



# Interpreting the biological relevance of bioinformatic analyses with T-DNA sequence for protein allergenicity

B. Harper\*, S. McClain, E.W. Ganko

Syngenta Crop Protection, LLC, Research Triangle Park, NC 27709, USA

## ARTICLE INFO

### Article history:

Received 3 January 2012

Available online 2 June 2012

### Keywords:

Transgenic  
Crops  
T-DNA  
Allergenicity  
Bioinformatics  
Sequence  
Biotechnology  
Protein  
Genome  
Proteome

## ABSTRACT

Global regulatory agencies require bioinformatic sequence analysis as part of their safety evaluation for transgenic crops. Analysis typically focuses on encoded proteins and adjacent endogenous flanking sequences. Recently, regulatory expectations have expanded to include all reading frames of the inserted DNA. The intent is to provide biologically relevant results that can be used in the overall assessment of safety. This paper evaluates the relevance of assessing the allergenic potential of all DNA reading frames found in common food genes using methods considered for the analysis of T-DNA sequences used in transgenic crops. FASTA and BLASTX algorithms were used to compare genes from maize, rice, soybean, cucumber, melon, watermelon, and tomato using international regulatory guidance. Results show that BLASTX for maize yielded 7254 alignments that exceeded allergen similarity thresholds and 210,772 alignments that matched eight or more consecutive amino acids with an allergen; other crops produced similar results. This analysis suggests that each nontransgenic crop has a much greater potential for allergenic risk than what has been observed clinically. We demonstrate that a meaningful safety assessment is unlikely to be provided by using methods with inherently high frequencies of false positive alignments when broadly applied to all reading frames of DNA sequence.

© 2012 Elsevier Inc. All rights reserved.

## 1. Introduction

Guidance for food safety assessment of transgenic crops includes a thorough bioinformatic analysis of the inserted DNA sequence and adjacent endogenous flanking sequences. Bioinformatics typically serve as one component in a larger weight-of-evidence safety assessment for identifying allergenic potential of transgenic crops (Codex Alimentarius Commission, 2009). The highest priority bioinformatic assessment of potential allergenicity is an alignment comparison between the novel transgenic protein(s) and known or putative allergens. Significant similarity between known allergen sequences can be identified by a high level of shared protein sequence; novel proteins that are highly similar to allergens may share allergy cross-reactivity. Comparisons of putative open reading frames (ORFs) spanning the genomic flanking DNA to T-DNA insert junctions are also performed for the purpose of identifying potential novel fusion proteins of concern. In this case, the ORF

sequences are putatively translated to amino acid sequence and the comparison to known allergens is made bioinformatically. Some regulatory agencies have expanded their expectations for bioinformatic analyses to also include all reading frames within the T-DNA (Fig. 1).

Screening novel proteins used in transgenic crops attempts to assess the risk of potential allergy cross-reactivity (Jenkins et al., 2005; Vieths et al., 2002). The regulatory guidance on threshold criteria for identifying relevant allergen similarity is greater than 35% shared amino acid identity over 80 or more amino acids (Codex Alimentarius Commission, 2009). This is considered conservative guidance and below the level of shared identity typically observed among cross-reacting allergens. While cross-reactivity can occur for allergens with a ~50% shared identity, it typically occurs at a level of shared identity of 70% or greater (Aalberse, 2000). Identifying similarity to allergens by searching for short linear sequence matches is another sequence comparison method typically included with the full length similarity assessment (typically 8 or more contiguous and exact matching amino acids) (Hileman et al., 2002; Silvanovich et al., 2009). However, there would only be potential biological relevance if the portion of the alignment identified were known to be an epitope in the allergen sequence. As it stands, there is neither an identified absolute minimum length for known allergen IgE binding epitopes nor a complete database of known epitopes (Bannon, 2004) and its utility for food

**Abbreviations:** EST, expressed sequence tag; FARRP, Food Allergy Research and Resource Program; ORF, open reading frame; T-DNA, transfer DNA; UTR, untranslated region.

\* Corresponding author. Address: Syngenta Biotech Inc., P.O. Box 12257, Research Triangle Park, NC 27709-2257, USA. Fax: +1 919 226 7464.

E-mail addresses: [brian.harper@Syngenta.com](mailto:brian.harper@Syngenta.com) (B. Harper), [scott.mcclain@Syngenta.com](mailto:scott.mcclain@Syngenta.com) (S. McClain), [eric.ganko@Syngenta.com](mailto:eric.ganko@Syngenta.com) (E.W. Ganko).

