Study Title

Bioinformatics Evaluation of DNA Sequences Flanking the 5′ and 3′ Junctions of Inserted DNA in MON 88302: Assessment of Putative Polypeptides

Authors

[redacted]

Study Completed On

November 16, 2010

Sponsor and Performing Laboratory

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

Laboratory Project ID

MSL Number: MSL0023088
Study Number: REG-10-526

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:_____

EE
HT
9-23-11

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

[redacted]
Sponsor Representative

[redacted]
Author

EE
HT
9-23-11

Quality Assurance Statement

Study Title: Bioinformatics Evaluation of DNA Sequences Flanking the 5′ and 3′ Junctions of Inserted DNA in MON 88302: Assessment of Putative Polypeptides

Study Number: REG-10-526

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
11/11/2010	Draft Report Review	11/15/2010	11/15/2010

[redacted]

Quality Assurance Unit **Date**
Monsanto Regulatory, Monsanto Company

EE
HT
9-23-11

Study Certification

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

Signatures of Final Report Approval:

[redacted]

Author

[redacted]

Lead, Regulatory Product Characterization Center

Study Information

Study Number: REG-10-526

Title: Bioinformatics Evaluation of DNA Sequences Flanking the 5′ and 3′ Junctions of Inserted DNA in MON 88302: Assessment of Putative Polypeptides

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

Protein Sciences Lead: [redacted]

Sponsor Representative: [redacted]

Authors: [redacted]

Study Start Date: September 15, 2010

Study Completion Date: November 16, 2010

Records Retention: The protocol, all raw data, documentation, records, and the final report for this study are retained at Monsanto Company.

Table of Contents

Figures

Tables

Appendices

Abbreviations and Definitions

AA	Amino acid
AD_2010	Allergen and gliadin protein sequence database (Release date January 22, 2010)
a.e.	Acid equivalents
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
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 a second generation herbicide-tolerant canola product, MON 88302, that is tolerant to in crop glyphosate application(s) from emergence to first flowering at a rate up to 1,800 g a.e. per hectare. With an increased window of application and higher spray rates, MON 88302 will provide superior weed control compared to the commercial first generation Roundup Ready® canola product RT73 (also referred to as GT73).

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′ canola genomic DNA- inserted DNA junctions. Besides the T-DNA sequence, a short intervening sequence was inserted into MON 88302 at the 3′ junction. Therefore, sequences spanning the 5′ canola genomic DNA-T-DNA and the 3′canola genomic DNA-intervening DNA and/or intervening DNA-T-DNA junctions were translated from stop codon to stop codon in all six reading frames. A total of 11 putative polypeptides of 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.

The bioinformatic analysis performed using the 11 putative sequences translated from junctions is theoretical as there is no reason to suspect, or evidence to indicate, the presence of transcripts spanning the flank junctions. The results of bioinformatic analysis indicate that no structurally relevant sequence similarities were observed between the 11 putative flank junction derived sequences and allergens, toxins or biologically active proteins. As a result, in the unlikely occurrence that any of the 11 peptides analyzed herein is found *in planta*, none would share significant similarity or identity to known allergens, toxins, or other biologically active proteins that could affect human or animal health.

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

2.0 Introduction

Monsanto Company has developed a second generation herbicide-tolerant canola product, MON 88302, that is tolerant to in crop glyphosate application(s) from emergence to first flowering at a rate up to 1,800 g a.e. per hectare. With an increased window of application and higher spray rates, MON 88302 will provide superior weed control compared to the commercial first generation Roundup Ready® canola product RT73 (also referred to as GT73).

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′ canola genomic DNA-inserted DNA junctions. Besides the T-DNA sequence, a short intervening sequence was inserted into MON 88302 at the 3′ junction. Therefore, sequences spanning the 5′ canola genomic DNA-T-DNA and the 3′canola genomic DNA-intervening DNA and/or intervening DNA-T-DNA junctions were translated from stop codon to stop codon in all six reading frames. A total of 11 putative polypeptides of 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. Additionally, 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 (Goodman et al., 2002; Hileman et al., 2002; Metcalfe et al., 1996). 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 canola 5′ genomic DNA-T-DNA and 3′canola genomic DNA-intervening DNA and/or intervening DNA-T-DNA junctions of MON 88302. 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 5′ genomic DNA-T-DNA and the 3′ genomic DNA-intervening DNA and/or intervening DNA-T-DNA junctions in MON 88302 to sequences in established databases. Sequences spanning these 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)[1] 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 88302 insertion site (Song et al., 2010) was analyzed for translational stop codons (TGA, TAG, TAA). All six reading frames spanning the 5′ genomic DNA-T-DNA and the 3′ 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 11 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).

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

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

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 (1×10^{-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 (Goodman et al., 2002; Hileman 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 1e-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 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 11 putative polypeptides deduced from DNA sequence spanning the 5′ and 3′ genomic DNA-inserted DNA junctions of MON 88302 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 88302 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 insertion site junction and was greater than or equal to eight amino acids in length. At each of the 5′ and the 3′ flanks, five and six 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 (see Figures 1 and 2 for reading

frame assignment). Supporting dataset output files for each putative 5′ polypeptide are contained in Appendices 1-5, while dataset output files for each putative 3′ polypeptide are contained in Appendices 6-11.

5.1 Assessment of Potential Allergenicity

The results of the allergenicity assessment are shown in Tables 2 and 3. Potential allergenicity of the 11 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 11 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 eight amino acid matches were identified in the sliding window search of the AD_2010 database. As a result, these 11 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 4 and 5. Potential toxicity of the 11 putative polypeptides was assessed using the FASTA algorithm. When searching the TOX_2010 database, no alignments with any of the 11 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 6 and 7. Potential untoward biological activity of the 11 putative polypeptides was assessed using the FASTA algorithm. When searching the PRT_2010 database, no alignments with any of the 11 query sequences generated an *E*-score of less than or equal to 1e-5.

6.0 Conclusions

The bioinformatic analysis performed using the 11 putative sequences translated from junctions is theoretical as there is no reason to suspect, or evidence to indicate the presence of transcripts spanning the flank junctions. The results of bioinformatic analysis indicate that no structurally relevant sequence similarities were observed between the 11 putative flank junction derived sequences and allergens, toxins or biologically active proteins. As a result, in the unlikely occurrence that any of the 11 putative polypeptides analyzed herein is found *in planta*, none would share significant similarity or identity to known allergens, toxins, or other biologically active proteins that could affect human or animal health.

7.0 References

Aalberse, R.C. 2000. Structural biology of allergens. Journal of Allergy and Clinical Immunology 106:228-238.

Aalberse, R.C., J. Akkerdaas, and R. von Ree. 2001. Cross-reactivity of IgE antibodies to allergens. Journal of Allergy and Clinical Immunology 56:478-490.

Codex Alimentarius. 2003. Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants.CAC/GL 45-2003, Codex Alimentarius Comission, Joint FAO/WHO Food Standards Programme, Food and Agriculture Organisation Rome, Italy ftp://ftp.fao.org/codex/Publications/Booklets/Biotech/Biotech_2003e.pdf [Accessed February 5, 2007].

FARRP. 2010. Allergen database. Food Allergy Research and Resource Program. www.allergenonline.org.

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., A. Silvanovich, R.E. Hileman, G.A. Bannon, E.A. Rice, and J.D. Astwood. 2002. Bioinformatic methods for identifying known or potential allergens in the safety assessment of genetically modified crops. Comments on Toxicology 8:251-269.

Henikoff, J.G., and S. Henikoff. 1996. Blocks database and its applications. Methods Enzymol 266:88-105.

Henikoff, S., and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915-10919.

Hileman, R.E., A. Silvanovich, R.E. Goodman, E.A. Rice, G. Holleschak, J.D. Astwood, and S.L. Hefle. 2002. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. Int Arch Allergy Immunol 128:280-291.

Kleter, G.A., and A.A. Peijnenburg. 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 Struct Biol 2:8.

Lipman, D.J., and W.R. Pearson. 1985. Rapid and sensitive protein similarity searches. Science Mar 227:1435-1441.

Metcalfe, D., J. Astwood, T. R., S. H., T.M. L., and F. R. 1996. Assessment of the allergenic potential of foods derived from genetically engineered crop plants. Critical Reviews in Food Science and Nutrition 36:165-186.

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

Pearson, W.R., and D.J. Lipman. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85:2444-2448.

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

Song, Z., S.M. Arackal, K.R. Lawry, K.A. Robinson, and Q. Tian. 2010. Molecular Analysis of Glyphosate-Tolerant Roundup Readyâ 2 (RR2) Canola MON 88302. Monsanto Technical Report MSL0022523. St. Louis, Missouri.

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

Thomas, K., G. Bannon, S. Hefle, C. Herouet, M. Holsapple, G. Ladics, S. MacIntosh, and L. Privalle. 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 A. Silvanovich. 2010. The Assembly of Databases Used for FASTA, BLAST and Sliding Window Searches in 2010. Monsanto Technical Report MSL0022498. St. Louis, Missouri.

This page was intentionally left blank.

[CBI Cross Reference Number 1]

Deleted Figures 1, 2 and 3 and Table 1

Deleted pages 21 – 23 are found in the Confidential Attachment, pages 5 – 7.

Table 2. Summary of alignments for the FASTA searches of the AD_2010 database using putative polypeptide sequences encoded by the genomic DNA-T-DNA 5′ junction in MON 88302

Appendix	Polypeptide	AD_2010 Sequence Database						
		Sliding Window	FASTA search					
		Hits	# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
1	5_1	No	1	66840998	5a2 protein [Triticum aest (94 aa)	71.429	7	0.62
2	5_2	No	1	57283137	villin 1 [Nicotiana tabacu (559 aa)	42.857	14	0.7
3	5_3	No	-	-	-	-	-	-
4	5_4	No	3	520	beta-lactoglobulin [Bos taurus] (178 aa)	56.250	16	0.99
5	5_6	No	-	-	-	-	-	-

Table 3. 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 88302

Appendix	Polypeptide	AD_2010 Sequence Database						
		Sliding Window	FASTA search					
		Hits	# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
6	3_1	No	-	-	-	-	-	-
7	3_2	No	-	-	-	-	-	-
8	3_3	No	-	-	-	-	-	-
9	3_4	No	14	114841629	pollen allergen [Cryptome (514 aa)	53.846	13	0.59
10	3_5	No	2	256429	Kunitz trypsin inhibitor [Gly (216 aa)	31.915	47	0.84
11	3_6	No	9	15384338	AF177030_1 acidic allergen C (244 aa)	26.087	23	0.26

Table 4. Summary of alignments for the FASTA searches of the TOX_2010 database using putative polypeptide sequences encoded by the genomic DNA-T-DNA 5´ junctions in MON 88302

Appendix	Polypeptide	FASTA search of TOX_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
1	5_1	-	-	-	-	-	-
2	5_2	-	-	-	-	-	-
3	5_3	-	-	-	-	-	-
4	5_4	-	-	-	-	-	-
5	5_6	-	-	-	-	-	-

Table 5. 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 88302

Appendix	Polypeptide	FASTA search of TOX_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
6	3_1	-	-	-	-	-	-
7	3_2	-	-	-	-	-	-
8	3_3	-	-	-	-	-	-
9	3_4	-	-	-	-	-	-
10	3_5	-	-	-	-	-	-
11	3_6	-	-	-	-	-	-

Table 6. Summary of alignments for the FASTA searches of the PRT_2010 database using putative polypeptide sequences encoded by the genomic DNA-T-DNA 5´ junctions in MON 88302

Appendix	Polypeptide	FASTA search of PRT_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
1	5_1	-	-	-	-	-	-
2	5_2	-	-	-	-	-	-
3	5_3	-	-	-	-	-	-
4	5_4	-	-	-	-	-	-
5	5_6	-	-	-	-	-	-

Table 7. Summary of alignments for the FASTA searches of the PRT_2010 database using putative polypeptide sequences encoded by the the genomic DNA-T-DNA 3´ junctions in MON 88302

Appendix	Polypeptide	FASTA search of PRT_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
6	3_1	-	-	-	-	-	-
7	3_2	-	-	-	-	-	-
8	3_3	-	-	-	-	-	-
9	3_4	-	-	-	-	-	-
10	3_5	-	-	-	-	-	-
11	3_6	-	-	-	-	-	-

[CBI Cross Reference Number 2]

Deleted Appendices 1-11

Deleted pages 27 – 60 are found in the Confidential Attachment, pages 8 – 41.

