



Regular Article

The use of *E*-scores to determine the quality of protein alignments

Andre Silvanovich *, Gary Bannon, Scott McClain

Monsanto Company, Regulatory Product Characterization Center, 800 North Lindbergh Blvd., St. Louis, MO 63167, USA

ARTICLE INFO

Article history:

Received 28 August 2008

Available online 24 February 2009

Keywords:

GMO

Allergy assessment

FASTA

ABSTRACT

In 2001, the FAO/WHO suggested a procedure for performing FASTA or BLAST searches, and a threshold of greater than 35% identity in 80 or greater amino acids to identify potential allergenic cross-reactivity of transgene encoded proteins in genetically enhanced crops. Transgene encoded proteins meeting or exceeding this threshold would require additional *in vitro* evaluation for allergy safety. In work described herein, a method to calculate an *E*-score threshold is proposed for utilizing the full capability of bioinformatics to accurately identify potential cross-reactive allergens. The threshold *E*-score, $3.9\text{E}-07$, was produced using a test dataset of 7695 corn protein sequences and a method that entailed FASTA searches of the FARRP 7 allergen database with each of the dataset sequences using a conventional full length and an 80 amino acid sliding window FASTA comparison followed by an evaluation of *E*-score distribution. The results show that this *E*-score threshold identifies known corn allergens and it displays a false positive rate for known allergens that is comparable to that obtained from the 2001 FAO/WHO guidance. Furthermore, the *E*-score threshold is of sufficient stringency that it rejects the majority of false positive, composition-based anomalies and is 100% effective at identifying Bet v 1 cross-reactive allergens.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

The comparison of novel proteins with known allergens using bioinformatic methods is an important component of the weight of evidence approach used to assess the safety of transgene encoded proteins contained in genetically enhanced crops. In January 2001, the FAO/WHO suggested a sequence similarity evaluation procedure that included the preparation of “a complete set of 80 amino acid length sequences” derived from the transgene encoded protein and the comparison of “each of the sequences” to an allergen database using FASTA. Any protein meeting the minimum threshold of greater than 35% identity over any 80 or greater amino acid length would have been labeled as having potential IgE cross-reactivity and require additional *in vitro* IgE-binding assessments to show lack of allergenic potential. In the bioinformatics analysis described herein, a method to calculate an expectation-score (*E*-score) threshold is proposed for determining a significant level of similarity with appropriate accuracy and sensitivity. The proposed use of *E*-scores maximizes the similarity alignment assessment of FASTA and improves on the accuracy of identifying homologies beyond the shared identity matching percentage criteria currently outlined in Codex Alimentarius (2003). Using a dataset of 7695 corn sequences, an *E*-score threshold $3.9\text{E}-07$ or lower was determined for conservatively identifying known allergens and other proteins with potential for allergenic cross-reactivity. This *E*-score threshold identifies all known corn allergens and displays a false

positive identification rate for known allergens comparable to that obtained from the 2001 FAO/WHO guidelines. Moreover, the threshold derived from this method is of sufficient stringency (i.e., sensitivity) that it rejects composition-based anomalies that erroneously lead to identifying sequences as homologous and is 100% effective at identifying a well described group of cross-reactive allergens, the homologous Bet v 1 proteins (Ebner et al., 1995).

In work described by Ladics et al. (2007) the reliability of performing a bioinformatic search using FASTA and the limiters of greater than 35% identity threshold over an 80 amino acid search segment was evaluated. The work described the difference between this 80 amino acid sliding window segment approach and a conventional or “global” full-length sequence search. The authors used a variety of query sequences including; (1) *bona fide* corn sequences, (2) hypothetical corn protein sequences, (3) randomly selected protein sequences from GenBank and (4) chimeric protein sequences assembled with segments of a known allergen placed into the sequence of non-allergens. The authors concluded 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 non-allergen proteins to those of known allergens.” The authors further stated that “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.”

While Ladics et al. (2007) suggest utilizing the entire protein sequence as a query for conducting the FASTA search, a question that remained unanswered was if a threshold criteria of greater than

* Corresponding author. Fax: +1 314 694 6931.

