

**Study Title**

**Bioinformatics Evaluation of DNA Sequences Flanking the 5' and 3' Junctions of  
Inserted DNA in MON 87708: Assessment of Putative Polypeptides**

**Authors**

**Haidi Tu, M.S.  
Andre Silvanovich, Ph.D.**

**Study Completed On**

**April 21, 2010**

**Sponsor and Performing Laboratory**

**Monsanto Company  
Regulatory Product Characterization Center  
800 North Lindbergh Blvd.  
St. Louis, MO 63167**

**Laboratory Project ID**

**MSL Number: MSL0022682  
Study Number: REG-10-161**

**The text below applies only to use of the data by the United States Environmental Protection Agency (U.S. EPA) in connection with the provisions of the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA).**

**The inclusion of this page in all reports is for quality assurance purposes and does not necessarily indicate that this report has been submitted to the U.S. EPA.**

### **Statement of Data Confidentiality Claim**

Information claimed confidential on the basis of its falling within the scope of FIFRA section 10(d)(1)(A), (B), or (C) has been removed to a confidential appendix, and is cited by cross-reference number in the body of the study.

We submit this material to the United States Environmental Protection Agency specifically under the requirements set forth in FIFRA as amended, and consent to the use and disclosure of this material by the EPA strictly in accordance with FIFRA. By submitting this material to the EPA in accordance with the method and format requirements contained in PR Notice 86-5, we reserve and do not waive any rights involving this material that are or can be claimed by the company notwithstanding this submission to the EPA.

Company: \_\_\_\_\_ Monsanto Company \_\_\_\_\_

Company Agent: \_\_\_\_\_

Title: \_\_\_\_\_

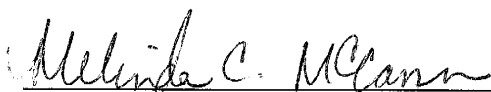
Signature: \_\_\_\_\_ Date: \_\_\_\_\_

**Statement of Compliance**

This project does not meet the U.S. EPA Good Laboratory Practice requirements as specified in 40 CFR Part 160.

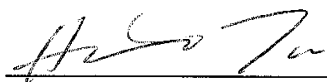
\_\_\_\_\_  
Submitter

Date: \_\_\_\_\_



Date: 21 Apr 2010

Melinda C. McCann, M.S.  
Sponsor Representative



Date: 4-21-10

Haidi Tu, M.S.  
Author

### Quality Assurance Statement


**Study Title:** Bioinformatics Evaluation of DNA Sequences Flanking the 5' and 3' Junctions of the Inserted DNA in MON 87708: Assessment of Putative Polypeptides

**Study Number:** REG-10-161

Reviews conducted by the Quality Assurance Unit confirm that the final report accurately describes the methods and standard operating procedures followed and accurately reflects the raw data of the study.

Following is a list of reviews conducted by the Monsanto Regulatory Quality Assurance Unit on the study reported herein.

Dates of Inspection/Audit	Phase	Date Reported to Study Director	Date Reported to Management
04/12/2010	Draft Report Review	04/21/2010	04/21/2010

  
\_\_\_\_\_  
**Quality Assurance Unit**  
**Monsanto Regulatory, Monsanto Company**

04/21/2010  
\_\_\_\_\_  
**Date**

### Study Certification

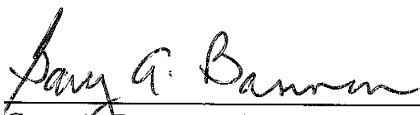
This report is an accurate and complete representation of the study/project activities.

#### Signatures of Final Report Approval:



Haidi Tu, M.S.  
Author

Date: 4-21-10



Gary A. Bannon, Ph.D.  
Lead, Regulatory Product Characterization Center

Date: April 21, 2010

### Study Information

**Study Number:** REG-10-161

**Title:** Bioinformatics Evaluation of DNA Sequences Flanking the 5' and 3' Junctions of the Inserted DNA in MON 87708: Assessment of Putative Polypeptides

**Facility:** Monsanto Company  
Regulatory Product Characterization Center  
800 North Lindbergh Blvd.  
St. Louis, MO 63167

**Protein Sciences Lead:** Gary A. Bannon, Ph.D.

**Sponsor Representative:** Melinda C. McCann, M.S.

**Authors:** Haidi Tu, M.S.  
Andre Silvanovich, Ph.D.

**Study Start Date:** March 26, 2010

**Study Completion Date:** April 21, 2010

**Records Retention:** All study specific raw data and final report will be retained at Monsanto-St. Louis.

**©2010 Monsanto Company. All Rights Reserved.**

This document is protected under copyright law. This document is for use only by the regulatory authority to which it has been submitted by Monsanto Company, and only in support of actions requested by Monsanto Company. Any other use of this material, without prior written consent of Monsanto, is strictly prohibited. By submitting this document, Monsanto does not grant any party or entity any right to license or to use the information of intellectual property described in this document.

