Study Title

Bioinformatics Evaluation of DNA Sequences Flanking the 5′ and 3′ Junctions of Inserted DNA in MON 87705: Assessment of Putative Polypeptides

Authors

Haidi Tu
Andre Silvanovich, Ph.D.

Study Completed On

May 1, 2009

Sponsor and Performing Laboratory

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

Laboratory Project ID

MSL Number: MSL0021929
Study Number: REG-09-088

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.

_____ Date: _____
Submitter

_____ Date: 5/1/09
Daniel J. Jenkins
Sponsor Representative

_____ Date: 5-1-09
Haidi Tu
Author

Summary of Quality Control Review

This report was checked to ensure that it accurately reflects the raw data of the study. The raw data was audited for compliance with the Monsanto Company Guidelines for Keeping Research Records (GRR September 2008, v.2), and where applicable, to Monsanto SOPs.

Date: 5/1/09

Quality Assurance Specialist
Monsanto Regulatory
Monsanto Company

Study Certification

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

Signatures of Final Report Approval:

Haidi Tu
Author

Date: 5-1-09

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

Date: May 1, 2009

Study Information

Study Number:	REG-09-088
Title:	Bioinformatics Evaluation of DNA Sequences Flanking the 5′ and 3′ Junctions of the Inserted DNA in MON 87705: Assessment of Putative Polypeptides
Facility:	Monsanto Company Product Characterization Center 800 North Lindbergh Blvd St. Louis, Missouri 63167
Protein Sciences Lead:	Gary A. Bannon, Ph.D.
Sponsor Representative:	Daniel J. Jenkins
Authors:	Haidi Tu Andre Silvanovich, Ph.D.
Supervisory Personnel:	Andre Silvanovich, Ph.D.
Study Start Date:	February 20, 2009
Study Completion Date:	May 1, 2009
Records Retention:	All study specific raw data and final report will be retained at Monsanto-St. Louis.

Table of Contents

Figures

Tables

Appendices

Abbreviations and Definitions

aa	Amino acid
AD_2009	Allergen, gliadin, and glutenin protein sequence database
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
E-Score	Expectation score
FAARP	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 sequence identification number
NCBI	National Center of Biotechnology Information at the National Institutes of Health, Bethesda, MD, USA
PRT_2009	A protein sequence database drived from GenBank realease 169
TOX_2009	Toxin protein sequence database

1.0 Summary

Monsanto Company has developed biotechnology-derived soybean, MON 87705, to generate nutritionally-improved soybean oil with decreased levels of saturated fats (16:0 palmitic acid and 18:0 stearic acid) and increased levels of oleic acid (18:1). Specifically, MON 87705 uses gene suppression technology to decrease the levels of two key oil biosynthetic enzymes, FATB and FAD2. Suppression of the FATB enzyme results in a decrease in the levels of saturated fats (16:0 palmitic acid and 18:0 stearic acid), while suppression of the FAD2 enzyme results in an increase of oleic acid (18:1).

In addition, MON 87705 also contains the 5-enolpyruvylshikimate-3-phosphate synthase gene derived from *Agrobacterium sp*. strain CP4 (*cp4 epsps*). Expression of the gene product (CP4 EPSPS) renders the plant tolerant to glyphosate, which is the active ingredient in the Roundup® family of agricultural herbicides. Glyphosate binds to the endogenous plant EPSPS enzyme and blocks the biosynthesis of shikimate-3-phosphate, thereby depriving plants of aromatic amino acids (Haslam, 1993; Steinrücken et al., 1984). The CP4 EPSPS protein is structurally similar and functionally identical to endogenous plant EPSPS enzymes, but has a much reduced affinity for glyphosate relative to endogenous plant EPSPS (Padgette et al., 1996). Introduction of the *cp4 epsps* gene into MON 87705 allows for the production of aromatic amino acids and other metabolites even in the presence of glyphosate (Padgette et al., 1996).

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′ soybean genomic DNA-inserted DNA junction and the 3′ inserted DNA-soybean genomic DNA junction 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_2009), toxin (TOX_2009), and public domain (PRT_2009) 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_2009, TOX_2009, and PRT_2009 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

® Roundup and Roundup Ready are registered trademarks of Monsanto Technology, LLC.

