15 October 2001

QRNA-1.1 documentation

Elena Rivas and Sean R. Eddy[1]

Department of Genetics,
Washington University, St. Louis, MO 63110 USA

Abstract

[1]To whom correspondence should be addressed. Tel: +1 314 362 7666; Fax: +1 314 362 7855; Email: eddy@genetics.wustl.edu.

The input file

- Start with a blastn output file **foo.blast**.
- Then you need to run the Perl scripts **blastn2qrna.pl**.
  (Perl scripts are in "~/qrna-1.1/scripts/").
- **"~/qrna-1.1/scripts/bastn2qrna.pl min_id min_len E-value_cutoff foo.blast "**
  - "min_id" is the minimum identity of blastn alignments allowed.
  - "min_len" is the minimum length of blastn alignments allowed.
  - "E-value_cutoff" removes blastn alignments with E values above this cutoff (e.g. 10 for default blastn E-value, 0.01 for a more stringent cutoff).
  - Even if you do not want to prune **foo.blast**, you have to run this script.
  - The script generates an output file named foo.blast.q
- **foo.blast.q** is the input file taken by **qrna**.
  It is a collection of sequences in fasta format, where two consecutive sequences are the two component of an alignment with the gaps in place.

To run the program

- **qrna** is the main program, written in C.
  Source code is in "~/qrna-1.1/src/".
- The "makefile" at "~/qrna-1.1/src/ can be modified to change the C compiler or to change compilation flags.
- To run qrna :
  ~/qrna-1.1/src/qrna -w ⟨winlen⟩ -x ⟨slide⟩ foo.blast.q
- Options are:

```
Usage: qrna [-options] <input_file.q> \n\
where options are:\n\
   -A             : do an all-to-all comparison between the two input files\n\
   -a             : print alignment \n\
   -b             : shuffle the alignment \n\
   -c <cfgfile>   : <cfgfile> to use to train the rna model (default = tRNA+rRNA)\n\
   -d             : log2 form (default = log2-odds space )\n\
   -e <num>       : number of sequen skipped in second file (for multiple comparisons). default 0. \n\
   -F             : change the overall base composition of the 3 models, based on nts frequencies in th
   -f             : use full dp for the probabilistic models (do not conserve the aligment-default is d
   -G             : change the overall base composition of the 3 models, based on nts frequencies for e
```

```
   -g             : do forward (default is viterbi)\n\
   -H             : include a file of Hexamer frequencies for the coding model\n\
   -h             : print short help and usage info\n\
   -i             : evolutionary time factor (default i=1)\n\
   -j             : use semi-full dp for the probabilistic model (use the alignment created by OTH)\n\
   -k             : allow pseudoknots (not implemented)\n\
   -l <minlenhit> : change the minlenhit parameter (default 0)\n\
   -L <maxlenhit> : change the maxlenhit parameter (default provided by longuest sequence)\n\
   -m             : do Forward and Viterbi Diagonal dp\n\
   -n             : do Forward and Viterbi Full     dp\n\
   -o <outfile>   : direct structure-annotated sequence to <outfile>\n\
   -p <pamfile>   : <pamfile> to use (default = BLOSUM62)\n\
   -P             : pedantic, check your evolutionary models for inconsistencies\n\
   -q             : do Forward and Viterbi semi-full dp\n\
   -r             : do Nussinov rna model (default is a 3-state model) \n\
   -s             : do global (not dp)\n\
   -S             : sweep a collection of motifs(seqfile1) across another bunch of sequences(seqfile2)\
   -t             : print traceback\n\
   -v             : verbose debugging output\n\
   -w <num>       : scanning window (default is full length)\n\
   -x <num>       : slide positions (default is 50)\n\
   -y             : grab sequences at random from the second data file\n\


   --cyk   : use CYK algorithm to calculate RNA score (default is Inside).
   --latte : I just called starbucks with your order...
   --rnass : print the alignment with the predicted RNA secondary structure.
   --shtoo : qrna the alignment and also give one shuffled score.
";
```

Example starting with a fasta file:

