Report Title

Updated Bioinformatics Evaluation of DNA Sequences Flanking the 5′, additional 5’ and 3′ Junctions of the Inserted DNA in MON 87705 Utilizing the AD_2010, TOX_2010 and PRT_2010 Databases: Assessment of Putative Polypeptides

Authors

Haidi Tu
Andre Silvanovich, Ph.D.

Report Completed On

March 2, 2010

Sponsor and Performing Laboratory

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

Laboratory Project ID

Study Number: RAR-10-051

Table of Contents

Page

Tables

Appendices

1.0 Summary

Monsanto Company has developed biotechnology-derived soybean MON 87705 with an improved fatty acid profile that results in enhanced nutritional characteristics. MON 87705 was developed to selectively supress two key enzymes, FATB and FAD2, involved in the soybean seed fatty acid biosynthetic pathway. As a result, MON 87705 soybean oil contains lower levels of saturated fatty acids (16:0 palmitic acid and 18:0 stearic acid) and higher levels of monounsaturated 18:1 oleic acid, with an associated decrease in polyunsaturated 18:2 linoleic acid levels relative to commodity soybean oil.

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

As part of a comprehensive safety assessment, bioinformatic analyses were performed to assess the potential for allergenicity, toxicity, or biological activity of putative polypeptides encoded by the 5′ and 3′ inserted DNA-soybean genomic DNA junctions (Tu and Silvanovich, 2009), and an additional junction that was created as part of a deletion and likely sequence duplication (Silvanovich and Tu, 2009). Periodically, the databases used to evaluate proteins are updated. Since the most recent report was completed, a new allergen (AD_2010), toxin (TOX_2010), and protein (PRT_2010) database have been assembled (Tu and Silvanovich, 2010). In order to determine if the putative polypeptides shared significant sequence similarity to new sequences contained in the updated allergen, toxin or protein databases, the 18 putative polypeptides from each reading frame of eight amino acids or greater in length were compared to AD_2010, TOX_2010, and PRT_2010 database sequences using bioinformatic tools.

The FASTA sequence alignment tool was used to assess structural relatedness between the query sequences and any protein sequences in the AD_2010, TOX_2010, and PRT_2010 databases. Structural similarities shared between each putative polypeptide with each sequence in the database were examined. The extent of structural relatedness was evaluated by detailed visual inspection of the alignment, the calculated percent identity, and the *E*-score. In addition to structural similarity, each putative polypeptide was screened for short polypeptide matches using a pair-wise comparison algorithm. In these analyses, eight contiguous and identical amino acids were defined as immunologically relevant, where eight represents the typical minimum sequence length likely to represent an immunological epitope.

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

No biologically relevant structural similarity to any sequence in the AD_2010, TOX_2010 or PRT_2010 databases were observed for any of the putative polypeptides. Furthermore, no short (eight amino acid) polypeptide matches were shared between any of the putative polypeptides and proteins in the AD_2010 database. These data demonstrate the lack of both structurally and immunologically relevant similarity to known allergens for all of the putative polypeptides analyzed. These data also demonstrate the lack of structurally relevant correlates to toxins or other biologically active proteins for the putative polypeptides analyzed.

This bioinformatics analysis is theoretical. No empirical evidence exists to suggest that transcription of DNA sequence at the additional junction described above for MON 87705 occurs. Rather, the results of these bioinformatic analyses indicate that in the highly unlikely occurrence that any of the junction sequences were to be transcribed and that a transcript were to be translated, the translation product would not share a sufficient degree of sequence similarity or identity to indicate that it would be potentially allergenic, toxic, or have other safety implications.

2.0 Sequence Database Preparation

The allergen, gliadin, and glutenin sequence database (AD_2010) was obtained from FARRP (2010)[1] and was used as provided. The AD_2010 database contains 1,471 sequences. A complete description of the AD_2010 database can be found in Tu and Silvanovich (2010).

GenBank protein database, release 175.0 (December 15, 2009), was downloaded from NCBI and formatted for use in these bioinformatic analyses. It is referred to herein as the PRT_2010 database and contains 17,815,538 sequences. A complete description of the PRT_2010 database can be found in Tu and Silvanovich (2010).

The toxin database is a subset of sequences derived from the PRT_2010 database that was selected using a keyword search and filtered to remove likely non-toxin proteins. It is referred to herein as the TOX_2010 database and contains 8,448 sequences. A complete description of the TOX_2010 database can be found in Tu and Silvanovich (2010).

3.0 Sequence Database Searches

The predicted sequence of polypeptides encoded by each reading frame at the 5´ and 3´ junctions (Table 1) of the MON 87705 insert were the same as those used in the previously conducted bioinformatics analysis described in Tu and Silvanovich (2009). The predicted sequence of polypeptides encoded by each reading frame at the additional 5´ junction (Table 2) contained in MON 87705 were the same as those used in the previously conducted bioinformatics analysis described in Silvanovich and Tu (2009).

FASTA analyses using the AD_2010, TOX_2010 and PRT_2010 databases were performed on a computer loaded with a SUSE LINUX version 10 operating system and FASTA version 3.4t 26 (July 7, 2006). The structural similarity of the translated protein sequences to sequences in each

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

database (AD_2010, TOX_2010 and PRT_2010) was assessed using the FASTA algorithm (Lipman and Pearson, 1985; Pearson and Lipman, 1988).

The structural similarity of translated sequence spanning the 5′ and 3′ junctions in MON 87705 to sequences in the AD_2010, TOX_2010 and PRT_2010 databases was assessed using the FASTA algorithm (Lipman and Pearson, 1985; Pearson and Lipman, 1988). FASTA comparisons are initiated by aligning the first match of a specific wordsize. The alignment is then extended based on the chosen scoring matrix. Specific FASTA comparison parameters used in this study included a FASTA assigned wordsize (*k-tuple*) of one or two, a gap creation penalty of 10, a gap extension penalty of two, and an expectation threshold (*E*-score) of one. FASTA comparisons were performed using the BLOSUM50 scoring matrix (Henikoff and Henikoff, 1992). The BLOSUM matrix series (Henikoff and Henikoff, 1992) was derived from a set of aligned, ungapped regions from protein families, called the BLOCKS database. Sequences from each block were clustered based on the percent of identical residues in the alignments (Henikoff and Henikoff, 1996). The BLOSUM50 matrix will identify blocks of conserved residues that are at least 50% identical. BLOSUM50 works well for identifying sequence similarities that include gaps, and thus recognizes distant evolutionary relationships (Pearson, 2000).

In addition to the FASTA comparisons of each putative polypeptide to allergens (to assess overall structural similarity), an eight amino acid sliding window search was performed. An algorithm was developed to identify whether or not a linearly contiguous match of eight amino acids existed between the query sequence and sequences within the allergen database (AD_2010). This program compares the query sequence to each protein sequence in the allergen database using a sliding-window of eight amino acids; that is, with a seven amino acid overlap relative to the preceding window.

4.0 Significance of the Alignments

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

5.0 Results

5.1 *Assessment of Potential Allergenicity*

No alignment met or exceeded the Codex (2003) FASTA alignment threshold for potential allergenicity of 35% identity over 80 amino acids when using the AD_2010 database to run a FASTA search (Tables 3, 6, and 9). Furthermore, no FASTA alignment displayed an *E*-score of less than or equal to 1e-5 and no eight amino acid matches were identified (Appendices 1-18).

5.2 *Assessment of Potential Toxicity*

None of the query sequences (Tables 4, 7, and 10) yielded alignments with *E*-scores of less than or equal to 1e-5 when using the TOX_2010 database to run a FASTA search (Appendices 1-18).

5.3 *Assessment of Potential Adverse Biological Activity*

Among all FASTA alignments between the 18 query sequences and the PRT_2010 database, one putative polypeptide (5_3) showed potentially significant alignments (Tables 5, 8, and 11; Appendices 1-18).

The putative polypeptide 5_3 showed 132 alignments with *E*-scores less than 1e-5 with the top alignment having an *E*-score of 2e-21 that corresponds to 47.22% identity in a window of 144 amino acids with an RNA-directed DNA polymerase from *Medicago truncatula* (GI number 124359710). Inspection of putative polypeptide 5_3 revealed that the soybean genome likely contains a RNA-directed DNA polymerase pseudo-gene on the 5' flank of the transgene insertion site. None of these alignments indicate that the putative polypeptide 5_3 possesses any bioactive function that would cause adverse effects to humans or animals. In the unlikely occurance that the putative polypeptide 5_3 sequence was to be translated, it does not share homology with a toxin or other bioactive protein.

6.0 Conclusions

The results of these bioinformatic analyses indicate that in the unlikely event that any of the peptides analyzed herein is found *in planta*, it would not share significant similarity or identity to known allergens, toxins or proteins of concern that could affect human or animal health.

7.0 References

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

FARRP, (Food Allergy Research and Resource Program Database) (2010). www.allergenonline.com. University of Nebraska.

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

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

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

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

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

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

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

Silvanovich, A. and Tu, H. 2009. Additional Bioinformatic Evaluation of DNA Sequences Flanking the 5′ Junction of the Inserted DNA in MON 87705: Assessment of Putative Polypeptides. Monsanto Technical Report MSL0022346, St. Louis, MO.

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

Tu, H. and Silvanovich, A. 2009. Bioinformatics Evaluation of DNA Sequences Flanking the 5′ and 3′ Junctions of Inserted DNA in MON 87705: Assessment of Putative Polypeptides. Monsanto Technical Report MSL0021929, St. Louis, MO.

Tu, H. and Silvanovich, A. 2010. The Assembly of Databases Used for FASTA, BLAST and Sliding Window Searches in 2010. Monsanto Technical Report MSL0022498, St. Louis, MO.

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.

Putative peptide ID	Putative peptide amino sequence
5_1a	CPNTHWWDSK STRGEQNVFY IINIQTKILK TVKQRIsgcc yhcglafgpr hrcpeknmrv vilakde
5_2a	NKGsqgvvit avwplgqgtv vlkki
5_3a	GSLVVPIMPK HPLVGLKIYK GRAECLLHHQ YPNQDSQDRE TKDlrvllsl rfglwakapl s
5_4a	DPLFHGLENL GLDIDDVEDI LLSPCRF
5_5a	qhpeILCFTV LRILVWILMM
5_6a	rvlqlsyffq dngalaqrpn rsdnntlrSF VSRS

Table 2. The predicted sequence of polypeptides encoded by each reading frame at the additional 5´ junction contained in MON 87705.
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 likely duplicated DNA.

Table 3. Summary of the best similarities for the FASTA search of the allergen sequence database (AD_2010) using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 5′ junction in MON 87705.

Appendix	Polypeptide	AD_2010 Sequence Database						
		Sliding Window	FASTA search					
		Hits	# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
1	5_1	No	-	-	-	-	-	-
2	5_2	No	-	-	-	-	-	-
3	5_3	No	3	2114497	30 kDa salivary gland allerg (253 aa)	32.075	53	0.97
4	5_4	No	-	-	-	-	-	-
5	5_5	No	1	75062228	ALL4_FELCA RecName: Full=Aller (186 aa)	29.730	37	0.88
6	5_6	No	-	-	-	-	-	-

Table 4. Summary of the best similarities for the FASTA search of the toxin sequence database (TOX_2010) using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 5′ junction in MON 87705.

Appendix	Polypeptide	TOX_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
1	5_1	6	225700708	exotoxin H precursor [Str (236 aa)	30.612	49	0.11
2	5_2	-	-	-	-	-	-
3	5_3	-	-	-	-	-	-
4	5_4	-	-	-	-	-	-
5	5_5	-	-	-	-	-	-
6	5_6	-	-	-	-	-	-

Table 5. Summary of the best similarities for the FASTA search of the protein sequence database (PRT_2010) using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 5′ junction in MON 87705.

Appendix	Polypeptide	PRT_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
1	5_1	-	-	-	-	-	-
2	5_2	-	-	-	-	-	-
3	5_3	219	124359710	RNA-directed DNA polymeras (1297 aa)	47.222	144	2e-21
4	5_4	-	-	-	-	-	-
5	5_5	-	-	-	-	-	-
6	5_6	-	-	-	-	-	-

Table 6. Summary of the best similarities for the FASTA search of the allergen sequence database (AD_2010) using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 3′ junction in MON 87705.

Appendix	Polypeptide	AD_2010 Sequence Database						
		Sliding Window	FASTA search					
		Hits	# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
7	3_1	No	-	-	-	-	-	-
8	3_2	No	-	-	-	-	-	-
9	3_3	No	2	71057064	thaumatin-like protein [Ac (225 aa)	40.000	25	0.83
10	3_4	No	2	169969	glycinin [Glycine max] (516 aa)	57.143	7	0.78
11	3_5	No	-	-	-	-	-	-
12	3_6	No	-	-	-	-	-	-

Table 7. Summary of the best similarities for the FASTA search of the toxin sequence database (TOX_2010) using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 3′ junction in MON 87705.

Appendix	Polypeptide	TOX_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
7	3_1	-	-	-	-	-	-
8	3_2	-	-	-	-	-	-
9	3_3	-	-	-	-	-	-
10	3_4	-	-	-	-	-	-
11	3_5	-	-	-	-	-	-
12	3_6	1	256378	neurotoxin Tx2-9 [Phoneutria (32 aa)	43.750	16	0.88

Table 8. Summary of the best similarities for the FASTA search of the protein sequence database (PRT_2010) using putative polypeptide sequences encoded by the genomic DNA-inserted DNA 3′ junction in MON 87705.

Appendix	Polypeptide	PRT_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
7	3_1	-	-	-	-	-	-
8	3_2	-	-	-	-	-	-
9	3_3	4	124360394	Peptidase aspartic, active (435 aa)	46.875	32	0.00047
10	3_4	-	-	-	-	-	-
11	3_5	-	-	-	-	-	-
12	3_6	-	-	-	-	-	-

Table 9. Summary of alignments for the FASTA searches of the AD_2010 database using putative polypeptide sequences encoded by the genomic DNA-likely duplicated DNA junction in MON 87705.

Appendix	Polypeptide	AD_2010 Sequence Database						
		Sliding Window	FASTA search					
		Hits	# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
13	5_1a	No	2	5059162	AF144060_1 alpha-amylase [Der (496 aa)	26.087	69	0.65
14	5_2a	No	-	-	-	-	-	-
15	5_3a	No	-	-	-	-	-	-
16	5_4a	No	-	-	-	-	-	-
17	5_5a	No	-	-	-	-	-	-
18	5_6a	No	2	729979	MAG_DERFA RecName: Full=Allergen (341 aa)	32.258	31	0.35

Table 10. Summary of alignments for the FASTA searches of the TOX_2010 database using putative polypeptide sequences encoded by the genomic DNA-likely duplicated DNA junction in MON 87705.

Appendix	Polypeptide	TOX_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
13	5_1a	-	-	-	-	-	-
14	5_2a	3	238624339	toxin of toxin-antitoxin s (134 aa)	36.842	19	0.15
15	5_3a	2	239523467	addiction module toxin [He (90 aa)	34.783	46	0.7
16	5_4a	-	-	-	-	-	-
17	5_5a	-	-	-	-	-	-
18	5_6a	-	-	-	-	-	-

Table 11. Summary of alignments for the FASTA searches of the PRT_2010 database using putative polypeptide sequences encoded by the genomic DNA-likely duplicated DNA junction in MON 87705.

Appendix	Polypeptide	PRT_2010 Sequence Database					
		# Hits	GI #	Description	% Identity	aa Overlap	*E*-score
13	5_1a	4	124360394	Peptidase aspartic, active (435 aa)	45.455	33	0.053
14	5_2a	-	-	-	-	-	-
15	5_3a	-	-	-	-	-	-
16	5_4a	-	-	-	-	-	-
17	5_5a	-	-	-	-	-	-
18	5_6a	-	-	-	-	-	-

Appendix 1. Bioinformatic analysis of polypeptide 5_1

```
>5_1
REKSLSLWFMSGKSTWISNEYDGQYGEKERVITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
```

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

Start time: Tue Jan 26 20:16:52 GMT 2010 Finish time: Tue Jan 26 20:16:53 GMT 2010

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

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

 5_1, 65 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     8     0:===
  22     0     0:            one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     0     2:*
  32     3     8:= *
  34     4    21:==    *
  36    16    44:======         *
  38    61    72:=====================  *
  40   119   101:===================================*======
  42   156   123:=========================================*===========
  44   121   136:=========================================     *
  46   128   138:==========================================  *
  48   114   132:=======================================     *
  50   114   121:=========================================  *
  52   136   106:===================================*==========
  54   115    91:==============================*========
  56    57    76:====================       *
  58    56    62:==================== *
  60    62    50:================*====
  62    23    40:=========     *
  64    34    32:==========*=
  66    25    25:=========*
  68    26    20:======*==
  70    22    16:=====*==
  72    13    12:===*=
  74    13    10:===*=
  76     9     7:==*
  78    16     6:=*====
  80     3     4:=*
  82     5     3:*=
  84     0     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:           *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.76170.0037; mu= 2.9576 0.193
 mean_var=53.599014.717, 0's: 8 Z-trim: 8  B-trim: 0 in 0/43
 Lambda= 0.175185
 Kolmogorov-Smirnov  statistic: 0.0432 (N=28) 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


65 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:16:52 2010 done: Tue Jan 26 20:16:52 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

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

 5_1, 65 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    70     0:=====
  22     0     0:            one = represents 17 library sequences
  24     0     0:
  26     0     0:
  28     2     2:*
  30     3    12:*
  32    22    45:==*
  34   103   122:=======*
  36   313   250:===============*====
  38   489   414:========================*====
  40   545   577:==================================*
  42   657   705:=======================================  *
  44   972   778:================================================*============
  46   779   793:================================================*
  48   629   759:=========================================       *
  50   792   692:=========================================*======
  52   502   609:==============================     *
  54   353   520:====================          *
  56   285   434:=================        *
```

```
  58   382   357:====================*==
  60   522   289:=================*==============
  62   259   232:==============*==
  64   180   184:==========*
  66   124   146:========*
  68    97   114:=======*
  70    61    90:==== *
  72    46    70:=== *
  74    56    55:===*
  76    55    43:==*=
  78    27    33:=*
  80    26    26:=*
  82    13    20:=*
  84    12    16:*
  86    16    12:*
  88    17     9:*          inset = represents 1 library sequences
  90    10     7:*
  92     5     6:*         :======*
  94    10     4:*         :===*======
  96     1     3:*         := *
  98     0     3:*         :  *
 100     0     2:*         : *
 102     1     2:*         :=*
 104     0     1:*         :*
 106     0     1:*         :*
 108     1     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     6     0:=         *======
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.65620.000642; mu= 4.1792 0.032
 mean_var=41.1635 9.082, 0's: 70 Z-trim: 76  B-trim: 473 in 2/60
 Lambda= 0.199902
 Kolmogorov-Smirnov  statistic: 0.0317 (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(8448)
gi|225700708|emb|CAW95321.1| exotoxin H precursor  ( 236)   83 30.1    0.11
gi|209540528|gb|ACI61104.1| Streptococcal pyrogeni ( 236)   83 30.1    0.11
gi|4838430|gb|AAD30989.1|AF124500_1 exotoxin H pre ( 236)   81 29.6    0.16
gi|94541990|gb|ABF32039.1| enterotoxin [Streptococ ( 236)   81 29.6    0.16
gi|134272104|emb|CAM30348.1| streptococcal exotoxi ( 236)   81 29.6    0.16
gi|13622161|gb|AAK33907.1| streptococcal exotoxin  ( 236)   81 29.6    0.16


>>gi|225700708|emb|CAW95321.1| exotoxin H precursor [Str  (236 aa)
 initn:  49 init1:  49 opt:  83  Z-score: 133.2  bits: 30.1 E(): 0.11
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|225 LYKHDSNLIEADSIKNSPDIVTSHMLKYSVKDKNLSVFF---EKDWISQEFK----DKEV
          50        60        70        80           90

               40        50        60
5_1    VITNFFSIQKCRCPQRYYKMKVHFDKTTNYDPSYL
        :  . . . :.:: . :.
gi|225 DIYALSAQEACECPGKRYEAFGGITLTNSEKKEIKVPINVWDKSKQHPPMFITVNKPKVT
       100       110       120       130       140       150

>>gi|209540528|gb|ACI61104.1| Streptococcal pyrogenic ex  (236 aa)
 initn:  49 init1:  49 opt:  83  Z-score: 133.2  bits: 30.1 E(): 0.11
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: 130.1  bits: 29.6 E(): 0.16
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 [Streptococcus   (236 aa)
 initn:  47 init1:  47 opt:  81  Z-score: 130.1  bits: 29.6 E(): 0.16
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|134272104|emb|CAM30348.1| streptococcal exotoxin H   (236 aa)
 initn:  47 init1:  47 opt:  81  Z-score: 130.1  bits: 29.6 E(): 0.16
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

>>gi|13622161|gb|AAK33907.1| streptococcal exotoxin H pr  (236 aa)
 initn:  47 init1:  47 opt:  81  Z-score: 130.1  bits: 29.6 E(): 0.16
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




65 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:16:53 2010 done: Tue Jan 26 20:16:53 2010
 Total Scan time:  0.110 Total Display time:  0.000

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

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

 5_1, 65 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 344415     0:==============
  22   133     0:=          one = represents 27318 library sequences
  24   258    17:*
  26   623   374:*
  28  2458  4039:*
  30 13139 24538:*
  32 61084 94880:===*
  34 212140 257303:======== *
  36 526960 528439:====================*
  38 919934 873314:==================================*==
  40 1377682 1218196:=============================================*======
  42 1587761 1489096:=========================================================*====
  44 1639038 1642612:============================================================*
  46 1583043 1673041:============================================================ *
  48 1496460 1601744:============================================================   *
  50 1317849 1461597:==============================================================    *
  52 1179947 1284988:==============================================   *
  54 995378 1097605:=====================================   *
  56 854922 916837:================================== *
  58 736512 752706:===========================*
  60 605088 609736:======================*
  62 498051 488827:==================*=
  64 399378 388761:===============*
  66 314001 307265:============*
  68 249637 241688:========*=
  70 198139 189401:=======*=
  72 152886 147999:======*
  74 128540 115390:=====*
  76 101900 89811:===*
  78 74503 69809:==*
  80 57404 54205:=*=
  82 44630 41466:=*
  84 33854 32846:=*
  86 25437 25415:*
  88 19374 19664:*          inset = represents 218 library sequences
  90 15119 15215:*
  92 10896 11773:*         :=========================================*
  94  8369  9109:*         :=========================================*
  96  6679  7048:*         :================================== *
  98  4732  5454:*         :======================   *
 100  3462  4220:*         :=================   *
 102  2373  3265:*         :============   *
 104  1796  2526:*         :==========  *
 106  1290  1955:*         :=======  *
 108  1156  1512:*         :=======*
 110   828  1170:*         :===== *
 112   542   905:*         :=== *
 114   493   701:*         :===*
 116   318   542:*         :==*
 118   179   419:*         :=*
>120   435   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810856 sequences
  Expectation_n fit: rho(ln(x))= 3.98400.000186; mu= 7.2664 0.010
 mean_var=46.0828 9.234, 0's: 1150 Z-trim: 1151  B-trim: 0 in 0/64
 Lambda= 0.188932
 Kolmogorov-Smirnov  statistic: 0.0241 (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


