GM CROPS & FOOD
2020, VOL. 11, NO. 2, 93–96
https://doi.org/10.1080/21645698.2019.1709021

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

Allergen false-detection using official bioinformatic algorithms

Rod A. Herman and Ping Song

Regulatory and Stewardship, Corteva Agriscience, Indianapolis, IN, USA

ABSTRACT
Bioinformatic amino acid sequence searches are used, in part, to assess the potential allergenic risk of newly expressed proteins in genetically engineered crops. Previous work has demonstrated that the searches required by government regulatory agencies falsely implicate many proteins from rarely allergenic crops as an allergenic risk. However, many proteins are found in crops at concentrations that may be insufficient to cause allergy. Here we used a recently developed set of high-abundance non-allergenic proteins to determine the false-positive rates for several algorithms required by regulatory bodies, and also for an alternative 1:1 FASTA approach previously found to be equally sensitive to the official sliding-window method, but far more selective. The current investigation confirms these earlier findings while addressing dietary exposure.

ARTICLE HISTORY
Received 27 August 2019
Revised 19 December 2019
Accepted 21 December 2019

KEYWORDS
Allergen; cross-reactivity; bioinformatics; selectivity; exposure

Introduction

Newly expressed proteins in genetically engineered (GE) foods are evaluated for allergenic risk. Multiple lines of evidence are used in a weight-of-evidence risk assessment. The most important factors to consider in this risk assessment include the allergenic status of the organism from which the transgene originates, the concentration of the protein in food, and the structural similarity of the protein to known allergens.[1] Regulatory agencies that assess the safety of GE crops also consider the heat and digestive stability of the expressed proteins, but these factors have been shown to be poorly associated with allergenic risk.[1–4] The structural similarity of a novel food protein to known allergens is typically assessed by comparing amino acid sequences. Previous work has shown that the official bioinformatic algorithms required by regulatory agencies for comparing the amino acid sequence of a newly expressed protein with that of known allergens falsely implicate many non-allergens as being an allergenic risk.[5–7] The most commonly used official bioinformatic method divides the newly expressed protein into overlapping 80-amino-acid contiguous sequences and then looks for >35% identity among aligning sequences within known allergen sequences (sliding-window approach).[8,9] Another standard approach looks for exact 8-amino-acid contiguous matches between the novel food protein and known allergens, but this latter method has been largely dismissed by scientists as not useful although most regulatory agencies still require such searches to be completed.[10,11] Short amino-acid identity matches have been shown to identify many false-positive sequences while not identifying any novel cross-reactive allergen pairs.[10] More recently, the European Food Safety Authority (EFSA) issued guidance on assessing proteins for non-IgE-mediated celiac-disease risk using short amino acid motifs and partial matches with 9-mer peptides known to cause celiac disease.[12] Predictably, these latter bioinformatic searches find a large number of random false-positive sequences derived from plant and animal proteins not associated with celiac disease.[13]

We and others have previously published on equally sensitive bioinformatic algorithms for detecting allergenic risk, but with substantially better selectivity for eliminating proteins with negligible risk.[5–7] These latter methods use conventional software (e.g. FASTA) for estimating amino acid sequence similarity rather than identity, and categorize risk using thresholds based on statistical measures of similarity (e.g. *E*-values). *E*-value calculations were initially developed to detect evolutionary relationships between sequences and organisms but have been found useful in detecting similar protein functions and structures, the latter

CONTACT Rod A. Herman Rod.herman@corteva.com Corteva Agriscience, 9330 Zionsville Road, Indianapolis, IN 46268, USA

of which might indicate cross-reactive binding to the IgE antibodies that are typically associated with allergy.[14] False-positive rates typically were estimated in these published investigations by determining the percentage of the full suite of proteins in one or more rarely allergenic food crops that are detected by various bioinformatic algorithms as representing an allergenic risk.

One weakness of using a large set of protein sequences from a non-allergenic food crop to assess false-positive rates is that actual dietary exposure to many of the proteins may be limited due to low concentrations in food. While relative comparisons among bioinformatic methods are still valid, the absolute false-positive rates might be skewed upward due to real allergens being expressed in non-allergenic crops at levels below which allergy is induced.