~/qrna-1.1/src/qrna ~/qrna1.1/Demos/trnas.fa
Output is:

```
#-------------------------------------------------------------------------------
#      qrna 1.1e (Wed Oct 10 17:36:25 CDT 2001) using squid 1.5m (Sept 1997)
#-------------------------------------------------------------------------------
#      PAM model =  BLOSUM62 scaled by 1.000
#-------------------------------------------------------------------------------
#      RNA model =  /mix_tied_linux.cfg
#-------------------------------------------------------------------------------
#      seq file  =  /nfs/wol2/people/elena/qrna-1.1e/Demos/trnas.fa
#                   #seqs: 2 (max_len = 124)
#-------------------------------------------------------------------------------
#      full length version:  -- length range = [0,1000]
#-------------------------------------------------------------------------------
# 1  [+ strand]
>DA0780 (124)
>DA0940 (124)

length alignment: 76 (id=72.37)
posX: 0-75 [0-72](73) -- (0.18 0.30 0.36 0.16)
posY: 0-75 [0-75](76) -- (0.14 0.34 0.37 0.14)
LOCAL_DIAG_VITERBI -- [Inside SCFG]
winner = RNA
OTH  ends = 73 0
COD  ends = 2 73
RNA  ends = 0 72
              OTH =       23.349             COD =       11.102             RNA =       45.313
   logoddspostOTH =        0.000  logoddspostCOD =      -12.247  logoddspostRNA =       21.964
```

- Every new blast alignment starts with two lines: ">Query_name" ">Subject_name"
- For a given scoring algorithm you get 5 numbers:
  - the three scores of the three models,
  - the two (COD and RNA) log-odds posterior probabilities respect to the OTH score.
  - Only 3 of these 5 numbers are independent. I like to report the OTH score, together with the two lododdsposteriors which determines the winningmodel in the following way:

$$\begin{aligned} \text{logoddspostCOD} > \text{logoddspostRNA} > 0 &\longrightarrow \text{winner} = \text{COD} \\ \text{logoddspostCOD} < \text{logoddspostRNA} \&\& \text{logoddspostRNA} > 0 &\longrightarrow \text{winner} = \text{RNA} \\ \text{logoddspostCOD} < 0 \&\& \text{logoddspostRNA} < 0 &\longrightarrow \text{winner} = \text{OTH} \end{aligned}$$

  - Option -a prints the scored alignment.

Some useful flags:

- For alignments that are too long, you can score them in chunks using -w ⟨windowsize⟩. The option -x ⟨slide⟩ to decide how many nucleotides

to move before you score another chunk of the given alignment. Every window analyzed starts with "length alignment:".

- Option -L determines the maximum length allowed for an alignment to be scored. The default is 1,000. I use this variable to control the memory usage of the program. If you are using the "window" version, memory is determined by the window so you can set -L as large as you want. If not using a window, I would not use a max length larger than $1,500$.
- There are three different scoring algorithms:

  global (-s)

  local viterbi (default)

  local forward (-g).

  You can report any of them or any desired combination of them. I would recommend to use the default.

Example using the scanning version with a window:

~qrna-1.1/src/qrna -w 150 -x 20 ~/qrna-1.1/Demos/briggsae-elegans_75.q
Output for the first two windows looks like:

```
#-------------------------------------------------------------------------------
#      qrna 1.1e (Wed Oct 10 17:36:25 CDT 2001) using squid 1.5m (Sept 1997)
#-------------------------------------------------------------------------------
#      PAM model =  BLOSUM62 scaled by 1.000
#-------------------------------------------------------------------------------
#      RNA model =  /mix_tied_linux.cfg
#-------------------------------------------------------------------------------
#      seq file  =  /nfs/wol2/people/elena/qrna-1.1e/Demos/briggsae-elegans_75.q
#                   #seqs: 31576 (max_len = 11666)
#-------------------------------------------------------------------------------
#      scanning version: window = 150   slide = 20 -- length range = [0,9999999]
#-------------------------------------------------------------------------------
# 1  [+ strand]
>R186-1- (575)
>G01A11-26286- (575)

length of whole alignment after removing common gaps: 575

length alignment: 150 (id=84.67)
posX: 0-149 [0-149](150) -- (0.29 0.21 0.14 0.35)
posY: 0-149 [0-149](150) -- (0.25 0.23 0.20 0.32)
LOCAL_DIAG_VITERBI -- [Inside SCFG]
winner = COD
OTH  ends = 0 149
COD  ends = 149 0
RNA  ends = 48 32
              OTH =       75.884             COD =      123.122             RNA =       67.625
   logoddspostOTH =        0.000  logoddspostCOD =       47.238  logoddspostRNA =       -8.259

length alignment: 150 (id=85.33)
posX: 20-169 [20-169](150) -- (0.29 0.22 0.13 0.36)
posY: 20-169 [20-169](150) -- (0.25 0.23 0.19 0.33)
```

```
LOCAL_DIAG_VITERBI -- [Inside SCFG]
winner = COD
OTH  ends = 20 169
COD  ends = 167 21
RNA  ends = 48 32
              OTH =       77.046             COD =      111.998             RNA =       69.070
   logoddspostOTH =        0.000  logoddspostCOD =       34.952  logoddspostRNA =       -7.976

length alignment: 150 (id=86.00)
posX: 40-189 [40-189](150) -- (0.26 0.22 0.14 0.38)
posY: 40-189 [40-189](150) -- (0.23 0.25 0.18 0.34)
LOCAL_DIAG_VITERBI -- [Inside SCFG]
winner = COD
OTH  ends = 40 189
COD  ends = 188 42
RNA  ends = 130 147
              OTH =       78.858             COD =      108.421             RNA =       68.126
   logoddspostOTH =        0.000  logoddspostCOD =       29.563  logoddspostRNA =      -10.732

length alignment: 150 (id=86.00)
posX: 60-209 [60-209](150) -- (0.27 0.25 0.14 0.35)
posY: 60-209 [60-209](150) -- (0.27 0.26 0.17 0.30)
LOCAL_DIAG_VITERBI -- [Inside SCFG]
winner = COD
OTH  ends = 60 209
COD  ends = 209 60
RNA  ends = 130 147
              OTH =       77.502             COD =      115.784             RNA =       66.943
   logoddspostOTH =        0.000  logoddspostCOD =       38.281  logoddspostRNA =      -10.559
```