65 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:16:53 2010 done: Tue Jan 26 20:21:40 2010
 Total Scan time: 240.890 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_2010
Query = 5_2

Start time: Tue Jan 26 20:21:40 GMT 2010 Finish time: Tue Jan 26 20:21:40 GMT 2010

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

# fasta34 5_2.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_2.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7, 2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_2, 25 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     2     2:*
  32     9     8:==*
  34    18    21:======*
  36    29    44:==========    *
  38    88    72:========================*======
  40    86   101:============================    *
  42   116   123:======================================= *
  44   126   136:============================================   *
  46   151   138:================================================*=====
  48   116   132:=======================================    *
  50   103   121:==================================     *
  52   135   106:===================================*=========
  54    72    91:========================      *
  56    71    76:======================== *
  58    53    62:==================  *
  60    57    50:================*==
  62    47    40:=============*==
  64    50    32:==========*======
  66    26    25:=========*
  68     9    20:===   *
  70    15    16:=====*
  72    58    12:===*================
  74    21    10:===*===
  76     7     7:==*
  78     0     6: *
  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:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.10580.0031; mu= 9.5201 0.161
 mean_var=29.2718 7.147, 0's: 2 Z-trim: 2  B-trim: 0 in 0/43
 Lambda= 0.237055
 Kolmogorov-Smirnov  statistic: 0.0390 (N=25) at  58

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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:21:40 2010 done: Tue Jan 26 20:21:40 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

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

       opt      E()
< 20    60     0:=====
  22     0     0:           one = represents 14 library sequences
  24     1     0:=
  26    13     0:=
  28     5     2:*
  30    14    12:*
  32   112    45:===*====
  34   238   122:========*========
  36   140   250:==========       *
  38   313   414:========================      *
  40   651   577:=============================================*=====
  42   585   706:============================================         *
  44   773   779:============================================================*
  46   792   793:============================================================*
  48   729   759:============================================================ *
  50   510   693:=======================================            *
  52   577   609:========================================== *
  54   763   520:====================================*=================
  56   415   435:============================== *
  58   406   357:=========================*===
  60   380   289:====================*=======
  62   175   232:=============   *
  64   130   184:==========   *
  66   120   146:========= *
```

```
  68    99   115:========*
  70    87    90:=======*
  72    54    70:====*
  74   102    55:===*====
  76    24    43:== *
  78    31    33:==*
  80    60    26:=*===
  82    32    20:=*=
  84    15    16:=*
  86     6    12:*
  88    10     9:*          inset = represents 1 library sequences
  90     7     7:*
  92     1     6:*         :=    *
  94     1     4:*         :=  *
  96     7     3:*         :==*====
  98     0     3:*         :  *
 100     0     2:*         : *
 102     1     2:*         :=*
 104     3     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:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 1.98470.0005; mu= 14.8131 0.025
 mean_var=25.7775 5.921, 0's: 60 Z-trim: 60  B-trim: 125 in 1/61
 Lambda= 0.252612
 Kolmogorov-Smirnov  statistic: 0.0368 (N=29) at  52

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


25 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:21:40 2010 done: Tue Jan 26 20:21:41 2010
 Total Scan time:  0.140 Total Display time:  0.000

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

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

 5_2, 25 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 276664     0:===========
  22    532     0:=          one = represents 27351 library sequences
  24    966    17:*
  26   3198   374:*
  28 11741  4039:*
  30 34560 24537:*=
  32 103403 94878:===*
  34 239757 257298:=========*
  36 460382 528430:================  *
  38 772491 873297:============================  *
  40 1070517 1218173:=========================================    *
  42 1340704 1489068:================================================    *
  44 1519924 1642581:============================================================   *
  46 1641002 1673010:=============================================================*
  48 1627345 1601714:============================================================*=
  50 1469522 1461570:=========================================================*
  52 1322470 1284964:==================================================*==
  54 1151633 1097585:============================================*==
  56 910165 916820:==================================*
  58 752869 752692:==========================*
  60 610547 609724:======================*
  62 514179 488818:=================*=
  64 406636 388754:==============*
  66 328540 307260:===========*=
  68 255705 241684:========*=
  70 212122 189398:======*=
  72 183690 147996:=====*=
  74 140223 115388:====*=
  76 106093 89809:===*
  78 78225 69807:==*
  80 60282 54204:=*=
  82 46482 41465:=*
  84 38233 32845:=*
  86 28792 25414:*=
  88 21562 19664:*          inset = represents 268 library sequences
  90 17498 15215:*
  92 13391 11773:*         :=======================================*
  94  9462  9109:*         :==================================*==
  96  6610  7048:*         :========================== *
  98  5924  5453:*         :====================*==
 100  3625  4220:*         :============== *
 102  2645  3265:*         :==========  *
 104  2574  2526:*         :=========*
 106  2411  1955:*         :=======*=
 108  1623  1512:*         :======*=
 110  1312  1170:*         :====*
 112   824   905:*         :===*
 114   587   701:*         :==*
 116   391   542:*         :==*
 118   416   419:*         :=*
>120   776   325:*         :=*=
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810527 sequences
  Expectation_n fit: rho(ln(x))= 3.70360.000175; mu= 6.5424 0.009
 mean_var=26.3173 5.324, 0's: 924 Z-trim: 925  B-trim: 6248 in 2/64
 Lambda= 0.250008
 Kolmogorov-Smirnov  statistic: 0.0346 (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
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:21:41 2010 done: Tue Jan 26 20:27:30 2010
 Total Scan time: 299.910 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
ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALSRRDEDKELQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSH
PQYTLECDILYFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREIYMDQQ
```

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

Start time: Tue Jan 26 20:27:31 GMT 2010 Finish time: Tue Jan 26 20:27:31 GMT 2010

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

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

 5_3, 161 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     1     0:=
  30     2     2:*
  32     2     8:= *
  34     7    21:===   *
  36    52    44:==============*===
  38    80    72:========================*===
  40    98   101:====================================*
  42   107   123:=======================================    *
  44   125   136:============================================    *
  46   106   138:====================================         *
  48   121   132:=============================================  *
  50   138   121:============================================*=====
  52   115   106:==================================*===
  54    78    91:==========================    *
  56    77    76:=========================*
  58    68    62:====================*==
  60    57    50:================*==
  62    61    40:=============*=======
  64    46    32:==========*=====
  66    31    25:========*==
  68    27    20:======*==
  70    16    16:=====*
  72    11    12:===*
  74     9    10:===*
  76     5     7:==*
  78     6     6:=*
  80     1     4:=*
  82     2     3:*
  84     3     3:*
  86     4     2:*=
  88     0     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     3     1:*         :*==
  94     0     1:*         :*
  96     3     1:*         :*==
  98     1     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:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 5.54430.00396; mu= 3.0405 0.205
 mean_var=58.634215.818, 0's: 2 Z-trim: 2  B-trim: 28 in 1/42
 Lambda= 0.167494
 Kolmogorov-Smirnov  statistic: 0.0517 (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(1471)
gi|2114497|gb|AAB58417.1| 30 kDa salivary gland al ( 253)   74 25.9    0.97
gi|232054|sp|P30575.1|ENO1_CANAL RecName: Full=Eno ( 440)   77 26.7    0.98
gi|94468552|gb|ABF18125.1| 30 kDa salivary gland a ( 258)   74 25.9    0.99


>>gi|2114497|gb|AAB58417.1| 30 kDa salivary gland allerg  (253 aa)
 initn:  66 init1:  66 opt:  74  Z-score: 102.6  bits: 25.9 E(): 0.97
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|232054|sp|P30575.1|ENO1_CANAL RecName: Full=Enolase  (440 aa)
 initn:  51 init1:  51 opt:  77  Z-score: 102.5  bits: 26.7 E(): 0.98
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|94468552|gb|ABF18125.1| 30 kDa salivary gland aller  (258 aa)
 initn:  66 init1:  66 opt:  74  Z-score: 102.5  bits: 25.9 E(): 0.99
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




161 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:27:31 2010 done: Tue Jan 26 20:27:31 2010
 Total Scan time:  0.030 Total Display time:  0.000

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

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

 5_3, 161 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    65     0:=====
  22     1     0:=          one = represents 13 library sequences
  24     1     0:=
  26     7     0:=
  28    18     2:*=
  30    44    12:*===
  32    47    45:===*
  34   129   122:==========*
  36   260   250:====================*
  38   557   414:================================*===========
  40   514   577:=========================================    *
  42   670   706:=====================================================  *
  44   703   779:=========================================================    *
  46   774   793:============================================================*
  48   664   759:==========================================================      *
  50   544   693:============================================           *
  52   734   609:=============================================*==========
  54   498   520:=======================================*
  56   339   435:==========================      *
  58   295   357:=======================    *
  60   573   289:======================*======================
  62   275   232:=================*====
  64   214   184:==============*==
  66   122   146:========== *
  68    94   115:========*
  70    42    90:====  *
  72    57    70:=====*
  74    58    55:====*
  76    17    43:== *
  78    29    33:==*
  80    26    26:=*
  82    16    20:=*
  84     3    16:=*
  86     9    12:*
  88    14     9:*=         inset = represents 1 library sequences
  90     6     7:*
  92     8     6:*         :=====*==
  94     2     4:*         :== *
  96     0     3:*         :  *
  98     1     3:*         := *
 100     7     2:*         :=*=====
 102     1     2:*         :=*
 104     0     1:*         :*
 106     3     1:*         :*==
 108     0     1:*         :*
 110     1     1:*         :*
 112     0     0:          *
 114     1     0:=         *=
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 5.57430.000598; mu= 3.7499 0.030
 mean_var=46.821210.058, 0's: 60 Z-trim: 62  B-trim: 0 in 0/62
 Lambda= 0.187436
 Kolmogorov-Smirnov  statistic: 0.0327 (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


161 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:27:31 2010 done: Tue Jan 26 20:27:32 2010
 Total Scan time:  0.180 Total Display time:  0.000

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

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

 5_3, 161 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 286286     0:===========
  22   311     0:=          one = represents 28050 library sequences
```

```
  24   453    17:*
  26  1207   374:*
  28  4158  4039:*
  30 20375 24537:*
  32 86844 94877:===*
  34 264816 257295:=========*
  36 606051 528423:==================*===
  38 951230 873286:==============================*==
  40 1319363 1218157:=============================================*====
  42 1533104 1489049:=======================================================*=
  44 1648113 1642559:============================================================*
  46 1682945 1672988:============================================================*
  48 1556835 1601693:============================================================ *
  50 1410100 1461551:================================================ *
  52 1209744 1284947:========================================== *
  54 1016691 1097570:==================================  *
  56 841127 916808:============================  *
  58 689639 752682:======================== *
  60 565846 609716:====================*
  62 460888 488811:================*
  64 360216 388749:=============*
  66 279232 307255:==========*
  68 230015 241681:========*
  70 184180 189395:======*
  72 144542 147994:=====*
  74 108118 115386:====*
  76 80600 89808:===*
  78 65900 69806:==*
  80 48884 54203:=*
  82 37683 41464:=*
  84 28416 32845:=*
  86 21015 25414:*
  88 16268 19664:*          inset = represents 192 library sequences
  90 11678 15215:*
  92  9566 11772:*         :=======================================*
  94  7428  9109:*         :=======================================*
  96  5486  7048:*         :============================       *
  98  4438  5453:*         :========================    *
 100  2996  4220:*         :================     *
 102  2324  3265:*         :=============    *
 104  1563  2526:*         :=========    *
 106  1065  1955:*         :======    *
 108   854  1512:*         :=====  *
 110   517  1170:*         :===   *
 112   410   905:*         :=== *
 114   331   701:*         :== *
 116   216   542:*         :==*
 118   159   419:*         := *
>120   999   325:*         :=*====
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810291 sequences
  Expectation_n fit: rho(ln(x))= 5.08540.000185; mu= 6.8690 0.010
 mean_var=62.719212.921, 0's: 986 Z-trim: 987  B-trim: 2359 in 1/63
 Lambda= 0.161947
 Kolmogorov-Smirnov  statistic: 0.0179 (N=29) at  46

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(17815538)
gi|124359710|gb|ABN06064.1| RNA-directed DNA polym (1297)  437 110.5   2e-21
gi|147807720|emb|CAN66553.1| hypothetical protein  (1448)  422 107.0 2.5e-20
gi|147854459|emb|CAN78588.1| hypothetical protein  (2232)  408 103.8 3.5e-19
gi|124360619|gb|ABD33394.2| FAR1; Polynucleotidyl  ( 657)  318 82.5 2.7e-13
gi|6466937|gb|AAF13073.1|AC011621_1 putative retro (1661)  319 83.0   5e-13
gi|208609060|dbj|BAG72152.1| hypothetical protein  (1369)  315 82.0 8.2e-13
gi|208609057|dbj|BAG72151.1| hypothetical protein  (1558)  315 82.0 9.1e-13
gi|208609055|dbj|BAG72150.1| hypothetical protein  (1558)  315 82.0 9.1e-13
gi|208609065|dbj|BAG72154.1| hypothetical protein  (1558)  315 82.0 9.1e-13
gi|208609051|dbj|BAG72148.1| hypothetical protein  (1558)  314 81.8 1.1e-12
gi|208609062|dbj|BAG72153.1| hypothetical protein  (1558)  314 81.8 1.1e-12
gi|208609053|dbj|BAG72149.1| hypothetical protein  (1520)  312 81.3 1.4e-12
gi|208609049|dbj|BAG72147.1| hypothetical protein  (1520)  312 81.3 1.4e-12
gi|147828709|emb|CAN66228.1| hypothetical protein  (1258)  308 80.3 2.4e-12
gi|147843077|emb|CAN83300.1| hypothetical protein  (1366)  302 79.0 6.7e-12
gi|147774273|emb|CAN76793.1| hypothetical protein  (1469)  301 78.7 8.3e-12
gi|147772855|emb|CAN73669.1| hypothetical protein  (1308)  299 78.2   1e-11
gi|147789424|emb|CAN66607.1| hypothetical protein  (2822)  302 79.1 1.2e-11
gi|8778789|gb|AAF79797.1|AC020646_20 T32E20.30 [Ar (1397)  297 77.8 1.5e-11
gi|170660047|gb|ACB28472.1| polyprotein [Ananas co ( 953)  292 76.5 2.5e-11
gi|222635621|gb|EEE65753.1| hypothetical protein O ( 645)  289 75.7   3e-11
gi|255677065|dbj|BAF19594.2| Os06g0493000 [Oryza s ( 866)  289 75.8 3.8e-11
gi|147823217|emb|CAN75278.1| hypothetical protein  ( 716)  285 74.8 6.2e-11
gi|116309032|emb|CAH66146.1| OSIGBa0114M03.4 [Oryz (1448)  285 75.0 1.1e-10
gi|155327954|gb|ABT83558.1| Sequence 171028 from p ( 161)  270 70.9   2e-10
gi|77551099|gb|ABA93896.1| retrotransposon protein ( 897)  278 73.2 2.3e-10
gi|77557165|gb|ABA99961.1| retrotransposon protein (1619)  279 73.6 3.2e-10
gi|147768278|emb|CAN60449.1| hypothetical protein  ( 647)  274 72.2 3.4e-10
gi|32488663|emb|CAE03590.1| OSJNBa0087O24.13 [Oryz (1311)  277 73.1 3.7e-10
gi|27764548|gb|AAO23078.1| polyprotein [Glycine ma (1552)  276 72.9   5e-10
gi|147786920|emb|CAN64437.1| hypothetical protein  ( 623)  271 71.5 5.3e-10
gi|108862596|gb|ABA97714.2| retrotransposon protei (1287)  272 71.9 8.2e-10
gi|147779107|emb|CAN73467.1| hypothetical protein  (1593)  272 72.0 9.8e-10
gi|215767834|dbj|BAH00063.1| unnamed protein produ ( 494)  264 69.8 1.4e-09
gi|147866223|emb|CAN81986.1| hypothetical protein  ( 672)  265 70.1 1.5e-09
gi|147864527|emb|CAN80491.1| hypothetical protein  (1412)  268 71.0 1.7e-09
gi|255678117|dbj|BAH94093.1| Os08g0123950 [Oryza s ( 783)  264 69.9   2e-09
gi|241943804|gb|EES16949.1| hypothetical protein S (1456)  267 70.8   2e-09
gi|108864659|gb|ABA95357.2| retrotransposon protei (2811)  270 71.7 2.2e-09
gi|241933637|gb|EES06782.1| hypothetical protein S (1414)  265 70.3 2.8e-09
gi|77554607|gb|ABA97403.1| retrotransposon protein ( 562)  259 68.7 3.4e-09
gi|62734016|gb|AAX96125.1| retrotransposon protein (1448)  263 69.9 3.9e-09
gi|77550700|gb|ABA93497.1| retrotransposon protein (1448)  263 69.9 3.9e-09
gi|241917803|gb|EER90947.1| hypothetical protein S (1450)  263 69.9 3.9e-09
gi|108706172|gb|ABF93967.1| retrotransposon protei (1461)  263 69.9 3.9e-09
gi|15451607|gb|AAK98731.1|AC090485_10 Putative ret (1461)  263 69.9 3.9e-09
gi|90399077|emb|CAJ86299.1| H0124B04.16 [Oryza sat (1265)  261 69.4 4.8e-09
gi|113611293|dbj|BAF21671.1| Os07g0510800 [Oryza s ( 517)  256 68.0 5.2e-09
gi|110289541|gb|AAP54937.2| retrotransposon protei (1477)  260 69.2 6.4e-09
gi|116310096|emb|CAH67116.1| H0502G05.7 [Oryza sat ( 642)  255 67.8 7.3e-09
gi|257741008|emb|CBC66937.1| unnamed protein produ ( 753)  255 67.8 8.3e-09
gi|219726516|emb|CAW33744.1| unnamed protein produ ( 753)  255 67.8 8.3e-09
gi|219904233|emb|CAW74225.1| unnamed protein produ ( 753)  255 67.8 8.3e-09
gi|38567865|emb|CAE03018.3| OSJNBa0091D06.3 [Oryza ( 753)  255 67.8 8.3e-09
gi|257319990|emb|CBC21422.1| unnamed protein produ ( 753)  255 67.8 8.3e-09
gi|219735497|emb|CAW51408.1| unnamed protein produ ( 753)  255 67.8 8.3e-09
gi|257292212|emb|CBB99778.1| unnamed protein produ ( 753)  255 67.8 8.3e-09
gi|15217201|gb|AAK92545.1|AC051624_3 Putative retr (1476)  257 68.5   1e-08
gi|113638940|dbj|BAF26245.1| Os10g0317000 [Oryza s (1476)  257 68.5   1e-08
gi|31431040|gb|AAP52878.1| retrotransposon protein (1476)  257 68.5   1e-08
gi|155282606|gb|ABT38210.1| Sequence 125680 from p ( 218)  246 65.4 1.3e-08
gi|241916679|gb|EER89823.1| hypothetical protein S (1414)  255 68.0 1.4e-08
gi|241936478|gb|EES09623.1| hypothetical protein S (1507)  255 68.0 1.5e-08
gi|225016158|gb|ACN78981.1| retrotransposon protei (1261)  254 67.7 1.5e-08
gi|225016150|gb|ACN78974.1| retrotransposon protei (1261)  254 67.7 1.5e-08
```