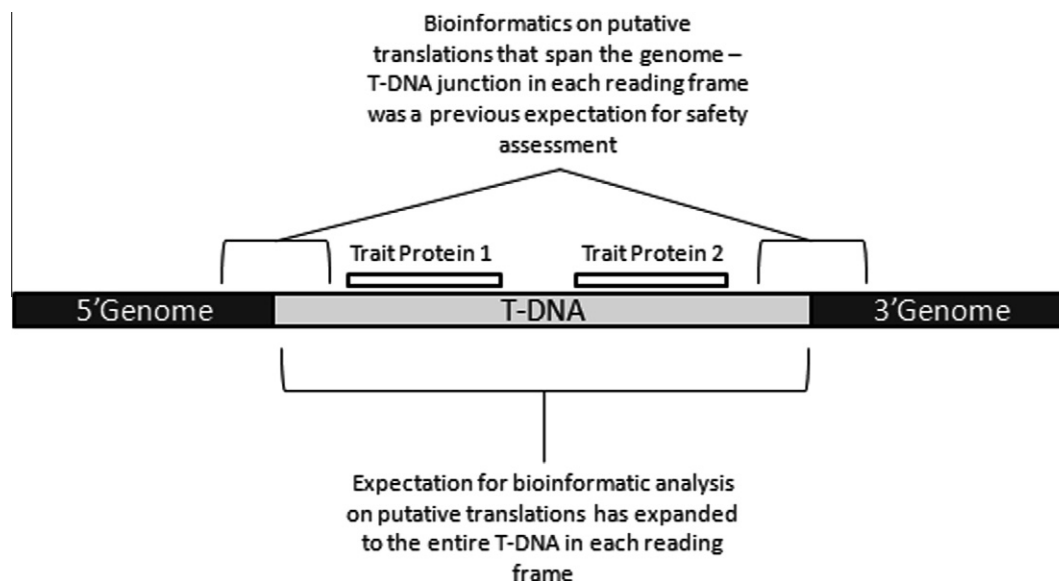


Fig. 1. Location of sequences used in T-DNA and T-DNA to genome junction bioinformatic analysis.

safety assessments has been questioned (Goodman et al., 2008). Despite the conservative nature of the threshold criteria, additional evaluation for potential allergic cross-reactivity is an expectation for novel proteins that have positive matches (as an example of a study design see Ladics et al., 2006).

In addition to threshold criteria, the source and relevance of the input (query) sequences are important to consider prior to generating and interpreting alignment output (Ladics et al., 2011). A basic premise behind bioinformatic comparisons is that the query sequence is a known protein with a biologically relevant level of expression (and thus possible exposure). For results to have a biologically meaningful context there should be evidence of *in planta* expression for the protein whose amino acid sequence is being evaluated for allergen similarity. Some regulatory agencies have indicated an expectation for bioinformatic assessment of allergen similarity for non-coding DNA sequences as part of the food safety assessment. This includes an assessment of alternative (all) reading frames of the T-DNA insert. Depending on regional regulatory requirements, ORFs may be defined by a minimum (or no) size, and bracketed by either start and stop codons or only stop codons. In this paper an ORF is defined as any sequence located between two stop codons in any reading frame. However, this oversimplified definition of an ORF ignores the complexity of eukaryotic gene expression and is therefore a very conservative approach for extrapolating any risk for putative amino acid sequences derived from these ORFs. Although putative ORFs can be theoretically translated to amino acid sequence in all potential reading frames *in silico*, the majority has a very low likelihood of *in planta* protein expression as they are often relatively short and lack start codons with the appropriate transcriptional and translational regulatory sequences that control gene expression. Therefore, a high rate of false positive bioinformatic alignments would be expected to overwhelm any safety value that may be gained from trying to evaluate DNA sequence that is not known or even likely to be transcribed and translated.

Attempting to assess the entire T-DNA sequence would require (1) a selection of theoretical translations from ORFs that may be based on a minimum (or no) size, location within the T-DNA, and location between start or stop codons, or (2) selection of the entire T-DNA sequence ignoring stop or start codons as in an alignment algorithm such as BLASTX. It becomes problematic to compare any translated ORFs against a database of allergen sequences using

current threshold criteria (Codex Alimentarius Commission, 2009) because the bioinformatic methods/criteria are themselves conservative with a high frequency of false positive alignments. Additionally, the potential for ORF translation to protein derived from all reading frames of the T-DNA is in most cases a purely *in silico* assumption. Therefore, it is questionable that the output from the bioinformatic search would provide biologically meaningful alignments from a food safety standpoint.

In this paper genomes from seven commonly consumed plant food sources (maize, rice, soybean, cucumber, melon, watermelon, and tomato) were used to evaluate the allergen alignment criteria of greater than 35% shared amino acid identity over 80 or more amino acids and 8 contiguous and exact matching amino acids. The premise is that application of bioinformatics currently used for well characterized novel proteins would likely over estimate the frequency of matches when assessing all reading frames. We conducted the bioinformatic analyses of plant food gene sets using either the FASTA or BLASTX algorithm (Fig. 2). FASTA and BLASTX can both be used to compare any two sets of sequences and results from each are discussed herein. The BLASTX algorithm inherently contains a straight forward approach for both translating DNA into proteins sequence and searching for protein sequence similarity and so was a good candidate for evaluation. FASTA is commonly used for bioinformatic allergenicity assessments and so was also included. Translations of gene sets from the food crops were compared against the allergen database from the FARRP AllergenOnline database (FARRP, 2011) in order to determine a total number of alignments to known allergens that exceed Codex Alimentarius guidance for novel proteins. Both BLAST and FASTA are discussed regarding the interpretation of the output in a side-by-side format, but our intent was not to provide an intensive review of the statistical features of BLAST and FASTA. Rather, our overall objective was to determine the level of false positive alignments and provide a context for the likelihood that relevant safety information can be gleaned from either approach.