## Table of Contents

Section	Page
Study Title.....	1
Statement of Data Confidentiality Claim.....	2
Statement of Compliance.....	3
Quality Assurance Statement.....	4
Study Certification .....	5
Study Information .....	6
Table of Contents.....	7
Abbreviations and Definitions .....	10
1.0 Summary.....	11
2.0 Introduction.....	12
3.0 Purpose.....	13
4.0 Methods.....	14
4.1 Sequence Database Preparation .....	14
4.2 Translation of Putative Polypeptides .....	14
4.3 Sequence Database Searches .....	15
4.4 Significance of the Alignment .....	16
5.0 Results and Discussion .....	17
5.1 Assessment of Potential Allergenicity .....	17
5.2 Assessment of Potential Toxicity.....	17
5.3 Assessment of Potential Adverse Biological Activity .....	18
6.0 Conclusions.....	18
7.0 References.....	19

## Figures

Figure 1. Reading frame assignment and DNA sequence at the 5' junctions of the MON 87708 insert. ....	21
--	----

Figure 2. Reading frame assignment and DNA sequence at the 3' junctions of the MON 87708 insert. ....	22
---	----

Figure 3. Graphic mapping of the flanking DNA sequences and putative polypeptides encoded by each reading frame at the 5' and 3' junctions of the MON 87708 insert. ....	23
--	----

## Tables

Table 1. The predicted sequence of polypeptides encoded by each reading frame at the 5' and 3' junctions of the MON 87708 insert. ....	24
--	----

Table 2. Summary of alignments for the FASTA searches of the AD_2010 database using putative polypeptide sequences encoded by the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 5' junctions in MON 87708. ....	25
--	----

Table 3. Summary of alignments for the FASTA searches of the TOX_2010 database using putative polypeptide sequences encoded by the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 5' junctions in MON 87708. ....	26
---	----

Table 4. Summary of alignments for the FASTA searches of the PRT_2010 database using putative polypeptide sequences encoded by the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 5' junctions in MON 87708. ....	27
---	----

Table 5. Summary of alignments for the FASTA searches of the AD_2010 database using putative polypeptide sequences encoded by the the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 3' junctions in MON 87708.....	28
---	----

Table 6. Summary of alignments for the FASTA searches of the TOX_2010 database using putative polypeptide sequences encoded by the the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 3' junctions in MON 87708.....	29
--	----

Table 7. Summary of alignments for the FASTA searches of the PRT_2010 database using putative polypeptide sequences encoded by the the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 3' junctions in MON 87708.....	30
--	----

## Appendices

Appendix 1. Bioinformatic analysis of polypeptide 5_1 .....	31
Appendix 2. Bioinformatic analysis of polypeptide 5_2a .....	33
Appendix 3. Bioinformatic analysis of polypeptide 5_2b .....	36
Appendix 4. Bioinformatic analysis of polypeptide 5_3a .....	39
Appendix 5. Bioinformatic analysis of polypeptide 5_3b .....	42



Appendix 6. Bioinformatic analysis of polypeptide 5_4a .....	44
Appendix 7. Bioinformatic analysis of polypeptide 5_4b .....	47
Appendix 8. Bioinformatic analysis of polypeptide 5_5a .....	49
Appendix 9. Bioinformatic analysis of polypeptide 5_5b .....	52
Appendix 10. Bioinformatic analysis of polypeptide 5_6a .....	55
Appendix 11. Bioinformatic analysis of polypeptide 3_1a .....	58
Appendix 12. Bioinformatic analysis of polypeptide 3_1b .....	61
Appendix 13. Bioinformatic analysis of polypeptide 3_2 .....	64
Appendix 14. Bioinformatic analysis of polypeptide 3_3a .....	67
Appendix 15. Bioinformatic analysis of polypeptide 3_3b .....	69
Appendix 16. Bioinformatic analysis of polypeptide 3_4 .....	72
Appendix 17. Bioinformatic analysis of polypeptide 3_5a .....	75
Appendix 18. Bioinformatic analysis of polypeptide 3_5b .....	78
Appendix 19. Bioinformatic analysis of polypeptide 3_6a .....	80
Appendix 20. Bioinformatic analysis of polypeptide 3_6b .....	83

