CONFIDENTIAL ATTACHMENT

Study Title

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

Authors

Haidi Tu
Andre Silvanovich, Ph.D.

Study Completed On

May 1, 2009

Sponsor and Performing Laboratory

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

Laboratory Project ID

MSL0021929
Study Number: REG-09-088

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
23-26, 30-83	Discloses manufacturing or quality control processes	10(d)

Table of Contents

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

Confidential Attachment **Page(s)**

Figures

Table

Appendices

Deleted pages are attached immediately following this page.

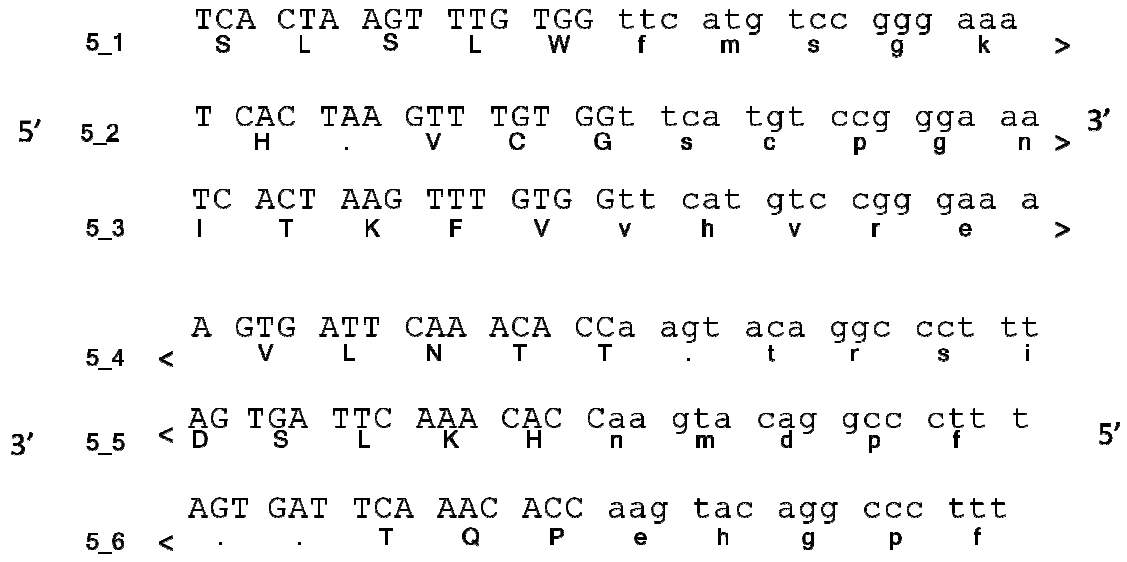

Figure 1. Reading frame alignment and DNA sequence at the 5′ junction of the MON 87705 insert.

Upper case characters refer to the genomic DNA and lower case characters refer to the inserted DNA found in MON 87705. Uppercase amino acids are translated from genomic sequence and lowercase amino acids are translated from the inserted DNA. The carat (> or <) points towards the carboxyl terminal of each polypeptide. Stop codons are denoted as periods (.).

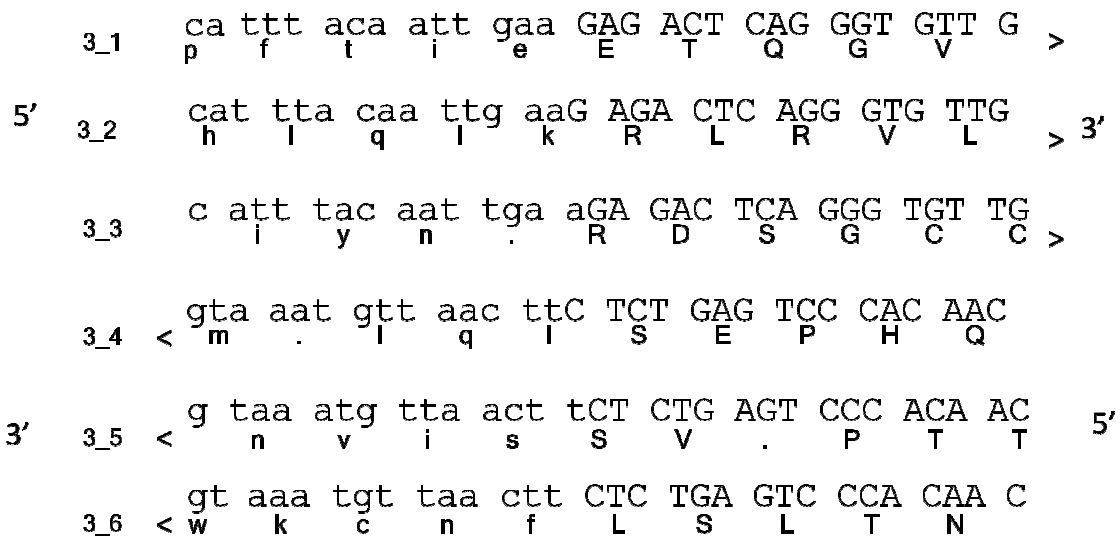

Figure 2. Reading frame alignment and DNA sequence at the 3′ junction of the MON 87705 insert.

Upper case characters refer to the genomic DNA and lower case characters refer to the inserted DNA found in MON 87705. Uppercase amino acids are translated from genomic sequence and lowercase amino acids are translated from the inserted DNA. The carat (> or <) points towards the carboxyl terminal of each polypeptide. Stop codons are denoted as periods (.).

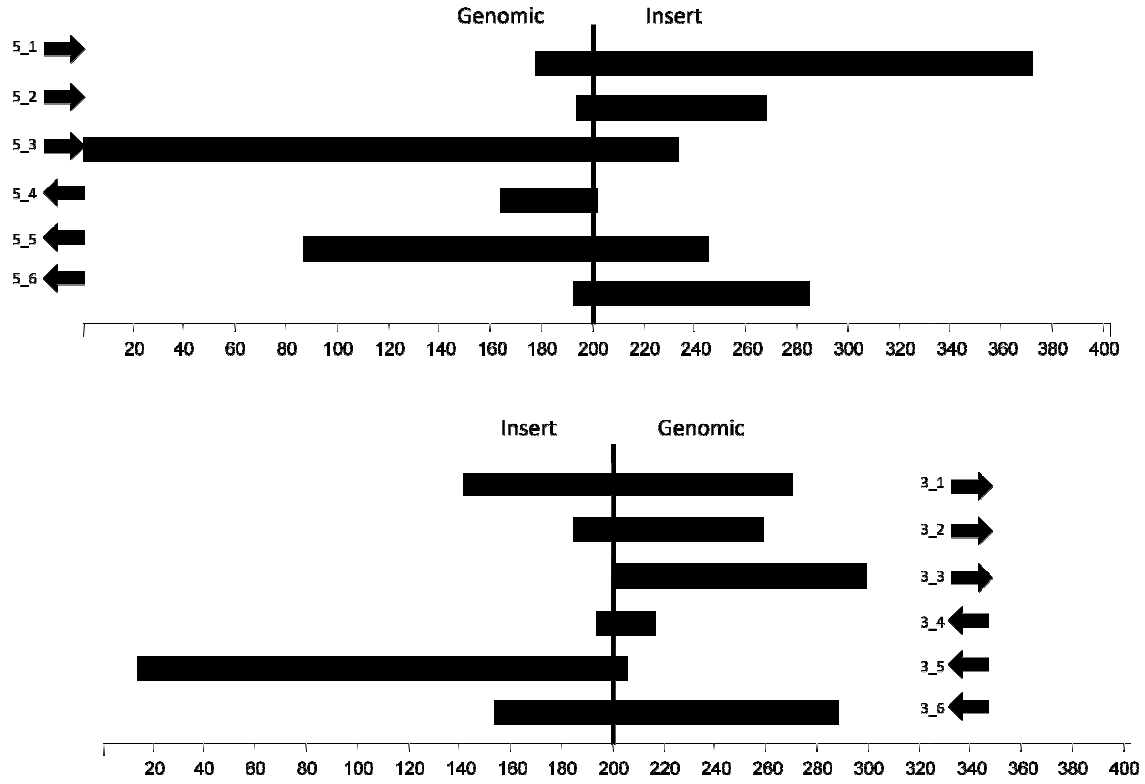

Figure 3. Graphic mapping of the flanking DNA sequences and putative polypeptides encoded by each reading frame at the 5′ and 3′ junctions of the MON 87705 insert.

The putative polypeptide coding sequences are mapped relative to the DNA sequence shown in Figures 1and 2, and the amino acid sequences are tabulated in Table 1. The scale at the bottom of each map refers to nucleotides. The arrow for each ORF points in the direction of the C-terminus.

Putative peptide ID	Putative peptide amino sequence
5_1	REKSLSLWfm sgkstwisne ydgqygeker vitnffsiqk crcpqryykm kvhfdkttny dpsyl
5_2	VCGscpgnlh gsamsmmvnm ekkke
5_3	ISELQLNQQY WVAKLLGYEF DIVYKVGASN KVVDALSRRD EDKELQGISR PFWKDITKIN EEVQKDPALA KIREELKDNL DSHPQYTLEC DILYFRGRLV LLASSLWIPK LLQEFQTSLM GGHSGIYITY RRITQSLYWI PIKGEITKFV vhvreiymdq q
5_4	TTNLVISPFI GIQ
5_5	psysllihvd fpdmnHKLSD FSLYWYPIKR LSDSSISYVN TRVPSHKRSL EFL
5_6	kkignysfff siltiiliad pcrfpghePQ T
3_1	psysllihvd fpdmkpftie ETQGVVITAV WPLGQGTVVL KKI
3_2	shlqlkRLRV LLSLRFGLWA KAPLS
3_3	RDSGCCYHCG LAFGPRHRCP EKNMRVVILA KDE
3_4	QHPESlql
3_5	VSsivngfms gkstwisney dgqygekerv itnffsiqkc rcpqryykmk vhfdkttnyd psyl
3_6	RVLQLSYFFQ DNGALAQRPN RSDNNTLSLf nckwlhvrei ymdqq

Table 1. The predicted sequence of polypeptides encoded by each reading frame at the 5′ and 3′ junctions of the MON 87705 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. Lowercase characters refer to sequence encoded by the insert DNA.

Appendix 1. Bioinformatic analysis of polypeptide 5_1

```
>5_1
REKSLSLWFMSGKSTWISNEYDGQYGEKERVITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
```

```
Sliding 8 amino acid window search
Database searched = AD_2009
Query = 5_1

Start time: Thu Mar 19 12:11:51 CDT 2009 Finish time: Thu Mar 19 12:11:51 CDT
2009

No 8 amino acid matches exist between 5_1 and the AD_2009 database
```

```
# fasta34 5_1.pep /home/ht/db/AD_2009 -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, 65 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20     9     0:===
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     0     2:*
  32     3     7:= *
  34     5    20:==    *
  36    20    41:=======      *
  38    53    68:==================    *
  40   132    95:==============================*============
  42   132   116:=========================================*=====
  44   117   128:======================================    *
  46   111   130:====================================      *
  48    98   125:================================        *
  50   114   114:=======================================*
  52   110   100:================================*===
  54   125    85:============================*=============
  56    53    71:==================     *
  58    53    59:================== *
  60    56    47:===============*===
  62    24    38:========    *
  64    27    30:=========*
  66    25    24:=======*=
  68    28    19:======*===
  70    17    15:====*=
  72    14    12:===*=
  74    11     9:==*=
  76    13     7:==*==
  78    14     5:=*===
  80     5     4:=*
  82     4     3:*=
  84     1     3:*
  86     3     2:*
  88     8     2:*==        inset = represents 1 library sequences
  90     0     1:*
  92     0     1:*         :*
  94     0     1:*         :*
  96     0     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))= 4.59450.00396; mu= 3.8330 0.205
 mean_var=51.692814.654, 0's: 9 Z-trim: 9  B-trim: 47 in 1/43
 Lambda= 0.178385
 Kolmogorov-Smirnov  statistic: 0.0444 (N=29) at  48

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


65 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:11:51 2009 done: Thu Mar 19 12:11:51 2009
 Total Scan time:  0.040 Total Display time:  0.000

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

```
# fasta34 5_1.pep /home/ht/db/TOX_2009 -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, 65 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    65     0:=====
  22     0     0:           one = represents 14 library sequences
  24     0     0:
  26     1     0:=
  28     1     2:*
  30     7    11:*
  32    18    41:==*
  34   119   110:=======*=
  36   253   227:================*==
  38   454   375:=========================*======
  40   561   523:====================================*===
  42   577   639:============================================   *
  44   782   705:==================================================*=====
  46   684   718:=================================================  *
  48   561   687:============================================        *
  50   648   627:============================================*==
  52   536   551:====================================*
  54   343   471:=========================        *
  56   272   393:====================         *
  58   266   323:==================    *
  60   491   262:==================*=================
  62   248   210:==============*===
  64   213   167:===========*====
  66   113   132:=========*
  68    80   104:====== *
  70    73    81:=====*
  72    48    63:====*
  74    41    50:===*
  76    47    39:==*=
  78    43    30:==*=
  80    15    23:=*
  82    14    18:=*
  84    14    14:*
  86    11    11:*
  88    11     8:*          inset = represents 1 library sequences
  90    10     7:*
  92    10     5:*         :====*=====
  94     3     4:*         :===*
  96     7     3:*         :==*====
  98     0     2:*         : *
 100     0     2:*         : *
 102     0     1:*         :*
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     1     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     5     0:=         *=====
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 4.72630.000656; mu= 3.5786 0.033
 mean_var=38.4883 8.456, 0's: 65 Z-trim: 70  B-trim: 83 in 1/61
 Lambda= 0.206733
 Kolmogorov-Smirnov  statistic: 0.0420 (N=29) at  58

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(7651)
gi|209540528|gb|ACI61104.1| Streptococcal pyrogeni ( 236)   83 30.7   0.066
gi|4838430|gb|AAD30989.1|AF124500_1 exotoxin H pre ( 236)   81 30.1     0.1
gi|94541990|gb|ABF32039.1| enterotoxin [Bacterioph ( 236)   81 30.1     0.1
gi|13622161|gb|AAK33907.1| streptococcal exotoxin  ( 236)   81 30.1     0.1
gi|134272104|emb|CAM30348.1| streptococcal exotoxi ( 236)   81 30.1     0.1


>>gi|209540528|gb|ACI61104.1| Streptococcal pyrogenic ex  (236 aa)
 initn:  49 init1:  49 opt:  83  Z-score: 136.4  bits: 30.7 E(): 0.066
Smith-Waterman score: 83;  30.612% identity (63.265% similar) in 49 aa
overlap (1-49:76-117)

                                              10        20        30
5_1                                       REKSLSLWFMSGKSTWISNEYDGQYGEKER
                                          ..:.::..:   .. :::.:.     .::
gi|209 LYKHDSNLIEADSIKNSPDIVTSHMLKYSVKDKNLSVFF---EKDWISQEFK----DKEV
          50        60        70        80           90

              40        50        60
5_1    VITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
        :  . . . :.:: . :.
gi|209 DIYALSAQEACECPGKRYEAFGGITLTNSEKKEIKVPINVWDKSKQQPPMFITVNKPKVT
      100       110       120       130       140       150

>>gi|4838430|gb|AAD30989.1|AF124500_1 exotoxin H precurs  (236 aa)
 initn:  47 init1:  47 opt:  81  Z-score: 133.2  bits: 30.1 E():  0.1
Smith-Waterman score: 81;  30.612% identity (63.265% similar) in 49 aa
overlap (1-49:76-117)

                                              10        20        30
5_1                                       REKSLSLWFMSGKSTWISNEYDGQYGEKER
```

```
                                    ..:.::..:   .. :::.:.     .::
gi|483 LYKHDSNLIEADSIKNSPDIVTSHMLKYSVKDKNLSVFF---EKDWISQEFK----DKEV
          50        60        70        80           90

               40        50        60
5_1    VITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
        :  . . . :.:: . :.
gi|483 DIYALSAQEVCECPGKRYEAFGGITLTNSEKKEIKVPVNVWDKSKQQPPMFITVNKPKVT
      100       110       120       130       140       150

>>gi|94541990|gb|ABF32039.1| enterotoxin [Bacteriophage   (236 aa)
 initn:  47 init1:  47 opt:  81  Z-score: 133.2  bits: 30.1 E():  0.1
Smith-Waterman score: 81;  30.612% identity (63.265% similar) in 49 aa
overlap (1-49:76-117)

                                         10        20        30
5_1                                 REKSLSLWFMSGKSTWISNEYDGQYGEKER
                                    ..:.::..:   .. :::.:.     .::
gi|945 LYKHDSNLIEADSIKNSPDIVTSHMLKYSVKDKNLSVFF---EKDWISQEFK----DKEV
          50        60        70        80           90

               40        50        60
5_1    VITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
        :  . . . :.:: . :.
gi|945 DIYALSAQEVCECPGKRYEAFGGITLTNSEKKEIKVPINVWDKSKQQPPMFITVNKPKVT
      100       110       120       130       140       150

>>gi|13622161|gb|AAK33907.1| streptococcal exotoxin H pr  (236 aa)
 initn:  47 init1:  47 opt:  81  Z-score: 133.2  bits: 30.1 E():  0.1
Smith-Waterman score: 81;  30.612% identity (63.265% similar) in 49 aa
overlap (1-49:76-117)

                                         10        20        30
5_1                                 REKSLSLWFMSGKSTWISNEYDGQYGEKER
                                    ..:.::..:   .. :::.:.     .::
gi|136 LYKHDSNLIEADSIKNSPDIVTSHMLKYSVKDKNLSVFF---EKDWISQEFK----DKEV
          50        60        70        80           90

               40        50        60
5_1    VITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
        :  . . . :.:: . :.
gi|136 DIYALSAQEVCECPGKRYEAFGGITLTNSEKKEIKVPVNVWDKSKQQPPMFITVNKPKVT
      100       110       120       130       140       150

>>gi|134272104|emb|CAM30348.1| streptococcal exotoxin H   (236 aa)
 initn:  47 init1:  47 opt:  81  Z-score: 133.2  bits: 30.1 E():  0.1
Smith-Waterman score: 81;  30.612% identity (63.265% similar) in 49 aa
overlap (1-49:76-117)

                                         10        20        30
```

```
5_1                                 REKSLSLWFMSGKSTWISNEYDGQYGEKER
                                    ..:.::..:   .. :::.:.     .::
gi|134 LYKHDSNLIEADSIKNSPDIVTSHMLKYSVKDKNLSVFF---EKDWISQEFK----DKEV
          50        60        70        80           90

               40        50        60
5_1    VITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
        :  . . . :.:: . :.
gi|134 DIYALSAQEVCECPGKRYEAFGGITLTNSEKKEIKVPVNVWDKSKQQPPMFITVNKPKVT
      100       110       120       130       140       150



65 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:11:51 2009 done: Thu Mar 19 12:11:51 2009
 Total Scan time:  0.200 Total Display time:  0.010

