Food and Chemical Toxicology 132 (2019) 110656

Contents lists available at ScienceDirect

Food and Chemical Toxicology

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

Validation of bioinformatic approaches for predicting allergen cross reactivity

Rod A. Herman*, Ping Song

Corteva Agriscience™, 9330 Zionsville Road, Indianapolis, IN, 46268, USA

ARTICLE INFO

Keywords:
Allergen
Cross-reactivity
Bioinformatics
Validation
Sensitivity
Selectivity

ABSTRACT

Part of the allergenicity assessment of newly expressed proteins in genetically engineered food crops involves an assessment of potential cross-reactivity with known allergens. Bioinformatic approaches are used to evaluate the amino acid sequence identity or similarity between newly expressed proteins and the sequences of known allergens. To be useful, such approaches must be sensitive to detecting cross-reactive potential, but also capable of excluding low-risk sequences. One difficulty in comparing the effectiveness of different bioinformatic approaches has been the lack of a standardized validation and evaluation method. Here, we propose a standardized method for evaluating the sensitivity of different bioinformatic algorithms using a comprehensive database of known allergen sequences. We combine this with a previously described method for evaluating selectivity using sequences from a crop not known to commonly cause food allergy (e.g. maize) to compare the standard " > 35% identity-criterion over sliding-window of ≥80 amino acids" bioinformatic approach with the previously described "one-to-one (1:1) FASTA" similarity approach using an *E*-value threshold of 1E-9. Results confirm the superiority of the 1:1 FASTA approach for selectively detecting cross-reactive allergens. The validation methods described here can be applied to other algorithms to select even better fit-for-purpose approaches for evaluating cross-reactive risk.

1. Introduction

One element of the weight-of-evidence assessment of newly expressed proteins in genetically engineered (GE) crops is a bioinformatic investigation for potential cross reactivity with known allergens (Ladics et al., 2011). Historically, the algorithms developed and required by the regulatory agencies that oversee the safety of GE crops were not formally validated as being fit for purpose (Ladics et al., 2007). This may stem from the formulators of the initial criteria being experts in clinical allergy rather than in bioinformatics or formal method validation, especially in relation to risk assessment. The sensitivity of the bioinformatic methods was intended to be controlled based on identification of disparate amino acid sequences among cross-reactive allergens (selected through expert knowledge), followed by identification of the minimum amino acid identity between pairs of these sequences (Goodman et al., 2008). The most commonly used criterion developed in this manner is > 35% identity over a sliding window of ≥80 amino acids using an alignment tool such as FASTA (Codex alimentarius commission, 2007; FAO/WHO, 2001). Such criteria can be useful if their selectivity for filtering out false positives is acceptable (minimal false identification of non-cross-reactive sequences). Unfortunately, the previously mentioned "identity-criterion over sliding-window" approach has poor selectivity, and alternative criteria based on sequence similarity measures, rather than identity, have been found to be more selective and equally sensitive for detection of known cross-reactive allergens (Cressman and Ladics, 2009; Herman et al., 2015; Hileman et al., 2002; Ladics et al., 2007; Silvanovich et al., 2009; Song et al., 2014).

A variety of suitable algorithms and tools based on advanced similarity searches have been described, but a common validation approach to identify the best fit-for-purpose method has not been formalized. Previous evaluations of sensitivity have mimicked the initial selection of disparate amino acid sequences from cross-reactive allergens identified based on expert knowledge, followed by selection of similarity thresholds that favor detection. Selectivity was then evaluated using a set of protein sequences from crop plants not known to commonly cause allergy (e.g. maize) (Song et al., 2014).

Here, we propose a complementary and standardized method for evaluating sensitivity using full-length amino acid sequences contained in the COMPARE allergen database (http://comparedatabase.org/). This approach makes use of a full suite of known allergen sequences as query proteins to examine how well a given criterion would have

[*] Corresponding author.
E-mail addresses: rod.herman@corteva.com (R.A. Herman), ping.song@corteva.com (P. Song).

https://doi.org/10.1016/j.fct.2019.110656
Received 29 March 2019; Received in revised form 24 June 2019; Accepted 1 July 2019
Available online 03 July 2019

detected each sequence if it was yet to be identified as an allergen. We used the previously described approach for evaluating selectivity based on querying an array of proteins from a crop not known to commonly cause allergy. We exemplified this validation approach by comparing a previously described one-to-one (1:1) FASTA approach with the commonly used regulatory approach based on $> 35\%$ identity over an 80-amino-acid sliding window (Song et al., 2014, 2015). Note that these two bioinformatic methods have established preexisting thresholds of similarity and identity, respectively, and are used here to exemplify our proposed approach for comparing candidate methods for sensitivity.

