



Further evaluation of the utility of “Sliding Window” FASTA in predicting cross-reactivity with allergenic proteins

Robert F. Cressman*, Gregory Ladics

Pioneer, A DuPont Company, Crop Genetics Regulatory Science and Registration, DuPont Experimental Station E353/317A, PO Box 80353, Wilmington, DE 19880-0353, USA

ARTICLE INFO

Article history:

Received 26 August 2008

Available online 11 December 2008

Keywords:

Protein allergenicity

Risk assessment

FASTA

Regulatory

Bioinformatics

Local similarity

ABSTRACT

FAO/WHO has recommended that IgE cross-reactivity between a transgenic protein and allergen be considered when there is greater than 35% identity over a sliding “window” of 80 amino acids. In a previous work, we evaluated the false positive and negative rates observed using the FAO/WHO criteria versus conventional, whole protein FASTA analyses [Ladics, G.S., Bannon, G.A., Silvanovich, A., Cressman, R.F., 2007. Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens. *Mol. Nutr. Food Res.* 51 (8), 985–998]. A number of protein sequence datasets were used as queries against the FARRP 7 allergen database. Results indicated that conventional FASTA analysis produced fewer false positives than the “sliding window” search proposed by FAO/WHO. Further, both methods were able to identify the potential for cross-reactivity between the Bet v 1 family of proteins, indicating that the conventional FASTA search possessed sufficient sensitivity. Recently, collections of protein sequences from multiple crop species (corn, soy, barley, lettuce, sugar beets, and spinach) were subjected to the same screen against the FARRP7 allergen dataset. In all cases, the conventional FASTA search returned fewer above threshold matches than the sliding window search. Examination of the matches not recognized by the conventional search revealed two scenarios: (1) “true” false positives consisting of low statistical significance (as measured by *E* score, i.e., a measure of the potential random occurrence of aligned sequences used to evaluate the significance of an observed alignment) alignments not contained in the conventional FASTA outputs, and (2) above-threshold sliding window alignments that fell below the 35% identity threshold with the conventional FASTA analysis. Although some alignments within this second group were between regions of low sequence complexity, where there was little/no change in *E* score, the majority of the alignments displayed more significance (lower *E* scores) under the conventional FASTA algorithm, yet did not meet the threshold defined by FAO/WHO. These data question the utility of the FAO/WHO recommended sliding window FASTA compared to the traditional whole sequence FASTA analysis coupled with appropriate statistical analysis.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

In silico assessment of novel proteins for identity to known allergens has come to represent a crucial step in the weight of evidence approach used to ascertain the safety of expressed proteins in genetically modified plant products. Beginning with the *Codex Alimentarius Commission's* recommendations in 2003, results of such assessments that demonstrate no above threshold similarities to known or putative allergens can eliminate the need to perform IgE-binding assays using allergic patient sera. Although IgE-binding is viewed as undeniable empirical evidence of a novel protein's potential for cross-reactivity, the enormous variation in allergenic responses even to well characterized allergens, the lack appropri-

ate characterization of sera, and frequent low background levels of IgE-binding to non-allergens all can contribute to erroneous results unless adequate measures are taken (Goodman, 2008).

Current bioinformatic analyses involve two recommended criteria; a search for continuous, identical stretches of amino acids 8 residues or greater in length, and an identity search using the FASTA local alignment algorithm (Pearson and Lipman 1988) to search for alignments of 80 residues or longer possessing a sequence identity of greater than 35% (FAO/WHO 2001; Codex 2003).

The value of short matching residue searches, originally intended to identify ‘theoretical’ shared epitopes between a query sequence and one or more allergenic proteins, has recently been called into question. The original FAO/WHO criteria, specifying a 6 residue continuous identity, has been largely discredited as producing too many false positives to be meaningful (Hileman et al., 2002; Stadler and Stadler, 2003; Thomas et al., 2005; Ladics

* Corresponding author. Fax: +1 302 695 9149.

