Food and Chemical Toxicology 71 (2014) 142–148

Contents lists available at ScienceDirect

Food and Chemical Toxicology

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

Evaluation of global sequence comparison and one-to-one FASTA local alignment in regulatory allergenicity assessment of transgenic proteins in food crops

Ping Song [*], Rod A. Herman, Siva Kumpatla

Dow AgroSciences, 9330 Zionsville Road, Indianapolis, IN 46268, United States

ARTICLE INFO

Article history:
Received 7 May 2014
Accepted 10 June 2014
Available online 19 June 2014

Keywords:
Bioinformatics
Allergen
Cross-reactivity
Transgenic
Protein

ABSTRACT

To address the high false positive rate using >35% identity over 80 amino acids in the regulatory assessment of transgenic proteins for potential allergenicity and the change of E-value with database size, the Needleman–Wunsch global sequence alignment and a one-to-one (1:1) local FASTA search (one protein in the target database at a time) using FASTA were evaluated by comparing proteins randomly selected from Arabidopsis, rice, corn, and soybean with known allergens in a peer-reviewed allergen database (http://www.allergenonline.org/). Compared with the approach of searching >35%/80aa+, the false positive rate measured by specificity rate for identification of true allergens was reduced by a 1:1 global sequence alignment with a cut-off threshold of ≧30% identity and a 1:1 FASTA local alignment with a cut-off E-value of ≦1.0E−09 while maintaining the same sensitivity. Hence, a 1:1 sequence comparison, especially using the FASTA local alignment tool with a biological relevant E-value of 1.0E−09 as a threshold, is recommended for the regulatory assessment of sequence identities between transgenic proteins in food crops and known allergens.

1. Introduction

Since the first commercial launch in 1996, the acreage of planted transgenic crops has consistently increased and reached 170 million hectares in 2012 worldwide (Clive, 2013). Transgenic crop products are subject to an intensive safety assessment process before commercial launch. Over the course of the past dozen years, a weight-of-evidence based transgenic-protein safety assessment approach has been developed and widely adopted. Comparison of transgenic proteins with known allergens for amino-acid sequence similarity is one of the critical components of the system to ensure the safety of the transgenic protein contained in transgenic food-crop products. In 2001, the Food and Agricultural Organization/World Health Organization suggested that >35% identity over an 80 amino acid stretch (>35%/80aa) be used as minimum threshold in such sequence comparisons to identify the potential IgE cross reactivity of a transgenic protein with known allergens (FAO/WHO, 2001). Ever since, a search for >35%/80aa has become the benchmark algorithm used in the regulatory assessment of transgenic protein safety.

According to the WHO/FAO guidelines, the transgenic protein (query) needs to be parsed into all the sequentially overlapping fragments of 80 amino acids, followed by a search of each fragment for >35% identity using a local alignment algorithm such as FASTA (Pearson and Lipman, 1988). One of the issues when using this >35%/80aa sliding window search methodology is the high rate of false positives. For instance, one study reported that 20.6% of known proteins in GenBank non-redundant (NR) Entrez Protein database (August 2008 update), 15.6% of rice proteins, and 12.6% of human proteins would be classified as potential cross-reactive “allergens” using this approach (Guarneri, 2010), while the expected percentage of real allergens at the time of the publication was only about 0.2% of the known proteins based on a peer reviewed allergen database (Version 9, http://www.allergenonline.org/). Similar results, 6.6%, 11.4%, and ~19%, were observed in another study using 1102 hypothetical corn ORFs (⩾80 aa), 907 randomly selected proteins, and 89 randomly selected corn proteins, respectively, as queries (Ladics et al., 2007). These authors concluded that a conventional FASTA search using the whole protein sequence is superior to use of a sliding window search (Ladics et al., 2007). To avoid high false positive rates associated with the >35%/80aa+ sequence identity, many alternative search routines have been explored, including a conventional FASTA search using the whole protein sequence as a query (Ladics et al., 2007;

[*] Corresponding author. Tel.: +1 (317)337 3434; fax: +1 (317)337 7054.
E-mail address: psong@dow.com (P. Song).

http://dx.doi.org/10.1016/j.fct.2014.06.008
0278-6915/© 2014 Elsevier Ltd. All rights reserved.