CONFIDENTIAL ATTACHMENT

Study Title

Bioinformatics Evaluation of DNA Sequences Flanking the 5′ and 3′ Junctions of Inserted DNA in MON 88302: Assessment of Putative Polypeptides

Authors

[redacted]

Study Completed On

November 16, 2010

Sponsor and Performing Laboratory

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

Laboratory Project ID

MSL0023088
Study Number: REG-10-526

Description of Confidential Attachment

The following sections of this report include *Confidential Business Information:*

Reason for deletion of each these sections listed above from the main body of report: *These sections disclose commercial information (description of product manufacturing or quality control processes) FIFRA reference 10(d)(1)(A).*

Deleted Pages	Reason for Deletion	FIFRA Reference
21-23, 27-60	Discloses manufacturing or quality control processes	10(d)

Table of Contents

The following sections of this report include *Confidential Business Information*:

Confidential Attachment **Page**

Figures

Table

Appendices

Deleted pages are attached immediately following this page

```
 5'   AAAACCTTTTAGTCATCATGTTGTACCACTtcaaacactgatagtttaaactgaaggcggg   3'
5_1     K  P  F  S  H  H  V  V  P  L  q  t  l  i  v  .  t  e  g  g    >
5_2   Q  N  L  L  V  I  M  L  Y  H  f  k  h  .  .  f  k  l  k  a  g   >
5_3    K  T  F  .  S  S  C  C  T  T  s  n  t  d  s  l  n  .  r  r    >
 3'   TTTTGGAAAATCAGTAGTACAACATGGTGAagtttgtgactatcaaatttgacttccgccc   5'
5_4     F  R  K  T  M  M  N  Y  W  K  l  c  q  y  n  l  s  f  a  p    <
5_5    F  G  K  L  .  .  T  T  G  S  .  v  s  i  t  .  v  s  p  p     <
5_6   L  V  K  .  D  D  H  Q  V  V  e  f  v  s  l  k  f  q  l  r  s   <
```

Figure 1. Reading frame assignment and DNA sequence at the 5´ junction of the MON 88302 insert

DNA sequences are translated in 6 reading frames. Upper case characters refer to the genomic DNA, lower case characters refer to the inserted T-DNA found in MON 88302. The carat (> or <) points towards the carboxyl terminal of each polypeptide. Stop codons are denoted as periods (.).

```
 5'   TTTCCCGGACATGAAGCCATTTACAATTGAccatcatacTCAACTTCAATTTTTTTTAATG   3'
3_1     F  P  D  M  K  P  F  T  I  D  h  h  t  Q  L  Q  F  F  L  M    >
3_2   I  S  R  T  .  S  H  L  Q  L  t  i  i  L  N  F  N  F  F  .  C   >
3_3    F  P  G  H  E  A  I  Y  N  .  p  s  y  S  T  S  I  F  F  N     >
 3'   AAAGGGCCTGTACTTCGGTAAATGTTAACTggtagtatgAGTTGAAGTTAAAAAAAATTAC   5'
3_4     E  R  V  H  L  W  K  C  N  V  m  m  s  L  K  L  K  K  .  H    <
3_5    K  G  S  M  F  G  N  V  I  S  w  .  v  .  S  .  N  K  K  I     <
3_6   N  G  P  C  S  A  M  .  L  Q  g  d  y  E  V  E  I  K  K  L  T   <
```

Figure 2. Reading frame assignment and DNA sequence at the 3´ junctions of the MON 88302 insert

Upper case characters refer to the genomic DNA, lower case characters refer to the intervening DNA, and bolded underlined uppercase characters refer to the inserted T-DNA found in MON 88302. The carat (> or <) points towards the carboxyl terminal of each polypeptide. Stop codons are denoted as periods (.).

Confidential Attachment

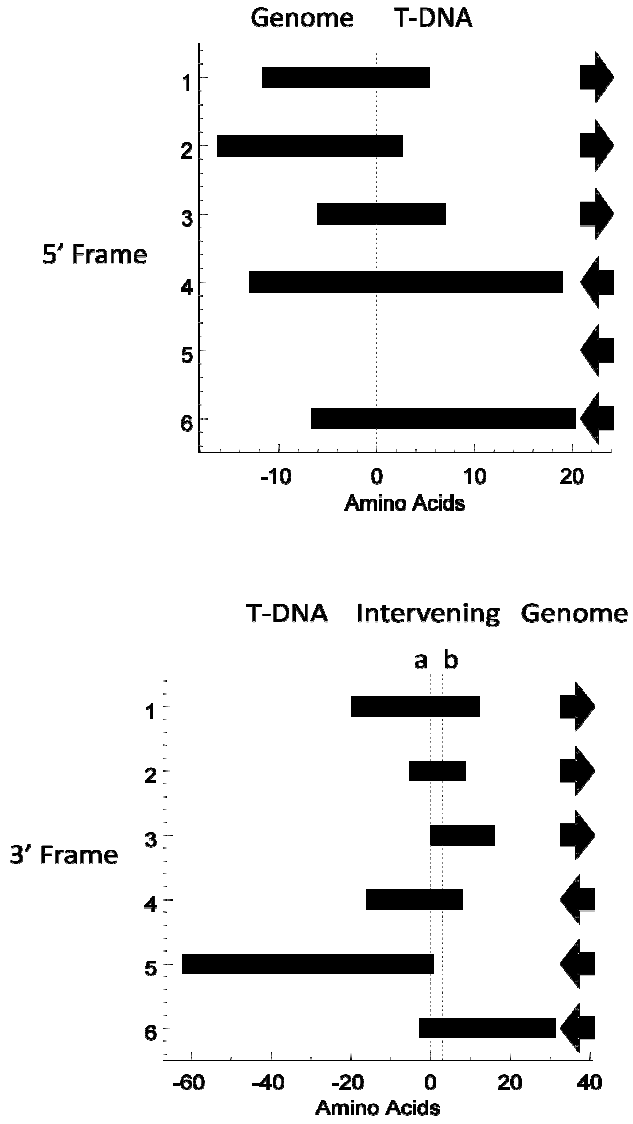

Figure 3. Graphic mapping of the flanking DNA sequences and putative polypeptides encoded by each reading frame at the 5′ and 3′ junctions of the MON 88302 insert

The putative polypeptide coding sequences are mapped relative to the DNA sequence shown in Figures 1 and 2, and the amino acid sequences are tabulated in Table 1. The scale at the bottom of each map refers to amino acids. The arrow for each ORF points in the direction of the C-terminus. Break point a refers to the intervening DNA-T-DNA junction and the break point b referes to the genomic DNA-intervening DNA junction.

Confidential Attachment

Table 1. The predicted sequence of polypeptides encoded by each reading frame at the 5′ and 3′ junctions of the MON 88302 insert

For display purposes, the predicted sequences are parsed into segments of ten amino acids in length. Uppercase characters refer to sequence encoded by genomic DNA. Underlined bolded uppercase characters refer to sequence encoded by the intervening DNA. Lowercase characters refer to sequence encoded by the T-DNA.

Putative Peptide ID	Putative peptide amino sequence
5_1	CTKPFSHHVV PLqtliv
5_2	TLFYSAQNLL VIMLYHfkh
5_3	SSCCTTsntd sln
5_4	slmgirlsfp afslnyqclK WYNMMTKRFC AL
5_6	leldgdqivv srlqfklsvf eVVQHDD
3_1	psysllihvd fpdmkpftid **HHT**QLQFFLM SL
3_2	shlql**TII**LN FNFF
3_3	**PSY**STSIFFN VIMIDE
3_4	KKLKL**SMM**vn ckwlhvreiy mdqq
3_5	**W**sivngfmsg kstwisneyd gqygekervi tnffsiqkcr cpqryykmkv hfdkttnydp syl
3_6	LLFYHYYGQQ DKKRNYSSII MTLKKIEVE**Y DG**ql

Confidential Attachment

Appendix 1. Bioinformatic analysis of polypeptide 5_1

```
>5_1
CTKPFSHHVVPLQTLIV
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_1

Start time: Fri Sep 17 20:40:17 GMT 2010 Finish time: Fri Sep 17 20:40:17 GMT
2010

No 8 amino acid matches exist between 5_1 and the AD_2010 database
```

```
# fasta34 5_1.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_1.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_1, 17 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 4 library sequences
  24     0     0:
  26     0     0:
  28     1     0:=
  30    26     2:*======
  32     3     8:=*
  34    28    21:=====*=
  36    19    44:=====     *
  38    54    72:==============    *
  40    48   101:============             *
  42    76   123:====================           *
  44   132   136:================================*
  46   217   138:================================*====================
  48   110   132:============================     *
  50   150   121:================================*=======
  52    85   106:======================     *
  54    84    91:==================== *
  56    76    76:==================*
  58    62    62:================*
  60    61    50:============*===
  62    54    40:=========*====
  64    24    32:====== *
  66    26    25:======*
  68    30    20:====*===
  70    11    16:===*
  72    23    12:==*===
  74    13    10:==*=
  76    14     7:=*==
  78    13     6:=*==
  80    14     4:*===
  82     4     3:*
  84     3     3:*
  86     2     2:*
  88     0     2:*          inset = represents 1 library sequences
  90     4     1:*
  92     1     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     0     0:          *
 100     0     0:          *
 102     0     0:          *
 104     0     0:          *
 106     1     0:=         *=
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.12080.0029; mu= 0.5331 0.150
 mean_var=21.0319 5.167, 0's: 2 Z-trim: 3  B-trim: 125 in 1/42
 Lambda= 0.279663
 Kolmogorov-Smirnov  statistic: 0.0814 (N=28) at  44

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(1471)
gi|66840998|emb|CAI64398.1| 5a2 protein [Triticum  (  94)   45 21.9    0.62


>>gi|66840998|emb|CAI64398.1| 5a2 protein [Triticum aest  (94 aa)
 initn:  45 init1:  45 opt:  45  Z-score: 106.1  bits: 21.9 E(): 0.62
Smith-Waterman score: 45;  71.429% identity (71.429% similar) in 7 aa overlap
(1-7:75-81)

                                          10
5_1                                    CTKPFSHHVVPLQTLIV
                                       : ::: :
gi|668 KANIPCLCAGVTKEKEKIYCMEKVAYVANFCKKPFPHGYKCGSYTFPPLA
           50        60        70        80        90
```

Confidential Attachment

```
17 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:40:17 2010 done: Fri Sep 17 20:40:17 2010
 Total Scan time:  0.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_1.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_1.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_1, 17 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    63     0:=====
  22     1     0:=          one = represents 14 library sequences
  24     7     0:=
  26    11     0:=
  28    35     2:*==
  30    34    12:*==
  32   139    45:===*======
  34   273   122:========*===========
  36   378   250:=================*==========
  38   466   414:==============================*====
  40   483   577:====================================      *
  42   692   706:=========================================================*
  44   838   779:============================================================*====
  46   479   793:====================================                        *
  48   481   759:====================================                      *
  50   597   693:==========================================      *
  52   481   609:====================================        *
  54   375   520:==========================          *
  56   470   435:==============================*==
  58   461   357:========================*=======
  60   639   289:====================*======================
  62   314   232:================*======
  64   225   184:=============*===
  66   125   146:========= *
  68   109   115:========*
  70    97    90:======*
  72    52    70:====*
  74    32    55:===*
  76     9    43:=  *
  78    28    33:==*
  80    11    26:=*
  82     6    20:=*
  84     2    16:=*
  86     5    12:*
  88     6     9:*          inset = represents 1 library sequences
  90     1     7:*
  92     2     6:*         :==   *
  94     7     4:*         :===*===
  96     8     3:*         :==*=====
  98     0     3:*         :  *
 100     0     2:*         : *
 102     1     2:*         :=*
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 2.60260.000479; mu= 9.7141 0.024
 mean_var=17.8268 3.748, 0's: 60 Z-trim: 60  B-trim: 427 in 1/61
 Lambda= 0.303765
 Kolmogorov-Smirnov  statistic: 0.0611 (N=29) at  54

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


17 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:40:17 2010 done: Fri Sep 17 20:40:18 2010
 Total Scan time:  0.130 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_1.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_1.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448
```

Confidential Attachment

```
 5_1, 17 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 278332     0:===========
  22  1136     0:=          one = represents 25986 library sequences
  24  2499    17:*
  26  6078   374:*
  28 19204  4039:*
  30 52015 24538:*==
  32 127334 94879:===*=
  34 265720 257301:=========*=
  36 489904 528436:=================== *
  38 757663 873308:==============================   *
  40 1071305 1218188:=========================================    *
  42    1324278    1489087:=================================================
*
  44                                                                  1474839
1642601:=========================================================  *
  46                                                                  1534536
1673030:==========================================================*
  48                                                                  1559106
1601734:==========================================================*
  50                                                                  1478464
1461589:=========================================================*
  52 1303725 1284980:=================================================*=
  54 1124232 1097598:=====================================*=
  56 977574 916831:=====================================*==
  58 860261 752701:============================*=====
  60 654725 609732:=======================*==
  62 519955 488824:==================*==
  64 407593 388759:===============*=
  66 331920 307263:===========*=
  68 259193 241687:==========*
  70 196386 189400:=======*
  72 169385 147998:=====*=
  74 134019 115389:====*=
  76 97065 89810:===*
  78 82289 69808:==*=
  80 60696 54205:==*
  82 47829 41466:=*
  84 33905 32846:=*
  86 28768 25414:*=
  88 20484 19664:*          inset = represents 249 library sequences
  90 14699 15215:*
  92 12415 11773:*         :=====================================*
  94  8504  9109:*         :================================== *
  96  5887  7048:*         :======================     *
  98  4445  5453:*         :=================   *
 100  3127  4220:*         :=============   *
 102  2254  3265:*         :==========   *
 104  1792  2526:*         :========  *
 106  1285  1955:*         :====== *
 108  1297  1512:*         :======*
 110   772  1170:*         :=====*
 112   722   905:*         :===*
 114   481   701:*         :==*
 116   345   542:*         :==*
 118   236   419:*         :=*
>120   547   325:*         :=*=
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810747 sequences
  Expectation_n fit: rho(ln(x))= 3.44070.000171; mu= 5.1611 0.009
 mean_var=20.7489 4.028, 0's: 946 Z-trim: 949  B-trim: 0 in 0/62
 Lambda= 0.281564
 Kolmogorov-Smirnov  statistic: 0.0412 (N=29) at  48

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


17 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:40:18 2010 done: Fri Sep 17 20:45:55 2010
 Total Scan time: 288.800 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 2. Bioinformatic analysis of polypeptide 5_2

