RepeatMatcher

http://repeatmasker.org/

INSTITUTE FOR
Systems
Biology

User Manual
version 1.0

Juan Caballero
jcaballero@systemsbiology.org
2012

What is RepeatMatcher?

RepeatMatcher is a simple analysis tool to help in the manual annotation of RepeatModeler sequences.

Motivation

After running RepeatModeler you probably will have a well identified sequences and a bunch of sequences to be manually annotated. RepeatMatcher is designed to help in this hard and repetitive task. We provide a simple pipeline to generate all the required files and a graphical user interface to manipulate and annotate the sequences. The GUI is intended to show all the information to make the decisions quick and easy for each sequence.

The pipeline

Installation

The pipeline is a Perl script (RepatMatcher.pl) that control all the processes, to be able to run the pipeline, first you need to resolve all the dependencies:

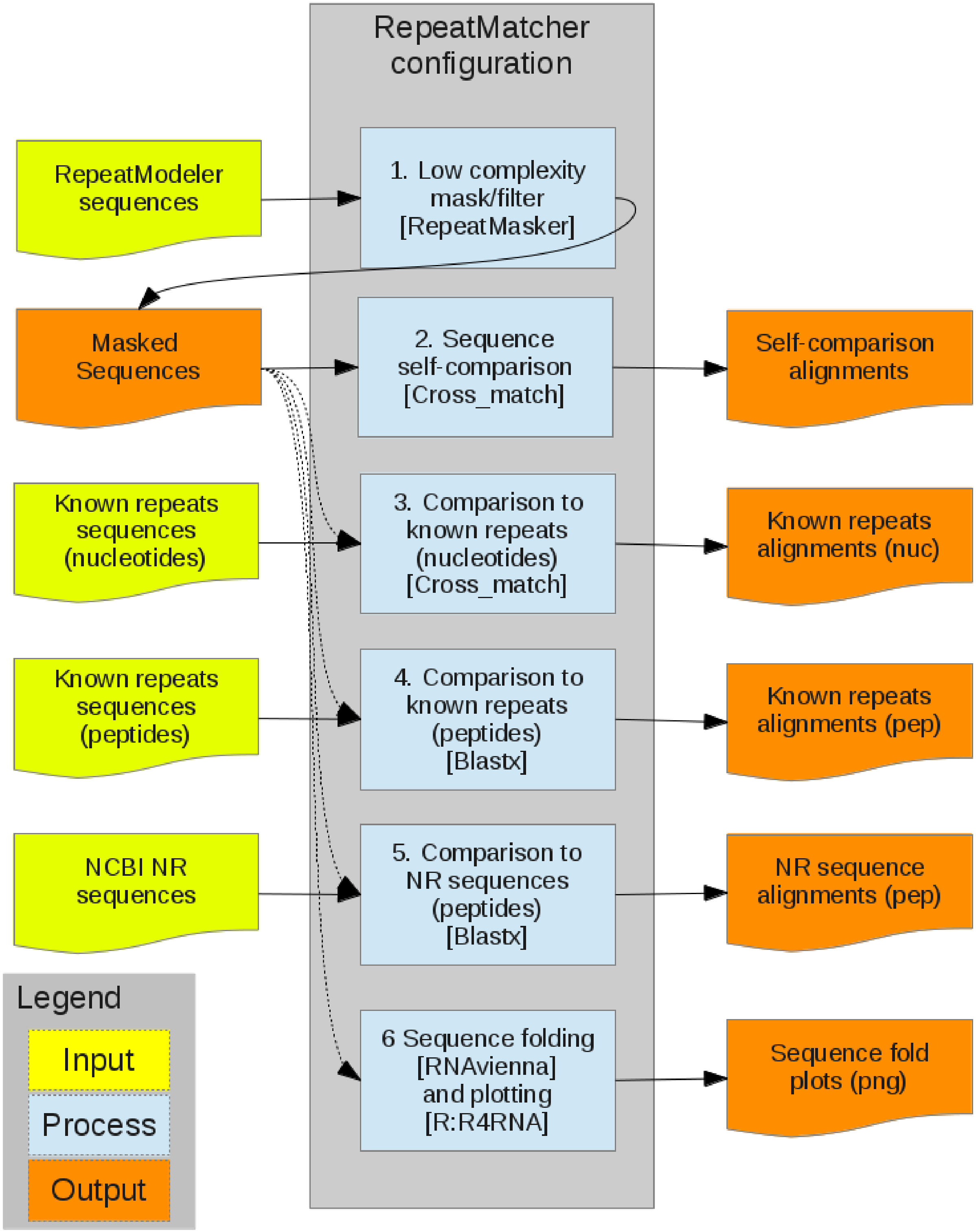
- RepeatMasker
- Blast
- Cross_match
- RNAVienna
- R
- R package R4RNA
- Databases: NCBI NR, known repeats (in nucleotides and proteins).