## 2. Methods

### 2.1. Query sequence sets

Seven crop gene sequence sets were used in the bioinformatic alignment searches for allergen similarity. The first was the filtered

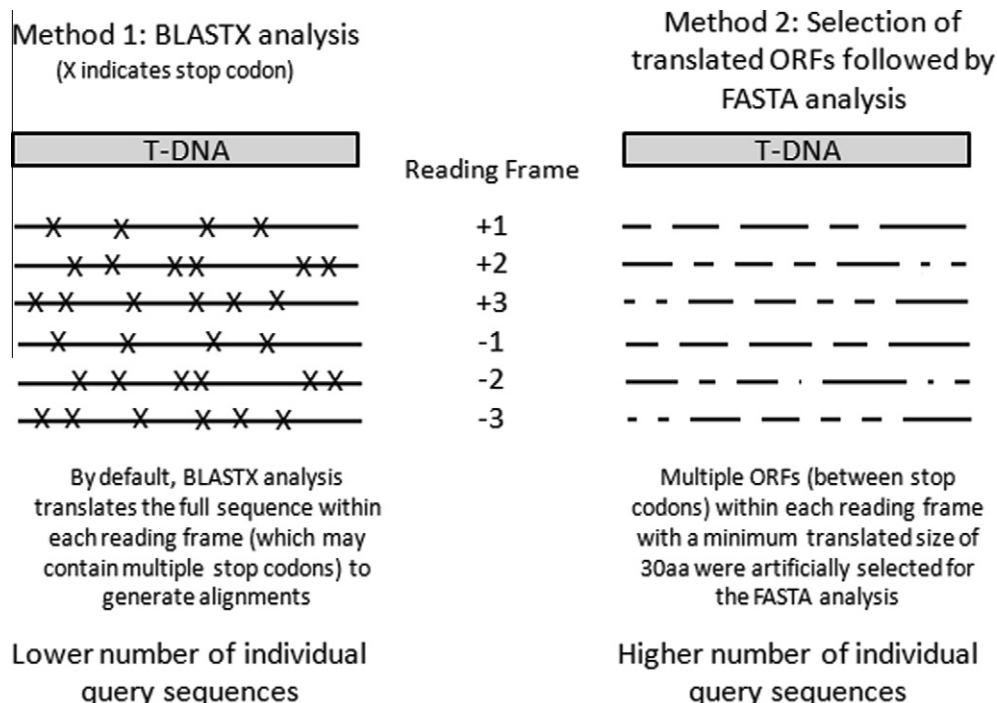


Fig. 2. BLASTX and FASTA methods used for bioinformatic analysis.

maize gene (DNA) set from [www.MaizeSequence.org](http://www.MaizeSequence.org) (version 5b.60/RefGen\_v2) consisting of 39,656 gene (nucleotide) sequences. The filtered set from [www.MaizeSequence.org](http://www.MaizeSequence.org) is a high quality gene set that has removed pseudogenes, transposable element encoded genes, and low confidence hypothetical models. This gene set contains gene exons, introns, and untranslated (but transcribed, e.g. 5' UTR) regions from the maize genome, but should not include promoters, terminators and other non-transcribed regions of the genome. Although this maize gene set does not include all gene regulatory elements as would be included in a full T-DNA analysis, it is relevant to this study because these sequences represent gene rich regions as would be contained in a T-DNA and would also contain sequences that are transcribed and contained in pre-mRNAs. In total, the query set for maize consists of each gene sequence and all five associated alternative reading frame sequences.

Similar gene sequence sets were collected for use in this analysis including rice, soybean, cucumber, melon, watermelon, and tomato. The rice (*Oryza sativa japonica*) gene set was obtained from the MSU Rice Genome Annotation Project Database, release 6.1, and consisted of 57,624 genes (Ouyang et al., 2007). The cucumber (*Cucumis sativus* Gy14) gene set was obtained from the Joint Genome Institute's Phytozome (<http://www.phytozome.net>). This sequence set consisted of 21,872 genes in the Fall 2010 release. The soybean (*Glycine max*) gene set was obtained from Phytozome, release Glyma1, and consisted of 46,367 genes (Schmutz et al., 2010). The melon (*Cucumis melo*) gene set was obtained from the Melonomics Consortium (Spain) and consisted of 31,925 genes. The watermelon (*Citrullus lanatus*) gene set was obtained from the Watermelon Sequencing Consortium (National Engineering Research Center for Vegetables, Beijing, China) and consisted of 23,435 genes. The tomato (*Solanum lycopersicum*) gene set was obtained from the International Tomato Genome Sequencing Project at the Sol Genomics Network (Bombarley et al., 2010), release SL2.31, and consisted of 35,802 gene sequences.

One additional sequence set of 63,299 maize proteins was analyzed separately from gene sets discussed above. Protein sequences

corresponding to the transcripts of the above maize gene set (5b.60/RefGen\_v2) were obtained ([www.maizesequence.org](http://www.maizesequence.org)) for the purposes of producing a sequence set that better represents known proteins in maize as opposed to putative translations using reading frames of a gene derived from a nucleotide database. Most of these protein sequences have EST evidence for expression *in planta* and may include isoforms and/or splice variants. The objective was to use these sequences in a FASTA analysis and compare the output to the maize gene sequence results.