```
>5_2
TLFYSAQNLLVIMLYHFKH
_____________________________________________________

Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_2

Start time: Fri Sep 17 20:45:56 GMT 2010 Finish time: Fri Sep 17 20:45:56 GMT
2010

No 8 amino acid matches exist between 5_2 and the AD_2010 database
_____________________________________________________

# fasta34 5_2.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_2.pep_ad.fasta
```

```
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_2, 19 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     3     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     8     0:===
  28     7     0:===
  30    20     2:*======
  32    25     8:==*======
  34    34    21:======*=====
  36    49    44:==============*==
  38    85    72:======================*=====
  40    68   101:=======================          *
  42   100   123:==================================      *
  44   128   136:=========================================  *
  46   118   138:=========================================     *
  48   106   132:====================================       *
  50   104   121:==================================     *
  52   173   106:==================================*====================
  54    95    91:==============================*=
  56    53    76:==================       *
  58    50    62:================    *
  60    51    50:================*
  62    59    40:=============*======
  64    15    32:=====     *
  66    32    25:========*==
  68    14    20:===== *
  70    27    16:=====*===
  72    15    12:===*=
  74     6    10:== *
  76     4     7:==*
  78     3     6:=*
  80     1     4:=*
  82     3     3:*
  84     4     3:*=
  86     6     2:*=
  88     0     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     1     1:*         :*
  94     0     1:*         :*
  96     2     1:*         :*=
  98     1     0:=         *=
 100     0     0:          *
 102     0     0:          *
 104     0     0:          *
 106     1     0:=         *=
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 2.59270.00215; mu= 7.6144 0.113
 mean_var=13.1270 3.204, 0's: 2 Z-trim: 2  B-trim: 125 in 1/42
 Lambda= 0.353991
 Kolmogorov-Smirnov  statistic: 0.0550 (N=28) at  38

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(1471)
gi|57283137|emb|CAE17316.1| villin 1 [Nicotiana ta ( 559)   44 24.4     0.7


>>gi|57283137|emb|CAE17316.1| villin 1 [Nicotiana tabacu  (559 aa)
 initn:  44 init1:  44 opt:  44  Z-score: 105.2  bits: 24.4 E():  0.7
Smith-Waterman  score:  44;   42.857%  identity  (78.571%  similar)  in  14  aa
overlap (3-16:27-40)

                                      10
5_2                                TLFYSAQNLLVIMLYHFKH
                                     :::..  .:.. ::
gi|572 EGGGKIEVWRINGSAKTPVPGDDIGKFYSGDCYIVLYTYHCNDRKEDYYLCWWIGKDSVE
               10        20        30        40        50        60

gi|572 EDQNMAAKLASTMCNSLKARPVLGRVYQGKEPPQFVAIFQPMLVLKGGLSSGYKSYIADK
               70        80        90       100       110       120



19 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:45:55 2010 done: Fri Sep 17 20:45:55 2010
 Total Scan time:  0.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_2.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_2.pep_tx.fasta
```

Confidential Attachment

```
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_2, 19 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    61     0:======
  22     0     0:           one = represents 12 library sequences
  24     4     0:=
  26     7     0:=
  28    17     2:*=
  30    80    12:*======
  32    59    45:===*=
  34   138   122:==========*=
  36   279   250:====================*===
  38   590   414:==================================*===============
  40   520   577:============================================     *
  42   712   706:============================================================*=
  44   657   779:=========================================================    *
  46   596   793:==========================================================   *
  48   555   759:=========================================================    *
  50   515   693:=========================================             *
  52   443   609:====================================             *
  54   716   520:=======================================*================
  56   460   435:====================================*==
  58   530   357:============================*===============
  60   559   289:========================*=====================
  62   194   232:=================  *
  64   159   184:============== *
  66   133   146:============*
  68   164   115:=========*====
  70    88    90:=======*
  72    48    70:==== *
  74    24    55:==  *
  76    20    43:== *
  78    23    33:==*
  80    18    26:==*
  82     8    20:=*
  84    20    16:=*
  86    29    12:*==
  88     5     9:*          inset = represents 1 library sequences
  90     4     7:*
  92     3     6:*         :===  *
  94     3     4:*         :===*
  96     0     3:*         :  *
  98     1     3:*         := *
 100     0     2:*         : *
 102     0     2:*         : *
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     1     0:=         *=
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 2.85010.000485; mu= 7.2330 0.025
 mean_var=15.5528 3.305, 0's: 60 Z-trim: 60  B-trim: 511 in 1/61
 Lambda= 0.325214
 Kolmogorov-Smirnov  statistic: 0.0703 (N=29) at  52

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


19 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:45:56 2010 done: Fri Sep 17 20:45:56 2010
 Total Scan time:  0.140 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_2.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_2.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_2, 19 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 281351     0:===========
  22   2335     0:=          one = represents 26585 library sequences
  24   4618    17:*
  26 12616    374:*
  28 35231   4039:*=
  30 88693 24538:*===
  32 204085 94880:===*====
  34 401182 257304:=========*======
  36 669120 528442:===================*======
  38 970625 873318:================================*====
```

Confidential Attachment

```
  40 1270254 1218202:==============================================*==
  42  1444696  1489104:============================================================
*
  44                                                                    1595060
1642621:============================================================*
  46                                                                    1583869
1673050:============================================================*
  48                                                                    1517772
1601752:============================================================ *
  50 1381315 1461606:=============================================================  *
  52 1202078 1284995:=============================================================  *
  54 1021478 1097611:=======================================  *
  56 868957 916842:================================== *
  58 717473 752710:=========================== *
  60 568633 609739:======================*
  62 442463 488830:================= *
  64 347511 388764:==============*
  66 276647 307267:===========*
  68 208857 241690:======== *
  70 173998 189402:=======*
  72 130553 148000:======*
  74 96733 115391:=====*
  76 76170 89811:===*
  78 55014 69809:==*
  80 39476 54205:==*
  82 31351 41466:=*
  84 29129 32846:=*
  86 16015 25415:*
  88 12506 19665:*          inset = represents 135 library sequences
  90  9153 15215:*
  92  6750 11773:*         :=======================================*
  94  5006  9109:*         :======================================== *
  96  3995  7048:*         :==============================           *
  98  2493  5454:*         :====================                     *
 100  1659  4220:*         :=============                  *
 102  1378  3265:*         :===========            *
 104   799  2526:*         :======             *
 106   525  1955:*         :====           *
 108   614  1512:*         :=====       *
 110   281  1170:*         :===      *
 112   235   905:*         :==     *
 114   111   701:*         :=     *
 116    93   542:*         :=    *
 118    79   419:*         :=  *
>120   190   325:*         :==*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810955 sequences
  Expectation_n fit: rho(ln(x))= 3.02640.000169; mu= 6.3962 0.009
 mean_var=19.8608 3.856, 0's: 947 Z-trim: 955  B-trim: 0 in 0/63
 Lambda= 0.287790
 Kolmogorov-Smirnov  statistic: 0.0375 (N=29) at  40

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


19 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:45:56 2010 done: Fri Sep 17 20:51:22 2010
 Total Scan time: 295.070 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 3. Bioinformatic analysis of polypeptide 5_3

```
>5_3
SSCCTTSNTDSLN
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_3

Start time: Fri Sep 17 20:51:23 GMT 2010 Finish time: Fri Sep 17 20:51:23 GMT
2010

No 8 amino acid matches exist between 5_3 and the AD_2010 database
```

```
# fasta34 5_3.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_3.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_3, 13 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     3     0:=
  22     0     0:           one = represents 4 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     1     2:*
  32     4     8:=*
```

```
  34    10    21:===  *
  36    12    44:===       *
  38    92    72:==================*=====
  40    70   101:==================       *
  42   143   123:==============================*=====
  44   107   136:==========================       *
  46   143   138:====================================*=
  48   107   132:==========================      *
  50   191   121:==============================*=================
  52   100   106:======================== *
  54    73    91:===================    *
  56    56    76:==============     *
  58    50    62:=============  *
  60    61    50:============*===
  62    24    40:======   *
  64    58    32:=======*=======
  66    49    25:======*======
  68    22    20:====*=
  70    36    16:===*=====
  72    11    12:==*
  74     9    10:==*
  76    12     7:=*=
  78     6     6:=*
  80     6     4:*=
  82     6     3:*=
  84     1     3:*
  86     5     2:*=
  88     0     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     1     1:*         :*
  94     2     1:*         :*=
  96     0     1:*         :*
  98     0     0:          *
 100     0     0:          *
 102     0     0:          *
 104     0     0:          *
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.74860.00266; mu= 0.0656 0.140
 mean_var=22.4745 5.632, 0's: 3 Z-trim: 3  B-trim: 15 in 1/42
 Lambda= 0.270538
 Kolmogorov-Smirnov  statistic: 0.0599 (N=28) at  48
```

```
FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


13 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:51:23 2010 done: Fri Sep 17 20:51:23 2010
 Total Scan time:  0.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_3.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_3.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_3, 13 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    60      0:===
  22     0      0:            one = represents 20 library sequences
  24    11      0:=
  26     0      0:
  28    36      2:*=
  30    52     12:*==
  32    79     45:==*=
  34   203    122:======*====
  36   282    250:============*==
  38   402    414:====================*
  40   248    577:=============                 *
  42   576    706:==============================      *
  44   938    779:=========================================*========
  46  1178    793:=========================================*====================
  48   700    759:====================================  *
  50   471    693:=========================           *
  52   467    609:========================      *
  54   509    520:=========================*
  56   530    435:====================*=====
  58   260    357:=============     *
  60   337    289:==============*==
  62   171    232:=========  *
  64   173    184:=========*
  66   196    146:========*==
  68   153    115:=====*==
  70   155     90:====*===
```

Confidential Attachment

```
  72    71    70:===*
  74    30    55:==*
  76    38    43:==*
  78    12    33:=*
  80    16    26:=*
  82     2    20:*
  84    10    16:*
  86    10    12:*
  88     0     9:*          inset = represents 1 library sequences
  90    22     7:*=
  92    41     6:*==       :=====*==================================
  94     3     4:*         :===*
  96     0     3:*         :  *
  98     0     3:*         :  *
 100     1     2:*         :=*
 102     0     2:*         : *
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 0.48980.000515; mu= 19.5260 0.027
 mean_var=39.341011.387, 0's: 60 Z-trim: 60  B-trim: 42 in 1/61
 Lambda= 0.204480
 Kolmogorov-Smirnov  statistic: 0.0364 (N=28) at  46

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


13 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:51:23 2010 done: Fri Sep 17 20:51:23 2010
 Total Scan time:  0.130 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_3.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_3.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_3, 13 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 283574     0:============
  22   466     0:=          one = represents 25163 library sequences
  24  1112    17:*
  26  2666   374:*
  28  9481  4039:*
  30 30248 24537:*=
  32 81825 94875:===*
  34 228258 257291:==========*
  36 385134 528415:================    *
  38 600804 873273:========================          *
  40 868387 1218139:===================================             *
  42     1188926     1489026:==================================================
*
  44  1382384  1642535:=========================================================
*
  46                                                                  1492301
1672962:=============================================================*
  48                                                                  1509743
1601669:=============================================================*
  50                                                                  1474379
1461529:=============================================================*
  52 1326395 1284928:=============================================================*=
  54 1169877 1097554:================================================*===
  56 1000329 916794:====================================*===
  58 861051 752670:==============================*=====
  60 714693 609707:========================*====
  62 632339 488804:====================*======
  64 514082 388743:================*=====
  66 413699 307251:============*====
  68 353092 241677:=========*=====
  70 277641 189392:=======*====
  72 221807 147992:=====*===
  74 168026 115384:====*==
  76 138189 89806:===*==
  78 102320 69805:==*==
  80 80184 54203:==*=
  82 67567 41464:=*=
  84 49436 32844:=*
  86 36494 25413:=*
  88 29653 19663:*=         inset = represents 348 library sequences
  90 23335 15215:*
  92 17397 11772:*         :==================================*======
  94 15296  9109:*         :============================*============
  96 10832  7048:*         :====================*===========
  98  8226  5453:*         :================*========
```

Confidential Attachment

```
 100  8632  4219:*         :============*============
 102  6868  3265:*         :=========*==========
 104  6378  2526:*         :=======*===========
 106  2918  1955:*         :=====*===
 108  4207  1512:*         :====*========
 110  2759  1170:*         :===*====
 112  2797   905:*         :==*======
 114  1917   701:*         :==*===
 116  1072   542:*         :=*==
 118   567   419:*         :=*
>120  1462   325:*         :*====
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810023 sequences
  Expectation_n fit: rho(ln(x))= 2.97050.000173; mu= 4.4679 0.009
 mean_var=21.4744 4.113, 0's: 890 Z-trim: 893  B-trim: 0 in 0/64
 Lambda= 0.276767
 Kolmogorov-Smirnov  statistic: 0.0927 (N=29) at  48

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


13 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:51:23 2010 done: Fri Sep 17 20:56:09 2010
 Total Scan time: 249.570 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 4. Bioinformatic analysis of polypeptide 5_4