Process description

We started with the RepeatModeler consensi output file (fasta), the steps are:

1. Masking all the low complexity regions with RepeatMasker and discard sequences with filters for a minimal composition.
2. All sequences will be compared with itself using Cross_match.
3. All sequences are compared in nucleotide space with known repeats consensi with Cross_match.
4. All sequences are compared in protein space with known repeats consensi with Blastx.
5. All sequences are compared in protein space with the sequences in NCBI-NR data base with Blastx.
6. All sequences are folded with RNA-Vienna and the principal structure is plotted with R:R4RNA.

RepeatMatcher Pipeline


RepeatMatcher configuration
RepeatModeler sequences
1. Low complexity mask/filter [RepeatMasker]
Masked Sequences
2. Sequence self-comparison [Cross_match]
Self-comparison alignments
Known repeats sequences (nucleotides)
3. Comparison to known repeats (nucleotides) [Cross_match]
Known repeats alignments (nuc)
Known repeats sequences (peptides)
4. Comparison to known repeats (peptides) [Blastx]
Known repeats alignments (pep)
NCBI NR sequences
5. Comparison to NR sequences (peptides) [Blastx]
NR sequence alignments (pep)
6 Sequence folding [RNAvienna] and plotting [R:R4RNA]
Sequence fold plots (png)
Legend
Input
Process
Output

Configuration file

RepeatMatcher uses a plain text file (`RepeatMatcher.conf`) to define some parameters, but if it's required, you can adjust the parameters before running the pipeline editing this file or activating the `-e` option.

The follow is an example of this file:

```
# RepeatMatcher configuration file
# Juan Caballero, Institute for Systems Biology @ 2012
# This is a simple configuration file, syntax is:
#          VARIABLE: VALUE
# One variable per line, blank lines are omitted.

# STEP 1. Masking low complexity sequences
# min_mask => minimal % of masked sequence
# min_size => minimal size after masking
min_mask: 90
min_size: 30

# STEP 2. Self-comparison
# crossmatch_self => cross_match parameters
crossmatch_self: -M data/nt_sub.matrix -gap_init -25 -gap_ext -5 -minscore 200 -
minmatch 9 -minscore 200

# STEP 3. Known annotation comparison
# crossmatch_comp => cross_match parameters
crossmatch_comp: -M data/nt_sub.matrix -gap_init -25 -gap_ext -5 -minscore 200 -
minmatch 9 -minscore 200

# STEP 4. Blastx to known repeat proteins
# blastx_rep => blast parameters
blastx_rep: -d data/repeats -W 2 -v 5 -b 5 -F F -e 0.01

# STEP 5. Blastx to NR database
# blastx_nr => blast parameters
blastx_nr: -d data/nr -W 2 -v 5 -b 5 -F F -e 1e-6

# STEP 6. Fold DNA
# fold => folding (RNAVienna) parameters
fold: --noconv --noGU --ImFeelingLucky
plot_w = 600
plot_h = 300
```

Launching the pipeline

The basic operation is:

```
perl RepeatMatcher.pl  -s sequences.fa -k knownrepeats.fa -o NewAnnotation
```

for a complete list of options:

```
perl RepeatMatcher.pl --help
```

THE GUI

Installation

The program is a Perl:Tk development, so you need to install it from CPAN.

