technol.* **46**(5), 141–144.

Taylor, S. 1992. Chemistry and detection of food allergens. *Food Tecnol.* **46**(5), 146–152.

Voelker, T., Worrell, A., Anderson, L., Bleibaum, J., Fan, C., Hawkins, D., Radke, S., and Davies, H. 1992. Fatty acid biosynthesis redirected to medium chains in transgenic oilseed Plants. *Science* **257,** 72–74.

Walsh, B., and Howden, M. 1991. Epitope mapping of allergens for rapid localization of continuous allergenic determinants. *Meth. Enzymol* **203,** 301–311.

Wilbur, W., and Lipman, D. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* **80,** 726–730.