Cressman and Ladics, 2009) and a combination of a FASTA search with E-values to evaluate the quality of an alignment between a query protein and known allergens (Cressman and Ladics, 2009). Results indicated that a conventional FASTA search provides sufficient sensitivity with reduced false positive rates. In one study, a threshold E-value of 3.9E−07 combined with a conventional full length FASTA search was determined to have sufficient stringency to reject the majority of false positive and composition-based anomalies (Silvanovich et al., 2009). A recent study indicated that a BLAST search of a full length protein sequence using an E-value upper limit of 0.1 achieved the correct recognition of all known allergen, reduced the false positive rate, and achieve 100% sensitivity (Guarneri, 2010).

At present, almost all searches of transgenic proteins to identify their sequence similarity with known allergens for assessment of cross reactivity use a local alignment algorithm. However, E-values change with changes in the database size (addition of newly discovered allergen sequences), which creates an issue for assigning a fixed E-value cutoff for assessing cross-reactive risk, especially in a regulatory context. For a cross-reaction to take place between a protein and a known allergen, it is likely that in excess of 50–70% sequence identity over a significant span of the target protein and allergen is needed (Aalberse, 2000). In a review of sequence identities among allergenic and non-allergenic homologs of pollen allergens, it was found that the prerequisite for allergenic cross-reactivity between proteins was a sequence identity of at least 50% across the length of the protein (Radauer and Breitender, 2006). With this in mind, we used the Needleman–Wunsch (Needleman and Wunsch, 1970) global alignment algorithm and a 1:1 (one protein in the target database at a time) conventional full length FASTA search to compare the sequences of randomly selected proteins from food and nonfood crops with each known allergen in a peer-reviewed database. This was followed by an evaluation of false positives and false negatives which allowed the specificity and sensitivity of the methods to be estimated, with the intent to evaluate the application of these two approaches and criteria in the bioinformatics analysis of transgenic proteins for potential cross-reactive allergenicity.

2. Materials and method

2.1. Plant protein sequences

Reviewed proteins from Arabidopsis, rice, corn, and soybean in the UniProKB (Protein Knowledgebase) (http://www.uniprot.org/, up to date as of December 20, 2012) were randomly selected and downloaded for analysis. According to the database documentation, reviewed proteins are the entries that have gone through manual annotation consisting of six major mandatory steps: (1) sequence curation, (2) sequence analysis, (3) literature curation, (4) family-based curation, (5) evidence attribution, and (6) quality assurance and integration of completed entries (http://www.uniprot.org/faq/45). At the time of this investigation, 436 Arabidopsis sequences were randomly selected from the reviewed Arabidopsis protein sequences, 191 rice sequences were randomly selected from the reviewed rice protein sequences, and 323 and 379 of total reviewed corn and soybean protein sequences were selected.

2.2. Allergen database

The peer-reviewed allergen database (1603 entries, Version 12, released in February 2012) from FARRP (Food Allergy Research and Resource Program, University of Nebraska, http://www.allergenonline.org/) was used for the entire study. In addition, allergens included in the Allergome (http://www.allergome.org/index.php), but not in FARRP database were treated as allergens in the evaluation of sensitivity and specificity.

2.3. Search algorithm and hit extraction

Designated Perl scripts were written to perform conventional FASTA searches, 1:1 global alignments (Needleman–Wunsch), and 1:1 FASTA searches, followed by extraction of hits that met the criteria set in each search algorithm. In all searches, hits from the exact matches (100% identity in full length) between a query protein and an allergen entry in the database were removed from the total number of hits per query protein so that sensitivity (ability to detect true unknown allergens) could be evaluated with a minimum of bias.