## 2.2. Allergen database

All sequence sets consisting only of non-transgenic, endogenous sequences were compared against the Food Allergy Research and Resource Program Protein AllergenOnline database, version 11.0 (FARRP, 2011) with the FASTA algorithm using a filter for minimum ORF size (30 aa) or the BLASTX algorithm. The FARRP AllergenOnline database contains the amino acid sequences of known and putative protein allergens. It is a curated, peer-reviewed database containing proteins identified as food allergens, respiratory allergens, allergenic venom proteins, contact allergens, gliadins, and glutenins. Entries were compiled primarily from searches of publicly available protein databases using the National Center for Biotechnology Information (NCBI) Entrez<sup>®</sup> search and retrieval system (NCBI, 2011). The NCBI dataset was screened for entries with text descriptions associated with allergy or celiac disease; duplicate entries were removed, and additional entries were identified from publications. The list of candidate entries was then curated by an international panel of allergy experts who reviewed published clinical and laboratory evidence to support the candidate sequences as allergens. Proteins are classified as known or putative allergens according to predetermined criteria set by the FARRP expert review panel. The 2011 version of the FARRP AllergenOnline database contained 1491 non-redundant entries. For all alignments in this paper known endogenous allergens were removed from the analyses to avoid endogenous allergen sequences from aligning to

themselves. For soybean, this included allergens from both *Glycine max* and *Glycine soja* (wild soybean).

### 2.3. Alignment comparison methods

Two separate sequence comparison methods were used with each of the gene sets to produce alignments against FARRP allergens. Amino acid alignments produced from both the BLASTX and FASTA (Pearson, 2000) methods were evaluated for the number of alignments that had greater than 35% shared identity over 80 or greater amino acids in length (termed “35%/80 aa” in the following text). A search with either translated ORFs (30 amino acid minimum) as prepared for FASTA, or entire gene sequences as prepared for BLASTX, was conducted to determine exact matches of eight contiguous amino acids to known or putative allergen sequences (termed “8 aa” in the following text). Since the native input sequence for BLASTX is nucleotide sequence and not amino acid sequence the gene sets were prepared by translating in each reading frame with a minimum size of eight amino acids prior to performing the search for eight or more amino acid alignments.

BLASTX (v2.2.19) was used with all gene sets against the FARRP AllergenOnline database with the BLOSUM62 scoring matrix (−11 gap creation penalty, −1 gap extension, with complexity filter). As is default in the BLASTX algorithm, stop codons in the query sequence were treated as gaps in the subject sequence alignment. Gaps and mismatches contributed to the overall length of the alignment for determination of percent identity. Full alignment lengths rather than individual ORFs (i.e. only sequences between stop codons) were considered for the BLASTX analysis.

The FASTA (v3.4t11/FASTA34) search algorithm was also used for all gene sets and the maize protein set with the BLOSUM50 scoring matrix (−11 gap creation penalty, −1 gap extension, no complexity filter). The (translated) sequence inputs for the FASTA analyses were created by selecting sequences in all reading frames between stop codons (TAA, TAG, TGA) within each gene set that had a putative ORF of 30 amino acids or greater. Although ORFs less than 80 aa were used, ORFs not producing an alignment that was greater than 80 aa in length could not break the dual 35%/80 aa threshold.

For both FASTA and BLASTX methods percent identity and alignment length statistics were pulled from each of the search algorithm outputs therefore no calculation for individual 80 aa “windows” within each alignment was made. Fig. 3 illustrates specifically how BLASTX alignments are assessed for the 35% shared identity over 80 or greater amino acids in length criteria. Though there are much more sophisticated profiling algorithms for identifying taxonomic relatedness (Altschul et al., 1997), the BLASTX and FASTA algorithms were selected because they are commonly used to assess novel food proteins and most directly support the global safety guideline requirements. BLASTX also offered the most

straight forward approach as the translation in each reading frame is inherent to the search algorithm.

## 3. Results

### 3.1. BLASTX allergen search

Table 1 illustrates the BLASTX results for maize, rice, soybean, cucumber, melon, watermelon, and tomato along with the number of unique genes with alignments exceeding the criteria and the number of known or putative allergens in the FARRP allergen database. BLASTX analysis of the maize gene sequence set produced 7254 alignments where the sequences translated *in silico* showed >35%/80 aa shared identity with an allergen. Out of the 7254 there were 1000 unique maize genes represented in coding or alternate reading frames which had one or more alignments. The maize gene sequence set also produced 210,772 matches of 8 aa representing 11,252 unique maize gene sequences (or 28.4% of all maize genes). Despite having no identified allergens in the FARRP AllergenOnline database, cucumber and watermelon produced 5297 and 5632 >35%/80 aa alignments and 127,796 and 106,307 matches of 8 aa, respectively. Soybean had the largest number of >35%/80 aa alignments, totaling 12,341.

### 3.2. FASTA allergen search

FASTA analysis using ORFs with a minimum translation size of 30 amino acids also produced a large number of bioinformatic alignments that exceeded criteria for allergen sequence similarity. Table 2 illustrates the FASTA results for maize, rice, soybean, cucumber, melon, watermelon, and tomato along with the number of unique genes having alignments that exceed the criteria and the number of known or putative allergens in the FARRP allergen database. Rice produced the greatest number of alignments exceeding the >35%/80 aa criteria with 27,669 total alignments derived from 10,402 unique rice genes. The rice gene sequence set also produced 308,040 matches of 8 aa derived from 15,200 unique rice gene sequences (or 26.4% of all rice genes). For the FASTA results, once again both cucumber and watermelon had a high number of sequences exceeding criteria without having any recorded allergens in the FARRP AllergenOnline database.