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 or toxins 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. The putative polypeptide 5_3 showed 217 alignments with *E*-scores less than 1×10^{-5} with the top alignment having an *E*-score of 2.7 X 10^{-21} that corresponds to 47.222% identity in a window of 144 amino acids for a RNA-directed DNA ploymerase from Medicago truncatula (GI number 124359710). Inspection of putative polypeptide 5_3 revealed that the soy genome likely contains a RNA-directed DNA polymerase pseudo-gene on the 5' flank of the transgene insertion site. None of these alignments indicate that the putative polypeptide 5_3 possesses any bioactive function that would cause adverse effects to animals or humans. In the unlikely occurance that the putative polypeptide 5_3 sequence was to be translated, it does not share homology with a toxin or other bioactive protein. 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 MON87705 occurs. Rather, the results of these bioinformatic analyses indicate that in the highly unlikely occurance 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 biotechnology-derived soybean, MON 87705, to generate nutritionally-improved soybean oil with decreased levels of saturated fats (16:0 palmitic acid and 18:0 stearic acid) and increased levels of oleic acid (18:1). Specifically, MON 87705 uses gene suppression technology to decrease the levels of two key oil biosynthetic enzymes, FATB and FAD2. Suppression of the FATB enzyme results in a decrease in the levels of saturated fats (16:0 palmitic acid and 18:0 stearic acid), while suppression of the FAD2 enzyme results in an increase of oleic acid (18:1).

In addition, MON 87705 also contains the 5-enolpyruvylshikimate-3-phosphate synthase gene derived from *Agrobacterium sp*. strain CP4 (*cp4 epsps*). Expression of the gene product (CP4 EPSPS) renders the plant tolerant to glyphosate, which is the active ingredient in the Roundup® family of agricultural herbicides. Glyphosate binds to the endogenous plant EPSPS enzyme and blocks the biosynthesis of shikimate-3-phosphate, thereby depriving plants of aromatic amino acids (Haslam, 1993; Steinrücken et al., 1980). The CP4 EPSPS protein is structurally similar and functionally identical to endogenous plant EPSPS enzymes, but has a much reduced affinity for glyphosate relative to endogenous plant EPSPS (Padgette et al., 1996). Introduction of the *cp4 epsps* gene into MON 87705 allows for the production of aromatic amino acids and other metabolites even in the presence of glyphosate (Padgette et al., 1996).

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′ soybean genomic DNA-inserted DNA junction and the 3′ inserted DNA-soybean genomic DNA junction 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_2009), toxin (TOX_2009), and public domain (PRT_2009) 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

® Roundup and Roundup Ready are registered trademarks of Monsanto Technology, LLC.

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 (Metcalfe 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 genomic DNA-inserted DNA 5′ junction and the inserted DNA-soybean genomic DNA 3′ junction of MON 87705. 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

The purpose of this study was to evaluate the amino acid sequences of putative polypeptides obtained from all reading frames that span the soybean genomic DNA inserted DNA T-DNA 5′ junction and the T-DNA inserted DNA soybean genomic DNA 3′ junction in MON 87705 to sequences in established databases. Sequences spanning these two 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 the soybean genomic DNA-inserted DNA 5′ junction and the inserted DNA-soybean genomic DNA 3′ junction, FASTA searches were performed on allergen (AD_2009), toxin (TOX_2009), and public domain (PRT_2009) 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_2009 database.

4.0 Methods

4.1 *Sequence Database Preparation:*

The allergen, gliadin, and glutenin sequence database (AD_2009) was obtained from FARRP (2009) [1] and was used as provided. The AD_2009 database contains 1,386 sequences. A complete description of the AD_2009 database can be found in Silvanovich (2009).

GenBank protein database, release 169.0 (December 16, 2008), was downloaded from NCBI and formatted for use in these bioinformatic analyses. It is refered to herein as the PRT_2009 database and contains 14,717,352 sequences. A complete description of the PRT_2009 database can be found in Silvanovich (2009).

The toxin database is a subset of sequences derived from the PRT_2009 database that was selected using a keyword search and filtered to remove likely non-toxin proteins. It is refered to herein as the TOX_2009 database and contains 7,651 sequences. A complete description of the TOX_2009 database can be found in Silvanovich (2009).

4.2 *Translation of Putative Polypeptides:*

DNA sequence spanning the 5′ and 3′ junctions of the MON 87705 insertion site (Skipwith et al., 2009) was analyzed for translational stop codons (TGA, TAG, TAA). All six possible reading frames originating or terminating within the MON 87705 insertion were translated using the standard genetic code from stop codon to stop codon.