Recently, a list of abundant food proteins with low allergenic potential (hereafter referred to as non-allergens) was published along with the methods used to determine their abundance and status as non-allergens.[15] This list can now be used to better assess the false-positive rates for different bioinformatic algorithms designed to selectively detect allergenic risk. Here we used this list of abundant non-allergenic food proteins to assess the false-positive rates for the official criteria of >35% identity over an 80-mer sliding-window and an 8-mer exact-match, and a previously reported 1:1 FASTA similarity approach.[7,8] Furthermore, we evaluate the selectivity of the recently implemented EFSA celiac peptide motif searches using these high-abundance non-allergens.

Methods and Materials

The 178 UniRef90 Cluster IDs listed in Table 4 of Krutz et al.[15], were used to search the UniProt database to obtain an amino acid sequence for each protein. Of these sequences, 169 returned current entries, and of those, 125 indicated the same source organism as listed in Table 4 of Krutz et al. The amino acid sequences for these 125 high-abundance non-allergens were compared with the allergen sequences in the COMPARE database version 2019 (http://db.comparedatabase.org/) using the standard search for >35% identity across 80-amino-acid windows and with the previously described 1:1 FASTA approach (with an *E*-value threshold of 1E-9 using FASTA version 35).[6] The percentage of non-allergens showing above threshold identity or similarity, respectively, was used to estimate the false-positive rate for each bioinformatic algorithm. In addition, 8-amino-acid exact matches between the non-allergens and allergens were determined. Finally, the number of sequences detected by the EFSA celiac-causing Q/EX_1PX_2 motif (Q = glutamine; E = glutamic acid; X_1 = L [leucine], Q, F [phenylalanine], S [serine], or E; *P* = proline; X_2 = Y [tyrosine], F, A [alanine], V [valine], or Q) and partial-match identity searches were determined (9-mer match allowing 3 mismatches with HLA-DQ8 restricted epitopes). The COMPARE database is used by the major registrants of genetically engineered crops when implementing the sliding-window, contiguous eight amino acid, and celiac peptide searches required by various regulatory bodies and thus represents current practice.

Results and Discussion

Of the 125 high-abundance non-allergenic food-crop proteins evaluated, 11 were implicated as an allergenic risk by the standard sliding-window bioinformatic approach and 1 was implicated by the 1:1 FASTA approach (Table 1). The 8-amino-acid search produced 3 hits and the EFSA celiac-peptide motif searches produced 13 hits (none of which could be excluded based on the presence of a proline duplex or based on positively-charged amino acids appearing in all 9-mer restricted-epitope matches at key positions as outlined by EFSA guidance). There were no 9-mer matches allowing 3 mismatches with the HLA-DQ8 restricted epitopes.

Previous work using 50,090 protein sequences from maize found the sliding-window bioinformatic approach to falsely implicate 19.9% of putative non-allergens as allergens, while the 1:1 FASTA approach falsely implicated 7.5% of proteins.[16] Using the 125 high-abundance non-allergens from food crops[15], false-positive rates were found to be 8.8% for the sliding-window approach and 0.8% for the 1:1 FASTA approach. The 8-amino acid exact match criterion falsely implicated 2.4% of the 125 high-abundance food-crop proteins and the celiac-peptide-motif search incorrectly found 10.4% of the 125 non-allergenic

Table 1. Bioinformatic matches (number of allergens or motifs) between non-allergens and allergens using different algorithms.

