

Study Title

**Updated Bioinformatics Evaluation of the CP4 EPSPS Protein Utilizing the
AD_2009 and TOX_2009 Databases**

Authors

Andre Silvanovich, Ph.D.

Study Completed On

February 23, 2009

Sponsor and Performing Laboratory

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

Laboratory Project ID

MSL number: MSL0021842

Study number: REG-09-021

**Previous MSL number for full report: 18752 dated: October 3, 2003
Previous MSL number for update with AD8: MSL0021228 dated: March 13, 2008**

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 No Data Confidentiality Claim

No claim of data confidentiality is made for any information contained in this study on the basis of its falling within the scope of FIFRA §10 (d)(1)(A), (B), or (C).

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

Company: _____ Monsanto Company _____

Company Agent: _____

Title: _____

Signature: _____ Date: _____

Statement of Compliance

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

Submitter

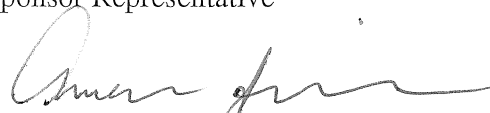
Date: _____



Natalia Bogdanova D. V. M.

Sponsor Representative

Date: 23 Feb 2009



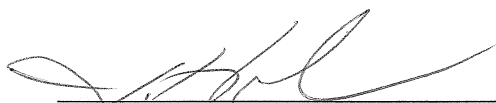
Andre Silvanovich, Ph.D.

Author

Date: Feb 23, 2009


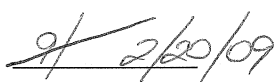
Summary of Quality Control Review

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



Quality Assurance Specialist
Monsanto Regulatory
Monsanto Company

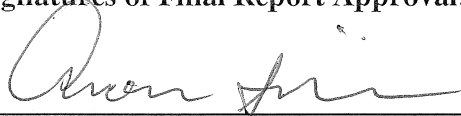
Date:

Study Certification

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

Signatures of Final Report Approval:



Andre Silvanovich, Ph.D.
Author

Date: Feb 23, 2009



Gary A. Bannon, Ph.D.
Lead, Protein and Molecular Sciences
Product Characterization Center

Date: 02-23-2009

Study Information

Study Number: REG-09-021

MSL Number: MSL0021842

Title: Updated Bioinformatics Evaluation of the CP4 EPSPS Protein Utilizing the AD_2009 and TOX_2009 Databases

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

Sponsor Representative: Natalia Bogdanova D. V. M.

Authors: Andre Silvanovich, Ph.D.

Study Start Date: January 22, 2009

Study Completion Date: February 23, 2009

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

© 2009 Monsanto Company. All Rights Reserved.

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

Table of Contents

	Page
Study Title.....	1
Statement of No Data Confidentiality Claim.....	2
Statement of Compliance.....	3
Summary of Quality Control Review	4
Study Certification	5
Study Information	6
Table of Contents	7
1.0 Summary	8
2.0 Sequence Database Preparation	8
3.0 Sequence Database Searches	8
4.0 Significance of the Alignment.	9
5.0 Results and Conclusions	9
6.0 References	10
Appendix 1. The CP4 EPSPS Protein Sequence	11
Appendix 2. Bioinformatic analysis of CP4 EPSPS	12

1.0 Summary

The bioinformatic evaluation of CP4 EPSPS has been conducted several times throughout the research and development process, with the most recent report (McClain and Silvanovich, 2008) concluding that the CP4 EPSPS protein was not similar to known allergens that may adversely affect animal or human health. Periodically, the databases used to evaluate proteins are updated. Since the most recent report was completed, the allergen database has been revised and published (FARRP, 2009 and Silvanovich, 2009) and a new toxin sequence database has been assembled (Silvanovich 2009). In order to determine if the CP4 EPSPS protein shared significant sequence similarity to new sequences contained in the updated allergen or toxin databases, the CP4 EPSPS protein sequence was used as a query for a FASTA and ALLERGENSEARCH search of the AD_2009 database and a FASTA search of the TOX_2009 database.

Results indicate there were no biologically relevant sequence similarities to allergens or toxins when the CP4 EPSPS protein sequence was used as a query for a FASTA search of the AD_2009 database. Furthermore, no short (eight amino acid) polypeptide matches were shared between the CP4 EPSPS protein sequence and proteins in the updated allergen database. These data indicate the lack of both structurally and immunologically relevant similarities between the CP4 EPSPS protein sequence and known allergens and toxins.

2.0 Sequence Database Preparation

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

The toxin database is a subset of sequences derived from the PRT_2009 database (Silvanovich 2009), that was selected using a keyword search and filtered to remove likely non-toxin proteins and proteins that are not relevant to human or animal health. It is referred to herein as the TOX_2009 database and contains 7,651 sequences. A complete description of the TOX_2009 database can be found in Silvanovich (2009).

3.0 Sequence Database Searches

FASTA analyses using the AD_2009 and TOX_2009 databases were performed on a desktop computer loaded with a SUSE LINUX version 10.1 operating system and FASTA version 3.4t26 and dated July 7, 2006. The structural similarity of the CP4 EPSPS protein sequence to sequences in the AD_2009 and TOX_2009 databases were assessed using the FASTA algorithm (Lipman and Pearson, 1985; Pearson and Lipman,

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

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. Specific FASTA comparison parameters used in this study included a wordsize (k-tuple) of two, a gap creation penalty of 12, a gap extension penalty of two, and an expectation threshold (*E*-score) of ten. FASTA comparisons were performed using the BLOSUM50 scoring matrix (Henikoff and Henikoff, 1992). Multiple alignments are made between the CP4 EPSPS sequence and each sequence in the database, and a score was calculated for each alignment. Only the best scoring alignment is extensively analyzed for each database sequence. The BLOSUM matrix series (Henikoff and Henikoff, 1992) was derived from a set of aligned, ungapped regions from protein families, called the BLOCKS database. Sequences from each block were clustered based on the percent of identical residues in the alignments (Henikoff and Henikoff, 1996). The BLOSUM50 matrix will identify blocks of conserved residues that are at least 50% identical. BLOSUM50 works well for identifying sequence similarities that include gaps, and thus recognizes distant evolutionary relationships (Pearson, 2000).

4.0 Significance of the Alignment.

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

5.0 Results and Conclusions

Potential structural similarities shared between the CP4 EPSPS protein (Appendix 1) and proteins in the allergen database were evaluated using the FASTA sequence alignment tool. Identified proteins were ranked according to their degree of similarity (Appendix 2). Results indicate the lack of sequence similarity between the CP4 EPSPS protein sequence and sequences in the allergen database. No alignment met nor exceeded the threshold of 35% identity over 80 amino acids recommended by Codex Alimentarius (2003). Furthermore, no eight contiguous amino acid identities were detected when the CP4 EPSPS protein sequence was compared to the AD_2009 sequence database (Appendix 2). When used to query the TOX_2009 database, no significant sequence alignments, those yielding an *E*-score of $1 \text{ e-}5$ or less, were observed (Appendix 2).