E-mail address: robert.f.cressman@pioneer.com (R.F. Cressman).

et al., 2006; Silvanovich et al., 2006; Goodman, 2008). Silvanovich et al. (2006), argued that, when using these short matching residue searches “matches with bona fide IgE-binding motifs are indistinguishable from false positive matches”. Additionally it is stated that continuous residue searches are “not designed to include a statistical analysis and do not assign a ranking to a match”. Because the number of annotated proteins sequences is increasing (from 1.9 million to over 5.5 million sequences in the NCBI Protein dataset from 2004 to 2007) as well as the number of IUIS named allergens (from 487 in the Allergome database in 2006 to 572 as of October 1st 2007, <http://www.allergome.org>), the number of false positives detected using the 8 residue criteria will likely increase as well. Recent work by Jenkins et al. (2005) has demonstrated that the majority of plant food allergens fall into a small number of functional groups, with members within a group displaying significant similarities over large regions of amino acid sequence. The lack of a built in assessment of significance, the increasing number of potential false positive matches, and the apparent functional/sequence relationship between allergens within a given family support reduced reliance upon the short matching component of the FAO/WHO allergenicity criteria.

The other criteria, involving the use of the FASTA local alignment algorithm (Pearson and Lipman, 1988) to identify regions between a query sequence and an allergen sequence displaying identity above 35% over an 80 or greater residue “threshold”, seems to be a more robust method. The algorithm in fact employs a word search as an initial step, identifying all matching words between two sequences (the default word size, or ktup, for protein sequences is 2). Matching words are joined, extended, and assigned a numerical “score” using both a scoring matrix to give appropriate weight to conserved regions, as well as pre-set penalties for the insertion/extension of gaps. The resulting scores generated from comparisons between the query and all dataset proteins are then used to establish a linear relationship between alignment score and protein length. The score of a particular alignment with respect to the calculated distribution of all scores gives an indication of whether the alignment is meaningful; this is reflected in the expect score (*E*, a measure of the potential random occurrence of aligned sequences used to evaluate the significance of an observed alignment) assigned to the alignment. The lower an *E* score for a particular alignment, the more likely the comparison between the proteins reflects a true structural similarity, although the value is influenced by protein length and database size (Baxeavanis and Ouellette, 1998). This internal estimate of alignment significance can be further enhanced by the use of the PRSS algorithm (included in the FASTA34 package) which compares specific results to those obtained with shuffled query protein sequences. These built in methods for checking the nature of alignment results make the FASTA package helpful in determining the significance of protein sequence matches.

When FAO/WHO published recommendations for the determination of potential cross-reactivity to known allergens, they also included a “standardized methodology” for performing the searches. The four step procedure includes a step where one is to “prepare a complete set of 80-amino acid length sequences derived from the expressed protein” (FAO/WHO, 2001) followed by subsequent comparison of each 80 residue protein to the allergen dataset. The potential for cross-reactivity was to be considered when there was “more than 35% identity in the amino acid sequence of the expressed protein (i.e. without the leader sequence, if any), using a window of 80 amino acids and a suitable gap penalty...” (FAO/WHO, 2001). As a result of this suggestion, scripts were written that generated all possible 80 amino acid fragments from a given query protein and submitted each one for FASTA analysis. Above threshold matches (>35% over 80 or greater amino acids) were automatically identified for closer scrutiny. This method,

along with the eight continuous amino acid search, became the *de facto* standard for performing *in silico* allergenicity assessments.

A number of issues were encountered when using this methodology. First and foremost, FASTA constructs the best local alignments between a query sequence and dataset proteins; the algorithm is designed to “scan” the entire length of both sequences and generate the most significant matches. By using only 80 residue “sub-peptides” with FASTA, the program is forced to treat each 80 residue peptide as an entire protein, resulting in “novel” statistical treatment and resulting *E* scores that do not always correlate with those obtained by whole protein FASTA analysis. This “alteration” resulted in above threshold alignments, which upon closer inspection, represented low significance matches not seen when the conventional FASTA analysis was used. Another troubling aspect was that producing an alignment greater than 80 residues is much more unlikely when using 80 residue peptides than when using the entire protein sequence. In addition to generating false positives, significant alignments could potentially be missed, because the length threshold wasn’t satisfied (i.e., a 99% identity between two protein sequences that spans 79 residues falls “below” the threshold, and would be considered “negative”).

