Study Title

Bioinformatics Evaluation of Genomic DNA Flanking CP4 EPSPS DNA Segments in Roundup Ready Soybean Event 40-3-2: Assessment of Putative Genetic Regulatory Elements and Putative Polypeptides

Authors

Study Completed On

June 6, 2000

Performing Laboratory

Monsanto Company
Product Safety Center
Biotechnology Regulatory Sciences
700 Chesterfield Parkway North
St. Louis, MO, USA

Laboratory Project ID

MSL Number: 16748

Study Number: 00-01-30-21

Table of Contents

Appendices

1.0 Summary

Development of DNA detection methods and further associated molecular characterization of the Roundup Ready® soybean event 40-3-2 has recently revealed two segments of DNA (a 72 bp and 250 bp segment) containing segments of the CP4 EPSPS coding region. To assess the potential for proximal genetic regulatory elements such as transcriptional promoters and terminators, the DNA sequences flanking these two segments were compared to known plant promoter and polyadenylation signals available in public domain databases. In addition, to assess the potential similarity towards allergens, toxins or other pharmacologically active proteins, putative polypeptides derived from the DNA sequences containing the 72 bp and 250 bp CP4 EPSPS DNA segments were translated and analyzed using bioinformatics tools. Each reading frame was translated, yielding deduced amino acid sequences spanning the CP4 EPSPS coding region and included contiguous flanking genomic sequences.

Bioinfomatics algorithms were used to assess the overall sequence identity to genetic regulatory elements and to assess the similarity of putative polypeptides encoded by potential open reading frames. For genetic elements, 100% sequence identity was considered biologically relevant. For polypeptides, a sequence similarity may indicate sequence homology (i.e., representing a sequence derived from a common ancestor gene with potentially homologous function). Sequences that share extensive amino acid sequence identity and/or similarity[1] throughout the entire alignment or identified domains are considered to be biologically relevant homologues.

In addition to structural similarity, each putative polypeptide was screened for immunologically relevant similarity using a pair-wise comparison algorithm. In these analyses, eight linearly contiguous and identical amino acids were defined as immunologically relevant, where eight represents the typical minimum sequence length likely to represent an IgE epitope.

No 100% identical DNA sequences were identified in the publicly available genetic element databases. Several polyadenylation signals were observed, but none were in the appropriate context of any putative open reading frames. These analyses demonstrated that both the 250 bp and 72 bp segments of CP4 EPSPS DNA are not likely to be proximal to the necessary genetic elements required to define a potentially active transcription unit, or gene. The necessary genetic elements include the presence of a 5' proximal promoter, a defined mRNA transcript containing a complete coding sequence with start and stop codons, and a transcriptional terminator which contains an mRNA polyadenylation signal. No such arrangement was observed in the soybean genomic DNA flanking the 250 and 72 bp CP4 EPSPS DNA segments. Furthermore, these data corroborate the previous conclusion obtained by northern blotting that

[1] A non-identical but physicochemically related amino acid is defined as a similar amino acid. Similar amino acids are structurally related (e.g. sharing polar, hydrophobic, or charged states). Such substitutions are referred to as "conservative" since they are unlikely to change the structures, and by inference the functions, of homologous proteins.

no mRNA transcripts containing either the 250 or 72 bp CP4 EPSPS DNA segments is expressed in RR soybean event 40-3-2.

No immunologically relevant sequence similarity is shared between putative polypeptides derived from the CP4 EPSPS segments (including those encoded by immediately proximal flanking genomic DNA) and the proteins in the allergen database. Further, no biologically relevant structural similarity to allergens, toxins or other proteins in public domain databases was observed. Similarities to homologous EPSPS proteins were identified when the FASTA alignment tool was used to assess the bona fide CP4 EPSPS coding frame of both the 72 and 250 bp segments.

The results of these bioinformatics analyses demonstrate the absence of a clearly defined transcriptional unit containing the 250 or 72 bp CP4 EPSPS DNA segments. In addition, these results demonstrate that polypeptides potentially encoded by the alternative reading frames of the 72 bp and 250 bp CP4 EPSPS DNA segments in combination with contiguous flanking genomic DNA do not share immunologically relevant sequences and are not structurally homologous to allergens, toxins or any proteins associated with animal and human health risks.