## Abbreviations and Definitions

AA	Amino acid
AD_2010	Allergen and gliadin protein sequence database (Release date January 22, 2010)
BLOCKS	A database of amino acid motifs found in protein families
BLOSUM	BLOcks SUBstitution Matrix, used to score similarities between pairs of distantly related protein or nucleotide sequences
DMO	Dicamba Mono-Oxygenase
<i>E</i> -Score	Expectation score
FARRP	Food Allergy Research and Resource Program Database
FASTA	Algorithm used to find local high scoring alignments between a pair of protein or nucleotide sequences
GenBank	A public genetic database maintained by the National Center for Biotechnology Information at the National Institutes of Health, Bethesda, MD, USA
GI	Gene Identification number
IgE	Immunoglobulin E
NCBI	National Center of Biotechnology Information at the National Institutes of Health, Bethesda, MD, USA
ORF	Open Reading Frame
PRT_2010	GenBank protein database, 175.0 (Release date January 22, 2010)
TOX_2010	Toxin protein sequence database (Release date January 22, 2010)

## 1.0 Summary

Monsanto Company has developed herbicide-tolerant soybean MON 87708 that is tolerant to dicamba (3,6-dichloro-2-methoxybenzoic acid) herbicide. MON 87708 contains a demethylase gene from *Stenotrophomonas maltophilia* that expresses the dicamba mono-oxygenase (DMO) protein to confer tolerance to dicamba herbicide.

Structurally, the MON 87708-produced DMO enzyme functions as a trimer and consists of two forms of the DMO protein, namely DMO and DMO+27. The DMO protein is the mature form of the protein while the DMO+27 protein contains an additional 27 amino acids on its amino terminus, originating from the pea Rubisco small subunit.

As part of a comprehensive safety assessment, bioinformatic analyses were performed to assess the potential for allergenicity, toxicity, or biological activity of putative polypeptides encoded by the 5' and 3' inserted DNA-soybean genomic DNA junctions. Besides the T-DNA sequence, two intervening sequences were inserted into MON 87708 at both the 5' and 3' junctions during transformation. Therefore, sequences spanning the 5' and 3' soybean genomic DNA-intervening DNA and intervening DNA-T-DNA junctions were translated from stop codon to stop codon in all six reading frames. Putative polypeptides from each reading frame, eight amino acids or greater in length were compared to allergen (AD\_2010), toxin (TOX\_2010), and all protein (PRT\_2010) database sequences using bioinformatic tools.

The FASTA sequence alignment tool was used to assess structural relatedness between the query sequences and protein sequences in the AD\_2010, TOX\_2010, and PRT\_2010 databases. Structural similarities shared between each putative polypeptide with each sequence in the database were examined. The extent of structural relatedness was evaluated by detailed visual inspection of the alignment, the calculated percent identity, and the *E*-score. In addition to structural similarity, each putative polypeptide was screened for short polypeptide matches using a pair-wise comparison algorithm. In these analyses, eight contiguous and identical amino acids were defined as immunologically relevant, where eight represents the typical minimum sequence length likely to represent an immunological epitope.

No biologically relevant structural similarity to known allergens, toxins, or biologically active proteins was observed for any of the putative polypeptides. Furthermore, no short (eight amino acid) polypeptide matches were shared between any of the putative polypeptides and proteins in the allergen database. These data demonstrate the lack of both structurally and immunologically relevant similarity to known allergens for all of the putative polypeptides analyzed.

This bioinformatics analysis is theoretical. No empirical evidence exists to suggest that transcription of DNA sequence at the 5' or 3' junctions of the DNA inserted in MON 87708 occurs. Rather, the results of these bioinformatic analyses indicate that in the highly unlikely occurrence that any of the junction sequences were to be transcribed and that a transcript were to be translated, the translation product would not share a sufficient degree of sequence similarity or identity to indicate that it would be potentially allergenic, toxic, or have other safety implications.

## **2.0 Introduction**

Monsanto Company has developed herbicide-tolerant soybean MON 87708 that is tolerant to dicamba (3,6-dichloro-2-methoxybenzoic acid) herbicide. MON 87708 contains a demethylase gene from *Stenotrophomonas maltophilia* that expresses the dicamba mono-oxygenase (DMO) protein to confer tolerance to dicamba herbicide.

Structurally, the MON 87708-produced DMO enzyme functions as a trimer and consists of two forms of the DMO protein, namely DMO and DMO+27. The DMO protein is the mature form of the protein while the DMO+27 protein contains an additional 27 amino acids on its amino terminus, originating from the pea Rubisco small subunit.