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

```
# fasta34 5_1.pep /home/ht/db/PRT_2009 -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

 5_1, 65 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 269868     0:============
  22   103     0:=          one = represents 22718 library sequences
  24   213    14:*
  26   423   309:*
  28  1533  3337:*
  30  7232 20272:*
  32 36242 78383:== *
  34 134809 212566:======   *
  36 368480 436561:=================  *
  38 685327 721472:================================*
  40 1077526 1006390:==============================================*===
  42 1311003
1230190:=========================================================*===
  44 1363029
1357014:============================================================*
  46 1319398
1382152:============================================================*
```

```
  48 1259093 1323251:=========================================================
*
  50 1122167 1207472:================================================    *
  52 997442 1061570:============================================  *
  54 842779 906767:=================================== *
  56 732741 757428:================================*
  58 624242 621834:============================*
  60 521060 503722:======================*
  62 427021 403836:=================*=
  64 347736 321168:==============*=
  66 272929 253842:===========*=
  68 216197 199667:========*=
  70 172448 156470:======*=
  72 132265 122267:=====*
  74 110561 95327:====*
  76 87509 74195:===*
  78 64715 57671:==*
  80 50504 44781:=*=
  82 39046 34256:=*
  84 29199 27135:=*
  86 22071 20996:*
  88 16763 16245:*          inset = represents 188 library sequences
  90 13197 12570:*
  92  9388  9726:*         :=======================================*
  94  7250  7525:*         :=======================================*
  96  5793  5823:*         :==============================*
  98  4089  4505:*         :====================== *
 100  2990  3486:*         :================  *
 102  2047  2697:*         :===========   *
 104  1515  2087:*         :==========  *
 106  1106  1615:*         :======  *
 108  1024  1249:*         :======*
 110   694   967:*         :==== *
 112   447   748:*         :===*
 114   398   579:*         :===*
 116   290   448:*         :==*
 118   157   347:*         :=*
>120   359   268:*         :=*
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714116 sequences
  Expectation_n fit: rho(ln(x))= 4.01400.000189; mu= 7.0975 0.010
 mean_var=45.5022 9.124, 0's: 1111 Z-trim: 1113  B-trim: 0 in 0/64
 Lambda= 0.190133
 Kolmogorov-Smirnov  statistic: 0.0308 (N=29) at  56

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

65 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:11:52 2009 done: Thu Mar 19 12:19:09 2009
 Total Scan time: 420.730 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
VCGSCPGNLHGSAMSMMVNMEKKKE
_____________________________________________________


Sliding 8 amino acid window search
Database searched = AD_2009
Query = 5_2

Start time: Thu Mar 19 12:19:11 CDT 2009 Finish time: Thu Mar 19 12:19:11 CDT
2009

No 8 amino acid matches exist between 5_2 and the AD_2009 database
_____________________________________________________


# fasta34 5_2.pep /home/ht/db/AD_2009 -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, 25 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20     3     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     4     2:*=
  32     9     7:==*
  34    19    20:======*
  36    28    41:==========   *
  38    81    68:=====================*====
  40    77    95:=========================     *
  42   115   116:======================================*
  44   117   128:=========================================   *
  46   132   130:============================================*
  48   112   125:=======================================   *
```

```
  50   104   114:==================================  *
  52   118   100:==================================*======
  54    72    85:========================    *
  56    64    71:====================== *
  58    58    59:==================*
  60    47    47:===============*
  62    50    38:============*====
  64    49    30:=========*=======
  66    23    24:=======*
  68    10    19:====  *
  70    14    15:====*
  72    49    12:===*=============
  74    21     9:==*====
  76     6     7:==*
  78     0     5: *
  80     1     4:=*
  82     1     3:*
  84     0     3:*
  86     0     2:*
  88     0     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     2     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:          *
 307888 residues in  1386 sequences
  Expectation_n fit: rho(ln(x))= 2.97520.0032; mu= 10.1836 0.166
 mean_var=29.0576 7.017, 0's: 3 Z-trim: 3  B-trim: 0 in 0/44
 Lambda= 0.237927
 Kolmogorov-Smirnov  statistic: 0.0373 (N=25) at  60

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


25 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:19:10 2009 done: Thu Mar 19 12:19:11 2009
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 5_2.pep /home/ht/db/TOX_2009 -Q -E 1 -O 5_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

 5_2, 25 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    59     0:=====
  22     1     0:=          one = represents 13 library sequences
  24     2     0:=
  26    15     0:==
  28     4     2:*
  30    32    11:*==
  32    99    41:===*====
  34   242   110:========*==========
  36   131   227:===========      *
  38   235   375:==================         *
  40   565   523:=========================================*===
  42   513   639:=========================================          *
  44   747   705:============================================================*===
  46   703   718:============================================================*
  48   666   688:==========================================================*
  50   463   627:====================================           *
  52   570   552:============================================*=
  54   728   471:====================================*==================
  56   281   394:======================         *
  58   458   323:=========================*===========
  60   269   262:====================*
  62   135   210:===========      *
  64   129   167:==========  *
  66   107   132:========= *
  68   118   104:=======*==
  70    64    81:===== *
  72    66    64:====*=
  74    67    50:===*==
  76    19    39:==*
  78    40    30:==*=
  80    66    23:=*====
  82    13    18:=*
  84    12    14:=*
  86     4    11:*
```

```
  88     7     8:*          inset = represents 1 library sequences
  90     3     7:*
  92     1     5:*         :=   *
  94     2     4:*         :== *
  96     4     3:*         :==*=
  98     1     2:*         :=*
 100     0     2:*         : *
 102     0     1:*         :*
 104     4     1:*         :*===
 106     0     1:*         :*
 108     1     1:*         :*
 110     0     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))= 1.81860.000521; mu= 15.9053 0.027
 mean_var=25.3868 5.993, 0's: 59 Z-trim: 59  B-trim: 74 in 1/61
 Lambda= 0.254549
 Kolmogorov-Smirnov  statistic: 0.0327 (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


25 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:19:11 2009 done: Thu Mar 19 12:19:11 2009
 Total Scan time:  0.160 Total Display time:  0.000

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


# fasta34 5_2.pep /home/ht/db/PRT_2009 -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, 25 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 218285     0:==========
  22   458     0:=          one = represents 22951 library sequences
  24   875    14:*
  26  3024   309:*
  28 10440  3337:*
  30 32534 20271:*=
  32 95686 78382:===*=
  34 214915 212562:=========*
  36 408437 436553:================== *
  38 668534 721460:============================ *
  40 921977 1006373:=======================================  *
  42 1136538 1230168:================================================   *
  44 1281171 1356990:============================================================
*
  46 1377060
1382128:============================================================*
  48 1342019
1323229:==========================================================*=
  50 1207708 1207451:==========================================================*
  52 1084398 1061551:==============================================*=
  54 926521 906751:=======================================*=
  56 734974 757415:================================*
  58 608095 621823:===========================*
  60 486716 503713:====================*
  62 412117 403829:=================*
  64 318606 321163:=============*
  66 261136 253837:===========*
  68 201476 199663:========*
  70 169331 156468:======*=
  72 141727 122265:=====*=
  74 105412 95326:====*
  76 82258 74194:===*
  78 60173 57670:==*
  80 46273 44780:=*=
  82 36686 34256:=*
  84 28600 27135:=*
  86 21938 20995:*
  88 16076 16245:*          inset = represents 200 library sequences
  90 13477 12570:*
  92  9977  9726:*         :=======================================*
  94  7173  7525:*         :====================================== *
  96  5164  5823:*         :=========================   *
  98  3943  4505:*         :====================  *
 100  2537  3486:*         :=============    *
 102  2008  2697:*         :===========  *
 104  1862  2087:*         :==========*
 106  1797  1615:*         :========*
 108  1191  1249:*         :=======*
 110   873   967:*         :=====*
 112   539   748:*         :===*
 114   441   579:*         :==*
 116   441   448:*         :==*
 118   205   347:*         :=*
```

```
>120   616   268:*         :=*==
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14713862 sequences
  Expectation_n fit: rho(ln(x))= 3.66060.000174; mu= 6.8184 0.009
 mean_var=26.6773 5.283, 0's: 922 Z-trim: 925  B-trim: 0 in 0/63
 Lambda= 0.248315
 Kolmogorov-Smirnov  statistic: 0.0205 (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


25 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:19:11 2009 done: Thu Mar 19 12:25:37 2009
 Total Scan time: 366.520 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
ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELK
DNLDSHPQYTLECDILYFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
EIYMDQQ
```

```
Sliding 8 amino acid window search
Database searched = AD_2009
Query = 5_3

Start time: Thu Mar 19 12:25:39 CDT 2009 Finish time: Thu Mar 19 12:25:40 CDT
2009

No 8 amino acid matches exist between 5_3 and the AD_2009 database
```

```
# fasta34 5_3.pep /home/ht/db/AD_2009 -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, 161 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
```

```
< 20     3     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     1     0:=
  30     1     2:*
  32     3     7:= *
  34     5    20:==    *
  36    51    41:=============*===
  38    75    68:======================*==
  40    86    95:============================  *
  42   106   116:===================================  *
  44   128   128:============================================*
  46    93   130:==============================           *
  48   109   125:====================================    *
  50   129   114:====================================*=====
  52   114   100:==================================*====
  54    73    85:=========================   *
  56    68    71:======================*
  58    65    59:====================*==
  60    54    47:================*==
  62    57    38:=============*======
  64    41    30:=========*====
  66    34    24:=======*====
  68    23    19:======*=
  70    14    15:====*
  72    14    12:===*=
  74     7     9:==*
  76     5     7:==*
  78     5     5:=*
  80     2     4:=*
  82     2     3:*
  84     1     3:*
  86     5     2:*=
  88     0     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     1     1:*         :*
  94     1     1:*         :*
  96     2     1:*         :*=
  98     2     0:=         *==
 100     0     0:          *
 102     6     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))= 5.36650.00416; mu= 3.9355 0.215
 mean_var=58.522515.847, 0's: 3 Z-trim: 3  B-trim: 26 in 1/43
 Lambda= 0.167654
 Kolmogorov-Smirnov  statistic: 0.0536 (N=28) at  48

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(1386)
gi|232054|sp|P30575.1|ENO1_CANAL RecName: Full=Eno ( 440)   77 26.7    0.89
gi|2114497|gb|AAB58417.1| 30 kDa salivary gland al ( 253)   74 25.9    0.89
gi|94468552|gb|ABF18125.1| 30 kDa salivary gland a ( 258)   74 25.9    0.91
gi|56417506|gb|AAV90694.1| GE-rich salivary protei ( 266)   74 25.9    0.94
gi|56417504|gb|AAV90693.1| 30 kDa salivary gland a ( 271)   74 25.9    0.95
gi|94468546|gb|ABF18122.1| 30 kDa salivary gland a ( 273)   74 25.9    0.96


>>gi|232054|sp|P30575.1|ENO1_CANAL RecName: Full=Enolase  (440 aa)
 initn:  51 init1:  51 opt:  77  Z-score: 102.8  bits: 26.7 E(): 0.89
Smith-Waterman score: 77;  30.508% identity (57.627% similar) in 59 aa
overlap (20-78:32-87)

                        10        20        30        40
5_3               ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGIS
                                  :  .   :::. : .::  :: ::  . ..
gi|232 SYATKIHARYVYDSRGNPTVEVDFTTDKGLFRSIVPSGASTGVHEALELRDGDKS-KWLG
             10        20        30        40        50         60

     50        60        70        80        90       100
5_3    RPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRGRLVLLASSLWIP
       .   : ....:. .   ::: : . .. :
gi|232 KGVLKAVANVNDIIA--PALIKAKIDVVDQAKIDEFLLSLDGTPNKSKLGANAILGVSLA
              70          80        90       100       110

>>gi|2114497|gb|AAB58417.1| 30 kDa salivary gland allerg  (253 aa)
 initn:  66 init1:  66 opt:  74  Z-score: 102.8  bits: 25.9 E(): 0.89
Smith-Waterman score: 74;  32.075% identity (64.151% similar) in 53 aa
overlap (32-82:177-229)

             10        20        30        40        50         60
5_3    SELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGISRPFWKDITK-IN
                                 ::.:..:  . ...::  . . ::. : :.
gi|211 LDKDTKVDHIQSEYLRSALNNDLQSEVRVPVVEAIGRIGDYSKIQGCFKSMGKDVKKVIS
        150       160       170       180       190       200

               70        80        90       100       110
5_3    EEVQK-DPALAKIREELKDNLDSHPQYTLECDILYFRGRLVLLASSLWIPKLLQEFQTSL
       :: .:    ..: . : . . ::
gi|211 EEEKKFKSCMSKKKSEYQCSEDSFAAAKSKLSPITSKIKSCVSSKGR

        210       220       230       240       250

>>gi|94468552|gb|ABF18125.1| 30 kDa salivary gland aller  (258 aa)
 initn:  66 init1:  66 opt:  74  Z-score: 102.6  bits: 25.9 E(): 0.91
Smith-Waterman score: 74;  32.075% identity (64.151% similar) in 53 aa
overlap (32-82:182-234)

             10        20        30        40        50         60
5_3    SELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGISRPFWKDITK-IN
                                 ::.:..:  . ...::  . . ::. : :.
gi|944 LDKDTKVDHIQSEYLRSALNNDLQSEVRVPVVEAIGRIGDYSKIQGCFKSMGKDVKKVIS
             160       170       180       190       200       210

               70        80        90       100       110
5_3    EEVQK-DPALAKIREELKDNLDSHPQYTLECDILYFRGRLVLLASSLWIPKLLQEFQTSL
       :: .:    ..: . : . . ::
gi|944 EEEKKFKSCMSKKKSEYQCSEDSFAAAKSKLSPITSKIKSCVSSKGR
              220       230       240       250

>>gi|56417506|gb|AAV90694.1| GE-rich salivary protein 30  (266 aa)
 initn:  70 init1:  70 opt:  74  Z-score: 102.4  bits: 25.9 E(): 0.94
Smith-Waterman score: 74;  29.091% identity (60.000% similar) in 55 aa
overlap (30-82:188-242)

               10        20        30        40        50
5_3     ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGISRPFWKDITKI
                                    : ::.:.::    ....:  . . .:. :.
gi|564 AILDKDTKVDNIQSEYLRSALNNDLQSEVRNPVVEAISRLGSFSKIEGCFKSMGSDVKKV
       160       170       180       190       200       210

      60          70        80        90       100       110
5_3    NEEVQK--DPALAKIREELKDNLDSHPQYTLECDILYFRGRLVLLASSLWIPKLLQEFQT
        .: ::     ..: . : . . ::
gi|564 IDEEQKAFKDCMTKKKSEYECSEDSFASAKGKLSPITSKIKSCVSSKGQ
       220       230       240       250       260

>>gi|56417504|gb|AAV90693.1| 30 kDa salivary gland aller  (271 aa)
 initn:  70 init1:  70 opt:  74  Z-score: 102.3  bits: 25.9 E(): 0.95
Smith-Waterman score: 74;  29.091% identity (60.000% similar) in 55 aa
overlap (30-82:193-247)

               10        20        30        40        50
5_3     ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGISRPFWKDITKI
                                    : ::.:.::    ....:  . . .:. :.
gi|564 AILDKDTKVDNIQSEYLRSALNNDLQSEVRNPVVEAISRLGSFSKIEGCFKSMGSDVKKV
            170       180       190       200       210       220

      60          70        80        90       100       110
5_3    NEEVQK--DPALAKIREELKDNLDSHPQYTLECDILYFRGRLVLLASSLWIPKLLQEFQT
        .: ::     ..: . : . . ::
```

```
gi|564 IDEEQKAFKDCMTKKKSEYECSEDSFASAKGKLSPITSKIKSCVSSKGQ
            230       240       250       260       270

>>gi|94468546|gb|ABF18122.1| 30 kDa salivary gland aller  (273 aa)
 initn:  66 init1:  66 opt:  74  Z-score: 102.2  bits: 25.9 E(): 0.96
Smith-Waterman score: 74;  32.075% identity (64.151% similar) in 53 aa
overlap (32-82:197-249)

             10        20        30        40        50         60
5_3    SELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGISRPFWKDITK-IN
                                    ::.:..:  . ...::  . . ::. : :.
gi|944 LDKDTKVDHIQSEYLRSALNNDLQSEVRVPVVEAIGRIGDYSKIQGCFKSMGKDVKKVIS
        170       180       190       200       210       220

                70        80        90       100       110
5_3    EEVQK-DPALAKIREELKDNLDSHPQYTLECDILYFRGRLVLLASSLWIPKLLQEFQTSL
       :: .:    ..: . : . . ::
gi|944 EEEKKFKSCMSKKKSEYQCSEDSFAAAKSKLSPITSKIKSCVSSKGR
        230       240       250       260       270




161 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:25:38 2009 done: Thu Mar 19 12:25:39 2009
 Total Scan time:  0.100 Total Display time:  0.010

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

```
# fasta34 5_3.pep /home/ht/db/TOX_2009 -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, 161 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    64     0:======
  22     2     0:=          one = represents 12 library sequences
  24     5     0:=
  26     5     0:=
  28    32     2:*==
  30    43    11:*===
  32    52    41:===*=
  34   115   110:==========*
  36   209   227:==================*
  38   494   375:==============================*==========
  40   518   523:=========================================*
  42   548   639:============================================       *
  44   664   705:=============================================================  *
  46   625   718:=========================================================      *
  48   602   688:=================================================== *
  50   542   627:============================================      *
  52   491   552:=========================================    *
  54   651   471:=======================================*===============
  56   293   394:========================       *
  58   242   323:====================     *
  60   445   262:====================*================
  62   326   210:=================*==========
  64   172   167:============*=
  66   138   132:==========*=
  68    71   104:======  *
  70    58    81:===== *
  72    47    64:==== *
  74    52    50:====*
  76    38    39:===*
  78    15    30:==*
  80    16    23:=*
  82    16    18:=*
  84     9    14:=*
  86     5    11:*
  88     9     8:*          inset = represents 1 library sequences
  90    11     7:*
  92     7     5:*         :====*==
  94     7     4:*         :===*===
  96     0     3:*         :  *
  98     1     2:*         :=*
 100     0     2:*         : *
 102     1     1:*         :*
 104     1     1:*         :*
 106     0     1:*         :*
 108     3     1:*         :*==
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     1     0:=         *=
>120     0     0:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 6.02050.000661; mu= 1.1038 0.033
 mean_var=43.2302 8.995, 0's: 59 Z-trim: 64  B-trim: 113 in 1/61
 Lambda= 0.195065
 Kolmogorov-Smirnov  statistic: 0.0359 (N=29) at  58

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(7651)
gi|163664913|gb|ABY32280.1| addiction module toxin (  93)   73 27.3    0.71


>>gi|163664913|gb|ABY32280.1| addiction module toxin, Re  (93 aa)
 initn:  34 init1:  34 opt:  73  Z-score: 117.8  bits: 27.3 E(): 0.71
Smith-Waterman score: 73;  28.571% identity (57.143% similar) in 63 aa
overlap (38-96:11-73)

        10        20        30        40        50         60
5_3    QQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGISRP-FWKDITKINEEVQKD
                                    .::  .:..:  :  .  :.: .  :.  :
gi|163                       MRTIERTTQFKRDYKREMKGRHRASLQADLTAVLTELAAD
                                     10        20        30        40

          70         80        90        100       110       120
5_3    -PALAKIREE-LKDNLDSHPQYTLECD-ILYFRGRLVLLASSLWIPKLLQEFQTSLMGGH
        :  :..:.. :  :  .: .  .. : .: .:
gi|163 RPLAARLRDHALTGNWADHRDCHVKPDLVLIYRLAGTETLQLVRLGSHAELGF
               50        60        70        80        90

           130       140       150       160
5_3    SGIYITYRRITQSLYWIPIKGEITKFVVHVREIYMDQQ




161 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:25:40 2009 done: Thu Mar 19 12:25:40 2009
 Total Scan time:  0.530 Total Display time:  0.000

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


# fasta34 5_3.pep /home/ht/db/PRT_2009 -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, 161 aa
 vs /home/ht/db/PRT_2009 library


       opt      E()
< 20 226248     0:==========
  22   304     0:=          one = represents 22881 library sequences
  24   385    14:*
  26  1092   309:*
  28  3540  3337:*
  30 17664 20271:*
  32 77861 78380:===*
  34 239826 212558:=========*=
  36 529432 436544:==================*====
  38 826908 721445:==============================*=====
  40 1104627 1006352:============================================*=====
  42 1260251 1230143:=================================================*==
  44 1341112
1356962:=============================================================*
  46 1372833
1382100:=============================================================*
  48 1263224 1323201:=============================================================
*
  50 1151844 1207426:================================================= *
  52 988720 1061529:============================================  *
  54 830563 906732:====================================  *
  56 685839 757399:==============================   *
  58 569681 621810:=========================  *
  60 465132 503703:==================== *
  62 379697 403820:=================*
  64 295708 321156:============ *
  66 231451 253832:===========*
  68 189646 199659:========*
  70 152884 156464:======*
  72 119032 122262:=====*
  74 90822 95324:====*
  76 67840 74193:===*
  78 55687 57669:==*
  80 41581 44779:=*
  82 32189 34255:=*
  84 24620 27134:=*
  86 18272 20995:*
  88 14034 16245:*          inset = represents 163 library sequences
  90 10120 12569:*
  92  8137  9726:*         :=======================================*
  94  6447  7525:*         :=======================================*
  96  4970  5823:*         :============================    *
  98  3849  4505:*         :=========================   *
 100  2760  3486:*         :=================    *
 102  2182  2697:*         :==============  *
 104  1305  2087:*         :==========   *
 106   940  1615:*         :======   *
 108   713  1249:*         :=====  *
 110   511   967:*         :==== *
 112   374   748:*         :=== *
 114   280   579:*         :== *
 116   205   448:*         :==*
 118   149   346:*         := *
```

```
>120   927   268:*         :=*====
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14713555 sequences
  Expectation_n fit: rho(ln(x))= 5.18760.000191; mu= 6.3487 0.010
 mean_var=63.556813.192, 0's: 977 Z-trim: 981  B-trim: 2132 in 1/63
 Lambda= 0.160877
 Kolmogorov-Smirnov  statistic: 0.0243 (N=29) at  42

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(14717352)
gi|124359710|gb|ABN06064.1| RNA-directed DNA polym (1297)  437 109.8 2.7e-21
gi|147807720|emb|CAN66553.1| hypothetical protein  (1448)  422 106.4 3.3e-20
gi|147854459|emb|CAN78588.1| hypothetical protein  (2232)  408 103.2 4.5e-19
gi|124360619|gb|ABD33394.2| FAR1; Polynucleotidyl  ( 657)  318 82.0 3.1e-13
gi|6466937|gb|AAF13073.1|AC011621_1 putative retro (1661)  319 82.5 5.8e-13
gi|208609060|dbj|BAG72152.1| hypothetical protein  (1369)  315 81.5 9.3e-13
gi|208609055|dbj|BAG72150.1| hypothetical protein  (1558)  315 81.6   1e-12
gi|208609057|dbj|BAG72151.1| hypothetical protein  (1558)  315 81.6   1e-12
gi|208609065|dbj|BAG72154.1| hypothetical protein  (1558)  315 81.6   1e-12
gi|208609062|dbj|BAG72153.1| hypothetical protein  (1558)  314 81.3 1.2e-12
gi|208609051|dbj|BAG72148.1| hypothetical protein  (1558)  314 81.3 1.2e-12
gi|208609053|dbj|BAG72149.1| hypothetical protein  (1520)  312 80.9 1.6e-12
gi|208609049|dbj|BAG72147.1| hypothetical protein  (1520)  312 80.9 1.6e-12
gi|147828709|emb|CAN66228.1| hypothetical protein  (1258)  308 79.9 2.7e-12
gi|147843077|emb|CAN83300.1| hypothetical protein  (1366)  302 78.5 7.5e-12
gi|147774273|emb|CAN76793.1| hypothetical protein  (1469)  301 78.3 9.4e-12
gi|147772855|emb|CAN73669.1| hypothetical protein  (1308)  299 77.8 1.2e-11
gi|147789424|emb|CAN66607.1| hypothetical protein  (2822)  302 78.7 1.4e-11
gi|8778789|gb|AAF79797.1|AC020646_20 T32E20.30 [Ar (1397)  297 77.4 1.7e-11
gi|170660047|gb|ACB28472.1| polyprotein [Ananas co ( 953)  292 76.1 2.8e-11
gi|147823217|emb|CAN75278.1| hypothetical protein  ( 716)  285 74.4 6.8e-11
gi|113595720|dbj|BAF19594.1| Os06g0493000 [Oryza s (1573)  289 75.5 6.9e-11
gi|116309032|emb|CAH66146.1| OSIGBa0114M03.4 [Oryz (1448)  285 74.6 1.2e-10
gi|155327954|gb|ABT83558.1| Sequence 171028 from p ( 161)  270 70.6 2.2e-10
gi|77551099|gb|ABA93896.1| retrotransposon protein ( 897)  278 72.8 2.5e-10
gi|77557165|gb|ABA99961.1| retrotransposon protein (1619)  279 73.2 3.5e-10
gi|147768278|emb|CAN60449.1| hypothetical protein  ( 647)  274 71.8 3.7e-10
gi|32488663|emb|CAE03590.1| OSJNBa0087O24.13 [Oryz (1311)  277 72.7 4.1e-10
gi|27764548|gb|AAO23078.1| polyprotein [Glycine ma (1552)  276 72.5 5.5e-10
gi|147786920|emb|CAN64437.1| hypothetical protein  ( 623)  271 71.1 5.7e-10
gi|108862596|gb|ABA97714.2| retrotransposon protei (1287)  272 71.5 8.9e-10
gi|147779107|emb|CAN73467.1| hypothetical protein  (1593)  272 71.6 1.1e-09
gi|215767834|dbj|BAH00063.1| unnamed protein produ ( 494)  264 69.4 1.5e-09
gi|147866223|emb|CAN81986.1| hypothetical protein  ( 672)  265 69.7 1.6e-09
gi|147864527|emb|CAN80491.1| hypothetical protein  (1412)  268 70.6 1.8e-09
gi|108864659|gb|ABA95357.2| retrotransposon protei (2811)  270 71.3 2.4e-09
gi|113645699|dbj|BAF28840.1| Os11g0686400 [Oryza s (2866)  270 71.3 2.4e-09
gi|77554607|gb|ABA97403.1| retrotransposon protein ( 562)  259 68.3 3.6e-09
gi|77550700|gb|ABA93497.1| retrotransposon protein (1448)  263 69.5 4.2e-09
gi|62734016|gb|AAX96125.1| retrotransposon protein (1448)  263 69.5 4.2e-09
gi|15451607|gb|AAK98731.1|AC090485_10 Putative ret (1461)  263 69.5 4.2e-09
gi|108706172|gb|ABF93967.1| retrotransposon protei (1461)  263 69.5 4.2e-09
gi|90399077|emb|CAJ86299.1| H0124B04.16 [Oryza sat (1265)  261 69.0 5.2e-09
gi|113611293|dbj|BAF21671.1| Os07g0510800 [Oryza s ( 517)  256 67.6 5.5e-09
gi|113639856|dbj|BAF27161.1| Os10g0551400 [Oryza s (1189)  260 68.7 5.8e-09
gi|113533835|dbj|BAF06218.1| Os01g0758700 [Oryza s ( 225)  251 66.2 6.1e-09
gi|110289541|gb|AAP54937.2| retrotransposon protei (1477)  260 68.8 6.9e-09
gi|116310096|emb|CAH67116.1| H0502G05.7 [Oryza sat ( 642)  255 67.4 7.7e-09
gi|38567865|emb|CAE03018.3| OSJNBa0091D06.3 [Oryza ( 753)  255 67.5 8.8e-09
gi|15217201|gb|AAK92545.1|AC051624_3 Putative retr (1476)  257 68.1 1.1e-08
gi|31431040|gb|AAP52878.1| retrotransposon protein (1476)  257 68.1 1.1e-08
gi|113638940|dbj|BAF26245.1| Os10g0317000 [Oryza s (1476)  257 68.1 1.1e-08
gi|155282606|gb|ABT38210.1| Sequence 125680 from p ( 218)  246 65.1 1.3e-08
gi|113645672|dbj|BAF28813.1| Os11g0677100 [Oryza s (1445)  255 67.6 1.5e-08
gi|116309424|emb|CAH66499.1| H0321H01.8 [Oryza sat (1602)  255 67.6 1.7e-08
gi|108710432|gb|ABF98227.1| retrotransposon protei (1160)  253 67.1 1.7e-08
gi|215737344|dbj|BAG96273.1| unnamed protein produ ( 459)  248 65.7 1.8e-08
gi|215695456|dbj|BAG90661.1| unnamed protein produ ( 459)  248 65.7 1.8e-08
gi|31712084|gb|AAP68389.1| putative polyprotein [O (1246)  253 67.1 1.9e-08
gi|18378611|gb|AAL68643.1|AF458767_1 polyprotein [ ( 775)  250 66.3   2e-08
gi|108707050|gb|ABF94845.1| retrotransposon protei ( 937)  250 66.3 2.4e-08
gi|38344579|emb|CAE05537.2| OSJNBa0053B21.11 [Oryz (1251)  250 66.4   3e-08
gi|147798109|emb|CAN73897.1| hypothetical protein  ( 269)  242 64.2   3e-08
gi|110289426|gb|AAP54661.2| retrotransposon protei ( 760)  247 65.6 3.2e-08
gi|77554308|gb|ABA97104.1| retrotransposon protein (1412)  250 66.4 3.3e-08
gi|52353389|gb|AAU43957.1| putative polyprotein [O (1263)  249 66.2 3.6e-08
gi|10122041|gb|AAG13430.1|AC051634_11 putative pla ( 921)  247 65.6 3.8e-08
gi|108706266|gb|ABF94061.1| retrotransposon protei (1159)  248 65.9 3.9e-08
gi|21397271|gb|AAM51835.1|AC105730_9 Putative plan (1159)  248 65.9 3.9e-08
gi|113639223|dbj|BAF26528.1| Os10g0419000 [Oryza s (1611)  249 66.2 4.4e-08
gi|31432119|gb|AAP53789.1| retrotransposon protein (1611)  249 66.2 4.4e-08
gi|18568269|gb|AAL76001.1|AF466646_9 putative gag- (2396)  251 66.8 4.4e-08
gi|38605839|emb|CAE02919.3| OSJNBb0108J11.11 [Oryz ( 815)  245 65.2 4.7e-08
gi|113563772|dbj|BAF14115.1| Os04g0191000 [Oryza s (1463)  248 66.0 4.7e-08
gi|38346036|emb|CAE01900.2| OSJNBa0059D20.8 [Oryza (1463)  248 66.0 4.7e-08
gi|167249400|gb|ABZ25960.1| Sequence 85 from paten (2437)  250 66.6 5.3e-08
gi|112064846|gb|ABH99846.1| Sequence 85 from paten (2437)  250 66.6 5.3e-08
gi|32492359|emb|CAE05990.1| OSJNBa0004L19.22 [Oryz (1586)  247 65.8 5.9e-08
gi|113564017|dbj|BAF14360.1| Os04g0316700 [Oryza s (1605)  247 65.8   6e-08
gi|31430458|gb|AAP52367.1| retrotransposon protein ( 415)  239 63.6   7e-08
gi|113536225|dbj|BAF08608.1| Os02g0323900 [Oryza s (1128)  244 65.0 7.2e-08
gi|38344156|emb|CAD41876.2| OSJNBa0041A02.23 [Oryz (1373)  245 65.3 7.3e-08
gi|38347666|emb|CAE05600.2| OSJNBa0054D14.1 [Oryza (1629)  245 65.3 8.4e-08
gi|38346992|emb|CAD40278.2| OSJNBb0062H02.17 [Oryz (1629)  245 65.3 8.4e-08
gi|22711553|gb|AAM01152.2|AC113336_4 Putative retr ( 575)  239 63.7 9.2e-08
gi|147802203|emb|CAN70510.1| hypothetical protein  ( 790)  240 64.0   1e-07
gi|38344463|emb|CAE04934.2| OSJNBa0017P10.11 [Oryz (1012)  241 64.3 1.1e-07
gi|113563897|dbj|BAF14240.1| Os04g0257000 [Oryza s (1012)  241 64.3 1.1e-07
```

gi|38347368|emb|CAE04958.2| OSJNBa0070D17.9 [Oryza (394)	236	62.9	1.1e-07
gi|147811718|emb|CAN77256.1| hypothetical protein (1365)	242	64.6	1.2e-07
gi|18378613|gb|AAL68644.1|AF458768_1 polyprotein [(933)	240	64.0	1.2e-07
gi|7800101|gb|AAF69810.1| pol protein [Picea abies (148)	230	61.3	1.3e-07
gi|38345562|emb|CAE03436.2| OSJNBa0032F06.19 [Oryz (1575)	240	64.2	1.8e-07
gi|113534177|dbj|BAF06560.1| Os01g0820300 [Oryza s (1333)	239	63.9	1.9e-07
gi|89887334|gb|ABD78322.1| polyprotein [Primula vu (1359)	239	63.9	1.9e-07
gi|10140673|gb|AAG13508.1|AC068924_13 putative gag (1608)	237	63.5	3e-07
gi|113610654|dbj|BAF21032.1| Os07g0195900 [Oryza s (1219)	235	62.9	3.3e-07
gi|32489310|emb|CAE03706.1| OSJNBa0060B20.14 [Oryz (3200)	240	64.3	3.3e-07
gi|31712076|gb|AAP68381.1| putative polyprotein [O (1118)	234	62.7	3.6e-07
gi|108710435|gb|ABF98230.1| retrotransposon protei (1118)	234	62.7	3.6e-07
gi|108864085|gb|ABA91843.2| retrotransposon protei (1411)	235	63.0	3.7e-07
gi|62733109|gb|AAX95226.1| retrotransposon protein (1513)	235	63.0	3.9e-07
gi|113578554|dbj|BAF16917.1| Os05g0241900 [Oryza s (1475)	234	62.7	4.5e-07
gi|53981172|gb|AAV24812.1| putative polyprotein [O (1475)	234	62.7	4.5e-07
gi|113611374|dbj|BAF21752.1| Os07g0528900 [Oryza s (1790)	235	63.0	4.5e-07
gi|113536180|dbj|BAF08563.1| Os02g0309600 [Oryza s (692)	230	61.6	4.6e-07
gi|77552522|gb|ABA95319.1| retrotransposon protein (955)	230	61.7	6e-07
gi|38346427|emb|CAD40214.2| OSJNBa0019J05.12 [Oryz (1817)	231	62.1	8.7e-07
gi|113610590|dbj|BAF20968.1| Os07g0184000 [Oryza s (1007)	227	61.0	1e-06
gi|77555174|gb|ABA97970.1| retrotransposon protein (1548)	227	61.1	1.5e-06
gi|108706171|gb|ABF93966.1| retrotransposon protei (1920)	228	61.4	1.5e-06
gi|15451608|gb|AAK98732.1|AC090485_11 Putative ret (1923)	228	61.4	1.5e-06
gi|147854038|emb|CAN83399.1| hypothetical protein (234)	217	58.4	1.5e-06
gi|4581164|gb|AAD24647.1| putative retroelement po (780)	223	60.0	1.6e-06
gi|12322948|gb|AAG51464.1|AC069160_10 gypsy/Ty3 el (1447)	226	60.9	1.6e-06
gi|78708171|gb|ABB47146.1| retrotransposon protein (813)	221	59.6	2.2e-06
gi|22725944|gb|AAN04954.1| Putative retroelement [(813)	221	59.6	2.2e-06
gi|62734404|gb|AAX96513.1| retrotransposon protein (605)	219	59.0	2.4e-06
gi|108864289|gb|ABA92870.2| retrotransposon protei (621)	219	59.1	2.5e-06
gi|194706326|gb|ACF87247.1| unknown [Zea mays] (360)	216	58.2	2.5e-06
gi|147783182|emb|CAN68669.1| hypothetical protein (1360)	222	59.9	2.9e-06
gi|113578926|dbj|BAF17289.1| Os05g0374700 [Oryza s (1301)	221	59.7	3.3e-06
gi|155323104|gb|ABT78708.1| Sequence 166178 from p (237)	210	56.7	4.7e-06
gi|78183243|emb|CAJ00274.1| hypothetical protein [(1508)	219	59.3	5.2e-06
gi|78183241|emb|CAJ00278.1| hypothetical protein [(1508)	218	59.0	6.1e-06
gi|155364082|gb|ABU19687.1| Sequence 207156 from p (211)	207	56.0	6.9e-06
gi|112065980|gb|ABI00028.1| Sequence 238 from pate (1499)	216	58.6	8.3e-06
gi|10998138|dbj|BAB03109.1| retroelement pol polyp (1499)	216	58.6	8.3e-06
gi|12322008|gb|AAG51046.1|AC069473_8 gypsy/Ty-3 re (1499)	216	58.6	8.3e-06
gi|113533208|dbj|BAF05591.1| Os01g0640000 [Oryza s (1660)	214	58.1	1.2e-05
gi|7800097|gb|AAF69808.1| pol protein [Picea abies (147)	200	54.3	1.6e-05
gi|5080762|gb|AAD39272.1|AC007203_4 Similar to ret (1264)	207	56.4	3.1e-05
gi|145012540|gb|EDJ97194.1| hypothetical protein M (1071)	206	56.2	3.1e-05
gi|29788873|gb|AAP03419.1| putative polyprotein [O (652)	201	54.9	4.6e-05
gi|31431768|gb|AAP53494.1| retrotransposon protein (999)	203	55.5	4.8e-05
gi|21672107|gb|AAM74469.1|AC124213_27 Putative ret (999)	203	55.5	4.8e-05
gi|18652523|gb|AAL77156.1|AC091732_7 Putative poly (999)	203	55.5	4.8e-05
gi|78183249|emb|CAJ00277.1| hypothetical protein [(1508)	204	55.8	5.8e-05
gi|18958673|gb|AAL82656.1|AC092387_4 retrotranspos (1338)	202	55.3	7.2e-05
gi|20270059|gb|AAM18147.1|AC092172_7 Putative gag- (1338)	202	55.3	7.2e-05
gi|78708062|gb|ABB47037.1| retrotransposon protein (1347)	202	55.3	7.2e-05
gi|116309666|emb|CAH66715.1| OSIGBa0118P15.5 [Oryz (1434)	201	55.1	8.9e-05
gi|147770944|emb|CAN69535.1| hypothetical protein (617)	195	53.5	0.00012
gi|78183245|emb|CAJ00275.1| hypothetical protein [(1112)	198	54.3	0.00012
gi|113537037|dbj|BAF09420.1| Os02g0633000 [Oryza s (1817)	200	54.9	0.00013
gi|113534021|dbj|BAF06404.1| Os01g0790200 [Oryza s (519)	193	53.0	0.00014
gi|38344578|emb|CAE05536.2| OSJNBa0053B21.10 [Oryz (582)	193	53.0	0.00015
gi|155300190|gb|ABT55794.1| Sequence 143264 from p (144)	184	50.6	0.0002
gi|155333463|gb|ABT89067.1| Sequence 176537 from p (144)	184	50.6	0.0002
gi|77551464|gb|ABA94261.1| retrotransposon protein (1369)	191	52.7	0.00043
gi|19881710|gb|AAM01111.1|AC098682_15 putative ret (1043)	189	52.2	0.00047
gi|116309348|emb|CAH66431.1| OSIGBa0096P03.5 [Oryz (413)	181	50.1	0.00079
gi|147841216|emb|CAN64356.1| hypothetical protein (1852)	188	52.1	0.0009
gi|113564206|dbj|BAF14549.1| Os04g0389000 [Oryza s (538)	181	50.2	0.00099
gi|7800092|gb|AAF69806.1| pol protein [Picea abies (83)	170	47.2	0.0012
gi|215704780|dbj|BAG94808.1| unnamed protein produ (218)	173	48.1	0.0017
gi|147855151|emb|CAN81740.1| hypothetical protein (771)	179	49.8	0.0018
gi|108862432|gb|ABA97314.2| retrotransposon protei (1219)	181	50.4	0.002
gi|190688747|gb|ACE86410.1| putative retroelement (1029)	179	49.9	0.0023
gi|57863925|gb|AAS55774.2| putative polyprotein [O (2108)	179	50.1	0.0042
gi|147768682|emb|CAN76063.1| hypothetical protein (1453)	176	49.3	0.005
gi|147769722|emb|CAN69702.1| hypothetical protein (1454)	176	49.3	0.0051
gi|147864892|emb|CAN79373.1| hypothetical protein (1439)	175	49.0	0.0059
gi|190688734|gb|ACE86397.1| putative retroelement (732)	170	47.7	0.0075
gi|147860462|emb|CAN82562.1| hypothetical protein (1384)	173	48.6	0.0079
gi|113611332|dbj|BAF21710.1| Os07g0518300 [Oryza s (1473)	173	48.6	0.0083
gi|77556305|gb|ABA99101.1| retrotransposon protein (1550)	173	48.6	0.0086
gi|38568032|emb|CAD40406.3| OSJNBa0065J03.2 [Oryza (1629)	173	48.6	0.009
gi|147810925|emb|CAN71787.1| hypothetical protein (1667)	173	48.6	0.0092
gi|147826806|emb|CAN63950.1| hypothetical protein (1545)	172	48.4	0.01
gi|147777812|emb|CAN64608.1| hypothetical protein (603)	167	47.0	0.01
gi|147783452|emb|CAN75210.1| hypothetical protein (1137)	169	47.6	0.013
gi|28209489|gb|AAO37507.1| retrotransposon protein (1155)	169	47.6	0.013
gi|155306694|gb|ABT62298.1| Sequence 149768 from p (158)	158	44.6	0.014
gi|18921316|gb|AAL82521.1|AC084766_7 putative poly (408)	162	45.7	0.017
gi|147775005|emb|CAN70471.1| hypothetical protein (1122)	167	47.1	0.017
gi|62733754|gb|AAX95863.1| retrotransposon protein (1126)	167	47.1	0.017
gi|77556354|gb|ABA99150.1| retrotransposon protein (1082)	165	46.7	0.023
gi|116310275|emb|CAH67280.1| OSIGBa0111L12.7 [Oryz (940)	164	46.4	0.024
gi|4235644|gb|AAD13304.1| polyprotein [Lycopersico (1542)	166	47.0	0.027
gi|108862641|gb|ABG22014.1| retrotransposon protei (1422)	164	46.5	0.034
gi|147772919|emb|CAN64786.1| hypothetical protein (1217)	163	46.2	0.035
gi|147819718|emb|CAN73574.1| hypothetical protein (1027)	161	45.7	0.042
gi|147840564|emb|CAN68329.1| hypothetical protein (1330)	162	46.0	0.045
gi|210072545|gb|EEA26632.1| retrovirus polyprotein (375)	155	44.1	0.048
gi|147810501|emb|CAN60890.1| hypothetical protein (1378)	161	45.8	0.054
gi|147845507|emb|CAN80599.1| hypothetical protein (1404)	161	45.8	0.055
gi|147810220|emb|CAN78060.1| hypothetical protein (1037)	159	45.3	0.059