Recently, we compared a number of protein datasets using both the sliding window and conventional FASTA searches (Ladics et al., 2007). An initial search using FGENESH-generated maize genomic ORFs resulted in an approximately 5-fold increase in the number of above threshold matches. Subsequent comparisons were performed on 1000 randomly selected protein sequences obtained from public datasets, then smaller datasets containing sequences specific to maize and maize seed. In every case the sliding window search generated more above threshold hits than did the conventional FASTA search. By removing all Bet v 1 cross reactive protein sequences from the allergen dataset, then spiking in single sequences, we were able to demonstrate that the conventional FASTA screen successfully identified the known Bet v 1 cross reactive proteins down to an identity of 38.1%. The relatively low degree of identity between members of this family of proteins (Scheurer et al., 1999) was considered the basis for the original FAO/WHO criteria

In this paper, the comparison of the sliding window and conventional FASTA searches have been further expanded to include protein datasets from soybean, barley, lettuce, spinach and sugar beet in an effort to understand possible effects related to the number of annotated allergens within a given plant species. In addition, a review of the alignment matches not present in the conventional search was undertaken to try to determine their significance based upon current criteria.

2. Materials and methods

2.1. Selection of protein sequences

Protein sequence files for each crop were retrieved from both the NCBI Protein and Uniprot datasets as FastA formatted records. Duplicate amino sequences were removed, as well as sequences less than 80 residues in length. The sequences were further curated to remove proteins characterized solely upon their similarity to other accessions (sequence records referring to proteins as hypothetical, unknown, putative, possible, or similar to other proteins). Because of the large number of provisional, computer annotated genome records characterized in the above manner, all RefSeq entries were removed. For the spinach, soy, and corn collections, 200 proteins were selected at random for comparison. For the lettuce dataset, only 184 proteins remained after curation; these were all incorporated into the test dataset. For the sugar beet dataset, 223 proteins remained after curation; all were used for the com-

parison. During the comparison, a barley protein was discovered that consisted of all undefined residues (X), and was removed.

2.2. Bioinformatic analysis

The FARRP7 Allergen Database (www.allergenonline.com; January, 2007) was employed for all comparisons, which were carried out on a SGI MIPS R14000 computer running the IRIX Version 64 software. Briefly the FASTA34 program (version 3.4t25) or a modified Perl script designed for sliding window searches (fastest34.pl) was run on all sequences in the query datasets. The sliding window script breaks a query protein sequence into all possible 80 residue sub-peptides, runs a FASTA search on each peptide, and returns alignments that equal or exceed the FAO/WHO (2001) threshold of 35% identity. For all FASTA 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.

3. Results

When the number of above threshold matches using either criteria were compared for each crop, in all cases there were more sliding window FASTA “positive” results than when using a conventional FASTA search on the entire protein (see Table 1). Although the percentage of both whole protein and sliding window “positives” appeared to be greater for crops considered “allergenic” (soy, corn, barley) than for putative “non-allergenic” crops (spinach, lettuce, and sugar beets), the difference could likely be due to the lack of annotated allergens within these crops. The percentage of sequences identified as above the threshold varied from 4.5% to 18.1% for the sliding window searches, and from 3% to 12.6% for the conventional searches. All of these percentages are higher than the expected percentage of real allergens (0.4%) based upon the SWISS-PROT index (Stadler and Stadler, 2003) and is likely due to the use of the very conservative identity threshold of 35%. Out of the total dataset of 1208 proteins, the sliding window search produced roughly one third more matches (134 vs. 102) than the whole protein conventional FASTA search. A total of 35 sequences reported as above threshold by the sliding window search were not returned by the conventional search; three above threshold sequences were returned by the conventional search that were not identified by the sliding window search. Alignments generated by both search types for these 38 accessions were retrieved and reviewed individually in order to try and understand the reason for their exclusion under one of the FASTA analyses.

The three above threshold matches that were returned exclusively from the whole protein FASTA search consisted of a barley cystatin (cysteine protease inhibitor), a 60S ribosomal protein from lettuce, and a DNA-binding protein from soybean. (Table 2) All three matches displayed between 35% and 36% identity in the con-

ventional search. The cystatin match appeared to be the most significant, with *E* scores of 1.1×10^{-6} and 1.6×10^{-8} for the conventional and sliding window searches, respectively. It seems obvious from the alignment that the 80 residue length of the sliding window peptide limited the alignment; using the whole protein query resulted in the algorithm aligning 17 extra residues, producing the above threshold match (Fig. 1). The situation is the same with the other two sequences; in both cases the alignments are extended above the 35% cutoff when a larger query sequence is used.