Example starting with a blast output:

File **"~/qrna-1.1/Demos/ecoli-trnas.blast"** is a typical BLASTN output file.
Typing:
~/qrna-1.1/scripts/blastn2qrna.pl 50 50 0.01 ~/qrna-1.1/Demos/ecoli-trnas.blast
creates file **"~/qrna-1.1/Demos/ecoli-trnas.blast.q"**
This file has selected those blastn alignment with at least 50% identity, at least 50 nucleotides, and E-values ≤ 0.01. The two aligned sequence are now ready to be sent to qrna.
Typing:
~qrna-1.1/src/qrna -a ~/qrna-1.1/Demos/ecoli-trnas.blast.q

produces an output that starts with:

```
#-------------------------------------------------------------------------------
#      qrna 1.1e (Wed Oct 10 17:36:25 CDT 2001) using squid 1.5m (Sept 1997)
#-------------------------------------------------------------------------------
#      PAM model =  BLOSUM62 scaled by 1.000
#-------------------------------------------------------------------------------
#      RNA model =  /mix_tied_linux.cfg
#-------------------------------------------------------------------------------
#      seq file  =  /nfs/wol2/people/elena/qrna-1.1e/Demos/ecoli-trna.blast.q
```

```
#                   #seqs: 94 (max_len = 78)
#-------------------------------------------------------------------------------
#      full length version:  -- length range = [0,1000]
#-------------------------------------------------------------------------------
# 1  [+ strand]
>DA0780-1- (76)
>ECOLI-225501- (76)

length alignment: 76 (id=77.63)
posX: 0-75 [0-72](73) -- (0.18 0.30 0.36 0.16)
posY: 0-75 [0-71](72) -- (0.17 0.29 0.33 0.21)


             DA0780-1- GGGCTCGTAGCTCAGCTGGAAGAGCGCG.GCGTTTGCAACGCCG.AGGCC
          ECOLI-225501- GGGCTA.TAGCTCAGCTGGGAGAGCGCCTGC.TTTGCA.CGCAGGAGGTC

             DA0780-1- TGGGGTTCAAATCCC.CACGGGTCCA
          ECOLI-225501- TGCGGTTCGA.TCCCGCATAGCTCCA

LOCAL_DIAG_VITERBI -- [Inside SCFG]
winner = OTH
OTH  ends = 0 75
COD  ends = 74 0
RNA  ends = 0 12
              OTH =       20.871             COD =       -4.957             RNA =       15.208
   logoddspostOTH =        0.000  logoddspostCOD =      -25.828  logoddspostRNA =       -5.663
```

Visualization of Results

There are several Perl scripts to parse through an output file:

- **gnuphase.pl**
  Visual identification of the scores. A sort of "phase diagram" for the scores. Displays the two posterior log-odds scores in a $2-D$ plot.
  usage: **~/qrna-1.1/scripts/gnuphase.pl 1 ~/qrna-1.1/Demos/blastn.manyhits.wm+fdust.q.out 1 LOD 5 -100 100 -100 100**

  gnuphase.pl creates file "~/qrna-1.1/Demos/blastn.manyhits.wm+fdust.q.out.ps".

- **phase_count.pl**
  For a collection of scores, gives numerical details of how many fall in which phase.
  usage: **~/qrna-1.1/scripts/phase_count.pl ~/qrna-1.1/Demos/blastn.manyhits.wm+fdust.q.out org1 org2 1 5 100 1 RNA >blastn.manyhits.wm+fdust.q.out.1.5.100.rnaloci**

  phase_count.pl creates file "~/qrna-1.1/Demos/blastn.manyhits.wm+fdust.q.out.1.5.100.loci".