gi|116309424|emb|CAH66499.1| H0321H01.8 [Oryza sat	(1602)	255	68.0	1.5e-08
gi|108710432|gb|ABF98227.1| retrotransposon protei	(1160)	253	67.5	1.6e-08
gi|215695456|dbj|BAG90661.1| unnamed protein produ	(459)	248	66.1	1.7e-08
gi|215737344|dbj|BAG96273.1| unnamed protein produ	(459)	248	66.1	1.7e-08
gi|31712084|gb|AAP68389.1| putative polyprotein [O	(1246)	253	67.5	1.7e-08
gi|18378611|gb|AAL68643.1|AF458767_1 polyprotein [	(775)	250	66.7	1.9e-08
gi|108707050|gb|ABF94845.1| retrotransposon protei	(937)	250	66.7	2.2e-08
gi|38344579|emb|CAE05537.2| OSJNBa0053B21.11 [Oryz	(1251)	250	66.8	2.8e-08
gi|147798109|emb|CAN73897.1| hypothetical protein	(269)	242	64.5	2.9e-08
gi|110289426|gb|AAP54661.2| retrotransposon protei	(760)	247	66.0	3e-08
gi|77554308|gb|ABA97104.1| retrotransposon protein	(1412)	250	66.8	3.1e-08
gi|52353389|gb|AAU43957.1| putative polyprotein [O	(1263)	249	66.6	3.3e-08
gi|10122041|gb|AAG13430.1|AC051634_11 putative pla	(921)	247	66.0	3.6e-08
gi|108706266|gb|ABF94061.1| retrotransposon protei	(1159)	248	66.3	3.7e-08
gi|21397271|gb|AAM51835.1|AC105730_9 Putative plan	(1159)	248	66.3	3.7e-08
gi|241943825|gb|EES16970.1| hypothetical protein S	(1462)	249	66.6	3.8e-08
gi|31432119|gb|AAP53789.1| retrotransposon protein	(1611)	249	66.6	4.1e-08
gi|18568269|gb|AAL76001.1|AF466646_9 putative gag-	(2396)	251	67.2	4.1e-08
gi|38346036|emb|CAE01900.2| OSJNBa0059D20.8 [Oryza	(1463)	248	66.4	4.4e-08
gi|113563772|dbj|BAF14115.1| Os04g0191000 [Oryza s	(1463)	248	66.4	4.4e-08
gi|38605839|emb|CAE02919.3| OSJNBb0108J11.11 [Oryz	(815)	245	65.5	4.5e-08
gi|167249400|gb|ABZ25960.1| Sequence 85 from paten	(2437)	250	67.0	4.9e-08
gi|112064846|gb|ABH99846.1| Sequence 85 from paten	(2437)	250	67.0	4.9e-08
gi|32492359|emb|CAE05990.1| OSJNBa0004L19.22 [Oryz	(1586)	247	66.1	5.6e-08
gi|31430458|gb|AAP52367.1| retrotransposon protein	(415)	239	63.9	6.8e-08
gi|38344156|emb|CAD41876.2| OSJNBa0041A02.23 [Oryz	(1373)	245	65.6	6.9e-08
gi|241916683|gb|EER89827.1| hypothetical protein S	(1437)	245	65.7	7.1e-08
gi|227438239|gb|ACP30609.1| disease resistance pro	(2726)	248	66.5	7.4e-08
gi|38347666|emb|CAE05600.2| OSJNBa0054D14.1 [Oryza	(1629)	245	65.7	7.9e-08
gi|38346992|emb|CAD40278.2| OSJNBb0062H02.17 [Oryz	(1629)	245	65.7	7.9e-08
gi|22711553|gb|AAM01152.2|AC113336_4 Putative retr	(575)	239	64.0	8.8e-08
gi|147802203|emb|CAN70510.1| hypothetical protein	(790)	240	64.3	9.8e-08
gi|38344463|emb|CAE04934.2| OSJNBa0017P10.11 [Oryz	(1012)	241	64.6	1e-07
gi|38347368|emb|CAE04958.2| OSJNBa0070D17.9 [Oryza	(394)	236	63.2	1.1e-07
gi|147811718|emb|CAN77256.1| hypothetical protein	(1365)	242	64.9	1.1e-07
gi|18378613|gb|AAL68644.1|AF458768_1 polyprotein [	(933)	240	64.4	1.1e-07
gi|7800101|gb|AAF69810.1| pol protein [Picea abies	(148)	230	61.6	1.2e-07
gi|38345562|emb|CAE03436.2| OSJNBa0032F06.19 [Oryz	(1575)	240	64.5	1.7e-07
gi|89887334|gb|ABD78322.1| polyprotein [Primula vu	(1359)	239	64.2	1.8e-07
gi|10140673|gb|AAG13508.1|AC068924_13 putative gag	(1608)	237	63.8	2.9e-07
gi|32489310|emb|CAE03706.1| OSJNBa0060B20.14 [Oryz	(3200)	240	64.7	3.1e-07
gi|108710435|gb|ABF98230.1| retrotransposon protei	(1118)	234	63.0	3.4e-07
gi|31712076|gb|AAP68381.1| putative polyprotein [O	(1118)	234	63.0	3.4e-07
gi|108864085|gb|ABA91843.2| retrotransposon protei	(1411)	235	63.3	3.5e-07
gi|62733109|gb|AAX95226.1| retrotransposon protein	(1513)	235	63.3	3.7e-07
gi|53981172|gb|AAV24812.1| putative polyprotein [O	(1475)	234	63.1	4.3e-07
gi|77552522|gb|ABA95319.1| retrotransposon protein	(955)	230	62.0	5.8e-07
gi|38346427|emb|CAD40214.2| OSJNBa0019J05.12 [Oryz	(1817)	231	62.4	8.3e-07
gi|77555174|gb|ABA97970.1| retrotransposon protein	(1548)	227	61.5	1.4e-06
gi|108706171|gb|ABF93966.1| retrotransposon protei	(1920)	228	61.8	1.4e-06
gi|15451608|gb|AAK98732.1|AC090485_11 Putative ret	(1923)	228	61.8	1.4e-06
gi|147854038|emb|CAN83399.1| hypothetical protein	(234)	217	58.6	1.5e-06
gi|4581164|gb|AAD24647.1| putative retroelement po	(780)	223	60.4	1.5e-06
gi|12322948|gb|AAG51464.1|AC069160_10 gypsy/Ty3 el	(1447)	226	61.2	1.6e-06
gi|22725944|gb|AAN04954.1| Putative retroelement [	(813)	221	59.9	2.2e-06
gi|78708171|gb|ABB47146.1| retrotransposon protein	(813)	221	59.9	2.2e-06
gi|62734404|gb|AAX96513.1| retrotransposon protein	(605)	219	59.4	2.4e-06
gi|108864289|gb|ABA92870.2| retrotransposon protei	(621)	219	59.4	2.4e-06
gi|194706326|gb|ACF87247.1| unknown [Zea mays]	(360)	216	58.5	2.5e-06
gi|147783182|emb|CAN68669.1| hypothetical protein	(1360)	222	60.3	2.8e-06
gi|155323104|gb|ABT78708.1| Sequence 166178 from p	(237)	210	57.0	4.7e-06
gi|78183243|emb|CAJ00274.1| hypothetical protein [	(1508)	219	59.6	5e-06
gi|78183241|emb|CAJ00278.1| hypothetical protein [	(1508)	218	59.4	5.9e-06
gi|155364082|gb|ABU19687.1| Sequence 207156 from p	(211)	207	56.3	6.9e-06
gi|12322008|gb|AAG51046.1|AC069473_8 gypsy/Ty-3 re	(1499)	216	58.9	8.1e-06
gi|10998138|dbj|BAB03109.1| retroelement pol polyp	(1499)	216	58.9	8.1e-06
gi|112065980|gb|ABI00028.1| Sequence 238 from pate	(1499)	216	58.9	8.1e-06
gi|7800097|gb|AAF69808.1| pol protein [Picea abies	(147)	200	54.6	1.6e-05
gi|5080762|gb|AAD39272.1|AC007203_4 Similar to ret	(1264)	207	56.7	3e-05
gi|145012540|gb|EDJ97194.1| hypothetical protein M	(1071)	206	56.5	3.1e-05
gi|29788873|gb|AAP03419.1| putative polyprotein [O	(652)	201	55.2	4.6e-05
gi|18652523|gb|AAL77156.1|AC091732_7 Putative poly	(999)	203	55.7	4.7e-05
gi|31431768|gb|AAP53494.1| retrotransposon protein	(999)	203	55.7	4.7e-05
gi|21672107|gb|AAM74469.1|AC124213_27 Putative ret	(999)	203	55.7	4.7e-05
gi|78183249|emb|CAJ00277.1| hypothetical protein [	(1508)	204	56.1	5.7e-05
gi|20270059|gb|AAM18147.1|AC092172_7 Putative gag-	(1338)	202	55.6	7.1e-05
gi|18958673|gb|AAL82656.1|AC092387_4 retrotranspos	(1338)	202	55.6	7.1e-05
gi|78708062|gb|ABB47037.1| retrotransposon protein	(1347)	202	55.6	7.1e-05
gi|116309666|emb|CAH66715.1| OSIGBa0118P15.5 [Oryz	(1434)	201	55.4	8.8e-05
gi|78183245|emb|CAJ00275.1| hypothetical protein [	(1112)	198	54.6	0.00012
gi|147770944|emb|CAN69535.1| hypothetical protein	(617)	195	53.8	0.00012
gi|38344578|emb|CAE05536.2| OSJNBa0053B21.10 [Oryz	(582)	193	53.3	0.00015
gi|155300190|gb|ABT55794.1| Sequence 143264 from p	(144)	184	50.8	0.00021
gi|155333463|gb|ABT89067.1| Sequence 176537 from p	(144)	184	50.8	0.00021
gi|77551464|gb|ABA94261.1| retrotransposon protein	(1369)	191	53.0	0.00043
gi|19881710|gb|AAM01111.1|AC098682_15 putative ret	(1043)	189	52.5	0.00047
gi|116309348|emb|CAH66431.1| OSIGBa0096P03.5 [Oryz	(413)	181	50.4	0.00081
gi|147841216|emb|CAN64356.1| hypothetical protein	(1852)	188	52.4	0.00089
gi|113564206|dbj|BAF14549.1| Os04g0389000 [Oryza s	(538)	181	50.4	0.001
gi|7800092|gb|AAF69806.1| pol protein [Picea abies	(83)	170	47.4	0.0013
gi|215704780|dbj|BAG94808.1| unnamed protein produ	(218)	173	48.3	0.0017
gi|147855151|emb|CAN81740.1| hypothetical protein	(771)	179	50.1	0.0019
gi|108862432|gb|ABA97314.2| retrotransposon protei	(1219)	181	50.7	0.002
gi|190688747|gb|ACE86410.1| putative retroelement	(1029)	179	50.1	0.0024
gi|57863925|gb|AAS55774.2| putative polyprotein [O	(2108)	179	50.3	0.0043
gi|147768682|emb|CAN76063.1| hypothetical protein	(1453)	176	49.5	0.0051
gi|147769722|emb|CAN69702.1| hypothetical protein	(1454)	176	49.5	0.0051
gi|147864892|emb|CAN79373.1| hypothetical protein	(1439)	175	49.3	0.006
gi|190688734|gb|ACE86397.1| putative retroelement	(732)	170	48.0	0.0077
gi|147860462|emb|CAN82562.1| hypothetical protein	(1384)	173	48.8	0.008
gi|77556305|gb|ABA99101.1| retrotransposon protein	(1550)	173	48.8	0.0088
gi|38568032|emb|CAD40406.3| OSJNBa0065J03.2 [Oryza	(1629)	173	48.9	0.0091
gi|147810925|emb|CAN71787.1| hypothetical protein	(1667)	173	48.9	0.0093
gi|147826806|emb|CAN63950.1| hypothetical protein	(1545)	172	48.6	0.01
gi|147777812|emb|CAN64608.1| hypothetical protein	(603)	167	47.2	0.011
gi|255673701|dbj|BAF06218.2| Os01g0758700 [Oryza s	(135)	159	45.0	0.011
gi|147783452|emb|CAN75210.1| hypothetical protein	(1137)	169	47.8	0.013
gi|28209489|gb|AAO37507.1| retrotransposon protein	(1155)	169	47.8	0.013
gi|155306694|gb|ABT62298.1| Sequence 149768 from p	(158)	158	44.8	0.015
gi|18921316|gb|AAL82521.1|AC084766_7 putative poly	(408)	162	45.9	0.017
gi|147775005|emb|CAN70471.1| hypothetical protein	(1122)	167	47.4	0.018
gi|62733754|gb|AAX95863.1| retrotransposon protein	(1126)	167	47.4	0.018
gi|223640087|emb|CAX44333.1| transposable element	(664)	163	46.3	0.022
gi|77556354|gb|ABA99150.1| retrotransposon protein	(1082)	165	46.9	0.024
gi|116310275|emb|CAH67280.1| OSIGBa0111L12.7 [Oryz	(940)	164	46.6	0.025
gi|4235644|gb|AAD13304.1| polyprotein [Solanum lyc	(1542)	166	47.2	0.027
gi|108862641|gb|ABG22014.1| retrotransposon protei	(1422)	164	46.7	0.035
gi|147772919|emb|CAN64786.1| hypothetical protein	(1217)	163	46.5	0.036
gi|147819718|emb|CAN73574.1| hypothetical protein	(1027)	161	45.9	0.044
gi|147840564|emb|CAN68329.1| hypothetical protein	(1330)	162	46.2	0.046
gi|270001171|gb|EEZ97618.1| hypothetical protein T	(622)	158	45.1	0.047
gi|210072545|gb|EEA26632.1| retrovirus polyprotein	(375)	155	44.3	0.05
gi|147810501|emb|CAN60890.1| hypothetical protein	(1378)	161	46.0	0.056
gi|147845507|emb|CAN80599.1| hypothetical protein	(1404)	161	46.0	0.056

```
gi|147810220|emb|CAN78060.1| hypothetical protein  (1037)  159 45.5   0.061
gi|147784791|emb|CAN75226.1| hypothetical protein  ( 422)  154 44.1   0.065
gi|158578546|gb|ABW74571.1| pol polyprotein [Boech ( 450)  154 44.1   0.069
gi|108708778|gb|ABF96573.1| retrotransposon protei (1481)  160 45.8   0.069
gi|147860532|emb|CAN81876.1| hypothetical protein  (1241)  159 45.5   0.071
gi|34452676|gb|AAQ72731.1| putative reverse transc ( 435)  153 43.9   0.079
gi|136637170|gb|EBP43218.1| hypothetical protein G ( 248)  150 43.0    0.08
gi|155286956|gb|ABT42560.1| Sequence 130030 from p ( 142)  147 42.2   0.083
gi|147804663|emb|CAN62382.1| hypothetical protein  ( 688)  154 44.2   0.097
gi|4539021|emb|CAB39733.1| protease, reverse trans (1147)  155 44.6    0.13
gi|270010054|gb|EFA06502.1| hypothetical protein T (2056)  156 44.9    0.17
gi|147835239|emb|CAN67792.1| hypothetical protein  (1414)  154 44.4    0.18
gi|147775377|emb|CAN69087.1| hypothetical protein  ( 611)  149 43.0     0.2
gi|147816473|emb|CAN64046.1| hypothetical protein  (1259)  152 43.9    0.22
gi|54291836|gb|AAV32204.1| putative polyprotein [O (1142)  151 43.6    0.24
gi|147775233|emb|CAN77085.1| hypothetical protein  (1520)  151 43.7     0.3
gi|198142740|gb|EDY71459.1| GA22199 [Drosophila ps (1555)  150 43.5    0.36
gi|34015213|gb|AAQ56407.1| putative gag-pol polypr (1619)  150 43.5    0.38
gi|270010053|gb|EFA06501.1| hypothetical protein T (1475)  148 43.0    0.48
gi|193907775|gb|EDW06642.1| GI21862 [Drosophila mo ( 565)  143 41.6    0.49
gi|241915630|gb|EER88774.1| hypothetical protein S (1609)  148 43.0    0.52
gi|15042811|gb|AAK82434.1|AC091247_1 putative poly ( 799)  144 41.9    0.56
gi|108711868|gb|ABF99663.1| retrotransposon protei (1087)  144 42.0    0.72
gi|18855070|gb|AAL79762.1|AC096687_26 putative pol (1087)  144 42.0    0.72
gi|108864565|gb|ABA94541.2| retrotransposon protei (1347)  145 42.3    0.73
gi|89355887|gb|ABD72267.1| pol polyprotein [Drosop (1143)  144 42.0    0.75
gi|193899715|gb|EDV98581.1| GH22324 [Drosophila gr ( 201)  135 39.5    0.77
gi|270016625|gb|EFA13071.1| hypothetical protein T (1488)  145 42.3    0.79
gi|147818980|emb|CAN67121.1| hypothetical protein  ( 938)  142 41.5    0.88
gi|147834477|emb|CAN63112.1| hypothetical protein  (1049)  142 41.5    0.96


>>gi|124359710|gb|ABN06064.1| RNA-directed DNA polymeras  (1297 aa)
 initn: 435 init1: 292 opt: 437  Z-score: 547.1  bits: 110.5 E(): 2e-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: 527.5  bits: 107.0 E(): 2.5e-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: 507.0  bits: 103.8 E(): 3.5e-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: 401.2  bits: 82.5 E(): 2.7e-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: 396.5  bits: 83.0 E(): 5e-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: 392.7  bits: 82.0 E(): 8.2e-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|208609057|dbj|BAG72151.1| hypothetical protein [Lot  (1558 aa)
 initn: 262 init1: 140 opt: 315  Z-score: 391.9  bits: 82.0 E(): 9.1e-13
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|208609055|dbj|BAG72150.1| hypothetical protein [Lot  (1558 aa)
 initn: 262 init1: 140 opt: 315  Z-score: 391.9  bits: 82.0 E(): 9.1e-13
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|208609065|dbj|BAG72154.1| hypothetical protein [Lot  (1558 aa)
 initn: 262 init1: 140 opt: 315  Z-score: 391.9  bits: 82.0 E(): 9.1e-13
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|208609051|dbj|BAG72148.1| hypothetical protein [Lot  (1558 aa)
 initn: 261 init1: 139 opt: 314  Z-score: 390.6  bits: 81.8 E(): 1.1e-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|208609062|dbj|BAG72153.1| hypothetical protein [Lot  (1558 aa)
 initn: 261 init1: 139 opt: 314  Z-score: 390.6  bits: 81.8 E(): 1.1e-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: 388.2  bits: 81.3 E(): 1.4e-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: 388.2 bits: 81.3 E(): 1.4e-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: 384.4 bits: 80.3 E(): 2.4e-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: 376.3 bits: 79.0 E(): 6.7e-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: 374.6 bits: 78.7 E(): 8.3e-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: 372.8  bits: 78.2 E(): 1e-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: 371.6  bits: 79.1 E(): 1.2e-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: 369.8  bits: 77.8 E(): 1.5e-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: 366.0  bits: 76.5 E(): 2.5e-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|222635621|gb|EEE65753.1| hypothetical protein OsJ_2  (645 aa)
 initn: 285 init1: 204 opt: 289  Z-score: 364.7  bits: 75.7 E(): 3e-11
Smith-Waterman score: 289;  36.806% identity (62.500% similar) in 144 aa overlap
(8-150:126-267)

                                     10        20        30
5_3                               ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :.  . :::.:.. : :: :  ::..::::
```

```
gi|222 KHYFLGTSLIIRTDQASLKYINEQRITEGVQHKLLIKLLSYDYKIEYKKGQENKAADALS
         100       110       120       130       140       150

        40         50        60        70        80        90
5_3    RRDEDKELQG-ISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFR
       : .. . :   .  : :  ::..    . ::   ... .:.   .::: :::.  :: ..
gi|222 RMQQLNALTTTVIVPQW--ITEVAASYSTDPKCHELESHLHIAPQSHPPYTLKGGILRYK
         160       170         180       190       200       210

        100       110       120       130       140       150
5_3    GRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREI
        ..:. :..    .::  :. : .:::::   ::.:. : .::  .:  .:.::
gi|222 DHIVVGAGNTLREQLLVSFHDSALGGHSGERATYQRMKQLFYWPGMKLAVTQFVKACPVC
           220       230       240       250       260       270

       160
5_3    YMDQQ

gi|222 QKNKTEHNMPAGLLQPLPLPEMAWSHITMDFVEGLPKSEGKDVIWVIVDRFTKYAHFIPL
           280       290       300       310       320       330

>>gi|255677065|dbj|BAF19594.2| Os06g0493000 [Oryza sativ  (866 aa)
 initn: 285 init1: 204 opt: 289  Z-score: 362.8  bits: 75.8 E(): 3.8e-11
Smith-Waterman score: 289;  36.806% identity (62.500% similar) in 144 aa overlap
(8-150:347-488)

                                    10        20        30
5_3                              ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                 :.  . :::.:.. : :: :  ::..:::: 
gi|255 KHYFLGTSLIIRTDQASLKYINEQRITEGVQHKLLIKLLSYDYKIEYKKGQENKAADALS
         320       330       340       350       360       370

        40         50        60        70        80        90
5_3    RRDEDKELQG-ISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDILYFR
       : .. . :   .  : :  ::..    . ::   ... .:.   .::: :::.  :: ..
gi|255 RMQQLNALTTTVIVPQW--ITEVAASYSTDPKCHELESHLHIAPQSHPPYTLKGGILRYK
         380       390         400       410       420       430

        100       110       120       130       140       150
5_3    GRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHVREI
        ..:. :..    .::  :. : .:::::   ::.:. : .::  .:  .:.::
gi|255 DHIVVGAGNTLREQLLVSFHDSALGGHSGERATYQRMKQLFYWPGMKLAVTQFVKACPVC
           440       450       460       470       480       490

       160
5_3    YMDQQ

gi|255 QKNKTEHNMPAGLLQPLPLPEMAWSHITMDFVEGLPKSEGKDVIWVIVDRFTKYAHFIPL
           500       510       520       530       540       550

>>gi|147823217|emb|CAN75278.1| hypothetical protein [Vit  (716 aa)
 initn: 280 init1: 280 opt: 285  Z-score: 359.0  bits: 74.8 E(): 6.2e-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|116309032|emb|CAH66146.1| OSIGBa0114M03.4 [Oryza sa  (1448 aa)
 initn: 247 init1: 157 opt: 285  Z-score: 354.5  bits: 75.0 E(): 1.1e-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: 349.6  bits: 70.9 E(): 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: 348.7  bits: 73.2 E(): 2.3e-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: 346.2  bits: 73.6 E(): 3.2e-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: 345.7  bits: 72.2 E(): 3.4e-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: 345.0  bits: 73.1 E(): 3.7e-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: 342.7  bits: 72.9 E(): 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: 342.2  bits: 71.5 E(): 5.3e-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: 338.8  bits: 71.9 E(): 8.2e-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: 337.4  bits: 72.0 E(): 9.8e-10
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: 334.9  bits: 69.8 E(): 1.4e-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: 334.1  bits: 70.1 E(): 1.5e-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: 333.2  bits: 71.0 E(): 1.7e-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|255678117|dbj|BAH94093.1| Os08g0123950 [Oryza sativ  (783 aa)
 initn: 172 init1:  97 opt: 264  Z-score: 331.9  bits: 69.9 E(): 2e-09
Smith-Waterman score: 264;  32.168% identity (66.434% similar) in 143 aa overlap
(8-146:187-327)

                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :.   .:::: .. : :. :.: ..:::: 
gi|255 RSYLQHVEFIILTDHHSLMHLSDQRLHTPWQHKAFTKLLGLQYKICYRKGSTNAAADALS
        160       170       180       190       200       210

        40          50          60        70        80        90
5_3    RRDE--DKELQGISR--PFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       :. .  ..:...::.  : :  . .. .  : ::  ...   :  : .:.:...:.  .:
gi|255 RKPQQSEEEFNAISQCVPQW--LMEVLQSYDTDPHATQLVAALTLNPNSKPHFSLQHGVL
        220       230         240       250       260       270

          100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
        ..: . .  :     :...:...: .:::::. .::::: : . :  .:..:
gi|255 RYKGNIWIGNSPDLQLKIINEMHASPVGGHSGFPVTYRRIKQLFAWNGMKSQIKETLANC
          280       290       300       310       320       330

           160
5_3    REIYMDQQ

gi|255 QICAQAKPDRSRYPRLLQPLPVPKGAWQTISLDFIEGLPRSSHYNCILVVVDKFSKYSHF
          340       350       360       370       380       390

>>gi|241943804|gb|EES16949.1| hypothetical protein SORBI  (1456 aa)
 initn: 221 init1: 160 opt: 267  Z-score: 331.7  bits: 70.8 E(): 2e-09
Smith-Waterman score: 267;  32.432% identity (66.216% similar) in 148 aa overlap
(8-150:887-1029)

                                    10        20        30
5_3                           ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :..:.::::..:.. :: :..: :.:::: 
gi|241 RPYLWGRRFVVKTDHYNLKYLLDQRLSTIPQHHWVGKLLGFDFSVEYKPGVTNTVADALS
        860       870       880       890       900       910

        40            50        60        70        80        90
5_3    RRDEDKELQG----ISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       ::: ..:  :    .: : .  ::.. .    :::.. .::.....  . :  ... :..
gi|241 RRDTEEETAGAVLALSAPRFDFITRLRQANAGDPAMVALREDISSGARGMPWSVVD-DMV
        920       930       940       950       960       970

          100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLM-GGHSGIYITYRRITQSLYWIPIKGEITKFVVH
        . ::: .  .:   : ::::.  ..   :: :.  : .:. .....  .:  . ...
gi|241 QYSGRLYIPPGS---P-LLQELVLAVHEDGHEGVQRTLHRLRRDFHFPNMKQVVQELIRV
          980           990      1000      1010      1020      1030
```

```
           160
5_3    VREIYMDQQ

gi|241 CVVCQRYKSEHLQPTGLLLPLSVPQGIWTDIALDFVEALPRVRGKTVILTVVDRFSKYCH
          1040      1050      1060      1070      1080      1090

>>gi|108864659|gb|ABA95357.2| retrotransposon protein, p  (2811 aa)
 initn: 252 init1: 161 opt: 270  Z-score: 331.3  bits: 71.7 E(): 2.2e-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|241933637|gb|EES06782.1| hypothetical protein SORBI  (1414 aa)
 initn: 160 init1: 160 opt: 265  Z-score: 329.4  bits: 70.3 E(): 2.8e-09
Smith-Waterman score: 265;  34.459% identity (66.216% similar) in 148 aa overlap
(8-150:849-991)

                                     10        20        30
5_3                             ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :..::.::::..: . :: :::: :.::::
gi|241 RPYLWGRRFIVKTDHYSLKYLLDQRLATIPQHHWVGKLLGFDFTVEYKSGASNTVADALS
      820       830       840       850       860       870

        40           50        60        70        80        90
5_3    RRDEDKE----LQGISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       ::: ...    ::.:: : .  :... .    ::::. :..... .  . : .:.   ..
gi|241 RRDTEEDSVGGLQAISAPRFDFIARLRHAQATDPALVTIHDDIRAGSRAAP-WTVADGMV
      880       890       900       910       920       930

          100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSL-MGGHSGIYITYRRITQSLYWIPIKGEITKFVVH
        . ::: .  ::   : :.::. ... . :: :.  : .:. .....  ..  .  ::
gi|241 LYDGRLYIPPSS---P-LVQEIVAAVHYDGHEGVQRTLHRLRRDFHFPQMRRIVQDFVRA
       940          950       960       970       980       990

           160
5_3    VREIYMDQQ
```

```
gi|241 CVTCQRYKSEHLLPAGLLLPLPVPTVVWADIGLDFVEALPRVQGKSVILTVVDRFSKYCH
          1000      1010      1020      1030      1040      1050

>>gi|77554607|gb|ABA97403.1| retrotransposon protein, pu  (562 aa)
 initn: 230 init1: 230 opt: 259  Z-score: 327.7  bits: 68.7 E(): 3.4e-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|62734016|gb|AAX96125.1| retrotransposon protein, pu  (1448 aa)
 initn: 196 init1: 105 opt: 263  Z-score: 326.7  bits: 69.9 E(): 3.9e-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|77550700|gb|ABA93497.1| retrotransposon protein, pu  (1448 aa)
```

```
 initn: 196 init1: 105 opt: 263  Z-score: 326.7  bits: 69.9 E(): 3.9e-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|241917803|gb|EER90947.1| hypothetical protein SORBI  (1450 aa)
 initn: 216 init1: 165 opt: 263  Z-score: 326.7  bits: 69.9 E(): 3.9e-09
Smith-Waterman score: 263;  30.612% identity (65.306% similar) in 147 aa overlap
(8-150:922-1064)

                                     10        20        30
5_3                               ISELQLNQQYWVAKLLGYEFDIVYKVGASNKVVDALS
                                    :..::.::::..:.. :: ::.: :.::::
gi|241 RPYLWGRRFVVKTDHYSLKYLLDQRLSTIPQHHWVGKLLGFDFSVEYKPGATNTVADALS
             900       910       920       930       940       950

        40           50        60        70         80        90
5_3    RRDEDKELQG----ISRPFWKDITKINEEVQKDPALAKIREELKDNLDSHPQYTLECDIL
       ::: ..:  :    .: : .  ::.. .    :::.. .::..... . :  ... :..
gi|241 RRDTEEETAGTVLALSAPRFDFITRLRQANAGDPAVVALREDISSGARGMPWSVVD-DMV
             960       970       980       990      1000       1010

          100       110       120       130       140       150
5_3    YFRGRLVLLASSLWIPKLLQEFQTSLMGGHSGIYITYRRITQSLYWIPIKGEITKFVVHV
        . ::: .  .:  . .:. : . .   :: :.    .:. .....  .:  . ...
gi|241 LYSGRLYIPPGSPLLQELVLEVHED---GHEGVQRMLHRLRRDFHFPNMKQVVQELIRVC
             1020      1030         1040      1050      1060

          160
5_3    REIYMDQQ

gi|241 VVCQRYKSEHLQPAGLLLPLPVPQGIWTDIALDFVEALPSVRGKTVILTVVDRFSKYCYF
      1070      1080      1090      1100      1110      1120

>>gi|108706172|gb|ABF93967.1| retrotransposon protein, p  (1461 aa)
 initn: 189 init1: 140 opt: 263  Z-score: 326.6  bits: 69.9 E(): 3.9e-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|15451607|gb|AAK98731.1|AC090485_10 Putative retroel  (1461 aa)
 initn: 189 init1: 140 opt: 263  Z-score: 326.6  bits: 69.9 E(): 3.9e-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|90399077|emb|CAJ86299.1| H0124B04.16 [Oryza sativa   (1265 aa)
 initn: 253 init1: 162 opt: 261  Z-score: 325.0  bits: 69.4 E(): 4.8e-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: 324.5  bits: 68.0 E(): 5.2e-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|110289541|gb|AAP54937.2| retrotransposon protein, p  (1477 aa)
 initn: 273 init1: 151 opt: 260  Z-score: 322.8  bits: 69.2 E(): 6.4e-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: 321.8  bits: 67.8 E(): 7.3e-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



161 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:27:32 2010 done: Tue Jan 26 20:35:23 2010
 Total Scan time: 412.780 Total Display time:  0.310

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_2010
Query = 5_4

Start time: Tue Jan 26 20:35:24 GMT 2010 Finish time: Tue Jan 26 20:35:24 GMT 2010

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