In summary, regardless of the sequence comparison method or query sequence set all results produced a large number of alignments relative to the number of known allergens from each plant species.

### 3.3. FASTA and BLASTX methods

The focus of our study was to determine the rate of false positives for both BLASTX and FASTA (with ORF translation) methods

```

Score = 203 bits (516), Expect = 3e-58
Identities = 107/120 (89%), Positives = 110/120 (92%), Gaps = 6/120 (5%)
Frame = +1

Query 1  TNPALREEMR*IQFNDMNSALTTAIP*LFAMVQNYQVPLLS*VYVQAANHDVHLSVLRDV 180
          TNPAL EEMR IQFNDMNSALTTAIP LF VQNYQVPLLS VYVQAAN +HLSVLRDV
Sbjct 122 TNPALTEEMR-IQFNDMNSALTTAIP-LFT-VQNYQVPLLS-VYVQAAN--LHLSVLRDV 175

Query 181 SVFGQRWGFDAATINSRYNDLTRLIGNYTDHAVRWYNTGLERVWGPDSRDWIRYNQFRRE 360
          SVFGQRWGFDAATINSRYNDLTRLIG YTD+AVRWYNTGLERVWGPDSRDW+RYNQFRRE
Sbjct 176 SVFGQRWGFDAATINSRYNDLTRLIGTYTDAVRWYNTGLERVWGPDSRDWVRYNQFRRE 235
  
```

**Fig. 3.** Assessment of BLASTX alignments for 35%/80 aa criteria. In the example BLASTX alignment below (between two non-allergens) there are 3 stop codons (\*) in the query sequence, each creating a gap in the subject sequence and extending the alignment length; likewise, there are 3 mismatches also extending the alignment length. Taken together, the total alignment length is 114 aa + 3 stop codons + 3 mismatches = 120 aa. The percent identity is 89% (107/120 as shown below) and therefore had this been an ORF alignment to an allergen it would be considered a hit as it is both greater than 35% identity and over 80 aa in length.

**Table 1**

Number of allergen alignments from food crops using BLASTX analysis against the FARRP allergen database (allergens specific to each crop are not counted<sup>a</sup>).

| Organism gene set | No. of alignments >35%/80 aa | Unique genes with alignments >35%/80 aa | No. of 8 aa alignments | Unique genes with 8 aa alignments | Allergen sequences in FARRP database |
|-------------------|------------------------------|-----------------------------------------|------------------------|-----------------------------------|--------------------------------------|
| Maize             | 7254                         | 1000                                    | 210,772                | 11,252                            | 24                                   |
| Rice              | 7343                         | 1065                                    | 330,552                | 16,578                            | 20                                   |
| Soybean           | 12,341                       | 1434                                    | 308,400                | 8607                              | 38                                   |
| Cucumber          | 5297                         | 604                                     | 127,796                | 4563                              | 0                                    |
| Melon             | 4830                         | 570                                     | 193,574                | 6022                              | 8                                    |
| Watermelon        | 5632                         | 610                                     | 106,307                | 4120                              | 0                                    |
| Tomato            | 7741                         | 856                                     | 244,113                | 4564                              | 6                                    |

<sup>a</sup> For all alignments known endogenous allergens were removed from the analyses to avoid endogenous allergen sequences from aligning to themselves.

**Table 2**

Number of allergen alignments from food crops and using FASTA analysis against the FARRP allergen database using putative open reading frames (between stop codons, minimum size 30aa) as an input (allergens specific to each crop are not counted<sup>a</sup>).

| Organism gene set | No. of alignments >35%/80 aa | Unique genes with alignments >35%/80 aa | No. of 8 aa alignments | Unique genes with 8 aa alignments | Allergen sequences in FARRP database |
|-------------------|------------------------------|-----------------------------------------|------------------------|-----------------------------------|--------------------------------------|
| Maize             | 20,326                       | 7120                                    | 186,522                | 10,332                            | 24                                   |
| Rice              | 27,669                       | 10,402                                  | 308,040                | 15,200                            | 20                                   |
| Soybean           | 17,011                       | 2811                                    | 269,434                | 7005                              | 38                                   |
| Cucumber          | 7706                         | 981                                     | 115,073                | 3809                              | 0                                    |
| Melon             | 7434                         | 980                                     | 169,466                | 4840                              | 8                                    |
| Watermelon        | 8275                         | 1057                                    | 99,019                 | 3477                              | 0                                    |
| Tomato            | 10,202                       | 1383                                    | 172,265                | 3640                              | 6                                    |

<sup>a</sup> For all alignments known endogenous allergens were removed from the analyses to avoid endogenous allergen sequences from aligning to themselves.

as two approaches for aligning gene datasets with protein databases. With the results presented here one is tempted to compare the two methods however, it is difficult to directly compare results between the BLASTX and FASTA algorithms due to variation of input sequence, reporting of multiple versus single local alignments, and scoring matrices. In particular for the 8 aa search, the minimum ORF size of the FASTA search (30 aa) would by default not contain all possible 8 aa segments as the BLASTX query set would.

