Research Article

Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens

Gregory S. Ladics[1], Gary A. Bannon[2], Andre Silvanovich[2] and Robert F. Cressman[1]

[1] Dupont/Pioneer Crop Genetics Regulatory Science and Registration, Wilmington, DE, USA
[2] Monsanto Co., St. Louis, MO, USA

Food and Agriculture Organization/World Health Organization (FAO/WHO) recommended that IgE cross-reactivity between a transgenic protein and allergen be considered when there is ≥35% identity over a sliding "window" of 80 amino acids. Our objective was to evaluate the false positive and negative rates observed using the FAO/WHO *versus* conventional FASTA analyses. Data used as queries against allergen databases and analyzed to assess false positive rates included: 1102 hypothetical corn ORFs; 907 randomly selected proteins; 89 randomly selected corn proteins; and 97 corn seed proteins. To evaluate false negative rates of both methods: Bet v 1a along with several crossreacting fruit/vegetable allergens and a bean α-amylase inhibitor were used as queries. Both methods were also evaluated for their ability to detect a putative nonallergenic test protein containing a sequence derived from Ara h 1. FASTA versions 3.3t0 and 3.4t25 were utilized. Data indicate a conventional FASTA analysis produced fewer false positives and equivalent false negative rates. Conventional FASTA *versus* sliding window derived *E* scores were generally more significant. Results suggest a conventional FASTA search provides more relevant identity to the query protein and better reflects the functional similarities between proteins. It is recommended that the conventional FASTA analysis be conducted to compare identities of proteins to allergens.

Keywords: Bioinformatics / FASTA analysis / Novel proteins / Protein allergen

Received: November 10, 2006; revised: February 1, 2007; accepted: March 5 2007

1 Introduction

Comparison of novel proteins for similarity to known allergens is a critical part of the weight of evidence approach used to ascertain the safety of expressed proteins in transgenic plant products. Food and Agriculture Organization/World Health Organization (FAO/WHO) recommended in January 2001 [1] that a similarity search be performed using the FASTA algorithm [2] to search for identities in amino acid sequence that may correspond to potential IgE crossreactivity to known allergens.

When the recommendations were published, they contained a suggested procedure of how this search should be performed. One of the steps in the procedure was to "prepare a complete set of 80 amino acid length sequences derived from the expressed protein" and "compare each of the sequences" to a dataset of allergens with FASTA using a 35% or greater identity threshold over any 80 amino acid length sequences to indicate the potential for IgE crossreactivity. Based upon this suggestion, algorithms have been developed to automatically generate all possible 80 residue subpeptides from a query protein and compare each peptide against a dataset of allergens ("sliding window search"). Any 80 amino acid peptide derived from the query protein that shows ≥35% identity to a known allergen triggers the need for additional testing (*i. e.*, an IgE screening study with sera from patients allergic to the identified protein) to establish the safety of the protein in the food supply and to

Correspondence: Dr. Gregory S. Ladics, Dupont/Pioneer Crop Genetics Regulatory Science and Registration, E400/4402, Route 141 & Henry Clay Road, Wilmington, DE 19880, USA
E-mail: gregory.s.ladics@usa.dupont.com
Fax: +1-302-695-3075

Abbreviations: FAO/WHO, Food and Agriculture Organization/World Health Organization; **FARRP,** Food Allergy Research and Resource Program; **NCBI,** National Center for Biotechnology Information

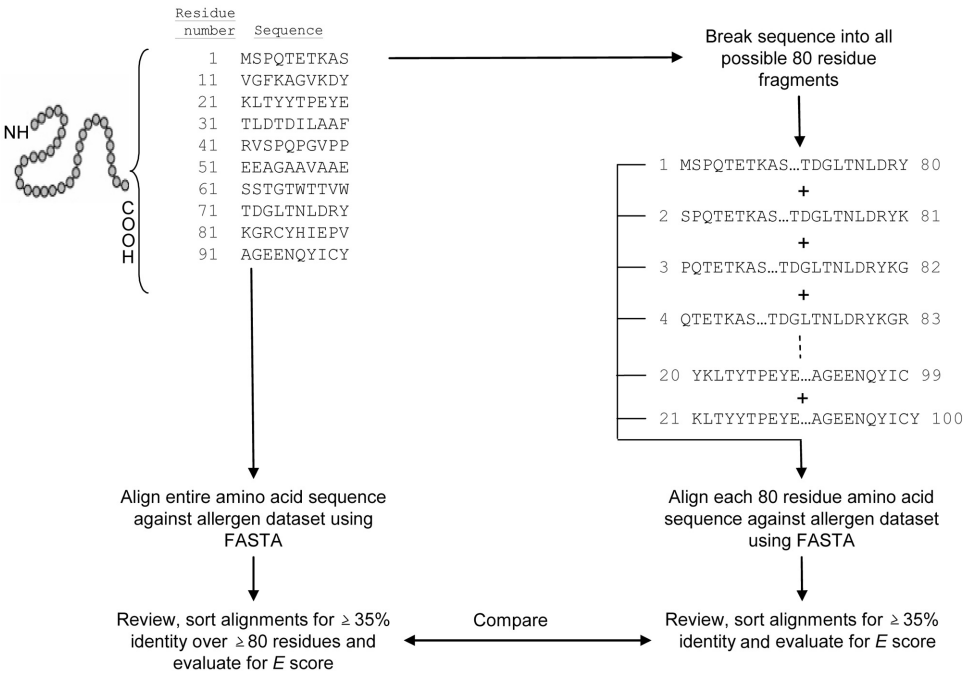

Figure 1. Schematic representation of conventional FASTA analysis of protein sequences *vs.* "sliding window" analysis. Sequences are aligned with dataset proteins conventionally using FASTA, or first broken down into all possible 80 residue "subpeptides"; each subpeptide is individually submitted for FASTA analysis. The resulting alignments are then sorted to reveal above threshold identities and *E* scores; these are then subsequently compared.

successfully register the transgenic plant product for sale in various geographies.

The use of 35% identity threshold, however, is considered to be overly conservative and likely results in a number of false positive findings. For example, others have reported that for crossreactivity to occur, a higher degree of similarity is needed, likely in excess of 50–70% sequence identity over significant spans of the target protein and allergen [3]. Radauer and Breiteneder [4] reviewed sequence identities among allergenic and nonallergenic homologs of pollen allergens and reported that the prerequisite for allergenic crossreactivity between proteins was a sequence identity of at least 50% across the length of the protein. In addition, after several years of experience in screening various proteins, we have come to believe that the use of a sliding window search in conjunction with a 35% identity threshold amplifies or exacerbates the number of potential false positive findings observed when comparing the amino acid sequences of novel proteins to those of known allergens. In order to help minimize false positive findings, a more scientifically valid approach would be to conduct the FASTA search in the conventional manner, utilizing the entire protein sequence as a query. Because the FASTA algorithm was designed to identify regions of local identity between proteins to generate an alignment, sequence along the entire length of the query protein is treated with equal weight when comparisons are made, negating the need to analyze protein sections independently. The length of sequence used to initiate and extend an aligned region is defined by the word size (k-tuple) and is set to a default value of two amino acids. Furthermore, using a sliding window search returns matches that are often inconsistent with the weight of the alignment as measured by the expectation (*E*) score. The *E* score reflects the potential random occurrence of aligned sequences and can be used to evaluate the significance of an observed alignment. The calculated *E* score depends on the overall length of gapped local sequence alignments, the quality (percent identity, similarity) of the overlap, and the size of the database [2, 5]. When comparing sequences, a very small *E* score may suggest a structurally relevant similarity, while large *E* scores (*i.e.*, >1.0) are typically associated with alignments that do not represent a biologically relevant similarity.

To test the effectiveness of these two methods, a series of *in silico* studies was performed using the Food Allergy Research and Resource Program (FARRP) Allergen Database (Version 6.0) along with a number of different protein datasets to compare the false positive and negative rates observed using a modified FASTA (*i.e.*, the 80 amino acid sliding window approach) *versus* conventional FASTA analysis at an identity threshold of 35% identity or greater (Fig. 1).