2. Methods and materials

Bioinformatic approaches: Identity over a sliding window of 80 amino acids was compared with a 1:1 FASTA approach using the amino acid sequences in the COMPARE 2018 database. The COMPARE database was initially constructed using sequences in the AllergenOnline database (Goodman et al., 2016). These two bioinformatic approaches have been previously described (Codex alimentarius commission, 2007; FAO/WHO, 2001; Song et al., 2015; Song et al., 2014). Briefly, the first method parses each query protein into sliding windows of 80 amino acids, each of which is then aligned with known allergen sequences, followed by identification of matches with $> 35\%$ identity. An adjustment was made for alignments under 80 amino acids where the number of identical amino acid matches was divided by 80 to calculate percent identity over 80 amino acids (Song et al., 2014). The second method uses the FASTA algorithm to search for local alignments between the query protein and each allergen placed singly into a database ensuring that the significance of the similarity (*E*-value) does not vary as the database size changes over time when sequences are added or removed from the allergen database (not controlled using conventional FASTA approach). It is noteworthy that the 1:1 FASTA approach is not equivalent to setting the database size to a fixed value because the 1:1 FASTA approach has a database size that varies with the length of the single sequence in the database during each query. Furthermore, the statistical methods used to generate the *E*-value are different compared with those typically used on a full database (Pearson, 2016). The previously proposed threshold *E*-value of $< 1\text{E-}9$ was used to indicate cross-reactive potential.

Sensitivity: Full-length amino acid sequences in the COMPARE allergen database were putatively identified by searching the "definition" field of each entry (GenBank format) within the database for the word "partial" (and eliminating these) and also eliminating additional sequences of < 29 amino acids (minimum for achieving $> 35\%$ identity over 80 amino acids and also likely not to be full length sequences) from the query sequence pool, but not the searched database. The current version of the COMPARE database does not consistently identify sequences as full length or partial. Only putative full-length sequences were selected as the query set because this mirrors the situation for proteins expressed in GE crops which all have the complete sequences known. These full-length sequences were used singly to query the sequences in the COMPARE database and the best-match was identified excluding the identical entry in the database (equivalent to removing the identical entry in the database before conducting the query) (Fig. 1). Note that identical sequences from different source organisms were not removed from the database, simulating a situation where the query sequence was newly identified from a previously unknown source organism. Different best-match protein pairs were then compared with one another to find those pairs with matches not meeting the threshold of $> 35\%$ for the "identity-criterion over sliding-window" approach or an *E*-value $< 1\text{E-}9$ for the 1:1 FASTA approach. The results from each approach were then compared to determine unique sequences detected by only one of the methods, followed by an investigation of these unique matches for evidence of cross reactivity among the source organisms.

Selectivity: Bioinformatic methods were evaluated for selectivity using a set of protein sequences *in silico* translated from the maize genome (ftp://ftp.ncbi.nih.gov/genomes/genbank/plant/Zea_mays/latest_assembly_versions/GCA_000005005.6_B73_RefGen_v4/GCA_000005005.6_B73_RefGen_v4_translated_cds.faa.gz), since maize is not known to commonly cause allergy. Although maize allergy is rare, thirty sequences in the COMPARE database are sourced from *Zea mays* (maize) and these entries were used to identify sequences for removal from the maize query list since the intent was to evaluate sequences not known to cause allergy for false-positive results. Based on the thirty putative maize allergen sequences in the COMPARE database, entries with the following text terms in the definition were removed from the maize query set: phospholipid transfer protein, lipid transfer protein, lipid-transfer protein, lipid binding protein, lipid-binding protein, LTP, lipid binding transfer protein, allerg, profilin, expansin, and endochitinase. Finally, the remaining maize sequences were queried against the sequences in the COMPARE database after the thirty putative maize allergen sequences in the COMPARE database were removed.

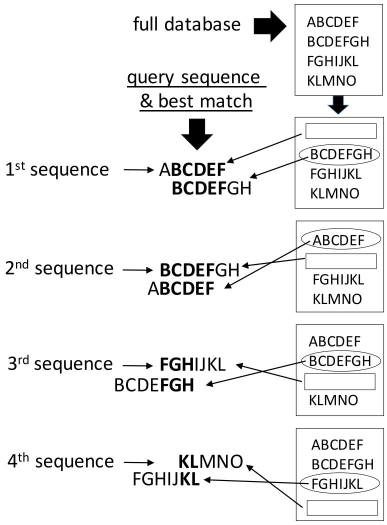

Fig. 1. Stylized illustration of sensitivity investigative approach. Each sequence is removed and used to query remaining sequences for the best match. Alphabetic symbols rather than amino acid symbols are used here to illustrate the generic process. Matches are for greatest identity or similarity depending on bioinformatic approach. Empty box indicates removed sequence and encircled sequence indicates best match.