As part of a comprehensive safety assessment, bioinformatic analyses were performed to assess the potential for allergenicity, toxicity, or biological activity of putative polypeptides encoded by the 5' and 3' inserted DNA-soybean genomic DNA junctions. Sequences spanning the 5' and 3' soybean genomic DNA-intervening DNA and intervening DNA-T-DNA junctions were translated from stop codon to stop codon in all six reading frames. Putative polypeptides from each reading frame, eight amino acids or greater in length were compared to allergen (AD\_2010), toxin (TOX\_2010), and all protein (PRT\_2010) database sequences using bioinformatic tools.

Exposure to allergens in foods may cause sudden, medically significant reactions in susceptible individuals. Gliadins and glutenins are suspected to cause celiac disease, a non-IgE mediated disorder (gluten-sensitive enteropathy), and are also considered important immunologically active proteins. Screening the amino acid sequences of proteins introduced into plants by modern biotechnology for similarity to sequences of known allergens, gliadins, and glutenins is one of many assessments performed to support product safety. Similarly, the amino acid sequences of introduced proteins are also screened against known toxins as well as all known proteins in publicly available genetic databases.

The FASTA algorithm can be used to evaluate the extent of sequence alignment between a query protein sequence and a database sequence. In principle, if two proteins share sufficient linear sequence similarity and identity, they will likely share three-dimensional structure and, therefore, functional homology. By definition, homologous proteins share

secondary structure and common three-dimensional folds (Pearson, 2000). Because the degree of relatedness between homologs varies widely, the data need to be carefully evaluated in order to maximize their potential predictive value. The allergenicity assessment is used to identify known allergens or potentially cross-reactive proteins. While related (homologous) proteins may share 25% amino acid identity in a 200 amino acid overlap (Pearson, 2000), this is not generally sufficient to indicate IgE mediated cross-reactivity (Aalberse et al., 2001). Indeed, allergenic cross-reactivity caused by proteins is rare at 50% identity and typically requires >70% amino acid identity across the full length of the protein sequences (Aalberse, 2000). A conservative approach is currently applied by which related protein sequences are identified as potentially cross-reactive if linear identity is 35% or greater in an 80 amino acid overlap (Thomas et al, 2005). Such levels of identity are readily detected using FASTA. Additionally, proteins closely related to gliadins or glutenins, the proteins that trigger celiac disease, can be easily identified using FASTA.

A second bioinformatics tool, an eight amino acid sliding window search, is used to specifically identify short linear polypeptide matches to known or suspected allergens. It is possible that proteins structurally unrelated to allergens, gliadins, and glutenins may still contain smaller immunologically significant epitopes. A query sequence may be considered allergenic if it has an exact sequence identity of at least eight contiguous amino acids with a potential allergen epitope (Metcalf et al., 1996; Hileman et al., 2002; Goodman et al., 2002). However, most allergen epitopes have not been confirmed and the amino acid length for those that have been identified can vary widely, thus the relevance of an exact match of eight amino acids may have limited immunological relevance (Thomas et al., 2005). The eight amino acid bioinformatic strategy is currently an *in silico* search that can produce matches containing significant uncertainty depending on the length of the query sequence (Silvanovich et al., 2006).

This report describes the bioinformatics assessment of putative polypeptides encoded at the soybean 5' and 3' genomic DNA-intervening DNA and intervening DNA-inserted DNA junctions of MON 87708. Inspection of the bioinformatic analysis data can be used to indicate whether the putative polypeptides have biologically relevant sequence similarity to known allergens, toxins, or other biologically active proteins.

### **3.0 Purpose**

Besides the T-DNA sequence, two intervening sequences were inserted into MON 87708 at both the 5' and 3' junctions during transformation. The purpose of this study was to evaluate the amino acid sequences of putative polypeptides obtained from all reading frames that span the 5' and 3' soybean genomic DNA-intervening DNA and intervening DNA-T-DNA junctions in MON 87708 to sequences in established databases. Sequences spanning these four junctions were translated from stop codon to stop codon in all reading frames. Structural relatedness between the putative polypeptides and known

allergens, toxins, and biologically active proteins was assessed using the FASTA sequence alignment tool. Using each putative polypeptide as a query sequence that was eight amino acids or greater in length and that spanned one of the junctions, FASTA searches were performed on allergen (AD\_2010), toxin (TOX\_2010), and all protein (PRT\_2010) sequence databases. Immunologically relevant correlates were assessed using the pairwise comparison algorithm using the putative polypeptide as a query sequence to search against the AD\_2010 database.