2 Methods

The following four datasets were analyzed to assess false positive rates:

(i) Hypothetical ORFs representing translations of maize genomic DNA: The ORFs were derived by using the FGENESH gene prediction program (Softberry, Mt. Kisco, NY) (data not shown). These predicted protein sequences have no homology to known proteins and there is no evidence of transcription or protein expression. For the 1102 hypothetical maize ORFs (168 sequences out of the original 1270 were less than 80 amino acids and could not generate positive hits at a length greater than 80 residues), the allergen dataset used for comparison was the Pioneer Hi-Bred International in-house allergen database. The database contained 2033 entries and was constructed by compiling protein allergen sequences identified by using keyword searches (*i.e.*, allergen(s); isoallergen(s) from published protein allergen databases [6 – 8] as well as the Swiss-Prot/TrEMBL, PIR, and GenPept, nr datasets). The FARRP6 Allergen Database (www.allergenonline.com; January, 2006) was employed for all other comparisons. All comparisons were carried out on an SGI MIPS R14000 computer running the IRIX Version 64 software. Briefly, either the FASTA33 or 34 programs, or a modified Perl script designed for sliding window searches (fastest33.pl, fastest34.pl) were run on all sequences in the query datasets. The sliding window scripts break a query protein sequence into all possible 80 residue subpeptides, run FASTA searches on each peptide, and return alignments that equal or exceed the FAO/WHO threshold. For all FASTA33 and 34 searches, the individual result files were concatenated into a single file; all relevant information was parsed into a summary file that was subsequently analyzed in Microsoft Excel. All matches displaying above threshold hits were identified and sorted by *E* score. Only the highest scoring match for each individual sequence was then used for the comparisons. Both conventional and sliding window FASTA searches generated additional above threshold alignments to other allergens at larger *E* scores; however, only the additional alignments generated using FASTA33 were compared in this study.

(ii) Randomly selected proteins from the Genpept dataset at National Center for Biotechnology Information (NCBI): A dataset of 1000 proteins was randomly selected from the GenePept dataset at NCBI (http://www.ncbi.nlm.nih.gov) (data not shown). Because 93 of the initial 1000 proteins were less than 80 amino acids and could not generate positive matches at a length greater than 80 residues only 907 proteins (containing both potential allergen homologs and sequences unrelated to allergens) were analyzed.

(iii) Randomly selected corn proteins: A set of amino acid sequences comprising all entries from maize (3989 accessions) were downloaded from the Uniprot dataset (http://www.pir.uniprot.org) and from this set 100 protein sequences (11 of these sequences were less than 80 residues long, the total number of sequences for percentage calculations were reduced to 89 accordingly; data not shown) were chosen at random for comparison of conventional and sliding window analysis using both FASTA33 and FASTA34.

(iv) Proteins specifically expressed in corn seed: To obtain corn protein sequences specific to seed, the edible part of the plant to which consumers are exposed, the NCBI database was searched for all proteins from corn. From this analysis, approximately 11 000 sequences were obtained. These proteins were further parsed to 248 proteins by removing hypothetical, predicted, putative, and unknown proteins, and then screening the remainder for protein sequences characterized from seed tissue. Of the 248 sequences, 14 were less than 80 amino acids while 133 were duplicates and were eliminated. This resulted in a dataset of 97 corn seed protein sequences (data not shown).

To evaluate false negatives, the following datasets were evaluated: (i) Bet v 1a and several crossreacting proteins, *i.e.*, carrot (Dau c 1); celery (Api g 1); apple (Mal d 1), cherry (Pru a 1), and pear (Pyr c 1): For the comparison of Bet v 1a and crossreacting fruit and vegetable proteins, allergen datasets were constructed that had most of the Bet v 1-like proteins removed. These datasets were then "spiked" with either the Bet v 1a protein, or the corresponding proteins from cherry (Pru a 1), celery (Api g 1), carrot (Dau c 1), apple (Mal d 1), or pear (Pyr c 1) and conventional and sliding window FASTA analysis conducted using FASTA33 and FASTA34 to determine whether key allergenic proteins would be missed using either method of FASTA analysis.

(ii) Evaluation of a bean α-amylase inhibitor transfected into pea: A nonallergenic bean α-amylase inhibitor (GI-47571317) that was transfected into pea was evaluated against the FARRP (6) Allergen Database. The bean α-amylase inhibitor expressed in the transformed pea was recently reported to display increased immunoreactivity in a nonvalidated animal model [9].

2.1 Analysis of a putative nonallergenic test protein containing target sequence from Ara h 1, a peanut allergen

To determine if, and how, a sliding window FASTA search differs from a conventional full-length FASTA search, a test protein containing a target sequence derived from Ara h 1 (GI-1168390), and a database composed of a subset of sequences derived from the FARRP (6) Allergen Database were assembled and used for comparison. The test protein sequence was composed of a single 20 amino acid segment, or a pair of variably spaced 10 amino acid segments from Ara h 1 inserted into the sequence of GI-2582631 (an acetate auxotroph from the bacteria, *Methanococcus maripaludis*). GI-2582631 was chosen due to its low degree of similarity to any sequence in the FARRP (6.0) Allergen Data-

base. The 20 amino acid target segment from Ara h 1 (amino acids 500–519) was inserted at position 60 of GI-2582631 and this sequence was used to query the FARRP Allergen Database. The 20 amino acid target segment was also split into two 10 amino acid segments (500–509 and 510–519), and these segments were inserted with variable spacing into GI-2582631 (*e.g.*, for a five amino acid spacing, one segment would be inserted at position 54 and the second at position 60) (Fig. 2). The database used for this analysis, AD6-1532, was identical to the FARRP (6.0) database with the exception that entries for Ara h 1 (GI-1168390 and -1168391) and conarachin (GI-46560472, -46560474, and -46560476) were removed. All searches were performed using FASTA version 3.3t05 on a Windows PC. Sliding window FASTA searches were implemented with FASTA Version 3.3t05 using DOS batch files.

Both older [Version 3.3t09 [FASTA33]; 2] and newer versions [Version 3.4t25 [FASTA34]; 10] of FASTA were utilized for the analysis of all datasets except for the 1270 maize ORFs and the putative nonallergenic test protein containing the target sequence from Ara h 1 (FASTA33 only).

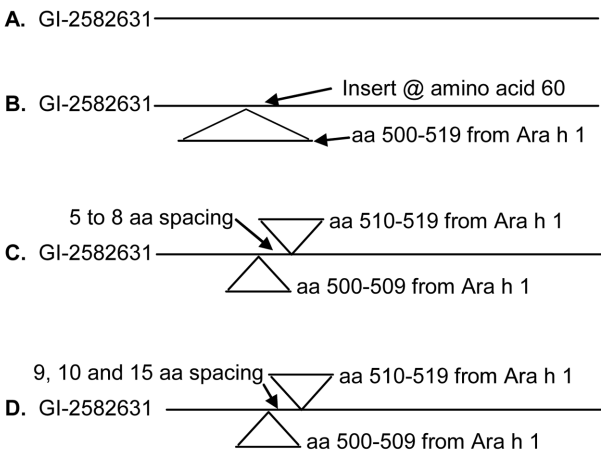

Figure 2. Analysis of a putative nonallergenic test protein containing a target sequence from the peanut allergen, Ara h 1. Construct a database consisting of AD6 minus Ara h 1 and conarachin. Conduct conventional FASTA search using A, B, and C. (A) No hits with *E* score <8. (B) Best hit is with pea vicilin. (C) Best hit with 5–8 amino acid spacing is pea vicilin. (D) Pea vicilin is no longer the top hit with conventional FASTA search. Test a sliding window search to determine if it is more sensitive than a conventional FASTA search. Seven structurally unrelated proteins including pea vicilin were identified as the top hit (depending upon the window) with a sliding window and a spacing of nine amino acids. Comparable results were obtained with 10 or 15 amino acid spacing. The sliding window search was no more reliable than the conventional FASTA search at identifying the target sequence.