3. Results and discussion

3.1. Sensitivity

Initial comparison of algorithms: There are 2038 amino acid sequences in the 2018 COMPARE allergen database. A total of 1553 putative full-length and 485 putative partial amino acid sequences were identified. Of the 1553 putative full-length sequences used to query the COMPARE database (after eliminating the identical entries from the database), 53 sequences with a best match of $\leq 35\%$ identity over a sliding window of ≥ 80 amino acids were identified, and 52 sequences were identified with an *E*-value $> 1\text{E-}9$ from the 1:1 FASTA comparison (Fig. 2). Both methods missed the same 42 sequences suggesting their uniqueness in the database (Tables 1 and 2).

3.2. Data cleansing and cross reactivity

Data cleansing: Data cleansing (or scrubbing) is the process of correcting datasets. A subset of sequences was selected in an automated manner (removal of those tagged "partial" and those < 29 amino acids long) from the COMPARE database as likely representing partial

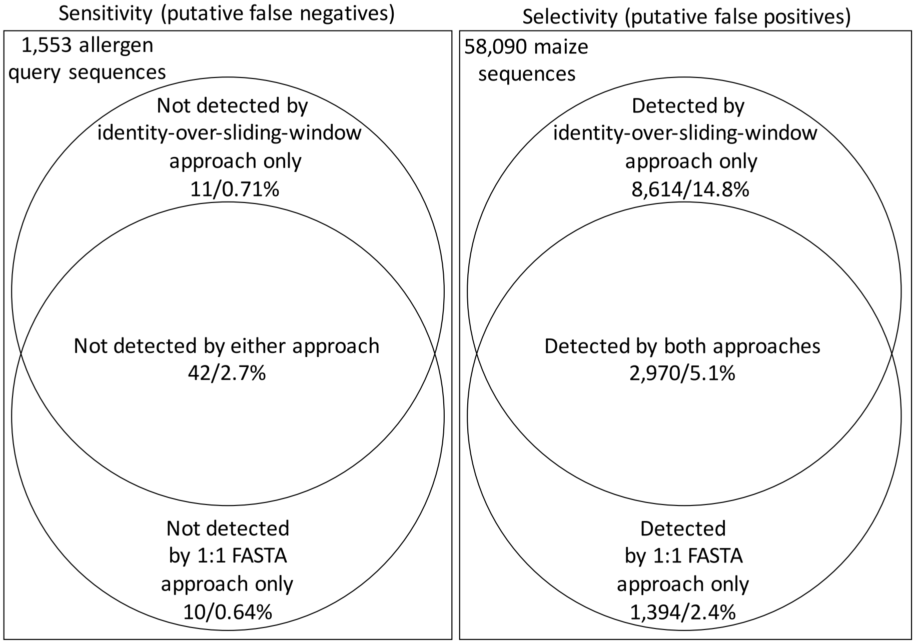
Sensitivity (putative false negatives)
1,553 allergen query sequences
Not detected by identity-over-sliding-window approach only 11/0.71%
Not detected by either approach 42/2.7%
Not detected by 1:1 FASTA approach only 10/0.64%
Selectivity (putative false positives)
58,090 maize sequences
Detected by identity-over-sliding-window approach only 8,614/14.8%
Detected by both approaches 2,970/5.1%
Detected by 1:1 FASTA approach only 1,394/2.4%

Fig. 2. Results comparing different bioinformatic approaches for detection of potential allergen cross reactivity. Sequences in COMPARE allergen database were used to evaluate sensitivity and maize protein sequences were used to evaluate selectivity. Sections in Venn diagram are not proportional to the number of sequences in each section.

sequences, but the remaining query sequences were not initially verified manually as being full length. For a sensitivity comparison between the two bioinformatic algorithms, one key measure is the number of full-length sequences detected by one method, but not the other. The use of full-length sequences as the query set is important because partial sequences may not actually include relevant IgE epitopes (or other key motifs) and thus could improperly skew the results of this investigation. Furthermore, partial sequences are not representative of newly expressed proteins in GE crops for which complete sequences are known. Therefore, select short sequences not detected by these bioinformatic methods were manually checked for their full-length status. In addition, the predicted cross reactivity between source organisms for best-match hits for missed sequences was investigated for literature support.

Uniquely missed by 1:1 FASTA approach: Of the ten sequences uniquely missed by the 1:1 FASTA approach, one query protein returned a best match subject from the same source organism as the query protein (Table 1). Query accession AAN73248.1 and subject accession CAA11266.1 share the fungus *Fusarium culmorum* as the source organism. However, the *E*-value for the 81 amino acid alignment is 22 suggesting that the aligned regions of the proteins do not share statistically significant similarity. For the nine remaining protein pairs detected only by the sliding window approach, no literature documenting cross reactivity between the source organisms was identified.