2.0 Methods

Data analysis. Nucletide bioinformatics analyses were carried out using software supplied by Genetics Computer Group (GCG), Wisconsin Package, version 10.0, Madison, Wisconsin, or other web-based tools. Promotor databases searched included PLACE (245 sequences, plant promotors), TRANSFEC (150 sequences, plant promotors) and DERWENT (442,744 sequences, patent database). All peptide bioinformatics analyses were carried out using Databases searched included the current version of ALLPEPTIDES (942,655 sequences representing SwissProt release 38.0, GenBank release 116.0 and a biweekly update of TrEMBL) as well as two subset databases that were assembled manually, TOXIN4 (4,677 sequences) and UPDATE2 (567 allergen sequences).

Promotor analysis. The complete DNA sequence containing both the 72 bp segment of CP4 EPSPS (Figure 1) and the 250 bp segment of CP4 EPSPS (Figure 2) were compared to databases containing plant promotor sequence elements. The DERWENT database was searched using the BLASTN sequence alignment tool and the PLACE and TRANSFEC databases were searched using a pattern match tool.

Polyadenylation signal analysis. The EditSeq program of DNAstar (version 3.99) was used to search for exact word matches to each genetic element described (Table 1).

Translation of putative peptides. DNA sequences spanning the 72 bp segment of CP4 EPSPS (Figure 1) and the 250 bp segment of CP4 EPSPS (Figure 2) were analyzed for start codons, stop

codons and open reading frames. All six possible reading frames originating or teminating within the 72 bp segment and extending into or from genomic sequence were translated using the standard genetic code. The three forward reading frames originating within the 250 bp segment and terminating in the 3'-flanking genomic sequence were translated using the standard genetic code (see Table 2 for details). To maximize the number of putative polypeptides, a start codon (ATG) was not used to define the usual N-termini of the translated polypeptides, rather a stop codon (TGA, TAG, TAA) was used to define both the N termini and the C-termini of each polypeptide.

Database preparation. The TOXIN4 database was previously assembled from public domain databases (Genbank and EMBL GenPept version 108, PIR and NRL3D version 56 and SwissProt version 36) and has been previously described (Hileman and Astwood, 1999). Protein sequences were retrieved using the STRINGSEARCH function (keyword = toxin) of GCG. Individual toxin sequences were compiled into a database using DATASET and the database named TOXIN4. The allergen database was also assembled from public domain databases STRINGSEARCH (keyword = allergen) and combined with sequences obtained from literature and Internet searches (Hileman and Astwood, 1999).

Database searches. The structural similarity of each putative peptide towards sequences in each database (ALLPEPTIDES, TOXIN4 and UPDATE2) was assessed using the FASTA algorithm (Pearson and Lipman, 1988). Although it was redundant to search both the TOXIN4 and ALLPEPTIDES databases for potential similarity towards protein toxins, the ALLPEPTIDES database search is performed to reveal potential similarity towards pharmacologically active proteins. In addition to the FASTA comparisons of each putative peptide towards allergens (to assess overall structural similarity), an 8-mer search was performed. The algorithm (IDENTITYSEARCH) was developed to identify whether or not a linearly contiguous match of 8 amino acids existed between the a query sequence and sequences within the allergen database (UPDATE2). The algorithm was run from a UNIX terminal window in GCG. This program compares the query sequence to each protein sequence in the allergen database using a sliding window of 8 amino acids. An epitope of 8 amino acids was chosen to represent the smallest typical immunologically significant IgE binding epitope (Metcalfe, et al., 1996).