3 Results

3.1 Analysis of 1102 hypothetical ORFs ≥80 amino acids from corn

A collection of ORFs (data not shown) encoding hypothetical maize proteins were subjected to allergen identity searches using an 80 amino acid sliding window FASTA33 search. A total of 73 hypothetical protein sequences out of 1102, or 6.7% of all the hypothetical protein sequences analyzed, exceeded the current threshold (*i.e.*, ≥35%) for allergenicity. When the conventional FASTA33 or FASTA34 search was used to examine the 1102 ORFs mentioned above, only 18 hypothetical protein sequences, or 1.7% of the total ORFs examined, exceeded the threshold for allergenicity (Fig. 3A). This represents approximately a five-fold decrease in the number of positive scores. These ORFs, representing translations of genomic DNA sequences using the FGENESH gene prediction program (Softberry), with no known matches to publicly available protein sequences, are unusually rich in low complexity sequences (*e.g.*, QQQQQ; PPPPPP). This could be one explanation for the dramatic difference between sliding window and conventional FASTA results when compared to the other datasets, as the sliding window search apparently magnifies the significance of low complexity sequences, as described below.

3.2 Analysis of 907 randomly selected protein sequences

When the protein sequences were subjected to the conventional FASTA analysis using either the FASTA33 or FASTA34 algorithm, 43 protein sequences (4.7%) exceeded the current recommended threshold for an identity match. In contrast, 104 protein sequences (11.5%) exceeded the threshold using a sliding window FASTA33 search, while 103 positives (11.4%) were returned when using a sliding FASTA34 search (Fig. 3B). Forty-one of the 43 sequences represented in the conventional search were present in the sliding search. *E* scores for the conventional FASTA searches were also generally much lower compared to those from the sliding window searches. For instance, the sliding window search had a larger number of *E* scores greater than 1.0 (*i.e.*, 21% of the total number of hits *vs.* 14% for the conventional FASTA33 search), which may suggest a higher number of alignments that do not represent a biologically relevant structural similarity (data not shown). These high *E* scores are reflected in the nature of the allergen hits returned using the sliding window search.

When the positive sequences from both the conventional and sliding window FASTA33 searches were analyzed for the presence of multiple above-threshold matches, approximately one half (52/104 for the sliding window search, and 19/43 for the conventional search) contained alignments to multiple FARRP sequences. The number of different allergens identified per query sequence ranged from 2 to 67, and all of the query proteins that were identified in the conventional FASTA search were contained in the sliding window

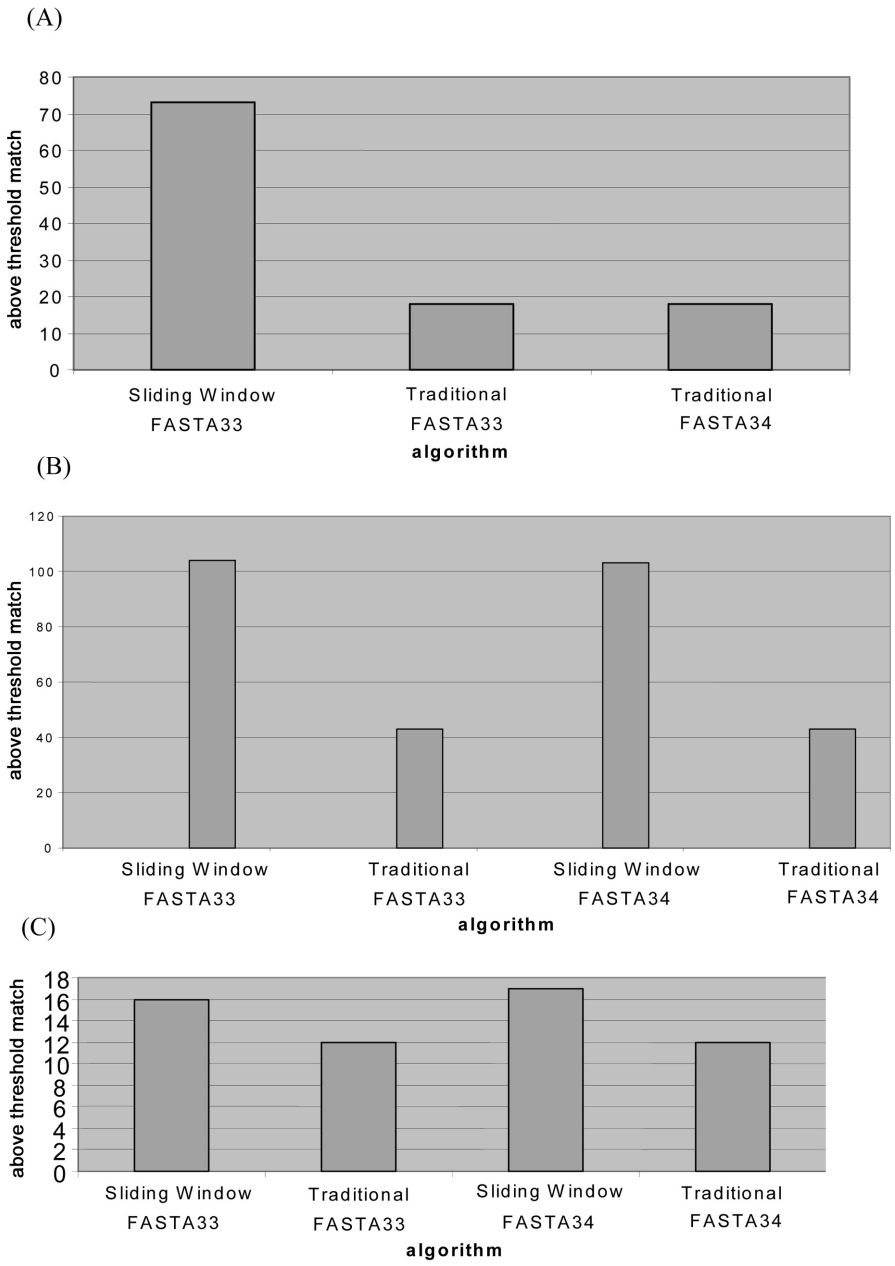

Figure 3. (A) Comparison of above threshold matches for 1102 hypothetical maize ORFs using different FASTA criteria. ORFs representing translations of maize genomic DNA encoding hypothetical maize proteins were derived by using the FGENESH gene prediction program (Softberry) and subjected to allergen identity searches using an 80 amino acid sliding window FASTA33 search or a conventional FASTA33 or FASTA34 search and the Pioneer in-house allergen database. (B) Comparison of above threshold matches for 907 randomly selected proteins using different FASTA criteria. Proteins randomly selected from the genpept dataset at NCBI were subjected to allergen identity searches using either an 80 amino acid sliding window or a conventional FASTA33 or FASTA34 search and the FARRP6 Allergen Database. (C) Comparison of above threshold matches for 89 randomly selected corn proteins using different FASTA criteria. Corn proteins randomly selected from the Uniprot dataset were subjected to allergen identity searches using either an 80 amino acid sliding window or a conventional FASTA33 or FASTA34 search and the FARRP6 Allergen Database. For all analysis, a 35% or greater identity threshold over any 80 or greater amino acid length sequence was utilized to indicate the potential for IgE crossreactivity.

Table 1. FASTA33 queries of the 907 randomly selected proteins producing multiple alignments to different allergen dataset accessions with large discrepancies between the sliding window and conventional searches

	Sliding FASTA33				Conventional FASTA33		
Query GI number	Number of matches	Avg identity (%)	Avg length (aa)	Query GI number	Number of matches	Avg identity (%)	Avg length (aa)
226743	14	36.04	85	226743	3	37.29	95
226896	12	37.20	82	226896	3	37.29	90
2267585	20	36.13	83	2267585	3	35.88	113
14268486	26	36.16	83	14268486	2	35.22	141
26985144	12	35.72	84	26985144	2	36.36	83

search. A list of proteins producing multiple alignments to different allergen dataset accessions with large discrepancies between the FASTA33 sliding window and conventional searches are listed in Table 1. For query sequences recognizing multiple allergens, the discrepency in *E* scores was even more pronounced, with only 3.1% of the conventional FASTA33 alignments producing *E* scores greater than or equal to 1, while 16% of the sliding window alignments generated to multiple allergen sequences using a single query generated *E* scores greater than or equal to 1.