```
# fasta34 5_4.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_4.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7, 2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 5_4, 13 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     1     0:=          one = represents 3 library sequences
  24     3     0:=
  26    11     0:====
  28     4     0:==
  30    23     2:*=======
  32    52     8:==*===============
  34    28    21:======*===
  36    31    44:===========   *
  38    48    72:================       *
  40    95   101:================================ *
  42    78   123:==========================               *
  44    85   136:=============================                 *
  46   131   138:============================================= *
  48   134   132:============================================*=
  50   122   121:===========================================*
  52    97   106:================================  *
  54   112    91:==============================*=======
  56    68    76:======================  *
  58    93    62:====================*==========
  60    66    50:================*=====
  62    37    40:=============*
  64    43    32:==========*====
  66    38    25:========*====
  68    16    20:=======*
  70    10    16:==== *
  72    14    12:===*=
  74     6    10:== *
  76     3     7:= *
  78     5     6:=*
  80     5     4:=*
  82     1     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     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:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.37740.00235; mu= 2.5045 0.123
 mean_var=13.8228 3.544, 0's: 2 Z-trim: 2  B-trim: 125 in 1/42
 Lambda= 0.344965
 Kolmogorov-Smirnov  statistic: 0.0616 (N=28) 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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:35:23 2010 done: Tue Jan 26 20:35:24 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

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

 5_4, 13 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    60     0:====
  22     2     0:=          one = represents 18 library sequences
  24     2     0:=
  26     6     0:=
  28    13     2:*
  30    76    12:*====
  32    64    45:==*=
  34   319   122:======*===========
  36   197   250:===========  *
  38   310   414:==================    *
  40   383   577:======================          *
  42   563   706:================================       *
  44  1037   779:=============================================*==============
  46   710   793:=========================================    *
  48   789   759:============================================*=
  50   489   693:===========================         *
  52   409   609:=======================          *
  54   534   520:============================*=
  56   543   435:========================*======
  58   503   357:==================*========
  60   234   289:=============   *
  62   268   232:============*==
  64   220   184:==========*==
  66   236   146:========*=====
  68   118   115:======*
```

```
  70   113    90:====*==
  72    79    70:===*=
  74    36    55:== *
  76    23    43:==*
  78    22    33:=*
  80     6    26:=*
  82    25    20:=*
  84    12    16:*
  86     2    12:*
  88    30     9:*=         inset = represents 1 library sequences
  90     4     7:*
  92     3     6:*         :===  *
  94     0     4:*         :   *
  96     1     3:*         := *
  98     1     3:*         := *
 100     0     2:*         : *
 102     1     2:*         :=*
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.14820.000456; mu= -0.2233 0.023
 mean_var=17.8910 3.940, 0's: 60 Z-trim: 60  B-trim: 200 in 1/61
 Lambda= 0.303219
 Kolmogorov-Smirnov  statistic: 0.0468 (N=29) at  52

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


13 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:35:24 2010 done: Tue Jan 26 20:35:24 2010
 Total Scan time:  0.120 Total Display time:  0.000

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

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

 5_4, 13 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 282225     0:===========
  22   1249     0:=          one = represents 26019 library sequences
  24   2958    17:*
  26   8713   374:*
  28 24415  4039:*
  30 69312 24538:*==
  32 165700 94881:===*===
  34 329470 257305:=========*===
  36 558342 528444:====================*=
  38 852718 873321:==================================*
  40 1122828 1218206:================================================  *
  42 1343261 1489108:=========================================================     *
  44 1485895 1642625:============================================================== *
  46 1561122 1673054:================================================================*
  48 1549501 1601757:================================================================*
  50 1437629 1461609:============================================================*
  52 1300538 1284998:==========================================*
  54 1082957 1097614:=========================================*
  56 933045 916844:==============================*
  58 784071 752712:============================*==
  60 632472 609741:=======================*=
  62 528607 488831:==================*==
  64 418391 388765:==============*==
  66 310748 307268:===========*
  68 251100 241690:=========*
  70 185287 189403:=======*
  72 142627 148000:=====*
  74 104482 115391:====*
  76 85667 89811:===*
  78 66583 69809:==*
  80 49022 54206:==*
  82 37793 41466:=*
  84 26096 32846:=*
  86 18463 25415:*
  88 14532 19665:*          inset = represents 185 library sequences
  90 11497 15215:*
  92  9247 11773:*         :=======================================*
  94  6961  9109:*         :======================================== *
  96  5466  7048:*         :==============================        *
  98  3225  5454:*         :==================           *
 100  1946  4220:*         :===========          *
 102  1625  3265:*         :=========        *
 104  1271  2526:*         :=======      *
 106   723  1955:*         :====      *
 108   506  1512:*         :===     *
 110   306  1170:*         :==    *
 112   199   905:*         :==  *
 114   125   701:*         :=  *
 116    76   542:*         := *
 118    58   419:*         := *
>120   175   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17811001 sequences
  Expectation_n fit: rho(ln(x))= 3.13750.000165; mu= 4.8472 0.009
 mean_var=16.8894 3.340, 0's: 952 Z-trim: 958  B-trim: 0 in 0/65
 Lambda= 0.312080
 Kolmogorov-Smirnov  statistic: 0.0203 (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
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:35:25 2010 done: Tue Jan 26 20:40:00 2010
 Total Scan time: 233.500 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_2010
Query = 5_5

Start time: Tue Jan 26 20:40:01 GMT 2010 Finish time: Tue Jan 26 20:40:01 GMT 2010

No 8 amino acid matches exist between 5_5 and the AD_2010 database
```

```
# fasta34 5_5.pep /genedata/1/db/AD_2010 -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 /genedata/1/db/AD_2010 library

       opt      E()
< 20    13     0:=====
  22     0     0:            one = represents 3 library sequences
  24     1     0:=
  26     0     0:
  28     0     0:
  30     1     2:*
  32    23     8:==*=====
  34    13    21:===== *
  36    36    44:============  *
  38    61    72:======================  *
  40    96   101:================================== *
  42   128   123:==========================================*==
  44   179   136:================================================*==============
  46   125   138:============================================    *
  48   147   132:============================================*=====
  50    99   121:==================================       *
  52    76   106:==========================          *
  54    88    91:==============================*
  56    65    76:======================    *
  58    54    62:==================  *
  60    55    50:================*==
  62    59    40:=============*======
  64    29    32:==========*
  66    31    25:========*==
  68    15    20:===== *
  70    19    16:=====*=
  72    19    12:===*===
  74     8    10:===*
  76     4     7:==*
  78     2     6:=*
  80     7     4:=*=
  82     4     3:*=
  84     3     3:*
  86     2     2:*
  88     2     2:*          inset = represents 1 library sequences
  90     1     1:*
  92     1     1:*         :*
  94     1     1:*         :*
  96     0     1:*         :*
  98     1     0:=         *=
 100     1     0:=         *=
 102     1     0:=         *=
 104     1     0:=         *=
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.27330.00339; mu= 3.0013 0.174
 mean_var=35.2571 9.866, 0's: 13 Z-trim: 13  B-trim: 52 in 1/40
 Lambda= 0.215999
 Kolmogorov-Smirnov  statistic: 0.0277 (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(1471)
gi|75062228|sp|Q5VFH6.1|ALL4_FELCA RecName: Full=A ( 186)   57 24.0    0.88


>>gi|75062228|sp|Q5VFH6.1|ALL4_FELCA RecName: Full=Aller  (186 aa)
 initn:  57 init1:  57 opt:  57  Z-score: 103.3  bits: 24.0 E(): 0.88
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




53 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:40:00 2010 done: Tue Jan 26 20:40:01 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 5_5.pep /genedata/1/db/TOX_2010 -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 /genedata/1/db/TOX_2010 library

       opt      E()
< 20    83     0:======
  22     1     0:=          one = represents 14 library sequences
  24     1     0:=
  26     1     0:=
  28     5     2:*
  30     6    12:*
  32    16    45:== *
  34    72   122:======  *
  36   126   250:=========        *
  38   642   414:==============================*================
  40   513   577:====================================     *
  42   576   706:============================================         *
  44   727   779:=====================================================    *
  46   750   793:======================================================   *
  48   775   759:=========================================================*=
  50   803   693:=================================================*========
  52   624   609:============================================*=
  54   392   520:============================         *
  56   348   435:=========================      *
  58   378   357:=========================*=
  60   219   289:================     *
  62   159   232:============    *
  64   176   184:=============*
  66   177   146:==========*==
  68   156   115:========*===
  70   120    90:======*==
  72   239    70:====*=============
  74    94    55:===*===
  76    46    43:===*
  78    70    33:==*==
  80    44    26:=*==
  82    19    20:=*
  84    13    16:=*
  86    19    12:*=
  88    10     9:*          inset = represents 1 library sequences
  90    14     7:*
  92     0     6:*         :     *
  94     2     4:*         :== *
  96    12     3:*         :==*=========
  98     9     3:*         :==*======
 100     1     2:*         :=*
 102     0     2:*         : *
 104     0     1:*         :*
 106     3     1:*         :*==
 108     0     1:*         :*
 110     1     1:*         :*
 112     0     0:          *
 114     1     0:=         *=
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 5.06840.000622; mu= -0.6615 0.030
 mean_var=38.0619 9.329, 0's: 82 Z-trim: 84  B-trim: 771 in 2/59
 Lambda= 0.207888
 Kolmogorov-Smirnov  statistic: 0.0560 (N=29) at  64

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
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:40:01 2010 done: Tue Jan 26 20:40:01 2010
 Total Scan time:  0.100 Total Display time:  0.000

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

```
# fasta34 5_5.pep /genedata/1/db/PRT_2010 -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 /genedata/1/db/PRT_2010 library

       opt      E()
< 20 369961     0:=============
  22   155     0:=          one = represents 28600 library sequences
  24   247    17:*
  26   763   374:*
  28  2393  4039:*
  30 12362 24538:*
  32 57898 94880:===*
  34 189977 257304:======= *
  36 452507 528442:================  *
  38 871606 873318:==============================*
  40 1315207 1218202:============================================*===
  42 1617340 1489104:=================================================*====
  44 1715973 1642620:=================================================================*==
  46 1635216 1673050:============================================================*
  48 1496910 1601752:=========================================================    *
  50 1293635 1461605:=============================================     *
  52 1138208 1284995:=========================================    *
  54 967825 1097611:==================================    *
  56 845620 916842:============================  *
  58 726596 752710:==========================*
  60 606586 609739:====================*
  62 547364 488830:=================*==
  64 411451 388763:=============*=
  66 328530 307267:==========*=
  68 257090 241690:========*
  70 200674 189402:======*=
  72 154663 148000:=====*
  74 130134 115390:====*
  76 118954 89811:===*=
  78 91923 69809:==*=
  80 77605 54205:=*=
  82 47649 41466:=*
  84 34383 32846:=*
  86 26026 25415:*
  88 18427 19665:*          inset = represents 203 library sequences
  90 14075 15215:*
  92 10135 11773:*         :=======================================*
  94  6623  9109:*         :==================================      *
  96  4997  7048:*         :========================         *
  98  3774  5454:*         :====================       *
```

```
 100  2566  4220:*         :=============       *
 102  1934  3265:*         :==========      *
 104  1503  2526:*         :========    *
 106  1096  1955:*         :======   *
 108   755  1512:*         :====   *
 110   554  1170:*         :===  *
 112   415   905:*         :=== *
 114   281   701:*         :== *
 116   179   542:*         := *
 118   161   419:*         := *
>120   319   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810952 sequences
  Expectation_n fit: rho(ln(x))= 4.35370.000184; mu= 4.1723 0.010
 mean_var=46.7049 9.593, 0's: 1262 Z-trim: 1262  B-trim: 3062 in 1/63
 Lambda= 0.187669
 Kolmogorov-Smirnov  statistic: 0.0335 (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


53 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:40:01 2010 done: Tue Jan 26 20:44:23 2010
 Total Scan time: 224.890 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_2010
Query = 5_6

Start time: Tue Jan 26 20:44:23 GMT 2010 Finish time: Tue Jan 26 20:44:23 GMT 2010

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


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

 5_6, 31 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     1     0:=
  28     2     0:=
  30    11     2:*===
  32    19     8:==*====
  34    56    21:======*===========
  36    76    44:==============*===========
  38    44    72:===============         *
  40    91   101:================================  *
  42   100   123:===================================       *
  44   107   136:=====================================          *
  46   112   138:=======================================        *
  48   116   132:=========================================    *
  50   111   121:======================================   *
  52   147   106:====================================*=============
  54   104    91:================================*====
  56    70    76:======================== *
  58    82    62:====================*=======
  60    45    50:=============== *
  62    32    40:===========  *
  64    55    32:==========*========
  66    19    25:======= *
  68    11    20:====  *
  70    13    16:=====*
  72    11    12:===*
  74     7    10:===*
  76     7     7:==*
  78     6     6:=*
  80     4     4:=*
  82     3     3:*
  84     0     3:*
  86     1     2:*
  88     2     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     0     1:*         :*
  94     2     1:*         :*=
  96     1     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:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.30430.00298; mu= 9.9265 0.155
 mean_var=22.4568 6.132, 0's: 2 Z-trim: 2  B-trim: 15 in 1/42
 Lambda= 0.270645
 Kolmogorov-Smirnov  statistic: 0.0612 (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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:44:23 2010 done: Tue Jan 26 20:44:23 2010
 Total Scan time:  0.020 Total Display time:  0.000
```

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

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

 5_6, 31 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    60     0:====
  22     1     0:=          one = represents 16 library sequences
  24     8     0:=
  26    21     0:==
  28    13     2:*
  30    27    12:*=
  32   106    45:==*====
  34   184   122:=======*====
  36   395   250:===============*=========
  38   286   414:==================       *
  40   481   577:================================     *
  42   562   706:====================================         *
  44   695   779:================================================    *
  46   905   793:================================================*=======
  48   488   759:================================                  *
  50   573   693:====================================       *
  52   701   609:============================================*=====
  54   501   520:================================*
  56   449   435:============================*=
  58   522   357:======================*==========
  60   207   289:==============     *
  62   364   232:==============*========
  64   146   184:========== *
  66    91   146:======   *
  68   145   115:=======*==
  70    71    90:=====*
  72    34    70:=== *
  74   213    55:===*==========
  76    36    43:==*
  78    24    33:==*
  80    58    26:=*==
  82    21    20:=*
  84    13    16:*
  86     7    12:*
  88    21     9:*=         inset = represents 1 library sequences
  90     6     7:*
  92     2     6:*          :==    *
  94     2     4:*          :== *
  96     0     3:*          :  *
  98     2     3:*          :==*
 100     1     2:*          :=*
 102     1     2:*          :=*
 104     0     1:*          :*
 106     0     1:*          :*
 108     0     1:*          :*
 110     0     1:*          :*
 112     0     0:           *
 114     0     0:           *
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 2.52130.000512; mu= 14.1090 0.026
 mean_var=22.4811 4.701, 0's: 60 Z-trim: 60  B-trim: 657 in 2/60
 Lambda= 0.270499
 Kolmogorov-Smirnov  statistic: 0.0486 (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


31 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:44:23 2010 done: Tue Jan 26 20:44:24 2010
 Total Scan time:  0.180 Total Display time:  0.000

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

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

 5_6, 31 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 276784     0:==========
  22    434     0:=          one = represents 27868 library sequences
  24    903    17:*
  26   2409   374:*
  28   8417  4039:*
  30 32742 24538:*=
  32 103865 94881:===*
  34 252667 257305:=========*
  36 502087 528443:==================*
  38 819729 873320:============================== *
  40 1165258 1218205:======================================== *
  42 1437422 1489107:================================================ *
  44 1598263 1642624:============================================================*
  46 1672054 1673053:============================================================*
  48 1634666 1601755:============================================================*=
  50 1479177 1461608:============================================================*=
  52 1272001 1284997:============================================*
  54 1069709 1097613:=======================================*
  56 893575 916843:============================*
  58 729994 752711:==========================*
  60 599045 609740:====================*
  62 470332 488831:=================*
  64 376077 388764:=============*
  66 302448 307268:===========*
  68 255232 241690:========*=
  70 197245 189403:======*=
  72 148887 148000:=====*
  74 118025 115391:====*
  76 96385 89811:===*
  78 67252 69809:==*
  80 54817 54206:=*
  82 43123 41466:=*
```