6.0 References

Codex Alimentarius Commission (2003). Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants. CAC/GL 45-2003.

FARRP (Food Allergy Research and Resource Program Database). (2009).
www.allergenonline.com. University of Nebraska, Lincoln, NE.

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

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

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

McClain, J.S. and Silvanovich, A. (2008). Updated bioinformatics evaluation of the CP4 EPSPS protein utilizing the AD8 database, MSL0021228, St. Louis, MO.

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

Pearson, W.R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* **132**:185-21.

Silvanovich A. (2009). The assembly of AD_2009, TOX_2009 and PRT_2009, MSL0021840, St. Louis, MO.

Appendix 1. The CP4 EPSPS Protein Sequence

1 MLHGASSRPA TARKSSGLSG TVRIPGDKSI SHRSFMFGGL ASGETRITGL
51 LEGEDVINTG KAMQAMGARI RKEGDTWIID GVGNGGLLAP EAPLDFGNAA
101 TGCRLTMGLV GVDYDFDSTFI GDASLTKRPM GRVLNPLREM GVQVKSEDGD
151 RLPVTLRGPK TPTPITYRVP MASAQVKS AV LLAGLNTPGI TTVIEPIMTR
201 DHTEKMLQGF GANLTVETDA DGVRTIRLEG RGKLTGQVID VPGDPSSTAF
251 PLVAALLVPG SDVTILNVLM NPTRTGLILT LQEMGADIEV INPRLAGGED
301 VADLRVRSST LKGVTVPEDR APSMIDEYPI LAVAAFAEG ATVMNGLEEL
351 RVKESDRLSA VANGLKLNGV DCDEGETSLV VRGRPDGKGL GNASGAAVAT
401 HLDHRIAMSF LVMGLVSENP VTVDDATMIA TSFPEFMDLM AGLGAKIELS
451 DTKAA

Appendix 2. Bioinformatic analysis of CP4 EPSPS

Sliding 8 amino acid window search

Database searched = AD_2009

Query = CP4_EPSPS

Start time: Thu Jan 22 17:47:21 CST 2009 Finish time: Thu Jan 22 17:47:21 CST 2009

No matches exist with the AD_2009 database

```
# fasta34 CP4_EPSPS.pep /home/andre/db/AD_2009 -Q -E 10 -O
CP4_EPSPS.pep.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

CP4_EPSPS, 455 aa

vs /home/andre/db/AD_2009 library

```

      opt      E()
< 20      4      0:==
      22      0      0:          one = represents 3 library sequences
      24      0      0:
      26      0      0:
      28      2      0:=
      30      2      2:*
      32      5      7:==*
      34      31     20:=====*====
      36      25     41:=====
      38      42     68:=====
      40      55     95:=====
      42      109    116:=====
      44      101    128:=====
      46      113    130:=====
      48      130    125:=====
      50      123    114:=====
      52      108    100:=====
      54      96     85:=====
      56      62     71:=====
      58      62     59:=====
      60      45     47:=====
      62      50     38:=====
      64      24     30:=====
      66      24     24:=====
      68      25     19:=====
      70      40     15:=====
      72      30     12:=====
      74      25     9:=====
      76      12     7:=====
      78      13     5:=====
```

```

80      4      4:==
82      2      3:*
84      4      3:*
86      4      2:*
88      1      2:*          inset = represents 1 library sequences
90      0      1:*
92      0      1:*          :*
94      5      1:*          :*====
96      7      1:*          :*=====
98      0      0:          *
100     1      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:          *
```

307888 residues in 1386 sequences

Expectation_n fit: rho(ln(x))= 5.35720.00497; mu= 9.9138 0.263

mean_var=72.534923.192, 0's: 4 Z-trim: 4 B-trim: 216 in 1/43

Lambda= 0.150592

Kolmogorov-Smirnov statistic: 0.0886 (N=29) at 46

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 2

join: 37, opt: 25, open/ext: -10/-2, width: 16

The best scores are:

	opt	bits	E(1386)
gi 21725588 emb CAD38375.1 unnamed protein produc (129)	78	25.8	1.4
gi 21725604 emb CAD38383.1 unnamed protein produc (129)	76	25.4	1.9
gi 21725602 emb CAD38382.1 unnamed protein produc (129)	76	25.4	1.9
gi 21725592 emb CAD38377.1 unnamed protein produc (129)	75	25.2	2.2
gi 21725594 emb CAD38378.1 unnamed protein produc (129)	75	25.2	2.2
gi 21725596 emb CAD38379.1 unnamed protein produc (129)	75	25.2	2.2
gi 21725600 emb CAD38381.1 unnamed protein produc (129)	75	25.2	2.2
gi 21725590 emb CAD38376.1 unnamed protein produc (129)	75	25.2	2.2
gi 1346568 sp P49372.1 ALL1 APIGR RecName: Full=Ma (154)	75	25.2	2.5
gi 21725586 emb CAD38374.1 unnamed protein produc (129)	74	24.9	2.5
gi 21725584 emb CAD38373.1 unnamed protein produc (129)	74	24.9	2.5
gi 21725582 emb CAD38372.1 unnamed protein produc (129)	74	24.9	2.5
gi 17978844 gb AAL47677.1 major Der f 2 isoform [(129)	73	24.7	2.9
gi 21920 emb CAA39099.1 CM2 protein [Triticum tur (145)	70	24.1	5.1
gi 9280360 gb AAF86369.1 major allergen I 18kDa a (150)	68	23.7	7
gi 21713 emb CAA35597.1 unnamed protein product [(168)	68	23.7	7.7
gi 100834 pir S16031 alpha-amylase inhibitor, tet (168)	68	23.7	7.7
gi 54039254 sp P67875.1 RNMG ASPFU RecName: Full=R (176)	68	23.7	8
gi 14285800 sp Q9N2R3.1 TPM_CHAFE RecName: Full=Tr (264)	69	24.1	9.6

```
>>gi|21725588|emb|CAD38375.1| unnamed protein product [D (129 aa)
initn: 57 init1: 57 opt: 78 Z-score: 99.4 bits: 25.8 E(): 1.4
Smith-Waterman score: 78; 25.882% identity (56.471% similar) in 85 aa
overlap (224-303:48-129)
```

```

      200      210      220      230      240      250
CP4_EP IEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFFPL
      : . . . . . : . . . . .
gi|217 VPGCHGNEPCIIGRGKPFQLEALFEANQNSKTAKIEIKASIDGLSVDVPGIDPNACHY--
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP VAALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      : . . . . : . . . . . : . . . . . : . . . . .
gi|217 MNCPLVNGQQYDIKYTNWVPKIAPNSENVVTVKVLG-DNGVLACAIATHAKIQD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP STLKGVTVPEDRAPSMIDEYPI LAVAAAFAGATVMNGLEELRVKESDRLSAVANGKLKLN