One difference between the FASTA33 and FASTA34 algorithms is that the gap creation penalty has been reduced from −12 to −10, respectively. However, neither the conventional nor the sliding window results obtained using the FASTA34 algorithm displayed any significant change in the number of above threshold alignments when compared to the FASTA33 algorithm. When the conventional FASTA33 *versus* FASTA34 algorithms were compared, each produced two unique matches. The unique matches returned by the FASTA33 algorithm were close to the threshold (35.7%/84 residues and 36.4%/88 residues) and produced very high *E* scores (7.5 and 9.9, respectively). The two matches unique to the FASTA34 algorithm were also near the threshold (37.5%/80 and 35.8%/81), but returned lower *E* scores (0.0042 and 6.2, respectively). The lower *E* scores obtained with the FASTA34 algorithm suggest that the matches returned are more likely to be significant, although they would need to be further investigated. Expectation values for the FASTA34 results appear slightly lower, but in most cases the identity and length of the alignments are identical. Comparison of the FASTA34 conventional search to a sliding window search yielded results similar to those with FASTA33. Both the FASTA33 and FASTA34 sliding window searches produced 2.5 times the number of positive matches compared to the conventional FASTA searches.

3.3 Analysis of 89 randomly selected corn protein sequences

The results of the conventional *versus* sliding window analysis of the corn protein sequences using both FASTA33 and FASTA34 are summarized in Tables 2–4. An increase in the number of positives was observed with the sliding window searches compared to the conventional analysis (~19% *vs.* 12%, respectively, Fig. 3C). When extrapolated to include the total number of Uniprot derived maize accessions (*i.e.*, 3989), approximately 758 and 479 corn proteins, respectively, would be deemed potentially allergenic based on using a ≥35% identity threshold in conjunction with a sliding window and conventional FASTA analysis.

As with the 907 protein sequences mentioned above, there are several proteins that return different allergen matches, among these 89 proteins depending upon the method used; however, unlike the previous set of 907 randomly selected sequences, all allergen matches returned were similar in nature. For example, accession Q6JBQ0_ZEAMP, a chitinase, returned hits to allergenic chitinases from *Cryptomeria japonica* using the sliding window FASTA33, as well as the conventional FASTA34 search, while the sliding window FASTA34 search returned a chitinase from *Castanea sativa*; the FASTA33 conventional search returned a chitinase from *Persea americana* as the top hit. As noted previously, the conventional FASTA searches in general produced lower *E* scores (*i.e.*, <1) when compared to the sliding window searches due to the extended length of the alignments.

3.4 Analysis of 97 corn seed protein sequences

For the conventional FASTA33 or FASTA34 analyses of the corn seed proteins, 39% of these proteins were identified as potentially crossreactive with known allergenic proteins, while 53–54% were identified as allergenic using the sliding window analysis. Raising the threshold to 50% decreased the number of putative positive findings with the corn seed proteins by approximately half (17 and 25%, respectively) using either the FASTA conventional or sliding window analysis. Increasing the threshold to 70% further diminished the number of putative positive findings (3 and 6%, respectively) using either the FASTA conventional or sliding window analysis (data not shown).

Table 2. Sliding window FASTA33 above threshold allergen matches for the randomly selected corn proteins

Protein	Description	Sliding FASTA33 positives Match accession number	Allergen	Identity	Length	*E* value
GLU2_MAIZE	Glutelin-2 precursor (Zein-gamma) (27 kDa zein) (Alcohol-soluble reduced glutelin) (ASG) (Zein ZC2).	gi|170734|gb|AAA34287.1	*Gamma gliadin B-III*[b]	40.2	87	0.00017
MOSA_MAIZE[a]	**Autonomous transposable element EN-1 mosaic protein (Suppressor – mutator system protein) (SPM)**	**gi|450239|gb|AAA53071.1|**	**PkIWl501**	**35**	**80**	**9.6**
O50018_MAIZE	Elongation factor 1-alpha	gi|21632054|gb|AAK85129.1|	Elongation factor (*Juniperus ashei*)	95	80	2.80E-37
Q2XXB6_ZEAMP	Pathogenesis-related protein 6	gi|62149372|dbj|BAD93486.1|	*Pollen allergen CJP38* (*C. japonica*)[b]	68.7	83	1.30E-24
Q41759_MAIZE[c]	Hypothetical protein	gi|1168391|sp|P43238|ALL12_ARAHY	Allergen Ara h 1, clone P41B precursor (Ara h 1)	35	80	0.14
Q41830_MAIZE	Mgp1 GTP-binding protein	gi|21217443|gb|AAM33785.1|	Rab11 (*Periplaneta americana*)	80	80	4.70E-33
Q41839_MAIZE	Polygalacturonase (Fragment)	gi|4826572|emb|CAB42886.1|	Polygalacturonase (*Phleum pratense*)	70	80	4.10E-30
Q41860_MAIZE	Transposable element Mu1 sequence	gi|42820661|emb|CAF31974.1|	Suppressor protein spt23-related, with ankyrin repeats (*Aspergillus fumigatus*)	35.6	90	0.75
Q4A1J1_MAIZE	cc10	gi|40807635|gb|AAR92223.1|	Phytocystatin (*Actinidia deliciosa*)	36.6	82	3.90E-05
Q64HB7_MAIZE	ASF/SF2-like pre-mRNA splicing factor SRP31	gi|63887|emb|CAA31942.1|	Vitellogenin (*Gallus gallus*)	40	80	1.50E-05
Q6JBQ0_ZEAMP	Chitinase	gi|56550550|dbj|BAD77932.1|	*Class IV chitinase* (*C. japonica*)[b]	55.4	83	7.30E-22
Q94FF5_MAIZE	Globulin 1 (Fragment)	gi|13183177|gb|AAK15089.1|AF240006_1	7S globulin (*Sesamum indicum*)	47.5	80	3.60E-18
Q9ATL1_MAIZE	Rretrotransposon gag protein	gi|736319|emb|CAA27052.1|	Glutenin (*Triticum aestivum*)	35.8	81	0.47
Q9SWU7_MAIZE	Receptor-like kinase (Fragment)	gi|22726221|gb|AAN05083.1|	Major antigen-like protein (*Salsola kali*)	43.2	81	2.70E-11
Q9ZTL2_MAIZE	Cell wall invertase Incw1 (EC 3.2.1.26)	gi|18542113|gb|AAL75449.1|AF465612_1	Minor allergen β-fructofuranosidase precursor (*Lycopersicon esculentum*)	65	80	2.50E-26
Q9ZTQ5_MAIZE	Cell wall invertase (EC 3.2.1.26)	gi|18542113|gb|AAL75449.1|AF465612_1	Minor allergen β-fructofuranosidase precursor (*L. esculentum*)	61.3	80	4.40E-25

a) Bold text highlights differences in above threshold sequences that vary with criteria.
b) Italicized text displays allergen matches that vary with criteria used.
c) Underlined text highlights Ara h 1 similarity.

3.5 Comparison of bet v 1a and crossreacting fruit and vegetable proteins

Both the conventional and sliding window FASTA analyses correctly identified above threshold similarities between Bet v 1a and the crossreacting fruit and vegetable proteins and the false negative rates of both methods of analysis were the same (Table 5). As noted previously, the conventional FASTA resulted in lower *E* scores compared to the sliding window analysis.