E-mail address: andre.silvanovich@monsanto.com (A. Silvanovich).

35% identity in a window of 80 amino acids or greater was sufficient or whether some other criterion, such as an *E*-score or a combination of both, is the most appropriate indicator of the potential for allergen cross-reactivity. The greater than 35% identity in a window of 80 amino acids or greater affords a relatively simple interpretation of alignment data with the only “gray areas” being whether alignments in segments containing gaps truly count towards the 80 amino acid window or if an alignment exceeding the threshold, but having an *E*-score of 1 or greater is meaningful. In contrast, without a predefined threshold, an interpretation of alignment data based upon *E*-scores would appear to be a more daunting task. Through the analysis of 7695 protein sequences derived from corn, a method to determine an *E*-score based threshold is proposed. Using the described method, alignments obtained using a full-length query sequence and default FASTA parameters that display an *E*-score of 3.9E–07 or lower are more likely to represent potential allergen cross-reactivity than alignments that display a higher *E*-score. The method yields an *E*-score threshold that is as conservative as a 35% identity in 80 amino acid window threshold in identifying known corn allergens and it displays a comparable predicted false positive identification rate of known allergens of ~95%. Moreover, the threshold derived from this method is of sufficient stringency that it rejects glutamine composition-based false positives and is 100% effective at identifying Bet v 1 cross-reactive allergens. Both of these characteristics are an indication that improving the precision and accuracy of identifying potentially cross-reactive protein allergens is possible using FASTA, a current standard in alignment algorithms.

2. Methods

In order to establish a method to determine an *E*-score-based threshold for sequences of any amino acid length beyond 79, a set of 11,482 protein sequences was downloaded from GenBank June 6, 2007 using the keywords “*Zea mays*”. The 11,482 sequences were then screened to remove sequences of fewer than 80 amino acids, and to remove sequences that contain the *Zea mays* in their annotation but are not actually derived from *Zea mays*. A total of 7695 sequences of 80 amino acids or greater whose source organism was “*Zea mays*” were then used as queries for a FASTA search of the Version 7 FARRP (FARRP, 2007) allergen database. FASTA searches were performed using a LINUX SUSE 10.1 operating system running FASTA Version 3.4t26. Default FASTA parameters were employed for all FASTA searches with the exception that either the top 40 alignments or all alignments yielding an *E*-score of less than 10, whichever is fewer, were recovered and analyzed. FASTA data for each query was evaluated using a LINUX shell script that identified alignments showing 35% or greater identity over 80 or more amino acids among any of the top 20 alignments, and that tabulated the *E*-score for the top alignment recovered with each query.

The relationship between *E*-score and amino acid composition was evaluated using 1000 random sequences per parent query. Random sequences were generated using the SHUFFLE algorithm contained in Version 10.3RDB-UNIX of the Wisconsin Package run with a Solaris 9 operating system. Shuffled sequences were exported to a LINUX environment and the FASTA analysis was performed as described above for the 7695 corn protein sequences derived from GenBank.

Each of the 7695 query sequences was also used to perform a sliding 80 amino acid window search. Sliding window searches were performed using a LINUX shell script that interfaced with FASTA Version 3.4t26 (Reese and Pearson, 2002). The top 40 alignments, regardless of *E*-score for each window were inspected to determine if any displayed greater than 35% identity in 80 amino acids. Results for each search were tabulated and the total number