[1] located at http://www.allergenonline.com

4.3 *Sequence Database Searches:*

FASTA analyses using the AD_2009, TOX_2009 and PRT_2009 databases were performed on a desktop computer loaded with a SUSE LINUX version 10.1 operating system and FASTA version 3.4t26 July 7, 2006. The DNA sequence was translated to the amino acid sequence with DNAStar, version 8.0.2 (13), 412 or SeqBuilder 8.0.2 (13) (Appendices 1-12). The structural similarity of the translated protein sequences to sequences in each database (AD_2009, TOX_2009, and PRT_2009) 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. Default FASTA comparison parameters for wordsize (*k-tuple*), gap creation penalty and gap extension penalty were used. The expectation threshold (*E*-score) limit was set to one. The *E*-score (expectation 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 less than 1e-5 (1×10^{-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 alignment is extensively analyzed for each database sequence. The BLOSUM matrix series (Henikoff and Henikoff, 1992) 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_2009). 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 (Metcalfe et al., 1996).

4.4 *Significance of the Alignment:*

An *E*-score of 1 e-5 (1×10^{-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 1 e-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 87705 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 87705 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 inserted DNA genomic DNA junctions and was greater than or equal to eight amino acids in length. At the 5′ and 3′ flanks, five deduced putative polypeptides spanned the genomic DNA inserted DNA 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 (see Figures 1 and 2 for reading frame assignment). Supporting dataset output files for each putative 5′ polypeptide are contained in Appendices 1-6, while dataset output files for each putative 3′ polypeptide are contained in Appendices 7-12.

5.1 *Assessment of Potential Allergenicity:*

The results of the allergenicity assessment are shown in Tables 2 and 5. Potential allergenicity of the twelve putative polypeptides was assessed using the FASTA and eight amino acid sliding window search algorithms. Using the FASTA algorithm to search the AD_2009 database, no alignments with any of the twelve query sequences generated an *E*-score of less than 1 e-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 with were found between any query sequence and the AD_2009 database. As a result, these twelve 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 twelve putative polypeptides was assessed using the FASTA algorithm. Using the FASTA algorithm to search the TOX_2009 database, no alignments with any of the twelve query sequences generated an *E*-score of less than 1 e-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 twelve putative polypeptides was assessed using the FASTA algorithm. Among all FASTA alignments between the twelve query sequences and the PRT_2009 database, one putative polypeptide (5_3) showed potentially significant alignments.

The putative polypeptide 5_3 showed 217 alignments with *E*-scores less than 1×10^{-5} with the top alignment having an *E*-score of 2.7 X 10^{-21} that corresponds to 47.222% identity in a window of 144 amino acids with an RNA-directed DNA polymerase from Medicago truncatula (GI number 124359710). Inspection of putative polypeptide 5_3 revealed that the soy genome likely contains a RNA-directed DNA polymerase pseudo-gene on the 5' flank of the transgene insertion site. None of these alignments indicate that the putative polypeptide 5_3 possesses any bioactive function that would cause adverse effects to animals or humans. In the unlikely occurance that the putative polypeptide 5_3 sequence was to be translated, it does not share homology with a toxin or other bioactive protein.

6.0 Conclusions

Analyses of putative polypeptides encoded by DNA spanning the 5′ and 3′ junctions of the MON 87705 inserted DNA were performed using bioinformatic tools. Results of the FASTA sequence alignments demonstrated a lack of structurally relevant similarity between any known allergenic or toxic proteins and the twelve putative polypeptides. The putative polypeptide 5_3 displayed 217 alignments with *E*-scores less than 1×10^{-5} . Inspection of the alignment with the lowest *E*-score revealed that the putative polypeptide 5_3 sequence does not share homology with toxic or other bioactive proteins. The results of these bioinformatic analyses demonstrate that even in the highly unlikely occurance 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) (2009). 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 Re*s **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.

Haslam, E. 1993. Shikimic Acid: Metabolism and Metabolites. John Wiley and Sons, Chichester, England.

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* Mar **227**:1435-1441.

Metcalfe, D.D., Astwood, J.D., Townsend, R., Sampson, HA., 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.

Padgette, S. R., Re, D. B., Barry, G. F., Eichholtz, D. E., Delannay, X., Fuchs, R. L., Kishore, G. M., and Fraley, R. T. 1996. New Weed Control Opportunities: Development of Soybeans with a Roundup Ready Gene. CRC Handbook. 4:53-84.

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.

Silvanovich, A. (2009). The assembly of AD_2009, TOX_2009 and PRT_2009. Monsanto Technical Report MSL-0021840, St. Louis, MO.

Skipwith, A., Lawry, K.R., Tian Q., and Masucci J.D. (2009). Molecular Analysis of Soybean MON 87705. Monsanto Technical Report MSL-0021371, St. Louis, MO.