After you ran the pipeline, you have all the files required for `RepeatMatcherGUI.pl`. The GUI uses a text file to keep all the parameters and annotations made (the LOG file). The first lines of that file record the files used, the after that, the rest of lines records all changes for each sequence.

Launching the GUI

```
perl RepeatMatcherGUI.pl -o OUT -i SEQS -s SELF -a ALIGN -b BLAST -n BLAST \
     -f FOLD -l LOG
```

Launching the GUI with a previous generated LOG:

```
perl RepeatMatcherGUI.pl -r LOG
```

for a complete list of options:

```
perl RepeatMatcherGUI.pl --help
```

Description of the GUI

The program will call a series of windows, this are:

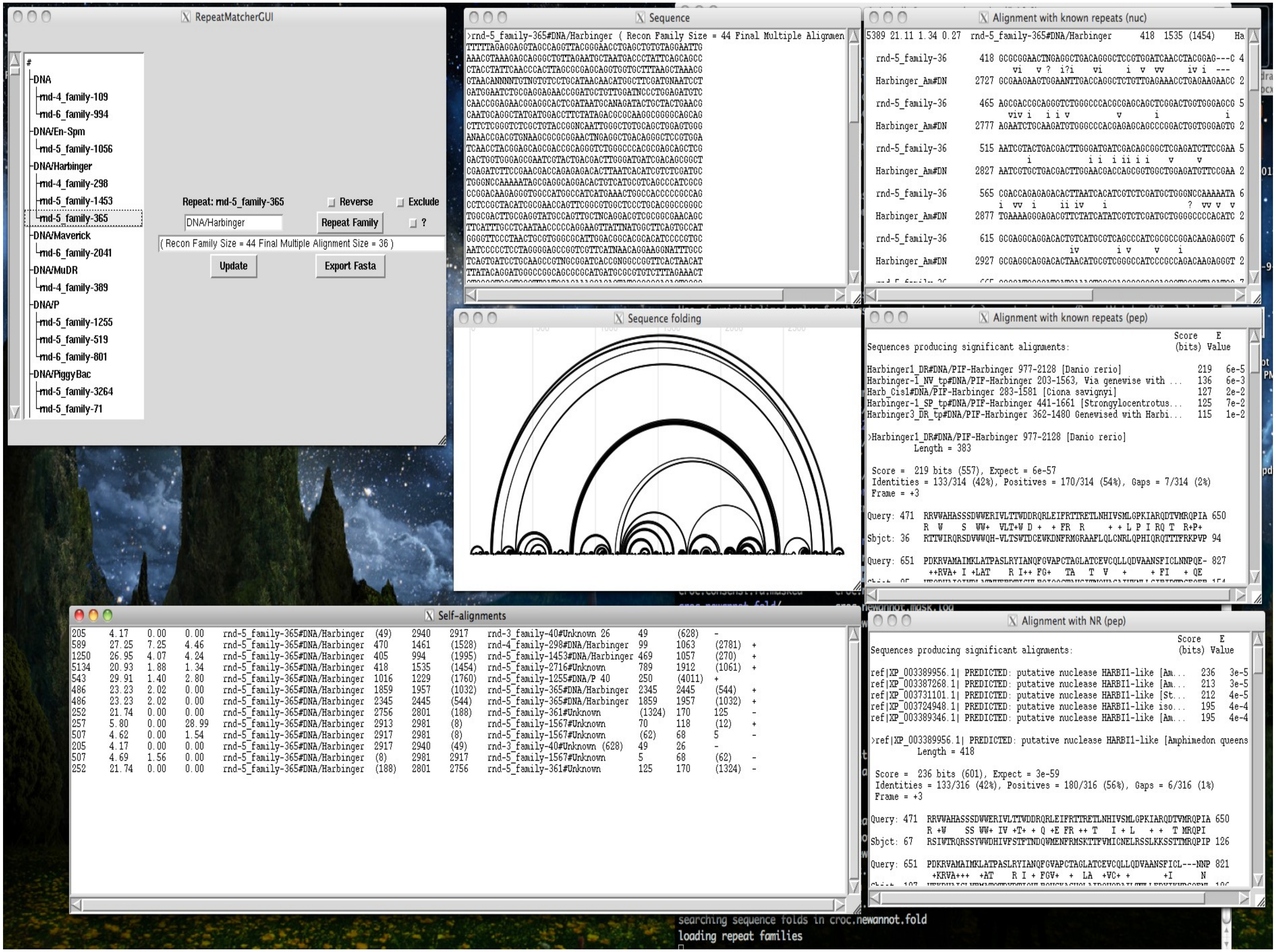
1. Annotation window
2. Sequence window
3. Self-alignments window
4. Alignments to known repeats (nucleotides) window
5. Alignments to known repeats (protein) window
6. Alignments to NR proteins
7. Sequence folding window