That said, while the FASTA method (using a minimum size of 30 for ORF sequences) resulted in more alignments exceeding the >35%/80 aa criteria than the BLASTX method, each method produced a high frequency of false positive alignments. Although the numbers may seem substantially lower for BLASTX compared with FASTA, the impact is very similar when translated to a false positive rate (99.7% for FASTA<sup>1</sup> and ~97.6% for BLASTX) and considering the fact that there are only 24 known maize allergen sequences (FARRP, 2011).

The difference in results between BLASTX and FASTA may additionally be due to the nature of individual BLASTX alignments which may span several stop codons. The BLASTX program simply takes a full length DNA sequence and translates on all six frames prior to aligning them to a database and does not recognize a minimum ORF size or location of stop codons. In contrast, the FASTA program treats each sequence bracketed by stop codons as a discrete sequence with a predefined length prior to the start of the alignment search process (Fig. 2); the user essentially prepares a ORF database from the parent six frame translation of DNA and sets the minimum ORF size prior to the FASTA actually performing the alignments with the database of interest. Regardless of how legitimate it is to compare the results from these two methods, it is clear that both produce a far greater number of alignments than the number of allergens for each respective plant species (Tables 1 and 2).

### 3.4. Observing the impact of working with predicted proteins; an example from maize

An additional and more biologically relevant alignment comparison using FASTA analysis was carried out for the maize protein set where only predicted protein sequences were compared to the allergen database. For both >35%/80 aa and the 8 aa criteria the total number of significant alignments was lower than the original gene set (Table 3) yet proportionately more alignments were identified with the maize protein set than with the gene set. The gene input sequences produced 20,326 alignments exceeding the >35%/80 aa criteria which roughly calculates as 3387 alignments per reading frame if each gene input sequence represents one of six possible ORFs; this compares to the 13,766 alignments for the protein set. This indicates the increased probability for an alignment exceeding alignment criteria with maize protein set which is not surprising given that real protein sequence could be expected to align at a low level more frequently than purely hypothetical translations (putative ORFs) from nearly random DNA. Also described in Table 3 are the number of unique maize genes or proteins that returned a significant alignment and the number of individual allergens that were hit in each search. The maize protein sequence set contains isoforms and splice variants and it is possible that some hits may have occurred to these thus increasing the number of significant alignments. However, it is evident that a wide range of individual allergens produced alignments in this search.

## 4. Discussion

The use of bioinformatics is an important component of the overall safety assessment of transgenic plants. Bioinformatic analysis helps to ensure that proteins likely to be cross-reactive with known allergens, or known allergens themselves, are not used in transgenic food crops. Routine updates to allergen databases can also help account for the discovery of new allergens. As transgenic

<sup>1</sup> E.g., 100% – (24 known allergens/7,120 FASTA alignments) × 100%.

**Table 3**

Total number of alignments<sup>a</sup> between maize gene and protein sets to known allergens. Alignments exceeding the >35%/80 aa and 8 aa criteria from a FASTA analysis using the FARRP allergen database.

| Sequence set                              | >35%/80 aa Alignments |                       |                      | 8aa Matches          |                       |                      |
|-------------------------------------------|-----------------------|-----------------------|----------------------|----------------------|-----------------------|----------------------|
|                                           | Number of alignments  | Unique genes/proteins | Individual allergens | Number of alignments | Unique genes/proteins | Individual allergens |
| Maize genes (all possible reading frames) | 20,326                | 7120                  | 876                  | 186,522              | 10,332                | 1312                 |
| Maize proteins (predicted)                | 13,766                | 2509                  | 799                  | 108,140              | 3578                  | 942                  |

<sup>a</sup> For all alignments known endogenous allergens were removed from the analyses to avoid endogenous allergen sequences from aligning to themselves.

crop safety assessment processes have matured, important aspects of bioinformatics have been recognized including the source organism of novel proteins, the alignment methods used, and the criteria (Silvanovich et al., 2009, 2006; Ladics et al., 2007; Thomas et al., 2005).

In this paper we have focused on assessing the biological relevance of DNA and protein sequence selection based on the interpretation of the output from bioinformatic analyses and the impact on assessing allergenic potential. The intent was to assess food genomes with methods similar to that considered for the inclusive T-DNA sequence search used in transgenic crops. As expected, the results suggest that each crop species has a much greater potential allergy risk than what is currently observed clinically. With the exception of soybean (considered one of the 8 most allergenic foods), the other crops in this study are not considered major allergen containing species, contrary to the number of alignments observed in this study exceeding allergen similarity criteria. In our analysis, soybean registers with 17,011 FASTA 35%/80aa alignments, but has only 38 proteins (from *Glycine max* and *Glycine soja*) listed in the FARRP AllergenOnline database. The 2811 aligned unique soybean sequences represent ~6% of the input sequences, yet the 38 identified or putative allergen proteins are ~0.08% of the total. Soybean can have 0.1% clinical allergy prevalence and rank as low as 11th place behind other foods (Hill et al., 1997). Therefore, the count of significantly aligned sequences in this study is displaying a high frequency of false positives when broadly comparing plant genes to known allergens. The conservative application of the bioinformatic search methods are certainly highlighted by the thousands of alignments to allergens for watermelon and cucumber (Tables 1 and 2), for which there are no identified allergens (FARRP, 2011). Another striking piece of data is that when compared to the maize genome (all reading frames, Table 3) 88% of all allergens had at least one 8aa identity with a translated ORF sequence. This general phenomenon of false positives can likely be extended to the endogenous genomic flanking sequences that are adjacent to transgenic DNA inserts as well. In these analyses flanking sequence is used to assess portions of a crop's endogenous DNA. Most often that endogenous DNA is assessed on all 6 reading frames for amino acid sequence similarity with known proteins using an algorithm such as BLASTX.