Table 3. Sliding window FASTA34 above threshold allergen matches for the randomly selected corn proteins

Peptide	Description	Match accession number	Allergen	Identity	Length	*E* value
GLU2_MAIZE	Glutelin-2 precursor (Zein-gamma) (27 kDa zein) (Alcohol-soluble reduced glutelin) (ASG) (Zein ZC2).	gi|62484809|emb|-CAI78902.1	*Putative gamma-gliadin*[b] (T. aestivum)	36.1	83	0.043
O50018_MAIZE	Elongation factor 1-alpha	gi|21632054|g-b|AAK85129.1|	Elongation factor (*J. ashei*)	93.8	80	1.20E-36
Q2XXB6_ZEAMP	Pathogenesis-related protein 6	gi|1184668|g-b|AAA87456.1	*β-1,3-Glucanase*[b]	56.8	81	1.70E-15
Q41759_MAIZE[c]	Hypothetical protein	gi|1168391|sp|-P43238|ALL12_ ARAHY	Allergen Ara h 1, clone P41B precursor (Ara h 1)	35	80	0.1
Q41830_MAIZE	Mgp1 GTP-binding protein	gi|21217443|g-b|AAM33785.1	Rab11 (*P. americana*)	75	80	2.40E-30
Q41839_MAIZE	Polygalacturonase (Fragment)	gi|4826572|emb|-CAB42886.1	Polygalacturonase (*P. pratense*)	70	80	1.40E-28
Q41860_MAIZE	Transposable element Mu1 sequence	gi|42820661|emb|-CAF31974.1	Suppressor protein spt23-related, with ankyrin repeats (*A. fumigatus*)	36.5	85	1.4
Q4A1J1_MAIZE	cc10	gi|40807635|g-b|AAR92223.1	Phytocystatin (*A. deliciosa*)	36.3	80	0.00012
Q64HB7_MAIZE	ASF/SF2-like pre-mRNA splicing factor SRP31	gi|63887|emb|-CAA31942.1	Vitellogenin (*G. gallus*)	40	80	8.50E-05
Q6JBQ0_ZEAMP	Chitinase	gi|1359600|emb|-CAA64868.1	*Chitinase Ib (C. sativa)*[b]	39.3	84	2.80E-08
Q94FF5_MAIZE	Globulin 1 (Fragment)	gi|13183177|gb|AA-K15089.1|AF240006_1	7S globulin (*S. indicum*)	42.7	82	3.40E-09
Q9ATL1_MAIZE	Rretrotransposon gag protein	gi|736319|emb|-CAA27052.1| glutenin (*T. aestivum*)	Glutenin (*T. aestivum*)	35.4	82	0.39
Q9ATN0_MAIZE[a]	**Plasma membrane integral protein ZmPIP1-6**	**gi|20502989|gb|AA-M22698.1|A-C098693_3**	**Putative pollen allergen (*Oryza sativa* (japonica cultivar-group))**	**35.3**	**85**	**1.9**
Q9LEE9_MAIZE	**OCL5 protein**	**gi|55859462|emb|-CAH92635.1| pollen allergen Hor v 4 H**	**Pollen allergen Hor v 4 (*Hordeum vulgare*)**	**35.3**	**85**	**4.8**
Q9SWU7_MAIZE	Receptor-like kinase (Fragment)	gi|22726221|g-b|AAN05083.1| major antigen-like protein	Major antigen-like protein (*S. kali*)	36.6	82	5.30E-08
Q9ZTL2_MAIZE	Cell wall invertase Incw1 (EC 3.2.1.26)	gi|18542113|gb|AA-L75449.1|AF465612_1 minor allergen be	Minor allergen β-fructo-furanosidase precursor (*L. esculentum*)	57.5	80	2.70E-19
Q9ZTQ5_MAIZE	Cell wall invertase (EC 3.2.1.26)	gi|18542115|gb|AA-L75450.1|AF465613_1 minor allergen be	Minor allergen β-fructo-furanosidase precursor (*L. esculentum*)	56.3	80	1.00E-19

a) Bold text highlights differences in above threshold sequences that vary with criteria.
b) Italicized text displays allergen matches that vary with criteria used.
c) Underlined text highlights Ara h 1 similarity.

3.6 Comparison of the bean α-amylase inhibitor transfected into pea to the FARP allergen database

Output from each of the two sliding window searches returned 3 above threshold identities. The highest, with identities of 35–51% over 80 residues, was to a soybean lectin (GI-170006), followed by a peanut agglutinin (GI-253289) with identities of 38–44%. Lastly, a glucose/mannose binding lectin from peanut (GI-951118) was reported, with identities between 35 and 39%. Both the conventional searches returned only the first match, with an alignment displaying 41% identity over 251–253 residues. Although not above the 35% over 80 residue threshold, the other two accessions were also captured in the conventional output.

Table 4. Randomly selected corn proteins displaying alignments to different allergen dataset accessions depending upon criteria used

Protein	FASTA33				FASTA34			
	Allergen match	% Identity	Length (aa)	*E* value	Allergen match	% Identity	Length (aa)	*E* value
Q2XXB6_ZEAMP	β-1,3-Glucanase-like protein (*Olea europaea*)	39.47	337	2.00E-16	β-1,3-Glucanase (*Hevea brasiliensis*)	53.15	333	3.8E-40
Q6JBQ0_ZEAMP	Endochitinase (*P. americana*)	44.05	311	3.20E-22	Class IV chitinase (*C. japonica*)	52.00	275	1.5E-52
Q94FF5_MAIZE	7S Globulin (*S. indicum*) cupin	36.89	225	8.50E-21	48-kDa Glycoprotein precursor (*C. avellana*) Cupin	37.17	191	1.2E-18

Though the number of hits returned varied, all four methods successfully identified the bean α-amylase inhibitor protein as a potential allergen.

3.7 Analysis of a putative nonallergenic test protein containing a target sequence from the peanut allergen, Ara h 1

When used as a query for a conventional FASTA search of AD6-1532 or FARRP (6.0) allergen databases, GI-2582631 (an acetate auxotroph from the bacteria, *M. maripaludis*) yielded no alignments with an *E* score <8 when a gap initiation penalty of 12 and gap extension penalty of 2 were employed. When a 20 amino acid target segment from Ara h 1 (amino acids 500–519 from GI-1168390) was inserted at position 60 of GI-2582631 and this sequence was used to query AD6-1532, the best alignment recovered was to *Pisum sativum* vicilin (GI-42414629). When the 20 amino acid target segment was split into two 10 amino acid segments (500–509 and 510–519), and these segments, inserted with variable spacing into GI-2582631 (for a five amino acid spacing, one segment would be inserted at position 54 and the second at position 60), a FASTA search returned *P. sativum* vicilin or a closely related homolog Len c 1.0102 (GI-29539111) as the best overall alignment for all segment spacing up to eight amino acids. Once a spacing of nine amino acids is placed between the two ten amino acid segments, a FASTA search identified a high molecular weight dust mite protein (GI-6492307) as the best alignment and *P. sativum* vicilin and Len c 1.0102 were the fifth and sixth best alignments, respectively. The conventional FASTA search was exceptionally sensitive and was able to identify the two target, ten amino acid sequences reliably with spacing of up to eight amino acids. Once the nine amino acid spacing was inserted, however, the full length FASTA search was no longer able to identify the target sequences in the tester protein.

In order to determine if a sliding window added to the sensitivity or reliability of a FASTA search, the tester protein was used as a query for 80 amino acid sliding window search. The AD6-1532 database was queried with the tester sequence that contained the ten amino acid target sequences separated by nine amino acids. The tester protein yielded a top alignment with seven structurally unrelated proteins in the AD6-1532 database depending upon the search window. These proteins included Len c 1.0102, high molecular weight dust mite protein, eosinophil granule major basic protein 2 precursor, thaumatin-like protein, ribosomal protein S12, MAG_DERFA (American house dust mite allergen), and Bos d 2.0102. If each top alignment is inspected and percent identity and alignment window size are examined, the most significant of the top alignments displays 33.333% identity (36.923% ungapped) in 72 amino acid overlap with the high molecular weight dust mite protein. When 10 amino acids were used to separate the target sequences in the tester protein, the sliding 80 amino acid window FASTA search also, depending upon the window identified 7 proteins as the top alignment. However, two of the seven proteins identified relative to the test protein with the ten amino acid spacing between the target sequences differed from those having the nine amino acid spacing (*i.e.*, high molecular weight dust mite protein is no longer identified as a top alignment by any window and the top alignment displaying 24.074% identity (24.074% ungapped) in a 54 amino acid overlap is with Bos d 2.0102). Although certain windows in the sliding window search were able to identify Len c 1.0102, the use of criteria such as alignment length, or combination of length and identity to select the most significant alignment in a series of sliding window searches is no more reliable than the conventional FASTA search at identifying the target sequence. When 15 amino acids were used to separate the target sequences in the tester protein, 50 of the 88 sliding search windows identified Len c 1.0102 as the top alignment. Of the 50 alignments with Len c 1.0102, 36 alignments were the product of the insertion of a 15 amino acid gap.