```
  84 33009 32846:=*
  86 26354 25415:*
  88 18926 19665:*          inset = represents 201 library sequences
  90 14733 15215:*
  92 10030 11773:*         :=======================================*
  94  7687  9109:*         :=======================================*
  96  5169  7048:*         :=========================         *
  98  3689  5454:*         :===================         *
 100  2631  4220:*         :===============      *
 102  2135  3265:*         :===========     *
 104  1569  2526:*         :========    *
 106  1077  1955:*         :======   *
 108   854  1512:*         :=====  *
 110   580  1170:*         :===  *
 112   497   905:*         :=== *
 114   277   701:*         :== *
 116   442   542:*         :==*
 118   269   419:*         :==*
>120   273   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810988 sequences
  Expectation_n fit: rho(ln(x))= 3.68250.000172; mu= 7.7373 0.009
 mean_var=28.6258 5.593, 0's: 937 Z-trim: 941  B-trim: 0 in 0/63
 Lambda= 0.239715
 Kolmogorov-Smirnov  statistic: 0.0167 (N=29) at  66

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
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:44:24 2010 done: Tue Jan 26 20:51:02 2010
 Total Scan time: 347.180 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_2010
Query = 3_1

Start time: Tue Jan 26 20:51:03 GMT 2010 Finish time: Tue Jan 26 20:51:03 GMT 2010

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

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

 3_1, 43 aa
 vs /genedata/1/db/AD_2010 library
```

```
       opt      E()
< 20     5     0:==
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     1     0:=
  28     1     0:=
  30     0     2:*
  32     5     8:==*
  34    36    21:======*=====
  36    40    44:==============*
  38    72    72:========================*
  40   116   101:================================*=====
  42   130   123:=========================================*===
  44   132   136:============================================= *
  46   132   138:============================================= *
  48   131   132:=============================================*
  50   123   121:=========================================*
  52    59   106:====================               *
  54    55    91:===================          *
  56    42    76:==============           *
  58    59    62:====================*
  60    41    50:==============  *
  62    51    40:=============*===
  64    56    32:==========*========
  66    39    25:========*====
  68    46    20:=======*=========
  70    18    16:=====*
  72    11    12:===*
  74    21    10:===*===
  76    13     7:==*==
  78     9     6:=*=
  80    11     4:=*==
  82     4     3:*=
  84     4     3:*=
  86     5     2:*=
  88     3     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:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.46830.00345; mu= 2.8920 0.182
 mean_var=44.624712.769, 0's: 5 Z-trim: 5  B-trim: 219 in 1/41
 Lambda= 0.191994
 Kolmogorov-Smirnov  statistic: 0.0735 (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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:51:02 2010 done: Tue Jan 26 20:51:02 2010
 Total Scan time:  0.000 Total Display time:  0.000

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

```
# fasta34 3_1.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 3_1.pep_tx.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7, 2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_1, 43 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    78     0:=====
  22     1     0:=          one = represents 18 library sequences
  24     1     0:=
  26     1     0:=
  28     4     2:*
  30    25    12:*=
  32    26    45:==*
  34   111   122:======*
  36   167   250:==========   *
  38   304   414:=================     *
  40   457   577:=========================      *
  42   792   706:========================================*====
  44  1032   779:============================================*==============
  46   690   793:=======================================     *
  48   990   759:============================================*============
  50   688   693:=======================================*
  52   554   609:==============================  *
  54   506   520:============================*
  56   325   435:==================     *
  58   299   357:=================  *
  60   182   289:==========     *
  62   211   232:============*
  64   124   184:=======   *
  66   127   146:========*
  68    90   115:===== *
  70    69    90:====*
  72    69    70:===*
  74   276    55:===*============
  76    39    43:==*
  78    54    33:=*=
  80    40    26:=*=
  82    23    20:=*
  84    16    16:*
  86     9    12:*
  88     9     9:*          inset = represents 1 library sequences
  90     9     7:*
  92    22     6:*=        :=====*================
  94     4     4:*         :===*
  96     3     3:*         :==*
  98     1     3:*         := *
 100     0     2:*         : *
 102    15     2:*         :=*=============
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.32270.000585; mu= 2.4697 0.029
 mean_var=31.9088 6.968, 0's: 78 Z-trim: 78  B-trim: 372 in 1/60
 Lambda= 0.227049
 Kolmogorov-Smirnov  statistic: 0.0423 (N=29) at  72

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
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:51:03 2010 done: Tue Jan 26 20:51:04 2010
 Total Scan time:  0.100 Total Display time:  0.000

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

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

 3_1, 43 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 361258     0:=============
  22   175     0:=          one = represents 28436 library sequences
  24   467    17:*
  26  1205   374:*
  28  4518  4039:*
  30 17330 24538:*
  32 63712 94880:===*
  34 200098 257304:======== *
  36 447093 528443:================  *
  38 816706 873319:============================== *
  40 1197382 1218204:============================================*
  42 1524439 1489106:=============================================================*=
  44 1683628 1642622:=============================================================*==
  46 1706113 1673052:=============================================================*=
  48 1547199 1601754:============================================================ *
  50 1377201 1461607:================================================  *
  52 1221624 1284996:=========================================  *
  54 1047687 1097612:=================================== *
  56 877848 916843:============================== *
  58 753286 752710:=========================*
  60 619302 609740:====================*
  62 498941 488830:=================*
  64 411845 388764:=============*=
  66 328075 307267:==========*=
```

```
  68 257680 241690:========*=
  70 206482 189402:=======*=
  72 155310 148000:======*
  74 119815 115391:=====*
  76 92858 89811:===*
  78 70900 69809:==*
  80 51821 54206:=*
  82 39754 41466:=*
  84 29244 32846:=*
  86 22410 25415:*
  88 15943 19665:*          inset = represents 169 library sequences
  90 11224 15215:*
  92  8406 11773:*         :=======================================*
  94  5764  9109:*         :====================================    *
  96  4256  7048:*         :==========================              *
  98  3395  5454:*         :=====================           *
 100  2698  4220:*         :=================         *
 102  1835  3265:*         :============         *
 104  1161  2526:*         :=======       *
 106   900  1955:*         :======     *
 108   748  1512:*         :=====   *
 110   478  1170:*         :===   *
 112   264   905:*         :==   *
 114   179   701:*         :==  *
 116   150   542:*         :=  *
 118   121   419:*         := *
>120   297   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810972 sequences
  Expectation_n fit: rho(ln(x))= 4.35300.000184; mu= 3.5266 0.010
 mean_var=40.3752 8.132, 0's: 1226 Z-trim: 1226  B-trim: 2 in 1/63
 Lambda= 0.201844
 Kolmogorov-Smirnov  statistic: 0.0249 (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


43 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:51:04 2010 done: Tue Jan 26 20:55:36 2010
 Total Scan time: 207.000 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_2010
Query = 3_2

Start time: Tue Jan 26 20:55:37 GMT 2010 Finish time: Tue Jan 26 20:55:37 GMT 2010

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

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

 3_2, 25 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     1     0:=
  26     0     0:
  28     0     0:
  30     0     2:*
  32     7     8:==*
  34    30    21:======*===
  36    71    44:==============*=========
  38    51    72:=================       *
  40    62   101:====================             *
  42    80   123:==========================               *
  44   115   136:=========================================      *
  46   136   138:===============================================*
  48   110   132:=====================================       *
  50   113   121:=====================================  *
  52    96   106:================================   *
  54   143    91:================================*=================
  56   108    76:=========================*==========
  58    78    62:====================*=====
  60    63    50:================*====
  62    40    40:=============*
  64    47    32:==========*=====
  66    22    25:========*
  68    32    20:======*====
  70     9    16:===  *
  72     7    12:===*
  74     7    10:===*
  76     9     7:==*
  78    11     6:=*==
  80    10     4:=*==
  82     3     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     0     0:          *
 102     1     0:=         *=
 104     0     0:          *
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
```

```
  Expectation_n fit: rho(ln(x))= 4.31910.00255; mu= -0.1473 0.134
 mean_var=18.6171 4.890, 0's: 2 Z-trim: 2  B-trim: 6 in 1/42
 Lambda= 0.297248
 Kolmogorov-Smirnov  statistic: 0.0897 (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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:55:36 2010 done: Tue Jan 26 20:55:37 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

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

 3_2, 25 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    61     0:=====
  22     2     0:=          one = represents 14 library sequences
  24     0     0:
  26     3     0:=
  28     7     2:*
  30    37    12:*==
  32    93    45:===*===
  34   117   122:========*
  36   260   250:=================*=
  38   479   414:==============================*=====
  40   450   577:================================        *
  42   749   706:================================================*===
  44   792   779:============================================================*=
  46   603   793:============================================          *
  48   707   759:================================================    *
  50   543   693:=======================================          *
  52   413   609:================================             *
  54   603   520:=======================================*======
  56   508   435:================================*=====
  58   555   357:=========================*==============
  60   302   289:====================*=
  62   269   232:=================*===
  64   218   184:=============*==
  66   125   146:========== *
  68    79   115:======  *
  70    71    90:======*
  72    39    70:=== *
  74   189    55:===*==========
  76    57    43:===*=
  78    12    33:= *
  80    13    26:=*
  82    35    20:=*=
  84    10    16:=*
  86    19    12:*=
  88     6     9:*          inset = represents 1 library sequences
  90     4     7:*
  92     1     6:*         :=    *
  94     1     4:*         :=  *
  96     0     3:*         :  *
  98     0     3:*         :  *
 100     2     2:*         :=*
 102     3     2:*         :=*=
 104     1     1:*         :*
 106     0     1:*         :*
 108     1     1:*         :*
 110     2     1:*         :*=
 112     2     0:=         *==
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 3.06310.000464; mu= 7.9103 0.024
 mean_var=17.5749 3.707, 0's: 60 Z-trim: 60  B-trim: 294 in 1/61
 Lambda= 0.305934
 Kolmogorov-Smirnov  statistic: 0.0604 (N=29) at  52

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


25 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:55:37 2010 done: Tue Jan 26 20:55:37 2010
 Total Scan time:  0.140 Total Display time:  0.000

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

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

 3_2, 25 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 280392     0:===========
  22   1009     0:=          one = represents 26520 library sequences
  24   2094    17:*
  26   6375   374:*
  28 20209  4039:*
  30 56471 24538:*==
  32 146602 94880:===*==
  34 319557 257304:=========*===
  36 589572 528441:==================*===
  38 844876 873317:================================*
  40 1146621 1218200:============================================= *
  42 1403159 1489101:=============================================================   *
  44 1561906 1642618:================================================================*
  46 1591188 1673047:================================================================*
  48 1554804 1601749:================================================================*
  50 1429642 1461603:============================================================ *
```

```
  52 1291715 1284993:================================================*
  54 1139160 1097609:=========================================*=
  56 897089 916840:==================================*
  58 736091 752708:============================*
  60 584920 609738:======================*
  62 456940 488829:==================*
  64 370039 388763:==============*
  66 291135 307266:===========*
  68 233253 241689:=========*
  70 192512 189402:=======*
  72 150095 148000:=====*
  74 112877 115390:====*
  76 98597 89811:===*
  78 77991 69809:==*
  80 53310 54205:==*
  82 41376 41466:=*
  84 30163 32846:=*
  86 24534 25415:*
  88 17625 19664:*          inset = represents 207 library sequences
  90 12655 15215:*
  92 10329 11773:*         :=======================================*
  94  8404  9109:*         :=======================================*
  96  5966  7048:*         :============================     *
  98  4662  5454:*         :======================   *
 100  5796  4220:*         :====================*=======
 102  2398  3265:*         :============   *
 104  1728  2526:*         :=========   *
 106  1355  1955:*         :=======  *
 108  1317  1512:*         :=======*
 110   933  1170:*         :=====*
 112   573   905:*         :=== *
 114   403   701:*         :== *
 116   268   542:*         :==*
 118   181   419:*         := *
>120   358   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810921 sequences
  Expectation_n fit: rho(ln(x))= 3.49020.000171; mu= 6.6867 0.009
 mean_var=23.0219 4.598, 0's: 927 Z-trim: 932  B-trim: 3436 in 1/63
 Lambda= 0.267303
 Kolmogorov-Smirnov  statistic: 0.0162 (N=29) at  68

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
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 20:55:37 2010 done: Tue Jan 26 21:02:07 2010
 Total Scan time: 338.560 Total Display time:  0.000

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_2010
Query = 3_3

Start time: Tue Jan 26 21:02:08 GMT 2010 Finish time: Tue Jan 26 21:02:08 GMT 2010

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

```
# fasta34 3_3.pep /genedata/1/db/AD_2010 -Q -E 1 -O 3_3.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7, 2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_3, 33 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:            one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     0     2:*
  32     3     8:= *
  34    10    21:====  *
  36    67    44:==============*========
  38    53    72:==================     *
  40    76   101:==========================       *
  42   133   123:=========================================*====
  44   108   136:====================================          *
  46   126   138:=========================================   *
  48   132   132:============================================*
  50   117   121:======================================== *
  52    91   106:==============================    *
  54   136    91:==============================*===============
  56   118    76:=========================*==============
  58    48    62:================    *
  60    38    50:=============   *
  62    57    40:=============*=====
  64    26    32:========= *
  66    34    25:========*===
  68    24    20:======*=
  70    12    16:==== *
  72    10    12:===*
  74     7    10:===*
  76     9     7:==*
  78     8     6:=*=
  80    10     4:=*==
  82     2     3:*
  84     4     3:*=
  86     2     2:*
  88     0     2:*          inset = represents 1 library sequences
  90     2     1:*
  92     2     1:*         :*=
  94     1     1:*         :*
  96     0     1:*         :*
  98     1     0:=         *=
 100     0     0:          *
 102     1     0:=         *=
 104     1     0:=         *=
 106     0     0:          *
```

```
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.34330.00277; mu= 3.7279 0.144
 mean_var=30.5542 8.286, 0's: 2 Z-trim: 3  B-trim: 4 in 1/42
 Lambda= 0.232027
 Kolmogorov-Smirnov  statistic: 0.0598 (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(1471)
gi|71057064|emb|CAI38795.2| thaumatin-like protein ( 225)   57 23.6    0.83
gi|146737976|gb|ABQ42566.1| thaumatin-like protein ( 201)   56 23.3    0.93