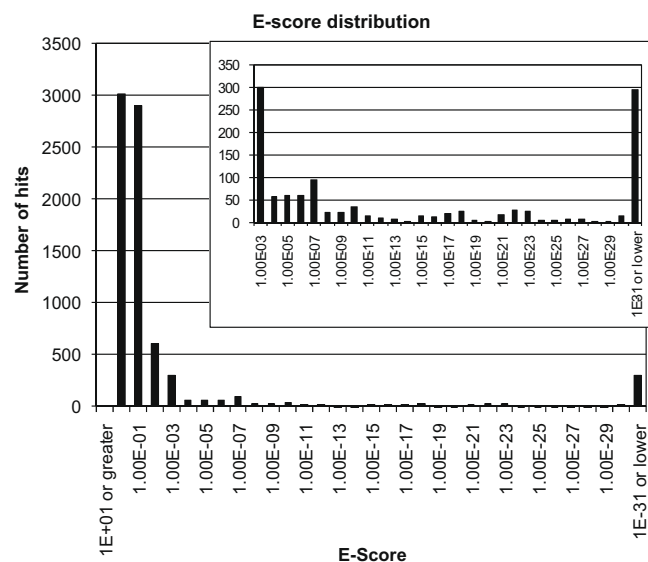


Fig. 1. 7695 corn protein sequences obtained from GenBank were used as queries for FASTA searches of the allergen database. The *E*-score of the top alignment obtained for each query was tabulated. The table of *E*-scores was sorted, and the number of *E*-scores per interval (i.e. 1.00 *E*-3, refer to *E*-scores of 9.9 *E*-3 to 1.00 *E*-3) was plotted. The inset graph shows *E*-scores of 1.00 *E*-3 or lower on an expanded Y-axis.

of queries yielding greater than 35% identity in 80 amino acids was determined.

To establish the premise that cross-reactive allergens can be identified using *E*-scores, the allergen database was tested with 151 Bet v 1 super family member proteins of 120 amino acids or greater in length that were extracted from the FARRP allergen database. Each extracted sequence was used as a query for a FASTA search of the FARRP allergen database. The top 160 alignments for each search were recovered and the 150th to 153rd best alignments for each query were inspected in order to identify the least significant (numerically greatest) *E*-score resulting from the alignment of two full-length Bet v 1 proteins.

3. Results

Sequences from *Zea mays* were chosen due to the availability of a large protein sequence dataset, its relevance to human health due to consumption of corn grain and corn grain containing products, and the low incidence of human corn allergy. Version 7 of the FARRP allergen database lists only 12 protein allergens that are derived from *Zea mays*. Two of the proteins are food allergens, profilin and non-specific lipid-transfer protein, while the remaining 10 are classified as aero allergens.

The *E*-score is a measure of the probability that a particular alignment was the product of chance versus an alignment that truly represents conserved structure, function and supports a conclusion the aligned proteins are homologous. The FASTA algorithm reports the degree of similarity for alignments on the basis of *E*-scores, similarity being a measure of the exact matches between amino acids as well as matches between biochemically similar amino acids, and includes other statistical measure of the alignment quality (Pearson and Lipman, 1988). For each query, alignments are reported in order from best to worst, with the best or top alignment having the lowest numerical *E*-score value while the least similar alignments have numerically greater *E*-score values. When the top alignment for each of the 7695 query sequences was identified, and the *E*-scores were tabulated and sorted on the basis of numerical value, a substantial number of the top align-

ments, 3418 (39.1%), yielded an *E*-score of 1 or greater (see Fig. 1) a measure of no significant similarity (Pearson, 2000). By definition, an *E*-score of 1 or greater indicates that an alignment is as likely to be the product of chance as it is a reflection of conserved structure or function (Pearson, 2000). This conclusion can be readily verified by randomizing the order of the amino acids in a query sequence (without changing the composition) and using this “shuffled” sequence as the query for the FASTA search. As shown later, shuffled sequences produced using a sequence yielding an *E*-score of 1 will also yield a shuffled sequence *E*-scores of ~ 1 in a FASTA search, an observance of no contribution of unique amino acid sequence to the *E*-score. If one were to then arbitrarily choose an *E*-score of 0.99 or less as a threshold for potential cross-reactivity, 4638 of the 7695 query sequences would be identified as potential allergens or cross-reactive allergens. This would be a gross overestimate since the FARRP allergen database contains 1251 proteins sequences and of those sequences, 12 are from corn (FARRP, 2007).

