Contents lists available at ScienceDirect

Toxicology Reports

journal homepage: www.elsevier.com/locate/toxrep

1:1 FASTA update: Using the power of *E*-values in FASTA to detect potential allergen cross-reactivity

Ping Song*, Rod Herman, Siva Kumpatla

Dow AgroSciences LLC, 9330 Zionsville Road, Indianapolis, IN 46268, USA

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form 11 August 2015
Accepted 11 August 2015
Available online 20 August 2015

Keywords:
Bioinformatics
Allergen
Transgenic crops

ABSTRACT

In the context of regulatory assessment of transgenic proteins for potential allergenicity, a previous investigation demonstrated that a 1:1 FASTA comparison using an *E*-value of 1.0E-09 as a criterion is superior to the conventional FASTA search (using the whole sequence as a query) for >35% identity over 80 amino acids, but with improved specificity. A further study, using groups of known cross-reactive peanut allergens, indicates the sensitivity of this approach is superior to the conventional FASTA search and equivalent to 80-mer sliding window FASTA search recommended by WHO/FAO. Specifically, the 1:1 FASTA approach eliminated the technical issues resulting from lack of identification of short query sequences with high identity to known allergens, or high identity over short amino acid stretches, and different *E*-value settings when searching for >35% identity over 80aa. Based on the performance of this simple application of existing bioinformatics tools, and its ease of implementation and interpretation in the context of a regulatory assessment, we advocate that adoption of this 1:1 FASTA approach as a supplement to the FAO/WHO/ CODEX criterion (>35% identity over 80aa) formulated 13 years ago. Adoption of this approach eliminates many biologically irrelevant homology hits generated by the FAO/WHO/CODEX criterion and improves the safety assessment of GM crops.

1. Introduction

1.1. Official guidance

The bioinformatics approaches for evaluating potential cross-reactivity between an introduced protein in a genetically modified (GM) food crop and known allergens were developed 13 years ago and documented in FAO/WHO/CODEX [1,2]. The intention was to have a conservative approach that is able to detect known cross-reactive allergens across even the most disparate amino acid sequences. The threshold for concern was set based on a diverse set of cross-reactive allergens in the birch-pollen/apple set of proteins [9]. The original FAO/WHO/CODEX method specified a threshold of >35% amino-acid identity over a stretch of 80 amino acids. While FAO/WHO guideline recommends that a FASTA or BLAST search be conducted in an 80-mer sliding-widow fashion, there is no specified approach in the CODEX guideline.

1.2. Method performance

While the FAO/WHO/CODEX criterion using the sliding-window approach indeed detected such cross reactive allergens, it also detected a high number of non-cross reactive proteins [4,6,7,12,13]. As such, a conventional FASTA (using the whole sequence as a query) with a criterion of >35% identity over 80 amino acids was proposed to eliminate the high false-positive rate of the sliding-window approach [4,13]. In retrospect, this high false-positive rate is not surprising because the sliding-window approach did not make full use of advanced local alignment algorithms (e.g., FASTA; [10]) available for differentiating true protein relationships from biologically insignificant relationships. The science of protein homology had recognized that percent identity alone was an inferior method of differentiating homology among proteins compared with modern bioinformatics tools [10,11].

1.3. Theoretical underpinning

For allergen cross-reactivity, similar epitopes (antibody binding sites) can enable the IgE antibodies that react with one allergen to cross react with another allergen. Protein function is also driven by local three-dimensional protein structures that interact with receptors or other compounds found in biological systems or the

* Corresponding author.
E-mail address: psong@dow.com (P. Song).

http://dx.doi.org/10.1016/j.toxrep.2015.08.005
2214-7500/

environment. It is both the detection of functional similarity and of evolutionary relationships that motivated the development of modern bioinformatics tools. Local alignment tools, like FASTA, make use of sophisticated bioinformatics algorithms to recognize local similarity among proteins. It would be surprising if such tools were not also superior to simple amino-acid identity percentage for differentiating similar IgE epitopes from those less similar [8].

1.4. Technical issues

In addition, the WHO/FAO/CODEX threshold of >35% identity over 80 amino acids, either achieved by sliding-window or conventional FASTA search, does not address the following technical issues: (1) lack of ability to detect a query sequence that has a much higher identity (e.g., 80% identity) over less than 80 amino acids, with a known allergen; (2) lack of ability to detect a query sequence that is less than 29 amino acids (29/80 = 35%), which is often the case when evaluating non-intended reading frames in a GM event; (3) the effect of variable *E*-value settings when running a local alignment search (e.g., FASTA or BLASTp) on the number of returned alignments containing >35% identity over 80 amino acids. In the latter case, for example, the default setting of *E*-value for a protein sequence similarity search in FASTA is 10, but the number of alignments (>35%/80aa+) can sometimes be higher for the same query protein if the *E*-value is set to 100.