The FASTA35 program was used to search for >35% identity over 80 amino acids or longer (>35%/80aa+) between a query (whole sequence) and known allergens in the aforementioned peer-reviewed allergen database with default search parameters (Matrix = BLOSUM50; Gap Penalties = −12/−2; ktup = 2; Expectation = 10) (Pearson and Lipman, 1988), followed by extraction of hits. To ensure that high identity matches over a short stretch (for example, 80% over 60 amino acids) were not overlooked, a calculation, (Identity% × number of overlapped amino acids)/80, was implemented as a conversion to check the criterion of >35%/80aa+ when the FASTA alignment (overlapped amino acids) were less than 80 amino acids (EFSA, 2010). If a query was less than 29 amino acids, a FASTA search was not performed since this is the minimum query length to achieve of >35%/80aa identity (28/80 = 35). Alignments with >35%/80aa+ were extracted as hits which were examined for biological relevance by manual review of the UniproKB database annotations on the query sequences.

For a 1:1 comparison using a local alignment algorithm, a FASTA search of each query against each single allergen in the database alone (1:1 FASTA search) was performed with default search parameters. Alignments with an E-value equal to or less than 1.0E−03, 1.0E−04, 1.0E−05, 1.0E−06, 3.9E−07, 1.0E−08, 1.0E−09, and 1.0E−10, were extracted as hits. All the extracted hits were examined for biological relevance by manual review of the UniproKB database annotations on the query sequences.

The Needleman–Wunsch program in Emboss suite (http://emboss.sourceforge.net/) was used for global alignments between a query and each allergen in the database with default settings (matrix = BLOSUM62; gap penalty = 10; extend penalty = 0.5) (Needleman and Wunsch, 1970). Alignments generating defined identity percentages of ≧30, 35, 40 or 50 were extracted and manually examined for biological relevance by manual review of the UniproKB database annotations on the query sequences.

2.4. Sensitivity and specificity

The sensitivity and specificity of each search method were calculated using the following formulas (Baldi et al., 2000):

Sensitivity (%) = [number of true positives/(number of true positives + number of false negatives)] × 100;
Specificity (%) = [number of true negatives/(number of true negatives + number of false positives)] × 100;

2.5. Test sequences

As described in a previous study (Ladics et al., 2007), birch pollen allergen Bet v 1a (GenBank gi: 159162097) and several cross-reacting allergens, i.e. carrot (Dau c 1; GenBank gi: 1335877); celery (Api g 1; GenBank gi: 1346568); apple (Mal d 1; GenBank gi: 1346478), cherry (Pru a 1; GenBank gi: 44409451), and pear (Pyr c 1; GenBank gi: 3044216), were used to evaluate false negatives using a cut-off E-value of ≦1.0E−09 in a 1:1 FASTA search and ≧30% identity in a global-alignment comparison as an indication of cross reactive proteins. The Bet v 1 protein sequence was compared with each of those proteins by a 1:1 FASTA and global alignment using Needleman–Wunsch algorithm as an assessment of known cross-reactive allergens. Another allergen, Vig r 6 (GenBank gi: 75217171), known to share weak sequence similarity but which is cross-reactive with Bet v 1a (Guhsl et al., 2013), was compared with 60 Bet v 1-related allergens by 1:1 FASTA and global alignment using Needleman–Wunsch algorithm as an addition assessment. Finally, 25 amino acids from position 500 to 524 from a peanut allergen Ara h 1 (GenBank gi: 1168390) were inserted into a putative non-allergenic protein (GenBank gi: 2582631; an acetate auxotroph from bacteria, *Methanococcus maripaludis*) by following an example in a previous publication (Ladics et al., 2007), and this sequence was tested using a cut-off E-value of ≦1.0E−09 in a 1:1 FASTA search.

3. Results

Among all the search results, a FASTA search for >35%/80aa+ and a 1:1 FASTA with cut off E-value of ≦1.0E−03 generated the highest percentage of proteins with hits across all the plant proteins used in the study (Table 1). A 1:1 global alignment, regardless of the identity cut-off value, produced a low number of proteins with hits, being even less than that observed in the 1:1 FASTA search with cutoff E-value of ≦1.0E−10. A 1:1 global alignment with ≧50% identity generated the lowest number of proteins with hits, but still higher than the number of true allergens among the searched plant proteins. Simply looking at the overall rate of hits is not sufficient for evaluating the value of a search algorithm.