To begin establishing a meaningful threshold for identifying potential cross-reactivity, the method must first relate the *E*-scores of known allergens to the *E*-scores of all sequences in the test dataset. Due to the redundant nature of the 7695 sequences in the query dataset, 34 queries (0.44%) yielded matches of 100% identity with the allergen database. These 34 queries are known allergens and any *E*-score threshold intended to identify cross-reactivity must by definition be able to identify all currently known allergens. In the corn dataset, the 34 known allergen sequences do not yield the 34 most significant *E*-scores; therefore it is likely that some

non-allergen and non-cross-reactive proteins will yield more significant *E*-scores, thereby resulting in false positives. If as described earlier an arbitrary *E*-score of 0.99 or less was selected, then 4638 sequences would be identified and would yield a false positive rate as great as $((4638-34)/4638) \times 100\% = 99.27\%$ for the identification of known allergens. This false positive rate assumes no other protein among the remaining 4604 (4638–34) was ever identified as being an allergen or was a cross-reactive allergen. If the *E*-score for the least significant alignment of a known allergen were employed as a threshold, $2.4E-31$, a total of 295 sequences would be identified in the dataset and result in a false positive rate as great as $((295-34)/295) \times 100 = 88.8\%$. This observation, that *E*-scores for likely non-allergens and potential cross-reactive allergens may be numerically lower (more significant) than for known allergens, is not unexpected since the *E*-score is influenced by the size of the database being searched, the length and amino acid composition of the aligned sequence and the identity of the aligned residues. Moreover, this observation reveals that some level of false positive identification is a characteristic of a FASTA-based method.

When selecting a threshold it is important to keep in mind that a small change in the false positive rate will have a significant impact on the total number of sequences that would need to be evaluated for potential allergenicity or cross-reactivity. As can be seen in Table 1, the number of proteins associated with a potential false positive rate of 99.29% (*E*-score 0.99 or less) is 4683, whereas with a potential false positive rate of 98.15% (*E*-score 0.099) the number of proteins is reduced to 1786. Since bioinformatic comparisons are not predictive unless displaying 100% shared identity along the entire length of the query and aligned allergen, they are best used as an indicator and guide for further biologically-based studies. However, false positives rates are useful benchmark for determining cut-off or threshold criteria.

When alignment data for the 7695 queries is examined using greater than 35% or identity over 80 amino acids or greater threshold using a full-length query sequence, 654 queries or 8.5% of the dataset were found to meet or exceed the threshold. Inspection of the tabulated and sorted *E*-score values revealed that the 654th most significant alignment yielded an *E*-score of $3.9E-07$. Of the 654 queries that met or exceeded the greater than 35% identity over 80 amino acids or greater threshold, 562 queries (including the 34 known allergens) or 86% also yielded an *E*-score of $3.9E-07$ or less. Not unexpectedly, this indicates that a 35% identity over 80 amino acids was 100% predictive for identifying known allergens but also yields a false positive rate as great as 94.8%, an indication of the conservative approach in using this guideline. Among the 92 alignments that yielded an *E*-score of $3.9E-07$ or less but did not meet the greater than 35% identity over 80 amino

Table 1
Threshold modeling estimates determined using the 7695 query sequences.

<i>E</i> -Score threshold	Number of alignments	Known allergens not identified	False positive identification rate of known allergens (%)
0.99	4683	0	99.29
0.099	1786	0	98.15
0.0099	1176	0	97.19
0.0043 ^a	1073	0	96.92
0.00099	876	0	96.23
9.90E–05	818	0	95.97
9.90E–06	757	0	95.64
9.90E–07	696	0	95.26
4.20E–07	660	0	95.00
3.90E–07 ^b	654	0	94.95
5.00E–11	519	0	93.64
2.4E–31	295	0	88.81
1.5E–55	165	19	80.00

^a This value is based upon the 35% over 80 amino acid window threshold using a sliding window yielding 1073 matches with the query dataset.

^b This value is based upon the 35% over 80 amino acid window threshold using a full-length query yielding 654 matches with the query dataset.

Table 2
Homologous sequences identified by *E*-score threshold of $3.9E-07$ but not a sliding 80 amino acid window FASTA search.