4 Discussion

Comparison of the amino acid sequence of novel proteins for similarity to known or putative allergens is an important

Table 5. Comparison of Bet v 1a to crossreacting proteins using different FASTA criteria

Comparison	FASTA33 sliding window			FASTA34 sliding window			Traditional FASTA33			Traditional FASTA34		
	Identity	Length	E	Identity	Length	E	Identity	Length	E	Identity	Length	E
Betv1 *vs.* dauc1	40	80	6.50E-10	40	80	1.90E-11	38.1	155	1.00E-18	38.1	155	1.90E-20
Dauc1 *vs.* betv1	40.7	81	5.20E-11	40.7	81	4.30E-10	38.1	155	5.60E-18	38.1	155	2.10E-19
Betv1 *vs.* apig1	45	80	1.40E-10	45	80	2.10E-12	41.9	155	3.50E-21	41.9	155	3.60E-23
Apig1 *vs.* betv1	45	80	7.90E-12	45	80	2.30E-12	41.9	155	2.60E-22	41.9	155	1.70E-24
Betv1 *vs.* mald1	61.3	80	8.10E-19	61.3	80	1.60E-19	56	159	2.50E-31	56	159	2.70E-34
Mald1 *vs.* betv1	61.3	80	1.50E-21	61.3	80	6.30E-24	56	159	4.50E-28	56	159	8.50E-33
Betv1 *vs.* pyrc1	62.5	80	1.50E-18	62.5	80	3.00E-19	57.5	160	1.00E-32	57.5	160	3.70E-35
Pyrc1 *vs.* betv1	62.5	80	3.50E-22	62.5	80	6.80E-24	57.5	160	1.90E-33	57.5	160	5.10E-37
Betv1 *vs.* prua1	62.5	80	1.50E-19	62.5	80	5.30E-20	59.4	160	4.00E-35	59.4	160	1.80E-38
Prua1 *vs.* betv1	62.5	80	2.70E-21	62.5	80	1.90E-23	59.4	160	4.60E-39	59.4	160	2.50E-43

part of the safety assessment of expressed proteins in transgenic plant products. Part of this analysis involves using the FASTA algorithm [2] to search for identities in amino acid sequences that may correspond to potential IgE cross-reactivity to known or putative allergenic proteins. The objective of this study was to compare the false positive and false negative rates for two FASTA methods (*i.e.*, the sliding window *vs.* a conventional FASTA analysis). To accomplish this, a number of data sets derived from hypothetical ORFs from corn, randomly selected proteins, and corn proteins, as well as Bet v 1a homologs, an α-amylase inhibitor from bean, and a putative nonallergenic test protein containing a target sequence from the peanut allergen, Ara h 1 were utilized. Both FASTA Version 33 and 34 were employed for this comparison.

One difference between the FASTA33 and FASTA34 algorithms is that the gap creation penalty has been reduced from −12 to −10, respectively. This reduction would be expected to increase the number of gaps inserted into an alignment and therefore, increase the likelihood of any given match exceeding the FAO/WHO criteria. However, neither the conventional nor the sliding window results using the FASTA34 algorithm displayed any significant change in the number of above threshold alignments compared to FASTA33.

When a collection of ORFs encoding hypothetical maize proteins were analyzed using the sliding window search, ∼7% of all sequences evaluated exceeded the current threshold of ≥35% identity, while the use of the conventional FASTA algorithm resulted in a five-fold decrease in the number of positive scores (*i.e.*, above threshold) observed with the dataset. Corn is not considered to be a major food allergen and has been classified as a "less common allergenic food" [11]. In addition, Moneret-Vautrin *et al.* [12] concluded that food allergy to corn is rare on the basis of a retrospective study on patients with histories of food allergy. The number of observed findings with the 1102 ORFs, therefore, undoubtedly reflects a large number of false positives. In addition to the five-fold increase in positive findings, the sliding window search also excluded ∼13% of the sequences from analyses because they were less than 80 amino acids in length. Because the ORFs examined were hypothetical, it was not possible to determine whether any of the positive results corresponded to cross-reacting allergens. Therefore, the FASTA analysis comparison was conducted on a series of 1000 randomly selected protein sequences.

Using the conventional FASTA analysis with either the FASTA33 or FASTA34 algorithm to evaluate the 907 randomly selected proteins resulted in 2.5-fold less positive matches compared to sliding window searches. The observed percentages (4.7 and 11.5, respectively) of positive matches for the conventional and sliding window analysis, however, were higher than the expected percentage of real allergens (*e.g.*, ∼0.4% for Swiss-Prot based on Swiss-Prot allergen index) [13]. This finding is likely due to the use of the currently recommended threshold of 35% [1, 14]. Data suggest that for two proteins to immunologically crossreact, a large degree of identity (in the order of 50–70%) is needed [3, 4]. *E* scores for the conventional FASTA searches were also generally much lower compared to those from the sliding window searches due to the extended

length of alignment that is possible with the conventional search. A lower *E* score may suggest a structurally relevant similarity, while large *E* scores (*e.g.*, >1.0) are typically associated with alignments that do not represent a biologically relevant structural similarity. Similar to the ORF analysis, the sliding window search excluded ~9% of the sequences from analysis because they were less than 80 amino acids in length. When the comparison between the conventional and sliding window FASTA33 analysis with the randomly selected proteins was examined further, most striking was the fact that the sliding window search resulted in 61 additional positive matches compared to the conventional analysis. Forty-one of the 43 sequences represented in the conventional search were present in the sliding search. Importantly, there were no instances identified where a sliding window search provided an informative result that differed from those obtained with a conventional search.

One of the protein sequences not present in the sliding window searches was a cytochrome oxidase from *Kradibia jacobsi*, which has similarity to a putative allergenic relative from *Sarcoptes scabiei*. This alignment generated an *E* score of 10^{-15} in the conventional searches, suggesting a high degree of potential significance. Although most proteins were represented in both the conventional and sliding window searches, in some cases the allergens matched were different. In one example, the conventional FASTA33 search identified a legume protein sequence from pea as very similar (47.5% identity over 519 residues) to an 11S globulin-like protein from *Corylus avellana*. The top match returned by the sliding window search using the same query protein sequence was to a glycinin subunit from wild soybean (*Glycine soja*). While also significant (75% identity over 80 residues), the ability to extend alignments beyond the 80 residue threshold using the conventional search generates an *E* score that is 20 orders of magnitude greater (*i.e.*, <1) than that for the sliding window search.

Although the highest scoring matches for both the sliding window and conventional searches largely matched the same accessions, analysis of the alignments produced to multiple distinct allergen dataset acccessions revealed five separate instances where a large number of positives were returned when compared to the conventional FASTA33 search (Table 1). A closer examination of the individual alignments reveals the additional matches generated by the sliding window search are based upon multiple stretches of low complexity sequence, such as QQQQ, PPPP, or EEEE. Within the context of a conventional search, these short sequences are part of a larger alignment window (the range for the examples is from 83 to 141 amino acids). In contrast, when the window size is reduced to 80 residues (range is from 82 to 85 amino acids for the sliding search), these regions are of greater influence, resulting in an increased number of above threshold alignments to sequences containing short stretches of matching sequence. These short repetive sequences are the hallmark of many celiac proteins, such as the gliadin and glutenins from wheat, which make up the majority of the additional positives returned. Sliding window positive alignments can also be generated based upon short matching "words" that are not repetitive. An example of this is found when a heat shock protein from *Bradyrhizobium* (GI 12642164) is used as a query protein. Both the sliding window and conventional searches return multiple hits including the allergens Cla h 4 and Pen c 19, both of which are heat shock proteins, but the sliding window search returns an additional alignment to a tropomyosin from cockroach. This single alignment has a much higher *E* score than the others (4.2 *vs.* 10^{-26} for the heat shock proteins) and appears to be due to the presence of the matching peptide AEADKK at the beginning of the alignment.