UniProt Entry	Bioinformatic Match (hits)				UniProt Entry	Bioinformatic Match (hits)				UniProt Entry	Bioinformatic Match (hits)			
Maize	Sliding Window	1:1 FASTA	8-mer Match	EFSA Celiac	Spinach	Sliding Window	1:1 FASTA	8-mer Match	EFSA Celiac	Potato	Sliding Window	1:1 FASTA	8-mer Match	EFSA Celiac
P28794					P80082					J7ENS8				
P06673					P10871					O24378				
P81009					A0A0K9QE98					Q43652				
P46517	1				P12301					Q9M3H3				
B6T8E4					P06003					P19595				
B6SGF3					P00833					O24379				
B6TTP4					P04160					P04045				
Q41881	1			2	P60128					M1D7J7				
B6UH99					P12355					M1AYK4				
P80639					A0A0K9QP00				1	Q93X17				
B6UH67					P12359					M0ZNV9				
B4FFZ9					P00455					Q00782				
B4FFK9			2		P12353					M1BPE5				
E9JVD4					Q41385					Q9AWA5				
P29518					P11402				1	Q3HRY7				1
B6SL97					P13788					P37829				
Q01526					O20252				1	Q9M4G5	25			
B4FUH2					P17353					Q9M4G4				
B4FPL1					P22418					C6F3B7	2			
B6UGJ4					A1XIR6					P33191	7		1	5
P55240					Q8RU73					P37830				
Q84J79					P09559					K7WJT8				
K7UNW7					P05435					Q948Z8				
B6T7B2					P06508					P23509				
P93804					Q02254									
B4F7S2					Q02060									
Rice					Tomato					Wheat				
Q6Z782					P14903					P33432		2		
A2XMB2					P38416					Q08000				
P37833					Q08655					P20158				
P46520					Q9SWF5					Q03968	2			
Q07661					P10967				1	W5BUF4				
P55142					P47921					P12299				
P0C5A4	1				Q6QLX4			5		P02276				1
Q6AVA8					Q40128					P30523				
Q69UI2					Q08451					P12783				
Q01L47	2				Q43497									
Q94JF2					P93207									
Q10LP5					P05116									
Q9AUV8				1	P46301									
Q5ZEL0					Q5NE21	1								
C7J0T2				1	K4B3I4									
Q6ZHP6				1	Q9ZR41									
P30298					O24024									
Q8H8B0	2			1	Q42876									
Q8H920					P38546				1					
A3AHG5	5				Q6QLU0									

proteins to represent a celiac-disease risk (only 1 of which originated from a crop known to cause celiac symptoms, wheat, but with no reports of this peptide causing celiac disease).

Together, the sliding window, 8-mer, and celiac-peptide-motif searches are required by some global government regulatory bodies and found 24 of the 125 non-allergenic proteins to present an allergenic risk (19.2%). Clearly, identifying nearly 1 in 5 putative high-abundance non-allergens as an allergenic risk demonstrates that these bioinformatic algorithms are not fit for purpose as they greatly overestimate risk and impede the use of safe proteins to develop improved crops. This is especially evident since this investigation in combination with previous investigations have found alternative bioinformatic algorithms, including the 1:1 FASTA approach, to be equally sensitive to the sliding-window search for detecting true allergens but with dramatically better selectivity for not falsely detecting low-risk protein sequences.[5–7] Similarly, previous investigations have suggested more selective methods for identifying peptides with potential risk of causing celiac disease.[13] Multifactor bioinformatic criteria have also been suggested with much improved selectivity for detecting known allergens and represent an additional avenue for evaluating the allergenic risk of novel food proteins.[17]

The current results evaluating the selectivity of bioinformatic searches using high-abundance nonallergenic food proteins support past investigations using a comprehensive list of proteins from crops with a low allergenic potential. Together, these findings give realistic estimates of relative false-positive rates and clearly support the superiority of alternative bioinformatic approaches using modern bioinformatic tools (e.g. 1:1 FASTA method).