The annotation window is used to check each repeat and corroborate or edit the annotation, the sequence also can be marked as “Excluded”, or can be flagged to be reverse complement with the “Reverse” mark. Each time the “Update” button is pushed, the LOG file is updated too, to close the session just close all windows.

The Annotation window also shows the status of the sequence using a color code in the labels:

- black – sequence is not saved or modified yet
- blue – sequence is saved without modifications
- red – sequence is saved with modifications

When the annotation is done, you can write the final sequences in a new fasta file pushing the “Export fasta” button.



Repeat: rnd-5_family-365

DNA/Harbinger

(Recon Family Size = 44 Final Multiple Alignment Size = 36)

```
>rnd-5_family-365#DNA/Harbinger ( Recon Family Size = 44 Final Multiple Alignmen
TTTTTAGAGGAGGTAGCCAGGTTACGGGAACCTGAGCTGTGTAGGAATTG
AAACGTAAAGAGCAGGGCTGTTAGAATGCTAATGACCCTATTCAGCAGCC
CTACCTATTCAACCCACTTAGCGCGAGCAGGTGGTGCTTTAAGCTAAACG
GTAACANNNNTGTNGTGTCCTGCATAACAACATGGCTTCGATGNAATCCT
GATGGAATCTGCGAGGAGAACCGGATGCTGTTGGATNCCCTGGAGATGTC
CAACCGGAGAACGGAGGCACTCGATAATGCANAGATACTGCTACTGAACG
CAATGCAGGCTATGATGGACCTTCTATAGACGCGCAAGGCGGGGCAGCAG
CTTCTCGGGTCTCGCTGTACCGGNCAATTGGGCTGTGCAGCTGGAGTGGG
ANAACCGACGTGNAAGCGCGCGGAACTNGAGGCTGACAGGGCTCCGTGGA
TCAACCTACGGAGCAGCGACCGCAGGGTCTGGGCCCACGCGAGCAGCTCG
GACTGGTGGGAGCGAATCGTACTGACGACTTGGGATGATCGACAGCGGCT
CGAGATCTTCCGAACGACCAGAGAGACACTTAATCACATCGTCTCGATGC
TGGGNCCAAAAATAGCGAGGCAGGACACTGTCATGCGTCAGCCCATCGCG
CCGGACAAGAGGGTGGCCATGGCCATCATGAAACTGGCCACCCCCGCCAG
CCTCCGCTACATCGCGAACCAGTTCGGCGTGGCTCCCTGCACGGCCGGGC
TGGCGACTTGCGAGGTATGCCAGTTGCTNCAGGACGTCGCGGCGAACAGC
TTCATTTGCCTCAATAACCCCCAGGAAGTTATTNATGGCTTCAGTGCCAT
GGGGTTCCCTAACTGCGTGGGCGCATTGGACGGCACGCACATCCCCGTGC
AATCCCCCTCCTAGGGGAGCCGGTCGTTCATNAACAGGAAGGNATTTGCC
TCAGTGATCCTGCAAGCCGTNGCGGATCACCGNGGCCGGTTCACTAACAT
TTATACAGGATGGGCCGGCAGCGCGCATGATGCGCGTGTCTTTAGAAACT
```

```
5389 21.11 1.34 0.27  rnd-5_family-365#DNA/Harbinger      418  1535 (1454)     Ha

  rnd-5_family-36        418 GCGCGGAACTNGAGGCTGACAGGGCTCCGTGGATCAACCTACGGAG---C 4
                                vi   v ?  i?i    vi      i  v  vv      iv i  ---
  Harbinger_Am#DN       2727 GCGAAGAAGTGGAANTTGACCAGGCTCTGTTGAGAAACCTGAGAAGAACC 2

  rnd-5_family-36        465 AGCGACCGCAGGGTCTGGGCCCACGCGAGCAGCTCGGACTGGTGGGAGCG 5
                               viv i   i i v           v       i              i
  Harbinger_Am#DN       2777 AGAATCTGCAAGATGTGGGCCCACGAGAGCAGCCCGGACTGGTGGGAGTG 2