>>gi|21725604|emb|CAD38383.1| unnamed protein product [D (129 aa)
  initn: 54 init1: 54 opt: 76 Z-score: 97.0 bits: 25.4 E(): 1.9
Smith-Waterman score: 76; 26.190% identity (55.952% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFFPLV
      : . . . . . : . . . . .
gi|217 PGCHGNEPCIHSGKPFQLEALFEANQNSATAKIEIKASIDGLSVDVPGIDPNACNY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS
      : . . . . : . . . . . : . . . . . : . . . . .
gi|217 KCPLVNGQQYDIKYTNWVPKIAPNSENVVTVKVLG-DNGVLACAIATHAKIQD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP TLKGVTVPEDRAPSMIDEYPI LAVAAAFAGATVMNGLEELRVKESDRLSAVANGKLKLN

>>gi|21725602|emb|CAD38382.1| unnamed protein product [D (129 aa)
  initn: 54 init1: 54 opt: 76 Z-score: 97.0 bits: 25.4 E(): 1.9
Smith-Waterman score: 76; 26.190% identity (55.952% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFFPLV
      : . . . . . : . . . . .
gi|217 PGCHGSEPCIHSGKPFQLEALFEANQNSATAKIEIKASIDGLSVDVPGIDPNACNY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      : . . . . : . . . . . : . . . . . : . . . . .
gi|217 KCPLVNGQQYDIKYTNWVPKIAPNSENVVTVKVLG-DNGVLACAIATHAKIQD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP TLKGVTVPEDRAPSMIDEYPI LAVAAAFAGATVMNGLEELRVKESDRLSAVANGKLKLN
```

```

>>gi|21725592|emb|CAD38377.1| unnamed protein product [D (129 aa)
  initn: 54 init1: 54 opt: 75 Z-score: 95.9 bits: 25.2 E(): 2.2
Smith-Waterman score: 75; 26.190% identity (55.952% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFFPLV
      : . . . . . : . . . . .
gi|217 PGCHGNEPCIIGRGKPFQLEALFEANQNSATAKIEIKASIDGLSVDVPGIDPNACHY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS
      : . . . . : . . . . . : . . . . . : . . . . .
gi|217 NCPLVNGQQYDIKYTNWVPKIAPNSENVVTVKVLG-DNGVLACAIATHAKIQD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP TLKGVTVPEDRAPSMIDEYPI LAVAAAFAGATVMNGLEELRVKESDRLSAVANGKLKLN

>>gi|21725594|emb|CAD38378.1| unnamed protein product [D (129 aa)
  initn: 54 init1: 54 opt: 75 Z-score: 95.9 bits: 25.2 E(): 2.2
Smith-Waterman score: 75; 26.190% identity (55.952% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFFPLV
      : . . . . . : . . . . .
gi|217 PGCHGNEPCIHSGKPFQLEALFEANQNSATAKIEIKASIDGLSVDVPGIDPNACNY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS
      : . . . . : . . . . . : . . . . . : . . . . .
gi|217 KCPLVNGQQYDIKYTNWVPKIAPNSENVVTVKVLG-DNGVLACAIATHAKIRD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP TLKGVTVPEDRAPSMIDEYPI LAVAAAFAGATVMNGLEELRVKESDRLSAVANGKLKLN

>>gi|21725596|emb|CAD38379.1| unnamed protein product [D (129 aa)
  initn: 54 init1: 54 opt: 75 Z-score: 95.9 bits: 25.2 E(): 2.2
Smith-Waterman score: 75; 26.190% identity (54.762% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFFPLV
      : . . . . . : . . . . .
gi|217 PGCHGNEPCIHSGKPFQLEALFEANQNSATAKIEIKASIDGLSVDVPGIDPNACNY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      : . . . . : . . . . . : . . . . . : . . . . .
```

```
gi|217 KCPLVNGQQYDIKYTNVVKIAPKSENVVTVKVLG-DNGVLACAIATHAKIQD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP TLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKNG

>>gi|21725600|emb|CAD38381.1| unnamed protein product [D (129 aa)
  initn: 35 init1: 35 opt: 75 Z-score: 95.9 bits: 25.2 E(): 2.2
Smith-Waterman score: 75; 26.190% identity (55.952% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFPLV
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|217 PGCHGNEPCIHSGKPFQLEALFEANQNSATAKIEIKASIDGLEVDVPGIDPNACNY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|217 KCPLVNGQQYDIKYTNVVKIAPNSENVVTVKVLG-DNGVLACAIATHAKIQD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP TLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKNG

>>gi|21725590|emb|CAD38376.1| unnamed protein product [D (129 aa)
  initn: 54 init1: 54 opt: 75 Z-score: 95.9 bits: 25.2 E(): 2.2
Smith-Waterman score: 75; 26.190% identity (55.952% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFPLV
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|217 PGCHGSEPCIIGRGKPFQLEALFEANQNSATAKIEIKASIDGLSDVDPGIDPNACHY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|217 NCPLVNGQQYDIKYTNVVKIAPNSENVVTVKVLG-DNGVLACAIATHAKIQD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP TLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKNG

>>gi|1346568|sp|P49372.1|ALL1_APIGR RecName: Full=Major (154 aa)
  initn: 38 init1: 38 opt: 75 Z-score: 94.7 bits: 25.2 E(): 2.5
Smith-Waterman score: 75; 22.059% identity (50.735% similar) in 136 aa
overlap (189-319:2-131)

      160      170      180      190      200      210
CP4_EP PKTPTPITYRVPMASQVKSALLAGLNTPGITTVIEPIMTRDHTKMLQGF--GANLTV
      : : . . . : : : : : : : : : : : : : : : : : : : : : : :
gi|134 MGVTQTHVLELTSSVSAEKIFQGFVIDVDVTL
```

```
      10      20      30

      220      230      240      250      260      270
CP4_EP ETDADGV-RTIRLEGRG-KLTGQVIDVP-GDPSSTAFPLVAALLVPGSDVTILNVLMNPT
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|134 PKAAPGAYKSVEIKGDGGPGTLKIITLPDGGPITT-----MTLRIDGVNKEALTFDYSVI
      40      50      60      70      80

      280      290      300      310      320      330
CP4_EP RTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSSTLKGVTVPEDRAPSMIDEYPILAV
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|134 DGDILLGFIESIENHVVLVPTADGG-SICKTTAIFHTKGDVAVPEENIKYANEQNTALFK
      90      100      110      120      130      140