Disclosure of Potential Conflicts of Interest

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

References

1. Herman RA, Ladics GS. Allergenic sensitization versus elicitation risk criteria for novel food proteins - short communication. Regul Toxicol Pharm. 2018;94:283–85. doi:10.1016/j.yrtph.2018.02.016.
2. Bøgh KL, Madsen CB. Food allergens: is there a correlation between stability to digestion and allergenicity? Crit Rev Food Sci Nutr. 2016;56:1545–67. doi:10.1080/10408398.2013.779569.
3. Privalle L, Bannon G, Herman R, Ladics G, McCLain S, Stagg N, Ward J, Herouet-Guicheney C. Heat stability, its measurement, and its lack of utility in the assessment of the potential allergenicity of novel proteins. Regul Toxicol Pharm. 2011;61:292–95. doi:10.1016/j.yrtph.2011.08.009.
4. Verhoeckx K, Bøgh KL, Dupont D, Egger L, Gadermaier G, Larré C, Mackie A, Menard O, Adel-Patient K, Picariello G. The relevance of a digestibility evaluation in the allergenicity risk assessment of novel proteins. Opinion of a joint initiative of COST action ImpARAS and COST action INFOGEST. Food Chem Toxicol. 2019;129:405–23. doi:10.1016/j.fct.2019.04.052.
5. Cressman RF, Ladics G. Further evaluation of the utility of "Sliding Window" FASTA in predicting cross-reactivity with allergenic proteins. Regul Toxicol Pharm. 2009;54:S20–S25. doi:10.1016/j.yrtph.2008.11.006.
6. Song P, Herman R, Kumpatla S. 1:1 FASTA update: using the power of *E*-values in FASTA to detect potential allergen cross-reactivity. Toxicol Rep. 2015;2:1145–48. doi:10.1016/j.toxrep.2015.08.005.
7. Song P, Herman RA, Kumpatla S. 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. 2014;71:142–48. doi:10.1016/j.fct.2014.06.008.
8. FAO/WHO, Evaluation of allergenicity of genetically modified foods. Report of Joint FAO/WHO Expert Consultation. Rome, Italy: Food and Agriculture Organization of the United Nations; 2001.
9. Ladics GS, Cressman RF, Herouet-Guicheney C, Herman RA, Privalle L, Song P, Ward JM, McClain S. Bioinformatics and the allergy assessment of agricultural biotechnology products: industry practices and recommendations. Regul Toxicol Pharm. 2011;60:46–53. doi:10.1016/j.yrtph.2011.02.004.
10. Herman R, Song P, ThirumalaiswamySekhar A. Value of eight-amino-acid matches in predicting the allergenicity status of proteins: an empirical bioinformatic investigation. Clin Mol Allergy. 2009;7:1–7. doi:10.1186/1476-7961-7-9.
11. Silvanovich A, Nemeth MA, Song P, Herman R, Tagliani L, Bannon GA. The value of short amino acid sequence matches for prediction of protein allergenicity. Toxicol Sci. 2006;90:252–58. doi:10.1093/toxsci/kfj068.
12. EFSA Panel on Genetically Modified Organisms, Naegeli H, Birch AN, Casacuberta J, De Schrijver A, Gralak MA, Guerche P, Jones H, Manachini B, Messéan A, et al. Guidance on allergenicity assessment of genetically modified plants. EFSA J. 2017;15:e04862.
13. Song P, Podevin N, Mirsky H, Anderson J, Delaney B, Mathesius C, Rowe L, Herman RA. Q-X1-P-X2 motif search for potential celiac disease risk has poor selectivity. Regul Toxicol Pharm. 2018;99:233–37. doi:10.1016/j.yrtph.2018.09.021.
14. Pearson WR. An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics. Chapter 3, Unit3.1-Unit3.1. 2013.
15. Krutz NL, Winget J, Ryan CA, Wimalasena R, Maurer-Stroh S, Dearman RJ, Kimber I, Gerberick GF. Proteomic and bioinformatic analyses for the identification of proteins with low allergenic potential for hazard assessment. Toxicol Sci. 2019;170:210–22. doi:10.1093/toxsci/kfz078.
16. Herman RA, Song P. Validation of bioinformatic approaches for predicting allergen cross reactivity. Food Chem Toxicol. 2019;132:110656. doi:10.1016/j.fct.2019.110656.
17. Mirsky HP, Cressman RF, Ladics GS. Comparative assessment of multiple criteria for the in silico prediction of cross-reactivity of proteins to known allergens. Regul Toxicol Pharm. 2013;67:232–39. doi:10.1016/j.yrtph.2013.08.001.