Uniquely missed by sliding window approach: Of the eleven sequences uniquely missed by the sliding-window approach, four query proteins returned best match subjects from the same source organism as the query protein (Table 2). However, further investigation found each of these four query sequences to be partial sequences and thus not representative of newly expressed proteins in GE crops (for which complete sequences are known) (Bulone et al., 1998; Coutos-Thevenot et al., 1993; Lind et al., 1988). The only other best-match pair with likely cross-reactive source organisms is CAA26038.1 from *Apis mellifera* and P01502.1 from *Apis dorsata*. However, the subject protein from *Apis dorsata* seems to have been placed in the database due to amino acid sequence homology with the query protein rather than experimental evidence of causing allergic reactions (Karamloo et al., 2005; Kemeny et al., 1983). For the six remaining protein pairs detected only by the 1:1 FASTA approach, no literature documenting cross reactivity between the source organisms was identified.

Shared missed sequences: Of the 42 sequences not detected by either bioinformatic approach, only two pairs of source organisms for the best-match sequence pairs appeared to have documented cross reactivity (Table 2). Query accession P86888.1 from peach and subject accession C0HKC0.1 from pomegranate did not meet the 1:1 FASTA threshold (*E*-value = 1.30E-7) or satisfy the > 35%-identity sliding-window criteria. These two source organisms are reported to show cross reactivity (Gaig et al., 1999) and the query sequence in the COMPARE database appears to be full length (63 amino acids long) (Tuppo et al., 2013). However, the subject sequence from pomegranate is only 20 amino acids long and represents approximately 30% of the putative full-length protein (Tuppo et al., 2017). In addition, query accession P82946.2 from orchard grass and subject accession cad54671.2 from timothy grass did not meet the 1:1 FASTA threshold (*E*-value = 1.50E-7) or satisfy the sliding window criteria, and their source organisms have known cross reactivity (Chakrabarty et al., 1981). However, the query protein is only 55 amino acids long (shortest of the 52 query proteins missed by the 1:1 FASTA approach) while the subject protein is 508 amino acids long. Upon investigation, it was found that the 55 amino acid orchard-grass sequence was partial, representing approximately 10% of the full-length sequence and thus is not representative of newly expressed proteins in GE crops (Leduc-Brodard et al., 1996). Four other query proteins returned best match subjects from the same source organism as the query protein (accessions BAV90601.1 from the dust mite *Dermatophagoides farinae*, AGL34967.1 from coffee *Coffea arabica*, NP_776,953.1 from cow's milk *Bos taurus*, and P06886.1 from the bacteria *Staphylococcus aureus*) which precludes an analysis of source-organism cross reactivity. For the 36 remaining protein pairs detected by neither approach, no literature documenting cross reactivity between the source organisms was identified.

Overall sensitivity: Both bioinformatic approaches performed similarly in terms of sensitivity, and neither uniquely identified known cross reactive allergens. Both methods appeared to detect any relevant amino acid homology that might confer allergenic cross reactivity.

3.3. *Selectivity*

The selectivity of the sliding-window and 1:1 FASTA bioinformatic approaches were compared using the *in silico* translated gene sequences for maize as query proteins since maize is a rarely allergenic crop (58,286 sequences). However, the known allergen amino acid sequences were first removed from the COMPARE allergen database, and sequences tagged with several text terms related to these sequences

Table 1
Protein sequences missed by one bioinformatic approach as a cross-reactive risk.