3.0 Results and Discussion

Development of DNA detection methods and further associated molecular characterization of the Roundup Ready® soybean event 40-3-2 has recently revealed two segments of DNA (a 72 bp and 250 bp segment) containing the CP4 EPSPS coding region (Lirette, et al., 2000). Bioinformatics analyses were performed on both the DNA to assess the potential for proximal genetic regulatory elements such as transcriptional promoters and terminators, the DNA sequences flanking these two segments were compared to known plant promoter and polyadenylation signals available in public domain databases. In addition, to assess the potential similarity towards allergens, toxins or other pharmacologically active proteins, putative polypeptides derived from the DNA sequences containing the 72 bp and 250 bp CP4 EPSPS DNA segments were translated and analyzed using bioinformatics tools. Each reading frame was translated, yielding deduced amino acid sequences spanning the CP4 EPSPS coding region and included contiguous flanking genomic sequences.

Promotor analysis. The complete output files are shown in Appendices 2-4. Using the entire DNA sequences shown in Figures 1A and 2 as query sequences, three databases were separately searched for similarity towards known genetic promotor elements. No promotor elements were identified (Appendix 4) to any of the 860 bp of DNA containing the 250 bp DNA segment of CP4 EPSPS in the TRANSFAC and PLACE databases. The 1103 bp segment of DNA containing the 72 bp DNA segment of CP4 EPSPS yielded two hits. One was a degenerate element sequence associated with phenylalanine ammonia-lyase (YTYYMMCMAMCMMC). The second element is associated with sucrose regulation (SURE2) of the patatin protein of *Solanum tuberosum* (potato). This sucrose regulated promoter element is always observed to have a cytosine residue at the fifth position in all patatin gene promoters (i.e. AATACTAAT) that have been described. Examination of the sequences analyzed from the genome of soybean event 40-3-2 do not show the presence of this cytosine residue at this position. Additionally, the SURE2 motif which regulates patatin expression in potato is located -184 to -156 bp from the start of transcription. Examination of the motifs in the DNA sequence being analyzed does not place them in a favorable context for any of the open reading frames that were identified. Finally, the SURE2 motif is present as a single copy. The SURE2 like motifs identified in the 1103 bp DNA segment from soybean event 40-3-2 were identified in multiple tandem copies. This is likely the result of non-specific sequence similarity due to the AT-rich nature of the DNA sequence being exanmined.

Possible polyadenylation signals were also analyzed and are summarized in Table 1. However, compared to our understanding of other eukaryotes, the polyadenylation signal(s) that occur in plants are not well defined (Rothnie, 1996). Bioinformatics analyses of the 3'-end of rice and arabidopsis expressed sequence tags indicates that plant polyadenylation signals consist of multiple elements (Graber, et al., 1999). Similar to yeast, plant polydenylation signals are generally comprised of three sequence regions upstream of the cleavage site (a uracil alternating purine-rich upstream element, the positioning element such as the canonical AAUAAA sequence

followed by a uracil-rich element). Following the cleavage site, a uracil-rich domain is present. Although the adenosine-rich hexamer AAUAAA is found in plants, it is present in less than 15% of polyadenylation signals in plants. Only the most dominant recently reported (Graber, et al., 1999) sequence elements were considered for this analysis. An upstream element (UUGUAU or UUGUAA) approximately 60 bp upstream of the positioning element (AAUAAA or AAUGAA) followed by a uracil-rich sequence.

A complete set of polyadenylation signals was not observed in either reading frame of the 72 bp DNA segment or the positive DNA strand of the 250 bp DNA segment of CP4 EPSPS.

Table 1. Polyadenylation signal analysis. Values reported represent the number of times the sequence occurred with the base positions shown in parentheses.

Upstream Element		Positioning Element		Uracil-rich Element		
UUGUAU	UUGUAA	AAUAAA	AAUGAA	UUUUCU	UUUUUU	Complete Signal
72 bp Segment (forward reading frames)						
0	0	3 (271-276) (819-824) (1059-164)	0	1 (310-315)	(813-818)	NO
72 bp Segment (reverse reading frames)						
0	1 (77-82)	1 (842-847)	2 (725-730) (793-798)	1 (876-881)	1 (874-879)	NO[1]
250 bp Segment (forward reading frames)						
0	0	4 (551-556) (582-587) (825-830) (832-837)	0	0	4 (571-576) (607-612) (608-613) (609-614)	NO

[1] Out of context. The upstream element was >>60 bp away from the positioning element(s).