Based on these data, it appears that subjecting a protein to a sliding window search is more likely to result in a match that, in some cases, is not functionally related to the query protein. Crossreactive allergens are typically functionally/structurally related. For example, Breiteneder and Ebner [15] reported that plant food allergens are either homologous to pathogenesis-related-type proteins or belong to a small number of protein classes, such as seed storage proteins or enzyme inhibitors. Mills *et al.* [16] further indicated that plant food allergens are members of three structurally related superfamilies that include: the prolamin superfamily (2S albumins, nonspecific lipid transfer proteins, and cereal α-amylase/trypsin inhibitors), the cupin superfamily (7S and 11S storage proteins from peanut, soy and tree nuts), and cysteine proteases (papain-like proteases). The majority of plant food allergens are either protective or storage proteins [17]. Jenkins *et al.* [18] further confirmed these findings by determining that the majority of plant food allergens belong to only four structural families (*i.e.*, prolamin, Betv 1 family, cupin, and profilins) accounting for over 65% of food allergens. Bredehors and David [19] also concluded that functional aspects of some allergens might play a role in the allergic response. To date, IgE crossreactivity between structurally unrelated allergens has not been demonstrated [20].

In order to compare the conventional and sliding window methodologies with a food not considered to be a major allergen [11], 89 random sequences from corn and 97 sequences specific to corn seed (*i.e.*, the edible part to which consumers are exposed) were obtained. Extrapolating the data from the 89 random sequences obtained from corn indicated that approximately 758 (19%) and 479 (12%) of Uniprot derived maize accessions, respectively would be identified as potentially allergenic based on the sliding window and conventional FASTA analysis. These data are similar to the percentage (*i.e.*, 18%) observed by Hileman *et al.* [21] in which 50 randomly selected corn proteins were evaluated using the conventional FASTA analysis. Similarly, 39% of corn seed proteins were identified as

```
>>gi|1168391|sp|P43238|ALL12_ARAHY Allergen Ara h 1, clo  (626 aa)
 initn:  75 init1:  50 opt:  88  Z-score: 117.0  bits: 28.8 E(): 0.16
Smith-Waterman score: 88; 35.000% identity (43.077% ungapped) in 80 aa
overlap (5-78:75-145)

gi|116 MRGRVSPLMLLLGILVLASVSATHAKSSPYQKKTENPCAQRCLQSCQQEPDDLKQKACES
               10        20        30        40        50        60

                                10        20        30        40
Q41759           SAASPRG------RRAPVLHRALRRHPRHVRADDIRRHGRRDTVDARHLR
                    |||      :|:|  :|:  |:|     || ||: ||:   :|
gi|116 RCTKLEYDPRCVYDPRGHTGTTNQRSPPGERTRGRQPGDY--DDDRRQPRREE-GGRW--
               70        80        90       100         110

           50        60        70        80
Q41759 EHAPAPRREGRLRHLPRVSRQDTRRPPRDTQRPRFL
         :||  || : ::  |  |:| ||| :  |:||
gi|116 --GPAGPREREREEDWRQPREDWRRPSH--QQPRKIRPEGREGEQEWGTPGSHVREETSR
           120       130       140         150       160       170
```

Figure 4. Maize protein Q41759 peptide 503 Alignment against peanut allergen Ara h 1 obtained following analysis of 89 randomly selected corn proteins using a sliding window FASTA33 search. The alignment was observed exclusively with the sliding window search. The identity match occurred exactly at the 35% threshold and involved only 74 residues. The sliding window algorithm inserted six gaps and extended the effective alignment length, thus triggering the identity match.

potentially allergenic with the conventional FASTA analysis, while 53 – 54% were identified as allergenic using the sliding window analysis. The number of observed findings with the selected corn proteins unquestionably indicates a large number of false positives and a gross overestimation of the number of allergenic proteins in corn. These data are again likely attributed to the use of the stringent threshold of 35% and clearly represents an unrealistic view of the potential allergenic proteins in corn, which is a less common allergenic food. For example, if the threshold is raised to 50% as suggested by Aalberse [3] and Radauer and Breiteneder [4], the number of positive findings with the corn seed proteins is decreased by approximately half (17 and 25%, respectively) using either FASTA conventional or sliding window analysis. Increasing the threshold to 70% further diminishes the number of positive findings (3 and 6%, respectively) using either FASTA conventional or sliding window analysis, providing a more realistic estimate of allergenic proteins in corn [22, 23].

One finding with the randomly selected maize proteins that warrants additional discussion involves the identity match of the maize protein Q41759 (a hypothetical corn protein) to the peanut allergen Ara h 1. This alignment was observed exclusively with the sliding window search. The identity match occurred exactly at the 35% threshold and involved only 74 residues. Although less than 80 amino acids in length, the sliding window algorithm inserted 6 gaps and extended the effective alignment length, thus triggering the identity match (Fig. 4). The nature of this alignment and *E* score obtained (*i.e.*, 0.16), coupled with its absence in the conventional FASTA searches suggests that this identity match is an artifact/false positive resulting from the use of the sliding window algorithm.

The main empirical data to support the establishment of an identity threshold of 35% in 2001 came from a paper analyzing the apparent cross-reactivity of the birch pollen allergen Bet v 1 with proteins from cherry, apple, pear, celery, and celery [24] The relatively low level of amino acid sequence similarity observed, particularly when comparing the celery allergen Api g 1 and the carrot allergen Dau c 1 to Bet v 1 (~40% identity), in conjunction with reported cross-reactivity, served as the basis for the establishment of a threshold (*i.e.*, 35%) that would identify such relationships. Because of the role of the Bet v 1-like allergens played in defining the criteria currently recommended, a crucial benchmark for any *in silico* analysis would be to recognize and identify similarity between these allergens. Based on the analysis with Bet v 1 homologs, there was no difference in false positive rate observed between the conventional *versus* sliding window FASTA analyses. Therefore, the conventional FASTA algorithm is appropriate for detecting identities at or near the current recommended threshold of 35%. Interestingly, the lowest above threshold identity observed was 38% (Bet v 1 *vs.* Dau c 1; Table 5).

Recently, a nonallergenic α-amylase inhibitor from bean, when transformed into pea, displayed increased immunoreactivity in a nonvalidated animal model [6]. While this protein has generated much interest of late, it is not likely the protein would have been commercialized based on the current weight of evidence approach due to its observed identity to several allergenic lectins. Nevertheless, any modification of the FASTA analysis procedure should be evaluated against this protein to ensure that a positive match was returned. The bean α-amylase inhibitor protein was identified as a potential allergen using either the FASTA33 or 34 sliding window or conventional analysis. This *in silico* finding, however, would require further analysis and testing with sera from appropriate allergic patients to further investigate whether it would constitute a risk for individuals with specific allergies.

The inability to reliably identify any further target sequence from Ara h 1 in a putative nonallergenic test protein using the sliding window search *versus* a conventional search was for several reasons not unexpected. By using a sliding window, sequence is removed from the context of the entire protein. In the absence of the entire protein sequence, FASTA will insert gaps and generate a "globalized" alignment where sequence in the window is "stretched" to fit across the length of the database sequence. The overall impact of a sliding window is in some respects comparable to lowering the gap initiation and gap extension penalties. This was clearly demonstrated with the test sequence that contained the 15 amino acid spacing of the target sequences. In those instances that a window from the test protein aligned with Len c 1.0102, the alignment always included a 15 amino acid gap. Such globalized alignments are unlikely to reflect *bona fide* structural homology as they may be excessively gapped.