Query				1:1 FASTA Subject				Alignment		Sliding window alignment and subject			
								1:1 FASTA					
accession	length	species	common name	accession	length	species	common name	*E*-value	overlap	% identity	accession	length	species	common name
(detected by 1:1 FASTA only)													
P16312.1	30[P]	***Dermatophagoides microceras***	dust mite	ABA39436.1	276	***Dermatophagoides farinae***	dust mite	6.50E-19	30	≤35	–	-	-	–
CAA26038.1	70	***Apis mellifera***	honey bee	P01502.1	26	***Apis dorsata***	giant honey bee	2.00E-14	26	≤35	–	-	-	–
CCW27997.1	70	*Hevea brasiliensis*	rubber tree (latex)	P82977.2	84	*Triticum aestivum*	wheat	9.90E-14	63	≤35	–	-	-	–
AHF71027.1	237	*Betula pendula*	birch	ACE82289.1	222	*Triticum aestivum*	wheat	9.70E-13	209	≤35	–	-	-	–
P33556.1	38[P]	***Vitis* sp.**	grape	P80274.1	37	***Vitis* sp.**	grape	2.50E-12	37	≤35	–	-	-	–
BAG93480.1	476	*Oryza sativa*	Asian rice	AAA32708.1	499	*Aspergillus oryzae*	fungus	4.30E-12	370	≤35	–	-	-	–
P80274.1	37[P]	***Vitis* sp.**	grape	P33556.1	38	***Vitis* sp.**	grape	5.00E-12	37	≤35	–	-	-	–
P81216.1	29[P]	***Equus caballus***	horse	P81217.1	19	***Equus caballus***	horse	2.30E-10	18	≤35	–	-	-	–
CAK50389.1	115	*Anisakis simplex*	human parasitic nematode	AAR92223.1	116	*Actinidia delicio*	kiwi	8.10E-10	87	≤35	–	-	-	–
P85524.1	150	*Actinidia deliciosa*	kiwi	ABZ81045.1	159	*Quercus alba*	white oak	1.00E-09	143	≤35	–	-	-	–
AAR92223.1	116	*Actinidia deliciosa*	kiwi	CAK50389.1	115	*Anisakis simplex*	parasitic fish worm	3.50E-09	87	≤35	–	-	-	–
(detected by sliding window only)													
AAP06493.1	129	*Schistosoma japonicum*	human blood fluke	CAA75506.1	133	*Helianthus annuus*	sunflower	1.20E-08	134	35.30	AIO08866.1	130	*Dermatophagoides farinae*	dust mite
ABA42918.1	274	*Cladosporium herbarum*	fungus	AAB26195.1	68	*Ascaris suum*	pig roundwork	8.00E-04	22	35.70	P56166	294	*Phalaris aquatica*	canary grass
AAN73248.1	450	***Fusarium culmorum***	fungus	AAA28303.1	203[P]	*Dolichovespula arenaria*	wasp	9.80E-04	95	35.80	CAA11266.1	302	***Fusarium culmorum***	fungus
XP_003,030,591.1	576	*Schizophyllum commune*	mushroom	BAF45320.1	65	*Cryptomeria japonica*	Japanese cedar	7.30E-04	20	36.60	AAC25998.1	82	*Phleum pratense*	timothy grass
BAI94503.1	165	*Cryptomeria japonica*	Japanese cedar	ABX56711.1	116	*Arachis hypogaea*	peanut	1.60E-07	119	37.00	ABX56711.1	116	*Arachis hypogaea*	peanut
CAA55854.1	205	*Betula pendula*	birch	BAA09634.1	79	*Brassica rapa*	brassica	3.10E-08	57	38.00	AAX77686.1	160	*Ambrosia artemisiifolia*	ragweed
BAA06905.1	731	*Cucumis melo*	muskmellon	P29600.1	269	*Bacillus lentus*	bacteria	5.60E-07	159	41.20	ADE74975.1	403	*Aspergillus versicolor*	fungus
NP_776,945.1	1364	*Bos taurus*	cattle (beef)	AAX77383.1	510	*Sinapis alba*	brassica	1.10E-05	75	47.50	AKF12278.1	156	*Parthenium hysterophorus*	aster
AAC49447.1	151	*Hevea brasiliensis*	rubber tree (latex)	BAB15802.1	517	*Glycine max*	soybean	4.40E-04	93	47.55	AAN73248.1	177	*Manihot esculenta*	cassava
P81729.1	91	*Brassica rapa*	brassica	1WKX_A	43	*Hevea brasiliensis*	rubber tree (latex)	3.00E-08	34	57.60	CAA05978.1	187	*Hevea brasiliensis*	rubber tree (latex)

[P] partial sequence; Bolded species entries do not exclude probable cross reactivity. Note that sliding-window software does not report matches of < 35% identity.

Table 2
Protein sequences not detected by either bioinformatic approach as a cross-reactive risk.