Polypeptide analyses. Putative peptides derived from DNA sequences containing the 72 bp and 250 bp CP4 EPSPS DNA segments were translated and analyzed using bioinformatics tools to assess potential similarity towards allergens, toxins or other pharmacologically active proteins. Each reading frame was translated, yielding deduced amino acid sequences spanning the CP4 EPSPS coding region into the flanking genomic sequence. Although some of these reading frames contained a start codon (ATG), translation was performed from stop-to-stop codon as shown in Table 1.

Table 2. Reading frames derived from the CP4 EPSPS segments including codons derived from contiguous flanking genomic DNA. Each deduced polypeptide sequence was evaluated using bioinformatics tools for sequence similarities towards known allergens, toxins and other pharmacologically active proteins. The shaded region of each sequence corresponds to amino acids deduced from the CP4 EPSPS sequence; the non-shaded regions correspond to amino acids deduced from the flanking soybean genomic DNA sequences.

72 bp CP4 EPSPS DNA Segment

Reading Frame	bp[a]	# start codons	Deduced Peptide Sequence
1	447-759	1	1 QRRISGGRRG LTAAAGTRSS IIEGARSSG TVTPFSVEER TRRVICFEIV 51 RPAPSPRSIC FNKNHTGVDL IAVISIETRC SCKDVNDTID LPLDKTNLID 101 QMF
2	482-541	0	1 LQRQARGARR SSKARGLPAP
	545-577	0	1 RPSAWRSERA G
3	360-587	1	1 SVLEGSLEGE RGEERQSAME RRKGSGSFLT TEDLRRETGI NCSGRHEELV 51 DHRRRAVFRH RDALQRGGAN AQGNLF
4	638-378	1	1 FLLKHILLGE GAGRTISKQI TLRVRSSTLK GVTVPEDRAP SMIDELLVPA 51 AAVNPRLPPE ILRCQEAAGT LAPLHCRLPL LAPLSLQ
5	628-554	0	1 SIYSSGRVQG GRFQNRLPCA FAPPR
	550-515	0	1 RASRCRKTAR LR
	511-317	1	1 STSSSCLPLQ LIPVSRLRSS VVKKLPEPLR RSIADCLSSP LSPSNEPSRT 51 LQVSLSLRDS VSHWM
6	762-481	0	1 FLKHLIYKVS FIKRQVNRVV HIFARAPRFY ANYCYQIYPC VIFVEAYTPR 51 GGCRADDFKT DYPARSLLHA EGRHGAGRPR AFDDRRAPRA CRCS

250 bp CP4 EPSPS DNA Segment

Reading Frame	bp[b]	# start codons	Deduced Peptide Sequence
1	139-516	6	1 DKLSRAVSSM LLDRGSIPHR SFMFGGLASG ETRITGLLEG EDVINTGKAM 51 QAMGARIRKE GDTWIIDGVG NGGLLAPEAP LDFGNAATGC RLTMGLVGVY 101 DFKRIMLGNF SEIISIFLGI SAVTGE
2	416-478	0	1 PWASSGSTIS SASCWEILAR L
3	372-477	0	1 GAARFRQCRH GLPPDHGPRR GLRFQAHHAG KF

[a] Positions as shown in Figure 4.
[b] Positions as shown in Figure 5.

The FASTA algorithm was used to assess overall structural similarity. A biologically relevant sequence similarity may indicate sequence homology (a sequence derived from a common ancestor gene and potentially homologous function). Sequences that share extensive amino acid sequence identity and/or similarity (a structurally related amino acid replacement, e.g. polar, hydrophobic, charged) throughout the entire alignment or domain are considered biologically relevant.

In addition to structural similarity, each putative peptide was screened for immunologically relevant similarity. Eight linearly contiguous amino acid identities were defined as an immunologically relevant sequence, the typical minimum sequence length likely to represent an IgE epitope.