      340      350      360      370      380      390
CP4_EP AAFAEGATVMNGLEELRVKESDRLSAVANGKLKNGVDCDEGETSLVVRGRPDGKGLGNA

gi|134 ALEAYLIAN
      150

>>gi|21725586|emb|CAD38374.1| unnamed protein product [D (129 aa)
  initn: 35 init1: 35 opt: 74 Z-score: 94.7 bits: 24.9 E(): 2.5
Smith-Waterman score: 74; 26.190% identity (55.952% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFPLV
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|217 PGCHGNEPCIIGRGKPFQLEALFEANQNSATAKIEIKASIDGLEVDVPGIDPNACHY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|217 NCPLVNGQQYDIKYTNVVKIAPNSENVVTVKVLG-DNGVLACAIATHAKIQD
      80      90      100      110      120

      310      320      330      340      350      360
CP4_EP TLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKNG

>>gi|21725584|emb|CAD38373.1| unnamed protein product [D (129 aa)
  initn: 54 init1: 54 opt: 74 Z-score: 94.7 bits: 24.9 E(): 2.5
Smith-Waterman score: 74; 26.190% identity (54.762% similar) in 84 aa
overlap (225-303:49-129)

      200      210      220      230      240      250
CP4_EP EPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPG-DPSSTAFPLV
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|217 PGCHGNEPCIIGRGKPFQLEALFEANQNSATAKIEIKASIDGLSDVDPGIDPNACHY--M
      20      30      40      50      60      70

      260      270      280      290      300
CP4_EP AALLVPGSDVTI---LNV-LMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|217 NCPLVNGQQYDIKYTNVVKIAPKSENVVTVKVLG-DNGVLACAIATHAKIQD
```

```

      210      220      230      240      250      260
CP4_EP QGFGANLTVETDADGVRTIRLEGRKLTGQVIDVPDPSSTAFPLVAALLVPG--SDVTI
                                         : : : : : . : : : : : : :
gi|100 CRCEALRYFIALPVPSPQVDPRSGNVGESGLIDLPGCPREMQWDFVRLLLVAPGQCNLATI
      100      110      120      130      140      150

```

```

      270      280      290      300      310      320
CP4_EP LNVLMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSSTLKGVTVDPEDRAPSMI
      :: ::
gi|100 HNVRYCPAVEQPLWI
      160

>>gi|54039254|sp|P67875.1|RNMG_ASPFU RecName: Full=Ribon (176 aa)
      initn: 57 init1: 57 opt: 68 Z-score: 85.7 bits: 23.7 E(): 8
      Smith-Waterman score: 68; 41.667% identity (70.833% similar) in 24 aa
      overlap (77-99:77-100)

      50      60      70      80      90      100
CP4_EP ITGLEGEDVINTGKAMQAMGARIRKEGDTWIIDGV-GNGGLLAPEAPLDFGNAATGCRL
      :: :: :: :: :: ::
gi|540 KRLLYSQAKAESNSHHAPLSDGKTGSSYPHWFTNGYDGNGLIKGRTPIKFGKADCDRPP
      50      60      70      80      90      100

      110      120      130      140      150      160
CP4_EP TMGLVGVDYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKSEGDRLPVTLRGPKTPTPI
gi|540 KHSQNGMGKDDHYLLEFPPTFPDGHYKFDSPKPKEDPGPARVIYTYPNKVFVCGIVAHQRG
      110      120      130      140      150      160

>>gi|14285800|sp|Q9N2R3.1|TPM_CHAFE RecName: Full=Tropom (264 aa)
      initn: 46 init1: 46 opt: 69 Z-score: 84.3 bits: 24.1 E(): 9.6
      Smith-Waterman score: 69; 30.952% identity (58.333% similar) in 84 aa
      overlap (286-358:174-256)

      260      270      280      290      300      310
CP4_EP LLVPGSDVTILNVLMNPTRTGLILTLQEMGADIEVINPRLAGGEDVA-----DLRVRSST
      :: :: :: :: :: ::
gi|142 LENQLKEARFLAEADRKYDEVARKLAMVEADLERAEERAEESGESKIVELEEELRVVGNN
      150      160      170      180      190      200

      320      330      340      350      360
CP4_EP LKGVTVPEDRAPSMIDIEY-----ILA--VAAFAEGATVMNGLEELRVKESDRLSAVANG
      :: :: :: :: :: ::
gi|142 LKSLEVSEKANQREETYKEQIKTLANKLKAEEARAFAERSVQKLQ-KEVDRLEDELVN
      210      220      230      240      250      260

      370      380      390      400      410      420
CP4_EP LKLVGVDCDEGETSLVVRGRPDGKGLGNASGAAVATHLDHRIAMSFLVMGLVSENPVTVD
gi|142 EK

455 residues in 1 query sequences
307888 residues in 1386 library sequences
Scomplib [34t26]
start: Thu Jan 22 17:47:20 2009 done: Thu Jan 22 17:47:21 2009
Total Scan time: 0.150 Total Display time: 0.050
```

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

```

# fasta34 CP4_EPSPS.pep /home/andre/db/TOX_2009 -Q -E 10 -O
CP4_EPSPS.pep.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

CP4_EPSPS, 455 aa
vs /home/andre/db/TOX_2009 library

      opt      E()
< 20      61      0:=====
22      8      0:= one = represents 11 library sequences
24      13      0:=
26      5      0:=
28      50      2:=====
30      34      11:=====
32      104      41:=====
34      325      110:=====*=====
36      254      227:=====*====
38      352      375:===== *
40      569      523:=====*=====
42      543      639:=====

*
44      467      705:=====
*
46      591      718:=====
*
48      602      688:=====
*
50      557      627:===== *
52      582      552:=====*=====
54      501      471:=====*=====
56      371      394:===== *
58      323      323:=====*
60      324      262:=====*=====
62      271      210:=====*=====
64      218      167:=====*=====
66      157      132:=====*=====
68      102      104:=====*
70      83      81:=====*
72      49      64:=====*
74      19      50:===== *
76      5      39:===== *
78      33      30:=====*
80      2      23:===== *
82      3      18:=====*
84      2      14:=====*
86      5      11:=====
88      4      8:===== *
90      11      7:===== *

inset = represents 1 library sequences
```



```
92      3      5:*      :== *
94      2      4:*      :== *
96     14     3:*      :==*=====
98     14     2:*      :==*=====
100     10     2:*      :==*=====
102      0     1:*      :*
104      0     1:*      :*
106      0     1:*      :*
108      0     1:*      :*
110      3     1:*      :==
112      0     0:*      :*
114      0     0:*      :*
116      0     0:*      :*
118      0     0:*      :*
>120     0     0:*      :*
1891534 residues in 7651 sequences
Expectation_n fit: rho(ln(x))= 7.59620.000782; mu= -1.7061 0.038
mean_var=60.926213.601, 0's: 59 Z-trim: 59 B-trim: 757 in 2/60
Lambda= 0.164313
Kolmogorov-Smirnov statistic: 0.0561 (N=29) at 40
```