1.5. Past findings and current objectives

With this in mind, previous work in our lab [14] compared a one-to-one (1:1) FASTA approach for detecting known cross-reactive allergens with the >35%/80aa+ criterion described in the FAO/WHO/CODEX guidelines [1,2]. The 1:1 FASTA combines all of the power of modern bioinformatics tools for recognizing local similarity with a novel approach of making each query against a database that contains only one allergen sequence at a time. The sequential isolation of each allergen sequence in the database stabilizes the statistical measure of sequence alignment (*E*-value) so that it does not change as the size of the allergen database changes. This stabilization of *E*-values allows for a threshold of concern to be identified that will not change purely as an artifact of the increased size of the database as newly characterized allergen sequences are added. It was previously found that the 1:1 FASTA approach was as sensitive as the conventional FASTA with a criterion of >35% identity over 80 amino acids at detecting true cross reactive allergens, while being far superior in eliminating false detections. In addition, the 1:1 FASTA approach allowed for detection of biologically meaningful homology over short stretches of amino acid sequence (<80 amino acids) [14]. Here, we further report on a sensitivity comparison of the 1:1 FASTA approach with the 80-mer sliding window search for >35% identity over 80 amino acid recommended by [2], and also the conventional FASTA search for >35% identity over 80 amino acids as proposed by others [4,13].

2. Materials and methods

2.1. Sequences of peanut allergens

Protein sequences of peanut allergens Ara h 1 (Accession: P43238), Ara h 2 (Accession: Q6PSU2), Ara h 3 (Accession: O82580; Q9SQH7), Ara h 6 (Accession: Q647G9), and Ara h 9 (Accession: B6CEX8; B6CG41) listed in a recent publication (Table 1, [3]) were downloaded from GenBank.

2.2. Sequences of cross-reactive allergens

As presented in Table 1, proteins sequences defined as vicilin, 7S globulin, legmin, 11S globulin, 2S albumin, and nsLTP (non-specific lipid transfer protein) were retrieved from the allergen online database (version 14) (http://www.allergenonline.org/). Sequences of vicilin and 7S globulin, legmin and 11S globulin, 2S albumin, and nsLTP were respectively grouped into four subsets of cross-reactive allergens.

2.3. FASTA searches

Each peanut allergen sequence was searched by FASTA against its corresponding cross-reactive allergen subset by three approaches: (1) 1:1 FASTA with a threshold *E*-value of 1.0E-09 as the criterion; (2) 80-mer sliding window search in which a peanut allergen sequence was parsed into sequentially overlapping 80-mers before FASTA search; and (3) the conventional FASTA described in the literature [4,13]. Both 80-mer sliding window and whole sequence FASTA used >35% identity over 80aa as the criterion. All the FASTA searches were conducted with default parameter setting (Matrix = BLOSUM50; Gap Penalties = −10/−2 ; ktup = 2; Expectation = 10) [10].

3. Results and discussion

3.1. Approach summary

In this investigation, peanut allergens, along with their known cross-reactive allergens, as recently reported (Table 1, [3]), were chosen to further evaluate the performance of the 1:1 FASTA approach. The sequence of each peanut allergen was queried against the corresponding cross-reactive allergens using the 1:1 FASTA with a threshold *E*-value of 1.0E-09 as the criterion, a whole sequence FASTA, or a 80-mer sliding window FASTA with >35% identity over ≥80 amino acids as the criterion. The intension of this additional work was to determine if the 1:1 FASTA approach is able to identify cross-reactive allergens detected by the FAO/WHO 80-mer sliding window search for >35%/80aa and the whole sequence FASTA search for >35%/80aa described in the literature [4,13] for this newly compiled set of cross-reactive allergen sequences [3].

3.2. Search results

When searching cross-reactive allergens of Ara h 6 (2S albumin; accession Q647G9) and Ara h 9 (nsLTP, non-specific lipid transfer protein; accession B6CG41), the performance (sensitivity and selectivity) of the 1:1 FASTA, 80-mer sliding window, or whole sequence conventional FASTA was the same (data not shown). However, the approach using whole sequence conventional FASTA coupled with the FAO/WHO/CODEX identity criterion (>35% over 80 or more amino acids) failed to detect some cross-reactive allergens of Ara h 1, Ara h 2, and Ara h 3, even though the *E*-values of those alignments were <1.0E-12 (Table 1). For these same comparisons, the performance of the 1:1 FASTA approach (*E*-Values of alignments <1.0E-17), was equivalent to the 80-mer sliding window search (Table 2), except for one search result for Ara h 2.

