Regulatory Toxicology and Pharmacology 99 (2018) 233–237

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Q-X1-P-X2 motif search for potential celiac disease risk has poor selectivity

Ping Song[a,*], Nancy Podevin[b], Henry Mirsky[b], Jennifer Anderson[b], Bryan Delaney[b], Carey Mathesius[b], Laura Rowe[b], Rod A. Herman[a]

[a] Corteva Agriscience™, Agriculture Division of DowDuPont™, Indianapolis, IN, USA
[b] Corteva Agriscience™, Agriculture Division of DowDuPont™, Johnston, IA, USA

ARTICLE INFO

Keywords:
Celiac disease
Bioinformatics
Motif
Selectivity
GM crops

ABSTRACT

The European Food Safety Authority (EFSA) recently published guidelines for assessment of potential celiac disease risk for newly expressed proteins in genetically modified (GM) crops. This novel step-wise approach prescribes, in part, how to conduct sequence identity searches between a newly expressed protein and known celiac disease peptides including a Q/E-X1-P-X2 amino acid motif. To evaluate the specificity of the recommended sequence identity searches in the context of risk assessment, protein sequences from celiac disease causing crops, as well as from crops not associated with celiac disease, were compared with known HLA-DQ restricted epitopes and searched for the presence of motifs followed by peptide analysis. Searches for the presence of the Q/E-X1-P-X2-motif were found to generate a high proportion of false-positive hits irrelevant to celiac disease risk. Identification of a 9mer exact match between a newly expressed protein and the known celiac disease peptides (recommended by the guideline) along with a supplementary sequence comparisons (suggested by FARRP/AllergenOnline) is considered better suited to more specifically capture the potential risk of a newly expressed protein for celiac disease.

1. Introduction

Genetically modified (GM) crops have been adopted by farmers in many countries at an unprecedented speed. In 2017, worldwide cultivation acreage of GM crops reached a record of 189.8 million hectares, a ~111-fold increase over a period of 22 years since their initial commercial launch in 1996 (ISAAA, 2017). The wide adoption of GM crop cultivation has brought economic and social benefits, and farm sustainability (National Research Council, 2010; Kathage and Qaim, 2012; Klümper and Qaim, 2014). Notably, there have been no safety incidents involving food/feed derived from GM crops, although the debate on the safety of GM crops has continued for many years. One of the most commonly voiced food safety concern is potential allergenicity of introduced proteins. To protect human food safety and public health, international organizations, including the United Nations Food and Agriculture Organization (FAO) and World Health Organization (WHO) and their joint Codex Alimentarius Commission, have issued guidelines on evaluation of allergenicity of foods derived from biotechnology (FAO/WHO, 2001; Codex Alimentarius Commission, 2009).

Despite the existence of rigorous regulations on food/feed derived through modern biotechnology and the absence of safety incidences involving food/feed derived from GM crops, certain regulatory authorities consider that adding requirements will increase the public trust in the risk assessment process. However, such regulation may inadvertently send the message that risks are now believed to be greater than previously thought, when in fact, experience and acquired knowledge has shown negligible risk exists. EFSA (European Food Safety Authority) recently published an additional guideline on the allergenicity assessment of GM crops (Naegeli et al., 2017). According to this guideline, newly expressed proteins in GM crops must be subjected to a step-wise risk assessment for potential risk of non-IgE-mediated allergic responses with particular emphasis on differentiating them from proteins known to provoke adverse effects in persons with celiac disease (CD). Fortunately, the proteins that cause celiac disease are not widely distributed in nature being largely restricted to gluten proteins from wheat (*Triticum aestivum*), gluten-like hordeins from barley (*Hordeum vulgare*), and secalins from rye (*Secale cereal*) and derived hybrids. Oats (*Avena sativa*) are generally considered relatively safe for celiac disease patients although this crop contains gluten-like avenins (Naegeli et al., 2017). Gluten is a composite of storage proteins localized together with starch in the endosperm of various cereal grains. “Gluten” proteins causing celiac disease are limited to the grains listed above. Gluten proteins are characterized by the abundant presence of glutamine and proline which together comprise over 50% of the amino

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

https://doi.org/10.1016/j.yrtph.2018.09.021
Received 30 July 2018; Received in revised form 19 September 2018; Accepted 21 September 2018
Available online 26 September 2018

acids (Wieser, 2007).