```
FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 2
join: 37, opt: 25, open/ext: -10/-2, width: 16
The best scores are:
opt bits E(7651)
gi|169246228|gb|ACA51202.1| toxin addiction system ( 101) 80 27.4 2
gi|197936227|gb|ACH73561.1| cytotoxic protein CcdB ( 101) 80 27.4 2
gi|16445235|gb|AAL23453.1| toxin addiction system: ( 101) 80 27.4 2
gi|68166314|gb|AAY88075.1| toxin addiction system ( 101) 73 25.7 6.4
gi|52854786|gb|AAU88265.1| cytolethal distending t ( 258) 80 27.0 6.5
gi|52854778|gb|AAU88259.1| cytolethal distending t ( 258) 80 27.0 6.5
gi|194358746|gb|ACF57188.1| cytotoxic protein CcdB ( 108) 73 25.7 6.9
gi|73476869|gb|AAZ76484.1| Post-segregation toxin ( 101) 72 25.5 7.5
gi|83308562|emb|CAI79534.1| cytotoxic protein CcdB ( 101) 72 25.5 7.5
gi|157076737|gb|ABV16446.1| cytotoxic protein CcdB ( 101) 72 25.5 7.5
gi|194358475|gb|ACF56919.1| cytotoxic protein CcdB ( 101) 72 25.5 7.5
gi|215267872|emb|CAS07542.1| post-segregation syte ( 101) 72 25.5 7.5
gi|157076511|gb|ABV16224.1| cytotoxic protein CcdB ( 101) 72 25.5 7.5
gi|115500692|dbj|BAF33923.1| cytotoxic protein Ccd ( 101) 72 25.5 7.5
gi|119951945|gb|ABM10846.1| cytolethal distending ( 257) 79 26.8 7.6
gi|23574038|emb|CAD48849.1| cytolethal distending ( 258) 79 26.8 7.7
gi|2218089|gb|AAC45442.1| cytolethal distending to ( 258) 79 26.8 7.7
gi|38154551|gb|AAR12200.1| cytolethal distending t ( 258) 79 26.8 7.7
gi|38154555|gb|AAR12203.1| cytolethal distending t ( 258) 79 26.8 7.7
gi|38154547|gb|AAR12197.1| cytolethal distending t ( 258) 79 26.8 7.7
gi|38154559|gb|AAR12206.1| cytolethal distending t ( 258) 79 26.8 7.7
gi|197239658|gb|ACH53455.1| cytolethal distending ( 134) 74 25.9 7.7
gi|197239660|gb|ACH53456.1| cytolethal distending ( 134) 74 25.9 7.7
gi|3337026|dbj|BAA31785.1| cytotoxic protein LetB ( 101) 71 25.2 8.8
gi|45359303|gb|AAS58890.1| cytotoxic protein CcdB ( 101) 71 25.2 8.8
gi|62550776|emb|CAH64699.1| toxin [uncultured bact ( 101) 71 25.2 8.8
gi|52854782|gb|AAU88262.1| cytolethal distending t ( 258) 78 26.6 9
```

```
>>gi|169246228|gb|ACA51202.1| toxin addiction system tox (101 aa)
initn: 53 init1: 53 opt: 80 Z-score: 109.8 bits: 27.4 E(): 2
```

Smith-Waterman score: 94; 31.633% identity (55.102% similar) in 98 aa
overlap (223-312:5-91)

```
200      210      220      230      240
CP4_EP VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGK----TGQVIDVPGDPSST
gi|169      MQFKVYTCKRESRYRLFVDVQSDIIDTPG--RRM
10      20      30

250      260      270      280      290      300
CP4_EP AFPLVAALL----VPGSDVTILNVLNPNTRTGLILTLQEMGADIEVINPRLAGGEDVADL
gi|169 AVPLVSARLLSEKVPRLDLPVMHIGDEPYR---LLT-----TDMTSV-PATVIGEEVADL
40      50      60      70      80

310      320      330      340      350      360
CP4_EP RVRSSTLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANG
gi|169 SLRENDIKNAINLMFRGI
90      100
```

```
>>gi|197936227|gb|ACH73561.1| cytotoxic protein CcdB [Sa (101 aa)
initn: 53 init1: 53 opt: 80 Z-score: 109.8 bits: 27.4 E(): 2
Smith-Waterman score: 94; 31.633% identity (55.102% similar) in 98 aa
overlap (223-312:5-91)
```

```
200      210      220      230      240
CP4_EP VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGK----TGQVIDVPGDPSST
gi|197      MQFKVYTCKRESRYRLFVDVQSDIIDTPG--RRM
10      20      30

250      260      270      280      290      300
CP4_EP AFPLVAALL----VPGSDVTILNVLNPNTRTGLILTLQEMGADIEVINPRLAGGEDVADL
gi|197 AVPLVSARLLSEKVPRLDLPVMHIGDEPYR---LLT-----TDMTSV-PATVIGEEVADL
40      50      60      70      80

310      320      330      340      350      360
CP4_EP RVRSSTLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANG
gi|197 SLRENDIKNAINLMFRGI
90      100
```

```
>>gi|16445235|gb|AAL23453.1| toxin addiction system: tox (101 aa)
initn: 53 init1: 53 opt: 80 Z-score: 109.8 bits: 27.4 E(): 2
Smith-Waterman score: 94; 31.633% identity (55.102% similar) in 98 aa
overlap (223-312:5-91)
```

```
200      210      220      230      240
CP4_EP VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGK----TGQVIDVPGDPSST
gi|164      MQFKVYTCKRESRYRLFVDVQSDIIDTPG--RRM
10      20      30
```

```

      250      260      270      280      290      300
CP4_EP AFPLVAALL----VPGSDVTILNVLNPNTRTGLILTLQEMGADIEVINPRLAGGEDVADL
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|164 AVPLVSARLLSEKVPRLDLPVPMHIGDEPYR---LLT-----TDMTSV-PATVIGEEVADL
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP RVRSSLKGVTPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANG
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|164 SLRENDIKNAINLMFRGI
      90      100

>>gi|68166314|gb|AAV88075.1| toxin addiction system toxi (101 aa)
      initn: 39 init1: 39 opt: 73 Z-score: 100.8 bits: 25.7 E(): 6.4
      Smith-Waterman score: 87; 30.208% identity (56.250% similar) in 96 aa
      overlap (225-312:7-91)

      200      210      220      230      240      250
CP4_EP EPIMTRDHEKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSSTAF
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|681      MQFKGYTCKRESRYRLFVDVQSDIIDTPG--RRMAV
      10      20      30