## **4.0 Methods**

### *4.1 Sequence Database Preparation*

The allergen, gliadin, and glutenin sequence database (AD\_2010) was obtained from FARRP (2010)<sup>1</sup> and was used as provided. The AD\_2010 database contains 1,471 sequences. A complete description of the AD\_2010 database can be found in Tu and Silvanovich (2010).

GenBank protein database, release 175.0, was downloaded from NCBI and formatted for use in these bioinformatic analyses. It is referred to herein as the PRT\_2010 database and contains 17,815,538 sequences. A complete description of the PRT\_2010 database can be found in Tu and Silvanovich (2010).

The toxin database is a subset of sequences derived from the PRT\_2010 database that was selected using a keyword search and filtered to remove likely non-toxin proteins. It is referred to herein as the TOX\_2010 database and contains 8,448 sequences. A complete description of the TOX\_2010 database can be found in Tu and Silvanovich (2010).

### *4.2 Translation of Putative Polypeptides*

DNA sequence spanning the 5' and 3' junctions of the MON 87708 insertion site (Song et al., 2009) was analyzed for translational stop codons (TGA, TAG, TAA). All six reading frames spanning the genomic DNA-intervening DNA and/or intervening DNA-T-DNA junctions were translated using the standard genetic code from stop codon to stop codon. A total of twenty sequences of eight amino acids or greater that spanned the junction(s) were analyzed. The DNA sequence was translated to the amino acid sequence with DNASTar, version 8.0.2 (13), 412 (Table 1).

---

<sup>1</sup> located at <http://www.allergenonline.com>

#### *4.3 Sequence Database Searches*

FASTA analyses using the AD\_2010, TOX\_2010 and PRT\_2010 databases were performed on a virtual machine loaded with a SUSE LINUX version 10 operating system and FASTA version 3.4t 26 (July 7, 2006). The structural similarity of the translated protein sequences to sequences in each database (AD\_2010, TOX\_2010, and PRT\_2010) was assessed using the FASTA algorithm (Lipman and Pearson, 1985; Pearson and Lipman, 1988).

FASTA comparisons are initiated by aligning the first match of a specific wordsize. The alignment is then extended based on the chosen scoring matrix. With the exception of expectation threshold (*E*-score) of one, default FASTA search parameters were used. The *E*-score is a statistical measure of the likelihood that the observed similarity score could have occurred by chance in a search. A larger *E*-score indicates a lower degree of similarity between the query sequence and the sequence from the database. Typically, alignments between two sequences will need to have an *E*-score of  $1e-5$  ( $1 \times 10^{-5}$ ) or smaller to be considered to have significant homology. FASTA comparisons were performed using the BLOSUM50 scoring matrix (Henikoff and Henikoff, 1992). Multiple alignments are made between the query sequence and each sequence in the database with a score calculated for each alignment. Only the top scoring alignments are extensively analyzed for each database sequence. The BLOSUM matrix series was derived from a set of aligned, ungapped regions from protein families, called the BLOCKS database. Sequences from each block were clustered based on the percent of identical residues in the alignments (Henikoff and Henikoff, 1996). The BLOSUM50 matrix will identify blocks of conserved residues that are at least 50% identical. BLOSUM50 works well for identifying sequence similarities that include gaps, and thus recognizes distant evolutionary relationships (Pearson, 2000).

If two proteins share sufficient linear sequence similarity and identity, they will also share three-dimensional structure and, therefore, functional homology. By definition, homologous proteins share secondary structure and common three-dimensional folds (Pearson, 2000). Because the degree of relatedness between homologs varies widely, the data need to be carefully evaluated in order to maximize their potential predictive value. The allergenicity assessment is used to identify known allergens or potentially cross-reactive proteins. While related (homologous) proteins may share 25% amino acid identity in a 200 amino acid overlap (Pearson, 2000), this is not generally sufficient to indicate IgE mediated cross-reactivity (Aalberse et al., 2001). Indeed, allergenic cross-reactivity caused by proteins is rare at 50% identity and typically requires >70% amino acid identity across the full length of the protein sequences (Aalberse, 2000). A conservative approach is currently applied by which related protein sequences are identified as

potentially cross-reactive if linear identity is 35% or greater in an 80 amino acid overlap (Thomas et al., 2005). Such levels of identity are readily detected using FASTA. Additionally, proteins closely related to gliadins or glutenins, the proteins that trigger celiac disease, can be easily identified using FASTA.