In one component of the step-wise approach outlined by EFSA, the sequence identity between a newly expressed protein and celiac disease peptides must be evaluated if knowledge of the protein including source organism, exposure, *etc*., is insufficient to support the history of safe use of that protein. Notably, this must be done without taking into account the source organism for the gene encoding the protein and whether or not it is a known gluten-containing plant (wheat, barley, rye, and oats) (Figure 1 in Naegeli et al., 2017). Such sequence identity evaluation includes a search for a Q/E-X1-P-X2 motif (Q = glutamine; E = glutamic acid; X1 = L [leucine], Q, F [phenylalanine], S [serine], or E; P = proline; X2 = Y [tyrosine], F, A [alanine], V [valine] or Q) present in the HLA-DQ2 restricted epitopes. The restricted epitopes associated with celiac disease, HLA-DQ2 and HLA-DQ8, are nine amino acids long, and the Q/E-X1-P-X2 is present in most, but not all of the HLA-DQ2 restricted epitopes (Sollid et al., 2012). One would expect a motif consisting of only four amino acids to be found in many protein sequences that are not related to celiac disease, thus creating a large number of 9mers requiring further evaluation by HLA-DQ peptide modeling according to the stepwise approach of the risk assessment for celiac disease as outlined in the EFSA guideline.

To fully understand the performance of the sequence identity searches for identifying potential celiac disease risk, especially the Q/E-X1-P-X2 motif search, whole protein amino acid sequences (*in silico* translated from genomic DNA from various food plants and animals) available in NCBI (ftp://ftp.ncbi.nih.gov/genomes/) and all the protein sequences of a given food plant or animal available in the UniProtKB (http://www.uniprot.org/) database were downloaded and used to investigate their sequence identity to HLA-DQ2/DQ8 restricted epitopes. The outcomes were evaluated to determine the specificity of the proposed approach.

2. Material and method

2.1. Protein sequences

Protein sequences from food plants and animals including apple (*Malus domestica*), sunflower (*Helianthus annuus*), tomato (*Solanum lycopersicum*), rice (*Oryza sativa*), potato (*Solanum tuberosum*), soybean (*Glycine max*), banana (*Musa acuminata*), maize (*Zea mays*), cattle (*Bos taurus*), and swine (*Sus scrofa*) were downloaded from the NCBI genome FTP site (ftp://ftp.ncbi.nih.gov/genomes/; as of December 2017). These proteins were *in silico* translated from the coding sequences in the whole genome, which reflects a variety of proteins to which a consumer might be exposed. Similar protein sequences are currently unavailable for wheat, oats, barley, and rye. However, the protein sequences from the genome of red wild einkorn wheat (*Triticum Urartu;* A genome), the contributor of one of the sub-genomes of bread wheat (which is hexaploid; AABBDD*)* was available from a recent sequencing effort, and was used in this investigation. In addition, all the protein sequences available in the UniProtKB database for wheat, barley, oats, rye, sunflower, apple, banana, peanut (*Arachis hypogaea*), potato, swine, chicken (*Gallus gallus*), and cattle were downloaded (as of January 2018). It is noteworthy that the protein sequences downloaded from the UniProtKB database might contain redundant sequences or bias toward certain types of proteins that do not reflect the whole protein set in a given species due to the nature of this database.