A summary of the total sequences identified in these bioinformatics analyses is shown in Appendix 1. The complete data output files for each analysis were assembled as Appendices 5 to 16 in this report. No immunologically relevant sequence was shared between a putative polypeptide and the allergen database. Further, no biologically relevant structural similarity was observed between a putative polypeptide and the allergen database (UPDATE2). The observed alignments were generally short (less than 40 amino acids), were gapped to achieve optimal alignment and had poor expectation scores (E()-scores) of approximately 1 or greater.

The best expectation score (E()-value of 0.044) was observed in reading frame 4 of the 72 bp CP4 EPSPS DNA segment (Appendix II-4) to bovine beta-casein (Accession No. M15132) and is shown in Figure 3. This protein shares 29.8% identical residues and 59.6% similar residues within a 47 amino acid overlap. The level of similarity is not biologically relevant (Doolittle, 1990) and does not indicate structural homology.

No biologically significant similarity was observed between a putative peptide sequence and sequences within the toxin database (TOXIN4). The best expectation score (E()-value of 0.021) was observed in reading frame 2 (bp 482-541) of the 72 bp CP4 EPSPS DNA segment (Figure 4) to the *Pasteurella haemolytica* coproporphyrinogen oxidase protein (Accession No. U46781). Although there are health concerns in patients with malfunctioning coproporphyrinogen oxidase enzyme, this protein is not itself a toxin. The TOXIN4 database was assembled using a keyword search and contains some irrelevant entries due to a reference to the word “toxin” in the annotation section of the file. Other entries identified were also irrelevant entries (Accession No. JC4049, polygalacturonidase, E-score of 0.25; Accession No. P27883, P27884, A38195 and M77235, ion channel proteins, E-scores of 0.21-0.56). The best expectation score to a known protein toxin was to the sea snake neurotoxic venom protein, erabutoxin (Accession No. N1LT1E, E-score of 0.45). Identified by a FASTA search to reading frame 1 of the 72 bp CP4 EPSPS DNA segment (Appendix II-1), the shared similar sequence in this entry was derived *entirely* from the soybean genomic DNA (see below, CP4 EPSPS-derived sequence is shaded). This protein shared 33.3% identical residues and 55.6% similar residues within the 54 amino acid overlap. Regardless of the origin of this putative peptide, the neurotoxin contains 4

disulfide bonds (Sato. et al., 1971), which are not observed in the putative peptide sequence and are necessary to maintain protein structure. Thus this level of similarity is not biologically relevant (Doolittle, 1990) and does not indicate structural homology.

No biologically significant similarity was observed between a putative peptide sequence and sequences found in the ALLPEPTIDES database, other than those expected. As previously observed (Hileman and Astwood, 1999), homologous EPSPS proteins were identified when the FASTA alignment tool is used in the proper CP4 EPSPS reading frame (Appendix 14). EPSPS proteins are ubiquitous, occurring in all plants and some algal and fungal species (Padgette, et al., 1996). The best expectation score (E()-value of 0.85) observed towards a protein other than an EPSPS was to reading frame 5 (bp 628-554) of the 72 bp CP4 EPSPS DNA segment (Figure 5). The human Ny-Ren-24 antigen (Accession No. Q9Y5A4) shared 48.0% identical residues and 76.0% similar residues within a 25 amino acid overlap (shown below). This protein was identified as a renal tumor carcinoma antigen (Scanlan, et al., 1999). The observed overlap is likely too short to represent a bona fide homology. Thus this level of similarity does not indicate structural homology (Doolittle, 1990).

4.0 Conclusions

The results of these bioinformatics analyses demonstrate that genetic elements necessary for transcription of a message are either not present or out of context for functionality based on our current understanding of plant molecular biology. Further, if a putative peptide were encoded from either the 72 bp or 250 bp CP4 EPSPS DNA segments, these polypeptides do not share immunologically relevant sequences and are not structurally homologous to known allergens, toxins or any proteins associated with animal and human health risks.

References

Doolittle, R. F. 1990. Searching through sequence databases. Methods in Enzymology 183, 99-110.