      260      270      280      290      300
CP4_EP PLVAALL----VPGSDVTILNVLNPNTRTGLILTLQEMGADIEVINPRLAGGEDVADLRV
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|681 PLVSARLLSEKVPRLDLPVPMHIGDEPYR---LLT-----TDMTSV-PATVIGKEVADLSL
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP RSSTLKGVTPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGLK
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|681 RKTDIKNAINLMFRGI
      90      100

>>gi|52854786|gb|AAU88265.1| cytolethal distending toxin (258 aa)
      initn: 46 init1: 46 opt: 80 Z-score: 100.6 bits: 27.0 E(): 6.5
      Smith-Waterman score: 80; 38.298% identity (51.064% similar) in 47 aa
      overlap (132-176:34-80)

      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVYDFDSTFIGDASLTKRPMGRVNLPLREMGVQVKSEGDRLPVTLRGPKT
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|528 KRTSIFIAGVLIPILLNGCSSGKNKAYLDPKVFPPQVEGGPTVPSPDESGLPLPGPGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMASAVLLAGLNTPGITTVIEPIMTRDHEKMLQGFGANLTVETD
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|528 PTNAPIPIPVPGTAPAVSLMNMDSVLTWMSRGAGSSSLWAYIISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|52854778|gb|AAU88259.1| cytolethal distending toxin (258 aa)
      initn: 46 init1: 46 opt: 80 Z-score: 100.6 bits: 27.0 E(): 6.5
```

```

      Smith-Waterman score: 80; 38.298% identity (51.064% similar) in 47 aa
      overlap (132-176:34-80)

      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVYDFDSTFIGDASLTKRPMGRVNLPLREMGVQVKSEGDRLPVTLRGPKT
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|528 KRTSIFIAGVLIPILLNGCSSGKNKAYLDPKVFPPQVEGGPTVPSPDESGLPLPGPGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMASAVLLAGLNTPGITTVIEPIMTRDHEKMLQGFGANLTVETD
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|528 PTNAPIPIPVPGTAPAVSLMNMDSVLTWMSRGAGSSSLWAYIISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|194358746|gb|ACF57188.1| cytotoxic protein CcdB [Sa (108 aa)
      initn: 39 init1: 39 opt: 73 Z-score: 100.1 bits: 25.7 E(): 6.9
      Smith-Waterman score: 73; 29.787% identity (57.447% similar) in 94 aa
      overlap (223-312:12-98)

      200      210      220      230      240
CP4_EP VIEPIMTRDHEKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|194      MRTGTGEMQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30

      250      260      270      280      290      300
CP4_EP AFPLVAALLVPGSDVTILNVLNPNTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVS
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|194 VIPLASARLL--SD-KVSRELYPVVHVGD-ESWRMTSDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80      90

      310      320      330      340      350      360
CP4_EP STLKGVTPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLIN
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|194 NDIKNAINLMFWGI
      100

>>gi|73476869|gb|AAZ76484.1| Post-segregation toxin [Esc (101 aa)
      initn: 39 init1: 39 opt: 72 Z-score: 99.5 bits: 25.5 E(): 7.5
      Smith-Waterman score: 72; 29.787% identity (57.447% similar) in 94 aa
      overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP VIEPIMTRDHEKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|734      MQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30

      250      260      270      280      290      300
CP4_EP AFPLVAALLVPGSDVTILNVLNPNTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVS
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|734 VIPLASARLL--SD-KVSRELYPVVHVGD-ESWRMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80
```

```

      310      320      330      340      350      360
CP4_EP  STLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKLN
      . . :
gi|734  NDIKNAINLMFWGI
      90      100

>>gi|83308562|emb|CAI79534.1| cytotoxic protein CcdB [Es (101 aa)
      initn: 39 init1: 39 opt: 72 Z-score: 99.5 bits: 25.5 E(): 7.5
Smith-Waterman score: 72; 29.787% identity (57.447% similar) in 94 aa
overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP  VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : :
gi|833  MQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30

      250      260      270      280      290      300
CP4_EP  AFPLVAALLVPGSDVTILNVLNMPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      . : : : : : :
gi|833  VIPLASARLL--SD-KVSRELYPVVHVGD-ESWRMMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP  STLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKLN
      . . :
gi|833  NDIKNAINLMFWGI
      90      100

>>gi|157076737|gb|ABV16446.1| cytotoxic protein CcdB [Es (101 aa)
      initn: 39 init1: 39 opt: 72 Z-score: 99.5 bits: 25.5 E(): 7.5
Smith-Waterman score: 72; 29.787% identity (57.447% similar) in 94 aa
overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP  VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : :
gi|157  MQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30

      250      260      270      280      290      300
CP4_EP  AFPLVAALLVPGSDVTILNVLNMPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      . : : : : : :
gi|157  VIPLASARLL--SD-KVSRELYPVVHVGD-ESWRMMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP  STLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKLN
      . . :
gi|157  NDIKNAINLMFWGI
      90      100

>>gi|194358475|gb|ACF56919.1| cytotoxic protein CcdB [Sa (101 aa)
      initn: 39 init1: 39 opt: 72 Z-score: 99.5 bits: 25.5 E(): 7.5
```

Smith-Waterman score: 72; 29.787% identity (57.447% similar) in 94 aa
overlap (223-312:5-91)

```

      200      210      220      230      240
CP4_EP  VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : :
gi|194  MQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30

      250      260      270      280      290      300
CP4_EP  AFPLVAALLVPGSDVTILNVLNMPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      . : : : : : :
gi|194  VIPLASARLL--SD-KVSRELYPVVHVGD-ESWRMMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP  STLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKLN
      . . :
gi|194  NDIKNAINLMFWGI
      90      100

>>gi|215267872|emb|CAS07542.1| post-segregation sytem cy (101 aa)
      initn: 39 init1: 39 opt: 72 Z-score: 99.5 bits: 25.5 E(): 7.5
Smith-Waterman score: 72; 29.787% identity (57.447% similar) in 94 aa
overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP  VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : :
gi|215  MQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30

      250      260      270      280      290      300
CP4_EP  AFPLVAALLVPGSDVTILNVLNMPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      . : : : : : :
gi|215  VIPLASARLL--SD-KVSRELYPVVHVGD-ESWRMMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP  STLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKLN
      . . :
gi|215  NDIKNAINLMFWGI
      90      100

>>gi|157076511|gb|ABV16224.1| cytotoxic protein CcdB [Es (101 aa)
      initn: 39 init1: 39 opt: 72 Z-score: 99.5 bits: 25.5 E(): 7.5
Smith-Waterman score: 72; 29.787% identity (57.447% similar) in 94 aa
overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP  VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : :
gi|157  MQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30
```