```
gi|147784791|emb|CAN75226.1| hypothetical protein  ( 422)  154 43.9   0.062
gi|158578546|gb|ABW74571.1| pol polyprotein [Boech ( 450)  154 43.9   0.065
gi|108708778|gb|ABF96573.1| retrotransposon protei (1481)  160 45.6   0.067
gi|113548823|dbj|BAF12266.1| Os03g0411900 [Oryza s (1492)  160 45.6   0.068
gi|147860532|emb|CAN81876.1| hypothetical protein  (1241)  159 45.3   0.068
gi|34452676|gb|AAQ72731.1| putative reverse transc ( 435)  153 43.7   0.075
gi|136637170|gb|EBP43218.1| hypothetical protein G ( 248)  150 42.8   0.076
gi|155286956|gb|ABT42560.1| Sequence 130030 from p ( 142)  147 42.0   0.077
gi|147804663|emb|CAN62382.1| hypothetical protein  ( 688)  154 44.0   0.093
gi|4539021|emb|CAB39733.1| protease, reverse trans (1147)  155 44.3    0.12
gi|147835239|emb|CAN67792.1| hypothetical protein  (1414)  154 44.2    0.17
gi|147775377|emb|CAN69087.1| hypothetical protein  ( 611)  149 42.8    0.19
gi|147816473|emb|CAN64046.1| hypothetical protein  (1259)  152 43.7    0.21
gi|54291836|gb|AAV32204.1| putative polyprotein [O (1142)  151 43.4    0.23
gi|113596006|dbj|BAF19880.1| Os06g0590100 [Oryza s (1399)  152 43.7    0.23
gi|147775233|emb|CAN77085.1| hypothetical protein  (1520)  151 43.5    0.29
gi|198142740|gb|EDY71459.1| GA22199 [Drosophila ps (1555)  150 43.3    0.35
gi|34015213|gb|AAQ56407.1| putative gag-pol polypr (1619)  150 43.3    0.36
gi|193907775|gb|EDW06642.1| GI21862 [Drosophila mo ( 565)  143 41.4    0.46
gi|15042811|gb|AAK82434.1|AC091247_1 putative poly ( 799)  144 41.7    0.53
gi|108711868|gb|ABF99663.1| retrotransposon protei (1087)  144 41.8    0.68
gi|18855070|gb|AAL79762.1|AC096687_26 putative pol (1087)  144 41.8    0.68
gi|113550241|dbj|BAF13684.1| Os03g0827200 [Oryza s (1087)  144 41.8    0.68
gi|113645444|dbj|BAF28585.1| Os11g0600000 [Oryza s (1347)  145 42.1    0.69
gi|108864565|gb|ABA94541.2| retrotransposon protei (1347)  145 42.1    0.69
gi|193899715|gb|EDV98581.1| GH22324 [Drosophila gr ( 201)  135 39.3    0.71
gi|89355887|gb|ABD72267.1| pol polyprotein [Drosop (1143)  144 41.8    0.71
gi|147818980|emb|CAN67121.1| hypothetical protein  ( 938)  142 41.3    0.83
gi|147834477|emb|CAN63112.1| hypothetical protein  (1049)  142 41.3    0.91


>>gi|124359710|gb|ABN06064.1| RNA-directed DNA polymeras  (1297 aa)
 initn: 435 init1: 292 opt: 437  Z-score: 543.5  bits: 109.8 E(): 2.7e-21
Smith-Waterman score: 437;  47.222% identity (73.611% similar) in 144 aa
overlap (8-150:761-904)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 :: :.:::::::.:.:.:: :  :: .::::
gi|124 RPYLLGRRFVVSTDQKSLKQLLQQRVVTAEQQNWAAKLLGYDFEIIYKPGKLNKGADALS
             740       750       760       770       780       790

       40        50        60        70        80        90
5_3    RRDEDKEL-QGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFR
       :  :: :: :::.   :::   . ::...:  : ::  .:. . .:.: : :.  .: ..
gi|124 RVREDGELCQGITSVQWKDEKLLREELSRDSQLQKIIGDLQRDASSRPGYMLKQGVLLYE
             800       810       820       830       840       850

       100       110       120       130       140       150
5_3    GRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREI

       ::::. ..:. :: :: ::... .::::::.: :::::. ..::. .:. . ..:
gi|124 GRLVVSSKSVMIPTLLAEFHSTPQGGHSGFYRTYRRLAANVYWVGMKNTVQEYVRSCDTC
             860       870       880       890       900       910

       160
5_3    YMDQQ

gi|124 QRQKYLASSPGGLLQPLPVPDRIWEDLSMDFIMGLPKSKGYEAVLVVVDRLSKYSHFILL
             920       930       940       950       960       970

>>gi|147807720|emb|CAN66553.1| hypothetical protein [Vit  (1448 aa)
 initn: 422 init1: 422 opt: 422  Z-score: 524.0  bits: 106.4 E(): 3.3e-20
Smith-Waterman score: 422;  45.070% identity (76.056% similar) in 142 aa
overlap (9-150:480-621)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSR
                                    : :::::.::.:.: .. : .::.::::::
gi|147 HYLLGRHFIVRTDQSSLKFLLEQRIVNESYQKWVAKLFGYDFEIQFRPGMKNKAVDALSR
     450       460       470       480       490       500

       40        50        60        70        80        90
5_3    RDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRGR
          . :: ..  :   : . :. .:. :: ::::...: :.::..:.:.:.  :: ..:
gi|147 IPISMELAALMVPSRIDTSLISSQVEADPCLAKIKQRLLDDLDAYPRYALDHGILIYKGC
     510       520       530       540       550       560

      100       110       120       130       140       150
5_3    LVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIYM
       ::.  .:  .: :::: ..:..::::::.  ::.:.:....:. .:..: .::
gi|147 LVFPKASPLVPALLQEGHASVVGGHSGFLWTYKRLTRDFFWVGMKNDIKEFVEKCLVCQQ
     570       580       590       600       610       620

      160
5_3    DQQ

gi|147 NKTLTLSLAGLLQPLPIPDKIWDDVTMDFIEGLPKSEGYNFILVVVDRLSKYAHFSLLKH
     630       640       650       660       670       680

>>gi|147854459|emb|CAN78588.1| hypothetical protein [Vit  (2232 aa)
 initn: 408 init1: 408 opt: 408  Z-score: 503.6  bits: 103.2 E(): 4.5e-19
Smith-Waterman score: 408;  43.662% identity (73.944% similar) in 142 aa
overlap (9-150:1619-1760)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSR
                                    : :::::.::.:.: .. :  ::..::::::
gi|147 HYLLGRHFIVRTDQSSLKFLLEQRIVNESYQKWVAKLFGYDFEIQFRPGXENKAADALSR
     1590      1600      1610      1620      1630      1640
```

```
       40        50        60        70        80        90
5_3    RDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRGR
          . :: ..  :   : . :. .:. :: : ::...:  . :..:.:.:.  :: ..::
gi|147 IPISMELXALMVPSRIDTXLISSQVEADPHLXKIKQRLLXDPDAYPRYSLDHGILLYKGR
     1650      1660      1670      1680      1690      1700

      100       110       120       130       140       150
5_3    LVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIYM
       :::  .:  .: :::: ..:..:::::.  ::.:.:....:. .:..: .::
gi|147 LVLPKASPLVPALLQEGHASVVGGHSGFLXTYKRLTRDFFWVGMKNDIKEFVEKCLVCQQ
     1710      1720      1730      1740      1750      1760

      160
5_3    DQQ

gi|147 NKTLTLSPAGLLQPLPIPDKIWDDVTMDFIEGLPKSEVIVTRYFLADFGQSYFGCRVLLS
     1770      1780      1790      1800      1810      1820

>>gi|124360619|gb|ABD33394.2| FAR1; Polynucleotidyl tran  (657 aa)
 initn: 294 init1: 155 opt: 318  Z-score: 398.7  bits: 82.0 E(): 3.1e-13
Smith-Waterman score: 318;  40.141% identity (66.197% similar) in 142 aa
overlap (6-146:244-376)

                                     10        20        30
5_3                                ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDA
                                   ..:: :.::::.::.:::.:: :  :: .::
gi|124 HWRPYLMGRKFTVCTDQKSLRQLLQQQITTMDQQNWAAKLFGYQFDILYKPGLENKGADA
           220       230       240       250       260       270

          40        50        60        70        80        90
5_3    LSRRDEDKELQG-ISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILY
       :::  .. ::.. .  : : .  .. :::.:: .:  :   :. .  ..: .. .  .:.
gi|124 LSRVHDSPELHSLVHYPEWLEGKEVLEEVHKDEVLKGIISALQKGEPTKPGFAYRRGVLF
           280       290       300       310       320       330

         100       110       120       130       140       150
5_3    FRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
       .. :::: :.: :::::: ::..: ..   :      :. :::   :. ..:
gi|124 YEDRLVLAANSKWIPKLLLEFHSSPQATTPG------RLLQSL---PVPSQIWEDVSKDF
           340       350       360             370          380

          160
5_3    EIYMDQQ

gi|124 ITGLPKSKGYEAVMVVVDRLSKYAHFIPLKHPYSAKTLADVFVKEVIRLHGVPISIIFMD
          390       400       410       420       430       440

>>gi|6466937|gb|AAF13073.1|AC011621_1 putative retroelem  (1661 aa)
 initn: 265 init1: 152 opt: 319  Z-score: 393.9  bits: 82.5 E(): 5.8e-13
Smith-Waterman score: 319;  36.301% identity (69.863% similar) in 146 aa
overlap (8-150:1094-1238)

                                     10        20        30
5_3                                ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                   :: :..:: : .. : :: :..:::.::::
gi|646 KHYLSSKEFIIKTDQRSLRHLLEQKSVSTIQQRWASKLSGLKYRIEYKPGVDNKVADALS
          1070      1080      1090      1100      1110      1120

          40        50        60        70        80        90
5_3    RRDEDKELQG--ISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQ-YTLECDILY
       ::   . :.   :. :   :.: .. :.:.:  :..: ..  .. : : . .:.   ..:
gi|646 RRPPTEALSQLTITGPPTIDLTALKAEIQQDHELSQILKNWAQG-DHHDSDFTVADGLIY
          1130      1140      1150      1160       1170      1180

          100       110       120       130       140       150
5_3    FRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
        .: ::. ..: .::::.:..:.:: .::: :   :..:.:. .::  .. ......
gi|646 RKGCLVIPVGSPFIPKMLEKFHTSPIGGHEGALKTFKRLTSEVYWRGLRKDVVNYIKGCQ
           1190      1200      1210      1220      1230      1240

          160
5_3    EIYMDQQ

gi|646 ICQENKYSTLSPAGLLSPLPIPQQIWSDVSLDFVEGLPSSNRFNCILVVVDRLSKYSHFI
           1250      1260      1270      1280      1290      1300

>>gi|208609060|dbj|BAG72152.1| hypothetical protein [Lot  (1369 aa)
 initn: 262 init1: 140 opt: 315  Z-score: 390.2  bits: 81.5 E(): 9.3e-13
Smith-Waterman score: 315;  36.364% identity (65.734% similar) in 143 aa
overlap (8-150:815-954)

                                     10        20        30
5_3                                ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                   :: :..::.::.:.: :: :  ::..::::
gi|208 RHYLLGSKFVIHTDQRSLRFLADQRIMGEEQQKWMSKLMGYDFEIKYKPGIENKAADALS
          790       800       810       820       830       840

         40        50        60        70        80        90
5_3    RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
       :. . . .....   : :.   . :. .:   :. .::  . .:   : :.   : ..
gi|208 RKLQFSAISSVQCAEWADL---EAEILEDERYRKVLQELATQGNSAVGYQLKRGRLLYKD
          850       860          870       880       890       900

       100       110       120       130       140       150
5_3    RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
       :.::  .:  :  .:.::. . .:::.::. ::.::.  .::  .: .: ..:
gi|208 RIVLPKGSTKILTVLKEFHDTALGGHAGIFRTYKRISALFYWEGMKLDIQNYVQKCEVCQ
            910       920       930       940       950       960
```

```
       160
5_3    MDQQ

gi|208 RNKYEALNPAGFLQPLPIPSQGWTDISMDFIGGLPKAMGKDTILVVVDRFTKYAHFIALS
             970       980       990      1000      1010      1020

>>gi|208609055|dbj|BAG72150.1| hypothetical protein [Lot  (1558 aa)
 initn: 262 init1: 140 opt: 315  Z-score: 389.3  bits: 81.6 E(): 1e-12
Smith-Waterman score: 315;  36.364% identity (65.734% similar) in 143 aa
overlap (8-150:1004-1143)

                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :: :..::.::.:.: :: :  ::..:::: 
gi|208 RHYLLGSKFVIHTDQRSLRFLADQRIMGEEQQKWMSKLMGYDFEIKYKPGIENKAADALS
           980       990      1000      1010      1020      1030

        40        50        60        70        80        90
5_3    RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
       :. . . .....   : :.   . :. .:    :. .::  . .:   : :.   : ..
gi|208 RKLQFSAISSVQCAEWADL---EAEILEDERYRKVLQELATQGNSAVGYQLKRGRLLYKD
          1040      1050         1060      1070      1080      1090

       100       110       120       130       140       150
5_3    RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
       :.::  .:  :  .:.::. . .:::.::. ::.::.  .::  .: .: ..:
gi|208 RIVLPKGSTKILTVLKEFHDTALGGHAGIFRTYKRISALFYWEGMKLDIQNYVQKCEVCQ
             1100      1110      1120      1130      1140      1150

       160
5_3    MDQQ

gi|208 RNKYEALNPAGFLQPLPIPSQGWTDISMDFIGGLPKAMGKDTILVVVDRFTKYAHFIALS
             1160      1170      1180      1190      1200      1210

>>gi|208609057|dbj|BAG72151.1| hypothetical protein [Lot  (1558 aa)
 initn: 262 init1: 140 opt: 315  Z-score: 389.3  bits: 81.6 E(): 1e-12
Smith-Waterman score: 315;  36.364% identity (65.734% similar) in 143 aa
overlap (8-150:1004-1143)

                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :: :..::.::.:.: :: :  ::..:::: 
gi|208 RHYLLGSKFVIHTDQRSLRFLADQRIMGEEQQKWMSKLMGYDFEIKYKPGIENKAADALS
           980       990      1000      1010      1020      1030

        40        50        60        70        80        90
5_3    RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
       :. . . .....   : :.   . :. .:    :. .::  . .:   : :.   : ..
gi|208 RKLQFSAISSVQCAEWADL---EAEILEDERYRKVLQELATQGNSAVGYQLKRGRLLYKD
          1040      1050         1060      1070      1080      1090

       100       110       120       130       140       150
5_3    RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
       :.::  .:  :  .:.::. . .:::.::. ::.::.  .::  .: .: ..:
gi|208 RIVLPKGSTKILTVLKEFHDTALGGHAGIFRTYKRISALFYWEGMKLDIQNYVQKCEVCQ
             1100      1110      1120      1130      1140      1150

       160
5_3    MDQQ

gi|208 RNKYEALNPAGFLQPLPIPSQGWTDISMDFIGGLPKTMGKDTILVVVDRFTKYAHFIALS
             1160      1170      1180      1190      1200      1210

>>gi|208609065|dbj|BAG72154.1| hypothetical protein [Lot  (1558 aa)
 initn: 262 init1: 140 opt: 315  Z-score: 389.3  bits: 81.6 E(): 1e-12
Smith-Waterman score: 315;  36.364% identity (65.734% similar) in 143 aa
overlap (8-150:1004-1143)

                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :: :..::.::.:.: :: :  ::..:::: 
gi|208 RHYLLGSKFVIHTDQRSLRFLADQRIMGEEQQKWMSKLMGYDFEIKYKPGIENKAADALS
           980       990      1000      1010      1020      1030

        40        50        60        70        80        90
5_3    RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
       :. . . .....   : :.   . :. .:    :. .::  . .:   : :.   : ..
gi|208 RKLQFSAISSVQCAEWADL---EAEILEDERYRKVLQELATQGNSAVGYQLKRGRLLYKD
          1040      1050         1060      1070      1080      1090

       100       110       120       130       140       150
5_3    RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
       :.::  .:  :  .:.::. . .:::.::. ::.::.  .::  .: .: ..:
gi|208 RIVLPKGSTKILTVLKEFHDTALGGHAGIFRTYKRISALFYWEGMKLDIQNYVQKCEVCQ
             1100      1110      1120      1130      1140      1150

       160
5_3    MDQQ

gi|208 RNKYEALNPAGFLQPLPIPSQGWTDISMDFIGGLPKAMGKDTILVVVDRFTKYAHFIALS
             1160      1170      1180      1190      1200      1210

>>gi|208609062|dbj|BAG72153.1| hypothetical protein [Lot  (1558 aa)
 initn: 261 init1: 139 opt: 314  Z-score: 388.1  bits: 81.3 E(): 1.2e-12
Smith-Waterman score: 314;  36.364% identity (65.734% similar) in 143 aa
overlap (8-150:1004-1143)

                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
```

```
                                     :: :..::.::.:.: :: :  ::..::::
gi|208 RHYLLGSKFVIHTDQRSLRFLADQRIMGEEQQKWMSKLMGYDFEIKYKPGIENKAADALS
           980       990       1000      1010      1020      1030