Query sequence	Alignment parameters ^a		Aligned FARRP AD7 sequence
	<i>E</i> -Score	% Identity/overlap	
Protein disulfide isomerase GI-59861283	1.30E–16	27.19/331	Protein disulfide isomerase GI-85701160
Alpha-expansin 2 GI-14193753	9.70E–15	32.03/281	Beta-expansin GI-8118421
Exoglucanase precursor GI-8809764	1.40E–14	25.31/561	Xylosidase GI-2181180
Cytoplasmic malate dehydrogenase GI-2286153	5.10E–12	25.77/326	Mitochondrial malate dehydrogenase homolog GI-4587985
Xyloglucan endotransglycosylase GI-1097378	1.70E–10	25.26/190	GH16_GPI-glucanotransferase GI-2879890
Xyloglucan endo-transglycosylase/hydrolase GI-57753593	1.00E–09	28.17/213	GH16_GPI-glucanotransferase GI-2879890
Lactate dehydrogenase GI-22353	1.10E–08	25.08/303	Mitochondrial malate dehydrogenase homolog GI-4587985
Acidic ribosomal protein GI-2431767	1.50E–08	74.47/47	60S acidic ribosomal protein P2 GI-19879657
Protein disulfide isomerase GI-59861277	2.60E–07	29.41/85	Thioredoxin-like protein gi-76666769
Putative aldehyde dehydrogenase GI-18449341	2.70E–07	25.29/340	Aldehyde dehydrogenase GI-76666769
FAD binding 4 FAD/FMN-containing dehydrogenases GI-38520863	5.20E–07	22.76/268	FAD binding protein GI-55859456

^a The alignment parameters refer to the alignment obtained using a full-length query sequence.

acids were six alignments (four if duplicate sequences are removed) that yielded *E*-scores numerically lower than $2.4\text{E}-31$, the least significant score observed for an alignment of a known allergen displaying 100% identity. The four proteins not identified by the 35% over 80 amino acid alignment criteria but identified by an alignment displaying an *E*-score less than $2.4\text{E}-31$ were a 3-glucanase (GI-48374979) that aligned with a 1,3-glucanase ($1.5\text{E}-44$), glyceraldehyde-3-phosphate dehydrogenase (GI-474408) that aligned with aldehyde dehydrogenase ($9.0\text{E}-43$), a cysteine protease protein (GI-643597) that aligned with cysteine protease ($1.7\text{E}-33$) and a fatty acid aldehyde dehydrogenase (GI-38489178) that aligned with aldehyde dehydrogenase, ($1.4\text{E}-31$). This observation reveals that in some instances, homologous sequences, (i.e., those sequences that have conserved structure and function) are not identified by a greater than 35% over 80 amino acids threshold using a full-length query, an indication that although generally conservative, simply using shared-identity scoring can lack necessary accuracy and precision.

When the analysis was repeated using a sliding 80 amino acid window FASTA search, 1073 query sequences (13.5%) met or exceeded the greater than 35% identity over 80 amino acids or greater threshold. Inspection of the tabulated and sorted *E*-score values revealed that the 1073th most significant alignment yielded an *E*-score of 0.0043 ($4.3\text{E}-03$). Of the 1073 queries that met or exceeded the threshold, 854 queries (including the 34 known allergens) or 80% also yielded an *E*-score of 0.0043 or less. As with the full-length query, the 34 known allergen sequences in the dataset were identified by the search and the search yielded a false positive

rate as great as 96.92%. When compared to searches using a full-length query sequences, the sliding window search resulted in the identification of 419 additional sequences. Among those additional sequences identified were all alignments that yielded an *E*-score less than $1.3\text{E}-16$. However, as with the greater than 35% identity over 80 amino acids or greater threshold described earlier for full-length protein query sequences, a number of likely structural homologs remain unidentified using a sliding window FASTA search, and a substantial number of alignments, 64, yield *E*-scores of 0.5 ($5\text{E}-1$) or greater. Examples of sequences, both unidentified and identified, by the sliding window FASTA search are shown in Tables 2 and 3, respectively. Although there is good agreement between the greater than 35% identity over 80 amino acids or greater threshold and an *E*-score threshold that displays the same false positive rate for identification of known allergens, there is 15–20% disagreement between the two thresholds. The *E*-score threshold can clearly identify known structural homologs (Table 2), whereas the greater than 35% identity over 80 amino acid threshold identifies unrelated sequences (Table 3). Table 3 shows the relationship between an arbitrary *E*-score threshold, the false positive rate for identification of known allergens, and known allergens not identified in the query sequence dataset.