>>gi|71057064|emb|CAI38795.2| thaumatin-like protein [Ac  (225 aa)
 initn:  41 init1:  41 opt:  57  Z-score: 103.8  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: 102.9  bits: 23.3 E(): 0.93
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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:02:07 2010 done: Tue Jan 26 21:02:07 2010
 Total Scan time:  0.030 Total Display time:  0.000
```

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

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

 3_3, 33 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    60      0:====
  22     0      0:            one = represents 18 library sequences
  24     0      0:
  26     0      0:
  28     6      2:*
  30    25     12:*=
  32    90     45:==*==
  34   131    122:======*=
  36   250    250:=============*
  38   572    414:======================*=========
  40   644    577:================================*===
  42   606    706:====================================     *
  44   512    779:============================                *
  46  1043    793:============================================*=============
  48   701    759:=========================================   *
  50   650    693:======================================== *
  52   774    609:================================*=========
  54   331    520:===================          *
  56   241    435:==============           *
  58   280    357:================    *
  60   346    289:================*===
  62   311    232:============*=====
  64   277    184:==========*=====
  66   127    146:========*
  68    58    115:====  *
  70    46     90:=== *
  72    61     70:===*
  74    48     55:===*
  76    36     43:==*
  78    48     33:=*=
  80    34     26:=*
  82    28     20:=*
  84     8     16:*
  86    29     12:*=
  88    16      9:*           inset = represents 1 library sequences
  90    40      7:*==
  92     9      6:*          :=====*===
  94     2      4:*          :== *
  96     1      3:*          := *
  98     0      3:*          :  *
 100     1      2:*          :=*
 102     0      2:*          : *
 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:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 0.82590.000591; mu= 23.6068 0.031
 mean_var=47.494711.084, 0's: 60 Z-trim: 60  B-trim: 586 in 2/60
 Lambda= 0.186102
 Kolmogorov-Smirnov  statistic: 0.0352 (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
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:02:08 2010 done: Tue Jan 26 21:02:08 2010
 Total Scan time:  0.160 Total Display time:  0.000

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

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

 3_3, 33 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 276182     0:==========
  22   245     0:=          one = represents 28083 library sequences
  24   617    17:*
  26  1360   374:*
  28  4793  4039:*
  30 18935 24536:*
  32 86170 94872:===*
  34 197687 257282:======== *
  36 451117 528397:================= *
  38 775444 873244:==========================   *
  40 1133002 1218099:=========================================  *
  42 1405120 1488978:================================================  *
  44 1605885 1642481:============================================================*
  46 1684965 1672908:============================================================*
  48 1632815 1601616:============================================================*=
  50 1495893 1461481:================================================*=
  52 1306756 1284886:============================================*=
  54 1078361 1097518:====================================*
  56 895222 916764:==============================*
  58 717945 752646:=========================*
  60 589353 609687:=====================*
  62 482581 488788:================*
  64 403424 388730:=============*=
  66 330195 307241:==========*=
  68 251574 241669:========*
  70 198629 189386:======*=
  72 161379 147987:=====*
  74 134808 115381:====*
  76 100114 89803:===*
  78 78905 69803:==*
  80 63352 54201:=*=
  82 54436 41463:=*
  84 41179 32843:=*
  86 32144 25412:*=
  88 25071 19663:*          inset = represents 327 library sequences
  90 19856 15214:*
  92 16318 11772:*         :===================================*====
  94 12344  9108:*         :============================*==========
  96  8960  7048:*         :======================*======
  98  6694  5453:*         :=================*=====
 100  5227  4219:*         :============*===
 102  5151  3265:*         :=========*======
 104  3877  2526:*         :=======*====
 106  3037  1955:*         :======*====
 108  2552  1512:*         :=====*===
 110  2532  1170:*         :===*=====
 112  2120   905:*         :==*====
 114  1679   701:*         :==*===
 116   920   542:*         :=*=
 118  2217   419:*         :=*=====
>120  2083   325:*         :*======
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17809440 sequences
  Expectation_n fit: rho(ln(x))= 3.79780.000177; mu= 7.3755 0.010
 mean_var=33.0312 6.380, 0's: 912 Z-trim: 915  B-trim: 0 in 0/63
 Lambda= 0.223158
 Kolmogorov-Smirnov  statistic: 0.0259 (N=29) at  62

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(17815538)
gi|124360394|gb|ABN08407.1| Peptidase aspartic, ac ( 435)  137 48.9 0.00047
gi|124360392|gb|ABN08405.1| Peptidase aspartic, ac ( 435)  137 48.9 0.00047
gi|124359710|gb|ABN06064.1| RNA-directed DNA polym (1297)  137 49.2  0.0012
gi|217073570|gb|ACJ85145.1| unknown [Medicago trun ( 185)  124 44.6  0.0042


>>gi|124360394|gb|ABN08407.1| Peptidase aspartic, active  (435 aa)
 initn: 137 init1: 137 opt: 137  Z-score: 235.4  bits: 48.9 E(): 0.00047
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: 235.4  bits: 48.9 E(): 0.00047
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: 228.2  bits: 49.2 E(): 0.0012
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: 218.4  bits: 44.6 E(): 0.0042
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



33 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:02:08 2010 done: Tue Jan 26 21:09:26 2010
 Total Scan time: 339.860 Total Display time:  0.000

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_2010
Query = 3_4

Start time: Tue Jan 26 21:09:27 GMT 2010 Finish time: Tue Jan 26 21:09:27 GMT 2010

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

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

 3_4, 8 aa
 vs /genedata/1/db/AD_2010 library

        opt      E()
< 20     2     0:=
  22     0     0:            one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     7     0:===
  30     4     2:*=
  32    22     8:==*=====
  34    18    21:======*
  36    30    44:==========    *
  38    48    72:================       *
  40    90   101:==============================   *
  42   107   123:=====================================    *
  44   154   136:=============================================*======
  46   102   138:==================================           *
  48   154   132:=============================================*========
  50   115   121:========================================== *
  52    92   106:==============================    *
  54    92    91:==============================*
  56    65    76:======================   *
  58    95    62:====================*===========
  60    45    50:=============== *
  62    55    40:=============*=====
  64    36    32:==========*=
  66    43    25:========*======
  68    26    20:======*==
  70    14    16:=====*
  72    12    12:===*
  74    13    10:===*=
  76     6     7:==*
  78     0     6: *
  80     7     4:=*=
  82     5     3:*=
  84     3     3:*
  86     3     2:*
  88     1     2:*          inset = represents 1 library sequences
  90     1     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:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.76610.00187; mu= -1.5052 0.099
 mean_var=10.9466 2.653, 0's: 2 Z-trim: 4  B-trim: 8 in 1/42
 Lambda= 0.387645
 Kolmogorov-Smirnov  statistic: 0.0478 (N=28) at  56

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(1471)
gi|169969|gb|AAA33964.1| glycinin [Glycine max]    ( 516)   40 22.9    0.78
gi|736002|emb|CAA55977.1| Gy5 [Glycine soja]       ( 517)   40 22.9    0.78


>>gi|169969|gb|AAA33964.1| glycinin [Glycine max]         (516 aa)
 initn:  40 init1:  40 opt:  40  Z-score: 104.3  bits: 22.9 E(): 0.78
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: 104.3  bits: 22.9 E(): 0.78
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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:09:26 2010 done: Tue Jan 26 21:09:26 2010
 Total Scan time:  0.010 Total Display time:  0.000

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

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

 3_4, 8 aa
 vs /genedata/1/db/TOX_2010 library
```

```
       opt      E()
< 20    60     0:=====
  22     0     0:           one = represents 14 library sequences
  24     0     0:
  26     0     0:
  28     5     2:*
  30    24    12:*=
  32    29    45:===*
  34    88   122:======= *
  36   319   250:==================*=====
  38   484   414:==============================*=====
  40   734   577:======================================*===========
  42   642   706:=============================================    *
  44   801   779:=========================================================*==
  46   751   793:=========================================================  *
  48   549   759:=========================================               *
  50   468   693:==================================               *
  52   717   609:============================================*========
  54   427   520:==============================      *
  56   438   435:==============================*
  58   282   357:====================    *
  60   384   289:====================*=======
  62   300   232:================*=====
  64   178   184:==============*
  66   194   146:==========*===
  68   138   115:========*=
  70   138    90:======*===
  72   109    70:====*===
  74    40    55:===*
  76    33    43:===*
  78    37    33:==*
  80    20    26:=*
  82    12    20:=*
  84     4    16:=*
  86     4    12:*
  88     3     9:*          inset = represents 1 library sequences
  90     8     7:*
  92    18     6:*=        :=====*============
  94     0     4:*         :   *
  96     3     3:*         :==*
  98     0     3:*         :  *
 100     2     2:*         :=*
 102     0     2:*         : *
 104     0     1:*         :*
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 2.41290.000443; mu= 5.2854 0.023
 mean_var=13.5473 2.924, 0's: 60 Z-trim: 60  B-trim: 200 in 1/61
 Lambda= 0.348456
 Kolmogorov-Smirnov  statistic: 0.0376 (N=29) at  58

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
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:09:27 2010 done: Tue Jan 26 21:09:27 2010
 Total Scan time:  0.100 Total Display time:  0.000

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

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

 3_4, 8 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 286675     0:============
  22   1130     0:=          one = represents 25427 library sequences
  24   2938    17:*
  26   9741   374:*
  28 30355  4039:*=
  30 73843 24538:*==
  32 164026 94881:===*===
  34 316835 257307:==========*==
  36 515436 528448:====================*
  38 752427 873328:==============================    *
  40 1019329 1218215:=========================================      *
  42 1227049 1489120:================================================         *
  44 1374550 1642638:=============================================================    *
  46 1525598 1673068:================================================================*
  48 1499390 1601769:================================================================*
  50 1481075 1461621:============================================================*=
  52 1357656 1285009:=========================================================*===
  54 1144494 1097623:==========================================*==
  56 989426 916851:====================================*==
  58 798439 752718:==============================*==
  60 663746 609745:========================*===
  62 552526 488835:====================*==
  64 465450 388768:================*===
  66 362148 307270:============*==
  68 300301 241692:=========*==
  70 223732 189404:========*=
  72 177803 148001:======*=
  74 121123 115392:=====*
  76 98253 89812:===*
  78 70310 69810:==*
  80 55900 54206:==*
  82 39811 41466:=*
  84 30881 32847:=*
  86 21702 25415:*
  88 15632 19665:*          inset = represents 166 library sequences
  90 11017 15216:*
  92  8276 11773:*         :=======================================*
  94  6330  9109:*         :=======================================*
  96  4095  7048:*         :=========================              *
  98  3075  5454:*         :===================           *
 100  2014  4220:*         :=============            *
 102  3026  3265:*         :====================*
 104  1147  2526:*         :=======         *
 106   636  1955:*         :====       *
 108   706  1512:*         :=====    *
 110   458  1170:*         :===    *
 112   177   905:*         :==   *
 114   136   701:*         :=   *
 116   240   542:*         :== *
 118    66   419:*         := *
>120    96   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17811143 sequences
  Expectation_n fit: rho(ln(x))= 3.12780.000165; mu= 2.5244 0.009
 mean_var=14.0979 2.757, 0's: 926 Z-trim: 928  B-trim: 0 in 0/65
 Lambda= 0.341583
 Kolmogorov-Smirnov  statistic: 0.0511 (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
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:09:27 2010 done: Tue Jan 26 21:13:22 2010
 Total Scan time: 200.460 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_2010
Query = 3_5

Start time: Tue Jan 26 21:13:23 GMT 2010 Finish time: Tue Jan 26 21:13:23 GMT 2010

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

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

 3_5, 64 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     7     0:===
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     1     0:=
  30     1     2:*
  32     2     8:= *
  34     4    21:==    *
```

```
  36    21    44:=======       *
  38    51    72:=================      *
  40   145   101:==================================*===============
  42   134   123:============================================*====
  44   128   136:=============================================  *
  46   129   138:=============================================  *
  48   123   132:=========================================  *
  50   103   121:====================================     *
  52   131   106:====================================*========
  54   101    91:==============================*===
  56    79    76:=========================*=
  58    59    62:====================*
  60    37    50:=============   *
  62    26    40:=========    *
  64    37    32:==========*==
  66    24    25:========*
  68    30    20:======*===
  70    20    16:=====*=
  72    13    12:===*=
  74    14    10:===*=
  76    12     7:==*=
  78    17     6:=*====
  80     4     4:=*
  82     3     3:*
  84     2     3:*
  86     6     2:*=
  88     5     2:*=         inset = represents 1 library sequences
  90     1     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:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.65330.00365; mu= 3.8297 0.191
 mean_var=51.840115.171, 0's: 7 Z-trim: 7  B-trim: 52 in 1/42
 Lambda= 0.178132
 Kolmogorov-Smirnov  statistic: 0.0459 (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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:13:22 2010 done: Tue Jan 26 21:13:22 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

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

 3_5, 64 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    73     0:=====
  22     0     0:           one = represents 15 library sequences
  24     0     0:
  26     0     0:
  28     1     2:*
  30     7    12:*
  32    13    45:= *
  34   191   122:========*====
  36   321   250:================*=====
  38   470   414:===========================*====
  40   484   577:================================     *
  42   795   706:================================================*=====
  44   899   779:=================================================*========
  46   763   793:================================================= *
  48   714   759:============================================  *
  50   604   693:=========================================     *
  52   381   609:=========================              *
  54   421   520:==============================     *
  56   333   435:=======================     *
  58   402   357:=======================*===
  60   597   289:==================*====================
  62   262   232:===============*==
  64   166   184:============*
  66   101   146:=======  *
  68    94   115:=======*
  70    69    90:=====*
  72   101    70:====*==
  74    37    55:===*
  76    21    43:==*
  78    38    33:==*
  80    15    26:=*
  82     9    20:=*
  84     7    16:=*
  86     9    12:*
  88    10     9:*          inset = represents 1 library sequences
  90     7     7:*
  92     2     6:*         :==   *
  94     4     4:*         :===*
  96     6     3:*         :==*===
  98     7     3:*         :==*====
 100     0     2:*         : *
 102     5     2:*         :=*===
 104     0     1:*         :*
 106     2     1:*         :*=
 108     1     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     1     0:=         *=
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 2069351 residues in  8448 sequences
```

```
  Expectation_n fit: rho(ln(x))= 4.16720.000604; mu= 6.7446 0.030
 mean_var=44.5817 9.804, 0's: 73 Z-trim: 73  B-trim: 0 in 0/62
 Lambda= 0.192086
 Kolmogorov-Smirnov  statistic: 0.0382 (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


64 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:13:23 2010 done: Tue Jan 26 21:13:23 2010
 Total Scan time:  0.120 Total Display time:  0.000

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

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

 3_5, 64 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 344904     0:=============
  22   105     0:=          one = represents 27633 library sequences
  24   223    17:*
  26   539   374:*
  28  2278  4039:*
  30 12454 24538:*
  32 61443 94880:===*
  34 211583 257302:======== *
  36 513212 528439:====================*
  38 919113 873312:==================================*==
  40 1371510 1218194:=============================================*=====
  42 1624506 1489094:============================================================*=====
  44 1657940 1642609:================================================================*
  46 1596670 1673039:============================================================== *
  48 1501591 1601742:=============================================================  *
  50 1333655 1461596:================================================   *
  52 1160756 1284986:============================================    *
  54 980943 1097604:====================================   *
  56 852031 916835:==============================  *
  58 735205 752705:============================*
  60 598201 609735:======================*
  62 492084 488826:==================*
  64 395533 388761:===============*
  66 321387 307265:============*
  68 249208 241688:=========*=
  70 195568 189401:=======*=
  72 153165 147999:======*
  74 119990 115390:=====*
  76 94154 89811:===*
  78 77750 69808:==*
  80 57431 54205:=*=
  82 43511 41466:=*
  84 32736 32846:=*
  86 23752 25414:*
  88 18762 19664:*          inset = represents 209 library sequences
  90 14494 15215:*
  92 10442 11773:*         :=======================================*
  94  8152  9109:*         :=======================================*
  96  6168  7048:*         :==============================    *
  98  5201  5454:*         :========================== *
 100  3334  4220:*         :================    *
 102  2342  3265:*         :============   *
 104  1732  2526:*         :==========   *
 106  1382  1955:*         :=======  *
 108  1254  1512:*         :======= *
 110   775  1170:*         :==== *
 112   592   905:*         :=== *
 114   443   701:*         :===*
 116   353   542:*         :==*
 118   204   419:*         := *
>120   464   325:*         :=*=
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810834 sequences
  Expectation_n fit: rho(ln(x))= 4.06700.000182; mu= 6.6846 0.010
 mean_var=45.6406 9.197, 0's: 1176 Z-trim: 1176  B-trim: 0 in 0/66
 Lambda= 0.189845
 Kolmogorov-Smirnov  statistic: 0.0227 (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
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:13:23 2010 done: Tue Jan 26 21:18:04 2010
 Total Scan time: 240.480 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_2010
Query = 3_6

Start time: Tue Jan 26 21:18:04 GMT 2010 Finish time: Tue Jan 26 21:18:04 GMT 2010

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

```
# fasta34 3_6.pep /genedata/1/db/AD_2010 -Q -E 1 -O 3_6.pep_ad.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7, 2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_6, 45 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
```

```
< 20    17     0:======
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     0     2:*
  32     1     8:= *
  34     9    21:===   *
  36    19    44:=======       *
  38    40    72:==============         *
  40    96   101:================================ *
  42   159   123:=========================================*============
  44   177   136:=============================================*============
  46   129   138:============================================  *
  48   110   132:====================================      *
  50   104   121:==================================     *
  52    98   106:================================  *
  54    83    91:============================  *
  56    89    76:=========================*=====
  58    70    62:====================*===
  60    44    50:================ *
  62    52    40:=============*====
  64    19    32:=======   *
  66    40    25:========*=====
  68    30    20:======*===
  70    20    16:=====*=
  72     4    12:== *
  74    10    10:===*
  76     9     7:==*
  78     1     6:=*
  80     6     4:=*
  82     3     3:*
  84     1     3:*
  86     1     2:*
  88     9     2:*==        inset = represents 1 library sequences
  90     0     1:*
  92     2     1:*         :*=
  94    17     1:*=====    :*================
  96     1     1:*         :*
  98     0     0:          *
 100     0     0:          *
 102     1     0:=         *=
 104     0     0:          *
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.33210.00375; mu= 6.7936 0.193
 mean_var=35.659010.283, 0's: 17 Z-trim: 17  B-trim: 58 in 1/41
 Lambda= 0.214778
 Kolmogorov-Smirnov  statistic: 0.0570 (N=29) 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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:18:04 2010 done: Tue Jan 26 21:18:04 2010
 Total Scan time:  0.010 Total Display time:  0.000

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

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

 3_6, 45 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    77     0:======
  22     0     0:            one = represents 15 library sequences
  24     1     0:=
  26     0     0:
  28     3     2:*
  30    23    12:*=
  32    59    45:==*=
  34   119   122:========*
  36   306   250:================*====
  38   370   414:=========================  *
  40   443   577:==============================         *
  42   629   706:==========================================     *
  44   841   779:================================================*=====
  46   690   793:=============================================      *
  48   746   759:================================================*
  50   681   693:=============================================*
  52   635   609:=========================================*==
  54   766   520:==================================*=================
  56   374   435:=========================   *
  58   232   357:================       *
  60   227   289:================   *
  62   326   232:===============*======
  64   192   184:============*
  66   118   146:======== *
  68   109   115:========*
  70    56    90:==== *
  72    75    70:====*
  74    56    55:===*
  76    79    43:==*===
  78    17    33:==*
  80    26    26:=*
  82    15    20:=*
  84    19    16:=*
  86    32    12:*==
  88    15     9:*           inset = represents 1 library sequences
  90    12     7:*
  92     7     6:*          :=====*=
  94    37     4:*==        :===*=================================
  96    15     3:*          :==*============
  98     1     3:*          := *
 100     2     2:*          :=*
 102     7     2:*          :=*=====
 104     1     1:*          :*
 106     0     1:*          :*
```

```
 108     0     1:*         :*
 110     0     1:*         :*
 112     3     0:=         *===
 114     0     0:          *
 116     1     0:=         *=
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.32300.000597; mu= 2.6989 0.029
 mean_var=31.0680 6.959, 0's: 77 Z-trim: 77  B-trim: 196 in 1/61
 Lambda= 0.230100
 Kolmogorov-Smirnov  statistic: 0.0289 (N=29) at  50

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(8448)
gi|256378|gb|AAB23447.1| neurotoxin Tx2-9 [Phoneut (  32)   55 23.7    0.88


>>gi|256378|gb|AAB23447.1| neurotoxin Tx2-9 [Phoneutria   (32 aa)
 initn:  55 init1:  55 opt:  55  Z-score: 117.0  bits: 23.7 E(): 0.88
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
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:18:05 2010 done: Tue Jan 26 21:18:05 2010
 Total Scan time:  0.100 Total Display time:  0.000

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

```
# fasta34 3_6.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 3_6.pep_prt.fasta
FASTA searches a protein or DNA sequence data bank version 3.4t26 July 7, 2006
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 3_6, 45 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 373323     0:==============
  22   366     0:=          one = represents 29048 library sequences
  24   298    17:*
  26   655   374:*
  28  1923  4039:*
  30  9547 24538:*
  32 42660 94879:== *
  34 152298 257301:======  *
  36 382989 528435:===============    *
  38 741050 873306:==========================    *
  40 1154550 1218185:======================================== *
  42 1530099 1489083:==================================================*=
  44 1742864 1642598:============================================================*===
  46 1693387 1673027:============================================================*=
  48 1557953 1601730:============================================================ *
  50 1341676 1461585:================================================    *
  52 1135311 1284977:============================================    *
  54 1009432 1097596:==================================  *
  56 869235 916829:============================== *
  58 750564 752699:==========================*
  60 644776 609730:====================*==
  62 519488 488823:================*=
  64 427184 388758:=============*=
  66 353024 307263:==========*==
  68 284407 241686:========*=
  70 232404 189400:======*==
  72 198322 147998:=====*=
  74 144885 115389:===*=
  76 114961 89810:===*
  78 92079 69808:==*=
  80 73058 54205:=*=
  82 57699 41465:=*
  84 45145 32846:=*
  86 31958 25414:*=
  88 25037 19664:*          inset = represents 286 library sequences
  90 18133 15215:*
  92 14280 11773:*         :=======================================*
  94 10942  9109:*         :================================*=======
  96  7973  7048:*         :========================*===
  98  5296  5453:*         :====================*
 100  4359  4220:*         :==============*=
 102  3409  3265:*         :============*
 104  2379  2526:*         :========*
 106  1685  1955:*         :=======*
 108  1200  1512:*         :======*
 110  1054  1170:*         :=====*
 112   883   905:*         :===*
 114  1886   701:*         :==*====
 116  2280   542:*         :=*======
 118   256   419:*         :=*
>120   603   325:*         :=*=
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810705 sequences
  Expectation_n fit: rho(ln(x))= 3.86510.000181; mu= 5.4815 0.010
 mean_var=36.5725 7.445, 0's: 1279 Z-trim: 1280  B-trim: 3361 in 1/64
 Lambda= 0.212079
 Kolmogorov-Smirnov  statistic: 0.0461 (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
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:18:05 2010 done: Tue Jan 26 21:22:29 2010
 Total Scan time: 205.530 Total Display time:  0.000

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

Appendix 13. Bioinformatic analysis of polypeptide 5_1a

```
>5_1a
CPNTHWWDSK STRGEQNVFY IINIQTKILK TVKQRIsgcc yhcglafgpr hrcpeknmrv vilakde
_____________________________________________________

Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_1a

Start time: Tue Jan 26 21:22:30 GMT 2010 Finish time: Tue Jan 26 21:22:30 GMT 2010

No 8 amino acid matches exist between 5_1a and the AD_2010 database
_____________________________________________________

# fasta34 5_1a.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_1a.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_1a, 67 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20    10     0:====
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     3     2:*
  32     6     8:==*
  34     9    21:===   *
  36    14    44:=====         *
  38    82    72:======================*====
  40    62   101:=====================           *
  42   144   123:=========================================*=======
  44   111   136:====================================        *
  46   135   138:================================================*
  48   103   132:==================================        *
  50    99   121:================================       *
  52   151   106:===================================*===============
  54   124    91:==============================*===========
  56    71    76:======================== *
  58    79    62:====================*======
  60    63    50:================*====
  62    40    40:=============*
  64    29    32:==========*
  66    23    25:=========*
  68    22    20:=======*=
  70    19    16:======*=
  72    23    12:===*====
  74     8    10:===*
  76    12     7:==*=
  78     4     6:=*
  80     2     4:=*
  82    11     3:*===
  84     2     3:*
  86     5     2:*=
  88     2     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     1     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     0     0:          *
 100     0     0:          *
 102     1     0:=         *=
 104     0     0:          *
 106     1     0:=         *=
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 5.65520.00361; mu= -2.7317 0.189
 mean_var=43.173011.435, 0's: 10 Z-trim: 11  B-trim: 3 in 1/42
 Lambda= 0.195195
 Kolmogorov-Smirnov  statistic: 0.0888 (N=29) at  50

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 36, opt: 24, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(1471)
gi|5059162|gb|AAD38942.1|AF144060_1 alpha-amylase  ( 496)   69 26.2    0.65
gi|481397|pir||S38584 allergen Phl p Vb - common t ( 280)   64 24.8    0.91


>>gi|5059162|gb|AAD38942.1|AF144060_1 alpha-amylase [Der  (496 aa)
 initn:  39 init1:  39 opt:  69  Z-score: 105.8  bits: 26.2 E(): 0.65
Smith-Waterman score: 69;  26.087% identity (57.971% similar) in 69 aa overlap (2-
60:401-467)

                                          10            20
5_1a                                CPNTHWWDSKS-----TRGEQNVFYIINIQT
                                     :  .:::. .     .::.. .:  ::.:
gi|505 DMTCNHEWICEHRWREIYNMVKFRMIAGQEPVHNWWDNGDYQIAFSRGNR-AFIAINLQ-
             380       390       400       410       420

         30          40          50        60
5_1a   KILKTVKQRI-SG--CCYHCGLAFGP--RHRCPEKNMRVVILAKDE
       :  ....:.. .:     .: .  :    ..:  :...:
gi|505 KNQQNLQQKLHTGLPAGTYCDIISGNLIDNKCTGKSIHVDKNGQADVYVGHDEFDAFVAY
      430       440       450       460       470       480

gi|505 HIGARIVS
      490

>>gi|481397|pir||S38584 allergen Phl p Vb - common timot  (280 aa)
 initn:  61 init1:  61 opt:  64  Z-score: 103.1  bits: 24.8 E(): 0.91
Smith-Waterman score: 64;  31.111% identity (60.000% similar) in 45 aa overlap (10-
54:36-80)

                                    10        20        30
5_1a                          CPNTHWWDSKSTRGEQNVFYIINIQTKILKTVKQRISGC
                                       :.:  ::...  ::.  :   ...:: ..
gi|481 RGPRGGPGRSYTADAGYAPATPAAAGAAAGKATTEEQKLIEDINVGFKAAVAARQRPAAD
         10        20        30        40        50        60

      40        50        60
5_1a   CYHCGLAFGPRHRCPEKNMRVVILAKDE
        ..   : .:::  :
gi|481 KFKTFEAASPRHPRPLRQGAGLVPKLDAAYSVAYKAAVGATPEAKFDSFVASLTEALRVI
          70        80        90       100       110       120
```

```
67 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:22:29 2010 done: Tue Jan 26 21:22:29 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 5_1a.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_1a.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_1a, 67 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    63     0:====
  22     0     0:           one = represents 16 library sequences
  24     0     0:
  26     0     0:
  28     2     2:*
  30    11    12:*
  32    65    45:==*==
  34   203   122:=======*=====
  36   455   250:===============*=============
  38   439   414:=========================*==
  40   362   577:=======================            *
  42   944   706:================================================*==============
  44   643   779:=========================================       *
  46   583   793:====================================             *
  48   442   759:============================                    *
  50   808   693:=============================================*=======
  52   652   609:======================================*==
  54   472   520:==============================  *
  56   466   435:============================*==
  58   410   357:=======================*===
  60   464   289:==================*==========
  62   179   232:============  *
  64   150   184:========== *
  66   160   146:=========*
  68   127   115:========*
  70    86    90:======*
  72    70    70:====*
  74    56    55:===*
  76    22    43:==*
  78    32    33:==*
  80    14    26:=*
  82    10    20:=*
  84     2    16:*
  86     8    12:*
  88     2     9:*          inset = represents 1 library sequences
  90     3     7:*
  92     2     6:*         :==   *
  94     1     4:*         :=  *
  96     1     3:*         := *
  98     0     3:*         :  *
 100    33     2:*==       :=*==============================
 102     0     2:*         : *
 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:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 1.30920.000623; mu= 22.2565 0.032
 mean_var=62.276414.296, 0's: 63 Z-trim: 63  B-trim: 598 in 1/61
 Lambda= 0.162522
 Kolmogorov-Smirnov  statistic: 0.0421 (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
!! No sequences with E() < 1.000000


67 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:22:30 2010 done: Tue Jan 26 21:22:30 2010
 Total Scan time:  0.120 Total Display time:  0.000