        40        50        60        70        80        90
5_3    RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
       :. . . .....   : :.   . :. .:    :. .::  . .:   : :.   : ..
gi|208 RKLQFSAISSVQCAEWADL---EAEILEDERYRKVLQELATQGNSAVGYQLKRGRLLYKD
          1040      1050          1060      1070      1080      1090

       100       110       120       130       140       150
5_3    RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
       :.::  .:  :  .:.::. . .:::.::. ::.::.  .::  .: .: ..:
gi|208 RIVLPKGSTKILTVLKEFHDTAIGGHAGIFRTYKRISALFYWEGMKLDIQNYVQKCEVCQ
             1100      1110      1120      1130      1140      1150

       160
5_3    MDQQ

gi|208 RNKYEALNPAGFLQPLPIPSQGWTDISMDFIGGLPKAMGKDTILVVVDRFTKYAHFIALS
             1160      1170      1180      1190      1200      1210

>>gi|208609051|dbj|BAG72148.1| hypothetical protein [Lot  (1558 aa)
 initn: 261 init1: 139 opt: 314  Z-score: 388.1  bits: 81.3 E(): 1.2e-12
Smith-Waterman score: 314;  36.364% identity (65.734% similar) in 143 aa
overlap (8-150:1004-1143)

                                      10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                     :: :..::.::.:.: :: :  ::..::::
gi|208 RHYLLGSKFVIHTDQRSLRFLADQRIMGEEQQKWMSKLMGYDFEIKYKPGIENKAADALS
           980       990       1000      1010      1020      1030

        40        50        60        70        80        90
5_3    RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
       :. . . .....   : :.   . :. .:    :. .::  . .:   : :.   : ..
gi|208 RKLQFSAISSVQCAEWADL---EAEILEDERYRKVLQELATQGNSAVGYQLKRGRLLYKD
          1040      1050          1060      1070      1080      1090

       100       110       120       130       140       150
5_3    RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
       :.::  .:  :  .:.::. . .:::.::. ::.::.  .::  .: .: ..:
gi|208 RIVLPKGSTKILTVLKEFHDTAIGGHAGIFRTYKRISALFYWEGMKLDIQNYVQKCEVCQ
             1100      1110      1120      1130      1140      1150

       160
5_3    MDQQ

gi|208 RNKYEALNPAGFLQPLPIPSQGWTDISMDFIGGLPKAMGKDTILVVVDRFTKYAHFIALS
             1160      1170      1180      1190      1200      1210

>>gi|208609053|dbj|BAG72149.1| hypothetical protein [Lot  (1520 aa)
 initn: 262 init1: 140 opt: 312  Z-score: 385.7  bits: 80.9 E(): 1.6e-12
Smith-Waterman score: 312;  36.364% identity (65.035% similar) in 143 aa
overlap (8-150:966-1105)

                                      10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                     :: :..::.::.:.: :: :  ::..::::
gi|208 RHYLLGSQFVIHTDQRSLRFLADQRIMGEEQQKWMSKLMGYDFEIKYKPGIENKAADALS
           940       950       960       970       980       990

        40        50        60        70        80        90
5_3    RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
       :. . . .....   : :.   . :.  :    :. .::  . .:   : :.   : ..
gi|208 RKLQFSAISSVQCAEWADL---EAEILGDERYRKVLQELATQGNSAIGYQLKRGRLLYKD
          1000      1010          1020      1030      1040      1050

       100       110       120       130       140       150
5_3    RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
       :.::  .:  :  .:.::. . .:::.::. ::.::.  .::  .: .: ..:
gi|208 RIVLPKGSTKILTVLKEFHDTALGGHAGIFRTYKRISALFYWEGMKLDIQNYVQKCEVCQ
             1060      1070      1080      1090      1100      1110

       160
5_3    MDQQ

gi|208 RNKYEALNPAGFLQPLPIPSQGWTDISMDFIGGLPKAMGKDTILVVVDRFTKYAHFIALS
             1120      1130      1140      1150      1160      1170

>>gi|208609049|dbj|BAG72147.1| hypothetical protein [Lot  (1520 aa)
 initn: 262 init1: 140 opt: 312  Z-score: 385.7  bits: 80.9 E(): 1.6e-12
Smith-Waterman score: 312;  36.364% identity (65.035% similar) in 143 aa
overlap (8-150:966-1105)

                                      10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                     :: :..::.::.:.: :: :  ::..::::
gi|208 RHYLLGSQFVIHTDQRSLRFLADQRIMGEEQQKWMSKLMGYDFEIKYKPGIENKAADALS
           940       950       960       970       980       990

        40        50        60        70        80        90
5_3    RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
       :. . . .....   : :.   . :.  :    :. .::  . .:   : :.   : ..
gi|208 RKLQFSAISSVQCAEWADL---EAEILGDERYRKVLQELATQGNSAIGYQLKRGRLLYKD
          1000      1010          1020      1030      1040      1050

       100       110       120       130       140       150
5_3    RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
       :.::  .:  :  .:.::. . .:::.::. ::.::.  .::  .: .: ..:
```

```
gi|208 RIVLPKGSTKILTVLKEFHDTALGGHAGIFRTYKRISALFYWEGMKLDIQNYVQKCEVCQ
          1060      1070      1080      1090      1100      1110

       160
5_3    MDQQ

gi|208 RNKYEALNPAGFLQPLPIPSQGWTDISMDFIGGLPKAMGKDTILVVVDRFTKYAHFIALS
          1120      1130      1140      1150      1160      1170
```

>>gi|147828709|emb|CAN66228.1| hypothetical protein [Vit (1258 aa)
initn: 301 init1: 301 opt: 308 Z-score: 381.9 bits: 79.9 E(): 2.7e-12
Smith-Waterman score: 308; 40.000% identity (68.889% similar) in 135 aa overlap (9-141:865-999)

```
                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSR
                                    : :::::.::.:.: .    .::..:::::
gi|147 HYLLGRHFIVQTDQSSLKFLLEQRVVNELYQKWVAKLFGYDFEIQFPPELKNKAADALSR
          840       850       860       870       880       890

       40        50        60        70        80        90
5_3    RDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRGR
          . .:     :   : . :: .:. :: :::: ..:  .:...:.:.:.  .: ..::
gi|147 IPISMDLATHMVPSRLDTSLINSQVEADPHLAKILQRLLVDLNAYPRYSLDHGVLLYKGR
          900       910       920       930       940       950

      100       110       120       130       140       150
5_3    LVLLASSLWIPKLLQEFQTSLMGGHSG--IYITYRRITQSLYWIPIKGEITKFVVHVREI
       :::  .:  .: :::: ..:..::::.  . ... . . .:. ::
gi|147 LVLPKASPLVPTLLQEGHASVVGGHSAQTVVVVFVQDVVKLHGIPHSIVSDRDKVFLSLF
          960       970       980       990      1000      1010

        160
5_3    YMDQQ

gi|147 WTKLFRLLGTSLCHSTAYHPQTDGQTEVVNRCVETYLRCFSYNKPRRWSTWLPWAKYWYN
          1020      1030      1040      1050      1060      1070
```

>>gi|147843077|emb|CAN83300.1| hypothetical protein [Vit (1366 aa)
initn: 280 init1: 156 opt: 302 Z-score: 373.9 bits: 78.5 E(): 7.5e-12
Smith-Waterman score: 302; 34.932% identity (64.384% similar) in 146 aa overlap (8-150:910-1055)

```
                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :: :..:::::::. :: :  :::.::::
gi|147 RSYLLGHNFKIQTDQQSLKYLLEEKMGTPLQQQWITKLLGYEFVVEYKQGKENKVADALS
     880       890       900       910       920       930

        40           50        60        70        80        90
```

```
5_3    RRDEDKE---LQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILY
       :. ::..   : .:. :    . ..      :: : .: ..:...  .  .:  .  .:.
gi|147 RKMEDQKEGKLYAITAPANTWLEQLRTSYAIDPKLQQIIKNLEQGSLASQNYKQRDGLLF
     940       950       960       970       980       990

          100       110       120       130       140       150
5_3    FRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
       ..::: . ::.    ..:  ...: .::::::.. : .:  . .::  .. :. .:.
gi|147 YKGRLYIPASKELREQILYLLHSSPQGGHSGFHKTLHRAKSEFYWEGMRKEVRRFIKECD
    1000      1010      1020      1030      1040      1050

          160
5_3    EIYMDQQ

gi|147 ICQQNKSENIHPAGLLQPLPIPTKSVIMVVVDRLSKYAHFIPISHPYTASKIAQVFLANI
    1060      1070      1080      1090      1100      1110
```

>>gi|147774273|emb|CAN76793.1| hypothetical protein [Vit (1469 aa)
initn: 279 init1: 155 opt: 301 Z-score: 372.1 bits: 78.3 E(): 9.4e-12
Smith-Waterman score: 301; 34.932% identity (64.384% similar) in 146 aa overlap (8-150:943-1088)

```
                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :: :..:::::::. :: :  :::.::::
gi|147 RSYLLGHNFKIQTDQXSLKYLLEQKMGTPLQQKWITKLLGYEFVVEYKQGKENKVADALS
          920       930       940       950       960       970

        40           50        60        70        80        90
5_3    RRDEDKE---LQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILY
       :. ::..   : .:. :    . ..      :: : .: ..:...  .  .:  .  .:.
gi|147 RKMEDQKEGKLYAITAPANTWLEQLRTXYAIDPKLQQIIKNLEQGSLASQNYKQRDGLLF
          980       990      1000      1010      1020      1030

          100       110       120       130       140       150
5_3    FRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
       ..::: . ::.    ..:  ...: .::::::.. : .:  . .::  .. :. .:.
gi|147 YKGRLYIPASKELREQILYLLHSSPQGGHSGFHKTLHRAKSEFYWEGMRKEVRRFIKECD
           1040      1050      1060      1070      1080      1090

          160
5_3    EIYMDQQ

gi|147 ICQQNKSENIHPAGLLQPLPIPTKVWTDISLDFIEGLPNSESYSVIMVVVDRLSKYAHFI
          1100      1110      1120      1130      1140      1150
```

>>gi|147772855|emb|CAN73669.1| hypothetical protein [Vit (1308 aa)
initn: 266 init1: 149 opt: 299 Z-score: 370.4 bits: 77.8 E(): 1.2e-11
Smith-Waterman score: 299; 35.669% identity (62.420% similar) in 157 aa overlap (6-159:801-951)

```
                                      10        20        30
5_3                                ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDA
                                   :.:: :::::::::...:....:  :...::
gi|147 LWRPYLLGRKFYIKTDQQSLKIFLDQHVATLEQQKWVAKLLGYDYEIIFRTGRENSAADA
            780       790       800       810       820       830

         40        50           60        70        80        90
5_3    LSRRDEDKELQGISRP---FWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDI
       ::::.:.  :  .      .::   .: :  ..:  .  . ..  :    : . : .  .
gi|147 LSRRQESPLLATLHFSEVDIWK---QIREAFKSDSYVQLLGKKAGD--PPHGNLTWHDGL
            840       850          860       870         880

          100       110       120       130       140       150
5_3    LYFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVH
       : ..:..:. :.     ::: : . : .:::::: ::::. :..::  ..  .  :: .
gi|147 LLYKGKVVVPADHSLRAKLLYEVHDSKVGGHSGILRTYRRLQQQFYWPKMHKAVQXFVQK
         890       900       910       920       930       940

           160
5_3    VREIYMDQQ
         :.. :
gi|147 C-EVWEDITLDFIEGLPTSHGKDTILVVVDRLSKFAHFIPLTHPFTAKVVVENFIEGVVK
          950       960       970       980       990      1000
```

>>gi|147789424|emb|CAN66607.1| hypothetical protein [Vit (2822 aa)
initn: 280 init1: 156 opt: 302 Z-score: 369.2 bits: 78.7 E(): 1.4e-11
Smith-Waterman score: 302; 34.932% identity (64.384% similar) in 146 aa
overlap (8-150:896-1041)

```
                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 :: :..:::::::: . :: :  :::.::::
gi|147 RSYLLGHNFKIQTDQQSLKYLLEQKMGTPLQQQWITKLLGYEFVVEYKQGKENKVADALS
         870       880       890       900       910       920

        40           50        60        70        80        90
5_3    RRDEDKE---LQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILY
       :. ::..   : .:. :    . ..      :: : .: ..:...  .  .:  .  .:.
gi|147 RKMEDQKEGKLYAITAPANTWLEQLRTSYAIDPKLQQIIKNLEQGSLASQNYKQRDGLLF
         930       940       950       960       970       980

          100       110       120       130       140       150
5_3    FRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
       ..::: . ::.    ..:  ...: .::::::.. : .:  . .::  .. :. .:.
gi|147 YKGRLYIPASKELREQILYLLHSSPQGGHSGFHKTLHRAKSEFYWEGMRKEVRRFIKECD
         990      1000      1010      1020      1030      1040

          160
5_3    EIYMDQQ
```

```
gi|147 ICQQNKSENIHPAGLLQPLPIPTKLAQDRMKKFANIKRTARSFNIGDLVYLRLQPYKQQS
        1050      1060      1070      1080      1090      1100
```

>>gi|8778789|gb|AAF79797.1|AC020646_20 T32E20.30 [Arabid (1397 aa)
initn: 309 init1: 180 opt: 297 Z-score: 367.5 bits: 77.4 E(): 1.7e-11
Smith-Waterman score: 297; 33.117% identity (67.532% similar) in 154 aa
overlap (3-150:863-1016)

```
                                         10        20        30
5_3                                   ISELQLNQQYWVAKLLGYEFDIVYKVGASNKV
                                      :.... : :..::: :::::.::.:..::.
gi|877 SIQKWKHYLMGRRFVLHTDQKSLKFLQEQREVSMDYQKWLTKLLHYEFDILYKLGVDNKA
            840       850       860       870       880       890

             40              50        60        70        80
5_3    VDALSRRDEDKE------LQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQY
       .:.:::  .          :.... :   ..  . ::....  : .. .:  .  ..   :
gi|877 ADGLSRMVQPTGSFSSMLLMAFTVPTVLQLHDLYEEIDSNAHLQHLVKECLSAKQGTSAY
            900       910       920       930       940       950

         90       100       110       120       130       140
5_3    TLECDILYFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEI
       :..   :. . ::..  .: ..: .: :....:.:::::.  :..:: ::..:  .  .:
gi|877 TVKEGRLWKKQRLIIPKDSKFLPLILAEYHSGLLGGHSGVLKTMKRIQQSFHWEGMMKDI
            960       970       980       990      1000      1010

        150       160
5_3    TKFVVHVREIYMDQQ
        :::
gi|877 QKFVAKCEMCQRQKYSTLSPAGLLQPLPIPTQVWEDISLDFVEGLPDRLSKYGHFIGLKH
           1020      1030      1040      1050      1060      1070
```

>>gi|170660047|gb|ACB28472.1| polyprotein [Ananas comosu (953 aa)
initn: 296 init1: 172 opt: 292 Z-score: 363.7 bits: 76.1 E(): 2.8e-11
Smith-Waterman score: 292; 34.194% identity (63.871% similar) in 155 aa
overlap (7-161:607-755)

```
                                      10        20        30
5_3                                ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDAL
                                   .:: :::::.::.....:: :  : :.:::
gi|170 WRPYLIGRHFKIKTDHQSLKYLMEQRVSTPSQQKWVAKLMGYDYELIYKKGQENVVADAL
        580       590       600       610       620       630

         40        50        60        70        80        90
5_3    SRRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFR
       ::   .  : ..:      . .:.   . :  : :: .. .....: :.::   : :  .
gi|170 SR---SPTLLAVSAIHTDLLDQIKWSWNVDDKLKKIIQQKQSDINSWPRYTWVQDQLRRK
          640       650       660       670       680       690
```

```
        100       110       120       130       140       150
5_3     GRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREI
        :.::. ..     .:...:..: .::::::  : :..  ..::  .. .. .::   ::
gi|170 GKLVVGSDPGLKLQLIHNFHASSIGGHSGMEATTRKLKGQFYWKGLRRDVEQFV---REC
            700       710       720       730       740           750

        160
5_3     YMDQQ
         . ::
gi|170 SVCQQNKYETTAPAGLLQPLPIPEGIWTEISMDFIEGLPNSQGKEVIMVVVDRLSKYAHF
               760       770       780       790       800       810

>>gi|147823217|emb|CAN75278.1| hypothetical protein [Vit  (716 aa)
 initn: 280 init1: 280 opt: 285  Z-score: 356.8  bits: 74.4 E(): 6.8e-11
Smith-Waterman score: 285;  43.750% identity (69.643% similar) in 112 aa
overlap (9-120:438-549)

                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSR
                                    : :::::.::.:.: .  :  ::..:::::
gi|147 HYQLGRHFIVQTDQSSLKFLLEQRVVNESYQKWVAKLFGYDFEIQFWPGLENKAADALSR
       410       420       430       440       450       460

       40        50        60        70        80        90
5_3    RDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRGR
          . :: ..      : . :  .:. .: ::::::.:  .::..:.:.:.  :: ..::
gi|147 IPISMELAALMVSSRLDTSLIXSQVEINPHLAKIRQKLLVDLDAYPRYSLDHGILLYKGR
       470       480       490       500       510       520

     100       110       120       130       140       150
5_3    LVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIYM
       :::  .:  .: :::: ..:..
gi|147 LVLPKASPLVPALLQEGHASMLKRWLLSLFVMLSNYMEFPVPLLVIVTKYFLVBFGRNYF
       530       540       550       560       570       580

>>gi|113595720|dbj|BAF19594.1| Os06g0493000 [Oryza sativ  (1573 aa)
 initn: 285 init1: 204 opt: 289  Z-score: 356.6  bits: 75.5 E(): 6.9e-11
Smith-Waterman score: 289;  36.806% identity (62.500% similar) in 144 aa
overlap (8-150:1054-1195)

                                     10        20        30
5_3                            ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                     :.  . :::.:.. : :: :  ::..:::::
gi|113 KHYFLGTSLIIRTDQASLKYINEQRITEGVQHKLLIKLLSYDYKIEYKKGQENKAADALS
          1030      1040      1050      1060      1070      1080

        40        50        60        70        80        90
5_3     RRDEDKELQG-ISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFR
        : .. . :   .  : :  ::..    . ::   ... .:.   .::: :::.  :: ..
gi|113 RMQQLNALTTTVIVPQW--ITEVAASYSTDPKCHELESHLHIAPQSHPPYTLKGGILRYK

           1090      1100        1110      1120      1130      1140

        100       110       120       130       140       150
5_3     GRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREI
         ..:. :..    .::  :. : .:::::   ::.:. : .::  .:  .:.::
gi|113 DHIVVGAGNTLREQLLVSFHDSALGGHSGERATYQRMKQLFYWPGMKLAVTQFVKACPVC
             1150      1160      1170      1180      1190      1200

        160
5_3     YMDQQ

gi|113 QKNKTEHNMPAGLLQPLPLPEMAWSHITMDFVEGLPKSEGKDVIWVIVDRFTKYAHFIPL
             1210      1220      1230      1240      1250      1260

>>gi|116309032|emb|CAH66146.1| OSIGBa0114M03.4 [Oryza sa  (1448 aa)
 initn: 247 init1: 157 opt: 285  Z-score: 352.2  bits: 74.6 E(): 1.2e-10
Smith-Waterman score: 285;  36.552% identity (68.276% similar) in 145 aa
overlap (8-150:891-1030)

                                     10        20        30
5_3                            ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                     :..::.::::..:.. :. :::: :.::::
gi|116 RPYLWGRRFIVKTDHYSLKYLLDQRLATIPQHHWVGKLLGFDFSVEYRSGASNTVADALS
              870       880       890       900       910       920

       40         50        60        70        80        90
5_3    RRD-EDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFR
       ::: .:  : .:: : .  ::.. .    ::::. :..:.. .  . : .:.  :.. .
gi|116 RRDVDDGALLAISAPRFDFITRLRHAQATDPALVAIHDEVRAGTRAAP-WTVVDDMVAYD
              930       940       950       960        970

       100       110       120       130       140       150
5_3    GRLVLLASSLWIPKLLQEFQTSLMG-GHSGIYITYRRITQSLYWIPIKGEITKFVVHVRE
       ::: .  .:   : ::::.....   :: :.. : .:. .....  ..  .  ::
gi|116 GRLYIPPTS---P-LLQEIMAAVHDDGHEGVHRTLHRLRRDFHFPNMRRLVQDFVRACTT
     980           990      1000      1010      1020      1030

        160
5_3     IYMDQQ

gi|116 CQRYKSEHLHPAGLLQPLPVPSIVWADIGIDFVEALPRVHGKTVILSVVDRFSKYCHFIP
        1040      1050      1060      1070      1080      1090

>>gi|155327954|gb|ABT83558.1| Sequence 171028 from paten  (161 aa)
 initn: 260 init1: 260 opt: 270  Z-score: 347.6  bits: 70.6 E(): 2.2e-10
Smith-Waterman score: 270;  34.752% identity (63.121% similar) in 141 aa
overlap (12-150:8-146)

               10        20        30        40          50
5_3     ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGIS--RPFWKDITK
```

```
                . ::.: .. : :: : .::..:::::::..:.:. .::  .: : :
gi|155     TPTSNKNIFKLMGLQYKIFYKPGITNKAADALSRREHDQEVYAISAVQPKWLD--T
                  10        20        30        40        50

       60        70        80        90       100       110
5_3    INEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRGRLVLLASSLWIPKLLQEFQTS
       . :  . ::   ..  ::. .  .   ..:   :. ...:. : . .     .:  ...:
gi|155 VIEGYRLDPKAQQLLTELSLSNTNDKGFSLVNGIIKYKNRVWLGTHKAAHAAVLLALHSS
           60        70        80        90       100       110

      120       130       140       150       160
5_3    LMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIYMDQQ
        ::::::.  ::.::.. . :  .: .: :::
gi|155 AMGGHSGFLATYHRIXKLFAWPHMKEDIQKFVQTCEVCQQAKTEHNK
          120       130       140       150       160

>>gi|77551099|gb|ABA93896.1| retrotransposon protein, pu  (897 aa)
 initn: 275 init1: 185 opt: 278  Z-score: 346.5  bits: 72.8 E(): 2.5e-10
Smith-Waterman score: 278;  30.612% identity (63.946% similar) in 147 aa
overlap (8-150:375-519)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 ::  ..:::: .. . :: : .::..:::
gi|775 DPKNVQAVQQWAIPSNVKEVRGDQRLATPWQQKAMTKLLGLQYKLQYKKGLDNKAADALS
          350       360       370       380       390       400

        40          50          60        70        80        90
5_3    RRDEDKELQ--GISR--PFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       :  .   .:  ..:   :.:  .....:  . .::  ..  ..  . :..:.::..  .:
gi|775 RYPQTDPMQICALSAVVPLW--LNEVQEGYKTEPATEQLLTQVLLSPDQYPHYTMQQGVL
          410       420         430       440       450       460

          100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
        ..::... ..     ..:  ...: .:::::: .:: :.   . :  .:  .  ::
gi|775 LYKGRILVGSNPALQHRILTALHASAIGGHSGIQVTYSRVKWLFAWTGLKRAVQTFVTDC
           470       480       490       500       510       520

          160
5_3    REIYMDQQ

gi|775 AVCKHAKSERMRYPGLLQPLPVPDQAWETVSLDFIEGLPKSSGFDCILVVVDKFSRYAHF
           530       540       550       560       570       580

>>gi|77557165|gb|ABA99961.1| retrotransposon protein, pu  (1619 aa)
 initn: 272 init1: 172 opt: 279  Z-score: 343.9  bits: 73.2 E(): 3.5e-10
Smith-Waterman score: 279;  35.333% identity (68.000% similar) in 150 aa
overlap (8-152:1040-1186)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 :.  ..:::: .. :.:: :. :...::::
gi|775 RPYLQHAEFSIRTDHRSLAFLDEQRLTTPWQHKALTKLLGLQYKILYKKGSENSAADALS
    1010      1020      1030      1040      1050      1060

        40          50          60        70        80        90
5_3    RRDEDKE---LQGISR--PFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDI
       :   :::   :...:   : : .   :.  .: . .:.:. . :  : .. :..::.  .
gi|775 RYP-DKETVVLSALSVCIPEWTQ-EVIEGYAQDSDSLSKV-QTLCINNSAIPDFTLKNGL
    1070       1080      1090       1100       1110      1120

           100       110       120       130       140       150
5_3    LYFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVH
       :::. .. .  . .   :.: ...:. .:::::: .::.:: : . :: ......::. :
gi|775 LYFKDKMWIGNNPIVQRKILANLHTAAIGGHSGITMTYQRIKQLFAWIGLRSDVVKFIQH
       1130      1140      1150      1160      1170      1180

           160
5_3    VREIYMDQQ

gi|775 CTICQQAKGEHVKYPGMLQPLPVPEQSWQIVSLDFIEGLPRSSTFNCILVVVDKFSKYAH
       1190      1200      1210      1220      1230      1240

>>gi|147768278|emb|CAN60449.1| hypothetical protein [Vit  (647 aa)
 initn: 280 init1: 148 opt: 274  Z-score: 343.6  bits: 71.8 E(): 3.7e-10
Smith-Waterman score: 274;  36.486% identity (65.541% similar) in 148 aa
overlap (6-150:98-240)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDA
                                 :.:: ::::::::...:.:: :  :..:::
gi|147 TWRPYLLGQKFYIQTNQRSLKYLLEQRVVTLEQQKWVAKLLGYDYEILYKPGRENSAVDA
        70        80        90       100       110       120

          40        50        60          70        80        90
5_3    LSRRDEDKELQGISRPFWKDITKINEEVQK---DPALAKIREELKDNLDSHPQYTLECDI
       :::   .  ....   : .. .:: ::..    : :      .:  .  . : :: .  .
gi|147 LSRVPSSLTFNAL---FVSQ-AKIWEEIKTAAADDAYMTCISKLAATKPGLP-YTNRQGL
       130       140          150       160       170       180

           100       110       120       130       140       150
5_3    LYFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVH
        ....:.:.  .:    .::.::. : .::::.  ::.::.:..::  .   ....:
gi|147 TFYKNRVVVPPQSHIPNQLLREFHDSPLGGHSGVLRTYKRIAQQFYWPSMYRMVNEYVSS
           190       200       210       220       230       240