```
>5_4
SLMGIRLSFPAFSLNYQCLKWYNMMTKRFCAL
--------------------------------------------------------


Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_4

Start time: Fri Sep 17 20:56:10 GMT 2010 Finish time: Fri Sep 17 20:56:10 GMT
2010

No 8 amino acid matches exist between 5_4 and the AD_2010 database
--------------------------------------------------------

# fasta34 5_4.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_4.pep_ad.fasta

FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_4, 32 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     4     0:==
  28    12     0:====
  30    19     2:*======
  32    38     8:==*==========
  34    35    21:======*=====
  36    54    44:==============*===
  38    51    72:==================      *
  40    71   101:========================          *
  42   104   123:===================================     *
  44    97   136:================================            *
  46   124   138:============================================    *
  48   132   132:============================================*
  50   151   121:============================================*==========
  52    88   106:==============================     *
  54   125    91:================================*===========
  56    55    76:====================      *
  58    76    62:====================*=====
  60    66    50:=================*=====
  62    38    40:=============*
  64    38    32:==========*==
  66    19    25:======== *
  68    16    20:=======*
  70    15    16:======*
  72    10    12:===*
  74     6    10:== *
  76     4     7:==*
  78     3     6:=*
  80     7     4:=*=
  82     0     3:*
  84     1     3:*
  86     3     2:*
  88     1     2:*          inset = represents 1 library sequences
  90     2     1:*
  92     0     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     0     0:          *
 100     0     0:          *
```

```
 102     4     0:==        *====
 104     0     0:          *
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 1.72470.00285; mu= 16.9232 0.149
 mean_var=26.7786 6.775, 0's: 2 Z-trim: 2  B-trim: 35 in 1/42
 Lambda= 0.247845
 Kolmogorov-Smirnov  statistic: 0.0591 (N=28) at  36

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(1471)
gi|520|emb|CAA32835.1| beta-lactoglobulin [Bos tau ( 178)   53 23.0    0.99
gi|195957138|gb|ACG59280.1| major allergen beta-la ( 178)   53 23.0    0.99
gi|125910|sp|P02754.3|LACB_BOVIN RecName: Full=Bet ( 178)   53 23.0    0.99


>>gi|520|emb|CAA32835.1| beta-lactoglobulin [Bos taurus]  (178 aa)
 initn:  53 init1:  53 opt:  53  Z-score: 102.4  bits: 23.0 E(): 0.99
Smith-Waterman  score:  53;   56.250%  identity  (68.750%  similar)  in  16  aa
overlap (3-18:161-176)

                                       10        20        30
5_4                                 SLMGIRLSFPAFSLNYQCLKWYNMMTKRFCAL
                                      : :::::   .:. ::
gi|520 QSLVCQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI
              140       150       160       170

>>gi|195957138|gb|ACG59280.1| major allergen beta-lactog  (178 aa)
 initn:  53 init1:  53 opt:  53  Z-score: 102.4  bits: 23.0 E(): 0.99
Smith-Waterman  score:  53;   56.250%  identity  (68.750%  similar)  in  16  aa
overlap (3-18:161-176)

                                       10        20        30
5_4                                 SLMGIRLSFPAFSLNYQCLKWYNMMTKRFCAL
                                      : :::::   .:. ::
gi|195 QSLVCQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI
              140       150       160       170

>>gi|125910|sp|P02754.3|LACB_BOVIN RecName: Full=Beta-la  (178 aa)
 initn:  53 init1:  53 opt:  53  Z-score: 102.4  bits: 23.0 E(): 0.99
Smith-Waterman  score:  53;   56.250%  identity  (68.750%  similar)  in  16  aa
overlap (3-18:161-176)
```

```
                                       10        20        30
5_4                                 SLMGIRLSFPAFSLNYQCLKWYNMMTKRFCAL
                                      : :::::   .:. ::
gi|125 QSLACQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI
              140       150       160       170



32 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:56:09 2010 done: Fri Sep 17 20:56:09 2010
 Total Scan time:  0.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_4.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_4.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_4, 32 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    60     0:===
  22     0     0:           one = represents 23 library sequences
  24     0     0:
  26     1     0:=
  28     0     2:*
  30    39    12:*=
  32    55    45:=*=
  34   123   122:=====*
  36   241   250:==========*
  38   387   414:=================*
  40   435   577:====================      *
  42   532   706:========================       *
  44   684   779:==============================    *
  46  1363   793:=================================*========================
  48   793   759:================================*==
  50   542   693:========================       *
  52   664   609:==========================*==
  54   555   520:======================*==
  56   366   435:================  *
  58   356   357:===============*
  60   168   289:========    *
```

Confidential Attachment

```
  62   138   232:======    *
  64   126   184:====== *
  66   100   146:===== *
  68   117   115:====*=
  70    86    90:===*
  72   110    70:===*=
  74   108    55:==*==
  76    59    43:=*=
  78    52    33:=*=
  80    22    26:=*
  82    33    20:*=
  84    28    16:*=
  86    44    12:*=
  88    27     9:*=          inset = represents 1 library sequences
  90    16     7:*
  92     1     6:*          :=     *
  94     5     4:*          :===*=
  96     4     3:*          :==*=
  98     0     3:*          :  *
 100     1     2:*          :=*
 102     1     2:*          :=*
 104     0     1:*          :*
 106     1     1:*          :*
 108     0     1:*          :*
 110     0     1:*          :*
 112     0     0:           *
 114     0     0:           *
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 3.10100.000519; mu= 9.6793 0.027
 mean_var=23.9338 5.436, 0's: 60 Z-trim: 60  B-trim: 722 in 2/60
 Lambda= 0.262161
 Kolmogorov-Smirnov  statistic: 0.0489 (N=29) at  44

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


32 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:56:10 2010 done: Fri Sep 17 20:56:10 2010
 Total Scan time:  0.190 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_4.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_4.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_4, 32 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 277311     0:==========
  22    755     0:=          one = represents 28082 library sequences
  24   1380    17:*
  26   3457   374:*
  28 11386  4039:*
  30 31015 24538:*=
  32 93738 94879:===*
  34 221599 257300:======== *
  36 438052 528433:================  *
  38 739445 873303:==========================    *
  40 1069907 1218180:======================================    *
  42 1384059 1489077:=================================================   *
  44                                                                1588267
1642591:============================================================ *
  46                                                                1684915
1673020:============================================================*
  48                                                                1640504
1601723:============================================================*=
  50 1486188 1461579:============================================================*
  52 1313971 1284972:=============================================*=
  54 1145326 1097591:=========================================*=
  56 936638 916825:================================*=
  58 748899 752696:=========================*
  60 606055 609728:=====================*
  62 501316 488821:=================*
  64 411563 388757:==============*=
  66 306034 307261:==========*
  68 252673 241685:========*
  70 197424 189399:======*=
  72 154278 147997:=====*
  74 124409 115388:====*
  76 98483 89810:===*
  78 74353 69808:==*
  80 61454 54205:=*=
  82 47854 41465:=*
  84 39089 32846:=*
  86 28274 25414:*=
  88 24994 19664:*          inset = represents 228 library sequences
  90 18112 15215:*
  92 11353 11773:*         :=======================================*
```

Confidential Attachment

```
  94  8733  9109:*         :======================================*
  96  8199  7048:*         :============================*=====
  98  4809  5453:*         :====================== *
 100  4310  4220:*         :==================*
 102  2853  3265:*         :============= *
 104  1825  2526:*         :=========  *
 106  1482  1955:*         :======= *
 108  1140  1512:*         :===== *
 110   794  1170:*         :==== *
 112   563   905:*         :===*
 114   602   701:*         :===*
 116   266   542:*         :==*
 118   371   419:*         :=*
>120   748   325:*         :=*==
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810633 sequences
  Expectation_n fit: rho(ln(x))= 3.56490.000176; mu= 7.7944 0.009
 mean_var=27.2830 5.374, 0's: 936 Z-trim: 939  B-trim: 658 in 1/64
 Lambda= 0.245543
 Kolmogorov-Smirnov  statistic: 0.0313 (N=29) at  44

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


32 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 20:56:10 2010 done: Fri Sep 17 21:03:28 2010
 Total Scan time: 389.260 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 5. Bioinformatic analysis of polypeptide 5_6

```
>5_6
LELDGDQIVVSRLQFKLSVFEVVQHDD
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_6

Start time: Fri Sep 17 21:03:29 GMT 2010 Finish time: Fri Sep 17 21:03:29 GMT
2010

No 8 amino acid matches exist between 5_6 and the AD_2010 database
```

```
# fasta34 5_6.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_6.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_6, 27 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2      0:=
  22     0      0:            one = represents 3 library sequences
  24     0      0:
  26     1      0:=
  28     3      0:=
  30    17      2:*=====
  32    12      8:==*=
  34    32     21:======*====
  36    37     44:============= *
  38    60     72:====================   *
  40    93    101:==============================  *
  42   106    123:====================================    *
  44   100    136:===================================           *
  46   104    138:===================================          *
  48   105    132:====================================        *
  50   136    121:============================================*=====
  52   127    106:====================================*=======
  54    90     91:==============================*
  56    83     76:=========================*==
  58    50     62:=================    *
  60    32     50:===========      *
  62    70     40:=============*==========
  64    33     32:==========*
  66    31     25:========*==
  68    21     20:=======*
  70    23     16:======*==
  72    23     12:===*====
  74    28     10:===*======
  76     6      7:==*
  78     9      6:=*=
  80    11      4:=*==
  82    16      3:*=====
  84     3      3:*
  86     3      2:*
  88     4      2:*=          inset = represents 1 library sequences
  90     0      1:*
  92     0      1:*          :*
  94     0      1:*          :*
```

Confidential Attachment

```
  96     0     1:*          :*
  98     0     0:           *
 100     0     0:           *
 102     0     0:           *
 104     0     0:           *
 106     0     0:           *
 108     0     0:           *
 110     0     0:           *
 112     0     0:           *
 114     0     0:           *
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.19990.00258; mu= 8.6381 0.135
 mean_var=25.1437 6.404, 0's: 2 Z-trim: 2  B-trim: 213 in 1/42
 Lambda= 0.255776
 Kolmogorov-Smirnov  statistic: 0.0728 (N=29) at  48

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


27 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:03:28 2010 done: Fri Sep 17 21:03:29 2010
 Total Scan time:  0.030 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_6.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_6.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_6, 27 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    60     0:====
  22     0     0:           one = represents 17 library sequences
  24     0     0:
  26     2     0:=
  28    11     2:*
  30    37    12:*==
  32    57    45:==*=
  34   145   122:=======*=
  36   233   250:==============*
  38   404   414:========================*
  40   460   577:============================      *
  42   567   706:==================================        *
  44   837   779:=============================================*====
  46   999   793:================================================*============
  48   638   759:======================================       *
  50   728   693:=========================================*==
  52   447   609:===========================         *
  54   538   520:==============================*=
  56   340   435:====================      *
  58   640   357:====================*=================
  60   279   289:================*
  62   192   232:============ *
  64   177   184:==========*
  66   126   146:========*
  68    97   115:======*
  70    74    90:=====*
  72    63    70:====*
  74    62    55:===*
  76    20    43:==*
  78    34    33:=*
  80    36    26:=*=
  82    27    20:=*
  84    34    16:*=
  86    11    12:*
  88    44     9:*==        inset = represents 1 library sequences
  90     9     7:*
  92     1     6:*          :=     *
  94     2     4:*          :== *
  96     1     3:*          := *
  98     3     3:*          :==*
 100     5     2:*          :=*===
 102     1     2:*          :=*
 104     0     1:*          :*
 106     0     1:*          :*
 108     0     1:*          :*
 110     1     1:*          :*
 112     0     0:           *
 114     1     0:=          *=
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.10640.000499; mu= 3.4215 0.025
 mean_var=22.2316 5.269, 0's: 60 Z-trim: 60  B-trim: 529 in 2/60
 Lambda= 0.272012
 Kolmogorov-Smirnov  statistic: 0.0326 (N=29) at  56
```

Confidential Attachment

```
FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


27 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:03:29 2010 done: Fri Sep 17 21:03:29 2010
 Total Scan time:  0.170 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 5_6.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_6.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_6, 27 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 280875     0:===========
  22  1848     0:=          one = represents 27749 library sequences
  24  3634    17:*
  26  9044   374:*
  28 28479  4039:*=
  30 76341 24538:*==
  32 183825 94879:===*===
  34 380219 257301:==========*====
  36 658190 528436:===================*====
  38 949847 873307:================================*===
  40 1231309 1218187:==========================================*=
  42 1482511 1489085:============================================================*
  44                                                                    1606108
1642600:============================================================ *
  46                                                                    1664882
1673029:============================================================*
  48                                                                    1564329
1601732:==========================================================*
  50 1397975 1461587:========================================================= *
  52 1226396 1284979:=========================================================== *
  54 1042345 1097597:====================================== *
  56 837269 916830:==============================  *
  58 665768 752700:=======================    *
  60 521632 609731:====================  *
  62 415887 488823:===============  *
  64 333937 388759:============= *
  66 274256 307263:========== *
  68 217994 241687:========*
  70 172356 189400:======*
  72 139502 147998:=====*
  74 104540 115389:====*
  76 82592 89810:===*
  78 68335 69808:==*
  80 48158 54205:=*
  82 37853 41466:=*
  84 26761 32846:=*
  86 20634 25414:*
  88 14060 19664:*          inset = represents 167 library sequences
  90 10651 15215:*
  92  8349 11773:*         :=======================================*
  94  6362  9109:*         :=========================================*
  96  4481  7048:*         :============================            *
  98  3004  5453:*         :==================              *
 100  2469  4220:*         :===============          *
 102  1708  3265:*         :===========         *
 104  1220  2526:*         :========       *
 106   932  1955:*         :======      *
 108   658  1512:*         :====     *
 110   628  1170:*         :====    *
 112   314   905:*         :==    *
 114   204   701:*         :==  *
 116   154   542:*         :=  *
 118   147   419:*         := *
>120   253   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810729 sequences
  Expectation_n fit: rho(ln(x))= 3.66710.00017; mu= 6.1631 0.009
 mean_var=23.6912 4.661, 0's: 912 Z-trim: 919  B-trim: 0 in 0/66
 Lambda= 0.263500
 Kolmogorov-Smirnov  statistic: 0.0298 (N=29) at  40

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


27 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:03:29 2010 done: Fri Sep 17 21:10:00 2010
 Total Scan time: 358.190 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Confidential Attachment

Appendix 6. Bioinformatic analysis of polypeptide 3_1