In addition to the FASTA comparisons of each putative polypeptide to known allergens (to assess overall structural similarity), an eight amino acid sliding window search was performed. An algorithm was developed to identify whether or not a linearly contiguous match of eight amino acids existed between the query sequence and sequences within the allergen database (AD\_2010). This program compares the query sequence to each protein sequence in the allergen database using a sliding-window of eight amino acids; that is, with a seven amino acid overlap relative to the preceding window. While there have been recommendations for using a shorter scanning window (Gendel, 1998; Kleter and Peijnenburg, 2002), only a few studies have actually investigated the ability of six, seven, or eight amino acid search windows to identify allergens (Hileman et al., 2002; Goodman et al., 2002; Stadler and Stadler, 2003). In these studies, randomly or specifically selected protein sequences were used as query sequences in FASTA and six, seven, and eight amino acid window searches against allergen databases. The results demonstrated that searches with six and seven amino acid windows led to high rates of false positive matches between non-allergenic query sequences and allergen database sequences. Additionally, searches with a six or seven amino acid window identified apparently random matches between totally unrelated proteins, such that the matched proteins were not likely to share any structural or sequence similarities that could act as cross-reactive epitopes. These studies concluded that six or seven amino acid sliding-window searches yielded such a high rate of false positive hits that they were of no predictive value. Furthermore, Silvanovich et al. (2006) recently demonstrated the lack of value of six or seven amino acid sliding-window searches in a comprehensive analysis of short peptide match frequencies by analyzing the match frequencies of peptides derived from ~1.95 million published protein sequences. In order to provide the best predictive capability to identify potentially cross-reactive proteins, a window of eight contiguous amino acids is used to represent the smallest immunologically significant sequential, or linear IgE binding epitope (Metcalf et al., 1996).

#### *4.4 Significance of the Alignment*

An *E*-score of  $1e-5$  ( $1 \times 10^{-5}$ ) was set as an initial high cut-off value for alignment significance. Although all alignments were inspected visually, any aligned sequence that yielded an *E*-score less than or equal to  $1e-5$  was analyzed further to determine if such an alignment represented significant sequence homology.



## 5.0 Results and Discussion

Bioinformatics analyses were performed on putative polypeptides deduced from DNA sequence spanning the 5' and 3' inserted DNA-genomic DNA junctions of MON 87708 to assess the potential for similarity towards known allergens, toxins, or other biologically active proteins. DNA sequence flanking the 5' (Figure 1) and 3' (Figure 2) junctions of the insertion site in MON 87708 were translated from stop codon to stop codon in all possible reading frames. Polypeptide sequence from each reading frame was then inspected to confirm that the sequence was both encoded by DNA spanning the genomic DNA-intervening DNA and/or intervening DNA-T-DNA junctions and was greater than or equal to eight amino acids in length. At each of the 5' and the 3' flanks, ten deduced putative polypeptides spanned the junctions (see Figure 3 and Table 1). Each putative polypeptide was designated as 5 or 3 (representing the 5' or 3' end, respectively), separated with an underscore by a numerical value 1 to 6 representing the respective reading frame, followed by a letter a or b (representing the junction between genomic-intervening DNA or intervening DNA-T-DNA, respectively; a or b is omitted for those putative polypeptides spanning the both junctions) (see Figures 1 and 2 for reading frame assignment). Supporting dataset output files for each putative 5' polypeptide are contained in Appendices 1-10, while dataset output files for each putative 3' polypeptide are contained in Appendices 11-20.

### 5.1 Assessment of Potential Allergenicity

The results of the allergenicity assessment are shown in Tables 2 and 5. Potential allergenicity of the twenty putative polypeptides was assessed using the FASTA and eight amino acid sliding window search algorithms. Using the FASTA algorithm to search the AD\_2010 database, no alignments with any of the twenty query sequences generated an *E*-score of less than or equal to  $1e-5$ . Likewise, no alignment met or exceeded the Codex Alimentarius (2003) FASTA alignment threshold for potential allergenicity of 35% identity over 80 amino acids. Finally, no alignments of eight or more consecutive identical amino acids were found between any query sequence and the AD\_2010 database. As a result, these twenty putative polypeptides are unlikely to contain any cross-reactive IgE binding epitopes with known allergens.

### 5.2 Assessment of Potential Toxicity

The results of the toxicity assessment are shown in Tables 3 and 6. Potential toxicity of the twenty putative polypeptides was assessed using the FASTA algorithm. Using the FASTA algorithm to search the TOX\_2010 database, no alignments with any of the twenty query sequences generated an *E*-score of less than or equal to  $1e-5$ .

### *5.3 Assessment of Potential Adverse Biological Activity*

The results of this assessment are shown in Tables 4 and 7. Potential untoward biological activity of the twenty putative polypeptides was assessed using the FASTA algorithm. Using the FASTA algorithm to search the PRT\_2010 database, no alignments with any of the twenty query sequences generated an *E*-score of less than or equal to  $1e-5$ .