  rnd-5_family-36        515 AATCGTACTGACGACTTGGGATGATCGACAGCGGCTCGAGATCTTCCGAA 5
                                  i            i i  i ii i i     v     v
  Harbinger_Am#DN       2827 AATCGTGCTGACGACTTGGAACGACCAGCGGTGGCTGGAGATGTTCCGAA 2

  rnd-5_family-36        565 CGACCAGAGAGACACTTAATCACATCGTCTCGATGCTGGGNCCAAAAATA 6
                             i  vv  i      ii iv    i                 ?  vv v  v
  Harbinger_Am#DN       2877 TGAAAAGGGAGACGTTCTATCATATCGTCTCGATGCTGGGGCCCCACATC 2

  rnd-5_family-36        615 GCGAGGCAGGACACTGTCATGCGTCAGCCCATCGCGCCGGACAAGAGGGT 6
                                                iv         i v     v    i
  Harbinger_Am#DN       2927 GCGAGGCAGGACACTAACATGCGTCGGGCCATCCCGCCAGACAAGAGGGT 2
```

```
                                                                   Score    E
Sequences producing significant alignments:                        (bits) Value

Harbinger1_DR#DNA/PIF-Harbinger 977-2128 [Danio rerio]                219   6e-5
Harbinger-1_NV_tp#DNA/PIF-Harbinger 203-1563, Via genewise with ...   136   6e-3
Harb_Cis1#DNA/PIF-Harbinger 283-1581 [Ciona savignyi]                 127   2e-2
Harbinger-1_SP_tp#DNA/PIF-Harbinger 441-1661 [Strongylocentrotus...   125   7e-2
Harbinger3_DR_tp#DNA/PIF-Harbinger 362-1480 Genewised with Harbi...   115   1e-2

>Harbinger1_DR#DNA/PIF-Harbinger 977-2128 [Danio rerio]
          Length = 383

 Score =  219 bits (557), Expect = 6e-57
 Identities = 133/314 (42%), Positives = 170/314 (54%), Gaps = 7/314 (2%)
 Frame = +3

Query: 471   RRVWAHASSSDWWERIVLTTWDDRQRLEIFRTTRETLNHIVSMLGPKIARQDTVMRQPIA 650
              R  W    S   WW+  VLT+W D +  + FR  R      + + L P I RQ T  R+P+
Sbjct: 36    RTTWIRQRSDVWWQH-VLTSWTDCEWKDNFRMGRAAFLQLCNRLQPHIQRQTTTFRKPVP 94

Query: 651   PDKRVAMAIMKLATPASLRYIANQFGVAPCTAGLATCEVCQLLQDVAANSFICLNNPQE- 827
               ++RVA+ I +LAT    R I++ FG+   TA   T  V   +      + FI   + QE
```

```
205     4.17    0.00    0.00    rnd-5_family-365#DNA/Harbinger  (49)    2940    2917    rnd-3_family-40#Unknown 26      49      (628)   -
589     27.25   7.25    4.46    rnd-5_family-365#DNA/Harbinger  470     1461    (1528)  rnd-4_family-298#DNA/Harbinger  99      1063    (2781)  +
1250    26.95   4.07    4.24    rnd-5_family-365#DNA/Harbinger  405     994     (1995)  rnd-5_family-1453#DNA/Harbinger 469     1057    (270)   +
5134    20.93   1.88    1.34    rnd-5_family-365#DNA/Harbinger  418     1535    (1454)  rnd-5_family-2716#Unknown       789     1912    (1061)  +