```
>3_1
PSYSLLIHVDFPDMKPFTIDHHTQLQFFLMSL
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 3_1

Start time: Fri Sep 17 21:10:00 GMT 2010 Finish time: Fri Sep 17 21:10:00 GMT
2010

No 8 amino acid matches exist between 3_1 and the AD_2010 database
```

```
# fasta34 3_1.pep /genedata/1/db/AD_2010 -Q -E 1 -O 3_1.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_1, 32 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     2     0:=          one = represents 3 library sequences
  24     0     0:
  26     1     0:=
  28     7     0:===
  30     6     2:*=
  32    29     8:==*=======
  34    38    21:======*======
  36    74    44:==============*==========
  38   128    72:=======================*==================
  40   129   101:================================*=========
  42   130   123:============================================*===
  44   170   136:=============================================*==========
  46    90   138:==============================               *
  48    79   132:==========================                  *
  50    86   121:============================              *
  52    68   106:=======================           *
  54    82    91:============================  *
  56    33    76:===========               *
  58    57    62:==================  *
  60    72    50:================*=======
  62    39    40:==============*
  64    32    32:==========*
  66    49    25:========*========
  68    18    20:======*
  70     9    16:===  *
  72    11    12:===*
  74     7    10:===*
  76     9     7:==*
  78     8     6:=*=
  80     2     4:=*
  82     1     3:*
  84     1     3:*
  86     0     2:*
  88     0     2:*          inset = represents 1 library sequences
  90     2     1:*
  92     0     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     0     0:          *
 100     0     0:          *
 102     0     0:          *
 104     0     0:          *
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.37560.00341; mu= 11.0095 0.176
 mean_var=37.113010.408, 0's: 2 Z-trim: 2  B-trim: 125 in 1/42
 Lambda= 0.210529
 Kolmogorov-Smirnov  statistic: 0.1412 (N=27) at  44

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


32 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:10:00 2010 done: Fri Sep 17 21:10:00 2010
 Total Scan time:  0.030 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_1.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 3_1.pep_tx.fasta
```

```
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_1, 32 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    64     0:=====
  22     0     0:           one = represents 14 library sequences
  24    27     0:==
  26    12     0:=
  28    16     2:*=
  30    38    12:*==
  32    59    45:===*=
  34   119   122:========*
  36   209   250:===============  *
  38   333   414:=======================     *
  40   482   577:==================================      *
  42   780   706:=================================================*=====
  44   793   779:==========================================================*=
  46   562   793:=========================================               *
  48   722   759:=====================================================  *
  50   839   693:==================================================*==========
  52   575   609:========================================== *
  54   594   520:=====================================*=====
  56   406   435:============================  *
  58   286   357:====================    *
  60   305   289:====================*=
  62   197   232:=============== *
  64   161   184:============ *
  66   155   146:==========*=
  68    85   115:======= *
  70    59    90:===== *
  72    43    70:====*
  74    51    55:===*
  76    87    43:===*===
  78   291    33:==*==================
  80    21    26:=*
  82    24    20:=*
  84    15    16:=*
  86    19    12:*=
  88     5     9:*          inset = represents 1 library sequences
  90     4     7:*
  92     3     6:*         :===  *
  94     2     4:*         :== *
  96     0     3:*         :  *
  98     0     3:*         :  *
 100     0     2:*         : *
 102     0     2:*         : *
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.97930.000562; mu= 0.3122 0.028
 mean_var=24.4911 5.209, 0's: 60 Z-trim: 64  B-trim: 372 in 1/61
 Lambda= 0.259161
 Kolmogorov-Smirnov  statistic: 0.0412 (N=29) at  74

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


32 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:10:01 2010 done: Fri Sep 17 21:10:01 2010
 Total Scan time:  0.170 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_1.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 3_1.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_1, 32 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 280519     0:===========
  22    854     0:=          one = represents 27507 library sequences
  24   1938    17:*
  26   4810   374:*
  28 12995  4039:*
  30 40276 24537:*=
  32 110787 94877:===*=
  34 256080 257296:=========*
  36 476702 528426:================== *
  38 788715 873292:============================  *
```

Confidential Attachment

```
  40 1119488 1218165:=========================================   *
  42 1413063 1489058:=================================================  *
  44                                                                1579267
1642570:========================================================== *
  46                                                                1645226
1672999:===========================================================*
  48                                                                1650402
1601703:============================================================*=
  50 1497374 1461561:============================================================*=
  52 1308629 1284956:=============================================*=
  54 1095733 1097577:=========================================*
  56 888233 916813:================================*
  58 740712 752687:==========================*
  60 578753 609720:======================*
  62 474071 488815:=================*
  64 386905 388752:==============*
  66 315642 307257:===========*
  68 269572 241682:========*=
  70 211273 189396:======*=
  72 170055 147995:=====*=
  74 129723 115387:====*
  76 91973 89808:===*
  78 67972 69807:==*
  80 51502 54204:=*
  82 39655 41465:=*
  84 28127 32845:=*
  86 21919 25414:*
  88 16277 19664:*          inset = represents 181 library sequences
  90 11842 15215:*
  92  9008 11772:*         :=======================================*
  94  6823  9109:*         :======================================== *
  96  4756  7048:*         :============================           *
  98  3938  5453:*         :=====================        *
 100  2687  4220:*         :===============         *
 102  1892  3265:*         :===========       *
 104  1389  2526:*         :========      *
 106   966  1955:*         :======     *
 108   684  1512:*         :====     *
 110   486  1170:*         :===    *
 112   548   905:*         :====*
 114   355   701:*         :== *
 116   212   542:*         :==*
 118   152   419:*         := *
>120   265   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810407 sequences
  Expectation_n fit: rho(ln(x))= 4.19470.000176; mu= 5.1007 0.009
 mean_var=28.4234 5.714, 0's: 924 Z-trim: 924  B-trim: 2360 in 1/62
 Lambda= 0.240567
 Kolmogorov-Smirnov  statistic: 0.0203 (N=29) at  46
```

```
FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


32 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:10:01 2010 done: Fri Sep 17 21:16:57 2010
 Total Scan time: 384.540 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 7. Bioinformatic analysis of polypeptide 3_2

```
>3_2
SHLQLTIILNFNFF
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 3_2

Start time: Fri Sep 17 21:16:57 GMT 2010 Finish time: Fri Sep 17 21:16:57 GMT
2010

No 8 amino acid matches exist between 3_2 and the AD_2010 database
```

```
# fasta34 3_2.pep /genedata/1/db/AD_2010 -Q -E 1 -O 3_2.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_2, 14 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     4     0:==         one = represents 3 library sequences
  24     2     0:=
  26     6     0:==
  28     6     0:==
  30     6     2:*=
  32    15     8:==*==
  34    32    21:======*====
```

Confidential Attachment

```
  36    59    44:==============*=====
  38    55    72:==================    *
  40    60   101:====================              *
  42   106   123:==================================    *
  44   103   136:====================================          *
  46   133   138:============================================*
  48   100   132:==================================          *
  50   178   121:=========================================*==================
  52   101   106:================================== *
  54    97    91:==============================*==
  56    94    76:=========================*======
  58    81    62:====================*======
  60    36    50:============    *
  62    46    40:=============*==
  64    38    32:==========*==
  66    16    25:======  *
  68    26    20:======*==
  70    20    16:=====*=
  72    20    12:===*===
  74    13    10:===*=
  76     7     7:==*
  78     0     6: *
  80     3     4:=*
  82     0     3:*
  84     2     3:*
  86     1     2:*
  88     2     2:*          inset = represents 1 library sequences
  90     1     1:*
  92     0     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     0     0:          *
 100     0     0:          *
 102     0     0:          *
 104     0     0:          *
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 2.05820.00237; mu= 8.4600 0.126
 mean_var=13.7522 3.106, 0's: 2 Z-trim: 2  B-trim: 0 in 0/43
 Lambda= 0.345851
 Kolmogorov-Smirnov  statistic: 0.0612 (N=27) at  48

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


14 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:16:57 2010 done: Fri Sep 17 21:16:57 2010
 Total Scan time:  0.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_2.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 3_2.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_2, 14 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    62     0:=====
  22     4     0:=          one = represents 15 library sequences
  24     3     0:=
  26    14     0:=
  28    36     2:*==
  30    74    12:*====
  32   132    45:==*======
  34   166   122:========*===
  36   216   250:================ *
  38   343   414:========================    *
  40   445   577:==============================        *
  42   581   706:============================================        *
  44   702   779:================================================    *
  46   895   793:============================================================*=======
  48   870   759:==========================================================*=======
  50   547   693:=========================================          *
  52   396   609:============================              *
  54   522   520:==================================*
  56   487   435:============================*====
  58   553   357:=======================*=============
  60   271   289:====================*
  62   317   232:===============*======
  64   138   184:==========  *
  66   100   146:=======  *
  68   220   115:=======*=======
  70   108    90:=====*==
  72    63    70:====*
```

Confidential Attachment

```
  74    72    55:===*=
  76    30    43:==*
  78    13    33:= *
  80     4    26:=*
  82    10    20:=*
  84     7    16:=*
  86     8    12:*
  88     1     9:*          inset = represents 1 library sequences
  90    11     7:*
  92     3     6:*         :===  *
  94     8     4:*         :===*====
  96    11     3:*         :==*========
  98     0     3:*         :  *
 100     0     2:*         : *
 102     0     2:*         : *
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 3.57100.000451; mu= 2.2305 0.023
 mean_var=12.1259 2.499, 0's: 60 Z-trim: 61  B-trim: 67 in 2/60
 Lambda= 0.368313
 Kolmogorov-Smirnov  statistic: 0.0403 (N=29) at  52

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


14 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:16:57 2010 done: Fri Sep 17 21:16:57 2010
 Total Scan time:  0.110 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_2.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 3_2.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_2, 14 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 287624     0:===========
  22   3549     0:=          one = represents 24390 library sequences
  24   8649    17:*
  26 24253   374:*
  28 47583  4039:*=
  30 99377 24538:=*===
  32 203696 94880:===*=====
  34 378922 257303:==========*=====
  36 622699 528439:====================*====
  38 900485 873313:==================================*=
  40 1160495 1218195:================================================== *
  42 1352510 1489095:============================================================
*
  44                                                                  1439061
1642610:============================================================*
  46                                                                  1463392
1673040:============================================================*
  48                                                                  1392280
1601742:============================================================ *
  50  1336468  1461596:============================================================
*
  52 1280128 1284987:==========================================================*
  54 1142575 1097604:=========================================================*=
  56 951954 916836:=======================================*==
  58 790992 752705:==============================*==
  60 669803 609735:========================*===
  62 527661 488827:====================*=
  64 432214 388761:================*==
  66 309841 307265:============*
  68 244272 241688:==========*=
  70 181264 189401:=======*
  72 135175 147999:======*
  74 101599 115390:====*
  76 91468 89811:===*
  78 62128 69808:==*
  80 43097 54205:==*
  82 33466 41466:=*
  84 23076 32846:=*
  86 18048 25415:=*
  88 14158 19664:*          inset = represents 142 library sequences
  90 10099 15215:*
  92  7065 11773:*         :=======================================*
  94  6154  9109:*         :=======================================*
  96  3691  7048:*         :=========================              *
  98  2762  5454:*         :====================                 *
 100  2176  4220:*         :================             *
```

Confidential Attachment

```
 102  1544  3265:*         :===========           *
 104  1130  2526:*         :========          *
 106   894  1955:*         :=======       *
 108   508  1512:*         :====       *
 110   362  1170:*         :===      *
 112   268   905:*         :==     *
 114   170   701:*         :==   *
 116   120   542:*         :=   *
 118    85   419:*         := *
>120   235   325:*         :==*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810844 sequences
  Expectation_n fit: rho(ln(x))= 2.73650.000168; mu= 6.3560 0.009
 mean_var=18.0663 3.521, 0's: 953 Z-trim: 959  B-trim: 0 in 0/64
 Lambda= 0.301745
 Kolmogorov-Smirnov  statistic: 0.0289 (N=29) at  38

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


14 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:16:57 2010 done: Fri Sep 17 21:21:44 2010
 Total Scan time: 252.880 Total Display time:  0.010

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 8. Bioinformatic analysis of polypeptide 3_3