## **6.0 Conclusions**

Analyses of putative polypeptides encoded by DNA spanning the 5' and 3' junctions of the MON 87708 inserted DNA were performed using bioinformatic tools. Results of the FASTA sequence alignments demonstrated a lack of structurally relevant similarity between any known allergens, toxins or biologically active proteins and the twenty putative polypeptides. The results of these bioinformatic analyses demonstrate that even in the highly unlikely occurrence that any of the junction polypeptides were translated, they would not share a sufficient degree of sequence similarity with other proteins to indicate that they would be potentially allergenic, toxic, or have other safety implications.

## 7.0 References

- Aalberse, R.C. 2000. Structural biology of allergens. *J Allergy Clin Immunol* **106**:228-38.
- Aalberse, R.C., Akkerdaas, J, and van Ree, R. 2001. Cross-reactivity of IgE antibodies to allergens. *Allergy* **56**:478-90.
- Codex Alimentarius 2003. Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants. CAC/GL 45-2003.
- FARRP, (Food Allergy Research and Resource Program Database) 2010.  
[www.allergenonline.com](http://www.allergenonline.com). University of Nebraska.
- Gendel, S.M., 1998. The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods. *Adv Food Nutr Res* **42**:45-62.
- Goodman, R.E., Silvanovich, A., Hileman, R.E., Bannon, G.A., Rice, E.A., and Astwood, J. D. 2002. Bioinformatic methods for identifying known or potential allergens in the safety assessment of genetically modified crops. *Comments on Toxicology*. **8**:251-269.
- Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**:10915-10919.
- Henikoff, J.G., and Henikoff, S. 1996. Blocks database and its applications. *Methods Enzymol* **266**:88-105.
- Hileman, R.E., Silvanovich, A., Goodman, R.E., Rice, E.A., Holleschak, G., Astwood, J.D., and Hefle, S.L. 2002. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int Arch Allergy Immunol* **128**:280-291.
- Kleter, G.A., and Peijnenburg, A.A.C.M. 2002. Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE – binding linear epitopes of allergens. *BMC Structl Biol* **2**:8-18.
- Lipman D.J. and Pearson W.R. 1985. Rapid and sensitive protein similarity searches. *Science* **227**:1435-1441.

Metcalfe, D.D., Astwood, J.D., Townsend, R., Sampson, H.A., Taylor, S.L., and Fuchs, R.L. 1996. Assessment of the allergenic potential of foods derived from genetically engineered crop plants. *Crit Rev Food Sci Nutr* **36**:S165-186.

Pearson, W.R., and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**:2440-2448.

Pearson, W.R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* **132**:185-219.

Silvanovich, A., Nemeth, M.A., Song, P., Herman, R., Taglianin, L., and Bannon, G.A. 2006. The value of short amino acid sequence matches for prediction of protein allergenicity. *Toxicol Sci* **90**:252-258.

Song, Z., Lawry, K. D., Rice, J. F., and Tian, Q. 2009. Amended report for MSL0021418: Molecular analysis of dicamba-tolerant soybean MON 87708. Monsanto Technical Report MSL0022109, St. Louis, MO.

Stadler, M.B., and Stadler, B.M. 2003. Allergenicity prediction by protein sequence. *FASEB J* **17**:1141-1143.

Thomas, K., Bannon, G., Hefle, S., Herouet, C., Holsapple, M., Ladics, G., MacIntosh, S., and 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.

Tu, H. and Silvanovich, A. 2010. The assembly of databases used for FASTA, BLAST and sliding window searches in 2010. Monsanto Technical Report MSL0022498, St. Louis, MO.

## **[CBI Cross Reference Number 1]**

Deleted Figures 1, 2 and 3 and Table 1

Deleted pages 21 – 24 are found in the Confidential Attachment, pages 5 – 8.