All but one of the sequences identified solely by the sliding window FASTA search displayed identity values below 40% (Table 3). Previous work (Ladics et al., 2007) had demonstrated the sensitivity of whole protein FASTA at identities as low as 38.1%, using the cross reacting allergen Bet v 1 and Dau c 1; many of these sequences are at or below this sensitivity. Additional sera screening studies utilizing other members of the Bet v 1 family of allergens could provide a basis for future bioinformatic comparison thresholds. Nine of the 35 amino acid sequences identified have no counterpart in the conventional FASTA search. Alignment identities within this group range from 35% to 36%, and the degree of significance as measured by *E* score, is very low (*E* scores ranging from 0.0065 to 5.8). None of the queries in this group appear to be functionally related to their allergen matches; for instance, a chloroplast encoded RNA polymerase beta' chain from lettuce produces an above threshold alignment to a legumin protein from cashew. The apparent absence of functional relation, in conjunction with the lack of whole protein alignments, low percent identities and high *E* scores, suggests that these nine represent true “false positives”. Two of the conventional alignments returned the same allergen match, but encompassed non-overlapping different regions of the allergen when compared to the sliding window search. These low significance matches (sliding window FASTA *E* scores are 2.6 and 3.6) are in essence false positives as well, because the above threshold alignments were not generated to the same region of the allergen. The remaining 24 alignments had corresponding matches encompassing the same region of allergen sequence returned from the conventional FASTA, but all of the conventional identities fell below the 35% identity threshold. Fourteen of these 24 sequences, with *E* scores ranging from 2.5×10^{-7} to 8.5, appear to represent alignments precipitated by regions of low sequence complexity, including runs of single amino acids (i.e. SSSSSS) as well as repetitive doublets, triplets and greater collections of amino acids, such as those belonging to the gliadins. The remaining 10 sliding FASTA only matches were all present in the whole protein FASTA outputs, but fell below the 35% threshold for identity due to extension of the alignment length (length of alignment ranged from 118 to 327 amino acids in the whole protein FASTA results, versus 80–85 amino acids for the sliding window results). Interestingly, all but one of the 10 sliding window positive alignments displayed greater significance (as indicated by a smaller *E* score) when using the whole protein search, despite not exceeding the established threshold identity.

Table 1
Summary of above threshold allergen matches by crop.

Crop	Proteins analyzed	Sliding window positives	Percent (%)	Conventional positives	Percent (%)
Spinach	200	9	4.5	6	3.0
Lettuce	184	12	6.5	8	4.3
Beets	223	25	11.2	23	10.3
Soy	200	24	12.0	20	10.0
Corn	200	28	14.0	20	10.0
Barley	199	36	18.1	25	12.6
Total	1208	134	11.1	102	8.4

Table 2

Above threshold matches unique to the conventional FASTA analysis.

gi Number	Description	Allergen match	Whole protein FASTA			Sliding window FASTA		
			Percent identity	Length	E score	Percent identity	Length	E score
109238657	cystatin Hv-CPI11 [Hordeum vulgare subsp. vulgare]	gi 40807635 gb AAR92223.1 phytocystatin	35.052	97	1.10E-06	36.765	68	1.60E-08
38503536	60S acidic ribosomal protein P3 [Lactuca sativa]	gi 1173074 sp P42038 RLA3_CLAHE 60S acidic ribosomal protein	35.577	104	0.019	42.86	21	6.2
1905785	G/HBF-1 [Glycine max]	gi 66849502 gb EAL89830.1 major allergen Asp F2	35.294	85	1.2	31.65	79	0.045