           160
5_3    VREIYMDQQ
```

```
gi|147 CDVCQRAKASTLSPAGLLQPLPIPCQVWDDITMDFIEGLPPSQGKNTILVVVDRLSKSAH
           250       260       270       280       290       300

>>gi|32488663|emb|CAE03590.1| OSJNBa0087O24.13 [Oryza sa  (1311 aa)
 initn: 248 init1: 248 opt: 277  Z-score: 342.8  bits: 72.7 E(): 4.1e-10
Smith-Waterman score: 277;  37.241% identity (59.310% similar) in 145 aa
overlap (8-150:791-933)

                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                :.  . :::.:.. : :: :  ::..::::
gi|324 KHYFLGTSLIIRTDQASLKYINEQRLTEGIQHKLLIKLLSYDYKIEYKKGKENKAADALS
             770       780       790       800       810       820

        40          50        60        70        80        90
5_3     RRDEDKEL--QGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
        :     .:    :  : :  ::.:      ::  . .. .:. . ..:: :::   :: .
gi|324 RIPSVAQLFSTTIIVPTW--ITEILASYATDPKCTALESQLRITPQGHPPYTLTSGILRY
              830         840       850       860       870

         100       110       120       130       140       150
5_3     RGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVRE
        ..:: . :..    :: : :. : .:::::   ::.:    .::  .: .:...:
gi|324 KNRLYVGAGTDLRAKLQQSFHDSALGGHSGERATYQRAKLLFYWPGMKKDIASYVKLCPV
      880       890       900       910       920       930

         160
5_3     IYMDQQ

gi|324 CQKNKSEHNLQPGLLHPLPIPEMAWTHISMDFIEGLPKSDNKDVIWVIVDRFTKYAHFVA
      940       950       960       970       980       990

>>gi|27764548|gb|AAO23078.1| polyprotein [Glycine max]    (1552 aa)
 initn: 258 init1: 150 opt: 276  Z-score: 340.4  bits: 72.5 E(): 5.5e-10
Smith-Waterman score: 276;  32.867% identity (64.336% similar) in 143 aa
overlap (8-150:976-1113)

                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                :: :. :.::::.: : :: : .:...::::
gi|277 RHYLLGNKFIIRTDQRSLKSLMDQSLQTPEQQAWLHKFLGYDFKIEYKPGKDNQAADALS
         950       960       970       980       990      1000

        40        50        60        70        80        90
5_3     RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
        :      . . :.:    . ..  .. .:: : .. :  :.. :.  .::..  .::..
gi|277 R----MFMLAWSEPHSIFLEELRARLISDPHLKQLMETYKQGADAS-HYTVREGLLYWKD
            1010      1020      1030      1040      1050      1060

       100       110       120       130       140       150

5_3     RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
        :.:. :    . :.::::..: .::::::  :  :.  ..::  .. ..  ..
gi|277 RVVIPAEEEIVNKILQEYHSSPIGGHAGITRTLARLKAQFYWPKMQEDVKAYIQKCLICQ
             1070      1080      1090      1100      1110      1120

       160
5_3     MDQQ

gi|277 QAKSNNTLPAGLLQPLPIPQQVWEDVAMDFITGLPNSFGLSVIMVVIDRLTKYAHFIPLK
             1130      1140      1150      1160      1170      1180

>>gi|147786920|emb|CAN64437.1| hypothetical protein [Vit  (623 aa)
 initn: 263 init1: 148 opt: 271  Z-score: 340.1  bits: 71.1 E(): 5.7e-10
Smith-Waterman score: 271;  34.247% identity (66.438% similar) in 146 aa
overlap (8-150:92-232)

                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                :: ::::::::::.:.:: .  :...::::
gi|147 RPYLLGQKFYIQTDQRSLKYLLEQRVVTPEQQKWVAKLLGYDYEILYKPSCENSTADALS
             70        80        90       100       110       120

        40          50        60        70        80        90
5_3     RRDEDKELQG--ISRP-FWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILY
        :   .  :..  .:.  .:..:  .  .   :  .:.: . :  .  . : :: .  . .
gi|147 RVPGSPTLNALFVSQAKIWEEIKTVAAD---DAYMARISK-LAATKPGLP-YTNRQGLTF
             130       140           150       160       170

          100       110       120       130       140       150
5_3     FRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
        ...:.:.  .:    .::.::. : .::::::  ::.::.:..::  .   .....
gi|147 YKNRVVVPPQSHIPNQLLREFHDSPLGGHSGVLRTYKRIAQQFYWPSMYRMVNEYISSCD
        180       190       200       210       220       230

         160
5_3     EIYMDQQ

gi|147 VCQRAKASTLSSTGLLQPLPIPCQVWDDITMDFIEGLPPSQGKNTILVVVDRLSKSAHFL
        240       250       260       270       280       290

>>gi|108862596|gb|ABA97714.2| retrotransposon protein, p  (1287 aa)
 initn: 196 init1: 149 opt: 272  Z-score: 336.6  bits: 71.5 E(): 8.9e-10
Smith-Waterman score: 272;  34.483% identity (68.276% similar) in 145 aa
overlap (8-150:724-863)

                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                :..::.::::..:.. :. ::.: :.::::
gi|108 AFFSRPVAPHHRALAAYECELIGLRLAMIPQHHWVGKLLGFDFSVEYRSGATNTVADALS
           700       710       720       730       740       750
```

```
        40         50         60         70         80         90
5_3     RRD-EDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFR
        ::: :. :: .:: : .  : .. .    ::.:. :..:.. .  . : ...  :.. .
gi|108 RRDIEEGELLAISAPRFDFIERLRHAQATDPSLVAIHDEVRAGTRAAP-WAVVTDMVTYD
            760        770        780        790        800        810

        100        110        120        130        140        150
5_3     GRLVLLASSLWIPKLLQEFQTSLMG-GHSGIYITYRRITQSLYWIPIKGEITKFVVHVRE
        ::: . ..:   : ::::. ...   :: :.. : .:. .....  ..  .  ::
gi|108 GRLYIPSAS---P-LLQEIVAAVHDDGHEGVHRTLHRLHRDFHFPHMRRLVQDFVKACVT
             820           830        840        850        860

         160
5_3     IYMDQQ

gi|108 CQRYKSEHLYPAGLLQPLPVPSIVWANIGLDFVEALPRVHAKTIILSVVDRFSKYCHFIP
       870        880        890        900        910        920

>>gi|147779107|emb|CAN73467.1| hypothetical protein [Vit  (1593 aa)
 initn: 246 init1: 134 opt: 272  Z-score: 335.2  bits: 71.6 E(): 1.1e-09
Smith-Waterman score: 272;  33.566% identity (65.734% similar) in 143 aa
overlap (8-150:812-952)

                                    10         20         30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :: :..::.::...:::: : .:...::::
gi|147 RPYLLGRRFTIQTDQRSLKYLLEQRIITPEQQKWMSKLVGYDYEIVYKPGKTNQAADALS
             790        800        810        820        830        840

        40         50         60         70         80         90
5_3     RRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFRG
        :   .  :. .  :  .  .: .:....: . .: . :  .   .: :  .  .. . .
gi|147 RNMTSPCLNVFFVPQVQVWDEIRHEANSNPYMQRIGQ-LATKQPRQP-YQWRNGLVCYNN
             850        860        870        880        890

       100        110        120        130        140        150
5_3     RLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIY
        :.:.  .:  :  ::.::. . ::::: :  ::.:..:..::  ..  . ..:
gi|147 RIVVPPGSPLIHCLLREFHDTPMGGHSRILRTYKRLSQQFYWPSMRRSVHQYVAACDVCQ
      900        910        920        930        940        950

       160
5_3     MDQQ

gi|147 KAKAETMSPAGLLQPLPIPCQVWDDITMDFIDGLPRSDGKTSIMVVVDRLSKSAHFIAIA
      960        970        980        990       1000       1010

>>gi|215767834|dbj|BAH00063.1| unnamed protein product [  (494 aa)
 initn: 264 init1: 186 opt: 264  Z-score: 332.8  bits: 69.4 E(): 1.5e-09
Smith-Waterman score: 264;  36.054% identity (63.946% similar) in 147 aa
overlap (8-150:92-236)

                                       10         20         30
5_3                                 ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                       ::  :.::::  . :::: :. : ..:::::
gi|215 RPYLQHAEFCIRTDHRSLSFLSDQRLSTPWQQKAVTKLLGLCYRIVYKKGTENGTADALS
             70         80         90        100        110        120

        40         50         60         70         80         90
5_3     RRDEDK--ELQGISR--PFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
        . ....   :...:   : :  . .: .  :  :   .  . :  :  . ::.::.  ::
gi|215 HVSDSSLVTLSALSVCIPEW--MQEIIDGYQLVPDACSNVQALCINNAALPQFTLKNGIL
             130        140         150        160        170

           100        110        120        130        140        150
5_3     YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
        ::. :. .  .:    :.: ...:. .:::::: .::.:: : . :  .......::
gi|215 YFQDRIWIGQNSSVQQKILANLHTAAIGGHSGIQVTYQRIKQLFAWPGLRADVVRFVQSC
     180        190        200        210        220        230

          160
5_3     REIYMDQQ

gi|215 DICQRAKSEHVRYPGFLQPLPVPDQYWQVVSLDFIEGLPRSASFNCILVVVDKFSKYAHF
     240        250        260        270        280        290

>>gi|147866223|emb|CAN81986.1| hypothetical protein [Vit  (672 aa)
 initn: 267 init1: 163 opt: 265  Z-score: 332.1  bits: 69.7 E(): 1.6e-09
Smith-Waterman score: 265;  35.669% identity (61.783% similar) in 157 aa
overlap (8-161:81-232)

                                       10         20         30
5_3                                 ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                       :: :::::. :...: :: :  : ..:::::
gi|147 LSDRFQTLSTYEKEMLAILMAIKKRVTTPTQQAWVAKLMQYDYEIRYKQGKENVAADALS
             60         70         80         90        100        110

        40         50          60         70         80         90
5_3     RRDEDKELQGISRPFWKD--ITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
        :  .  ::  .:  . .      :.:    :: : :: .  . . ...:.:. . . :
gi|147 RI-QPAELFVLSTTILNTQLYDLIKESWGVDPELQKIIKAKEADPSAYPKYSWRGEELRR
               120        130        140        150        160

         100        110        120        130        140        150
5_3     RGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVV-HVR
        .:.::. ..     ..:. :. :  :::::.:.: .::.  .::   :: . :::  .::
gi|147 KGKLVVGVNEQLRREILNSFHDSPTGGHSGVYVTTKRISAVIYW---KG-LRKFVREYVR
     170        180        190        200        210           220
```

```
         160
5_3    EIYMDQQ
       . .. :.
gi|147 NCFVCQRFKPENKPYSGLLQPLPVPEGVFTDITMDFIEGLPKSNGKTAIFVVVDRLTKYG
         230       240       250       260       270       280

>>gi|147864527|emb|CAN80491.1| hypothetical protein [Vit  (1412 aa)
 initn: 266 init1: 161 opt: 268  Z-score: 331.0  bits: 70.6 E(): 1.8e-09
Smith-Waterman score: 268;  32.468% identity (60.390% similar) in 154 aa
overlap (8-159:941-1093)

                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                :: :::::. :...: :: :  : ..:::: 
gi|147 ESYLVDRHFVIKTDHQSLKYLLEQRVTTPTQQAWVAKLMQYDYEISYKQGKENVAADALS
              920       930       940       950       960       970

       40        50          60        70        80        90
5_3    RRDEDKELQGISRPFWKD--ITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
       :  .  ::  .:  . .      :.:    :: : :: .  . . ...:.:. . . :
gi|147 RI-QPAELFVLSTTILNTQLYDLIKESWGVDPELQKIIKAKEADPSAYPKYSWRGEELRR
               980       990      1000      1010      1020

        100       110       120       130       140       150
5_3    RGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVRE
       .:.::. ..     ..:. :. :  :::::.:.: .::.  .::  ..  . ..: .
gi|147 KGKLVVGVNEQLRREILNSFHDSPTGGHSGVYVTTKRISAIVYWKGLRKFVREYVRNCYG
    1030      1040      1050      1060      1070      1080

        160
5_3    IYMDQQ
       .. :
gi|147 VFTDITMDFIEGLPKSNGKTAIFVVVDRLTKYGHFMLLPHPYTAKMVAQVFLDSVYKLHG
    1090      1100      1110      1120      1130      1140

>>gi|108864659|gb|ABA95357.2| retrotransposon protein, p  (2811 aa)
 initn: 252 init1: 161 opt: 270  Z-score: 329.0  bits: 71.3 E(): 2.4e-09
Smith-Waterman score: 270;  33.557% identity (63.758% similar) in 149 aa
overlap (8-152:2240-2386)

                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                ::  ..:::: .. : :: :  : ..:::: 
gi|108 RSYLQHDEFVIRTDHRSLSFLTNQRLSTPWQQKALTKLLGLRYKICYKKGLENGAADALS
    2210      2220      2230      2240      2250      2260

       40           50        60        70        80        90
5_3    RRDED--KELQGIS--RPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       ::. :  .:...::   : :  . .. .  :.:    :. . :. .  .  .. .. ::
gi|108 RRQSDGLEEVSAISICLPDW--LQELIDGYQSDSEAQKLLQALSVSGAGPSNFEVQNGIL
    2270      2280      2290      2300      2310      2320

          100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
       ::. :. .  ..:   :.: ...:. .:::::: .::.:. : . :  .. .. .:: :
gi|108 YFKHRIWIGHNKLLQQKILANLHTAAVGGHSGILVTYQRVKQLFSWPGMRKDVQEFVQHC
      2330      2340      2350      2360      2370      2380

          160
5_3    REIYMDQQ

gi|108 DICQRAKSEHVKYPGLLQPLEVPSQSWQVITMDFIEGLPKSASFDCILVIVDKFSKFAHF
      2390      2400      2410      2420      2430      2440

>>gi|113645699|dbj|BAF28840.1| Os11g0686400 [Oryza sativ  (2866 aa)
 initn: 252 init1: 161 opt: 270  Z-score: 328.9  bits: 71.3 E(): 2.4e-09
Smith-Waterman score: 270;  33.557% identity (63.758% similar) in 149 aa
overlap (8-152:2295-2441)

                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                ::  ..:::: .. : :: :  : ..:::: 
gi|113 RSYLQHDEFVIRTDHRSLSFLTNQRLSTPWQQKALTKLLGLRYKICYKKGLENGAADALS
        2270      2280      2290      2300      2310      2320

       40            50        60        70        80        90
5_3    RRDED--KELQGIS--RPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       ::. :  .:...::   : :  . .. .  :.:    :. . :. .  .  .. .. ::
gi|113 RRQSDGLEEVSAISICLPDW--LQELIDGYQSDSEAQKLLQALSVSGAGPSNFEVQNGIL
         2330      2340      2350      2360      2370      2380

          100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
       ::. :. .  ..:   :.: ...:. .:::::: .::.:. : . :  .. .. .:: :
gi|113 YFKHRIWIGHNKLLQQKILANLHTAAVGGHSGILVTYQRVKQLFSWPGMRKDVQEFVQHC
          2390      2400      2410      2420      2430      2440

          160
5_3    REIYMDQQ

gi|113 DICQRAKSEHVKYPGLLQPLEVPSQSWQVITMDFIEGLPKSASFDCILVIVDKFSKFAHF
          2450      2460      2470      2480      2490      2500

>>gi|77554607|gb|ABA97403.1| retrotransposon protein, pu  (562 aa)
 initn: 230 init1: 230 opt: 259  Z-score: 325.7  bits: 68.3 E(): 3.6e-09
Smith-Waterman score: 259;  35.862% identity (58.621% similar) in 145 aa
overlap (8-150:204-346)

                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
```

```
                                      :.  . :::.:.. : :: :  ::..::::
gi|775 NYTQPFILLADASDQASLKYINEQRLTKGIQHKLLIKLLNYNYKIEYKKGKENKAADALS
           180       190       200       210       220       230

        40          50        60        70        80        90
5_3    RRDEDKEL--QGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
       :     .     :  : :  : .:      ::  . .. .:. . ..:::.:::   :: .
gi|775 RIPFVAQWFSTTIIVPTW--IIEILASYATDPKCTALESQLRITPQGHPSYTLTGGILRY
           240       250         260       270       280       290

         100       110       120       130       140       150
5_3    RGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVRE
       ..:: . :..    :: : :. : .:::::   ::.:    .::  .: .:...:
gi|775 KNRLYVGAGTDLRAKLQQSFHDSALGGHSGERATYQRAKLLFYWPGMKKDIASYVKLCPV
              300       310       320       330       340       350

         160
5_3    IYMDQQ

gi|775 CQKNKSEHNLQPGLLHPLPVPEMAWTHISMDFIEGLPKSDNKDVIWVIVDRFTKYAHFVA
              360       370       380       390       400       410

>>gi|77550700|gb|ABA93497.1| retrotransposon protein, pu  (1448 aa)
 initn: 196 init1: 105 opt: 263  Z-score: 324.6  bits: 69.5 E(): 4.2e-09
Smith-Waterman score: 263;  34.591% identity (60.377% similar) in 159 aa
overlap (8-161:862-1014)

                                       10        20        30
5_3                                ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                   ::   .:::: .: : :: :  : ..::::
gi|775 RSYLQYAEFLILTDHKSLMNLTDQRLHTSWQQKAYTKLLGLKFKIYYKKGIHNGAADALS
              840       850       860       870       880       890

        40            50        60        70         80        90
5_3    RRDEDKELQ----GISRPFWKDITKINEEVQKDPALAKIREELK-DNLDSHPQYTLECDI
       : :. . :.    .: .: :  . .. :   .:   ::.  .:. .::.. : : :.  .
gi|775 RCDHGEPLEVAAISICKPTW--LEEVAEGYLHDAKSAKLLAQLSIQNLEDSP-YKLKDGL
              900       910         920       930       940

            100       110       120       130       140       150
5_3    LYFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVH
       . ..::. :  .:    :... .. : .:::::. .:::::   . :  .: .:    ..
gi|775 IRYKGRIWLGHNSALQNKIFSALHDSSIGGHSGFPVTYRRIKTLFAWPDMKKQIK---LK
      950       960       970       980       990      1000

           160
5_3    VREIYMDQQ
       :.:  . ::
gi|775 VKECSICQQAKPDRFRYPGLLQPLLVPSGAWQTVTMDFIEGLPKSRRFNCIMVVVDKLSR
        1010      1020      1030      1040      1050      1060

>>gi|62734016|gb|AAX96125.1| retrotransposon protein, pu  (1448 aa)
 initn: 196 init1: 105 opt: 263  Z-score: 324.6  bits: 69.5 E(): 4.2e-09
Smith-Waterman score: 263;  34.591% identity (60.377% similar) in 159 aa
overlap (8-161:862-1014)

                                       10        20        30
5_3                                ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                   ::   .:::: .: : :: :  : ..::::
gi|627 RSYLQYAEFLILTDHKSLMNLTDQRLHTSWQQKAYTKLLGLKFKIYYKKGIHNGAADALS
              840       850       860       870       880       890

        40            50        60        70         80        90
5_3    RRDEDKELQ----GISRPFWKDITKINEEVQKDPALAKIREELK-DNLDSHPQYTLECDI
       : :. . :.    .: .: :  . .. :   .:   ::.  .:. .::.. : : :.  .
gi|627 RCDHGEPLEVAAISICKPTW--LEEVAEGYLHDAKSAKLLAQLSIQNLEDSP-YKLKDGL
              900       910         920       930       940

            100       110       120       130       140       150
5_3    LYFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVH
       . ..::. :  .:    :... .. : .:::::. .:::::   . :  .: .:    ..
gi|627 IRYKGRIWLGHNSALQNKIFSALHDSSIGGHSGFPVTYRRIKTLFAWPDMKKQIK---LK
      950       960       970       980       990      1000

           160
5_3    VREIYMDQQ
       :.:  . ::
gi|627 VKECSICQQAKPDRFRYPGLLQPLLVPSGAWQTVTMDFIEGLPKSRRFNCIMVVVDKLSR
        1010      1020      1030      1040      1050      1060

>>gi|15451607|gb|AAK98731.1|AC090485_10 Putative retroel  (1461 aa)
 initn: 189 init1: 140 opt: 263  Z-score: 324.5  bits: 69.5 E(): 4.2e-09
Smith-Waterman score: 263;  34.932% identity (63.699% similar) in 146 aa
overlap (8-150:896-1036)

                                       10        20        30
5_3                                ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                   :..::.::::..: . :: ::.: :.::::
gi|154 RPYLWGRHFTVKTDHYSLKYLLDQRLSTIPQHHWVGKLLGFDFTVEYKPGAANTVADALS
         870       880       890       900       910       920

        40          50        60        70        80        90
5_3    RRD--EDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
       :::  ::  .  .: : .  : .. .  . ::::. .. :....  . : ...   .. :
gi|154 RRDTTEDASVLVLSAPRFDFIERLRQAQDVDPALVALQAEIRSGTRAGP-WSMADGMVLF
         930       940       950       960       970       980

         100       110       120       130       140       150
5_3    RGRLVLLASSLWIPKLLQEFQTSLMG-GHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
       ::: :  .:   : ::::   ..   :: :.  : .:. .....  .:. .  ::
```

```
gi|154 AGRLYLPPAS---P-LLQEVLRAVHEEGHEGVQRTLHRLRRDFHFPNMKSVVQDFVRTCE
          990          1000      1010      1020      1030      1040

         160
5_3    EIYMDQQ

gi|154 VCQRYKAEHLQPAGLLLPLPVPQGVWTDVALDFVEALPRVRGKSVILTVVDRFSKYCHFI
             1050      1060      1070      1080      1090      1100

>>gi|108706172|gb|ABF93967.1| retrotransposon protein, p  (1461 aa)
 initn: 189 init1: 140 opt: 263  Z-score: 324.5  bits: 69.5 E(): 4.2e-09
Smith-Waterman score: 263;  34.932% identity (63.699% similar) in 146 aa
overlap (8-150:896-1036)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 :..::.::::..: . :: ::.: :.::::
gi|108 RPYLWGRHFTVKTDHYSLKYLLDQRLSTIPQHHWVGKLLGFDFTVEYKPGAANTVADALS
         870       880       890       900       910       920

        40          50        60        70        80        90
5_3    RRD--EDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
       :::  ::  .  .: : .  : .. .  . ::::. .. :.... . : ...   .. :
gi|108 RRDTTEDASVLVLSAPRFDFIERLRQAQDVDPALVALQAEIRSGTRAGP-WSMADGMVLF
         930       940       950       960       970       980

         100       110       120       130       140       150
5_3    RGRLVLLASSLWIPKLLQEFQTSLMG-GHSGIYITYRRITQSLYWIPIKGEITKFVVHVR
        ::: :  .:   : ::::   ..   :: :.  : .:. .....  .:. .  ::
gi|108 AGRLYLPPAS---P-LLQEVLRAVHEEGHEGVQRTLHRLRRDFHFPNMKSVVQDFVRTCE
          990          1000      1010      1020      1030      1040

         160
5_3    EIYMDQQ

gi|108 VCQRYKAEHLQPAGLLLPLPVPQGVWTDVALDFVEALPRVRGKSVILTVVDRFSKYCHFI
             1050      1060      1070      1080      1090      1100

>>gi|90399077|emb|CAJ86299.1| H0124B04.16 [Oryza sativa   (1265 aa)
 initn: 253 init1: 162 opt: 261  Z-score: 322.9  bits: 69.0 E(): 5.2e-09
Smith-Waterman score: 261;  33.333% identity (62.585% similar) in 147 aa
overlap (8-152:1026-1172)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 ::  ..:::: .. : :: :  : ..::::
gi|903 RSYLQHDEFVIRTDHRSLSFLINQRLSTPWQQKALTKLLGLRYKICYKKGLENGAADALS
        1000      1010      1020      1030      1040      1050

        40          50        60        70        80        90

5_3    RRDED--KELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
       ::. :  .::..::  .   . ..    :.:    :. . :. .  .  .. .   ::::
gi|903 RRQSDGLEELSAISICLSDWLQELIAGYQSDSEAQKLLQALSVSGAGPSNFEVLNGILYF
        1060      1070      1080      1090      1100      1110

         100       110       120       130       140       150
5_3    RGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVRE
       . :. .  ..:   :.: ...:. .:::::: .::.:. : . :  .. .. .:: :
gi|903 KHRIWIGHNKLLQQKILANLHTTAVGGHSGILVTYQRVKQLFSWPGLRKDVQEFVQHCDI
        1120      1130      1140      1150      1160      1170

         160
5_3    IYMDQQ

gi|903 CQRAKSEHVKYPGLLQPLEVPSQSWQVITMDFIEGLPRSASFDCILVIVDKFSKFALFFT
        1180      1190      1200      1210      1220      1230

>>gi|113611293|dbj|BAF21671.1| Os07g0510800 [Oryza sativ  (517 aa)
 initn: 216 init1: 131 opt: 256  Z-score: 322.5  bits: 67.6 E(): 5.5e-09
Smith-Waterman score: 256;  28.571% identity (65.306% similar) in 147 aa
overlap (8-150:358-502)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 ::  ..:::: .. .::: : .:...:::
gi|113 RSYLQFGEFVIRTDQRSLIHLGDQKLATPWQQKAMTKLLGLNYRLVYKRGLDNRAADALS
       330       340       350       360       370       380

        40            50        60        70        80        90
5_3    RRDEDKELQG----ISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       : . : ..:.    .  : :  .:...:   .::  : .  ..  . .. :...:.  ::
gi|113 RCSTDDHIQACALSVCIPDW--LTEVQEGYLSDPYSADLLSKVTLQASAVPNFSLRDGIL
       390       400         410       420       430       440

           100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
       ...... .  ..    ..:. .... .:::::. .:: :: . . :  .:  . .:.
gi|113 HYKNKIWIGNNTTLQHRILSALHSGAIGGHSGVQVTYSRIKKLFAWQGLKKSVLEFIDQC
         450       460       470       480       490       500

          160
5_3    REIYMDQQ

gi|113 SVCKQAKAERVH
         510

>>gi|113639856|dbj|BAF27161.1| Os10g0551400 [Oryza sativ  (1189 aa)
 initn: 273 init1: 151 opt: 260  Z-score: 322.1  bits: 68.7 E(): 5.8e-09
Smith-Waterman score: 260;  31.034% identity (66.207% similar) in 145 aa
overlap (8-148:934-1076)
```