In short, the bioinformatic results do not correlate with what is generally observed by clinicians and allergen databases (FARRP, 2011) regarding the allergenicity of these foods. For each plant species we observed a high frequency of false positive alignments regardless of the search algorithm used. Although various allergen isoforms are represented in the database which may account for some alignments, each individual positive alignment would normally be investigated further upon its identification. Using maize as an example, there were 206,848 FASTA alignments that exceeded at least one of the (Codex Alimentarius Commission, 2009) criteria for bioinformatics. If one asserts that at least one of the 20,326 35%/80aa FASTA alignments is a relevant indication of concern then one must also assume that the aligned endogenous gene is either (a) an unidentified real allergen or potentially

cross-reactive with IgE specific to the allergen, or (b) or genetic sequence that is not translated into a protein. To check whether query sequences from the prospective proteome differ from derived gene sets, an analysis of predicted maize proteins was conducted. The results in Table 3 show that even endogenous non-allergen proteins display a disproportionately high number of alignments (13,766) when in fact there are as few as 24 maize allergen proteins in the proteome (Table 3). Real proteins appear to produce more false positive alignments per query sequence than when using all putative reading frames of DNA sequence.

A conservative approach used in setting the alignment criteria for allergens has been a focus of novel food safety development (Codex Alimentarius Commission, 2009, 2003; FAO/WHO, 2001). Conservative in this context means that alignment criteria produce mostly false positive alignments (Hileman et al., 2002; Ladics et al., 2007), but are valuable for screening the identity of potentially relevant safety issues for novel proteins. Bioinformatics applied this way have been very successful in avoiding the introduction of allergens and cross-reactive proteins into novel foods (Delaney et al., 2008; Goodman and Tetteh, 2011) and this has been highlighted by the fact that the featured bioinformatic analysis has focused on the trait proteins themselves as intended for expression in transgenic crops. However, an attempt to provide an even more thorough analysis for transgenic crop safety by analyzing all reading frames has resulted in an inability to identify meaningful results when applying these conservative metrics and methods as observed here.

Bioinformatic analysis when applied to a plant genome produces essentially a snapshot in terms of the ongoing dynamic nature of genome sequence changes that occur during routine crop breeding. Genome changes from breeding would be expected to impact transgenic and non-transgenic crop genomes alike (Cellini et al., 2004). In essence, if genomic sequence changes were to be observed using bioinformatics as the viewing window, results would show mostly irrelevant, false positive alignments to allergens and would change incrementally even though the relative allergenic potential is unlikely to change. The false positive frequency is expected to be at least ~95% (Silvanovich et al., 2009) for the FASTA methodology and appears even greater in this study (>98%; e.g. soybean). A very small number of real allergens will always remain difficult to filter and identify from a very large pool of non-allergen sequences using conservative criteria. The net effect is that bioinformatics seems to be substantially reduced in its utility for evaluating allergy safety when assessing a large pool of sequences created with alternative reading frames and non-coding DNA sequence. This would certainly be expected to impact the outcome of a transgenic crop's T-DNA analysis if it was assessed for its non-coding sequence and alternative reading frames because the T-DNA includes a significant portion of non-coding DNA.

The frequency of false positive alignments between a plant's genes and allergens do not correlate with known allergen proteins or food allergy risk. In fact, when attempting to assess a genome-sized number of sequences it is virtually impossible to separate out the small number of known allergens from non-allergens with

bioinformatics alone. Additionally, evaluation methods (e.g. stop-to-stop definition of an ORF) ignore the complexity of eukaryotic gene expression. For food safety, bioinformatics is considered a first step in identifying relevant similarity between a novel protein and an allergen. If significant shared identity were observed then subsequent antibody binding studies with IgE from the sera of specific allergic individuals would likely follow. In essence, bioinformatics does not possess the capacity to predict or act as an end-determinant of allergenicity for proteins, especially with regard to very short (eight amino acids) sequences (Goodman et al., 2008; Bannon and Ogawa, 2006). However, it would be impractical and add little value to the safety assessment to run potentially dozens of irrelevant IgE binding studies if prompted by a bioinformatics analysis of a T-DNA and the entirety of its non-coding sequence. It is proposed that the current strategy that thoroughly assesses trait proteins intended for expression in transgenic crops along with putative genomic DNA/T-DNA junction sequences using the current Codex Alimentarius criteria for bioinformatics is sufficient for most of the risk for inserting a known allergen or cross-reactive novel protein into transgenic crops. The combined effect of applying bioinformatics to an overly broad, untargeted T-DNA sequence selection along with conservative bioinformatic methods and criteria reveal here that the output is likely irrelevant from an allergy safety perspective.