4. Discussion

A comparison of the 80 residue “sliding window” search (FAO/WHO2001; Codex, 2003) with the traditional FASTA search conducted on entire protein sequences (Pearson and Lipman, 1988) was performed using randomly selected collections of protein sequences from plant food crops containing well characterized allergens (soybean, barley, and corn), as well as those considered less allergenic (sugar beet, lettuce, and spinach). For every crop, the sliding window search produced more above threshold (greater than 35% over 80 residues) matches than did the whole sequence FASTA search. The percentage of sequences with above threshold matches to allergens ranged from 4.5–18.1% using the sliding window approach and 3.0–12.6% when using whole sequence FASTA,

reflecting large numbers of false positives when compared to the percentage of allergens in the SWISS-PROT dataset (Stadler and Stadler, 2003). Previous work (Ladics et al., 2007) had suggested that the discrepancies were due to an increased number of false positive results associated with the sliding window search. Of interest is the observation that all but one of the unique sliding window positive results listed in Table 3 produced a FASTA alignment below 40% identity. Given that the threshold of 35% identity specified by FAO/WHO in 2001 for sliding window searches was arrived at without any apparent empirical testing, the raising of the sliding window threshold to 40% would appear to eliminate most of the discrepancy between the two FASTA approaches. A significant number of the sliding window positive alignments (9/35) could in fact be attributed to false positives, as evidenced by the

Sliding Window Strategy

```
>>gi|40807635|gb|AAR92223.1| phytocystatin [Actinidia de (116 aa)
  initn: 146 initl: 82 opt: 147 Z-score: 231.3 bits: 47.6 E(): 5.6e-08
Smith-Waterman score: 147; 35.443% identity (69.620% similar) in 79 aa overlap (2-80:10-86)
```

	10	20	30	40	50
109238	TLFLAAIVIIHVHVVATPTTAIPGGWVN	KNIADPHIQELGKWAVLEHTQLGG			
		::: :	:	::: : :: :::	:: :
gi 408	MVPKPLSLLFL-LALSAAVVGGRKLVAAGGWRPIESLNSAEVQDVAQFAVSEHNKQA-				
	10	20	30	40	50
	60	70	80		
109238	NDGLRFVKVVSAAEQIVNGVNYQLVIDA				
		::: :	: :	::: ::	::
gi 408	NDELQYQSVVGRGTQVVAGTNYRLVIAAKDGAVVGNYEAVVWDKPMHFRNLTSFRKV				
	60	70	80	90	100 110

Whole Protein Strategy

```
>>gi|40807635|gb|AAR92223.1| phytocystatin [Actinidia de (116 aa)
  initn: 146 initl: 82 opt: 157 Z-score: 208.1 bits: 43.8 E(): 1.1e-06
Smith-Waterman score: 157; 35.052% identity (69.072% similar) in 97 aa overlap (5-100:10-102)
```

	10	20	30	40	50
109238	MRTTLFLAAIVIIHVHVVATPTTAIPGGWVN	KNIADPHIQELGKWAVLEHTQLGG			
		::: :	:	::: : :: :::	:: :
gi 408	MVPKPLSLLFL-LALSAAVVGGRKLVAAGGWRPIESLNSAEVQDVAQFAVSEHNKQA-				
	10	20	30	40	50
	60	70	80	90	100 110
109238	NDGLRFVKVVSAAEQIVNGVNYQLVIDALRLDGS	SHR-TYKAVLFEKDSSNAKTRKFISFT			
		::: :	: :	::: ::	::
gi 408	NDELQYQSVVGRGTQVVAGTNYRLVIAA--KDGAVVGNYEAVVWDKPMHFRNLTSFRKV				
	60	70	80	90	100 110

109238 PAN

Fig. 1. Comparison of alignment strategies using barley cystatin Hv-CPI11 (gi|109238657).

Table 3

Above threshold matches unique to the FASTA sliding window analysis.