3.3. Ara h 2 discussion

The FAO/WHO 80-mer sliding window search of Ara h 2 (accession Q6PSU2) detected a >35% over 80 alignment with a 2S albumin from Brazil nut tree (*Bertholletia excelsa*) (accession CAA38363; 154aa). This entry is a 2S albumin large subunit homologue of Ber e 1 (2S albumin 1 small unit from *Bertholletia excels*; accession

Table 1
Cross-reactivity of peanut allergens [3].

Protein super family	Cupin		Prolamin			
Protein family	Vicilin or 7S globulin	Legumin or 11S globulin	2S albumin			nsLTP
Allergen	Ara h 1	Ara h 3	Ara h 2	Ara h 6	Ara h 7	Ara h 9
Isoallergen (UniProt accession)	Ara h 1.0101 (P43238)	Ara h 3.0101 (O82580) Ara h 3.0201 (Q9SQH7)	Ara h 2.0101 (Q6PSU2) Ara h 2.0201 (Q6PSU2)	Ara h 6 (Q647G9)	Ara h 7.0101 (Q9SQH1) Ara h 7.0201 (B4XID4)	Ara h 9.0101 (B6CEX8) Ara h 9.0201 (B6CG41)
Cross-reactivity	With other legume and tree nut vicilins and Ara h 2 and Ara h 3	With other legumes and tree nut legumins and Ara h 1, 2, and 6	With 2S albumins from almond and Brazil nut, and Ara h 1, 3 and 6	With Ara h 1–3	Not known	With peach and hazelnut nsLTP (Pru p 3 and Cor a 8)
Number of sequences in allergen online database V14	27	36	12	9		32

Table 2
Hits detected by 1:1 FASTA with *E*-value of $\leq$1.0E-09 as a threshold or 80-mer sliding window FASTA but not by conventional whole sequence FASTA with >35% identity over 80 amino acids or longer as threshold.[a]

Peanut allergen (UniPro accession)	1:1 FASTA		80-mer sliding window FASTA			Whole sequence FASTA		
	Hit accession[b] (GenBank accession)	*E*-Value	Hit accession (GenBank accession)	Identity (%)	Alignment Overlap	Identity (%)	Alignment overlap	*E*-Value[c]
Ara h 1 (P43238)	AAK15089.1	3.80E-025	AAK15089.1	37.5	80	31.7	619	1.2E-18
	AAM73729.1	6.20E-017	AAM73729.1	45.1	82	28.8	579	3.7E-12
	AAM73730.2	1.60E-018	AAM73730.2	45.1	82	28.7	579	4.0E-12
Ara h 1 (P43237)	AAK15089.1	5.50E-025	AAK15089.1	55.1	78	33.2	561	7.4E-18
	AAM73729.1	1.80E-019	AAM73729.1	45.1	82	28.9	599	2.0E-13
	AAM73730.2	1.60E-019	AAM73730.2	37.5	80	28.7	599	2.0E-18
	AAF18269.1	1.10E-017	AAF18269.1	45.1	82	35.0	625	2.2E-13
Ara h 3 (ADQ53859)	O23878.1	2.50E-020	O23878.1	43.6	78	33.4	575	9.8E-19
	Q9XFM4.1	9.10E-021	Q9XFM4.1	43.6	78	34.3	545	4.3E-19
Ara h 2 (Q6PSU2)	CAA38363	1.80E-05	CAA38363	35.8	81	27.9	179	9.0E-05

[a] FASTA (v35.04) searches were conducted using default setting (Matrix = BLOSUM50; Gap Penalties = −10/−2; ktup = 2; *E*-value = 10).

[b] AAK15089.1—7S globulin from *Sesamum indicum* (sesame); AAM73729.1 and AAM73730.2—vilcilin-like protein from *Anacardium occidentale* (cashew); AAF18269.1—vilcilin-like protein precursor from Juglans regia; O23878.1—13S globulin seed storage protein (legumin-like protein) from *Fagopyrum esculentum* (common buckwheat); Q9XFM4.1—13S globulin seed storage protein (legumin-like protein) from *Fagopyrum esculentum* (common buckwheat); CAA38363—2S albumin from *Bertholletia* excels Brazil nut).

[c] *E*-Values from whole sequence FASTA search of Ara h1, Ara h 2, and Ara h 3 were generated based on a database with 27, 12, and 36 sequence entries, respectively. The *E*-values of those alignments generated by whole sequence conventional FASTA search are significant except the one between Ara h 2 and the 2S albumin large subunit (accession CAA38363), but they won't be classified as hits according to the criterion of >35% identity over 80aa or longer.