```

      250      260      270      280      290      300
CP4_EP AFPLVAALLVPGSDVTILNVLNMPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      ..... : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|157 VIPLASARLL--SD-KVSRELYPVVHVGD-ESWRMMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP STLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKLN
      . : :
gi|157 NDIKNAINLMFWGI
      90      100

>>gi|115500692|dbj|BAF33923.1| cytotoxic protein CcdB (L (101 aa)
  initn: 39 init1: 39 opt: 72 Z-score: 99.5 bits: 25.5 E(): 7.5
Smith-Waterman score: 72; 29.787% identity (57.447% similar) in 94 aa
overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|115                               MQFKVITYKRESRYRLFVDVQSDIIDTPG--RRM
                               10      20      30

      250      260      270      280      290      300
CP4_EP AFPLVAALLVPGSDVTILNVLNMPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVRS
      ..... : : : : : : : : : : : : : : : : : : : : : : :
gi|115 VIPLASARLL--SD-KVSRELYPVVHVGD-ESWRMMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP STLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGKLKLN
      . : :
gi|115 NDIKNAINLMFWGI
      90      100

>>gi|119951945|gb|ABM10846.1| cytolethal distending toxi (257 aa)
  initn: 47 init1: 47 opt: 79 Z-score: 99.4 bits: 26.8 E(): 7.6
Smith-Waterman score: 79; 36.170% identity (48.936% similar) in 47 aa
overlap (132-176:33-79)

      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVDYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKSEGDRLPVTLRGPKT
      . : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|119 KCTSILIVGILIPILLNGCSSRKNAHLDPKVFPPQVEGGPTIPSPDEPGLPLPGAGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMSAQVKSAVLLAGLNTPGITTVIEPIMTRDHTKMLQGFGANLTVETD
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|119 PTNAPIPIPVPGTAPAVSLMNMDSVLTMWSRGAGSSSLWAYIISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|23574038|emb|CAD48849.1| cytolethal distending toxi (258 aa)
  initn: 47 init1: 47 opt: 79 Z-score: 99.4 bits: 26.8 E(): 7.7
```

```

Smith-Waterman score: 79; 36.170% identity (48.936% similar) in 47 aa
overlap (132-176:34-80)

      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVDYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKSEGDRLPVTLRGPKT
      . : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|235 KYTPIFIAGILLPILLNGCSSGKNKAHLDPKVFPPQVEGGPTIPSPDEPGLPLPGAGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMSAQVKSAVLLAGLNTPGITTVIEPIMTRDHTKMLQGFGANLTVETD
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|235 PTNAPIPIPVPGTAPAVSLMNMDSVLTMWSRGAGSSSLWAYIISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|2218089|gb|AAC45442.1| cytolethal distending toxin- (258 aa)
  initn: 47 init1: 47 opt: 79 Z-score: 99.4 bits: 26.8 E(): 7.7
Smith-Waterman score: 79; 36.170% identity (48.936% similar) in 47 aa
overlap (132-176:34-80)

      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVDYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKSEGDRLPVTLRGPKT
      . : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|221 KCTSILIVGILIPILLNGCSSRKNAHLDPKVFPPQVEGGPTIPSPDEPGLPLPGAGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMSAQVKSAVLLAGLNTPGITTVIEPIMTRDHTKMLQGFGANLTVETD
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|221 PTNAPIPIPVPGTAPAVSLMNMDSVLTMWSRGAGSSSLWAYIISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|38154551|gb|AAR12200.1| cytolethal distending toxin (258 aa)
  initn: 47 init1: 47 opt: 79 Z-score: 99.4 bits: 26.8 E(): 7.7
Smith-Waterman score: 79; 36.170% identity (48.936% similar) in 47 aa
overlap (132-176:34-80)

      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVDYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKSEGDRLPVTLRGPKT
      . : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|381 KYTPIFIAGILLPILLNGCSSGKNKAHLDPKVFPPQVEGGPTIPSPDEPGLPLPGAGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMSAQVKSAVLLAGLNTPGITTVIEPIMTRDHTKMLQGFGANLTVETD
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|381 PTNAPIPIPVPGTAPAVSLMNMDSVLTMWSRGAGSSSLWAYIISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|38154555|gb|AAR12203.1| cytolethal distending toxin (258 aa)
  initn: 47 init1: 47 opt: 79 Z-score: 99.4 bits: 26.8 E(): 7.7
Smith-Waterman score: 79; 36.170% identity (48.936% similar) in 47 aa
overlap (132-176:34-80)

      110      120      130      140      150      160
```

```
CP4_EP GCRLTMGLVGVDYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKSSEDGDRLPVTLRGPKT
      . . . . .
gi|381 KCTSILIVGILIPILLNGCSSRKNAHLDPKVFPQVEGGPTIPSPDEPGLPLPGAGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMASAQVKSASVLLAGLNTPGITTVIEPIMTRDHTKMLQGFGANLTVETD
      . . . . .
gi|381 PTNAPIPIPVPGTAPAVSLMNDGSLTMWSRGAGSSWLWAYIYISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|38154547|gb|AAR12197.1| cytolethal distending toxin (258 aa)
  initn: 47 init1: 47 opt: 79 Z-score: 99.4 bits: 26.8 E(): 7.7
Smith-Waterman score: 79; 36.170% identity (48.936% similar) in 47 aa
overlap (132-176:34-80)

      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVDYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKSSEDGDRLPVTLRGPKT
      . . . . .
gi|381 KYTPIFIAGILLPILLNGCSSGKNKAHLDPKVFPQVEGGPTIPSPDEPGLPLPGAGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMASAQVKSASVLLAGLNTPGITTVIEPIMTRDHTKMLQGFGANLTVETD
      . . . . .
gi|381 PTNAPIPIPVPGTAPAVSLMNDGSLTMWSRGAGSSWLWAYIYISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|38154559|gb|AAR12206.1| cytolethal distending toxin (258 aa)
  initn: 47 init1: 47 opt: 79 Z-score: 99.4 bits: 26.8 E(): 7.7
Smith-Waterman score: 79; 36.170% identity (48.936% similar) in 47 aa
overlap (132-176:34-80)

      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVDYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKSSEDGDRLPVTLRGPKT
      . . . . .
gi|381 KYTPIFIAGILLPILLNGCSSGKNKAHLDPKVFPQVEGGPTIPSPDEPGLPLPGAGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMASAQVKSASVLLAGLNTPGITTVIEPIMTRDHTKMLQGFGANLTVETD
      . . . . .
gi|381 PTNAPIPIPVPGTAPAVSLMNDGSLTMWSRGAGSSWLWAYIYISDSNSFGELRNWQIMPG
      70      80      90      100      110      120

>>gi|197239658|gb|ACH53455.1| cytolethal distending toxi (134 aa)
  initn: 45 init1: 45 opt: 74 Z-score: 99.3 bits: 25.9 E(): 7.7
Smith-Waterman score: 74; 22.449% identity (59.184% similar) in 98 aa
overlap (235-327:17-113)