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

Steinrücken, H. and Amrhein, N. 1984. 5-Enolpyruvylshikimate-3-Phosphate Synthase of *Klebsiella* Pneumoniae. Eur.J.Biochem. 143:351-357.

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.

[CBI Cross Reference Number 1]

Deleted Figures 1, 2 and 3 and Table 1

Deleted pages 23 – 26 are found in the Confidential Attachment, pages 5 – 8.

Table 2. Summary of alignments for the FASTA searches of the AD_2009 database using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 5′ junction in MON 87705.

Appendix	Polypeptide	AD_2009 Sequence Database						
		Sliding Window	FASTA search					
		# Hits	# Hits	GI #	Description	*E*-score	% Identity	aa Overlap
1	5_1	No	-	-	-	-	-	-
2	5_2	No	-	-	-	-	-	-
3	5_3	No	6	232054	ENO1_CANAL RecName: Full=Enolase (440 aa)	0.89	30.508	59
4	5_4	No	-	-	-	-	-	-
5	5_5	No	2	75062228	ALL4_FELCA RecName: Full=Aller (186 aa)	0.76	29.730	37
6	5_6	No	-	-	-	-	-	-

Table 3. Summary of alignments for the FASTA searches of the TOX_2009 database using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 5′ junction in MON 87705.

Appendix	Polypeptide	FASTA search of TOX_2009 Sequence Database					
		# Hits	GI #	Description	*E*-score	% Identity	aa Overlap
1	5_1	5	209540528	Streptococcal pyrogenic ex (236 aa)	0.066	30.612	49
2	5_2	-	-	-	-	-	-
3	5_3	1	163664913	addiction module toxin, Re (93 aa)	0.71	28.571	63
4	5_4	-	-	-	-	-	-
5	5_5	-	-	-	-	-	-
6	5_6	-	-	-	-	-	-

Table 4. Summary of alignments for the FASTA searches of the PRT_2009 database using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 5′ junction in MON 87705.

Appendix	Polypeptide	FASTA search of PRT_2009 Sequence Database					
		# Hits	GI #	Description	*E* score	% Identity	aa Overlap
1	5_1	-	-	-	-	-	-
2	5_2	-	-	-	-	-	-
3	5_3	217[1]	124359710	RNA-directed DNA polymeras (1297 aa)	2.7e-21	47.222	144
4	5_4	-	-	-	-	-	-
5	5_5	-	-	-	-	-	-
6	5_6	-	-	-	-	-	-

Table 5. Summary of alignments for the FASTA searches of the AD_2009 database using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 3′ junction in MON 87705.

Appendix	Polypeptide	AD_2009 Sequence Database						
		Sliding Window	FASTA search					
		# Hits	# Hits	GI #	Description	*E*-score	% Identity	aa Overlap
7	3_1	No	-	-	-	-	-	-
8	3_2	No	-	-	-	-	-	-
9	3_3	No	2	71057064	thaumatin-like protein [Ac (225 aa)	0.83	40.000	25
10	3_4	No	2	169969	glycinin (516 aa)	0.69	57.143	7
11	3_5	No	-	-	-	-	-	-
12	3_6	No	-	-	-	-	-	-

[1] Only the top 50 alignments are shown in the appendix.

Table 6. Summary of alignments for the FASTA searches of the TOX_2009 database using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 3´ junction in MON 87705.

Appendix	Polypeptide	FASTA search of TOX_2009 Sequence Database					
		# Hits	GI #	Description	*E*-score	% Identity	aa Overlap
7	3_1	-	-	-	-	-	-
8	3_2	-	-	-	-	-	-
9	3_3	-	-	-	-	-	-
10	3_4	-	-	-	-	-	-
11	3_5	1	158635887	cytotoxic polypeptide [Mi (222 aa)	0.75	31.250	48
12	3_6	1	256378	neurotoxin Tx2-9 [Phoneutria (32 aa)-	0.66	43.750	16

Table 7. Summary of alignments for the FASTA searches of the PRT_2009 database using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 3´ junction in MON 87705.

Appendix	Polypeptide	FASTA search of PRT_2009 Sequence Database					
		# Hits	GI #	Description	*E* score	% Identity	aa Overlap
7	3_1	-	-	-	-	-	-
8	3_2	-	-	-	-	-	-
9	3_3	6	124360394	Peptidase aspartic, active (435 aa)	0.0004	46.875	32
10	3_4	-	-	-	-	-	-
11	3_5	-	-	-	-	-	-
12	3_6						

[CBI Cross Reference Number 2]

Deleted Appendices 1-12

Deleted pages 30 – 83 are found in the Confidential Attachment, pages 9 – 62.