2.2. Search for match of HLA-DQ2 and HLA-DQ8 restricted epitopes

The 22 HLA-DQ2 and Four HLA-DQ8 restricted epitopes (9 amino acids long) listed in the EFSA guideline (Naegeli et al., 2017) along with four HLA-DQ2 restricted epitopes downloaded from the ProPepper website (https://propepper.net/epitope-list) were used for sequence identity searches (Table 1). Since the Q/E-X1-P-X2 motif is not present in one of the 22 HLA-DQ2 epitopes in the list provided in the EFSA guideline, only 21 HLA-DQ2 restricted epitopes were used as patterns to scan all the protein sequences for exact 9mer matches using the FUZZPRO program in Emboss Package V6.4.0 (http://emboss.sourceforge.net/). The remaining one plus the four HLA-DQ2 restricted epitopes from ProPepper were used as patterns to scan all the protein sequences for 9mer matches but allowing three mismatches as recommended by the EFSA guideline. The four HLA-DQ8 restricted epitopes were also used as patterns to scan all the protein sequences for 9mer matches allowing three mismatches but excluding "E" (glutamic acid) or "Q" (glutamine) at the first and last position of a 9mer match as recommended by the EFSA guideline.

Table 1
List of HLA-DQ restricted epitopes for sequence identity search.
(Source: Naegeli et al., 2017; ProPepper: https://propepper.net/epitope-list)

Epitope Name	Sequence	Search Pattern
DQ2.5-glia-α1a	PFPQPQLPY	9mer exact match
DQ2.5-glia-α1b	PYPQPQLPY	
DQ2.5-glia-α2	PQPQLPYPQ	
DQ2.5-glia-α3	FRPQQPYPQ	
DQ2.5-glia-γ1	PQQSFPQQQ	9mer match allowing three mismatches
DQ2.5-glia-γ2	IQPQQPAQL	9mer exact match
DQ2.5-glia-γ3	QQPQQPYPQ	
DQ2.5-glia-γ4a	SQPQQQFPQ	
DQ2.5-glia-γ4b	PQPQQQFPQ	
DQ2.5-glia-γ4c	QQPQQPFPQ	
DQ2.5-glia-γ4d	PQPQQPFCQ	
DQ2.5-glia-γ5	QQPFPQQPQ	
DQ2.5-glia-ω1	PFPQPQQPF	
DQ2.5-glia-ω2	PQPQQPFPW	
DQ2.2-glut-L1	PFSQQQQPV	
DQ2.5-glut-L2	FSQQQQSPF	
DQ2.5-hor-1	PFPQPQQPF	
DQ2.5-hor-2	PQPQQPFPQ	
DQ2.5-sec-1	PFPQPQQPF	
DQ2.5-sec-2	PQPQQPFPQ	
DQ2.5-ave-1	PYPEQQEPF	
DQ2.5-ave-1b	PYPEQQQPF	
DQ2-LMW-glutenin	FSQQQQQPL	9mer match allowing three mismatches
DQ2.5-hor-2, DQ2.5-sec-2	PQPQQPFPQ	
DQ2.2-glut-L1, DQ2.5-glut-L1	PFSQQQQPV	
DQ2-sec	PQQSFPQQP	
DQ8-glia-α1	QGSFQPSQQ	9mer match allowing three mismatches
DQ8-glia-γ1a	QQPQQPFPQ	excluding Q/E at first and last position
DQ8-glia-γ1b	QQPQQPYPQ	
DQ8-glut-H1	QGYYPTSPQ	

2.3. Search for the presence of HLA-DQ2 motif and peptide analysis

The Q/E-X1-P-X2 motif present in the HLA-DQ2 restricted epitopes was used as a pattern to identify its presence in all the protein sequences using the FUZZPRO program. Since a HLA-DQ2 restricted epitope is typically nine amino acids long with the motif generally starting at position 4 or 6, all of the 9mers containing motifs starting at position 4 or 6 were extracted from all the protein sequences containing the motif, followed by scanning for the presence of a proline duplex (PP), a null motif "ELPF", positively charged amino acids (lysine, histidine, arginine) at positions 1, 4, 6, 7, and/or 9, glycine (G) or alanine (A) before the motif, serine after the motif (only applicable to the 9mers with motif starting at position 4), and a trypsin site (Naegeli et al., 2017). The presence of these amino acids or a trypsin site dramatically reduces the risk that the 9mer is a potential restricted epitope for celiac disease.

3. Results

When the HLA-DQ2 restricted epitopes were used to search for exact matches in the proteins from food plants and animals, there were no