RIsearch: User manual
version 1.2

Updated by: Giulia I. Corsi[1,2]; Includes material from the publication of RIsearch, authored by Anne Wenzel[1,2], Erdinç Akbaşli[3], Jan Gorodkin[1,2]

[1]Center for non-coding RNA in Technology and Health, [2]Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark, [3] Software Development Group, University of Copenhagen, Rued Langgaards Vej 7, DK-2300 Copenhagen S, Denmark

June 15, 2021

Installation and execution

After download, unpack and compile with the following commands:

```
tar -xzvf RIsearch-1.1.tar.gz
cd RIsearch-1.1
make
```

This will create a standalone executable 'RIsearch'. For more convenient use, add the installation folder to PATH or copy the binary to a location that is in `$PATH`, i.e.

```
cp RIsearch /usr/local/bin/.
```

Usage

Required Parameters

RIsearch expects at least two parameters: query and target sequences. These can be given as fasta formatted files:

```
RIsearch -q query.fa -t target.fa
```

or directly on the command line:

```
RIsearch -Q acguacgu -T cgauagcuguagacugaugcau
```

A mixture of both is allowed as well. With multiple sequences in fasta files, RIsearch does an all-against-all comparison.

Optional Parameters

-s $< int >$ To include suboptimals, set some threshold score s, i.e.:

```
./RIsearch -q query.fa -t target.fa -s 2400
```

NB, the threshold is on the score not the energy (subject to change). The first hit is the 'best' one from before, followed by the suboptimals (includes the top hit again). Here, we first get subsets of the top hit, followed by some alternative duplexes (and subsets). With the additional flag -n 20 the latter spurious hits are avoided (only report the highest-scoring hit from a neighborhood of 20).

```
./RIsearch -q query.fa -t target.fa -s 2400 -n 20
```

-d $< int >$ This option sets the per-nucleotide extension penalty in dacal/mol (30 seems to be a good value), which favors short stable interactions. This is especially important if larger query sequences are used and one does not expect the whole sequence to be part of the interaction.

p[123]

- p1 : gives a shorter output (one line per interaction), the 'top hit' is always first and contained a second time in the list.
- p2 : does not produce a header per query/target pair, but instead repeats sequence names on each line, for an easy to parse tsv file with the following order: 'Qname Qbeg Qend Tname Tbeg Tend score energy' (Q for query, T for target) The 'top hit' is only printed once and not necessarily first.
- p3: produces one line per pair with number of hits that would have been printed, tab seperated 'Qname Tname hit-count'

-e $< num >$ Energy threshold. Is checked after backtrack, only print interactions with energies lower than or equal to this.

-m $< str >$ Specity the scoring matrix of nearest neighbor parameters, default is Turner 2004 (RNA-RNA). The previous version of RIsearch (v.1.1) computes RNA-RNA binding free energies using the scoring schemes specified in the Turner 1999 (t99) or Turner 2004 (t04) energy models [4, 3]. In addition to these, RIsearch v.1.2 includes the energy model for RNA-DNA hybrids of Sugimoto (1995 and 2000) [6, 7] and Watkins [8], in matrices su95 and su95_noGU (no GU wobble base pairs). The scoring scheme for DNA-DNA duplexes were derived from SantaLucia and Allawi [5, 2] and the corresponding scoring scheme is named sl04_noGU. The parameters were imported from RIsearch v.2.1, modified as described in [1], and simply reformatted.

-n $< int >$ **and -l** $< int >$ Parameter `-n` sets the size of the neighborhood, hence allowing backtrack only from the best position within this range to omit many overlapping results; default is 0, backtrack all. Parameter `-l` sets the max trace back length (default: 40)

		target 3'-5'									
		-	**A**	**C**	**C**	**C**	**C**	**G**	**G**	**G**	**G**
	-	0	-8	-8	-8	-8	-8	-8	-8	-8	-8
	G	-8	0	150	150	150	150	0	0	0	0
	G	-8	0	150	480	480	480	110	0	0	0
query	**G**	-8	0	150	480	810	810	440	218	178	138
5'-3'	**G**	-8	0	150	480	810	1140	770	548	508	468
	C	-8	0	0	0	320	650	1480	900	860	820
	C	-8	0	0	0	218	548	900	1810	1240	1200
	C	-8	0	0	0	178	508	860	1240	2140	1570
	C	-8	0	0	0	138	468	820	1200	1570	2470

Table 1: RIsearch v.1.1 match and mismatch (M) matrix computed for query GGGGCCCC and target AGGGGCCCCA