gi Number	Description	Allergen match	Sliding window FASTA			Whole protein FASTA		
			Percent identity	Length	E score	Percent identity	Length	E score
1709619	PDI_MAIZE protein disulfide-isomerase precursor (PDI)	PDI_ALTAL Protein disulfide-isomerase (PDI) (Allergen Alt a 4)	59.036	83	8.10E–21	30.887	327	5.00E–22
154816295	Short-chain dehydrogenase/reductase protein [Lactuca sativa]	MTDH_CLAHE Probable NADP-dependent mannitol dehydrogenase (MtDH) (Mannitol 2-dehydrogenase [NADP+]) (Allergen Cla h 8)	35.366	82	0.016	30.682	264	3.0E–21
2815493	CBP1_HORVU serine carboxypeptidase 1 precursor	Serine carboxypeptidase II [Triticum aestivum]	37.5	80	4.10E–14	28.521	284	7.20E–19
84579420	Lactuca sativa short-chain dehydrogenase/reductase 2	MTDH_CLAHE Probable NADP-dependent mannitol dehydrogenase (MtDH) (Mannitol 2-dehydrogenase [NADP+]) (Allergen Cla h 8)	36.905	84	0.00082	30.534	262	1.0E–16
10835710	A Chain A, crystal structure of thioredoxin F from spinach chloroplast (short form)	Thioredoxin h1 protein [Zea mays]	36.25	80	2.10E–12	33.962	106	1.50E–12
47115349	putative xyloglucan endotransglucosylase/hydrolase [Lactuca sativa]	rAsp f 9 [Aspergillus fumigatus]	35.294	85	3.50E–08	30.435	161	1.90E–12
116054	Calcium-dependent protein kinase SK5 (CDPK)	Pollen allergen Jun o 4 [Juniperus oxycedrus]	36.25	80	6.30E–08	32.639	144	5.70E–11
27464177	Expansin [Glycine max]	MPAL1_LOLPR Pollen allergen Lol p 1 precursor (Lol p 1) (Allergen R7)	39.506	81	1.50E–07	29.876	241	7.50E–11
109238649	Cystatin Hv-CPI10 [Hordeum vulgare subsp. vulgare]	phytolectin [Actinidia deliciosa]	37.5	80	4.90E–11	34.746	118	7.60E–11
126034	LDHB_HORVU L-lactate dehydrogenase B (LDH-B)	AF084828_1 major allergenic protein Mal f4 [Malassezia furfur]	35	80	6.70E–07	25.806	310	2.80E–10
22179	22-kD Alpha-zein [Zea mays]	O22116_WHEAT LMM glutenin 3	35	80	0.002	30.705	241	2.50E–07
82408517	Pathogenesis-related protein PR-1 [Glycine max]	VA5_VESMA Venom allergen 5 (Antigen 5) (Ag5) (Allergen Vesp m 5)	35.366	82	6.50E–05	34.615	104	2.60E–06
14594821	Alanine and glutamic acid rich protein [Beta vulgaris]	Latex allergen	35.714	84	1.80E–05	33.987	153	5.80E–05
595681	2-oxoglutarate/malate translocator	Putative gamma-gliadin [Triticum aestivum]	35.366	82	0.0012	32	100	3.40E–04
45357045	Coatomer alpha subunit [Hordeum vulgare subsp. vulgare]	30-kDa Salivary gland allergen variant 2 [Aedes aegypti]	36.471	85	4.60E–05	34.314	102	9.50E–04
15866410	DWARF8 [Zea mays]	CRF1_ASPFU Probable glycosidase crf1 precursor (Crh-like protein 1) (Allergen Asp f 9)	35.366	82	0.0028	27.363	201	0.0038
4585621	HMG1/Y protein [Zea mays]	S38584 allergen Phl p Vb—common timothy	35	80	0.015	34.831	89	0.0092
38679405	Absciscic stress ripening-like protein [Glycine max]	30-kDa Salivary gland allergen 30k-3 [Aedes albopictus]	36.905	84	0.22	27.928	111	0.073
4240037	ZmGR2a [Zea mays]	Latex allergen	39.506	81	0.05	27.083	96	0.25
18182311	GT-2 factor [Glycine max]	ALL7_ASPFU Allergen Asp f 7 precursor	35	80	2.6	29.6	125	0.53
6017938	AF181456_1 dehydrin; DHN6 [Hordeum vulgare]	Cu/Zn super-oxide dismutase [Olea europaea]	36.145	83	3.6	31.469	143	0.62
17017263	Methionine synthase [Zea mays]	AF208981_1 paramyosin isoform [Anisakis simplex]	35	80	0.24	21.714	175	0.83
148473091	dof zinc finger protein 2 [Hordeum vulgare subsp. vulgare]	MPA53_PHAHQ Major pollen allergen Pha a 5.3 precursor (Pha a 5) (Clone 29)	35	80	0.049	24.057	212	0.9
122248967	CYF_LACSA Apocytochrome f precursor	Unnamed protein product [Glycine max]	35.294	85	0.24	34.483	87	1.1
20513874	AF427791_27 hv711N16.15 [Hordeum vulgare]	AF510854_1 allergen Arah3/Arah4 [Arachis hypogaea]	35.366	82	2.3	32.432	74	1.4
120680	G3PX_HORVU Glyceraldehyde-3-phosphate dehydrogenase, cytosolic	ABA-1 allergen [Ascaris lumbricoides]	35.632	87	4.2	31.915	94	8.5
39725278	salt tolerance protein 2 [Beta vulgaris]	GE-rich salivary protein 30k-4 [Aedes albopictus]	35.366	82	0.0065	Not Present		
1039332	Glutamyl tRNA reductase [Hordeum vulgare subsp. vulgare]	Mag44 [Dermatophagoides farinae]	35.294	85	0.69	Not Present		
2495156	HEM11_HORVU Glutamyl-tRNA reductase 1, chloroplast precursor (GluTR)	Mag44 [Dermatophagoides farinae]	35.294	85	0.69	Not Present		
1938256	Phytase [Zea mays]	Non-specific lipid transfer protein [Solanum lycopersicum]	35.802	81	1	Not Present		
24370984	Quinone-oxidoreductase homolog [Spinacia oleracea]	S32101 PHLP5A protein—common timothy (fragment)	35.714	84	1.4	Not Present		
85700398	RPOC1_LACSA DNA-directed RNA polymerase beta' chain (PEP)	AF453947_1 allergen Ana o 2 [Anacardium occidentale]	36.25	80	2.3	Not Present		
78354955	plastidial ADP-glucose transporter [Hordeum vulgare subsp. vulgare]	MPA92_POAPR Pollen allergen KBG 41 precursor (Pollen allergen Poa p 9) (Poa p IX)	35.294	85	4.9	Not Present		
12002139	AF112148_1 MADS box protein 1 [Zea mays]	TPM_LEPDS Tropomyosin (Allergen Lep d 10)	35	80	5.7	Not Present		
34922469	LOX21_HORVU Lipoygenase 2.1, chloroplast precursor	AF240006_1 7S globulin [Sesamum indicum]	35	80	5.8	Not Present		