Query				1:1 FASTA Subject				Alignment		Sliding window
								1:1 FASTA		
accession	length	species	common name	accession	length	species	common name	*E*-value	overlap	% identity
Q6R4B4.1	231	*Alternaria alternata*	fungus	AAX33729.1	216	*Periplaneta americana*	cockroach	1.50E-08	119	≤35
BAG88472.1	221	*Oryza sativa*	Asian rice	AAL73404.1	515	*Corylus avellana*	hazelnut	2.10E-08	138	≤35
P13080.1	579	*Aedes aegypti*	mosquito	AAD38942.1	496^{P}	*Dermatophagoides pteronyssinus*	dust mite	2.30E-08	225	≤35
L7UZ85.1	885	*Dermatophagoides farinae*	dust mite	AAF31151.1	171	*Olea europaea*	olive	3.20E-08	160	≤35
AAB22817.1	273	*Arachis hypogaea*	peanut	AK068307.1	764	*Oryza sativa*	Asian rice	8.10E-08	276	≤35
P86888.1	63	***Prunus persica***	peach	C0HKC0.1	20^{P}	***Punica granatum***	pomegranate	1.30E-07	20	≤35
AAL49391.1	98	*Felis catus*	house cat	CAK50389.1	115	*Anisakis simplex*	human parasitic nematode	1.50E-07	60	≤35
P82946.1	55^{P}	***Dactylis glomerata***	orchard grass	cad54671.2	508	***Phleum pratense***	timothy grass	1.50E-07	20	≤35
AAF07903.2	169	*Triatoma protracta*	kissing bug	ACF53837.1	190	*Blattella germanica*	cockroach	3.40E-07	171	≤35
CAD56944.1	1770	*Apis mellifera*	honey bee	vitellogenin^{M}	284	*Gallus gallus*	red junglefowl	5.00E-07	323	≤35
AAC67308.1	191	*Schistosoma japonicum*	human blood fluke	AAT45383.1	109	*Lates calcarifer*	seabass	1.10E-06	58	≤35
P81943.3	86	*Apium graveolens*	celery	CAH92637.1	423	*Lolium perenne*	perennial ryegrass	1.70E-06	35	≤35
BAJ04354.1	472	*Cryptomeria japonica*	Japanese cedar	P00791.3	385	*Sus scrofa*	pig (pepsin)	1.80E-06	362	≤35
ADK47876.1	126	*Thaumetopoea pityocampa*	moth	P02224.2	162	*Chironomus thummi thummi*	midge	6.40E-06	129	≤35
P24337.1	80	*Glycine max*	soybean	ACE07189.1	117	*Artemisia vulgaris*	mugwort	3.40E-05	79	≤35
ACD65081.1	325	*Forcipomyia taiwana*	midge	P14947.1	97	*Lolium perenne*	perennial ryegrass	4.60E-05	31	≤35
P06886.1	234	***Staphylococcus aureus***	bacteria	P20723.1	258	***Staphylococcus aureus***	bacteria	5.80E-05	186	≤35
Q28050.1	101	*Bos taurus*	cattle (amniotic fluid)	ADD19989.1	222	*Glossina morsitans morsitans*	tsetse fly	5.90E-05	51	≤35
AGL34968.1	65	*Coffea arabica*	Arabian coffee	CCW27997.1	70	*Hevea brasiliensis*	rubber tree (latex)	8.00E-05	43	≤35
AAR17475.1	228	*Penicillium citrinum*	Penicillium fungus	AAT95010.1	227	*Polistes dominula*	wasp	8.20E-05	131	≤35
ABI26088.1	169	*Alternaria alternata*	Alternaria fungus	P80207.1	129	*Brassica juncea*	brassica	9.30E-05	19	≤35
AAK67492.1	108	*Curvularia lunata*	fungi	AAC48795.1	180	*Canis lupus familiaris*	dog	1.00E-04	59	≤35
AKJ77985.1	89	*Triticum aestivum*	wheat	AHF71027.1	237	*Betula pendula*	European white birch	1.40E-04	23	≤35
CAM54066.1	185	*Aspergillus fumigatus*	fungus	P86745.1	108	*Merluccius australis australis*	southern hake (fish)	1.50E-04	98	≤35
CAA57342.1	350	*Candida albicans*	yeast	CAA52194.1	607	*Equus caballus*	horse	1.70E-04	240	≤35
NP_776,953.1	222	***Bos taurus***	cattle (milk)	AAA30429.1	214	***Bos taurus***	cattle (milk)	2.10E-04	158	≤35
CAA65313.1	137	*Triticum aestivum*	wheat	AAT37679.1	342^{P}	*Rhodotorula mucilaginosa*	yeast	2.30E-04	82	≤35
ABB89950.1	733	*Penicillium citrinum*	fungus	P81729.1	91	*Brassica rapa*	brassica	2.80E-04	49	≤35
NP_001,036,878.1	227	*Bombyx mori*	silkworm	P49148.1	110	*Alternaria alternata*	fungus	2.80E-04	56	≤35
AGL34967.1	80	***Coffea arabica***	Arabian coffee	AGL34968.1	65	***Coffea arabica***	Arabian coffee	2.90E-04	79	≤35
P18153.2	321	*Aedes aegypti*	mosquito (saliva)	ABX26138.1	152	*Olea europaea*	olive	3.80E-04	25	≤35
AAN11300.1	236	*Candida albicans*	yeast	AAW29810.1	507	*Juglans regia*	English walnut	4.50E-04	147	≤35
P00304.2	101	*Ambrosia artemisiifolia*	ragweed	P84296.1	161	*Chironomus thummi thummi*	migde	5.10E-04	42	≤35
CAA09886.2	179	*Malassezia sympodialis*	Malassezia	P02221.2	158	*Chironomus thummi thummi*	midge	5.10E-04	80	≤35
BAW32535.1	225	*Scleronephthya gracillimum*	soft coral	AAD03608.1	367	*Juniperus ashei*	Ozark white cedar	5.90E-04	128	≤35
AF084828_1	342	*Malassezia furfur*	fungus	NLTP1_PEA	120	*Pisum sativum*	pea	6.70E-04	97	≤35
CAD42710.1	105	*Cladosporium herbarum*	fungus	ABH06350.1	129	*Blomia tropicalis*	storage mite	6.90E-04	34	≤35
CAA65341.1	350	*Malassezia sympodialis*	skin fungi	P27357.1	173	*Triticum aestivum*	wheat	9.20E-04	86	≤35
BAV90601.1	128	***Dermatophagoides farinae***	dust mite	ACK76291.1	259	***Dermatophagoides farinae***	dust mite	1.10E-03	32	≤35
CAI43283.4	618	*Malassezia sympodialis*	skin fungi	AAA34280.1	286	*Triticum aestivum*	wheat	2.10E-03	46	≤35
AAP94213.1	155	*Humulus japonicus*	Japanese hop	NP_001,037,083.1	195	*Bombyx mori*	silkworm	2.50E-03	27	≤35
AKJ77987.1	108^{H}	*Triticum aestivum*	wheat	AAM43909.1	392	*Aspergillus fumigatus*	fungus	3.30E-03	18	≤35