```
>3_3
PSYSTSIFFNVIMIDE
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 3_3

Start time: Fri Sep 17 21:21:45 GMT 2010 Finish time: Fri Sep 17 21:21:45 GMT
2010

No 8 amino acid matches exist between 3_3 and the AD_2010 database
```

```
# fasta34 3_3.pep /genedata/1/db/AD_2010 -Q -E 1 -O 3_3.pep_ad.fasta

FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_3, 16 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     2     0:=
  26     0     0:
  28     1     0:=
  30     1     2:*
  32    22     8:==*=====
  34     6    21:==    *
  36    30    44:==========    *
  38    47    72:================       *
  40   100   101:================================*
  42    99   123:================================       *
  44   147   136:============================================*===
  46   100   138:================================            *
  48   130   132:============================================*
  50   121   121:=========================================*
  52   109   106:==================================*=
  54    78    91:==========================     *
  56    98    76:=========================*=======
  58    78    62:====================*=====
  60    72    50:================*=======
  62    34    40:============ *
  64    59    32:==========*=========
  66    25    25:========*
  68    28    20:======*===
  70    18    16:=====*
  72    18    12:===*==
  74    22    10:===*====
  76     2     7:= *
  78     4     6:=*
  80     6     4:=*
  82     4     3:*=
  84     0     3:*
  86     0     2:*
  88     1     2:*          inset = represents 1 library sequences
  90     1     1:*
  92     2     1:*         :*=
  94     1     1:*         :*
  96     1     1:*         :*
  98     1     0:=         *=
 100     1     0:=         *=
```

Confidential Attachment

```
 102     0     0:          *
 104     0     0:          *
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.81070.00244; mu= 0.3813 0.128
 mean_var=15.7813 3.828, 0's: 2 Z-trim: 2  B-trim: 21 in 1/42
 Lambda= 0.322852
 Kolmogorov-Smirnov  statistic: 0.0691 (N=27) at  54

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


16 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:21:44 2010 done: Fri Sep 17 21:21:44 2010
 Total Scan time:  0.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_3.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 3_3.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_3, 16 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    60     0:=====
  22     1     0:=          one = represents 14 library sequences
  24     1     0:=
  26     3     0:=
  28     9     2:*
  30    27    12:*=
  32    54    45:===*
  34   125   122:========*
  36   183   250:==============   *
  38   260   414:==================          *
  40   827   577:=========================================*==================
  42   739   706:================================================*==
  44   647   779:=============================================          *
  46   683   793:================================================       *
  48   819   759:============================================================*====
  50   778   693:=================================================*======
  52   650   609:==============================================*===
  54   341   520:=========================           *
  56   382   435:============================    *
  58   402   357:=========================*===
  60   327   289:====================*===
  62   343   232:================*========
  64   113   184:=========     *
  66   139   146:==========*
  68    88   115:======= *
  70    48    90:====  *
  72    44    70:====*
  74    19    55:== *
  76    39    43:===*
  78    56    33:==*=
  80    18    26:=*
  82    13    20:=*
  84    36    16:=*=
  86     6    12:*
  88     4     9:*           inset = represents 2 library sequences
  90    22     7:*=
  92    72     6:*=====     :==*================================
  94    54     4:*===       :=*=========================
  96     6     3:*          :=*=
  98     5     3:*          :=*=
 100     0     2:*          :*
 102     0     2:*          :*
 104     0     1:*          :*
 106     0     1:*          :*
 108     0     1:*          :*
 110     0     1:*          :*
 112     0     0:           *
 114     0     0:           *
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.05920.000508; mu= 0.9990 0.026
 mean_var=21.8404 4.426, 0's: 60 Z-trim: 60  B-trim: 511 in 1/61
 Lambda= 0.274438
 Kolmogorov-Smirnov  statistic: 0.0244 (N=29) at  76

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000
```

Confidential Attachment

```
16 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:21:45 2010 done: Fri Sep 17 21:21:45 2010
 Total Scan time:  0.140 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_3.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 3_3.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_3, 16 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 282400     0:============
  22  1337     0:=          one = represents 25166 library sequences
  24  3189    17:*
  26  8552   374:*
  28 26510  4039:*=
  30 75181 24538:*==
  32 172357 94878:===*===
  34 333178 257298:==========*===
  36 576622 528431:====================*==
  38 846542 873299:====================================*
  40 1125723 1218175:================================================    *
  42   1325154   1489071:============================================================
*
  44                                                                    1480058
1642584:=================================================================*
  46                                                                    1509952
1673013:=================================================================*
  48                                                                    1474679
1601717:=================================================================*
  50                                                                    1415438
1461573:========================================================== *
  52 1299801 1284967:=============================================================*
  54 1127183 1097587:=============================================*=
  56 950446 916821:====================================*=
  58 803632 752693:==============================*==
  60 660958 609726:========================*==
  62 547576 488819:====================*==
  64 412679 388755:===============*=
  66 324326 307260:============*
  68 255169 241684:=========*=
  70 190170 189398:=======*
  72 146631 147997:=====*
  74 115627 115388:====*
  76 83719 89809:===*
  78 60202 69807:==*
  80 45496 54204:==*
  82 34188 41465:=*
  84 27053 32845:=*
  86 18710 25414:=*
  88 13623 19664:*          inset = represents 144 library sequences
  90 10363 15215:*
  92  7155 11773:*         :=======================================*
  94  5613  9109:*         :=======================================*
  96  4132  7048:*         :============================           *
  98  3091  5453:*         :======================             *
 100  1976  4220:*         :==============               *
 102  1463  3265:*         :===========           *
 104   906  2526:*         :=======          *
 106   757  1955:*         :======       *
 108   518  1512:*         :====      *
 110   400  1170:*         :===     *
 112   203   905:*         :==    *
 114   152   701:*         :==  *
 116   100   542:*         :=  *
 118    95   419:*         := *
>120   240   325:*         :==*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810561 sequences
  Expectation_n fit: rho(ln(x))= 3.02890.000167; mu= 6.1767 0.009
 mean_var=18.5316 3.629, 0's: 936 Z-trim: 944  B-trim: 0 in 0/64
 Lambda= 0.297933
 Kolmogorov-Smirnov  statistic: 0.0282 (N=29) at  50

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


16 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:21:45 2010 done: Fri Sep 17 21:26:47 2010
 Total Scan time: 268.860 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 9. Bioinformatic analysis of polypeptide 3_4

```
>3_4
KKLKLSMMVNCKWLHVREIYMDQQ
________________________________________________________

Sliding 8 amino acid window search
Database searched = AD_2010
Query = 3_4

Start time: Fri Sep 17 21:26:48 GMT 2010 Finish time: Fri Sep 17 21:26:48 GMT
2010

No 8 amino acid matches exist between 3_4 and the AD_2010 database
________________________________________________________

# fasta34 3_4.pep /genedata/1/db/AD_2010 -Q -E 1 -O 3_4.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_4, 24 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2      0:=
  22     0      0:           one = represents 3 library sequences
  24     0      0:
  26     2      0:=
  28    11      0:====
  30    13      2:*====
  32    12      8:==*=
  34    20     21:======*
  36    20     44:=======       *
  38    46     72:================       *
  40    79    101:==========================      *
  42   103    123:===================================     *
  44   106    136:====================================          *
  46   156    138:================================================*======
  48   149    132:=============================================*======
  50   138    121:=========================================*=====
  52   107    106:==================================*
  54   105     91:==============================*====
  56    40     76:==============           *
  58    85     62:====================*========
  60    41     50:==============  *
  62    34     40:============ *
  64    41     32:==========*===
  66    36     25:========*===
  68    44     20:======*========
  70    14     16:=====*
  72    21     12:===*===
  74     7     10:===*
  76     7      7:==*
  78     2      6:=*
  80     9      4:=*=
  82     1      3:*
  84     3      3:*
  86     2      2:*
  88     1      2:*          inset = represents 1 library sequences
  90     0      1:*
  92     0      1:*         :*
  94     0      1:*         :*
  96     0      1:*         :*
  98     0      0:          *
 100     0      0:          *
 102     0      0:          *
 104     0      0:          *
 106    14      0:=====     *==============
 108     0      0:          *
 110     0      0:          *
 112     0      0:          *
 114     0      0:          *
 116     0      0:          *
 118     0      0:          *
>120     0      0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.70040.00325; mu= 3.7436 0.170
 mean_var=23.0867 5.737, 0's: 2 Z-trim: 2  B-trim: 58 in 1/42
 Lambda= 0.266927
 Kolmogorov-Smirnov  statistic: 0.0643 (N=29) at  44

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
The best scores are:                                          opt bits E(1471)
gi|114841629|dbj|BAF32116.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|114841635|dbj|BAF32119.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|24898904|dbj|BAC23082.1| allergen Cry j 2 [Cryp ( 514)   54 24.9    0.59
gi|24898906|dbj|BAC23083.1| allergen Cry j 2 [Cryp ( 514)   54 24.9    0.59
gi|24898908|dbj|BAC23084.1| allergen Cry j 2 [Cryp ( 514)   54 24.9    0.59
gi|114841607|dbj|BAF32105.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|114841617|dbj|BAF32110.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|114841671|dbj|BAF32137.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|114841657|dbj|BAF32130.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|114841641|dbj|BAF32122.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|114841653|dbj|BAF32128.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|1171004|sp|P43212.1|PGLR2_CRYJA RecName: Full=P ( 514)   54 24.9    0.59
gi|114841663|dbj|BAF32133.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
gi|114841665|dbj|BAF32134.1| pollen allergen [Cryp ( 514)   54 24.9    0.59
```

```
>>gi|114841629|dbj|BAF32116.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                    10        20
3_4                          KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841635|dbj|BAF32119.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                    10        20
3_4                          KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|24898904|dbj|BAC23082.1| allergen Cry j 2 [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                    10        20
3_4                          KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|248 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|248 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|24898906|dbj|BAC23083.1| allergen Cry j 2 [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                    10        20
3_4                          KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|248 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|248 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|24898908|dbj|BAC23084.1| allergen Cry j 2 [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                    10        20
3_4                          KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|248 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|248 GLRLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841607|dbj|BAF32105.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                    10        20
3_4                          KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841617|dbj|BAF32110.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                    10        20
3_4                          KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841671|dbj|BAF32137.1| pollen allergen [Cryptome  (514 aa)
```

Confidential Attachment

```
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                10        20
3_4                         KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841657|dbj|BAF32130.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                10        20
3_4                         KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLRLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841641|dbj|BAF32122.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                10        20
3_4                         KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841653|dbj|BAF32128.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                10        20
3_4                         KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|1171004|sp|P43212.1|PGLR2_CRYJA RecName: Full=Polyg  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                10        20
3_4                         KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|117 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|117 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841663|dbj|BAF32133.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                10        20
3_4                         KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250

>>gi|114841665|dbj|BAF32134.1| pollen allergen [Cryptome  (514 aa)
 initn:  54 init1:  54 opt:  54  Z-score: 106.5  bits: 24.9 E(): 0.59
Smith-Waterman  score:  54;   53.846%  identity  (76.923%  similar)  in  13  aa
overlap (11-23:168-180)

                                10        20
3_4                         KKLKLSMMVNCKWLHVREIYMDQQ
                                    :::.. :::  :.
gi|114 IWLQFAKLTGFTLMGKGVIDGQGKQWWAGQCKWVNGREICNDRDRPTAIKFDFSTGLIIQ
       140       150       160       170       180       190

gi|114 GLKLMNSPEFHLVFGNCEGVKIIGISITAPRDSPNTDGIDIFASKNFHLQKNTIGTGDDC
       200       210       220       230       240       250
```

Confidential Attachment

```
24 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:26:47 2010 done: Fri Sep 17 21:26:47 2010
 Total Scan time:  0.040 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_4.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 3_4.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_4, 24 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    61     0:====
  22     0     0:           one = represents 16 library sequences
  24     0     0:
  26     7     0:=
  28    11     2:*
  30    32    12:*=
  32    99    45:==*====
  34   243   122:=======*========
  36   228   250:===============*
  38   408   414:=========================*
  40   619   577:====================================*==
  42   539   706:=====================================          *
  44   623   779:==========================================          *
  46   694   793:============================================      *
  48   532   759:==================================              *
  50   941   693:=============================================*===============
  52   750   609:=========================================*========
  54   683   520:==============================*==========
  56   357   435:=======================     *
  58   251   357:================      *
  60   197   289:=============     *
  62   272   232:==============*==
  64   258   184:===========*=====
  66   167   146:=========*=
  68    70   115:=====  *
  70    90    90:=====*
  72    69    70:====*
  74    36    55:===*
  76    53    43:==*=
  78    18    33:==*
  80    27    26:=*
  82    21    20:=*
  84    20    16:*=
  86    11    12:*
  88     4     9:*          inset = represents 1 library sequences
  90    22     7:*=
  92    13     6:*         :=====*=======
  94     2     4:*         :== *
  96     7     3:*         :==*====
  98     0     3:*         :  *
 100     1     2:*         :=*
 102     5     2:*         :=*===
 104     2     1:*         :*=
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 2.54920.000501; mu= 10.6243 0.026
 mean_var=18.8026 4.008, 0's: 60 Z-trim: 61  B-trim: 514 in 1/61
 Lambda= 0.295778
 Kolmogorov-Smirnov  statistic: 0.0506 (N=29) at  48

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


24 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:26:48 2010 done: Fri Sep 17 21:26:48 2010
 Total Scan time:  0.140 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_4.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 3_4.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_4, 24 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
```