lack of a corresponding alignment when using whole sequence FASTA along with higher (less significant) *E* scores. However, the majority of the sequences unique to either of the analysis strategies resulted from the extension of alignments by the whole sequence FASTA analysis that either dropped or raised the percent identity of the alignment with regard to the established threshold. For many of the alignments, the significance (as measured by *E* score) of the alignment was greater when using the whole sequence FASTA method; this suggests that the threshold itself, based upon arbitrarily defined values for identity and length may be less accurate than reliance upon the embedded statistical calculations within the FASTA algorithm. From Table 3, a single sliding window above threshold alignment whose corresponding whole protein FASTA score was below the 35% identity threshold was returned. Although this alignment, between an allergenic fungal disulfide isomerase (Alt a 4) and a maize homolog, displays only 30.9% identity when analyzed using the entire sequence, the *E* score for the whole sequence alignment is two orders of magnitude greater. This is undoubtedly due to the length of the alignment (327 aa); such long alignments often reflect conserved tertiary structure associated with cross-reactivity (Jenkins et al., 2005). The adoption of search criteria based upon *E* score rather than arbitrary identity and length thresholds could identify alignments that span long regions of amino acid sequence yet contain low overall amino acid identity. Using the current criteria, these associations are only identified if an 80 residue region of the sequence exceeds 35% identity. An *E* based threshold would confer a number of additional advantages to bioinformatic analysis for potential allergenicity, including allowing the presence of conserved amino acid residues between two protein sequences to be considered. The current threshold is based upon amino acid identity. Additionally, an *E* based threshold could also identify significant regions less than 80 residues in length (i.e., a 95% identity over 60 amino acids). The current FASTA criteria do not reveal these matches, as they do not exceed the 80 residue threshold; any shorter, non-contiguous alignments (conceivably, contiguous alignments of 8 or greater residues would be identified by a linear epitope search) are ignored. Finally, an *E* based threshold would readily allow for the identification of insignificant alignments, regardless of length or identity. As an example, by setting an *E* score threshold of 0.01, below which significant similarity between protein sequences is generally recognized (Pearson, 2000), all but one of the 9 sliding window only FASTA alignments identified as false positives above would be eliminated, while all 10 of the sliding window only FASTA alignments that fell below threshold in the whole protein FASTA search would be included as positives. Even if this cutoff were to be used to evaluate the sliding window results from this study, 21 of the 35 sliding window unique sequences would be eliminated; this represents only 1.6% of the total above threshold hits (134), yet 60% of the sliding window only positives. The whole sequence FASTA data screened using this threshold would result in the removal of 9/102 (0.88%), yet 2/3 (66%) of the whole protein only records.