543     29.91   1.40    2.80    rnd-5_family-365#DNA/Harbinger  1016    1229    (1760)  rnd-5_family-1255#DNA/P 40      250     (4011)  +
486     23.23   2.02    0.00    rnd-5_family-365#DNA/Harbinger  1859    1957    (1032)  rnd-5_family-365#DNA/Harbinger  2345    2445    (544)   +
486     23.23   2.02    0.00    rnd-5_family-365#DNA/Harbinger  2345    2445    (544)   rnd-5_family-365#DNA/Harbinger  1859    1957    (1032)  +
252     21.74   0.00    0.00    rnd-5_family-365#DNA/Harbinger  2756    2801    (188)   rnd-5_family-361#Unknown        (1324)  170     125     -
257     5.80    0.00    28.99   rnd-5_family-365#DNA/Harbinger  2913    2981    (8)     rnd-5_family-1567#Unknown       70      118     (12)    +
507     4.62    0.00    1.54    rnd-5_family-365#DNA/Harbinger  2917    2981    (8)     rnd-5_family-1567#Unknown       (62)    68      5       -
205     4.17    0.00    0.00    rnd-5_family-365#DNA/Harbinger  2917    2940    (49)    rnd-3_family-40#Unknown (628)   49      26      -
507     4.69    1.56    0.00    rnd-5_family-365#DNA/Harbinger  (8)     2981    2917    rnd-5_family-1567#Unknown       5       68      (62)    -
252     21.74   0.00    0.00    rnd-5_family-365#DNA/Harbinger  (188)   2801    2756    rnd-5_family-361#Unknown        125     170     (1324)  -
```

```
                                                                   Score    E
Sequences producing significant alignments:                        (bits) Value

ref|XP_003389956.1| PREDICTED: putative nuclease HARBI1-like [Am...   236   3e-5
ref|XP_003387268.1| PREDICTED: putative nuclease HARBI1-like [Am...   213   3e-5
ref|XP_003731101.1| PREDICTED: putative nuclease HARBI1-like [St...   212   4e-5
ref|XP_003724948.1| PREDICTED: putative nuclease HARBI1-like iso...   195   4e-4
ref|XP_003389346.1| PREDICTED: putative nuclease HARBI1-like [Am...   195   4e-4

>ref|XP_003389956.1| PREDICTED: putative nuclease HARBI1-like [Amphimedon queens
          Length = 418

 Score =  236 bits (601), Expect = 3e-59
 Identities = 133/316 (42%), Positives = 180/316 (56%), Gaps = 6/316 (1%)
 Frame = +3

Query: 471   RRVWAHASSSDWWERIVLTTWDDRQRLEIFRTTRETLNHIVSMLGPKIARQDTVMRQPIA 650
              R +W    SS WW+ IV +T+ + Q +E FR ++ T   I + L   + +  T MRQPI
Sbjct: 67    RSIWTRQRSSYWWDHIVFSTFTNDQWMENFRMSKTTFVMICNELRSSLKKSSTTMRQPIP 126

Query: 651   PDKRVAMAIMKLATPASLRYIANQFGVAPCTAGLATCEVCQLLQDVAANSFICL---NNP 821
               +KRVA+++   +AT    R I + FGV+  +  LA  +VC+ +         +I      N
```

```
searching sequence folds in croc.newannot.fold
loading repeat families
```