Searches using a sliding window will also tend to exaggerate the effect of low complexity regions on an alignment, as illustrated with the query sequences that generated multiple matches (Table 1). Low-complexity sequences yield alignments that are statistically significant but have little biological relevance. Although the Ara h 1 segment containing test proteins described herein did not contain any regions of low complexity, the use of criteria such as a 35% identity in 80 amino acids for the assessment of the significance of an alignment does not take advantage of the sophisticated statistical analyses (*i. e.*, a histogram of the identity scores and an *E* score) performed by the FASTA algorithm. These analyses include a histogram of the similarity scores and an *E* score. Inspection of the similarity histogram permits one to determine if the query sequence contains regions of low complexity. The *E* score is a statistical measure of the likelihood that the alignment is reliable. An *E* score of 1 or greater indicates that the alignment generated between the query and the database protein is no more meaningful than the alignment that would be obtained if the query sequence were shuffled prior to conducting the search. Therefore, the use of criteria such as alignment length (*i. e.*, 80 amino acid window), or combination of length and identity to select the most significant alignment in a series of sliding window searches is unreliable.

In summary, the data indicate that a conventional FASTA analysis compared to the sliding window analysis using the currently recommended threshold criteria of 35% or greater identity results in fewer potential false positive findings, while providing an equivalent false negative rate. The positive results obtained with the conventional FASTA analysis, however, still exceeded what would be predicted based on the expected percentage of real or true allergens in the clinic. This finding is likely attributed to the use of the currently recommended threshold criteria of 35%. For example, when the threshold was raised to 50% when evaluating corn seed protein sequences, the number of positive findings decreases by half using either the conventional or traditional FASTA analysis. In addition, the *E* values associated with the use of a conventional FASTA analysis were in general greater (<1) than those observed with the sliding window analysis and may suggest a more relevant identity to the query protein. Data further indicate that the use of criteria such as alignment length or a combination of length and identity to select the most significant alignment in a series of sliding window searches is unreliable. This is due to the following: (i) a sliding window search takes what should be a local alignment and makes it a global alignment by removing the sequence from the context of the entire protein. In the absence of the entire protein sequence, FASTA will insert gaps, and generate a "global" alignment where sequence in the window is stretched to fit across the length of the database sequence; (ii) the sliding window scoring regime does not take advantage of the statistical analysis performed by the FASTA algorithm (*i. e.*, a histogram of the similarity scores and an *E* score). Finally, the conventional FASTA analysis resulted in identity matches that better reflected functional similarities between proteins. In some cases, the sliding window analysis resulted in identity matches to a variety of proteins from different families with diverse functions. These data indicate that the 80 amino acid sliding window approach results in a greater number of potential false positive findings, as there appears to be little scientific justification for many of the matches (*i. e.*, matches occur between functionally divergent proteins). Therefore, it is recommended that the conventional FASTA analysis be conducted to compare the identity of a protein to known allergens.

5 References

[1] Food and Agriculture Organization, *Evaluation of Allergenicity of Genetically Modified Foods: Report of a Joint FAO/WHO Expert Consultation*, Rome 2001.

[2] Pearson, W. R., Lipman, D. J., Improved tools for biological comparison, *Proc. Natl. Acad. Sci. USA* 1988, *85*, 2440 – 2448.

[3] Aalberse, R. C., Structural biology of allergens, *J. Allergy Clin. Immunol.* 2000, *106*, 228 – 238.

[4] Radauer, C., Breiteneder, H., Pollen allergens are restricted to few protein families and show distinct patterns of species distribution, *J. Allergy Clin. Immunol.* 2006, *117*, 141 – 147.

[5] Baxevanis, A. D., Ouellette, B. F. F. (Eds.), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, John Wiley & Sons, Inc., New York 1998.

[6] Gendel, S. *Adv.* Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods, *Food Nutr. Res.* 1998, *42*, 63 – 92.

[7] King, T. P., Hoffman, D., Lowenstein, H., Marsh, D. G., *et al.*, Allergen nomenclature, *Int. Arch. Allergy Immunol.* 1994, *105*, 224 – 233.

[8] Metcalfe, D. D., Astwood, J. D., Townsend, R., Sampson, H. A., *et al.* Assessment of the allergenic potential of foods derived from genetically engineered crop plants, *Crit. Rev. Food Sci. Nutr.* 1996, *36*, S165 – S186.

[9] Prescott, V. E., Campbell, P. M., Moore, A., Mattes, J., *et al.*, Transgenic expression of bean α-amylase inhibitor in peas results in altered structure and immunogenicity, *J. Agric. Food Chem.* 2005, *53*, 9023 – 9030.

[10] Reese, J. T., Pearson, W. R., Empirical determination of effective gap penalties for sequence comparison, *Bioinformatics* 2002, *18*, 1500 – 1507.

[11] Hefle, S. L., Nordlee, J. A., Taylor, S. L., Allergenic foods, *Crit. Rev. Food Sci. Nutri.* 1996, *36*, S69 – S89.

[12] Moneret-Vautrin, D. A., Kanny, G., Beaudouin, E., L'allergie alimentaire au mais existe-t-elle?, *Allerg. Immunol.* 1998, *30*, 230.

[13] Stadler, M. B., Stadler, B. M., Allergenicity prediction by protein sequence, *FASEB J.* 2003, *17*, 114 – 1143.

[14] Codex Alimentarius Commission, Alinorm 03/34: Joint FAO/WHO Food Standard Programme, Codex Alimentarius Commission, Twenty-Fifth Session, Rome, Italy, June 30 – July 5, Appendices III and IV, 2003, pp. 47 – 60.

[15] Breiteneder, H., Ebner, C., Molecular and biochemical classification of plant-derived food allergens, C., *J. Allergy Clin. Immunol.* 2000, *106*, 27 – 36.

[16] Mills, E. N. C., Madsen, C., Shewry, P. R., Wichers, H. J., Food allergens of plant origin-their molecular and evolutionary relationships, *Trends Food Sci. Technol.* 2003, *14*, 145 – 156.

[17] Mills, E. N. C., Jenkins, J. A., Alcocer, M. J. C., Shewry, P. R., Structural, biological, and evolutionary relationships of plant food allergens sensitizing via the gastrointestinal tract, *Crit. Rev. Food Sci. Nutri.* 2004, *44*, 379 – 407.

[18] Jenkins, J. A., Griffiths-Jones, S., Shewry, P. R., Breiteneder, H., Mills, E. N. C., Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: An analysis, *J. Allergy Clin. Immunol.* 2005, *115*, 163 – 170.

[19] Bredehorst, R., David, K. J., What establishes a protein as an allergen?, *Chromatogr. B. Biomed. Sci. Appl.* 2001, *756*, 33 – 40.

[20] Breiteneder, H., Mills, C., Structural bioinformatic approaches to understand cross-reactivity, *Mol. Nutr. Food Res.* 2006, *50*, 628 – 632.

[21] Hileman, R. E., Silvanovich, A., Goodman, R. E., Rice, E. A. *et al.*, Bioinformatic Methods for Allergenicity Assessment Using a Comprehensive Allergen Database, *Int. Arch. Allergy Immunol.* 2002, *128*, 280 – 291.

[22] Pasini, G., Simonato, B., Curioni, A., Vincenzi, S., *et al.*, IgE-mediated allergy to corn: a 50 kDa protein, belonging to the Reduced Soluble Proteins, is a major allergen, *Allergy* 2002, *57*, 98 – 106.

[23] Pastorello, E. A., Pompei, C., Pravettoni, V., Farioli, L., *et al.*, Lipid-transfer protein is the major maize allergen maintaining IgE-binding activity after cooking at 100 degrees C, as demonstrated in anaphylactic patients and patients with positive double-blind, placebo-controlled food challenge results, *J. Allergy Clin. Immunol.* 2003, *112*, 775 – 783.

[24] Scheurer, S., Son, D. Y., Boehm, M., Karamloo, F., *et al.*, Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen, *Mol. Immunol.* 1999, *36*, 155 – 167.