^{H}hypothetical sequence; ^{P}partial sequence; ^{M}manual entry; Bolded species entries do not exclude probable cross reactivity.
Note that sliding-window software does not report matches of < 35% identity.

were removed from the maize query set. This data cleansing was designed to reduce the number of potentially true allergens from the query set, so that a better absolute rate of false positive results could be obtained to compare bioinformatic methods.

A total of 58,090 putative non-allergen amino acid sequences were identified from maize (Fig. 2). Of these sequences, the sliding-window approach identified 11,584 (19.9%) as putative allergens, while the 1:1 FASTA approach identified 4363 (7.5%) as putative allergens (both approaches identified the same 2970 subset of sequences). This indicates that the 1:1 FASTA approach is much more selective at identifying potential allergenic cross reactivity (2.7 fold fewer false-positive hits) compared with the sliding-window approach while having almost identical sensitivity.

4. Conclusions

Previous sensitivity analyses for different bioinformatic approaches designed to detect potential cross reactivity with known allergens were often compared using disparate allergen sequences known to show allergic cross reactivity. This type of investigation made use of expert clinical allergy knowledge, but often resulted in inconsistent approaches and criteria among studies investigating the best fit-for-purpose bioinformatic algorithms. Here, we present a standardized approach for comparing the sensitivity of different bioinformatic approaches using a database of allergen sequences. Rather than starting with particular groups of known cross-reactive allergens, all putative full-length sequences in the COMPARE allergen database were used as query proteins. Unique sequences not detected by each bioinformatic approach were investigated for known cross reactivity to evaluate the comparative sensitivity of each approach. Furthermore, selectivity was evaluated using protein sequences from maize.

As previously reported, the 1:1 FASTA approach to identifying potential allergen cross reactivity was found to be superior to the 80-amino-acid sliding-window/identity approach (Song et al., 2014, 2015). While this result is not surprising since the physiochemical properties of mismatched amino acids are considered by similarity searches and not by identity searches (Herman et al., 2015), it is important to document the superior performance of the former approach since regulatory guidelines continue to be based on inferior identity criteria.

It is hoped that this systematic approach for comparing the bioinformatic algorithms and thresholds for sensitivity can be combined with a selectivity approach, based on a set of amino acid sequences from sources not commonly causing allergy, to identify the best available fit-for-purpose algorithms for detecting allergic cross-reactive risk while maintaining good selectivity. Although the selectivity of the 1:1 FASTA approach is much improved over the 80-amino-acid sliding-window/identity approach, while maintaining almost identical sensitivity, clearly a 7.5% false-positive rate can be improved upon. In fact, since none of the alignments detected by either criterion alone indicate documented cross reactivity, requiring alignments to simultaneously meet both sets of the bioinformatic criteria discussed here appears to maintain excellent sensitivity while producing only a 5.1% false-positive rate (Fig. 2). Thus, this investigation indicates that > 35% identity matches can be further filtered through the 1:1 FASTA criterion to remove many false positives without sacrificing detection of cross-reactive risk.

This validation approach would be aided by comprehensively tagging verified full-length and partial sequences in the COMPARE database. The availability of a comprehensive and curated list of cross-reactive protein sequences and source organisms would also improve the efficiency and consistency of the validation process. In addition, the 63 sequences in the COMPARE allergen database not detected in this investigation by either the 1:1 FASTA similarity and/or sliding window approach might be reinvestigated for the strength of the evidence supporting their allergenicity as it is possible they were initially included in the database based on overly conservative selection criteria, and perhaps lack experimental evidence of allergy.

Conflicts of interest

The authors are employed by a company that develops and markets genetically engineered seed.

Declaration of interests

The authors are employed by a company that develops and markets transgenic seed.

References

Bulone, V., Krogstad-Johnsen, T., Smestad-Paulsen, B., 1998. Separation of horse dander allergen proteins by two-dimensional electrophoresis: molecular characterisation and identification of Equ c 2.0101 and Equ c 2.0102 as lipocalin proteins. Eur. J. Biochem. 253, 202–211.

Chakrabarty, S., Loewenstein, H., Ekramoddoullah, A., Kisil, F., Sehon, A., 1981. Detection of cross-reactive allergens in Kentucky bluegrass pollen and six other grasses by crossed radioimmunoelectrophoresis. Int. Arch. Allergy Immunol. 66, 142–157.