**Table 2. Summary of alignments for the FASTA searches of the AD\_2010 database using putative polypeptide sequences encoded by the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 5' junctions in MON 87708.**

Appendix	Polypeptide	AD_2010 Sequence Database						
		Sliding Window	FASTA search					
		# Hits	# Hits	GI #	Description	% Identity	aa Overlap	E-score
1	5_1	No	1	25361513	helix-loop-helix protein [F (450 aa)	52.174	23	0.2
2	5_2a	No	-	-	-	-	-	-
3	5_2b	No	2	145559502	NLTP3_VITSX RecName: Full=Non (91 aa)	41.176	17	0.11
4	5_3a	No	-	-	-	-	-	-
5	5_3b	No	-	-	-	-	-	-
6	5_4a	No	-	-	-	-	-	-
7	5_4b	No	-	-	-	-	-	-
8	5_5a	No	-	-	-	-	-	-
9	5_5b	No	-	-	-	-	-	-
10	5_6a	No	-	-	-	-	-	-

**Table 3. Summary of alignments for the FASTA searches of the TOX\_2010 database using putative polypeptide sequences encoded by the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 5' junctions in MON 87708.**

Appendix	Polypeptide	FASTA search of TOX_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	E-score
1	5_1	-	-	-	-	-	-
2	5_2a	-	-	-	-	-	-
3	5_2b	-	-	-	-	-	-
4	5_3a	-	-	-	-	-	-
5	5_3b	-	-	-	-	-	-
6	5_4a	-	-	-	-	-	-
7	5_4b	-	-	-	-	-	-
8	5_5a	-	-	-	-	-	-
9	5_5b	1	255922165	Ricin B lectin [Actinosynn (597 aa)	42.857	42	0.066
10	5_6a	-	-	-	-	-	-

**Table 4. Summary of alignments for the FASTA searches of the PRT\_2010 database using putative polypeptide sequences encoded by the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 5' junctions in MON 87708.**

Appendix	Polypeptide	FASTA search of PRT_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	E-score
1	5_1	-	-	-	-	-	-
2	5_2a	-	-	-	-	-	-
3	5_2b	-	-	-	-	-	-
4	5_3a	-	-	-	-	-	-
5	5_3b	-	-	-	-	-	-
6	5_4a	-	-	-	-	-	-
7	5_4b	-	-	-	-	-	-
8	5_5a	-	-	-	-	-	-
9	5_5b	-	-	-	-	-	-
10	5_6a	-	-	-	-	-	-



**Table 5. Summary of alignments for the FASTA searches of the AD\_2010 database using putative polypeptide sequences encoded by the the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 3' junctions in MON 87708.**

Appendix	Polypeptide	AD_2010 Sequence Database						
		Sliding Window	FASTA search					
		# Hits	# Hits	GI #	Description	% Identity	aa Overlap	E-score
11	3_1a	No	-	-	-	-	-	-
12	3_1b	No	1	51315784	EST_HEVBR RecName: Full=Estera (391 aa)	28.571	21	0.48
13	3_2	No	2	21413	aspartic proteinase inhibitor (217 aa)	50.000	18	0.094
14	3_3a	No	-	-	-	-	-	-
15	3_3b	No	-	-	-	-	-	-
16	3_4	No	-	-	-	-	-	-
17	3_5a	No	-	-	-	-	-	-
18	3_5b	No	-	-	-	-	-	-
19	3_6a	No	-	-	-	-	-	-
20	3_6b	No	-	-	-	-	-	-

**Table 6. Summary of alignments for the FASTA searches of the TOX\_2010 database using putative polypeptide sequences encoded by the the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 3' junctions in MON 87708.**

Appendix	Polypeptide	FASTA search of TOX_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	E-score
11	3_1a	4	38636512	heat stable enterotoxin I (72 aa)	66.667	12	0.21
12	3_1b	-	-	-	-	-	-
13	3_2	-	-	-	-	-	-
14	3_3a	-	-	-	-	-	-
15	3_3b	-	-	-	-	-	-
16	3_4	2	125997004	Aerolysin/hemolysin/leukoc (430 aa)	30.233	86	0.62
17	3_5a	-	-	-	-	-	-
18	3_5b	-	-	-	-	-	-
19	3_6a	-	-	-	-	-	-
20	3_6b	-	-	-	-	-	-

**Table 7. Summary of alignments for the FASTA searches of the PRT\_2010 database using putative polypeptide sequences encoded by the the genomic DNA-intervening DNA and/or intervening DNA-T-DNA 3' junctions in MON 87708.**

Appendix	Polypeptide	FASTA search of PRT_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	E-score
11	3_1a	-	-	-	-	-	-
12	3_1b	-	-	-	-	-	-
13	3_2	-	-	-	-	-	-
14	3_3a	-	-	-	-	-	-
15	3_3b	-	-	-	-	-	-
16	3_4	-	-	-	-	-	-
17	3_5a	-	-	-	-	-	-
18	3_5b	-	-	-	-	-	-
19	3_6a	-	-	-	-	-	-
20	3_6b	-	-	-	-	-	-

## **[CBI Cross Reference Number 2]**

Deleted Appendices 1-20

Deleted pages 31 – 86 are found in the Confidential Attachment, pages 9 – 64.