Graber, J. H., Cantor, C. R., Mohr, S. C. and Smith, T. F. 1999. *In silico* detection of control signals: mRNA 3'-end processing signal in diverse species. Proc Natl Acad Sci, USA 96, 14055-14060.

Hileman, R. E. and Astwood, J. D. 1999. Bioinformatics Analysis of CP4 EPSPS Protein Sequence Utilizing an Allergen Database. Study number 99-01-62-07, MSL-16267, an unpublished study conducted by Monsanto Company.

Hileman, R. E. and Astwood, J. D. 1999. Bioinformatics Analysis of CP4 EPSPS Protein Sequence Utilizing Toxin and Public Domain Genetic Databases. Study number 99-01-62-08, MSL-16268, an unpublished study conducted by Monsanto Company.

Lirette, R. P., 2000. Further Characterization of Roundup Ready® Soybean Event 40-3-2. Study number 99-01-30-22, MSL-16646, an unpublished study conducted by Monsanto Company.

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, S165-86.

Padgette, S., Re, D., Eichholtz, D., Delannay, X., Fuchs, R., Kishore, G. and Fraley, R. 1996. New weed control opportunities: Development of soybeans with a Roundup Ready™ gene, Duke, S. O., ed. CRC, Boca Raton, FL.

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

Rothnie, H. M. 1996. Plant mRNA 3'-end formation. Plant Mol Biol 32, 43-61.

Scanlan, M. J., Gordan, J. D., Williamson, B., Stockert, E., Bander, N. H., Jongeneel, V., Gure, A. O., Jager, D., Knuth, A., Chen, Y. T. and Old, L. J. 1999. Antigens recognized by antibody in patients with renal-cell carcinoma. Int J Cancer 83.

Sato Y. E., , S., Ishii, S. and Tamiya, N. 1971. The disulfide bonds of erabutoxin a, a neurotoxic protein of a sea snake (*Laticauda semifasciata*) venom. Biochem J 122, 463-467.

[

]

[

]

```
  1 TTTCTGTTGA ATACGTTAAG CATGTAATAA TTAACATGTA ATGCATGACG
 51 TTATTTATGA GATGGGTTTT TATGATTAGA GTCCCGCAAT TATACATTTA
101 ATACGCGATA GAAAACAAAA TATAGCCGCG CAAACTAGGA TAAATTATCG
151 CGCGCGGTGT CATCTATGTT ACTAGATCGG GGATCGATCC CCCACCGGTC
201 CTTCATGTTC GGCGGTCTCG CGAGCGGTGA AACGCGCATC ACCGGCCTTC
251 TGGAAGGCGA GGACGTCATC AATACGGGCA AGGCCATGCA GGCCATGGGC
301 GCCAGGATCC GTAAGGAAGG CGACACCTGG ATCATCGATG GCGTCGGCAA
351 TGGCGGCCTC CTGGCGCCTG AGGCGCCGCT CGATTTCGGC AATGCCGCCA
401 CGGGCTGCCG CCTGACCATG GGCCTCGTCG GGGTCTACGA TTTCAAGCGC
451 ATCATGCTGG GAAATTTTAG CGAGATTATA AGTATCTTCC TGGGGATCTC
501 TGCTGTTACT GGTGAATAGT GAGACAGAGT CTTCTGAGCT CATAGGATAA
551 AATAAATTAT AATTAGTAAA TTTTTTAATT AAATAAATCA ATTACTTCAT
601 AAATAATTTT TTTTATAGAA TATGTTGACA TTCTAGCTGG ATATAGAACT
651 AATATAAAGA AACCTTAAAA ATTTTGTTTG GAAGAATATG TTATTGAAAG
701 ACAAATCTAA TTAAGTTTAT CAGGGTCATT TGTTGAAGAT AGGAAACCTT
751 CAGCAATTTG AATATTAAGT AACTGCTTCT CCCAGAATGA TCGGAGTTTC
801 TCCTCCTGCT ATTACATGAA AAAAAATAAA AAATAAAAAA AAGATAAGAT
851 TAAGCTTCAA
```

Figure 2. DNA sequence of the 250 bp CP4 EPSPS segment and the 3'-flanking genomic soybean sequence. The shaded sequence (195-444) corresponds to the 250 bp segment of CP4 EPSPS. Bases 1-194 corresponds to the 3' portion of the NOS 3' transcriptional termination element and plasmid PV-GMGT04 sequence. Bases 445-860 corresponds to the 3'-flanking genomic soybean sequence.