      210      220      230      240      250      260
CP4_EP KMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPGDPSSTAFLVAALLVPGSDVT
      . . . . .
gi|197 ILAVQEAGSPSTAVDTGRVIPSPGIPVRELIWNLSNRSR--QQVY
      10      20      30      40
```

```
      270      280      290      300      310
CP4_EP ILNVLNMNPT--RTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS---TLKGVTVPED
      . . . . .
gi|197 IYFSAVDALGGRVNLALVSNRRADEVFVLSVPRQGGPRLGIRIGNDAFFTAHAIAMRNN
      50      60      70      80      90      100

      320      330      340      350      360      370
CP4_EP RAPSMIDEYPILAVAAAFEGATVMNGLEELRVKESDRLSAVANGLKLVGDCDEGETSL
      . . . . .
gi|197 DAPALVEEVYNFFRDSRDPVHQALNWMIL
      110      120      130

>>gi|197239660|gb|ACH53456.1| cytolethal distending toxi (134 aa)
  initn: 45 init1: 45 opt: 74 Z-score: 99.3 bits: 25.9 E(): 7.7
Smith-Waterman score: 74; 22.449% identity (59.184% similar) in 98 aa
overlap (235-327:17-113)

      210      220      230      240      250      260
CP4_EP KMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPGDPSSTAFLVAALLVPGSDVT
      . . . . .
gi|197 ILAVQEAGSPSTAVDTGRVIPSPGIPVRELIWNLSNRSR--QQVY
      10      20      30      40

      270      280      290      300      310
CP4_EP ILNVLNMNPT--RTGLILTLQEMGADIEVINPRLAGGEDVADLRVRSS---TLKGVTVPED
      . . . . .
gi|197 IYFSAVDALGGRVNLALVSNRRADEVFVLSVPRQGGPRLGIRIGNDAFFTAHAIAMRNN
      50      60      70      80      90      100

      320      330      340      350      360      370
CP4_EP RAPSMIDEYPILAVAAAFEGATVMNGLEELRVKESDRLSAVANGLKLVGDCDEGETSL
      . . . . .
gi|197 DAPALVEEVYNFFRDSRDPVHQALNWMIL
      110      120      130

>>gi|3337026|dbj|BAA31785.1| cytotoxic protein LetB [Esc (101 aa)
  initn: 39 init1: 39 opt: 71 Z-score: 98.2 bits: 25.2 E(): 8.8
Smith-Waterman score: 71; 29.787% identity (56.383% similar) in 94 aa
overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPGDPSST
      . . . . .
gi|333 MQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30

      250      260      270      280      290      300
CP4_EP AFPLVAALLVPGSDVTILNVLNMNPTRTGLILTLQEMGADIEVINPRLAGGEDVADLRVS
      . . . . .
gi|333 VIPLASARLL--SD-KVSRELYPVVHIGD-ESWRMMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP STLKGVTVPEDRAPSMIDEYPILAVAAAFEGATVMNGLEELRVKESDRLSAVANGLKLN
```

```
. . .
gi|333 NDIKNAINLMFWGI
    90      100

>>gi|45359303|gb|AAS58890.1| cytotoxic protein CcdB [Sal (101 aa)
  initn: 53 init1: 53 opt: 71 Z-score: 98.2 bits: 25.2 E(): 8.8
Smith-Waterman score: 85; 30.612% identity (54.082% similar) in 98 aa
overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|453      MQFKVYTCKRESRYRLFVDVQSDIIDTP--ERRM
      10      20      30

      250      260      270      280      290      300
CP4_EP AFPLVAALL----VPGSDVTILNVLNPNRTGLILTLQEMGADIEVINPRLAGGEDVADL
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|453 AVPLVSARLLSEKVPRLDLYPVMHIGDEPYR---LLT-----TDMTSV-PATVIGEEVADL
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP RVRSSTLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANG
      . : . : .
gi|453 SLRENDIKNAINLMFRGI
    90      100

>>gi|62550776|emb|CAH64699.1| toxin [uncultured bacteriu (101 aa)
  initn: 39 init1: 39 opt: 71 Z-score: 98.2 bits: 25.2 E(): 8.8
Smith-Waterman score: 71; 29.787% identity (56.383% similar) in 94 aa
overlap (223-312:5-91)

      200      210      220      230      240
CP4_EP VIEPIMTRDHTKMLQGFGANLTVETDADGVRTIRLEGRGKL----TGQVIDVPGDPSST
      : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|625      MQFKVYTYKRESRYRLFVDVQSDIIDTPG--RRM
      10      20      30

      250      260      270      280      290      300
CP4_EP AFPLVAALLVPGSDVTILNVLNPNRTGLILTLQEMGADIEVINPRLAGGEDVADLRVS
      . : : : : : : : : : : : : : : : : : : : : : : : : : : : :
gi|625 VIPLASARLL--SD-KVSRELYPVVHIGD-ESWRMMTTDMASV-PVSVIGEEVADLSHRE
      40      50      60      70      80

      310      320      330      340      350      360
CP4_EP STLKGVTVPEDRAPSMIDEYPILAVAAFAEGATVMNGLEELRVKESDRLSAVANGLKLN
      . : . : .
gi|625 NDIKNAINLMFWGI
    90      100

>>gi|52854782|gb|AAU88262.1| cytolethal distending toxin (258 aa)
  initn: 46 init1: 46 opt: 78 Z-score: 98.1 bits: 26.6 E(): 9
Smith-Waterman score: 78; 36.170% identity (48.936% similar) in 47 aa
overlap (132-176:34-80)
```

```
      110      120      130      140      150      160
CP4_EP GCRLTMGLVGVYDFDSTFFIGDASLTKRPMGRVLNPLREMGVQVKSSEDGDRLPVTLRGPKT
      . : . : : : . : . : : :
gi|528 KRTSIFIAGVLIPILLNGCSSGKNKAYLDPKVFPQVEGGPTIPSPDEPGLPLPGPGPAL
      10      20      30      40      50      60

      170      180      190      200      210
CP4_EP PT--PITYRVPMASQVKSAYLLAGLNTPGITTVIEPIMTRDHTKMLQGFANLTVETD
      : : : : : : : : : : : : : : : : : : : : : :
gi|528 PTNAPIPIPVPGTAPAVSLMNMDSVLTMWSRGAGSSSLWAYIISDSNSFGELRNWQIMPG
      70      80      90      100      110      120
```

455 residues in 1 query sequences
1891534 residues in 7651 library sequences
Scomplib [34t26]
start: Thu Jan 22 17:47:49 2009 done: Thu Jan 22 17:47:50 2009
Total Scan time: 0.780 Total Display time: 0.090

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