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

```
# fasta34 5_1a.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_1a.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_1a, 67 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 327800     0:=============
  22   121     0:=          one = represents 26485 library sequences
  24   258    17:*
  26   601   374:*
  28  2156  4039:*
  30 13043 24538:*
  32 64821 94878:===*
  34 209848 257299:======== *
  36 493964 528431:====================*
  38 942181 873299:==================================*===
  40 1298818 1218176:================================================*====
  42 1524892 1489072:============================================================*=
  44 1589074 1642585:==============================================================*
  46 1540665 1673014:==============================================================*
  48 1442666 1601718:============================================================    *
  50 1323306 1461574:================================================     *
  52 1181922 1284967:============================================   *
  54 1037192 1097587:======================================== *
  56 917027 916822:====================================*
  58 772061 752693:============================*=
  60 649680 609726:======================*=
  62 513483 488819:==================*=
  64 427193 388755:==============*==
```

```
  66 335517 307260:===========*=
  68 264598 241685:==========*
  70 204916 189398:========*
  72 160537 147997:======*=
  74 132966 115388:=====*=
  76 101482 89809:===*
  78 78236 69807:==*
  80 60111 54204:==*
  82 47171 41465:=*
  84 35512 32845:=*
  86 25907 25414:*
  88 21164 19664:*          inset = represents 241 library sequences
  90 15286 15215:*
  92 12026 11773:*         :=======================================*
  94  9919  9109:*         :========================================*==
  96  7397  7048:*         :============================*=
  98  6076  5453:*         :=======================*===
 100  5027  4220:*         :==================*===
 102  3585  3265:*         :=============*=
 104  3376  2526:*         :==========*====
 106  2682  1955:*         :========*===
 108  1293  1512:*         :======*
 110   996  1170:*         :====*
 112   672   905:*         :===*
 114   613   701:*         :==*
 116   344   542:*         :==*
 118   301   419:*         :=*
>120   743   325:*         :=*==
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810568 sequences
  Expectation_n fit: rho(ln(x))= 4.06710.000187; mu= 7.1124 0.010
 mean_var=48.2153 9.552, 0's: 1051 Z-trim: 1051  B-trim: 0 in 0/66
 Lambda= 0.184706
 Kolmogorov-Smirnov  statistic: 0.0335 (N=29) at  54

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(17815538)
gi|124360394|gb|ABN08407.1| Peptidase aspartic, ac ( 435)  135 43.2   0.053
gi|124360392|gb|ABN08405.1| Peptidase aspartic, ac ( 435)  135 43.2   0.053
gi|124359710|gb|ABN06064.1| RNA-directed DNA polym (1297)  137 44.1   0.083
gi|217073570|gb|ACJ85145.1| unknown [Medicago trun ( 185)  127 40.7    0.12


>>gi|124360394|gb|ABN08407.1| Peptidase aspartic, active  (435 aa)
 initn: 107 init1: 107 opt: 135  Z-score: 198.6  bits: 43.2 E(): 0.053
Smith-Waterman score: 135;  45.455% identity (75.758% similar) in 33 aa overlap
(34-66:58-90)

            10        20        30        40        50        60
5_1a   THWWDSKSTRGEQNVFYIINIQTKILKTVKQRISGCCYHCGLAFGPRHRCPEKNMRVVIL
                                     .: .: :..::  . :::.::.::. :..:
gi|124 NVGQNKTHTINTANWRDKNVRSLSSQEIADRRQKGLCFKCGGPYHPRHQCPDKNLSVMVL
        30        40        50        60        70        80


5_1a   AKDE
         :
gi|124 EDDSEDENEVRVLNDEDVDTGAEELQLNVLTFENALTFDRQTEYYQDRFQCIRFQGKVRE
        90       100       110       120       130       140

>>gi|124360392|gb|ABN08405.1| Peptidase aspartic, active  (435 aa)
 initn: 107 init1: 107 opt: 135  Z-score: 198.6  bits: 43.2 E(): 0.053
Smith-Waterman score: 135;  45.455% identity (75.758% similar) in 33 aa overlap
(34-66:58-90)

            10        20        30        40        50        60
5_1a   THWWDSKSTRGEQNVFYIINIQTKILKTVKQRISGCCYHCGLAFGPRHRCPEKNMRVVIL
                                     .: .: :..::  . :::.::.::. :..:
gi|124 NVGQNKTHTINTANWRDKNVRSLSSQEIADRRQKGLCFKCGGPYHPRHQCPDKNLSVMVL
        30        40        50        60        70        80


5_1a   AKDE
         :
gi|124 EDDSEDENEVRVLNDEDVDTGAEELQLNVLTFENALTFDRQTEYYQDRFQCIRFQGKVRE
        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: 195.1  bits: 44.1 E(): 0.083
Smith-Waterman score: 137;  41.667% identity (80.556% similar) in 36 aa overlap
(32-67:105-140)

              10        20        30        40        50        60
5_1a   PNTHWWDSKSTRGEQNVFYIINIQTKILKTVKQRISGCCYHCGLAFGPRHRCPEKNMRVV
                                     .... .: :..::  : : :.::.:..::.
gi|124 GPKGEKQAQYDKKKSGPRDRSFTHLSYNELMERKQKGLCFKCGGPFHPMHQCPDKQLRVL
            80        90       100       110       120       130


5_1a   ILAKDE
       .: .::
gi|124 VLEEDEEGEPEGKLLAVEVDDEEEGDGEMCMMEFFHLGHSRPQSIKLMGVIKEVPVVVLV
          140       150       160       170       180       190

>>gi|217073570|gb|ACJ85145.1| unknown [Medicago truncatu  (185 aa)
 initn: 106 init1:  68 opt: 127  Z-score: 192.1  bits: 40.7 E(): 0.12
Smith-Waterman score: 127;  44.444% identity (77.778% similar) in 36 aa overlap
(33-67:3-38)

             10        20        30        40        50         60
5_1a   NTHWWDSKSTRGEQNVFYIINIQTKILKTVKQRISGCCYHCGLAFGPR-HRCPEKNMRVV
                                     ..: .: :..::  . :  :.::::..::.
gi|217                               MAERRAKGLCFKCGGKYHPTLHKCPEKSLRVL
                                             10        20        30


5_1a   ILAKDE
       ::.. :
gi|217 ILGEGEGVNEEGEIVSLETQEVLEEEEEEIESECKVIGVLGSMGEYNTMKIGGKLENIDV
              40        50        60        70        80        90




67 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:22:30 2010 done: Tue Jan 26 21:27:43 2010
 Total Scan time: 249.480 Total Display time:  0.000

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

Appendix 14. Bioinformatic analysis of polypeptide 5_2a

```
>5_2a
NKGsqgvvit avwplgqgtv vlkki
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_2a

Start time: Tue Jan 26 21:27:44 GMT 2010 Finish time: Tue Jan 26 21:27:44 GMT 2010

No 8 amino acid matches exist between 5_2a and the AD_2010 database
```

```
# fasta34 5_2a.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_2a.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_2a, 25 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:            one = represents 3 library sequences
  24     3     0:=
  26    15     0:=====
  28     3     0:=
  30    16     2:*=====
  32    21     8:==*====
  34    16    21:======*
  36    50    44:==============*==
  38    35    72:============           *
  40    65   101:=======================            *
  42    85   123:==============================             *
  44    97   136:===================================             *
  46   114   138:========================================        *
  48   148   132:============================================*======
  50   169   121:============================================*================
  52   130   106:===================================*========
  54    80    91:===========================    *
  56    74    76:=========================*
  58    50    62:==================    *
  60    84    50:================*===========
  62    42    40:==============*
  64    14    32:=====      *
  66    16    25:======  *
  68    18    20:======*
  70    16    16:=====*
  72    23    12:===*====
  74    19    10:===*===
  76     8     7:==*
  78    18     6:=*====
  80    11     4:=*==
  82     8     3:*==
  84     2     3:*
  86     3     2:*
  88     8     2:*==        inset = represents 1 library sequences
  90     5     1:*=
  92     2     1:*          :*=
  94     1     1:*          :*
  96     0     1:*          :*
  98     0     0:           *
 100     0     0:           *
 102     0     0:           *
 104     0     0:           *
 106     0     0:           *
 108     0     0:           *
 110     0     0:           *
 112     0     0:           *
 114     0     0:           *
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.58170.00295; mu= 6.2337 0.156
 mean_var=23.8448 6.383, 0's: 2 Z-trim: 2  B-trim: 219 in 1/42
 Lambda= 0.262650
 Kolmogorov-Smirnov  statistic: 0.0849 (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
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:27:44 2010 done: Tue Jan 26 21:27:44 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 5_2a.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_2a.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_2a, 25 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    61     0:====
  22     0     0:            one = represents 18 library sequences
  24     6     0:=
  26     6     0:=
  28    11     2:*
  30    23    12:*=
  32    43    45:==*
  34   166   122:======*===
  36   284   250:==============*==
  38   314   414:==================     *
  40   428   577:=========================         *
  42   581   706:==================================       *
  44   741   778:============================================ *
  46  1032   793:=============================================*==============
  48   648   759:=======================================       *
  50   794   693:=======================================*======
  52   764   609:==================================*=========
  54   486   520:=========================== *
  56   323   434:==================       *
```

```
  58   392   357:====================*==
  60   384   289:=================*=====
  62   152   232:==========   *
  64   132   184:=========  *
  66    97   146:=======  *
  68   100   115:=======*
  70   120    90:=====*==
  72    43    70:===*
  74    47    55:===*
  76    38    43:==*
  78    45    33:=*=
  80    52    26:=*=
  82    35    20:=*
  84    13    16:*
  86     5    12:*
  88    15     9:*          inset = represents 1 library sequences
  90     7     7:*
  92     5     6:*         :======*
  94     6     4:*         :===*==
  96    10     3:*         :==*=======
  98    14     3:*         :==*============
 100     0     2:*         : *
 102     0     2:*         : *
 104    13     1:*         :*============
 106     4     1:*         :*===
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     3     0:=         *===
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 3.25940.000512; mu= 7.6593 0.026
 mean_var=19.1246 3.789, 0's: 60 Z-trim: 64  B-trim: 598 in 1/61
 Lambda= 0.293277
 Kolmogorov-Smirnov  statistic: 0.0362 (N=29) at  44

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(8448)
gi|238624339|gb|ACR47045.1| toxin of toxin-antitox ( 134)   59 27.5    0.15
gi|228021815|gb|ACP53523.1| Toxin of toxin-antitox ( 134)   59 27.5    0.15
gi|67004657|gb|AAY61583.1| Toxin of toxin-antitoxi ( 134)   55 25.8    0.48


>>gi|238624339|gb|ACR47045.1| toxin of toxin-antitoxin s  (134 aa)
 initn:  59 init1:  59 opt:  59  Z-score: 130.9  bits: 27.5 E(): 0.15
Smith-Waterman score: 59;  36.842% identity (78.947% similar) in 19 aa overlap (4-
22:31-49)

                                      10        20
5_2a                               NKGSQGVVITAVWPLGQGTVVLKKI
                                     ..:... :.: :  :.:.:
gi|238 MVSFVLDSSIALSWLMPDEVASLDILDKTITEGAIVPAIWGLEIGNVLLCAERAKRLTAN
               10        20        30        40        50        60

gi|238 QRHQAIYTLKDLYIKIDQITLEHIWFETMDLAVQYGLTLYDASYLELVLRCGLPIATLDK
               70        80        90       100       110       120

>>gi|228021815|gb|ACP53523.1| Toxin of toxin-antitoxin (  (134 aa)
 initn:  59 init1:  59 opt:  59  Z-score: 130.9  bits: 27.5 E(): 0.15
Smith-Waterman score: 59;  36.842% identity (78.947% similar) in 19 aa overlap (4-
22:31-49)

                                      10        20
5_2a                               NKGSQGVVITAVWPLGQGTVVLKKI
                                     ..:... :.: :  :.:.:
gi|228 MVSFVLDSSIALSWLMPDEVASLDILDKTITEGAIVPAIWGLEIGNVLLCAERAKRLTAN
               10        20        30        40        50        60

gi|228 QRHQAIYTLKDLYIKIDQITLEHIWFETMDLAVQYGLTLYDASYLELVLRCGLPIATLDK
               70        80        90       100       110       120

>>gi|67004657|gb|AAY61583.1| Toxin of toxin-antitoxin sy  (134 aa)
 initn:  55 init1:  55 opt:  55  Z-score: 121.7  bits: 25.8 E(): 0.48
Smith-Waterman score: 55;  31.579% identity (78.947% similar) in 19 aa overlap (4-
22:31-49)

                                      10        20
5_2a                               NKGSQGVVITAVWPLGQGTVVLKKI
                                     ..:... ..: :  :.:.:
gi|670 MVSFVLDSSIALSWLMPDEVASLDILDKTITEGTIVPSIWGLEIGNVLLCAERAKRLTSN
               10        20        30        40        50        60

gi|670 QRHQAIYTLKDIYIKVDQITLEHTWFETMDLALQYGLTLYDASYLELALRCGLPIATLDK
               70        80        90       100       110       120




25 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:27:44 2010 done: Tue Jan 26 21:27:44 2010
 Total Scan time:  0.160 Total Display time:  0.000

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

```
# fasta34 5_2a.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_2a.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_2a, 25 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 279408     0:===========
  22   1371     0:=          one = represents 26443 library sequences
  24   2946    17:*
  26   7044   374:*
  28 20244  4039:*
  30 53411 24538:*==
  32 133560 94881:===*==
  34 282631 257305:=========*=
  36 498862 528444:====================*
  38 792315 873320:==============================   *
  40 1073864 1218205:=========================================     *
  42 1342902 1489107:=================================================     *
  44 1512359 1642624:============================================================ *
  46 1586569 1673054:============================================================*
  48 1545794 1601756:============================================================*
```

```
  50 1471049 1461609:=========================================================*
  52 1345689 1284998:=================================================*==
  54 1179878 1097614:============================================*===
  56 963068 916844:=================================*==
  58 768147 752711:============================*=
  60 608858 609740:=======================*
  62 496221 488831:==================*
  64 405326 388764:===============*=
  66 319685 307268:============*=
  68 248166 241690:==========*
  70 193675 189403:========*
  72 158738 148000:======*=
  74 115674 115391:====*
  76 96818 89811:===*
  78 73919 69809:==*
  80 63480 54206:==*
  82 43643 41466:=*
  84 32794 32846:=*
  86 23487 25415:*
  88 18816 19665:*          inset = represents 209 library sequences
  90 13665 15215:*
  92 10411 11773:*         :=======================================*
  94  7436  9109:*         :====================================   *
  96  5376  7048:*         :==========================       *
  98  4049  5454:*         :====================      *
 100  2819  4220:*         :==============      *
 102  2103  3265:*         :===========    *
 104  1328  2526:*         :=======     *
 106  1082  1955:*         :======   *
 108   880  1512:*         :=====  *
 110   538  1170:*         :===  *
 112   322   905:*         :==  *
 114   215   701:*         :== *
 116   186   542:*         := *
 118   130   419:*         := *
>120   274   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810994 sequences
  Expectation_n fit: rho(ln(x))= 3.94940.000172; mu= 4.9929 0.009
 mean_var=25.3950 4.973, 0's: 948 Z-trim: 951  B-trim: 0 in 0/65
 Lambda= 0.254507
 Kolmogorov-Smirnov  statistic: 0.0316 (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
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:27:44 2010 done: Tue Jan 26 21:33:56 2010
 Total Scan time: 325.310 Total Display time:  0.000

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

Appendix 15. Bioinformatic analysis of polypeptide 5_3a

```
>5_3a
GSLVVPIMPK HPLVGLKIYK GRAECLLHHQ YPNQDSQDRE TKDlrvllsl rfglwakapl s
_____________________________________________________
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_3a

Start time: Tue Jan 26 21:33:57 GMT 2010 Finish time: Tue Jan 26 21:33:57 GMT 2010

No 8 amino acid matches exist between 5_3a and the AD_2010 database
_____________________________________________________

# fasta34 5_3a.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_3a.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_3a, 61 aa
 vs /genedata/1/db/AD_2010 library


       opt      E()
< 20     8      0:==
  22     0      0:            one = represents 4 library sequences
  24     0      0:
  26     1      0:=
  28     0      0:
  30     0      2:*
  32     1      8:=*
  34    13     21:==== *
  36    13     44:====       *
  38    68     72:==================*
  40    85    101:======================    *
  42   130    123:==============================*==
  44   111    136:============================      *
  46   113    138:============================      *
  48   195    132:================================*================
  50    94    121:========================       *
  52   117    106:=========================*===
  54    76     91:===================    *
  56    67     76:================== *
  58    61     62:===============*
  60    49     50:============*
  62    55     40:=========*====
  64    31     32:=======*
  66    34     25:======*==
  68    20     20:====*
  70    27     16:===*===
  72    25     12:==*====
  74    18     10:==*==
  76    10      7:=*=
  78     8      6:=*
  80     2      4:*
  82     5      3:*=
  84    21      3:*=====
  86     3      2:*
  88     6      2:*=          inset = represents 1 library sequences
  90     1      1:*
  92     2      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:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 4.87040.00339; mu= 0.2688 0.176
 mean_var=35.2020 9.401, 0's: 8 Z-trim: 8  B-trim: 113 in 1/41
 Lambda= 0.216168
 Kolmogorov-Smirnov  statistic: 0.0750 (N=29) at  46

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


61 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:33:56 2010 done: Tue Jan 26 21:33:57 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 5_3a.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_3a.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_3a, 61 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    82     0:=====
  22     0     0:           one = represents 17 library sequences
  24    13     0:=
  26     4     0:=
  28     4     2:*
  30    10    12:*
  32    28    45:==*
  34   138   122:=======*=
  36   228   250:==============*
  38   557   414:========================*========
  40   612   577:==================================*==
  42   589   706:=======================================      *
  44   688   779:============================================     *
  46  1012   793:================================================*=============
  48   589   759:====================================         *
  50   614   693:=========================================   *
  52   397   609:=========================           *
  54   460   520:============================  *
  56   542   435:=========================*======
  58   407   357:====================*===
  60   323   289:=================*==
  62   328   232:=============*======
  64   251   184:==========*=====
  66   125   146:========*
  68    73   115:===== *
  70    61    90:==== *
  72    50    70:=== *
  74    47    55:===*
  76    53    43:==*=
  78    26    33:=*
  80    21    26:=*
  82    13    20:=*
  84    11    16:*
  86    48    12:*==
  88     8     9:*          inset = represents 1 library sequences
  90     7     7:*
  92     6     6:*         :======*
  94     1     4:*         :=  *
  96     7     3:*         :==*====
  98     3     3:*         :==*
 100     1     2:*         :=*
 102     0     2:*         : *
 104     1     1:*         :*
 106     0     1:*         :*
 108     3     1:*         :*==
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     1     0:=         *=
 118     1     0:=         *=
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.18970.000534; mu= 5.6546 0.027
 mean_var=38.2201 8.354, 0's: 82 Z-trim: 83  B-trim: 193 in 1/61
 Lambda= 0.207457
 Kolmogorov-Smirnov  statistic: 0.0405 (N=29) at  54

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(8448)
gi|239523467|gb|EEQ63333.1| addiction module toxin (  90)   67 26.0     0.7
gi|63256641|gb|AAY37737.1| Ricin B lectin [Pseudom ( 803)   75 28.8     0.9


>>gi|239523467|gb|EEQ63333.1| addiction module toxin [He  (90 aa)
 initn:  31 init1:  31 opt:  67  Z-score: 118.7  bits: 26.0 E():  0.7
Smith-Waterman score: 67;  34.783% identity (54.348% similar) in 46 aa overlap (9-
53:41-83)

                                     10         20        30
5_3a                          GSLVVPIMPKHPLVGLKI-YKGRAECLLHHQYPNQDSQ
                                     :::   .::  : :  ::   :  :.
gi|239 KKEYKRAIRQGKQDKIDSLIEKLANDEALEPKHKDHALKGEYTGYREC---HIEPDLLLI
               20        30        40        50           60

         40        50        60
5_3a   DRETKDLRVLLSLRFGLWAKAPLS
        .. .:. ::. .:.:
gi|239 YKKQEDILVLVCFRLGSHSELFS
         70        80        90

>>gi|63256641|gb|AAY37737.1| Ricin B lectin [Pseudomonas  (803 aa)
 initn:  81 init1:  57 opt:  75  Z-score: 116.8  bits: 28.8 E():  0.9
Smith-Waterman score: 75;  36.842% identity (57.895% similar) in 38 aa overlap (3-
40:441-473)

                                          10        20        30
```

```
5_3a                              GSLVVPIMPKHPLVGLKIYKGRAECLLHHQYP
                                    :.:: .:..:.::      :   ::: .:
gi|632 VGDNEVLGFAGNTIQMVPESRSKAVTWNFELAVPKIPRRPIVG-----DRKVLLLHTRYS
              420       430       440       450            460

             40        50        60
5_3a   NQDSQDRETKDLRVLLSLRFGLWAKAPLS
       ..   : :
gi|632 DKPETDFEATREAFFGAPGDNTSFVNAVSFSSGDHLKFIGDAVTGLDLGPRPSGCPSTEL
         470       480       490       500       510       520



61 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:33:57 2010 done: Tue Jan 26 21:33:57 2010
 Total Scan time:  0.110 Total Display time:  0.000

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

```
# fasta34 5_3a.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_3a.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_3a, 61 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 340415     0:==============
  22   214     0:=          one = represents 26602 library sequences
  24   438    17:*
  26  1206   374:*
  28  4186  4039:*
  30 18343 24538:*
  32 76445 94879:===*
  34 236272 257301:==========*
  36 546707 528436:====================*=
  38 931287 873308:================================*===
  40 1306209 1218188:==============================================*====
  42 1543562 1489087:=========================================================*===
  44 1596080 1642602:=============================================================*
  46 1557039 1673031:=============================================================*
  48 1443217 1601734:=========================================================    *
  50 1312879 1461589:=================================================    *
  52 1173791 1284980:============================================   *
  54 1033793 1097599:=========================================  *
  56 876357 916831:================================== *
  58 764907 752701:============================*
  60 627478 609732:======================*=
  62 507872 488824:====================*=
  64 423963 388759:==============*=
  66 337697 307263:===========*=
  68 268126 241687:==========*=
  70 199419 189400:========*
  72 156507 147998:======*
  74 122669 115389:=====*
  76 98201 89810:===*
  78 77057 69808:==*
  80 55352 54205:==*
  82 44095 41466:=*
  84 31346 32846:=*
  86 24136 25414:*
  88 17847 19664:*          inset = represents 213 library sequences
  90 14969 15215:*
  92 10601 11773:*         :=========================================*
  94  8560  9109:*         :=========================================*
  96  5831  7048:*         :============================     *
  98  4467  5453:*         :====================    *
 100  3133  4220:*         :===============    *
 102  2371  3265:*         :============   *
 104  1639  2526:*         :========   *
 106  1190  1955:*         :======   *
 108   860  1512:*         :=====  *
 110   685  1170:*         :==== *
 112   458   905:*         :=== *
 114   362   701:*         :== *
 116   268   542:*         :==*
 118   170   419:*         :=*
>120   549   325:*         :=*=
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810750 sequences
  Expectation_n fit: rho(ln(x))= 4.26090.000183; mu= 5.7977 0.010
 mean_var=42.8297 8.545, 0's: 1112 Z-trim: 1112  B-trim: 0 in 0/64
 Lambda= 0.195975
 Kolmogorov-Smirnov  statistic: 0.0292 (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


61 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:33:58 2010 done: Tue Jan 26 21:38:41 2010
 Total Scan time: 244.640 Total Display time:  0.000

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

Appendix 16. Bioinformatic analysis of polypeptide 5_4a

```
>5_4a
DPLFHGLENL GLDIDDVEDI LLSPCRF
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_4a

Start time: Tue Jan 26 21:38:41 GMT 2010 Finish time: Tue Jan 26 21:38:42 GMT 2010