Inspection of the 274 alignments displaying *E*-scores having values of $9.9\text{E}-04$ to $1.1\text{E}-07$ revealed that ~59.9% of the alignments were the product of query sequences aligning with wheat gliadin and glutenin. In contrast, less than ~6.2% of alignments outside of this *E*-score range (*E*-scores having either numerically larger or smaller values) were the product of alignments between

Table 3

Non-homologous sequences identified by a sliding 80 amino acid window FASTA search.

Query sequence	Alignment parameters ^a		Aligned FARRP AD7 sequence
	<i>E</i> -Score	% Identity/overlap	
Isoamylase-type starch debranching enzyme GI-29126649	1.1	27.84/97	Alpha-amylase inhibitor GI-18955
Actin GI-1498390	1.2	24.68/235	HSP70 Heat shock protein GI-729764
Protein kinase GI-7766821	1.2	25.33/150	Allergen <i>Periplaneta americana</i> GI-1580792
NADH dehydrogenase subunit 6 GI-114151645	1.3	27.45/51	Venom allergen 5 GI-85681830
Glutathione S-transferase GI-11385503	1.8	32.69/52	Beta-lactoglobulin GI-162748
Actin GI-1498382	2.2	24.04/208	Putative luminal binding protein GI-10944737
Histone acetyltransferase GI-5579441	2.6	29.35/92	Major allergen Mal d GI-60280839
Protein kinase CK2 catalytic subunit GI-11527006	3.1	27.17/92	Alpha-s2-like casein precursor GI-162929
Gag and Pol GI-20502803	3.2	25.69/109	Paramyosin isoform GI-8453086
Sucrose phosphate synthase GI-168626	4.1	26.47/68	2s albumin GI-24473800
Bronze-2 protein GI-55741044	5.1	21.61/162	Ovalbumin GI-129293

^a The alignment parameters refer to the alignment obtained using a full-length query sequence.

Table 4The impact of high glutamine content and query sequence shuffling on *E*-scores.

Unshuffled query <i>E</i> -score	Likely non-allergen GI-16305323 1.0E+00	Potential Allergen GI-1330254 3.40E-07	Allergen GI-255569 4.10E-75	Allergen GI-898927277 3.40E-154	Glutamine-rich GI-16305142 3.70E-06
<i>Shuffled sequence E-scores</i>					
>10	3	0	2	8	1
>=1.0 <10.0	522	489	525	543	71
>=0.1 <1.0	407	423	411	384	351
>=0.01 <0.1	63	80	59	61	339
>=0.001 <0.01	4	5	3	9	162
>=0.0001 <0.001	1	3	0	1	57
>=1E-05 <0.0001	0	0	0	0	16
>=1E-06 <1E-05	0	0	0	0	2
>=1E-07 <1E-06	0	0	0	0	1
Number of unique database sequences yielding a best match	403	421	453	425	114 ^a
Number of 35% over 80 amino acid matches with shuffled queries	1	3	1	6	7

^a Of the 114 unique sequences identified by the shuffled queries, 26 of the unique proteins were described as being gliadins and glutenins. These 26 gliadins and glutenins accounted for 844 of the 1000 top alignments for the shuffled query.

query sequences and gliadins and glutenins. The reason for this disproportionate distribution of alignments with gliadins and glutenins is that these proteins have an unusual amino acid composition that is rich in glutamine when compared to other proteins. When glutamine-rich proteins are used to query the allergen database, they will invariably identify gliadins and glutenins, and yield alignments where the amino acid glutamine is the dominant identity. The resultant *E*-scores from such alignments appear to reflect potential structural similarity. However, this glutamine-based bias is easily confirmed as such by shuffling the query sequence, and using the shuffled sequence as a query. When shuffled, glutamine-rich sequences are used as queries in a FASTA search, they too will identify gliadins and yield *E*-scores comparable to the unshuffled query.