### Conflict of interest statement

The authors are employed by Syngenta Crop Protection, LLC, which develops and markets transgenic seed.

### References

- Aalberse, R.C., 2000. Structural biology of allergens. *J. Allergy Clin. Immunol.* 106, 228–238.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Bannon, G., 2004. What makes a food protein an allergen? *Curr. Allergy Asthma Rep.* 4, 43–46.
- Bannon, G.A., Ogawa, T., 2006. Evaluation of available ige-binding epitope data and its utility in bioinformatics. *Mol. Nutr. Food Res.* 50, 638–644.
- Bombarley, A., Menda, N., Tecle, I., Buels, R., Stricker, S., Fischer-York, T., Pujar, A., Leo, J., Gosselin, J., Mueller, L., 2010. The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res.* 39, D1149–D1155.
- Cellini, F., Chesson, A., Colquhoun, I., Constable, A., Davies, H.V., Engel, K.H., Gatehouse, A.M.R., Karenlampi, S., Kok, E.J., Leguay, J.-J., Lehesranta, S., Noteborn, H.P.J.M., Pedersen, J., Smith, M., 2004. Unintended effects and their detection in genetically modified crops. *Food Chem. Toxicol.* 42, 1089–1125.
- Codex Alimentarius Commission, 2003. Alinorm 03/34: Joint FAO/WHO Food Standard Programme, Codex Alimentarius Commission, Twenty-Fifth Session, Rome, 30 June–5 July, 2003. Appendix III, Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants and Appendix IV, Annex on the assessment of possible allergenicity, pp. 47–60.
- Codex Alimentarius Commission. CAC/GL 45–2003. Foods derived from modern biotechnology. Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants (2009).
- Delaney, B., Astwood, J.D., Cunney, H., Conn, R.E., Herouet-Guicheney, C., MacIntosh, S., Meyer, L.S., Privalle, L., Gao, Y., Mattsson, J., Levine, M., 2008. Evaluation of protein safety in the context of agricultural biotechnology. *Food Chem. Toxicol.* 46, S71–S97.
- FAO/WHO 2001. Evaluation of allergenicity of genetically modified foods. Report of a joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology. (Food and Agriculture Organization of the United Nations (FAO), Rome).
- FARRP (Food allergy research and resource program), 2011. Located at <http://www.allergenonline.org>, University of Nebraska, Lincoln, NE.
- 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., Tetteh, A., 2011. Suggested improvements for the allergenicity assessment of genetically modified plants used in foods. *Curr. Allergy Asthma Rep.* 11, 317–324.
- 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.
- Hill, D.J., Hosking, C.S., Zhie, C.Y., Leung, R., Baratwidjaja, K., Ikura, Y., Iyngkaran, N., Gonzalez-Andaya, A., Wah, L.B., Hsieh, K.H., 1997. The frequency of food allergy in Australia and Asia. *Environ. Toxicol. Pharm.* 4, 101–110.
- Jenkins, J.A., Griffiths-Jones, S., Shewry, P.R., Breiteneder, H., Mills, C., 2005. Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: an *in silico* analysis. *J. Allergy Clin. Immunol.* 115, 163–170.
- Ladics, G.S., Bardina, L., Cressman, R.F., Mattsson, J.L., Sampson, H.A., 2006. Lack of cross-reactivity between cry1f protein in herculex<sup>tm</sup> maize and dust mite der p 7 protein with human sera positive for der p 7-ige. *Reg. Toxicol. Pharmacol.* 44, 136–143.
- 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 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. *Reg. Toxicol. Pharmacol.* 60, 46–53.
- National center for biotechnology information (NCBI), 2011. Blast program selection guide, located at <http://www.blast.ncbi.nlm.nih.gov>.
- Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., Buell, C.R., 2007. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* 35, D883–D887. <http://www.rice.plantbiology.msu.edu/index.shtml>.
- Pearson, W.R., 2000. Flexible sequence similarity searching with the fasta3 program package. In: Misener, S., Krawetz, S.A. (Eds.), *Methods Mol. Biol.* Humana Press, Totowa, NJ, pp. 185–219.
- Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., Jackson, S.A., 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463 (7278), 178–183.
- Silvanovich, A., Bannon, G., McClain, S., 2009. The use of e-scores to determine the quality of protein alignments. *Reg. Toxicol. Pharmacol.* 54, S26–S31.
- Silvanovich, A., Nemeth, M.A., Song, P., Herman, R., Tagliani, L., Bannon, G.A., 2006. The value of short amino acid sequence matches for the prediction of protein allergenicity. *Toxicol. Sci.* 90, 252–258.
- Thomas, K., Bannon, G., Hefle, S., Herouet, C., Holsapple, M., Ladics, G., MacIntosh, S., Privalle, L., 2005. In silico methods for evaluating human allergenicity to novel proteins: International bioinformatics workshop meeting report, 23–24 February 2005. *Toxicol. Sci.* 88, 307–310.
- Vieths, S., Scheurer, S., Ballmer-Weber, B., 2002. Current understanding of cross-reactivity of food allergens and pollen. *Ann. NY Acad. Sci.* 964, 47–68.