In summary, this comparison of the sliding window FASTA search using query proteins against a dataset of known and putative allergenic proteins with the traditional, whole sequence FASTA

analysis revealed that, while the whole sequence FASTA search appeared to generate fewer, more significant alignments (as measured by *E* score), closer inspection of alignments unique to either methodology demonstrate that both sliding window and whole protein FASTA analysis are constrained by their reliance on rigid identity/length thresholds. By switching to a “significance” threshold measured by *E* score rather than predefined length and identity thresholds, the number of false positives may be reduced, while additional significant alignments may be uncovered. Data with additional proteins, however, are needed to confirm these preliminary data. Additionally, any criteria for the establishment of cross-reactivity between an allergen and a transgenic protein should ideally be verified by empirical testing using sera from allergic patients; such verification would give support to any model for evaluation, including the current criteria. Future studies will focus on further investigating the effects of adoption of *E* based criterion upon the results of such bioinformatic searches, including potential comparisons with motif and structural prediction algorithms.

References

- Baxeianis, A.D., Ouellette, B.F.F., 1998. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons, Inc., New York.
- Codex Alimentarius Commission, Alinorm 03/34: Joint FAO/WHO Food Standard Programme, Codex Alimentarius Commission, Twenty-Fifth Session, Rome, Italy, June 30–July 5, 2003. Appendix III, Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants, and Appendix IV, Annex of the assessment of possible allergenicity, pp. 47–60.
- FAO/WHO, 2001. Evaluation of allergenicity of genetically modified foods. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, January 22–25, 2001, Rome, Italy.
- Goodman, R.E., 2008. Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. Food and Chemical Toxicology 46, S24–S34.
- Hileman, R.E., Silvanovich, A., Goodman, R.E., Rice, E.A., Holleschak, G., Astwood, J.D., Hefle, S.L., 2002. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. Int. Arch. Allergy Immunol. 128, 280–291.
- Ladics, G.S., Bannon, G.A., Silvanovich, A., Cressman, R.F., 2007. Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens. Mol. Nutr. Food Res. 51 (8), 985–998.
- Ladics, G.S., Bardina, L., Cressman, R.F., Mattsson, J.L., Sampson, H.A., 2006. Lack of cross-reactivity between the *Bacillus thuringiensis* derived protein Cry 1F in maize grain and dust mite Der p7 protein with human sera positive for Der p7-IgE. Regul. Toxicol. Pharmacol. 44 (2), 136–143.
- Jenkins, J.A., Griffiths-Jones, S., Shewry, P.R., Breiteneder, H., Mills, E., 2005. Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: an *in silico* analysis. J. Allergy Clin. Immunol. 115, 163–170.
- Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological comparison. Proc. Natl. Acad. Sci. USA 85, 2440–2448.
- Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. 132, 185–219.
- Scheurer, S., Son, D.Y., Boehm, M., Karamloo, F., Franke, S., Hoffmann, A., Hausteil, D., Veiths, S., 1999. Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen. Mol. Immunol. 36, 155–167.
- Silvanovich, A., Nemeth, M.A., Song, P., Herman, R., Tagliani, L., Bannon, G.A., 2006. The value of short amino acid sequence matches for prediction of protein allergenicity. Toxicol. Sci. 90 (1), 252–258.
- Stadler, M.B., Stadler, B.M., 2003. Allergenicity prediction by protein sequence. FASEB J. 17, 1141–1143.
- Thomas, K., Bannon, G., Hefle, S., Herouet, C., Holsapple, M.P., Ladics, G.S., MacIntosh, S., Privalle, L., 2005. In: Silico Methods for Evaluating Human Allergenicity to Novel Proteins: International Bioinformatics Workshop Meeting Report—Mallorca, Spain, February 24, 25, 2005. Toxicol. Sci. 88 (2), 307–310.