To demonstrate the relationship between the *E*-score of a query protein sequence versus shuffled query sequences, two known allergens, a sequence that yielded a top alignment with a glutamine-rich protein, and query proteins that yielded *E*-scores of 1.0 and $3.4\text{E}-07$ were each shuffled 1000 times and each shuffled sequence was used as a query for a FASTA search of the FARRP allergen database. Based upon *E*-score, it is highly likely that the sequence that yielded an *E*-score of 1 is a non-allergen, while the sequence that yielded an *E*-score of $3.4\text{E}-07$ is worthy of further analysis as it may constitute a potential allergen or cross-reacting protein. The results of the analysis are shown in Table 4.

The *E*-score distribution for the query sequences obtained by shuffling both of the known allergens, and the non-allergen and potential allergen were as expected. Of the 4000 shuffled sequences, all but five yielded *E*-scores of greater than $9.9\text{E}-04$. Of the five shuffled sequences whose *E*-score was below $9.9\text{E}-04$, the shuffled sequences were derived from an allergen, a potential allergen and a likely non-allergen. The distribution of *E*-scores derived from the shuffled sequences of known allergens, the non-allergen and predicted allergen were indistinguishable. Consistent with the random nature of the query sequences, each group of 1000 shuffled sequences identified an average of approximately 425 unique proteins from the allergen database as a top match. In contrast, shuffled sequences derived from the glutamine-rich query yielded a bell curve like distribution of *E*-scores with 22.8% of the alignments having an *E*-score $< 9.9\text{E}-03$, and the majority of sequences having *E*-scores in the range of $9.9\text{E}-01$ to $1.0\text{E}-02$. Unlike the shuffled sequences from allergens, a potential allergen and a likely non-allergen, shuffled sequences from a glutamine-rich query identified fewer unique database sequences and 84.4% of the top alignments were with proteins described as being gliadins or glutenins. Both random glutamine-rich query sequences and *bona fide* sequences that are glutamine-rich but that may have a relationship to gliadins or glutenins yield *E*-scores in the $9.9\text{E}-00$ to $1\text{E}-07$ range. Any method to determine an *E*-score threshold should not yield a threshold that is subject to glutamine-based bias. Based upon the analysis described herein, the method should not yield a threshold value that is greater than $1.0\text{E}-06$.

The Bet v super family of allergens is currently the largest and most diverse group of identified or suspected cross-reacting allergens and includes representatives from a range of plants that includes birch, oak, hornbeam, hazel, apple, peanut, mung bean, pear, cherry, carrot and strawberry (FARRP, 2007). Although it is not known if all members of the Bet v family contained in the FARRP database are involved in human allergic disease, taken as a whole, the Bet v family affords the opportunity to estimate *E*-scores for homologous protein sequences associated with known biological cross-reactivity. The 151 members of the Bet v super family of allergens, each at least 120 amino acids in length, were used as full-length queries for the FASTA searches of the FARRP allergen database. For each query, the top 160 alignments were

recovered and the alignments were inspected to identify the *E*-score associated with the poorest alignment with a Bet v 1 homolog of 120 amino acids or greater in length. In all instances, using a full-length Bet v 1 super family member as a query for a FASTA search of the FARRP 7 allergen database yielded *E*-scores not greater than $5.0\text{E}-11$ when the alignment was with a full-length Bet v 1 protein family member sequence contained in the allergen database. The alignment yielding an *E*-score of $5.0\text{E}-11$ (39.655% identity, 75% similarity in a 116 amino acid overlap) was the product of a GI-18652047 (Dau c 1.0201) query sequence aligning with GI-4006947 (Bet v 1, isoform 14). When the alignment data for the reciprocal search using GI-4006947 as the query sequence were inspected, as expected, the alignment was identical but the observed *E*-score was $1.3\text{E}-13$.