Forced interactions (*-f* $<$ *int* $>$) Option *-f* forces interactions to start at the 3' target end and to end at the 3' and 5' ends of the query and the target, respectively.
NOTE: Options -d -s -n -l -e -p are not available in combination with option -f.
To force the interaction to start at the 3' target end a large number is placed in the first column of the Smith-Waterman-like matrix that stores the scores of matches and mismatches in RIsearch (the "M" matrix). In this way starting at the first nucleotide of the target (3' end) will always be preferential compared to starting at another column in the M matrix or anywhere else, including any of the two scoring matrices that keep track of gapped bindings. To force the interaction to end at the 3' and 5' ends of the query and the target, respectively, the maximum among the bottom-right values in any of the 3 scoring matrices is always considered the optimal one, and is used to start the backtracking. Ending with a bulge is allowed in RIsearch v.1.2. The penalty for closing the interaction with a bulge is set to the same cost as expanding a bulge in all of the nearest neighbor matrices. Note that this will not affect non-forced interactions, as the maximum score is never the one ending with a bulge, which implies a penalty.

For instance, executing RIsearch with default settings on query `GGGGCCCC` and target `AGGGGCCCCA` produces the M matrix reported in Table 1. Note that the 5' nucleotide "A" of the target is omitted in the output format of RIsearch, as it only penalizes the binding. If present, it would be displayed in Table 1 as an additional column, with scores strictly lower than those in the current last column. The best score is 2470 and the binding pattern can symbolized as:

```
GGGGCCCC
||||||||
CCCCGGGG
```

With option *-f*, RIsearch v.1.2 can be executed on the same query (Q) and target (T) as follows:

```
./RIsearch -Q GGGGCCCC -T AGGGGCCCCA -f 2500 -w noweights
```

This means that the number 2500 will be added to the first score column, instead of -8 (which stands for $-\infty$). If the number given to *-f* is too low and the backtracking does not end at the first nucleotide at the 3' end of the target (first column in M), then an error is given, suggesting to increase the parameter:

Force start option did not work: try to increase the number given to -f.

Try with a number > 5000

The suggested number, 5000 in this case, is the maximum between 2000 and the length of the query * 500 or the length of the target * 500. The number 500 is greater than any penalty for a valid interaction in the nearest neighbor matrices. However, because in this example the match is almost perfect, a lower value is sufficient to force the interaction to start with a mismatch. Option *-w* must be specified in combination with *-f*. To avoid weights, the flag "noweights" can be activated to set all weights to 1. In this case the M matrix will look as in Table 2. If the interaction was not forced with *-f*, the best score would be 4419, but instead the score is taken from the bottom right cell of matrix Bt (Table 2), which contains the scores relatives to interactions with bulges in the target and that in this case is 4179. The lower score will result in a higher binding energy, which is expected as the binding is forced to follow a non-spontaneous pattern. The binding pattern is symbolized as:

Query: `M|||||||`
Target: `MB|||||||B`

In which M is a mismatch, B a bulge and `|` is a match. This can be re-written as follows using a single line of symbols:

```
G-GGGCCCC-
MB|||||||B
ACCCCGGGGA
```

However, the formulation with two lines of symbols was selected for this case, because it specifies in which of the two sequences the bulges are located, without having to read or modify the sequences themselves.

Weighted stacking interactions (*-w* < *str* >) Option *-w* was introduced to weight CRISPR/-Cas9 gRNA-target-DNA interactions. With this option, given a predefined array of weights (defined in file `weights.c` and compiled together with RIsearch v.1.2) the contribution of a stacking base pair, mismatch or bulge is multiplied by a corresponding weight. The array of weights must be at least as long as the query sequence minus 1. Thus, for a CRISPR/Cas9 interaction in which the gRNA is made of 20 nt, 19 weights are necessary. This is because only stacking interactions are weighted. If the provided array of weights is longer than the query size -1, then it is shortened from left to right. The $-w$ option requires $-f$ to be set.
NOTE: Options -d -s -n -l -e -p are not available in combination with option -w.
Weights are relative to positions on the query sequence, as shown in the following examples:

- If query and target are complementary with each other, weights are distributed from the second base pair.

```
query sequence    GGGGCCCC
weight positions:  1234567
target sequence   CCCCGGGG
```

- If the query contains a bulge (insertion), the corresponding weight is "consumed" at that position to weight the energy penalty related to the bulge, as in the example below.