No 8 amino acid matches exist between 5_4a and the AD_2010 database
```

```
# fasta34 5_4a.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_4a.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_4a, 27 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     1     0:=           one = represents 3 library sequences
  24    10     0:====
  26    16     0:======
  28     0     0:
  30     4     2:*=
  32    11     8:==*=
  34    11    21:====  *
  36    26    44:=========     *
  38    44    72:===============        *
  40   108   101:==================================*==
  42    70   123:========================                *
  44   169   136:================================================*===========
  46   108   138:====================================          *
  48   146   132:============================================*=====
  50   145   121:=========================================*========
  52   135   106:====================================*=========
  54    74    91:=========================     *
  56    79    76:=========================*=
  58    62    62:====================*
  60    44    50:================ *
  62    55    40:=============*=====
  64    24    32:========  *
  66    25    25:========*
  68    26    20:======*==
  70    12    16:==== *
  72    10    12:===*
  74    10    10:===*
  76     7     7:==*
  78     7     6:=*=
  80    19     4:=*=====
  82     0     3:*
  84     4     3:*=
  86     2     2:*
  88     0     2:*           inset = represents 1 library sequences
  90     1     1:*
  92     0     1:*          :*
  94     2     1:*          :*=
  96     2     1:*          :*=
  98     0     0:           *
 100     0     0:           *
 102     0     0:           *
 104     0     0:           *
 106     0     0:           *
 108     0     0:           *
 110     0     0:           *
 112     0     0:           *
 114     0     0:           *
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.7975 0.003; mu= 6.3825 0.155
 mean_var=24.2195 6.393, 0's: 2 Z-trim: 2  B-trim: 82 in 2/41
 Lambda= 0.260610
 Kolmogorov-Smirnov  statistic: 0.0475 (N=27) at  42

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


27 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:38:41 2010 done: Tue Jan 26 21:38:41 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 5_4a.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_4a.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_4a, 27 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    61     0:=====
  22     0     0:            one = represents 13 library sequences
  24     1     0:=
  26     2     0:=
  28     6     2:*
  30    36    12:*==
  32    89    45:===*===
  34   129   122:=========*
  36   313   250:====================*=====
  38   340   414:============================    *
  40   493   577:======================================      *
  42   750   706:============================================================*===
  44   744   779:============================================================ *
  46   660   793:=============================================          *
  48   658   759:=========================================================       *
  50   673   693:============================================== *
  52   479   609:=====================================          *
  54   648   520:=========================================*==========
  56   616   435:================================*==============
  58   329   357:========================== *
  60   294   289:======================*
  62   283   232:=================*====
  64   182   184:==============*
  66   172   146:===========*==
  68    83   115:======= *
  70    62    90:===== *
  72    66    70:=====*
  74    63    55:====*
  76    47    43:===*
  78    28    33:==*
  80    16    26:=*
  82    22    20:=*
  84    39    16:=*=
  86    15    12:*=
  88    15     9:*=          inset = represents 1 library sequences
  90    14     7:*=
  92     2     6:*          :==   *
  94     1     4:*          :=  *
  96     4     3:*          :==*=
  98     5     3:*          :==*==
```

```
 100     1     2:*         :=*
 102     0     2:*         : *
 104     2     1:*         :*=
 106     0     1:*         :*
 108     0     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 3.56470.000505; mu= 7.4322 0.025
 mean_var=20.8799 4.359, 0's: 60 Z-trim: 60  B-trim: 130 in 1/61
 Lambda= 0.280679
 Kolmogorov-Smirnov  statistic: 0.0464 (N=29) at  52

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


27 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:38:42 2010 done: Tue Jan 26 21:38:42 2010
 Total Scan time:  0.160 Total Display time:  0.000

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

```
# fasta34 5_4a.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_4a.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_4a, 27 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 277886     0:===========
  22   609     0:=          one = represents 27380 library sequences
  24  1278    17:*
  26  3181   374:*
  28 10798  4039:*
  30 37080 24538:*=
  32 101985 94879:===*
  34 243421 257301:==========*
  36 478781 528435:================== *
  38 788471 873306:============================  *
  40 1106521 1218185:=========================================   *
  42 1402737 1489083:================================================  *
  44 1579438 1642597:============================================================ *
  46 1642766 1673026:============================================================*
  48 1620025 1601730:============================================================*=
  50 1487799 1461585:=================================================*=
  52 1274760 1284977:=============================================*
  54 1110195 1097596:=========================================*
  56 925990 916829:==============================*
  58 748619 752699:==========================*
  60 602938 609730:======================*
  62 484172 488823:=================*
  64 396719 388758:==============*
  66 323790 307263:===========*
  68 257199 241686:========*=
  70 209962 189400:======*=
  72 169908 147998:=====*=
  74 126914 115389:====*
  76 96755 89810:===*
  78 73927 69808:==*
  80 57445 54205:=*=
  82 42251 41465:=*
  84 33636 32846:=*
  86 22653 25414:*
  88 17450 19664:*          inset = represents 227 library sequences
  90 12997 15215:*
  92 11337 11773:*         :=======================================*
  94  8382  9109:*         :=====================================  *
  96  5465  7048:*         :=========================      *
  98  3901  5453:*         :==================      *
 100  2882  4220:*         :=============     *
 102  2063  3265:*         :==========    *
 104  1566  2526:*         :=======    *
 106  1287  1955:*         :======  *
 108   902  1512:*         :====  *
 110   584  1170:*         :===  *
 112   474   905:*         :===*
 114   325   701:*         :== *
 116   285   542:*         :==*
 118   177   419:*         :=*
>120   539   325:*         :=*=
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810703 sequences
  Expectation_n fit: rho(ln(x))= 4.25660.000171; mu= 4.9545 0.009
 mean_var=28.7628 5.716, 0's: 888 Z-trim: 893  B-trim: 0 in 0/63
 Lambda= 0.239143
 Kolmogorov-Smirnov  statistic: 0.0233 (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


27 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:38:42 2010 done: Tue Jan 26 21:45:14 2010
 Total Scan time: 345.030 Total Display time:  0.000

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

Appendix 17. Bioinformatic analysis of polypeptide 5_5a

```
>5_5a
qhpeILCFTV LRILVWILMM
```

```
Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_5a

Start time: Tue Jan 26 21:45:15 GMT 2010 Finish time: Tue Jan 26 21:45:15 GMT 2010
```

```
No 8 amino acid matches exist between 5_5a and the AD_2010 database
```

```
# fasta34 5_5a.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_5a.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_5a, 20 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     1     0:=
  28     2     0:=
  30     6     2:*=
  32     7     8:==*
  34    13    21:===== *
  36    47    44:==========*=
  38    75    72:======================*=
  40    55   101:====================              *
  42    58   123:====================                    *
  44    96   136:==================================              *
  46   130   138:=========================================== *
  48   155   132:============================================*========
  50   169   121:=========================================*================
  52   111   106:====================================*=
  54    82    91:============================  *
  56    99    76:=========================*=======
  58    77    62:====================*=====
  60    61    50:=================*====
  62    26    40:=========    *
  64    50    32:==========*======
  66    48    25:========*=======
  68    28    20:======*===
  70    13    16:=====*
  72    10    12:===*
  74     9    10:===*
  76     7     7:==*
  78    19     6:=*=====
  80     0     4: *
  82     0     3:*
  84     4     3:*=
  86     0     2:*
  88     3     2:*          inset = represents 1 library sequences
  90     0     1:*
  92     0     1:*         :*
  94     0     1:*         :*
  96     4     1:*=        :*===
  98     4     0:==        *====
 100     0     0:          *
 102     0     0:          *
 104     0     0:          *
 106     0     0:          *
 108     0     0:          *
 110     0     0:          *
 112     0     0:          *
 114     0     0:          *
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.24930.00246; mu= 4.2469 0.129
 mean_var=19.3617 5.160, 0's: 2 Z-trim: 2  B-trim: 21 in 1/42
 Lambda= 0.291476
 Kolmogorov-Smirnov  statistic: 0.1053 (N=26) 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


20 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:45:14 2010 done: Tue Jan 26 21:45:14 2010
 Total Scan time:  0.020 Total Display time:  0.000

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

```
# fasta34 5_5a.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_5a.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_5a, 20 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    63     0:=====
  22     3     0:=          one = represents 14 library sequences
  24     3     0:=
  26     5     0:=
  28    35     2:*==
  30    50    12:*===
  32    92    45:===*===
  34   228   122:========*========
  36   267   250:=================*==
  38   276   414:====================         *
  40   654   577:=========================================*=====
  42   592   706:============================================       *
  44   823   779:============================================================*===
  46   668   793:================================================        *
  48   639   759:============================================         *
  50   529   693:=======================================           *
  52   576   609:========================================== *
  54   552   520:======================================*==
  56   578   435:==============================*==========
  58   414   357:=========================*=====
  60   318   289:====================*==
  62   214   232:================*
  64   255   184:=============*=====
  66   160   146:==========*=
  68    74   115:======  *
  70    39    90:===   *
  72    66    70:====*
  74    58    55:===*=
  76    69    43:===*=
  78    28    33:==*
  80     8    26:=*
  82    24    20:=*
```

```
  84    16    16:=*
  86    13    12:*
  88    25     9:*=         inset = represents 1 library sequences
  90     1     7:*
  92     1     6:*         :=    *
  94     2     4:*         :== *
  96     2     3:*         :==*
  98     1     3:*         := *
 100     4     2:*         :=*==
 102     1     2:*         :=*
 104     1     1:*         :*
 106    10     1:*         :*=========
 108     5     1:*         :*====
 110     0     1:*         :*
 112     0     0:          *
 114     1     0:=         *=
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 1.30110.000473; mu= 15.9113 0.024
 mean_var=19.0208 3.940, 0's: 60 Z-trim: 62  B-trim: 100 in 1/61
 Lambda= 0.294077
 Kolmogorov-Smirnov  statistic: 0.0384 (N=29) at  52

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


20 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:45:15 2010 done: Tue Jan 26 21:45:15 2010
 Total Scan time:  0.120 Total Display time:  0.000

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

```
# fasta34 5_5a.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_5a.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_5a, 20 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 278809     0:===========
  22   843     0:=          one = represents 27172 library sequences
  24  1718    17:*
  26  4556   374:*
  28 13418  4039:*
  30 40325 24538:*=
  32 106877 94880:===*
  34 238766 257304:==========*
  36 445447 528443:=================  *
  38 734988 873319:============================    *
  40 1087783 1218203:==========================================   *
  42 1377297 1489105:==================================================   *
  44 1564157 1642622:============================================================ *
  46 1630308 1673051:============================================================*
  48 1606183 1601754:============================================================*=
  50 1488059 1461607:==========================================================*=
  52 1315706 1284996:======================================================*=
  54 1135693 1097612:==================================================*=
  56 963427 916842:================================*==
  58 799606 752710:============================*==
  60 654695 609740:======================*==
  62 517442 488830:==================*==
  64 400646 388764:===============*
  66 306630 307267:===========*
  68 249250 241690:========*=
  70 194333 189402:======*=
  72 150645 148000:=====*
  74 116033 115391:====*
  76 90975 89811:===*
  78 74840 69809:==*
  80 52076 54206:=*
  82 39443 41466:=*
  84 28284 32846:=*
  86 29850 25415:*=
  88 16709 19665:*          inset = represents 203 library sequences
  90 11881 15215:*
  92 10105 11773:*         :=======================================*
  94  7212  9109:*         :====================================   *
  96  4966  7048:*         :=========================          *
  98  3822  5454:*         :===================       *
 100  3107  4220:*         :================    *
 102  1887  3265:*         :==========      *
 104  3184  2526:*         :============*===
 106  4483  1955:*         :==========*=============
 108  2248  1512:*         :========*====
 110  1010  1170:*         :======*
 112   603   905:*         :=== *
 114   267   701:*         :== *
 116   141   542:*         := *
 118   151   419:*         := *
>120   341   325:*         :=*
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810970 sequences
  Expectation_n fit: rho(ln(x))= 3.19950.000173; mu= 6.9222 0.009
 mean_var=22.8350 4.529, 0's: 955 Z-trim: 959  B-trim: 1777 in 2/62
 Lambda= 0.268395
 Kolmogorov-Smirnov  statistic: 0.0319 (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


20 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:45:15 2010 done: Tue Jan 26 21:50:33 2010
 Total Scan time: 276.250 Total Display time:  0.000

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

Appendix 18. Bioinformatic analysis of polypeptide 5_6a

```
>5_6a
rvlqlsyffq dngalaqrpn rsdnntlrSF VSRS
```

```
_____________________________________________________

Sliding 8 amino acid window search
Database searched = AD_2010
Query = 5_6a

Start time: Tue Jan 26 21:50:33 GMT 2010 Finish time: Tue Jan 26 21:50:33 GMT 2010

No 8 amino acid matches exist between 5_6a and the AD_2010 database
_____________________________________________________

# fasta34 5_6a.pep /genedata/1/db/AD_2010 -Q -E 1 -O 5_6a.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_6a, 34 aa
 vs /genedata/1/db/AD_2010 library

       opt      E()
< 20     2     0:=
  22     0     0:           one = represents 3 library sequences
  24     2     0:=
  26     0     0:
  28     3     0:=
  30     4     2:*=
  32    17     8:==*===
  34    43    21:======*========
  36    39    44:============= *
  38    78    72:=======================*==
  40    58   101:====================             *
  42   124   123:=========================================*=
  44   103   136:====================================          *
  46   135   138:=============================================*
  48   164   132:=========================================*==========
  50   139   121:=========================================*======
  52   105   106:====================================*
  54   109    91:==============================*======
  56    76    76:========================*
  58    48    62:================    *
  60    37    50:=============   *
  62    40    40:=============*
  64    23    32:========  *
  66    40    25:========*=====
  68    21    20:======*
  70    12    16:==== *
  72     9    12:===*
  74     5    10:== *
  76     5     7:==*
  78     6     6:=*
  80     5     4:=*
  82     2     3:*
  84     0     3:*
  86     1     2:*
  88     0     2:*          inset = represents 1 library sequences
  90     2     1:*
  92     7     1:*==        :*======
  94     5     1:*=         :*====
  96     0     1:*          :*
  98     0     0:           *
 100     0     0:           *
 102     1     0:=          *=
 104     0     0:           *
 106     0     0:           *
 108     0     0:           *
 110     1     0:=          *=
 112     0     0:           *
 114     0     0:           *
 116     0     0:           *
 118     0     0:           *
>120     0     0:           *
 331323 residues in  1471 sequences
  Expectation_n fit: rho(ln(x))= 3.66260.00266; mu= 6.9013 0.139
 mean_var=22.5879 7.025, 0's: 2 Z-trim: 3  B-trim: 50 in 1/42
 Lambda= 0.269859
 Kolmogorov-Smirnov  statistic: 0.0264 (N=27) at  38

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 1
 join: 42, opt: 30, open/ext: -10/-2, width:  16
The best scores are:                                      opt bits E(1471)
gi|729979|sp|P39673.1|MAG_DERFA RecName: Full=Alle ( 341)   57 25.5    0.35
gi|158121995|gb|ABW17159.1| Ara h 7 allergen precu ( 164)   51 23.2    0.87


>>gi|729979|sp|P39673.1|MAG_DERFA RecName: Full=Allergen  (341 aa)
 initn:  55 init1:  55 opt:  57  Z-score: 110.5  bits: 25.5 E(): 0.35
Smith-Waterman score: 57;  32.258% identity (70.968% similar) in 31 aa overlap (3-
33:169-198)

                                        10        20        30
5_6a                                 RVLQLSYFFQDNGALAQRPNRSDNNTLRSFVS
                                      .. .: ..: : :  .: .:.:. :.. .:
gi|729 DIDGTLIEGHAQGTIREGKIHIKGRQTDFEIESNYRYED-GKLIIEPVKSENGKLEGVLS
      140       150       160       170       180       190


5_6a   RS
       :
gi|729 RKVPSHLTLETPRVKMNMKYDRYAPVKVFKLDYDGIHFEKHTDIEYEPGVRYKIIGNGKL
       200       210       220       230       240       250

>>gi|158121995|gb|ABW17159.1| Ara h 7 allergen precursor  (164 aa)
 initn:  51 init1:  51 opt:  51  Z-score: 103.5  bits: 23.2 E(): 0.87
Smith-Waterman score: 51;  36.364% identity (72.727% similar) in 22 aa overlap (4-
25:72-93)

                                       10        20        30
5_6a                                RVLQLSYFFQDNGALAQRPNRSDNNTLRSFVSR
                                       :  : . . :. .:::..::..
gi|158 DQCQRQLQRANLRPCEEHIRQRVEKEQEQEQDEYPYIQRGSRGQRPGESDEDQEQRCCNE
            50        60        70        80        90       100


5_6a   S

gi|158 LNRFQNNQRCMCQALQQILQNQSFRFQQDRSQLHQMERELRNLPQNCGFRSPSRCDLSSR
           110       120       130       140       150       160




34 residues in 1 query   sequences
331323 residues in 1471 library sequences
 Scomplib [34t26]
```

```
 start: Tue Jan 26 21:50:33 2010 done: Tue Jan 26 21:50:33 2010
 Total Scan time:  0.030 Total Display time:  0.000

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

```
# fasta34 5_6a.pep /genedata/1/db/TOX_2010 -Q -E 1 -O 5_6a.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_6a, 34 aa
 vs /genedata/1/db/TOX_2010 library

       opt      E()
< 20    60     0:====
  22     3     0:=          one = represents 17 library sequences
  24     2     0:=
  26     6     0:=
  28    23     2:*=
  30    43    12:*==
  32    59    45:==*=
  34   162   122:=======*==
  36   182   250:===========   *
  38   363   414:=====================  *
  40   416   577:=========================        *
  42   635   706:======================================   *
  44   638   779:======================================       *
  46   770   793:================================================*
  48   813   759:=============================================*===
  50   691   693:==========================================*
  52  1011   609:===================================*=======================
  54   642   520:==============================*=======
  56   351   435:====================    *
  58   259   357:================    *
  60   190   289:============    *
  62   173   232:===========  *
  64   173   184:==========*
  66   155   146:========*=
  68   101   115:======*
  70    65    90:==== *
  72    55    70:====*
  74    90    55:===*==
  76    42    43:==*
  78    54    33:=*==
  80    23    26:=*
  82    76    20:=*===
  84    17    16:*
  86    27    12:*=
  88    10     9:*          inset = represents 1 library sequences
  90    14     7:*
  92     5     6:*         :=====*
  94     3     4:*         :===*
  96     3     3:*         :==*
  98     7     3:*         :==*====
 100    21     2:*=        :=*====================
 102     5     2:*         :=*===
 104     0     1:*         :*
 106     3     1:*         :*==
 108     1     1:*         :*
 110     0     1:*         :*
 112     0     0:          *
 114     1     0:=         *=
 116     0     0:          *
 118     0     0:          *
>120     0     0:          *
2069351 residues in  8448 sequences
  Expectation_n fit: rho(ln(x))= 4.72170.000562; mu= 1.8456 0.028
 mean_var=19.1764 4.026, 0's: 60 Z-trim: 60  B-trim: 478 in 2/60
 Lambda= 0.292881
 Kolmogorov-Smirnov  statistic: 0.0475 (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


34 residues in 1 query   sequences
2069351 residues in 8448 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:50:33 2010 done: Tue Jan 26 21:50:34 2010
 Total Scan time:  0.190 Total Display time:  0.000

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

```
# fasta34 5_6a.pep /genedata/1/db/PRT_2010 -Q -E 1 -O 5_6a.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_6a, 34 aa
 vs /genedata/1/db/PRT_2010 library

       opt      E()
< 20 279849     0:===========
  22   558     0:=          one = represents 27321 library sequences
  24   957    17:*
  26  2523   374:*
  28  8572  4039:*
  30 31819 24538:*=
  32 104504 94879:===*
  34 251637 257301:=========*
  36 469507 528436:================== *
  38 763470 873309:============================   *
  40 1095975 1218189:=========================================   *
  42 1348977 1489087:=================================================    *
  44 1532024 1642602:============================================================  *
  46 1639249 1673031:==============================================================*
  48 1584812 1601734:=============================================================*
  50 1473594 1461589:============================================================*
  52 1298680 1284981:================================================*
  54 1104814 1097599:=========================================*
  56 920804 916831:==================================*
  58 748141 752701:==========================*
  60 622742 609732:======================*
  62 505069 488824:=================*=
  64 424266 388759:==============*=
  66 331614 307264:===========*=
  68 284436 241687:========*==
  70 214439 189400:======*=
  72 177579 147998:=====*=
  74 138802 115389:====*=
  76 104559 89810:===*
```

```
  78 81605 69808:==*
  80 65482 54205:=*=
  82 54474 41466:=*
  84 34868 32846:=*
  86 26907 25414:*
  88 20655 19664:*          inset = represents 241 library sequences
  90 15950 15215:*
  92 12027 11773:*         :=======================================*
  94  9100  9109:*         :======================================*
  96  6812  7048:*         :============================*
  98  4845  5453:*         :====================  *
 100  3605  4220:*         :===============  *
 102  2801  3265:*         :============ *
 104  2094  2526:*         :========== *
 106  1679  1955:*         :======= *
 108  1089  1512:*         :===== *
 110   941  1170:*         :====*
 112   641   905:*         :===*
 114   451   701:*         :==*
 116   384   542:*         :==*
 118   299   419:*         :=*
>120   544   325:*         :=*=
4761287459 residues in 17815538 sequences
 statistics sampled from 60000 to 17810755 sequences
  Expectation_n fit: rho(ln(x))= 4.09510.000171; mu= 5.4676 0.009
 mean_var=25.6998 5.090, 0's: 938 Z-trim: 940  B-trim: 0 in 0/63
 Lambda= 0.252993
 Kolmogorov-Smirnov  statistic: 0.0327 (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


34 residues in 1 query   sequences
4761287459 residues in 17815538 library sequences
 Scomplib [34t26]
 start: Tue Jan 26 21:50:34 2010 done: Tue Jan 26 21:57:50 2010
 Total Scan time: 378.620 Total Display time:  0.000

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

```
Database checksum values:

Tue Jan 26 21:57:51 GMT 2010      a184245745a6ed8c6ecde45b26637bba
/genedata/1/db/AD_2010

Tue Jan 26 21:57:51 GMT 2010      17c3a19148dfb0163e270cf41e2aa437
/genedata/1/db/TOX_2010

Tue Jan 26 21:58:48 GMT 2010      e657d3127c1aad11f9f7df8dcc5e448c
/genedata/1/db/PRT_2010
```