Confidential Attachment

```
< 20 278649     0:==========
  22  1165     0:=          one = represents 26716 library sequences
  24  2466    17:*
  26  5967   374:*
  28 16467  4039:*
  30 44130 24537:*=
  32 114371 94877:===*=
  34 244882 257296:=========*
  36 463424 528425:================== *
  38 752489 873289:============================   *
  40 1100737 1218162:=========================================   *
  42 1366678 1489055:=====================================================   *
  44                                                                  1509864
1642566:=========================================================  *
  46                                                                  1602933
1672994:==========================================================*
  48                                                                  1555643
1601699:==========================================================*
  50 1435694 1461557:==========================================================*
  52 1282550 1284952:==================================================*
  54 1113757 1097575:=======================================*
  56 926509 916811:==================================*
  58 766657 752685:============================*
  60 624218 609719:=======================*=
  62 507796 488813:==================*=
  64 417711 388751:==============*=
  66 337880 307257:===========*=
  68 265070 241682:==========*
  70 215084 189396:=======*=
  72 171122 147995:======*=
  74 144847 115387:====*=
  76 105832 89808:===*
  78 97272 69807:==*=
  80 71820 54204:==*
  82 60017 41465:=*=
  84 54052 32845:=*=
  86 38271 25414:*=
  88 28428 19664:*=         inset = represents 276 library sequences
  90 19822 15215:*
  92 13780 11772:*         :=======================================*
  94 10320  9109:*         :================================*====
  96 10585  7048:*         :========================*=============
  98  5611  5453:*         :===================*=
 100  5443  4220:*         :===============*=====
 102  3973  3265:*         :===========*===
 104  3292  2526:*         :=========*==
 106  2321  1955:*         :=======*=
 108  1508  1512:*         :======*
 110  3144  1170:*         :====*=======
 112  1076   905:*         :===*
 114   586   701:*         :==*
 116  3729   542:*         :=*============
 118   635   419:*         :=*=
>120   948   325:*         :=*==
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810361 sequences
  Expectation_n fit: rho(ln(x))= 3.31270.000172; mu= 7.0715 0.009
 mean_var=24.1122 4.920, 0's: 923 Z-trim: 931  B-trim: 5815 in 2/61
 Lambda= 0.261190
 Kolmogorov-Smirnov  statistic: 0.0373 (N=29) at  52

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


24 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:26:48 2010 done: Fri Sep 17 21:32:50 2010
 Total Scan time: 327.200 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 10. Bioinformatic analysis of polypeptide 3_5

```
>3_5
WSIVNGFMSGKSTWISNEYDGQYGEKERVITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
_____________________________________________________________

Sliding 8 amino acid window search
Database searched = AD_2010
Query = 3_5

Start time: Fri Sep 17 21:32:51 GMT 2010 Finish time: Fri Sep 17 21:32:51 GMT
2010

No 8 amino acid matches exist between 3_5 and the AD_2010 database
_____________________________________________________________

# fasta34 3_5.pep /genedata/1/db/AD_2010 -Q -E 1 -O 3_5.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_5, 63 aa
 vs /genedata/1/db/AD_2010 library
```

```
       opt      E()
< 20     8     0:===
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     2     2:*
  32     2     8:= *
  34     3    21:=     *
  36    21    44:=======       *
  38    49    72:=================      *
  40   139   101:==================================*=============
  42   141   123:=========================================*======
  44   125   136:============================================   *
  46   133   138:=============================================*
  48   123   132:=========================================  *
  50   104   121:==================================     *
  52   136   106:==================================*==========
  54    95    91:==============================*=
  56    83    76:========================*==
  58    55    62:=================== *
  60    35    50:============    *
  62    29    40:==========   *
  64    38    32:==========*==
  66    23    25:========*
  68    28    20:======*===
  70    18    16:======*
  72    13    12:===*=
  74    14    10:===*=
  76    12     7:==*=
  78    18     6:=*====
  80     3     4:=*
  82     3     3:*
  84     3     3:*
  86     9     2:*==
  88     3     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     0     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     1     0:=         *=
 100     0     0:          *
 102     0     0:          *
 104     2     0:=         *==
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.70970.00364; mu= 3.4834 0.190
 mean_var=53.153715.479, 0's: 8 Z-trim: 8  B-trim: 67 in 1/42
 Lambda= 0.175917
 Kolmogorov-Smirnov  statistic: 0.0480 (N=29) at  38

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 36, opt: 24, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(1471)
gi|256429|gb|AAB23464.1| Kunitz trypsin inhibitor  ( 216)   68 24.5    0.84
gi|18770|emb|CAA45777.1| trypsin inhibitor subtype ( 217)   68 24.5    0.84


>>gi|256429|gb|AAB23464.1| Kunitz trypsin inhibitor [Gly  (216 aa)
 initn:  46 init1:  46 opt:  68  Z-score: 103.8  bits: 24.5 E(): 0.84
Smith-Waterman  score:  68;   31.915%  identity  (61.702%  similar)  in  47  aa
overlap (1-44:117-163)

                                          10        20
3_5                                     WSIVNGFMSGKSTWISNEYDGQYG--EKER
                                        ::.:. .  : .. :... :.. :  . ::
gi|256 RIRFIAEGHPLSLKFDSFAVIMLCVGIPTEWSVVEDLPEGPAVKIGENKDAMDGWFRLER
         90       100       110       120       130       140

       30        40        50        60
3_5    VITNFFSIQK-CRCPQRYYKMKVHFDKTTNYDPSYL
       :  . :.  :   :::.
gi|256 VSDDEFNNYKLVFCPQQAEDDKCGDIGISIDHDDGTRRLVVSKNKPLVVQFQKLDKESLA
        150       160       170       180       190       200

>>gi|18770|emb|CAA45777.1| trypsin inhibitor subtype A [  (217 aa)
 initn:  46 init1:  46 opt:  68  Z-score: 103.7  bits: 24.5 E(): 0.84
Smith-Waterman  score:  68;   31.915%  identity  (61.702%  similar)  in  47  aa
overlap (1-44:118-164)

                                          10        20
3_5                                     WSIVNGFMSGKSTWISNEYDGQYG--EKER
                                        ::.:. .  : .. :... :.. :  . ::
gi|187 RIRFIAEGHPLSLKFDSFAVIMLCVGIPTEWSVVEDLPEGPAVKIGENKDAMDGWFRLER
        90       100       110       120       130       140

       30        40        50        60
3_5    VITNFFSIQK-CRCPQRYYKMKVHFDKTTNYDPSYL
       :  . :.  :   :::.
gi|187 VSDDEFNNYKLVFCPQQAEDDKCGDIGISIDHDDGTRRLVVSKNKPLVVQFQKLDKESLA
       150       160       170       180       190       200
```

Confidential Attachment

```
63 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:32:50 2010 done: Fri Sep 17 21:32:50 2010
 Total Scan time:  0.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_5.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 3_5.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_5, 63 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    73     0:=====
  22     0     0:           one = represents 16 library sequences
  24     0     0:
  26     1     0:=
  28     4     2:*
  30     5    12:*
  32    35    45:==*
  34   141   122:=======*=
  36   257   250:===============*=
  38   507   414:=========================*======
  40   564   577:====================================*
  42   773   706:============================================*====
  44   903   779:=================================================*========
  46   788   793:=================================================*
  48   708   759:============================================  *
  50   713   693:============================================*=
  52   314   609:====================                  *
  54   333   520:====================          *
  56   391   435:========================  *
  58   542   357:=====================*===========
  60   468   289:==================*===========
  62   263   232:==============*==
  64   150   184:========== *
  66    91   146:======   *
  68    86   115:====== *
  70    58    90:==== *
  72    82    70:====*=
  74    45    55:===*
  76    26    43:==*
  78    37    33:==*
  80    14    26:=*
  82     8    20:=*
  84    11    16:*
  86     7    12:*
  88    13     9:*          inset = represents 1 library sequences
  90     5     7:*
  92     1     6:*         :=    *
  94    10     4:*         :===*======
  96     0     3:*         :  *
  98     7     3:*         :==*====
 100     1     2:*         :=*
 102     4     2:*         :=*==
 104     0     1:*         :*
 106     2     1:*         :*=
 108     0     1:*         :*
 110     1     1:*         :*
 112     0     0:          *
 114     1     0:=         *=
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.64210.000636; mu= 4.6756 0.031
 mean_var=43.2328 9.569, 0's: 73 Z-trim: 73  B-trim: 372 in 1/61
 Lambda= 0.195060
 Kolmogorov-Smirnov  statistic: 0.0338 (N=29) at  44

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 36, opt: 24, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


63 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:32:51 2010 done: Fri Sep 17 21:32:51 2010
 Total Scan time:  0.220 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_5.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 3_5.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_5, 63 aa
```

Confidential Attachment

```
 vs /genedata/1/db/PRT_2010 library

      opt      E()
< 20 350022     0:=============
  22   102     0:=          one = represents 27932 library sequences
  24   219    17:*
  26   552   374:*
  28  2211  4039:*
  30 12083 24538:*
  32 58834 94880:===*
  34 208821 257302:======== *
  36 514420 528437:==================*
  38 908604 873310:================================*=
  40 1372028 1218191:============================================*======
  42                                                                1628176
1489090:=========================================================*=====
  44                                                                1675864
1642605:==========================================================*=
  46                                                                1596810
1673034:============================================================ *
  48  1501850  1601737:=============================================================
*
  50 1340695 1461592:================================================    *
  52 1150002 1284983:============================================    *
  54 974795 1097601:==================================    *
  56 857875 916833:=============================== *
  58 735855 752703:=========================*
  60 594524 609733:======================*
  62 488465 488825:=================*
  64 393595 388760:==============*=
  66 319497 307264:============*
  68 245159 241687:=========*
  70 195544 189400:=======*=
  72 155193 147998:======*
  74 127457 115389:=====*
  76 92557 89810:===*
  78 73699 69808:==*
  80 57365 54205:=*=
  82 43816 41466:=*
  84 33112 32846:=*
  86 24290 25414:*
  88 18419 19664:*          inset = represents 217 library sequences
  90 14764 15215:*
  92 10802 11773:*         :=======================================*
  94  8356  9109:*         :=======================================*
  96  6339  7048:*         :============================  *
  98  5196  5453:*         :======================== *
 100  3391  4220:*         :================   *
 102  2418  3265:*         :============   *
 104  1786  2526:*         :=========  *
 106  1429  1955:*         :=======  *
 108  1197  1512:*         :======*
 110   903  1170:*         :======*
 112   590   905:*         :=== *
 114   467   701:*         :===*
 116   345   542:*         :==*
 118   222   419:*         :=*
>120   510   325:*         :=*=
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810786 sequences
  Expectation_n fit: rho(ln(x))= 3.91300.000185; mu= 7.4896 0.010
 mean_var=45.0437 8.989, 0's: 1143 Z-trim: 1144  B-trim: 0 in 0/64
 Lambda= 0.191099
 Kolmogorov-Smirnov  statistic: 0.0227 (N=29) at  62

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 36, opt: 24, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