M matrix

	target 3'-5'										
		-	**A**	**C**	**C**	**C**	**C**	**G**	**G**	**G**	**G**	**A**
query 5'-3'	**-**	2500	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8
	G	2500	2500	150	150	150	150	0	0	0	0	0
	G	2500	2500	2580	2429	2389	2349	2287	2247	2207	2167	2127
	G	2500	2500	2580	2910	2759	2719	2309	2265	2225	2185	2145
	G	2500	2500	2580	2910	3240	3089	2679	2568	2528	2488	2448
	C	2500	2500	2478	2420	2750	3080	3429	2960	2920	2880	2818
	C	2500	2500	2478	2456	2648	2978	3080	3759	3290	3250	3087
	C	2500	2500	2478	2456	2608	2938	3009	3410	4089	3620	3457
	C	2500	2500	2478	2456	2568	2898	2969	3339	3740	4419	3827

Bt matrix

	target 3'-5'											
		-	**A**	**C**	**C**	**C**	**C**	**G**	**G**	**G**	**G**	**A**
query 5'-3'	**-**	0	0	0	0	0	0	0	0	0	0	0
	G	-8	-8	2429	2389	2349	2309	2269	2229	2189	2149	2109
	G	-8	-8	2429	2389	2349	2309	2269	2229	2189	2149	2109
	G	-8	-8	2429	2389	2670	2630	2590	2550	2510	2470	2430
	G	-8	-8	2429	2389	2670	3000	2960	2920	2880	2840	2800
	C	-8	-8	2429	2407	2367	2679	3009	3189	3149	3109	3069
	C	-8	-8	2429	2407	2385	2577	2907	2867	3519	3479	3439
	C	-8	-8	2429	2407	2385	2537	2867	2827	3170	3849	3809
	C	-8	-8	2429	2407	2385	2497	2827	2787	3099	3500	4179

Table 2: RIsearch v.1.2 match and mismatch (M) and target bulges (Bt) matrices computed for query GGGGCCCC and target AGGGGCCCCA with option *-f 2500* and *-w noweights*

```
query sequence    GGAGGCCCC
weight positions:  12345678
target sequence   CC-CCGGGG
```

- If the target contains a bulge (insertion), no weight is consumed and the bulge is weighted by the average between the current and the next weight, in this case weights 1 and 2. The average is symbolized with `w` below.

```
query sequence    GG-GGCCCC
weight positions:  1w234567
target sequence   CCACCGGGG
```

- If the target contains a bulged and the whole query sequence is already employed in the binding, then the bulge is weighted with the last available weight. The reason for this is the effect of option *-f*, which forces the interaction to terminate at the last available nucleotide of both the query and the target, even though this means to terminate at a bulge.

```
query sequence    GGGGCCCC-
weight positions:  12345677
target sequence   CCCCGGGGA
```

Note that while RIsearch v.1.1 solely works with integers in its scoring matrices, RIsearch v.1.2 allows for float values as well, thus the weights can be in floating-point. The array of weights

for CRISPR interactions is named `CRISPR_20nt_5p_3p` in RIsearch v.1.2 and contains 19 weights $[w_1, w_2, ..., w_{18}, w_{19}]$, where w_{19} is the weight assigned to the first PAM-proximal interaction. For a query gRNA and target DNA binding site the array `CRISPR_20nt_5p_3p` can be used as follows:

```
./RIsearch -Q ACAAATTTGGGGAGCTCTTC -T GAAGAGCTCCCCAAATTTGT \
-m su95 -f 5000 -w CRISPR_20nt_5p_3p
```

The scoring matrix here is su95 (RNA-DNA interactions with wobble base pairs allowed). The related free energy is -50.45 $kcal/mol$. gRNAs shorter than 20 nt can be scored with the same array of weights. For instance, in the case of a 19 nt gRNA only the last 18 weights $[w_2, w_3, ..., w_{18}, w_{19}]$ will be used.

References

[1] F. Alkan, A. Wenzel, C. Anthon, J. H. Havgaard, and J. Gorodkin. CRISPR-cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biology*, 19(1), Oct. 2018.

[2] H. T. Allawi and J. SantaLucia. Thermodynamics and nmr of internal g·t mismatches in dna. *Biochemistry*, 36(34):10581–10594, Aug 1997.

[3] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19):7287–7292, May 2004.

[4] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, May 1999.

[5] J. SantaLucia and D. Hicks. The thermodynamics of DNA structural motifs. *Annual Review of Biophysics and Biomolecular Structure*, 33(1):415–440, June 2004.

[6] N. Sugimoto, S. ichi Nakano, M. Katoh, A. Matsumura, H. Nakamuta, T. Ohmichi, M. Yoneyama, and M. Sasaki. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, 34(35):11211–11216, Sept. 1995.

[7] N. Sugimoto, M. Nakano, and S.-i. Nakano. Thermodynamics-Structure Relationship of Single Mismatches in RNA/DNA Duplexes. *Biochemistry*, 39(37):11270–11281, Sep 2000.

[8] N. E. Watkins, W. J. Kennelly, M. J. Tsay, A. Tuin, L. Swenson, H.-R. Lee, S. Morosyuk, D. A. Hicks, and J. SantaLucia. Thermodynamic contributions of single internal rA·dA, rC·dC, rG·dG and rU·dT mismatches in RNA/DNA duplexes. *Nucleic Acids Research*, 39(5):1894–1902, Nov. 2010.