```
                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                ::   .:::: .. :::. :..:...:::: 
gi|113 RSYLQHQEFMILTDHHSLTHLSDQRLHTPWQQKAFTKLLGLQYRIVYRKGSANSAADALS
          910       920       930       940       950       960

        40          50          60        70        80        90
5_3    RRD--EDKELQGISR--PFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       :.:  .. .. ..:   : :  . .. .  ..:   ...  ::. :  .. .:::.  ..
gi|113 RKDLGDSAQILAVSSCSPSW--LQEVIQGYEQDKFSSQLLAELSLNPKAREHYTLQQGLI
          970       980         990      1000      1010      1020

          100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
        ..::. .  ..    ::.:.. .  :::::. .::::: . . :. .: .: .
gi|113 RYKGRIWVGNNTDLQLKLIKELHDNPAGGHSGFPVTYRRIKHLFAWLGMKQQIQQQLKQC
          1030      1040      1050      1060      1070      1080

          160
5_3    REIYMDQQ

gi|113 QICQQAKPERVKYPGLLQPLPVPKGAWQVISMDFIEGLPTSDKYNCILVVVDKFSKYAHF
          1090      1100      1110      1120      1130      1140
```

>>gi|113533835|dbj|BAF06218.1| Os01g0758700 [Oryza sativ (225 aa)
initn: 244 init1: 169 opt: 251 Z-score: 321.6 bits: 66.2 E(): 6.1e-09
Smith-Waterman score: 251; 32.886% identity (67.114% similar) in 149 aa
overlap (8-150:32-177)

```
                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                :.  ..:::: .. :.:: :. : ..:::: 
gi|113 WIIGAHAEFFIRTDHRSLSFLDDQRLTTPWQHKALTKLLGLRYKIIYKKGTENGAADALS
             10        20        30        40        50        60

        40          50          60        70        80        90
5_3    R-RDEDK--ELQGISR--PFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDI
       :  . ::  :: .::   : :  : .. .  ..::  ..  . :  .  . :..::.  .
gi|113 RCPSADKVFELTAISSVIPNW--IQEVVDGYMSDPEASSKVQTLCISPAAVPDFTLKDGV
             70        80          90       100       110

          100       110       120       130       140       150
5_3    LYFRGRLVLLASSLWIP-KLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVV
       :::.. .. ..... .  :.: ...:. .:::::::..::.:. : . :  ... . .::
gi|113 LYFKN-IMWIGNNVQVQQKILANLHTAPVGGHSGIHVTYQRVKQLFAWPHLRSTVMQFVN
     120       130       140       150       160       170

            160
5_3    HVREIYMDQQ
```

```
gi|113 SCSICQQAKSEHVKYPGMLQPLPVPEHAWFWPKLEKWLINWLCRPLV
      180       190       200       210       220
```

>>gi|110289541|gb|AAP54937.2| retrotransposon protein, p (1477 aa)
initn: 273 init1: 151 opt: 260 Z-score: 320.7 bits: 68.8 E(): 6.9e-09
Smith-Waterman score: 260; 31.034% identity (66.207% similar) in 145 aa
overlap (8-148:975-1117)

```
                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                ::   .:::: .. :::. :..:...:::: 
gi|110 RSYLQHQEFMILTDHHSLTHLSDQRLHTPWQQKAFTKLLGLQYRIVYRKGSANSAADALS
          950       960       970       980       990      1000

        40          50          60        70        80        90
5_3    RRD--EDKELQGISR--PFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       :.:  .. .. ..:   : :  . .. .  ..:   ...  ::. :  .. .:::.  ..
gi|110 RKDLGDSAQILAVSSCSPSW--LQEVIQGYEQDKFSSQLLAELSLNPKAREHYTLQQGLI
         1010      1020        1030      1040      1050      1060

           100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
        ..::. .  ..    ::.:.. .  :::::. .::::: . . :. .: .: .
gi|110 RYKGRIWVGNNTDLQLKLIKELHDNPAGGHSGFPVTYRRIKHLFAWLGMKQQIQQQLKQC
           1070      1080      1090      1100      1110      1120

           160
5_3    REIYMDQQ

gi|110 QICQQAKPERVKYPGLLQPLPVPKGAWQVISMDFIEGLPTSDKYNCILVVVDKFSKYAHF
           1130      1140      1150      1160      1170      1180
```

>>gi|116310096|emb|CAH67116.1| H0502G05.7 [Oryza sativa (642 aa)
initn: 254 init1: 190 opt: 255 Z-score: 319.8 bits: 67.4 E(): 7.7e-09
Smith-Waterman score: 255; 33.103% identity (59.310% similar) in 145 aa
overlap (8-150:37-181)

```
                                    10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                :.   .::::  . : :. :. :  .:::: 
gi|116 RSYLQMGEFIILTDHHSLMHLSDQRLHTPWQHKAFTKLLGLSYRICYRKGTCNGPADALS
          10        20        30        40        50        60

          40        50        60        70        80        90
5_3    RR--DEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
       :.  : . ::  ::      : ....  ..::  ...  ::  :  .. ..::.  .. :
gi|116 RKFQDTEDELCHISACTLTWIQEVTDGYKQDPFSTQLLTELAVNATGRKHFTLNSGLIRF
          70        80        90       100       110       120
```

```
         100       110       120       130       140       150
5_3    RGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVRE
       .::. .  .     ::: :...: .:::::. .:::.. : . :  .:     ::
gi|116 KGRVWIGDNPTLQSKLLTELHSSPIGGHSGFPVTYRKLKQLFAWPKMKKMTKTFVQQCQI
        130       140       150       160       170       180

         160
5_3    IYMDQQ

gi|116 CLQAKPDRARYPGLLQPLPVPEGAWQVISLDFIEGLPRSDHSNCILVVVDKFSKYAHFLP
        190       200       210       220       230       240

>>gi|38567865|emb|CAE03018.3| OSJNBa0091D06.3 [Oryza sat  (753 aa)
 initn: 254 init1: 190 opt: 255  Z-score: 318.8  bits: 67.5 E(): 8.8e-09
Smith-Waterman score: 255;  33.103% identity (59.310% similar) in 145 aa
overlap (8-150:37-181)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 :.   .::::  . : :. :. :  .::::
gi|385 RSYLQMGEFIILTDHHSLMHLSDQRLHTPWQHKAFTKLLGLSYRICYRKGTCNGPADALS
         10        20        30        40        50        60

          40        50        60        70        80        90
5_3    RR--DEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYF
       :.  : . ::  ::      : ....  ..::  ...  ::  :  .. ..:::.  .. :
gi|385 RKFQDTEDELCHISACTLTWIQEVTDGYKQDPFSTQLLTELAVNATGRKHFTLNSGLIRF
         70        80        90       100       110       120

         100       110       120       130       140       150
5_3    RGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVRE
       .::. .  .     ::: :...: .:::::. .:::.. : . :  .:     ::
gi|385 KGRVWIGDNPTLQSKLLTELHSSPIGGHSGFPVTYRKLKQLFAWPKMKKMTKTFVQQCQI
        130       140       150       160       170       180

         160
5_3    IYMDQQ

gi|385 CLQAKPDRARYPGLLQPLPVPEGAWQVISLDFIEGLPRDKVFTSQLWEHLFEGAVQATPF
        190       200       210       220       230       240

>>gi|15217201|gb|AAK92545.1|AC051624_3 Putative retroele  (1476 aa)
 initn: 182 init1: 160 opt: 257  Z-score: 316.9  bits: 68.1 E(): 1.1e-08
Smith-Waterman score: 257;  33.333% identity (64.583% similar) in 144 aa
overlap (8-150:911-1050)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 :..::.::::..: . :: :::: :.::::
gi|152 RPYLWGRAFVVRTDHYSLKYLLDQRLATIPQHHWVGKLLGFDFTVEYKSGASNVVADALS
              890       900       910       920       930       940

         40        50        60        70        80        90
5_3    RRDEDK-ELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFR
       ::: :.  . ..: : .  : ..      .:::. ::. .. .  : : ..:.  ...:
gi|152 RRDTDEGAVLALSAPRFDYIERLRAAQTTEPALVAIRDAIQAGTRSAP-WALRDGMVMFD
              950       960       970       980       990

         100       110       120       130       140       150
5_3    GRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREI
       .:: .  ::  . ..:  ..:.   :: :.  : .:. ....   ..  . .::
gi|152 SRLYIPPSSPLLHEILAAIHTD---GHEGVQRTLHRLRRDFHSPAMRRVVQEFVRACDTC
    1000      1010      1020         1030      1040      1050

         160
5_3    YMDQQ

gi|152 QRNKSEHLHPGGLLLPLPVPTTVWADIGLDFVEALPRVGGKTVILTVVDRFSKYCHFIPL
       1060      1070      1080      1090      1100      1110



161 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:25:40 2009 done: Thu Mar 19 12:43:12 2009
 Total Scan time: 1029.940 Total Display time:  1.420

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

Appendix 4. Bioinformatic analysis of polypeptide 5_4

```
>5_4
TTNLVISPFIGIQ
```

```
Sliding 8 amino acid window search
Database searched = AD_2009
Query = 5_4

Start time: Thu Mar 19 12:43:14 CDT 2009 Finish time: Thu Mar 19 12:43:14 CDT
2009

No 8 amino acid matches exist between 5_4 and the AD_2009 database
```

```
# fasta34 5_4.pep /home/ht/db/AD_2009 -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, 13 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20     3     0:=
  22     0     0:           one = represents 3 library sequences
  24     3     0:=
  26     4     0:==
  28     9     0:===
  30    19     2:*======
  32    45     7:==*============
  34    31    20:======*====
  36    30    41:==========   *
  38    38    68:=============         *
  40    91    95:================================*
  42    71   116:=======================               *
  44    74   128:=========================                  *
  46   141   130:============================================*===
  48   119   125:========================================== *
  50   126   114:===================================*====
  52    92   100:==============================  *
  54    96    85:============================*===
  56    71    71:======================*
  58    72    59:==================*====
  60    74    47:===============*=========
  62    37    38:============*
  64    40    30:=========*====
  66    30    24:=======*==
  68    17    19:======*
  70    10    15:====*
  72    13    12:===*=
  74     5     9:==*
  76     6     7:==*
  78     4     5:=*
  80     6     4:=*
  82     0     3:*
  84     0     3:*
  86     3     2:*
  88     4     2:*=         inset = represents 1 library sequences
  90     0     1:*
  92     1     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     1     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:          *
 307888 residues in  1386 sequences
  Expectation_n fit: rho(ln(x))= 3.30100.0025; mu= 2.8031 0.130
 mean_var=13.9285 3.540, 0's: 3 Z-trim: 3  B-trim: 106 in 1/43
 Lambda= 0.343654
 Kolmogorov-Smirnov  statistic: 0.0588 (N=27) at  34

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
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:43:13 2009 done: Thu Mar 19 12:43:14 2009
 Total Scan time:  0.040 Total Display time:  0.000

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

```
# fasta34 5_4.pep /home/ht/db/TOX_2009 -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, 13 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    59     0:====
  22     3     0:=          one = represents 16 library sequences
  24     1     0:=
  26     8     0:=
  28    14     2:*
  30    21    11:*=
  32    63    41:==*=
  34   240   110:======*========
  36   200   227:============ *
  38   285   375:==================     *
```

```
  40   343   523:======================          *
  42   530   639:================================     *
  44   954   705:================================================*===============
  46   692   718:=============================================*
  48   654   688:============================================ *
  50   527   627:================================      *
  52   397   552:=========================          *
  54   295   471:===================          *
  56   585   394:=======================*============
  58   488   323:====================*==========
  60   206   262:=============    *
  62   125   210:========     *
  64   253   167:==========*=====
  66   172   132:========*==
  68   205   104:======*======
  70    66    81:=====*
  72    87    64:===*==
  74    60    50:===*
  76    24    39:==*
  78    16    30:=*
  80    19    23:=*
  82     5    18:=*
  84     4    14:*
  86    10    11:*
  88     1     8:*           inset = represents 1 library sequences
  90    23     7:*=
  92     6     5:*          :====*=
  94     4     4:*          :===*
  96     0     3:*          :  *
  98     0     2:*          : *
 100     1     2:*          :=*
 102     0     1:*          :*
 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:           *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 4.18870.000497; mu= -0.6248 0.025
 mean_var=16.2088 3.503, 0's: 59 Z-trim: 59  B-trim: 426 in 1/61
 Lambda= 0.318566
 Kolmogorov-Smirnov  statistic: 0.0609 (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


13 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:43:14 2009 done: Thu Mar 19 12:43:14 2009
 Total Scan time:  0.130 Total Display time:  0.000

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

```
# fasta34 5_4.pep /home/ht/db/PRT_2009 -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, 13 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 223245     0:===========
  22   982     0:=          one = represents 21567 library sequences
  24  2182    14:*
  26  6447   309:*
  28 17567  3337:*
  30 49911 20272:*==
  32 119414 78384:===*==
  34 246596 212567:=========*==
  36 429009 436564:====================*
  38 666955 721477:==============================  *
  40 891996 1006397:=============================================    *
  42 1078962 1230198:=====================================================
*
  44 1215886
1357023:=========================================================  *
  46 1294018
1382162:============================================================*
  48 1289252
1323260:============================================================*
  50 1206254 1207480:=============================================================*
  52 1088638 1061577:=================================================*=
  54 921557 906773:=========================================*
  56 796707 757433:====================================*=
  58 670059 621838:============================*===
  60 545445 503725:=======================*==
  62 449356 403838:==================*==
  64 356018 321170:===============*==
  66 267530 253843:===========*=
  68 214297 199668:=========*
```

```
  70 159553 156471:=======*
  72 121782 122268:=====*
  74 89818 95328:====*
  76 73624 74196:===*
  78 58174 57671:==*
  80 41874 44781:==*
  82 33001 34256:=*
  84 22396 27135:=*
  86 15728 20996:*
  88 12787 16246:*          inset = represents 163 library sequences
  90  9723 12570:*
  92  8102  9726:*         :=======================================*
  94  6029  7525:*         :====================================  *
  96  4569  5823:*         :============================      *
  98  2724  4505:*         :=================          *
 100  1687  3486:*         :===========          *
 102  1443  2697:*         :=========       *
 104  1129  2087:*         :=======     *
 106   684  1615:*         :=====    *
 108   437  1249:*         :===    *
 110   279   967:*         :==   *
 112   192   748:*         :==  *
 114   109   579:*         :=  *
 116    72   448:*         := *
 118    62   347:*         := *
>120   157   268:*         :=*
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714215 sequences
  Expectation_n fit: rho(ln(x))= 3.17800.000168; mu= 4.6625 0.009
 mean_var=16.6209 3.292, 0's: 911 Z-trim: 919  B-trim: 0 in 0/65
 Lambda= 0.314591
 Kolmogorov-Smirnov  statistic: 0.0320 (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


13 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:43:14 2009 done: Thu Mar 19 12:48:10 2009
 Total Scan time: 272.920 Total Display time:  0.000

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

Appendix 5. Bioinformatic analysis of polypeptide 5_5

```
>5_5
PSYSLLIHVDFPDMNHKLSDFSLYWYPIKRLSDSSISYVNTRVPSHKRSLEFL
_____________________________________________________


Sliding 8 amino acid window search
Database searched = AD_2009
Query = 5_5

Start time: Thu Mar 19 12:48:12 CDT 2009 Finish time: Thu Mar 19 12:48:12 CDT
2009

No 8 amino acid matches exist between 5_5 and the AD_2009 database
_____________________________________________________


# fasta34 5_5.pep /home/ht/db/AD_2009 -Q -E 1 -O 5_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

 5_5, 53 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20    14     0:=====
  22     0     0:           one = represents 3 library sequences
  24     1     0:=
  26     0     0:
  28     0     0:
  30     5     2:*=
  32    20     7:==*====
  34    16    20:======*
  36    26    41:=========    *
  38    70    68:======================*=
  40    98    95:================================*=
  42   131   116:=========================================*=====
  44   135   128:============================================*==
  46   117   130:=========================================    *
  48   143   125:============================================*======
  50    92   114:==============================       *
  52    73   100:=========================        *
  54    84    85:============================*
  56    56    71:==================    *
  58    51    59:================  *
  60    53    47:===============*==
  62    54    38:============*=====
  64    31    30:=========*=
  66    26    24:=======*=
  68    15    19:===== *
  70    20    15:====*==
  72    12    12:===*
```

```
  74    13     9:==*==
  76     4     7:==*
  78     2     5:=*
  80     4     4:=*
  82     6     3:*=
  84     2     3:*
  86     2     2:*
  88     3     2:*          inset = represents 1 library sequences
  90     1     1:*
  92     1     1:*         :*
  94     0     1:*         :*
  96     1     1:*         :*
  98     1     0:=         *=
 100     0     0:          *
 102     1     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:          *
 307888 residues in  1386 sequences
  Expectation_n fit: rho(ln(x))= 4.21420.00353; mu= 3.4229 0.181
 mean_var=34.0969 9.817, 0's: 14 Z-trim: 14  B-trim: 78 in 2/40
 Lambda= 0.219643
 Kolmogorov-Smirnov  statistic: 0.0314 (N=29) at  58

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(1386)
gi|75062228|sp|Q5VFH6.1|ALL4_FELCA RecName: Full=A ( 186)   57 24.1    0.76
gi|729979|sp|P39673.1|MAG_DERFA Allergen Mag       ( 341)   59 24.8    0.86


>>gi|75062228|sp|Q5VFH6.1|ALL4_FELCA RecName: Full=Aller  (186 aa)
 initn:  57 init1:  57 opt:  57  Z-score: 104.0  bits: 24.1 E(): 0.76
Smith-Waterman score: 57;  29.730% identity (67.568% similar) in 37 aa
overlap (3-39:36-72)

                                        10        20        30
5_5                                PSYSLLIHVDFPDMNHKLSDFSLYWYPIKRLS
                                     ::.:.  :  .  .. ... ..   :: :.
gi|750 LCLGLILVCAHEEENVVRSNIDISKISGEWYSILLASDVKEKIEENGSMRVFVEHIKALD
          10        20        30        40        50        60

             40        50
5_5    DSSISYVNTRVPSHKRSLEFL
       .::.:.:
gi|750 NSSLSFVFHTKENGKCTEIFLVADKTKDGVYTVVYDGYNVFSIVETVYDEYILLHLLNFD
          70        80        90       100       110       120

>>gi|729979|sp|P39673.1|MAG_DERFA Allergen Mag            (341 aa)
 initn:  34 init1:  34 opt:  59  Z-score: 103.1  bits: 24.8 E(): 0.86
Smith-Waterman score: 59;  27.500% identity (60.000% similar) in 40 aa
overlap (8-46:164-203)

                                       10        20        30
5_5                               PSYSLLIHVDFP-DMNHKLSDFSLYWYPIKRLSDSSI
                                          ..::  . :..  : .:   :.:  . .
gi|729 DQINMDIDGTLIEGHAQGTIREGKIHIKGRQTDFEIESNYRYEDGKLIIEPVKSENGKLE
           140       150       160       170       180       190

         40        50
5_5    SYVNTRVPSHKRSLEFL
       . .. .:::: 
gi|729 GVLSRKVPSHLTLETPRVKMNMKYDRYAPVKVFKLDYDGIHFEKHTDIEYEPGVRYKIIG
           200       210       220       230       240       250



53 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:48:12 2009 done: Thu Mar 19 12:48:12 2009
 Total Scan time:  0.040 Total Display time:  0.000

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

```
# fasta34 5_5.pep /home/ht/db/TOX_2009 -Q -E 1 -O 5_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

 5_5, 53 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    81     0:=======
  22     0     0:          one = represents 13 library sequences
  24     1     0:=
  26     1     0:=
  28     3     2:*
  30     6    11:*
  32    40    41:===*
```

```
  34    53   110:=====   *
  36   140   227:===========      *
  38   444   375:============================*======
  40   594   523:=======================================*=====
  42   466   639:==================================             *
  44   574   705:============================================         *
  46   648   718:=================================================     *
  48   645   688:================================================  *
  50   731   627:================================================*========
  52   705   552:=========================================*============
  54   390   471:==============================      *
  56   268   394:====================         *
  58   337   323:========================*=
  60   288   262:====================*==
  62   185   210:================ *
  64   141   167:=========== *
  66   121   132:==========*
  68   153   104:=======*====
  70    92    81:======*=
  72    83    64:====*==
  74   197    50:===*============
  76    66    39:==*===
  78    34    30:==*
  80    39    23:=*=
  82    42    18:=*==
  84     4    14:=*
  86    12    11:*
  88    12     8:*          inset = represents 1 library sequences
  90    14     7:*=
  92    10     5:*         :====*=====
  94     4     4:*         :===*
  96     3     3:*         :==*
  98     6     2:*         :=*====
 100     7     2:*         :=*=====
 102     2     1:*         :*=
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     3     1:*         :*==
 112     1     0:=         *=
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 4.72560.000702; mu= 0.7612 0.034
 mean_var=34.5258 8.105, 0's: 80 Z-trim: 81  B-trim: 736 in 2/59
 Lambda= 0.218274
 Kolmogorov-Smirnov  statistic: 0.0560 (N=29) at  48
```

```
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


53 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:48:12 2009 done: Thu Mar 19 12:48:13 2009
 Total Scan time:  0.160 Total Display time:  0.000

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

```
# fasta34 5_5.pep /home/ht/db/PRT_2009 -Q -E 1 -O 5_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

 5_5, 53 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 289725     0:=============
  22   136     0:=          one = represents 23586 library sequences
  24   217    14:*
  26   641   309:*
  28  1735  3337:*
  30  8380 20272:*
  32 39500 78383:== *
  34 135115 212566:======   *
  36 332948 436560:===============   *
  38 661101 721472:============================= *
  40 1029531 1006389:=========================================*=
  42 1308969 1230188:============================================================*===
  44 1415113
1357012:============================================================*==
  46 1363568
1382151:==========================================================*
  48 1248323 1323250:==========================================================
*
  50 1077194 1207471:=============================================     *
  52 952854 1061569:=========================================    *
  54 811191 906766:==================================   *
  56 713919 757427:============================== *
  58 618428 621833:=========================*
  60 516844 503722:======================*
  62 456183 403835:=================*==
  64 361492 321168:=============*==
```

```
  66 290503 253841:==========*==
  68 226466 199666:========*=
  70 181040 156470:======*=
  72 139513 122267:=====*
  74 113703 95327:====*
  76 101379 74195:===*=
  78 85784 57671:==*=
  80 60233 44781:=*=
  82 51475 34256:=*=
  84 30304 27135:=*
  86 25379 20996:*=
  88 17794 16245:*          inset = represents 199 library sequences
  90 12861 12570:*
  92  9914  9726:*         :=====================================*
  94  6762  7525:*         :================================   *
  96  4698  5823:*         :=======================     *
  98  3757  4505:*         :==================   *
 100  2658  3486:*         :==============   *
 102  1908  2697:*         :==========   *
 104  1360  2087:*         :=======   *
 106  1045  1615:*         :======  *
 108   792  1249:*         :====  *
 110   540   967:*         :=== *
 112   417   748:*         :===*
 114   303   579:*         :==*
 116   198   448:*         := *
 118   150   347:*         :=*
>120   375   268:*         :=*
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714102 sequences
  Expectation_n fit: rho(ln(x))= 4.32410.000189; mu= 4.4250 0.010
 mean_var=47.2757 9.685, 0's: 1224 Z-trim: 1226  B-trim: 0 in 0/63
 Lambda= 0.186533
 Kolmogorov-Smirnov  statistic: 0.0422 (N=29) at  58

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


53 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:48:13 2009 done: Thu Mar 19 12:54:48 2009
 Total Scan time: 378.370 Total Display time:  0.000

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

Appendix 6. Bioinformatic analysis of polypeptide 5_6

```
>5_6
KKIGNYSFFFSILTIILIADPCRFPGHEPQT
_______________________________________________________


Sliding 8 amino acid window search
Database searched = AD_2009
Query = 5_6

Start time: Thu Mar 19 12:54:51 CDT 2009 Finish time: Thu Mar 19 12:54:51 CDT
2009

No 8 amino acid matches exist between 5_6 and the AD_2009 database
_______________________________________________________


# fasta34 5_6.pep /home/ht/db/AD_2009 -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, 31 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20     3     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     1     0:=
  28     5     0:==
  30    14     2:*====
  32    23     7:==*=====
  34    50    20:======*==========
  36    70    41:==============*==========
  38    45    68:================       *
  40    80    95:============================    *
  42    93   116:================================       *
  44    89   128:================================            *
  46   113   130:=========================================     *
  48   112   125:=========================================    *
  50   104   114:==================================  *
  52   140   100:==================================*=============
  54   102    85:============================*=====
  56    57    71:====================    *
  58    75    59:===================*=====
  60    42    47:============== *
  62    31    38:=========== *
  64    46    30:=========*======
  66    27    24:=======*=
  68    11    19:====  *
```

Confidential Attachment

```
  70    11    15:====*
  72    12    12:===*
  74     7     9:==*
  76     5     7:==*
  78     5     5:=*
  80     4     4:=*
  82     4     3:*=
  84     0     3:*
  86     1     2:*
  88     2     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     0     1:*         :*
  94     1     1:*         :*
  96     0     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))= 3.17790.00311; mu= 10.6022 0.162
 mean_var=22.0923 5.799, 0's: 3 Z-trim: 3  B-trim: 14 in 1/43
 Lambda= 0.272868
 Kolmogorov-Smirnov  statistic: 0.0667 (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
!! No sequences with E() < 1.000000


31 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:54:50 2009 done: Thu Mar 19 12:54:50 2009
 Total Scan time:  0.040 Total Display time:  0.000

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


# fasta34 5_6.pep /home/ht/db/TOX_2009 -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, 31 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    59     0:====
  22     1     0:=          one = represents 15 library sequences
  24     6     0:=
  26    10     0:=
  28    15     2:*
  30    35    11:*==
  32   123    41:==*======
  34   241   110:=======*=========
  36   277   227:===============*===
  38   268   375:==================      *
  40   403   523:===========================       *
  42   372   639:=========================                 *
  44   850   705:================================================*==========
  46   601   718:=========================================      *
  48   486   688:================================            *
  50   565   627:====================================   *
  52   505   552:================================  *
  54   720   471:================================*================
  56   416   394:=========================*=
  58   379   323:====================*====
  60   246   262:================*
  62   307   210:=============*=======
  64    90   167:======     *
  66    87   132:======  *
  68    58   104:====  *
  70    82    81:=====*
  72    72    64:====*
  74    20    50:== *
  76   207    39:==*==========
  78    15    30:=*
  80    25    23:=*
  82    59    18:=*==
  84    21    14:*=
  86     0    11:*
  88     5     8:*          inset = represents 1 library sequences
  90    15     7:*
  92     1     5:*         :=   *
  94     0     4:*         :   *
  96     2     3:*         :==*
  98     1     2:*         :=*
 100     1     2:*         :=*
 102     0     1:*         :*
 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:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 2.01410.000553; mu= 16.5964 0.028
 mean_var=21.6467 4.655, 0's: 59 Z-trim: 59  B-trim: 550 in 2/60
 Lambda= 0.275663
 Kolmogorov-Smirnov  statistic: 0.0606 (N=28) 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


31 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:54:51 2009 done: Thu Mar 19 12:54:51 2009
 Total Scan time:  0.210 Total Display time:  0.000

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