P04403; 146aa) with an allergen status of "unassigned" in the Version 14 Allergen Online database. The sequence identity between this 2S albumin (accession CAA38363) and the authentic allergen Ber e 1 (2S albumin from Brazil nut; accession P04403) is ∼70% overall, but their cross reactivity with Ara h 2 has not yet been clearly demonstrated, although allergenic cross-reactivity between Ara h 2 and Brazil nut proteins exists [5]. Interestingly, none of the three approaches detected hits between Ara h 2 and the authentic Ber e 1 (accession P04403), a well defined 2S albumin allergen from Brazil nut.

3.4. Summary

Several studies indicated that the 80-mer sliding window search for >35% identity over 80 amino acids is extremely conservative with high sensitivity, but low specificity (high false positive rate) [4,6,7,12,13]. However, more recent work calls into question the ability of the FAO/WHO/CODEX criterion (>35% identity over 80 amino acids) to detect sequences with very high similarity but low identity [8]. The present work further indicates that the sensitivity of 1:1 FASTA approach is similar to the 80-mer sliding window search for >35%/80aa but, as described previously [14], with significantly improved specificity.

3.5. Recommendation

Based on the performance of this simple application of existing modern bioinformatics tools, and its ease of implementation and interpretation in the context of a regulatory assessment, we advocate that the 1:1 FASTA approach be used to augment the antiquated FAO/WHO/CODEX amino-acid identity approach formulated 13 years ago. Adoption of this approach as a supplement to the FAO/WHO/CODEX criterion (>35% identity over 80 amino acids), and as a second tier to eliminate many biologically irrelevant homology hits generated by the FAO/WHO/CODEX criterion, should improve the safety assessment of GM crops as well as improve the assessment of the cross-reactive risk for other novel food proteins.

Acknowledgement

The authors thank Dr. Steve Evans for his critical review of this manuscript.

References

[1] Codex Alimentarius Commission, Proposed draft annex on the assessment of possible allergenicity of the draft guideline for the conduct of food safety assessment of foods derived from recombinant DNA plants. Joint FAO/WHO Food Standard Program Appendix IV (2003); 57–60.

[2] FAO/WHO, Evaluation of allergenicity of genetically modified foods, in: Report of a Joint FAO/WHO Expert Consulation on Allergenicity of Foods Derived from Biotechnology, FAO/WHO, Rome Italy, 2001, pp. 22–25, January.

[3] M. Bublin, H. Breitender, Cross-reactivity of peanut allergens, Curr. Allergy Asthma Rep. 14 (2014) 426, http://dx.doi.org/10.1007/s11882-014-0426-8.

[4] R.F. Cressman, G.S. Ladics, Further evaluation of the utility of sliding window FASTA in predicting cross-reactivity with allergenic proteins, Reg. Toxicol. Pharmacol. 54 (2009) 20–25.

[5] M.P. de Leon, A.c. Drew, I.N. Glaspole, C. Suphioglu, R.E. O'Hehir, J.M. Rolland, IgE cross-reactivity between the major peanuts allergen Ara h 2 and tree nut allergens, Mol. Immunol. 44 (2007) 463–471.

[6] F. Guarneri, In silico allergen identification: proposal for a revision of FAO/WHO guideline, Atti Accad. Pelorit. Pericol. Cl. Sci. Fis. Mat. Nat. 88 (2010) C1A1002006.

[7] G.S. Ladics, G.A. Bannon, A. Silvanovich, R.R. Cressman, Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergend, Mol. Nutr. Food Res. 51 (2007) 985–998.

[8] R.A. Herman, P. Song, S. Kumpatla, Percent amino-acid identity thresholds are not necessarily conservative for predicting allergenic cross-reactivity, Food Chem. Toxicol. 81 (2015) 141–142.

[9] R.A. Herman, P. Song, A. ThirumalaiswamySekhar, Value of eight-amino-acid matches in predicting the allergenicity status of proteins: an empirical bioinformatic investigation, Clin. Mol. Allergy 7 (2009) 9, http://dx.doi.org/10.1186/1476-7961-7-9.

[10] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 2444–2448.

[11] W.R. Pearson, Flexible Sequence similarity searching with the FASTA3 program package, Methods Mol. Biol. 132 (1999) 185–219.

[12] M.B. Stadler, B.M. Stadler, Allergenicity prediction by protein sequence, FASEB J. 17 (2003) 1141–1143.

[13] A. Silvanovich, G. Bannon, S. McClain, The use of *E*-score to determine the quality of protein alignment, Regul. Toxicol. Pharmacol. 54 (2009) 26–31.

[14] P. Song, R.A. Herman, S. Kumpatla, Evaluation of global sequence comparison and 1:1 FASTA local alignment in regulatory allergenicity assessment of transgenic proteins in food crops, Food Chem.Toxicol. 71 (2014) 142–148.