4. Discussion

Through the analysis of *E*-score distributions produced by FASTA analysis of 7695 corn protein sequences and 151 Bet v super family members used to search the FARRP 7 allergen protein database, a continuum of *E*-score values were produced. This continuum spanned a range of values from 15 to $1.6\text{E}-193$ and reflected alignments no better than an alignment that would be produced using a random amino acid sequence to alignments that reflected the positive identification of a known allergen. Through the examination of *E*-score distribution, false positive rates for the identification of known allergens were calculated and compared to the false positive rate for the greater than 35% identity over 80 amino acid threshold. An *E*-score of $3.9\text{E}-7$ was found to display a false positive identification rate for identification of known allergens of 94.95%, the same rate that was observed with a 35% over 80 amino acid threshold using a full-length query. Whereas an *E*-score of $4.3\text{E}-3$ was found to display a false positive identification rate of known allergens of 96.92%, the same rate that was observed with a 35% over 80 amino acid threshold using a sliding window query. In order to establish context for the two *E*-scores ($3.9\text{E}-07$ and $4.3\text{E}-03$) several additional analyses were performed. By shuffling the sequences of two known allergens, a potential allergen (*E*-score $3.4\text{E}-07$) and a likely non-allergen (*E*-score $1.0\text{E}-00$) and using the shuffled sequences as queries for a FASTA search, 0.65% of the shuffled sequences yielded *E*-scores of $1.0\text{E}-02$ or less. When glutamine-rich shuffled sequences were used as queries, 23.8% of the queries yielded *E*-scores of $1.0\text{E}-02$ or less. Therefore, an *E*-score threshold that has a false positive identification rate of known allergens comparable to that of a sliding 80 amino acid window FASTA search, $4.3\text{E}-03$, would be unable to exclude glutamine composition bias, for random glutamine-rich sequences. In order to determine if $3.9\text{E}-07$ was a suitable threshold to identify cross-reacting allergens, the *E*-scores associated with alignments among Bet v 1 family members revealed that the least significant *E*-score among all possible alignments of full-length Bet v 1 super family members and the FARRP 7 allergen database was $5\text{E}-11$. Since alignments among Bet v 1 super family members were the basis for the 35% over 80 amino acid threshold and the *E*-scores associated with these alignments are lower than $3.9\text{E}-07$, the statistical measure of this threshold's accuracy and precision indicates it is sufficiently conservative to identify allergen sequence homology. The use of $3.9\text{E}-07$ as a threshold has a potential false positive rate as high as 95% for the identification of known allergens. Yet, it is of sufficient sensitivity that it will discriminate against the majority of alignments that are the product of glutamine-rich protein composition bias. The use of an *E*-score based threshold for a full-length query sequence by itself, or in combination with a sliding window search has the added advantage of being fully consistent with the recommended

method for the use of FASTA as described by the algorithm developers (Reese and Pearson, 2002).

References

- Codex Alimentarius Commission, 2003. Joint FAO/WHO Food Standard Programme: Appendix III, Guideline for the conduct of food safety assessment of foods derived from recombinant DNA plants and Appendix IV, Annex on the assessment of possible allergenicity, 25th Session, Rome, Italy, 30 June–5 July, 2003, pp. 47–60.
- Ebner, C., Hirschwehr, R., Bauer, L., Breiteneder, H., Valenta, R., Ebner, H., Kraft, D., Scheiner, O., 1995. Identification of allergens in fruits and vegetables: IgE cross-reactivities with the important birch pollen allergens Bet v 1 and Bet v 2 (birch profilin). *J. Allergy Clin. Immunol.* 95 (5), 962–969.
- 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, Rome, Italy, 2001.
- FARRP (Food Allergy Research and Resource Program Database), 2007. University of Nebraska, Lincoln, NE. Available from: www.allergenonline.com.
- 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, 985–998.
- Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence 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.
- Reese, J.T., Pearson, W.R., 2002. Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics* 18, 1500–1507.