```
# fasta34 5_6.pep /home/ht/db/PRT_2009 -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, 31 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 218471     0:==========
  22   357     0:=          one = represents 22888 library sequences
  24   737    14:*
  26  1919   309:*
  28  6155  3337:*
  30 22478 20272:*
  32 72839 78384:===*
  34 184630 212567:=========*
  36 373991 436563:=================  *
  38 633481 721476:===========================   *
  40 923074 1006395:=========================================  *
  42 1149613 1230196:=================================================  *
  44 1308103
1357021:=================================================== *
  46 1373223
1382159:====================================================*
  48 1344232
1323258:=====================================================*=
  50 1241140 1207478:=========================================================*==
  52 1079846 1061575:==============================================*=
  54 910943 906771:=======================================*
  56 764813 757432:================================*
  58 630266 621837:===========================*
  60 514415 503725:======================*
  62 407494 403838:=================*
  64 325537 321170:==============*
  66 260643 253843:===========*
  68 218222 199668:========*=
  70 169567 156471:======*=
  72 130682 122267:=====*
  74 103122 95328:====*
  76 84785 74196:===*
  78 60041 57671:==*
  80 46897 44781:=*=
  82 38281 34256:=*
  84 27614 27135:=*
  86 22373 20996:*
  88 17471 16245:*          inset = represents 184 library sequences
  90 13106 12570:*
  92  9154  9726:*         :=======================================*
  94  6882  7525:*         :======================================= *
  96  4648  5823:*         :===========================     *
  98  3466  4505:*         :====================     *
 100  2517  3486:*         :===============    *
 102  1883  2697:*         :===========   *
 104  1419  2087:*         :========   *
 106  1013  1615:*         :======  *
 108   699  1249:*         :====  *
 110   623   967:*         :==== *
 112   448   748:*         :=== *
 114   253   579:*         :== *
 116   303   448:*         :==*
 118   276   347:*         :=*
>120   273   268:*         :=*
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714192 sequences
  Expectation_n fit: rho(ln(x))= 3.69880.000175; mu= 7.5707 0.009
 mean_var=28.1007 5.548, 0's: 859 Z-trim: 865  B-trim: 0 in 0/64
 Lambda= 0.241945
 Kolmogorov-Smirnov  statistic: 0.0275 (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


31 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 12:54:51 2009 done: Thu Mar 19 13:01:59 2009
 Total Scan time: 411.660 Total Display time:  0.000

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

Appendix 7. Bioinformatic analysis of polypeptide 3_1

```
>3_1
PSYSLLIHVD FPDMKPFTIE ETQGVVITAV WPLGQGTVVL KKI
```

```
Sliding 8 amino acid window search
Database searched = AD_2009
Query = 3_1

Start time: Thu Mar 19 13:02:00 CDT 2009 Finish time: Thu Mar 19 13:02:00 CDT
2009

No 8 amino acid matches exist between 3_1 and the AD_2009 database
```

```
# fasta34 3_1.pep /home/ht/db/AD_2009 -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, 43 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20     6     0:==
  22     0     0:           one = represents 3 library sequences
  24     1     0:=
  26     1     0:=
  28     0     0:
  30     0     2:*
  32     7     7:==*
  34    34    20:======*=====
  36    49    41:=============*===
  38    64    68:======================*
  40   101    95:================================*==
  42   118   116:====================================*=
  44   122   128:========================================== *
  46   132   130:============================================*
  48   126   125:===========================================*
  50   109   114:====================================*
  52    55   100:==================             *
  54    46    85:================           *
  56    43    71:===============        *
  58    59    59:==================*
  60    42    47:============== *
  62    41    38:============*=
  64    59    30:=========*==========
  66    36    24:=======*====
  68    43    19:======*========
  70    15    15:====*
  72    12    12:===*
  74    19     9:==*====
  76    14     7:==*==
  78     7     5:=*=
  80     9     4:=*=
  82     4     3:*=
  84     4     3:*=
  86     6     2:*=
  88     2     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     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))= 4.70550.00348; mu= 1.8202 0.183
 mean_var=44.817312.745, 0's: 6 Z-trim: 6  B-trim: 216 in 1/42
 Lambda= 0.191581
 Kolmogorov-Smirnov  statistic: 0.0722 (N=29) at  60

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
```

```
43 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:02:00 2009 done: Thu Mar 19 13:02:00 2009
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 3_1.pep /home/ht/db/TOX_2009 -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, 43 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    76     0:======
  22     1     0:=          one = represents 15 library sequences
  24     1     0:=
  26     2     0:=
  28     3     2:*
  30    22    11:*=
  32    26    41:==*
  34   108   110:=======*
  36   151   227:===========    *
  38   329   375:======================  *
  40   446   523:==============================    *
  42   620   639:============================================*
  44   825   705:============================================*========
  46   684   718:============================================= *
  48   850   688:============================================*==========
  50   691   627:=========================================*=====
  52   618   552:==================================*=====
  54   418   471:============================    *
  56   286   394:====================      *
  58   252   323:=================    *
  60   181   262:=============    *
  62   190   210:=============*
  64   136   167:========== *
  66   111   132:========*
  68    69   104:===== *
  70    56    81:==== *
  72    69    64:====*
  74   126    50:===*=====
  76   125    39:==*======
  78    32    30:=*=
  80    18    23:=*
  82    41    18:=*=
  84    18    14:*=
  86    10    11:*
  88     4     8:*          inset = represents 1 library sequences
  90     6     7:*
  92     7     5:*         :====*==
  94    18     4:*=        :===*==============
  96     3     3:*         :==*
  98     1     2:*         :=*
 100     1     2:*         :=*
 102     0     1:*         :*
 104    15     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:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 4.12540.000655; mu= 3.3444 0.033
 mean_var=29.5307 6.632, 0's: 76 Z-trim: 76  B-trim: 319 in 1/60
 Lambda= 0.236014
 Kolmogorov-Smirnov  statistic: 0.0377 (N=29) at  70

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


43 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:02:00 2009 done: Thu Mar 19 13:02:01 2009
 Total Scan time:  0.140 Total Display time:  0.000

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

```
# fasta34 3_1.pep /home/ht/db/PRT_2009 -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, 43 aa
 vs /home/ht/db/PRT_2009 library
```

```
       opt      E()
< 20 284847     0:============
  22   139     0:=          one = represents 23914 library sequences
  24   286    14:*
  26   837   309:*
  28  2639  3337:*
  30 11398 20272:*
  32 44029 78384:== *
  34 146879 212568:======= *
  36 362039 436564:================  *
  38 682989 721478:============================ *
  40 1014830 1006398:=========================================*
  42 1307552 1230199:=====================================================*===
  44 1434817
1357024:=========================================================*===
  46 1421402
1382163:=========================================================*==
  48 1281019 1323261:========================================================= *
  50 1127517 1207481:=============================================  *
  52 998041 1061577:=========================================  *
  54 852253 906773:================================== *
  56 716310 757434:============================== *
  58 614845 621839:=========================*
  60 504804 503726:====================*
  62 409044 403839:================*=
  64 333196 321171:=============*
  66 268247 253843:==========*=
  68 209838 199668:=========*
  70 166993 156471:======*
  72 126770 122268:=====*
  74 97396 95328:===*=
  76 75472 74196:===*
  78 56991 57671:==*
  80 41164 44781:=*
  82 32382 34256:=*
  84 23568 27135:=*
  86 17792 20996:*
  88 12959 16246:*          inset = represents 134 library sequences
  90  8926 12570:*
  92  6651  9726:*         :=======================================*
  94  4581  7525:*         :==================================    *
  96  3503  5823:*         :==========================            *
  98  2504  4505:*         :===================              *
 100  2121  3486:*         :================          *
 102  1446  2697:*         :===========         *
 104   938  2087:*         :=======        *
 106   733  1615:*         :======      *
 108   579  1249:*         :=====    *
 110   359   967:*         :===    *
 112   213   748:*         :==   *
 114   131   579:*         :=   *
 116   126   448:*         :=  *
 118    90   347:*         := *
>120   233   268:*         :=*
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714225 sequences
  Expectation_n fit: rho(ln(x))= 4.32530.000189; mu= 3.6413 0.010
 mean_var=39.6819 8.045, 0's: 1144 Z-trim: 1145  B-trim: 0 in 0/68
 Lambda= 0.203600
 Kolmogorov-Smirnov  statistic: 0.0213 (N=29) at  58

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


43 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:02:01 2009 done: Thu Mar 19 13:07:28 2009
 Total Scan time: 310.850 Total Display time:  0.000

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

Appendix 8. Bioinformatic analysis of polypeptide 3_2

```
>3_2
SHLQLKRLRV LLSLRFGLWA KAPLS
_____________________________________________________


Sliding 8 amino acid window search
Database searched = AD_2009
Query = 3_2

Start time: Thu Mar 19 13:07:30 CDT 2009 Finish time: Thu Mar 19 13:07:30 CDT
2009

No 8 amino acid matches exist between 3_2 and the AD_2009 database
_____________________________________________________


# fasta34 3_2.pep /home/ht/db/AD_2009 -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, 25 aa
 vs /home/ht/db/AD_2009 library
```

```
       opt      E()
< 20     3     0:=
  22     0     0:           one = represents 3 library sequences
  24     1     0:=
  26     0     0:
  28     0     0:
  30     0     2:*
  32     4     7:==*
  34    30    20:======*===
  36    64    41:=============*========
  38    50    68:=================     *
  40    57    95:===================            *
  42    73   116:=========================             *
  44   113   128:======================================    *
  46   133   130:=============================================*=
  48    89   125:==============================           *
  50   101   114:==================================   *
  52    96   100:================================= *
  54   134    85:============================*===============
  56   104    71:======================*===========
  58    74    59:==================*=====
  60    61    47:===============*=====
  62    47    38:============*===
  64    33    30:=========*=
  66    30    24:=======*==
  68    25    19:======*==
  70    11    15:====*
  72     7    12:===*
  74     7     9:==*
  76     9     7:==*
  78    11     5:=*==
  80     9     4:=*=
  82     2     3:*
  84     2     3:*
  86     1     2:*
  88     0     2:*          inset = represents 1 library sequences
  90     2     1:*
  92     1     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     1     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))= 4.27580.00276; mu= 0.0647 0.144
 mean_var=19.1863 5.044, 0's: 3 Z-trim: 3  B-trim: 7 in 1/43
 Lambda= 0.292805
 Kolmogorov-Smirnov  statistic: 0.0975 (N=28) 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


25 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:07:29 2009 done: Thu Mar 19 13:07:30 2009
 Total Scan time:  0.040 Total Display time:  0.000

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


# fasta34 3_2.pep /home/ht/db/TOX_2009 -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, 25 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    60     0:=====
  22     2     0:=          one = represents 12 library sequences
  24     0     0:
  26     2     0:=
  28     6     2:*
  30     8    11:*
  32    76    41:===*===
  34   155   110:=========*===
  36   252   227:==================*==
  38   538   375:================================*=============
  40   349   523:==============================             *
  42   651   639:=========================================================*=
  44   719   705:============================================================*=
  46   636   718:=========================================================       *
  48   483   688:=========================================                   *
  50   556   627:================================================     *
  52   411   552:==================================          *
  54   377   471:================================       *
```

Confidential Attachment

```
  56   555   394:================================*==============
  58   458   323:=========================*============
  60   291   262:=====================*===
  62   227   210:=================*=
  64   231   167:==============*======
  66   156   132:==========*==
  68    90   104:========*
  70    54    81:===== *
  72    40    64:==== *
  74    30    50:=== *
  76   126    39:===*=======
  78    37    30:==*=
  80     9    23:=*
  82    14    18:=*
  84    21    14:=*
  86     3    11:*
  88    10     8:*          inset = represents 1 library sequences
  90     4     7:*
  92     4     5:*         :====*
  94     1     4:*         :=  *
  96     1     3:*         := *
  98     0     2:*         : *
 100     0     2:*         : *
 102     1     1:*         :*
 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:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 2.97290.000515; mu= 7.8512 0.026
 mean_var=16.9523 3.633, 0's: 59 Z-trim: 59  B-trim: 228 in 1/61
 Lambda= 0.311501
 Kolmogorov-Smirnov  statistic: 0.0617 (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


25 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:07:30 2009 done: Thu Mar 19 13:07:31 2009
 Total Scan time:  0.180 Total Display time:  0.000
```

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


# fasta34 3_2.pep /home/ht/db/PRT_2009 -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, 25 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 221631     0:===========
  22   713     0:=          one = represents 22018 library sequences
  24  1355    14:*
  26  4014   309:*
  28 12027  3337:*
  30 33699 20272:*=
  32 92723 78383:===*=
  34 207211 212566:=========*
  36 402145 436561:==================*
  38 639503 721473:============================  *
  40 870307 1006391:=========================================     *
  42 1099441 1230191:=================================================     *
  44 1255755
1357015:============================================================ *
  46 1321023
1382153:============================================================*
  48 1298037
1323253:============================================================*
  50 1212749 1207473:============================================================*=
  52 1096462 1061570:=============================================*=
  54 992986 906767:=========================================*====
  56 811456 757429:=================================*==
  58 656791 621835:============================*=
  60 521458 503723:====================*=
  62 414084 403836:==================*
  64 325669 321168:==============*
  66 266362 253842:===========*=
  68 202457 199667:=========*
  70 171487 156470:=======*
  72 129572 122267:=====*
  74 102954 95327:====*
  76 82066 74195:===*
  78 63588 57671:==*
  80 49796 44781:==*
  82 37432 34256:=*
  84 27474 27135:=*
  86 21465 20996:*
```

```
  88 16502 16245:*          inset = represents 176 library sequences
  90 11868 12570:*
  92  8757  9726:*         :=====================================*
  94  7742  7525:*         :=====================================*
  96  5733  5823:*         :================================*
  98  3878  4505:*         :======================  *
 100  4776  3486:*         :====================*========
 102  2583  2697:*         :===============*
 104  1598  2087:*         :========== *
 106  1252  1615:*         :======== *
 108  1212  1249:*         :=======*
 110   874   967:*         :=====*
 112   575   748:*         :====*
 114   389   579:*         :===*
 116   253   448:*         :==*
 118   160   347:*         :=*
>120   374   268:*         :=*=
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714129 sequences
  Expectation_n fit: rho(ln(x))= 3.48500.000174; mu= 6.6673 0.009
 mean_var=22.5966 4.543, 0's: 875 Z-trim: 883  B-trim: 3403 in 1/62
 Lambda= 0.269807
 Kolmogorov-Smirnov  statistic: 0.0368 (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


25 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:07:31 2009 done: Thu Mar 19 13:14:10 2009
 Total Scan time: 384.470 Total Display time:  0.010

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

Appendix 9. Bioinformatic analysis of polypeptide 3_3

```
>3_3
RDSGCCYHCG LAFGPRHRCP EKNMRVVILA KDE
_____________________________________________________

Sliding 8 amino acid window search
Database searched = AD_2009
Query = 3_3

Start time: Thu Mar 19 13:14:12 CDT 2009 Finish time: Thu Mar 19 13:14:13 CDT
2009

No 8 amino acid matches exist between 3_3 and the AD_2009 database
_____________________________________________________

# fasta34 3_3.pep /home/ht/db/AD_2009 -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, 33 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20     3     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     2     2:*
  32     1     7:= *
  34    14    20:===== *
  36    62    41:=============*=======
  38    45    68:===============       *
  40    69    95:=======================         *
  42   120   116:=======================================*=
  44   102   128:==================================         *
  46   125   130:============================================ *
  48   137   125:============================================*====
  50    91   114:================================       *
  52    90   100:==============================    *
  54   136    85:============================*=================
  56   111    71:=======================*=============
  58    42    59:==============      *
  60    39    47:=============  *
  62    46    38:============*===
  64    28    30:=========*
  66    35    24:========*====
  68    21    19:=======*
  70    12    15:====*
  72     7    12:===*
  74     9     9:==*
  76     8     7:==*
  78     5     5:=*
  80    10     4:=*==
  82     2     3:*
  84     3     3:*
  86     2     2:*
  88     0     2:*          inset = represents 1 library sequences
  90     3     1:*
```

```
  92     3     1:*          :*==
  94     0     1:*          :*
  96     0     1:*          :*
  98     1     0:=          *=
 100     0     0:           *
 102     2     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))= 4.58810.00284; mu= 2.5429 0.147
 mean_var=30.7520 8.237, 0's: 3 Z-trim: 4  B-trim: 3 in 1/43
 Lambda= 0.231280
 Kolmogorov-Smirnov  statistic: 0.0635 (N=28) 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(1386)
gi|71057064|emb|CAI38795.2| thaumatin-like protein ( 225)   57 23.6    0.83
gi|146737976|gb|ABQ42566.1| thaumatin-like protein ( 201)   56 23.2    0.92


>>gi|71057064|emb|CAI38795.2| thaumatin-like protein [Ac  (225 aa)
 initn:  41 init1:  41 opt:  57  Z-score: 103.4  bits: 23.6 E(): 0.83
Smith-Waterman score: 57;  40.000% identity (64.000% similar) in 25 aa
overlap (2-21:174-198)

                                          10          20
3_3                                 RDSGCCY--HCGLA-FGP--RHRCPEKNMRV
                                     :. ::   .:::. :.   . :::.
gi|710 IKCTADINGQCPNELRAPGGCNNPCTVFKTDQYCCNSGNCGLTNFSKFFKDRCPDAYSYP
          150       160       170       180       190       200

        30
3_3    VILAKDE

gi|710 KDDQTSTFTCPAGTNYKVVFCP
          210       220

>>gi|146737976|gb|ABQ42566.1| thaumatin-like protein [Ac  (201 aa)
 initn:  41 init1:  41 opt:  56  Z-score: 102.5  bits: 23.2 E(): 0.92
Smith-Waterman score: 56;  40.000% identity (64.000% similar) in 25 aa
overlap (2-21:150-174)
```

```
                                          10          20
3_3                                 RDSGCCY--HCGLA-FGP--RHRCPEKNMRV
                                     :. ::   .:::. :.   . :::.
gi|146 IKCTADINGQCPNELRAPGGCNNPCTVFKTDQFCCNSGNCGLTNFSKFFKDRCPDAYSYP
     120       130       140       150       160       170

        30
3_3    VILAKDE

gi|146 KDDQTSTFTCPAGTNYKVVFCP
     180       190       200



33 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:14:12 2009 done: Thu Mar 19 13:14:12 2009
 Total Scan time:  0.030 Total Display time:  0.010

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

```
# fasta34 3_3.pep /home/ht/db/TOX_2009 -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, 33 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    59     0:=====
  22     0     0:           one = represents 14 library sequences
  24     0     0:
  26     0     0:
  28     4     2:*
  30    17    11:*=
  32    32    41:==*
  34    57   110:=====  *
  36   270   227:================*===
  38   399   375:==========================*==
  40   802   523:====================================*====================
  42   594   639:============================================  *
  44   469   705:==================================                *
  46   726   718:================================================*
  48   788   688:================================================*=======
  50   494   627:====================================        *
```

```
  52   738   552:=======================================*============
  54   324   471:=======================         *
  56   224   394:================             *
  58   238   323:================       *
  60   273   262:==================*=
  62   279   210:==============*=====
  64   318   167:===========*==========
  66   120   132:=========*
  68    63   104:=====  *
  70    30    81:===  *
  72    56    64:====*
  74    43    50:===*
  76    37    39:==*
  78    48    30:==*=
  80    32    23:=*=
  82    29    18:=*=
  84     7    14:*
  86    27    11:*=
  88     4     8:*          inset = represents 1 library sequences
  90    33     7:*==
  92     9     5:*         :====*====
  94     1     4:*         :=  *
  96     0     3:*         :  *
  98     0     2:*         : *
 100     1     2:*         :=*
 102     0     1:*         :*
 104     1     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:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 0.73390.000608; mu= 23.6282 0.032
 mean_var=49.785911.815, 0's: 59 Z-trim: 59  B-trim: 499 in 1/61
 Lambda= 0.181770
 Kolmogorov-Smirnov  statistic: 0.0386 (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


33 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:14:13 2009 done: Thu Mar 19 13:14:13 2009
 Total Scan time:  0.200 Total Display time:  0.000

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

# fasta34 3_3.pep /home/ht/db/PRT_2009 -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, 33 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 217875     0:==========
  22   216     0:=          one = represents 23559 library sequences
  24   565    14:*
  26  1251   309:*
  28  4313  3337:*
  30 17531 20270:*
  32 78202 78378:===*
  34 181826 212552:======== *
  36 404778 436532:==================*
  38 678658 721425:============================ *
  40 981788 1006325:=========================================*
  42 1209560 1230110:=========================================================*
  44 1365320
1356926:=========================================================*
  46 1413510
1382062:=========================================================*=
  48 1356604
1323165:=========================================================*=
  50 1227509 1207394:=========================================================*=
  52 1062237 1061501:=========================================================*
  54 858628 906708:=================================== *
  56 711798 757379:============================== *
  58 571304 621794:======================== *
  60 466857 503689:==================== *
  62 387565 403809:=================*
  64 314015 321147:=============*
  66 252598 253825:==========*
  68 195535 199654:========*
  70 151920 156460:======*
  72 126031 122259:=====*
  74 101649 95321:====*
  76 75790 74191:===*
  78 61453 57667:==*
  80 49467 44778:=*=
  82 41237 34254:=*
```

```
  84 31564 27133:=*
  86 24391 20994:*=
  88 18867 16244:*          inset = represents 238 library sequences
  90 15084 12569:*
  92 11874  9725:*         :======================================*
  94  9031  7525:*         :==============================*======
  96  6644  5822:*         :========================*===
  98  4914  4505:*         :==================*==
 100  4301  3486:*         :===============*====
 102  3875  2697:*         :===========*=====
 104  2936  2087:*         :========*====
 106  2328  1615:*         :======*===
 108  1967  1249:*         :=====*===
 110  1900   967:*         :====*===
 112  1955   748:*         :===*=====
 114   959   579:*         :==*==
 116  1681   448:*         :=*======
 118  1094   346:*         :=*===
>120  1463   268:*         :=*=====
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14713160 sequences
  Expectation_n fit: rho(ln(x))= 3.82010.000183; mu= 7.1989 0.010
 mean_var=33.1512 6.354, 0's: 913 Z-trim: 915  B-trim: 0 in 0/64
 Lambda= 0.222754
 Kolmogorov-Smirnov  statistic: 0.0187 (N=29) at  70

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(14717352)
gi|124360394|gb|ABN08407.1| Peptidase aspartic, ac ( 435)  137 48.9  0.0004
gi|124360392|gb|ABN08405.1| Peptidase aspartic, ac ( 435)  137 48.9  0.0004
gi|124359710|gb|ABN06064.1| RNA-directed DNA polym (1297)  137 49.1   0.001
gi|217073570|gb|ACJ85145.1| unknown [Medicago trun ( 185)  124 44.5  0.0035
gi|125599582|gb|EAZ39158.1| hypothetical protein O ( 325)  104 38.2    0.49
gi|157351425|emb|CAO41650.1| unnamed protein produ ( 219)  101 37.2    0.68


>>gi|124360394|gb|ABN08407.1| Peptidase aspartic, active  (435 aa)
 initn: 137 init1: 137 opt: 137  Z-score: 235.1  bits: 48.9 E(): 0.0004
Smith-Waterman score: 137;  46.875% identity (78.125% similar) in 32 aa
overlap (1-32:59-90)

                                          10        20        30
3_3                                    RDSGCCYHCGLAFGPRHRCPEKNMRVVILA
                                       :..: :..::  . :::.::.::. :..:
gi|124 VGQNKTHTINTANWRDKNVRSLSSQEIADRRQKGLCFKCGGPYHPRHQCPDKNLSVMVLE
       30        40        50        60        70        80
```

```
3_3    KDE
        :
gi|124 DDSEDENEVRVLNDEDVDTGAEELQLNVLTFENALTFDRQTEYYQDRFQCIRFQGKVREI
       90        100       110       120       130       140

>>gi|124360392|gb|ABN08405.1| Peptidase aspartic, active  (435 aa)
 initn: 137 init1: 137 opt: 137  Z-score: 235.1  bits: 48.9 E(): 0.0004
Smith-Waterman score: 137;  46.875% identity (78.125% similar) in 32 aa
overlap (1-32:59-90)

                                          10        20        30
3_3                                    RDSGCCYHCGLAFGPRHRCPEKNMRVVILA
                                       :..: :..::  . :::.::.::. :..:
gi|124 VGQNKTHTINTANWRDKNVRSLSSQEIADRRQKGLCFKCGGPYHPRHQCPDKNLSVMVLE
       30        40        50        60        70        80


3_3    KDE
        :
gi|124 DDSEDENEVRVLNDEDVDTGAEELQLNVLTFENALTFDRQTEYYQDRFQCIRFQGKVREI
       90        100       110       120       130       140

>>gi|124359710|gb|ABN06064.1| RNA-directed DNA polymeras  (1297 aa)
 initn: 137 init1: 137 opt: 137  Z-score: 227.9  bits: 49.1 E(): 0.001
Smith-Waterman score: 137;  45.455% identity (81.818% similar) in 33 aa
overlap (1-33:108-140)

                                          10        20        30
3_3                                    RDSGCCYHCGLAFGPRHRCPEKNMRVVILA
                                       ...: :..::  : : :.::.:..::..:
gi|124 GEKQAQYDKKKSGPRDRSFTHLSYNELMERKQKGLCFKCGGPFHPMHQCPDKQLRVLVLE
        80        90       100       110       120       130