63 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:32:51 2010 done: Fri Sep 17 21:37:38 2010
 Total Scan time: 254.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Appendix 11. Bioinformatic analysis of polypeptide 3_6

```
>3_6
LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
_______________________________________________

Sliding 8 amino acid window search
Database searched = AD_2010
Query = 3_6

Start time: Fri Sep 17 21:37:39 GMT 2010 Finish time: Fri Sep 17 21:37:39 GMT
2010

No 8 amino acid matches exist between 3_6 and the AD_2010 database
_______________________________________________

# fasta34 3_6.pep /genedata/1/db/AD_2010 -Q -E 1 -O 3_6.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
```

```
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_6, 34 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     1     0:=
  30    11     2:*===
  32    36     8:==*=========
  34    33    21:======*====
  36    28    44:==========    *
  38    54    72:==================     *
  40    77   101:==========================       *
  42    84   123:============================             *
  44   128   136:=============================================  *
  46   114   138:=======================================       *
  48   130   132:============================================*
  50   145   121:============================================*========
  52    79   106:==========================        *
  54   163    91:==============================*========================
  56    61    76:====================    *
  58    44    62:===============     *
  60    53    50:================*=
  62    58    40:=============*======
  64    28    32:==========*
  66    27    25:========*
  68    28    20:======*===
  70    22    16:======*==
  72    18    12:===*==
  74     5    10:== *
  76    10     7:==*=
  78     3     6:=*
  80     6     4:=*
  82     4     3:*=
  84     4     3:*=
  86     1     2:*
  88     1     2:*          inset = represents 1 library sequences
  90     1     1:*
  92     2     1:*         :*=
  94     1     1:*         :*
  96     0     1:*         :*
  98     0     0:          *
 100     0     0:          *
 102     0     0:          *
 104     0     0:          *
 106     2     0:=         *==
 108     0     0:          *
 110     2     0:=         *==
 112     5     0:==        *=====
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.00470.00311; mu= 5.3176 0.160
 mean_var=23.7078 5.723, 0's: 2 Z-trim: 9  B-trim: 121 in 1/42
 Lambda= 0.263408
 Kolmogorov-Smirnov  statistic: 0.0571 (N=29) at  52

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(1471)
gi|15384338|gb|AAK96255.1|AF177030_1 acidic allerg ( 244)   58 25.5    0.26
gi|10314021|gb|AAF80379.2| acidic Cyn d 1 isoaller ( 244)   58 25.5    0.26
gi|16076693|gb|AAL14077.1| acidic Cyn d 1 isoaller ( 262)   58 25.5    0.28
gi|16076695|gb|AAL14078.1| acidic Cyn d 1 isoaller ( 262)   58 25.5    0.28
gi|16076697|gb|AAL14079.1|AF177380_1 acidic Cyn d  ( 262)   58 25.5    0.28
gi|89892723|gb|ABD79095.1| Zea m 1 allergen [Zea m ( 252)   57 25.1    0.34
gi|115502167|sp|Q1ZYQ8.2|EXB10_MAIZE RecName: Full ( 270)   57 25.1    0.37
gi|14423757|sp|O04701.1|MPAC1_CYNDA RecName: Full= ( 246)   55 24.4    0.57
gi|168419914|gb|ACA23876.1| Pas n 1 allergen precu ( 265)   55 24.4    0.62


>>gi|15384338|gb|AAK96255.1|AF177030_1 acidic allergen C  (244 aa)
 initn:  58 init1:  58 opt:  58  Z-score: 113.0  bits: 25.5 E(): 0.26
Smith-Waterman  score:  58;   26.087%  identity  (78.261%  similar)  in  23  aa
overlap (8-30:120-142)

                                      10        20        30
3_6                                LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                      :..:: :. . ... ..... ::
gi|153 LIKITDKNYEHIAAYHFDLSGKAFGAMAKKGEEDKLRKAGELMLQFRRVKCEYPSDTKIA
      90        100       110       120       130       140

gi|153 FHVEKGSSPNYLALLVKYAAGDGNIVGVDIKPKGSDEFLPMKQSWGAIWRIDPPKPLKGP
     150       160       170       180       190       200

>>gi|10314021|gb|AAF80379.2| acidic Cyn d 1 isoallergen   (244 aa)
 initn:  58 init1:  58 opt:  58  Z-score: 113.0  bits: 25.5 E(): 0.26
Smith-Waterman  score:  58;   26.087%  identity  (78.261%  similar)  in  23  aa
overlap (8-30:120-142)

                                      10        20        30
3_6                                LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                      :..:: :. . ... ..... ::
gi|103 LIKITDKNYEHIAAYHFDLSGKAFGAMAKKGEEDKLRKAGELMLQFRRVKCEYPSDTKIA
```

Confidential Attachment

```
      90        100       110       120       130       140

gi|103 FHVEKGSSPNYLALLVKYAAGDGNIVGVDIKPKGSDEFLPMKQSWGAIWRMDPPKPLKGP
     150       160       170       180       190       200

>>gi|16076693|gb|AAL14077.1| acidic Cyn d 1 isoallergen   (262 aa)
 initn:  58 init1:  58 opt:  58  Z-score: 112.4  bits: 25.5 E(): 0.28
Smith-Waterman score:  58;   26.087% identity (78.261% similar) in 23 aa
overlap (8-30:138-160)

                                    10        20        30
3_6                            LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                   :..:: :. . ... ..... ::
gi|160 LIKITDKNYEHIAAYHFDLSGKAFGAMAKKGEEDKLRKAGELMLQFRRVKCEYPSDTKIA
       110       120       130       140       150       160

gi|160 FHVEKGSNPNYLALLVKYAAGDGNIVSVDIKSKGSDEFLPMKQSWGAIWRIDPPKPLKGP
       170       180       190       200       210       220

>>gi|16076695|gb|AAL14078.1| acidic Cyn d 1 isoallergen   (262 aa)
 initn:  58 init1:  58 opt:  58  Z-score: 112.4  bits: 25.5 E(): 0.28
Smith-Waterman score:  58;   26.087% identity (78.261% similar) in 23 aa
overlap (8-30:138-160)

                                    10        20        30
3_6                            LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                   :..:: :. . ... ..... ::
gi|160 LIKITDKNYEHIAAYHFDLSGKAFGAMAKKGEEDKLRKAGELMLQFRRVKCEYPSDTKIT
       110       120       130       140       150       160

gi|160 FHVEKGSSPNYLALLVKYAAGDGNIVGVDIKPKGSDVFLPMKLSWGAIWRMDPPKPLKGP
       170       180       190       200       210       220

>>gi|16076697|gb|AAL14079.1|AF177380_1 acidic Cyn d 1 is  (262 aa)
 initn:  58 init1:  58 opt:  58  Z-score: 112.4  bits: 25.5 E(): 0.28
Smith-Waterman score:  58;   26.087% identity (78.261% similar) in 23 aa
overlap (8-30:138-160)

                                    10        20        30
3_6                            LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                   :..:: :. . ... ..... ::
gi|160 LIKITDKNYEHIAAYHFDLSGKAFGAMAKKGEEDKLRKAGELMLQFRRVKCEYPSDTKIA
       110       120       130       140       150       160

gi|160 FHVEKGSNPNYLALLVKYAAGDGNIVSVDIKSKGSDEFLPMKQSWGAIWRIDPPKPLKGP
       170       180       190       200       210       220

>>gi|89892723|gb|ABD79095.1| Zea m 1 allergen [Zea mays]  (252 aa)
 initn:  57 init1:  57 opt:  57  Z-score: 110.7  bits: 25.1 E(): 0.34
Smith-Waterman score:  57;   25.926% identity (77.778% similar) in 27 aa
overlap (8-34:133-159)

                                        10        20        30
3_6                                LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                       :...: :. . : : ..... .::...
gi|898 VVHITDMNYEPIAAYHFDLAGTAFGAMAKKGEEEKLRKAGIIDMQFRRVKCKYDSKVTFH
           110       120       130       140       150       160

gi|898 LEKGCGPNYLALLVKYVDGDGDIVAVDVKEKGSDTYEPLKHSWGAIWRKDSDKPLKGPLT
           170       180       190       200       210       220

>>gi|115502167|sp|Q1ZYQ8.2|EXB10_MAIZE RecName: Full=Exp  (270 aa)
 initn:  57 init1:  57 opt:  57  Z-score: 110.1  bits: 25.1 E(): 0.37
Smith-Waterman score:  57;   25.926% identity (77.778% similar) in 27 aa
overlap (8-34:151-177)

                                        10        20        30
3_6                                LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                       :...: :. . : : ..... .::...
gi|115 VVHITDMNYEPIAAYHFDLAGTAFGAMAKKGEEEKLRKAGIIDMQFRRVKCKYDSKVTFH
             130       140       150       160       170       180

gi|115 LEKGCGPNYLALLVKYVDGDGDIVAVDVKEKGSDTYEPLKHSWGAIWRKDSDKPLKGPLT
             190       200       210       220       230       240

>>gi|14423757|sp|O04701.1|MPAC1_CYNDA RecName: Full=Majo  (246 aa)
 initn:  55 init1:  55 opt:  55  Z-score: 106.8  bits: 24.4 E(): 0.57
Smith-Waterman score:  55;   26.087% identity (73.913% similar) in 23 aa
overlap (8-30:120-142)

                                        10        20        30
3_6                                LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                       ::.:: :. . . . ..... .:
gi|144 LVKITDKNYEHIAAYHFDLSGKAFGAMAKKGQEDKLRKAGELTLQFRRVKCKYPSGTKIT
       90        100       110       120       130       140

gi|144 FHIEKGSNDHYLALLVKYAAGDGNIVAVDIKPRDSDEFIPMKSSWGAIWRIDPKKPLKGP
     150       160       170       180       190       200

>>gi|168419914|gb|ACA23876.1| Pas n 1 allergen precursor  (265 aa)
 initn:  55 init1:  55 opt:  55  Z-score: 106.1  bits: 24.4 E(): 0.62
Smith-Waterman score:  55;   28.000% identity (76.000% similar) in 25 aa
overlap (8-32:142-166)

                                        10        20        30
3_6                                LLFYHYYGQQDKKRNYSSIIMTLKKIEVEYDGQL
                                       : .:: :.:. . . ..... .:.:
gi|168 TVFITDMNYEPIAPYHFDLSGKAFGAMAKPGLNDKLRHYGIFDLEFRRVRCKYQGGQKIV
           120       130       140       150       160       170
```

Confidential Attachment

```
gi|168 FHVEKGSNPNYLAMLVKFVADDGDIVLMELKEKSSDWKPMKLSWGAIWRMDTPKALVPPF
              180       190       200       210       220       230




34 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:37:39 2010 done: Fri Sep 17 21:37:39 2010
 Total Scan time:  0.030 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_6.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 3_6.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_6, 34 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    61     0:=====
  22     0     0:           one = represents 13 library sequences
  24     0     0:
  26     2     0:=
  28     4     2:*
  30    30    12:*==
  32    28    45:===*
  34    67   122:======   *
  36   261   250:====================*=
  38   567   414:================================*============
  40   544   577:=========================================  *
  42   739   706:=========================================================*==
  44   736   779:============================================================  *
  46   633   793:================================================           *
  48   730   759:============================================================ *
  50   644   693:================================================    *
  52   565   609:=========================================  *
  54   490   520:======================================== *
  56   578   435:================================*===========
  58   313   357:========================  *
  60   267   289:==================== *
  62   215   232:=================*
  64   163   184:============= *
  66   162   146:===========*=
  68   137   115:========*==
  70    56    90:===== *
  72   177    70:=====*========
  74    82    55:====*==
  76    31    43:===*
  78    31    33:==*
  80    30    26:=*=
  82    20    20:=*
  84    13    16:=*
  86    25    12:*=
  88    10     9:*          inset = represents 1 library sequences
  90    15     7:*=
  92     1     6:*          :=    *
  94    13     4:*          :===*=========
  96     2     3:*          :==*
  98     0     3:*          :  *
 100     0     2:*          : *
 102     0     2:*          : *
 104     0     1:*          :*
 106     0     1:*          :*
 108     0     1:*          :*
 110     0     1:*          :*
 112     1     0:=          *=
 114     0     0:           *
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.70850.000518; mu= 1.9831 0.026
 mean_var=29.1920 6.977, 0's: 60 Z-trim: 60  B-trim: 598 in 1/61
 Lambda= 0.237379
 Kolmogorov-Smirnov  statistic: 0.0289 (N=29) at  54

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


34 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:37:39 2010 done: Fri Sep 17 21:37:40 2010
 Total Scan time:  0.190 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

```
# fasta34 3_6.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 3_6.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7,
2006
```

Confidential Attachment

```
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_6, 34 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 279732     0:===========
  22   1187     0:=          one = represents 27195 library sequences
  24   2637    17:*
  26   7899   374:*
  28 26204  4039:*
  30 76298 24538:*==
  32 187410 94880:===*===
  34 399501 257303:=========*=====
  36 640071 528440:==================*====
  38 960752 873314:================================*===
  40 1240161 1218196:============================================*=
  42 1478774 1489097:=========================================================*
  44                                                                  1631696
1642612:=============================================================*
  46                                                                  1601697
1673042:=============================================================*
  48                                                                  1544039
1601744:========================================================== *
  50 1351201 1461598:=================================================    *
  52 1175514 1284989:============================================   *
  54 987957 1097606:=======================================   *
  56 845256 916837:================================ *
  58 693631 752706:========================== *
  60 557757 609736:===================== *
  62 455065 488827:=================*
  64 372957 388762:==============*
  66 290767 307265:===========*
  68 230143 241689:========*
  70 180872 189401:=======*
  72 140718 147999:======*
  74 109925 115390:====*
  76 85845 89811:===*
  78 63738 69809:==*
  80 47628 54205:=*
  82 36035 41466:=*
  84 27844 32846:=*
  86 20538 25415:*
  88 15799 19664:*          inset = represents 173 library sequences
  90 11606 15215:*
  92  8621 11773:*         :=====================================*
  94  6835  9109:*         :=====================================*
  96  4595  7048:*         :=========================           *
  98  3276  5454:*         :==================            *
 100  2402  4220:*         :==============          *
 102  1823  3265:*         :===========       *
 104  1412  2526:*         :=========     *
 106   861  1955:*         :=====       *
 108   734  1512:*         :=====   *
 110   461  1170:*         :===   *
 112   328   905:*         :==   *
 114   264   701:*         :==  *
 116   218   542:*         :== *
 118   234   419:*         :==*
>120   307   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810866 sequences
  Expectation_n fit: rho(ln(x))= 3.70950.000172; mu= 7.5713 0.009
 mean_var=30.5735 6.034, 0's: 953 Z-trim: 955  B-trim: 2 in 1/64
 Lambda= 0.231954
 Kolmogorov-Smirnov  statistic: 0.0309 (N=29) at  40

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


34 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Fri Sep 17 21:37:40 2010 done: Fri Sep 17 21:44:22 2010
 Total Scan time: 378.020 Total Display time:  0.000

Function used was FASTA [version 3.4t26 July 7, 2006]
```

Database checksum values:

```
Fri Sep 17 21:44:22 GMT 2010      a184245745a6ed8c6ecde45b26637bba
/genedata/1/db/AD_2010

Fri Sep 17 21:44:22 GMT 2010      17c3a19148dfb0163e270cf41e2aa437
/genedata/1/db/TOX_2010

Fri Sep 17 21:46:21 GMT 2010      e657d3127c1aad11f9f7df8dcc5e448c
/genedata/1/db/PRT_2010
```

Confidential Attachment