Figure 3 Best similarity to a protein in the allergen database (UPDATE2).

```
SCORES   Init1: 36    Initn: 36    Opt: 79    z-score: 119.2 E(): 0.044
Smith-Waterman score: 79;    29.8% identity in 47 aa overlap

                 10        20        30        40        50        60
frame4.pep   KHILLGEGAGRTISKQITLRVRSSTLKGVTVPEDRAPSMIDELLVPAAAVNPRLPPEILR
                                          :| :  |: |  |    ::| | | ||:|
BOVCASB_1    EEQQQTEDELQDKIHPFAQTQSLVYPFPGPIP-NSLPQNIPPLTQTPVVVPPFLQPEVLG
                    60        70        80         90        100

                 70        80
frame4.pep   CQEAAGTLAPLHCRLPLLAPLSLQ
              :::  ::|| | ::|:
BOVCASB_1    VSKVKEAMAPKHKEMPFPKYPVEPFTESQSLTLTDVENLHLPPLLLQSWMHQPHQPLPPT
           110       120       130       140       150       160
```

Figure 4 Best Similarity to a protein in the toxin database (TOXIN4).

```
SCORES   Init1: 45    Initn: 45    Opt: 74    z-score: 117.6 E(): 0.45
Smith-Waterman score: 74;    33.3% identity in 54 aa overlap

                  20        30        40                50        60
frame1.pep   AGTRSSSIIEGARSSGTVTPFSVEERTRRVICF-------EIVRPAPSPRSICFNKNHTG
                                          |||       : ::  || :| |:||: :
N1LT1E                                  RICFNQHSSQPQTTKTCPSGESSCYNKQWS-
                                                10        20        30

               70        80        90       100
frame1.pep   VDLIAVISIETRCSCKDVNDTIDLPLDKTNLIDQMF
              |: ::| ||  |:|  |:  | |
N1LT1E       -DFRGTI-IERGCGCPTVKPGIKLSCCESEVCNN
                     40        50        60
```

Figure 5 Best similarity to a protein in the GenBank or SwissProt (ALLPEPTDES).

```
SCORES   Init1: 45    Initn: 45    Opt: 82    z-score: 154.0 E(): 0.85
Smith-Waterman score: 82;    48.0% identity in 25 aa overlap

                                             10        20
frame5_1.pep                              SIYSSGRVQGGRFQNRLPCA--FAPPR
                                            :: ||::| |||:| | :  :||||
Q9Y5A4       IFYPDLIDKRSTPEYFLEACADNKDFAILRFTRGRLRGHRFQDRQPRVGILAPPRLPLPV
              150       160       170       180       190       200
```

Appendix 1. Number of sequences identified in each bioinformatics analysis.

72 bp DNA Segment				
	Allergens		Structural Similarity	
Reading Frame (bp)	Epitope Matching	Structural Similarity	Toxins	All
1 (447-759)	0	11	21	0
2 (482-541)	0	28	22	0
2 (545-577)	0	14	20	5
3 (360-587)	0	28	3	5
4 (638-378)	0	18	19	0
5 (628-554)	0	20	19	3
5 (550-515)	0	14	19	9
5 (511-317)	0	11	10	0
6 (762-481)	0	14	14	0
250 bp DNA Segment				
	Allergens		Structural Similarity	
Reading Frame (bp)	Epitope Matching	Structural Similarity	Toxins	All
1 (139-516)	0	7	26	81
2 (416-478)	0	9	10	0
3 (372-477)	0	18	38	0

APPENDICES 2-16