3_3    KDE
       .::
gi|124 EDEEGEPEGKLLAVEVDDEEEGDGEMCMMEFFHLGHSRPQSIKLMGVIKEVPVVVLVDSG
       140       150       160       170       180       190

>>gi|217073570|gb|ACJ85145.1| unknown [Medicago truncatu  (185 aa)
 initn: 123 init1:  81 opt: 124  Z-score: 218.2  bits: 44.5 E(): 0.0035
Smith-Waterman score: 124;  47.059% identity (76.471% similar) in 34 aa
overlap (1-33:5-38)

                  10        20        30
3_3        RDSGCCYHCGLAFGPR-HRCPEKNMRVVILAKDE
           : .: :..::  . :  :.:::::..::.::.. :
gi|217 MAERRAKGLCFKCGGKYHPTLHKCPEKSLRVLILGEGEGVNEEGEIVSLETQEVLEEEEE
               10        20        30        40        50        60
```

```
gi|217 EIESECKVIGVLGSMGEYNTMKIGGKLENIDVVVLVDSGATHNFISAKLTSALGLTITPM
               70        80        90       100       110       120

>>gi|125599582|gb|EAZ39158.1| hypothetical protein OsJ_0  (325 aa)
 initn:  98 init1:  98 opt: 104  Z-score: 179.8  bits: 38.2 E(): 0.49
Smith-Waterman score: 104;  45.161% identity (64.516% similar) in 31 aa
overlap (1-31:250-279)

                                        10        20        30
3_3                                  RDSGCCYHCGLAFGPRHRCPEKNMRVVILA
                                     :  : :..::  .:: :.::.   :.: :
gi|125 GILGAAPEDNKKAEEKSKWEEQFDSLKAARRARGECFKCGEKYGPGHKCPNFR-RIVNLK
     220       230       240       250       260       270


3_3    KDE
       :
gi|125 KGMQKVQRRNWYCLNVQLQARWAGKPSSFMGNTKTTVAYLSGLRELK
      280       290       300       310       320

>>gi|157351425|emb|CAO41650.1| unnamed protein product [  (219 aa)
 initn:  85 init1:  85 opt: 101  Z-score: 177.2  bits: 37.2 E(): 0.68
Smith-Waterman score: 101;  42.424% identity (66.667% similar) in 33 aa
overlap (1-33:50-80)

                                        10        20        30
3_3                                  RDSGCCYHCGLAFGPRHRCPEKNMRVVILA
                                     ::.: ::.:   ..: :.:  :. :. :.
gi|157 TRLAIGPPSPPEKRAIVPVQRLSPSQMKERRDKGLCYNCDDKWAPGHKC--KSARLFIME
      20        30        40        50        60          70


3_3    KDE
        ::
gi|157 CDESSDDEVPKSEIEPGISIHALVGSPNPKTMRFLGHICGRAVVILVDTGSTHNFMDPSV
        80        90       100       110       120       130




33 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:14:13 2009 done: Thu Mar 19 13:21:52 2009
 Total Scan time: 440.630 Total Display time:  0.010

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

Appendix 10. Bioinformatic analysis of polypeptide 3_4

```
>3_4
QHPESLQL
________________________________________________________

Sliding 8 amino acid window search
Database searched = AD_2009
Query = 3_4

Start time: Thu Mar 19 13:21:55 CDT 2009 Finish time: Thu Mar 19 13:21:55 CDT
2009

No 8 amino acid matches exist between 3_4 and the AD_2009 database
________________________________________________________

# fasta34 3_4.pep /home/ht/db/AD_2009 -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, 8 aa
 vs /home/ht/db/AD_2009 library

        opt      E()
< 20     3     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     7     0:===
  30     2     2:*
  32    22     7:==*=====
  34    15    20:===== *
  36    27    41:=========    *
  38    53    68:==================    *
  40    72    95:=======================       *
  42   107   116:====================================  *
  44   140   128:============================================*====
  46   102   130:====================================          *
  48   147   125:============================================*=======
  50   115   114:====================================*=
  52    82   100:============================     *
  54    75    85:=========================   *
  56    76    71:=======================*==
  58    74    59:==================*=====
  60    55    47:===============*===
  62    42    38:============*=
  64    33    30:=========*=
  66    46    24:=======*========
  68    26    19:======*==
```

Confidential Attachment

```
  70    14    15:====*
  72     8    12:===*
  74    12     9:==*=
  76     8     7:==*
  78     0     5: *
  80     6     4:=*
  82     6     3:*=
  84     3     3:*
  86     2     2:*
  88     0     2:*          inset = represents 1 library sequences
  90     2     1:*
  92     0     1:*         :*
  94     1     1:*         :*
  96     0     1:*         :*
  98     0     0:          *
 100     1     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:          *
 307888 residues in  1386 sequences
  Expectation_n fit: rho(ln(x))= 3.68250.00195; mu= -1.1072 0.103
 mean_var=10.9225 2.642, 0's: 3 Z-trim: 5  B-trim: 8 in 1/43
 Lambda= 0.388074
 Kolmogorov-Smirnov  statistic: 0.0472 (N=27) at  54

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 45, opt: 33, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(1386)
gi|169969|gb|AAA33964.1| glycinin                  ( 516)   40 23.0    0.69
gi|736002|emb|CAA55977.1| Gy5 [Glycine soja]       ( 517)   40 23.0    0.69


>>gi|169969|gb|AAA33964.1| glycinin                        (516 aa)
 initn:  40 init1:  40 opt:  40  Z-score: 104.8  bits: 23.0 E(): 0.69
Smith-Waterman score: 40;  57.143% identity (100.000% similar) in 7 aa
overlap (1-7:196-202)


3_4                                 QHPESLQL
                                    .::::.:
gi|169 VAISPLDTSNFNNQLDQNPRVFYLAGNPDIEHPETMQQQQQQKSHGGRKQGQHRQQEEEG
         170       180       190       200       210       220

gi|169 GSVLSGFSKHFLAQSFNTNEDTAEKLRSPDDERKQIVTVEGGLSVISPKWQEQEDEDEDE
         230       240       250       260       270       280

>>gi|736002|emb|CAA55977.1| Gy5 [Glycine soja]              (517 aa)
 initn:  40 init1:  40 opt:  40  Z-score: 104.8  bits: 23.0 E(): 0.69
Smith-Waterman score: 40;  57.143% identity (100.000% similar) in 7 aa
overlap (1-7:196-202)


3_4                                 QHPESLQL
                                    .::::.:
gi|736 VAISLLDTSNFNNQLDQNPRVFYLAGNPDIEHPETMQQQQQQKSHGGRKQGQHQQQEEEG
         170       180       190       200       210       220

gi|736 GSVLSGFSKHFLAQSFNTNEDTAEKLRSPDDERKQIVTVEGGLSVISPKWQEQEDEDEDE
         230       240       250       260       270       280



8 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:21:54 2009 done: Thu Mar 19 13:21:54 2009
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 3_4.pep /home/ht/db/TOX_2009 -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, 8 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    59     0:=====
  22     0     0:           one = represents 13 library sequences
  24     0     0:
  26     0     0:
  28     5     2:*
  30     9    11:*
  32    32    41:===*
  34    99   110:========*
  36   316   227:==================*=======
  38   398   375:============================*==
  40   648   523:=========================================*=========
```

```
  42   482   639:====================================           *
  44   635   705:================================================     *
  46   769   718:============================================================*====
  48   648   688:================================================  *
  50   516   627:=========================================        *
  52   574   552:=========================================*==
  54   457   471:====================================*
  56   395   394:==============================*
  58   209   323:================       *
  60   204   262:==============    *
  62   298   210:===============*======
  64   152   167:============*
  66   170   132:==========*===
  68   120   104:=======*==
  70   122    81:======*===
  72    89    64:====*==
  74    67    50:===*==
  76    36    39:==*
  78    54    30:==*==
  80    26    23:=*
  82    10    18:=*
  84     8    14:=*
  86    29    11:*==
  88     4     8:*          inset = represents 1 library sequences
  90     1     7:*
  92     2     5:*         :==  *
  94     1     4:*         :=  *
  96     0     3:*         :  *
  98     1     2:*         :=*
 100     0     2:*         : *
 102     1     1:*         :*
 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:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 2.63570.000506; mu= 4.0691 0.026
 mean_var=12.4760 2.720, 0's: 59 Z-trim: 59  B-trim: 499 in 1/61
 Lambda= 0.363108
 Kolmogorov-Smirnov  statistic: 0.0358 (N=29) at  60

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 45, opt: 33, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000
```

```
8 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:21:55 2009 done: Thu Mar 19 13:21:55 2009
 Total Scan time:  0.100 Total Display time:  0.000

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

```
# fasta34 3_4.pep /home/ht/db/PRT_2009 -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, 8 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 226999     0:===========
  22   808     0:=          one = represents 20826 library sequences
  24  2010    14:*
  26  6239   309:*
  28 20907  3337:*=
  30 53828 20272:*==
  32 122663 78384:===*==
  34 246489 212569:==========*=
  36 402212 436567:====================*
  38 590834 721483:==============================     *
  40 807917 1006406:=========================================          *
  42 995963 1230209:==================================================
*
  44 1119604 1357035:=========================================================
*
  46 1238333
1382173:============================================================*
  48 1249520
1323272:============================================================*
  50 1233716
1207491:==========================================================*==
  52 1134922 1061586:=====================================================*====
  54 983863 906780:=========================================*====
  56 836754 757440:=======================================*====
  58 690625 621844:============================*====
  60 556651 503730:========================*==
  62 469491 403842:===================*===
  64 393253 321173:================*===
  66 309887 253846:============*==
  68 249987 199670:=========*===
```

```
  70 194000 156473:=======*==
  72 152333 122269:=====*==
  74 104448 95329:====*=
  76 84404 74197:===*=
  78 62210 57672:==*
  80 45786 44781:==*
  82 34615 34257:=*
  84 24422 27136:=*
  86 19201 20996:=*
  88 13487 16246:*          inset = represents 147 library sequences
  90  9600 12570:*
  92  7349  9726:*         :=======================================*
  94  5367  7525:*         :=====================================  *
  96  3562  5823:*         :=========================              *
  98  2782  4505:*         :====================           *
 100  1852  3486:*         :==============          *
 102  2356  2697:*         :================= *
 104  1072  2087:*         :=========      *
 106   570  1615:*         :====       *
 108   575  1249:*         :====    *
 110   380   967:*         :===   *
 112   143   748:*         :=    *
 114   119   579:*         :=  *
 116   164   448:*         :== *
 118    54   347:*         := *
>120    92   268:*         :=*
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714342 sequences
  Expectation_n fit: rho(ln(x))= 3.18190.000171; mu= 2.2128 0.009
 mean_var=13.9914 2.820, 0's: 883 Z-trim: 885  B-trim: 1749 in 1/64
 Lambda= 0.342882
 Kolmogorov-Smirnov  statistic: 0.0631 (N=29) at  48

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 45, opt: 33, open/ext: -10/-2, width:  16
!! No sequences with E() < 1.000000


8 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:21:55 2009 done: Thu Mar 19 13:26:04 2009
 Total Scan time: 223.920 Total Display time:  0.000

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

Appendix 11. Bioinformatic analysis of polypeptide 3_5

```
>3_5
VSSIVNGFMS GKSTWISNEY DGQYGEKERV ITNFFSIQKC RCPQRYYKMK VHFDKTTNYD PSYL
_____________________________________________________________

Sliding 8 amino acid window search
Database searched = AD_2009
Query = 3_5

Start time: Thu Mar 19 13:26:06 CDT 2009 Finish time: Thu Mar 19 13:26:06 CDT
2009

No 8 amino acid matches exist between 3_5 and the AD_2009 database
_____________________________________________________________

# fasta34 3_5.pep /home/ht/db/AD_2009 -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, 64 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20     8     0:===
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     1     0:=
  30     1     2:*
  32     1     7:= *
  34     5    20:==    *
  36    23    41:========     *
  38    45    68:===============       *
  40   140    95:================================*===============
  42   132   116:=========================================*=====
  44   113   128:=======================================     *
  46   120   130:========================================    *
  48   110   125:======================================     *
  50   100   114:===================================    *
  52   128   100:================================*=========
  54    89    85:============================*=
  56    76    71:========================*==
  58    54    59:================== *
  60    34    47:============   *
  62    25    38:==========   *
  64    37    30:==========*===
  66    23    24:========*
  68    27    19:=======*==
  70    19    15:====*==
  72    13    12:===*=
```

```
  74    13     9:==*==
  76    13     7:==*==
  78    13     5:=*===
  80     5     4:=*
  82     3     3:*
  84     3     3:*
  86     8     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     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))= 4.60190.00387; mu= 4.1477 0.201
 mean_var=52.513315.272, 0's: 8 Z-trim: 8  B-trim: 47 in 1/43
 Lambda= 0.176986
 Kolmogorov-Smirnov  statistic: 0.0455 (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
!! No sequences with E() < 1.000000


64 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:26:05 2009 done: Thu Mar 19 13:26:05 2009
 Total Scan time:  0.040 Total Display time:  0.000

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

```
# fasta34 3_5.pep /home/ht/db/TOX_2009 -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, 64 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    66     0:=====
  22     0     0:           one = represents 14 library sequences
  24     0     0:
  26     4     0:=
  28     2     2:*
  30    13    11:*
  32    24    41:==*
  34   259   110:=======*===========
  36   217   227:================*
  38   403   375:=========================*==
  40   520   523:====================================*
  42   553   639:=========================================     *
  44   838   705:==================================================*=========
  46   568   718:=========================================         *
  48   677   688:==================================================*
  50   557   627:=========================================    *
  52   391   552:============================           *
  54   377   471:===========================      *
  56   296   394:======================       *
  58   341   323:=======================*=
  60   532   262:==================*==================
  62   244   210:==============*===
  64   191   167:===========*==
  66   123   132:=========*
  68   119   104:=======*=
  70    68    81:=====*
  72    55    64:====*
  74    52    50:===*
  76    28    39:==*
  78    16    30:==*
  80    43    23:=*==
  82    16    18:=*
  84     8    14:*
  86     7    11:*
  88     6     8:*          inset = represents 1 library sequences
  90    12     7:*
  92     2     5:*         :==  *
  94     3     4:*         :===*
  96     6     3:*         :==*===
  98     0     2:*         : *
 100     0     2:*         : *
 102     2     1:*         :*=
 104     0     1:*         :*
 106     4     1:*         :*===
 108     1     1:*         :*
 110     0     1:*         :*
```

```
 112     1     0:=         *=
 114     0     0:          *
 116     0     0:          *
 118     1     0:=         *=
>120     0     0:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 4.88070.000709; mu= 2.8593 0.036
 mean_var=40.1636 8.806, 0's: 66 Z-trim: 67  B-trim: 867 in 2/60
 Lambda= 0.202375
 Kolmogorov-Smirnov  statistic: 0.0506 (N=29) at  56

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(7651)
gi|158635887|dbj|BAF91371.1| cytotoxic polypeptide ( 222)   72 27.1    0.75


>>gi|158635887|dbj|BAF91371.1| cytotoxic polypeptide [Mi  (222 aa)
 initn:  39 init1:  39 opt:  72  Z-score: 117.5  bits: 27.1 E(): 0.75
Smith-Waterman score: 72;  31.250% identity (62.500% similar) in 48 aa
overlap (7-49:119-166)

                                     10          20          30
3_5                               VSSIVNGFMSGKSTW--ISNEYDGQYG---EKERVI
                                        :.  : :::   .:.:::...   ..  :.
gi|158 CPLGETIKSIGSIHDNHYEDRQWDIDCKPAGYTMGISTWSPYANDYDGSMNFECNEGSVV
       90       100       110       120       130       140

             40        50        60
3_5    TNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
       :.. ::.     .: :..
gi|158 TGMSSIHDNYYEDRRYQLMCSYLNNWKRGSCAWTSYTTYDASFVELTPTGKFLVGMKSQH
      150       160       170       180       190       200




64 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:26:06 2009 done: Thu Mar 19 13:26:06 2009
 Total Scan time:  0.280 Total Display time:  0.000

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

```
# fasta34 3_5.pep /home/ht/db/PRT_2009 -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, 64 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 270819     0:============
  22    96     0:=          one = represents 23069 library sequences
  24   198    14:*
  26   462   309:*
  28  1795  3337:*
  30  9331 20272:*
  32 45502 78383:== *
  34 158091 212566:=======  *
  36 398078 436560:==================*
  38 725262 721471:==============================*
  40 1108484 1006389:=============================================*=====
  42 1342177
1230188:=========================================================*=====
  44 1384124
1357012:============================================================*=
  46 1328054
1382150:============================================================ *
  48 1250960 1323249:============================================================
*
  50 1107747 1207470:================================================   *
  52 965611 1061568:=============================================     *
  54 821926 906765:=====================================   *
  56 712671 757427:================================ *
  58 613513 621833:===========================*
  60 502508 503721:======================*
  62 413320 403835:==================*
  64 334136 321168:=============*=
  66 271603 253841:===========*
  68 209252 199666:========*=
  70 165238 156470:======*=
  72 129125 122267:=====*
  74 101162 95327:====*
  76 79649 74195:===*
  78 65376 57671:==*
  80 49046 44781:=*=
  82 36967 34256:=*
  84 27520 27135:=*
  86 20293 20996:*
  88 16018 16245:*          inset = represents 180 library sequences
  90 12287 12570:*
  92  8968  9726:*         :=======================================*
  94  6768  7525:*         :======================================== *
  96  5149  5823:*         :============================   *
  98  4340  4505:*         :=========================*
```

```
 100  2801  3486:*         :================   *
 102  1930  2697:*         :===========   *
 104  1455  2087:*         :=========  *
 106  1147  1615:*         :======= *
 108  1071  1249:*         :======*
 110   690   967:*         :==== *
 112   489   748:*         :=== *
 114   342   579:*         :== *
 116   294   448:*         :==*
 118   176   347:*         :=*
>120   397   268:*         :=*=
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714094 sequences
  Expectation_n fit: rho(ln(x))= 3.98620.000186; mu= 7.0983 0.010
 mean_var=44.7755 9.186, 0's: 1171 Z-trim: 1173  B-trim: 2898 in 1/62
 Lambda= 0.191670
 Kolmogorov-Smirnov  statistic: 0.0245 (N=29) at  60

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


64 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:26:07 2009 done: Thu Mar 19 13:33:39 2009
 Total Scan time: 435.140 Total Display time:  0.000

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

Appendix 12. Bioinformatic analysis of polypeptide 3_6

```
>3_6
RVLQLSYFFQ DNGALAQRPN RSDNNTLSLF NCKWLHVREI YMDQQ
_____________________________________________________


Sliding 8 amino acid window search
Database searched = AD_2009
Query = 3_6

Start time: Thu Mar 19 13:33:41 CDT 2009 Finish time: Thu Mar 19 13:33:42 CDT
2009

No 8 amino acid matches exist between 3_6 and the AD_2009 database
_____________________________________________________

# fasta34 3_6.pep /home/ht/db/AD_2009 -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, 45 aa
 vs /home/ht/db/AD_2009 library

       opt      E()
< 20    18     0:======
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     0     2:*
  32     0     7:  *
  34     2    20:=     *
  36    24    41:========     *
  38    36    68:============          *
  40    88    95:============================== *
  42   155   116:=======================================*=============
  44   161   128:============================================*===========
  46   126   130:============================================ *
  48   108   125:====================================     *
  50   103   114:==================================  *
  52    94   100:================================ *
  54    63    85:====================       *
  56    81    71:=======================*===
  58    62    59:==================*=
  60    46    47:================*
  62    53    38:============*=====
  64    18    30:======   *
  66    39    24:=======*=====
  68    31    19:======*====
  70    18    15:====*=
  72     7    12:===*
  74     8     9:==*
  76     7     7:==*
  78     2     5:=*
  80     5     4:=*
  82     3     3:*
  84     1     3:*
  86     1     2:*
  88     6     2:*=         inset = represents 1 library sequences
  90     0     1:*
  92     3     1:*         :*==
  94    16     1:*=====    :*===============
  96     0     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))= 3.19610.00392; mu= 7.4012 0.202
 mean_var=35.947310.223, 0's: 18 Z-trim: 18  B-trim: 49 in 1/42
 Lambda= 0.213915
 Kolmogorov-Smirnov  statistic: 0.0611 (N=28) at  40

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


45 residues in 1 query   sequences
307888 residues in 1386 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:33:41 2009 done: Thu Mar 19 13:33:41 2009
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 3_6.pep /home/ht/db/TOX_2009 -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, 45 aa
 vs /home/ht/db/TOX_2009 library

       opt      E()
< 20    77     0:======
  22     1     0:=          one = represents 13 library sequences
  24     0     0:
  26     1     0:=
  28     0     2:*
  30    24    11:*=
  32    39    41:===*
  34   103   110:========*
  36   232   227:=================*
  38   360   375:============================*
  40   449   523:==================================     *
  42   593   639:=============================================   *
  44   760   705:=========================================================*====
  46   617   718:================================================       *
  48   630   688:================================================   *
  50   635   627:================================================*
  52   580   552:=========================================*==
  54   714   471:===================================*==================
  56   327   394:=========================    *
  58   219   323:=================       *
  60   184   262:===============     *
  62   307   210:================*=======
  64   152   167:============*
  66    91   132:=======   *
  68    99   104:=======*
  70    72    81:======*
  72    65    64:====*
  74    43    50:===*
  76    70    39:==*===
  78    23    30:==*
  80    22    23:=*
  82     9    18:=*
  84    17    14:=*
  86    39    11:*==
  88     7     8:*          inset = represents 1 library sequences
  90    13     7:*
  92     5     5:*         :====*
  94    18     4:*=        :===*==============
  96    35     3:*==       :==*=================================
  98     1     2:*         :=*
 100     1     2:*         :=*
 102     7     1:*         :*======
 104     1     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     2     0:=         *==
 114     1     0:=         *=
 116     0     0:          *
 118     1     0:=         *=
>120     0     0:          *
1891534 residues in  7651 sequences
  Expectation_n fit: rho(ln(x))= 4.45540.000653; mu= 1.9335 0.033
 mean_var=30.2304 6.717, 0's: 77 Z-trim: 77  B-trim: 169 in 1/61
 Lambda= 0.233266
 Kolmogorov-Smirnov  statistic: 0.0303 (N=29) at  48

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(7651)
```

```
gi|256378|gb|AAB23447.1| neurotoxin Tx2-9 [Phoneut (  32)   55 24.0    0.66


>>gi|256378|gb|AAB23447.1| neurotoxin Tx2-9 [Phoneutria   (32 aa)
 initn:  55 init1:  55 opt:  55  Z-score: 118.4  bits: 24.0 E(): 0.66
Smith-Waterman score: 55;  43.750% identity (75.000% similar) in 16 aa
overlap (18-33:7-22)

               10        20        30        40
3_6    RVLQLSYFFQDNGALAQRPNRSDNNTLSLFNCKWLHVREIYMDQQ
                        .: .::.:  . :.::
gi|256            SFCIPFKPCKSDENCCKKFKCKTTGIVKLCRW
                         10        20        30




45 residues in 1 query   sequences
1891534 residues in 7651 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:33:42 2009 done: Thu Mar 19 13:33:42 2009
 Total Scan time:  0.140 Total Display time:  0.000

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


# fasta34 3_6.pep /home/ht/db/PRT_2009 -Q -E 1 -O 3_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

 3_6, 45 aa
 vs /home/ht/db/PRT_2009 library

       opt      E()
< 20 295115     0:=============
  22   351     0:=          one = represents 24275 library sequences
  24   249    14:*
  26   531   309:*
  28  1505  3337:*
  30  7008 20271:*
  32 31267 78383:== *
  34 115989 212565:=====   *
  36 299175 436558:=============     *
  38 598995 721468:=========================    *
  40 952263 1006384:======================================== *
  42 1273663 1230182:==================================================*==
  44 1456453
1357005:==========================================================*====
  46 1424260
1382144:==========================================================*==
  48 1310842 1323243:============================================================*
  50 1123829 1207465:=============================================  *
  52 948847 1061563:=======================================   *
  54 843153 906761:==================================  *
  56 716933 757423:============================== *
  58 615908 621830:=========================*
  60 526609 503719:====================*=
  62 427950 403833:================*=
  64 350305 321166:=============*=
  66 289249 253840:==========*=
  68 233236 199665:========*=
  70 189420 156469:======*=
  72 156278 122266:=====*=
  74 116277 95327:===*=
  76 93424 74195:===*
  78 73263 57671:==*=
  80 59558 44780:=*=
  82 44757 34256:=*
  84 34596 27135:=*
  86 25069 20996:*=
  88 18932 16245:*          inset = represents 219 library sequences
  90 14208 12570:*
  92 10909  9726:*         :=======================================*
  94  8558  7525:*         :==================================*=====
  96  5888  5823:*         :=========================*
  98  4349  4505:*         :====================*
 100  3316  3486:*         :===============*
 102  2643  2697:*         :============*
 104  1785  2087:*         :=========*
 106  1213  1615:*         :====== *
 108   929  1249:*         :=====*
 110  1021   967:*         :====*
 112   493   748:*         :===*
 114  2943   579:*         :==*==========
 116   258   448:*         :==*
 118   192   347:*         :=*
>120   454   268:*         :=*=
3787527556 residues in 14717352 library sequences
 statistics sampled from 60000 to 14714025 sequences
  Expectation_n fit: rho(ln(x))= 3.86920.000184; mu= 5.4731 0.010
 mean_var=36.1245 7.298, 0's: 1130 Z-trim: 1130  B-trim: 0 in 0/65
 Lambda= 0.213390
 Kolmogorov-Smirnov  statistic: 0.0420 (N=29) at  58

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
```

```
45 residues in 1 query   sequences
3787527556 residues in 14717352 library sequences
 Scomplib [34t26]
 start: Thu Mar 19 13:33:42 2009 done: Thu Mar 19 13:39:02 2009
 Total Scan time: 303.020 Total Display time:  0.000

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

```
Database checksum values:

Thu Mar 19 13:39:05 CDT 2009      5c91759664b0377022cf35e7d5a0d23c
/home/ht/db/AD_2009

Thu Mar 19 13:39:05 CDT 2009      fa2987b358e662a0d03802f0cfba9676
/home/ht/db/TOX_2009

Thu Mar 19 13:40:59 CDT 2009      a155ebc7632842e917ba5dd4cce1dc09
/home/ht/db/PRT_2009
```