Codex alimentarius commission, 2007. Report of the Sixth Session of the CODEX Ad Hoc Intergovernmental Task Force on Foods Derived from Biotechnology. Chiba, Japan 27 November – 1 December 2006. ALINORM 07/30/34.

Coutos Thevenot, P., Jouenne, T., Maes, O., Guerbette, F., Grosbois, M., Le Caer, J.P., Boulay, M., Deloire, A., Kader, J.C., Guern, J., 1993. Four 9-kDa proteins excreted by somatic embryos of grapevine are isoforms of lipid-transfer proteins. Eur. J. Biochem. 217, 885–889.

Cressman, R.F., Ladics, G., 2009. Further evaluation of the utility of "Sliding Window" FASTA in predicting cross-reactivity with allergenic proteins. Regul. Toxicol. Pharmacol. 54, S20–S25.

FAO/WHO, 2001. Evaluation of Allergenicity of Genetically Modified Foods. Report of Joint FAO/WHO Expert Consultation. Food and Agriculture Organization of the United Nations, Rome.

Gaig, P., Bartolome, B., Lleonart, R., García-Ortega, P., Palacios, R., Richart, C., 1999. Allergy to pomegranate (Punica granatum). Allergy 54, 287–288.

Goodman, R.E., Vieths, S., Sampson, H.A., Hill, D., Ebisawa, M., Taylor, S.L., van Ree, R., 2008. Allergenicity assessment of genetically modified crops - what makes sense? Nat. Biotechnol. 26, 73–81.

Goodman, R.E., Ebisawa, M., Ferreira, F., Sampson, H.A., van Ree, R., Vieths, S., Baumert, J.L., Bohle, B., Lalithambika, S., Wise, J., 2016. AllergenOnline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. Mol. Nutr. Food Res. 60, 1183–1198.

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

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.

Karamloo, F., Schmid-Grendelmeier, P., Kussebi, F., Akdis, M., Salagianni, M., von Beust, B.R., Reimers, A., Zumkehr, J., Soldatova, L., Housley-Markovic, Z., 2005. Prevention of allergy by a recombinant multi-allergen vaccine with reduced IgE binding and preserved T cell epitopes. Eur. J. Immunol. 35, 3268–3276.

Kemeny, D.M., Harries, M.G., Youlten, L.J., Mackenzie-Mills, M., Lessof, M.H., 1983. Antibodies to purified bee venom proteins and peptides: I. Development of a highly specific RAST for bee venom antigens and its application to bee sting allergy. J. Allergy Clin. Immunol. 71, 505–514.

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.

Ladics, G.S., Cressman, R.F., Herouet-Guicheney, C., Herman, R.A., Privalle, L., Song, P., Ward, J.M., McClain, S., 2011. Bioinformatics and the allergy assessment of agricultural biotechnology products: industry practices and recommendations. Regul. Toxicol. Pharmacol. 60, 46–53.

Leduc-Brodard, V., Inacio, F., Jaquinod, M., Forest, E., David, B., Peltre, G., 1996. Characterization of Dac g 4, a major basic allergen from Dactylis glomerata pollen. J. Allergy Clin. Immunol. 98, 1065–1072.

Lind, P., Hansen, O., Horn, N., 1988. The binding of mouse hybridoma and human IgE antibodies to the major fecal allergen, Der p I, of Dermatophagoides pteronyssinus. Relative binding site location and species specificity studied by solid-phase inhibition assays with radiolabeled antigen. J. Immunol. 140, 4256–4262.

Pearson, W.R., 2016. Finding protein and nucleotide similarities with FASTA. Current protocols in bioinformatics 53 3.9. 1-3.9. 25.

Silvanovich, A., Bannon, G., McClain, S., 2009. The use of E-scores to determine the quality of protein alignments. Regul. Toxicol. Pharmacol. 54, S26–S31.

Song, P., Herman, R.A., Kumpatla, S., 2014. Evaluation of global sequence comparison and one-to-one FASTA local alignment in regulatory allergenicity assessment of transgenic proteins in food crops. Food Chem. Toxicol. 71, 142–148.

Song, P., Herman, R., Kumpatla, S., 2015. 1:1 FASTA update: using the power of *E*-values in FASTA to detect potential allergen cross-reactivity. Toxicology Reports 2, 1145–1148.

Tuppo, L., Alessandri, C., Pomponi, D., Picone, D., Tamburrini, M., Ferrara, R., Petriccione, M., Mangone, I., Palazzo, P., Liso, M., 2013. Peamaclein–a new peach allergenic protein: similarities, differences and misleading features compared to Pru p 3. Clin. Exp. Allergy 43, 128–140.

Tuppo, L., Alessandri, C., Pasquariello, M.S., Petriccione, M., Giangrieco, I., Tamburrini, M., Mari, A., Ciardiello, M.A., 2017. Pomegranate cultivars: identification of the new IgE-binding protein pommaclein and analysis of antioxidant variability. J. Agric. Food Chem. 65, 2702–2710.