RIsearch2

Large-scale RNA–RNA interaction prediction

Anne Wenzel

Center for non-coding RNA in Technology and Health
Department of Veterinary Clinical and Animal Sciences
Faculty of Health and Medical Sciences
University of Copenhagen

31st TBI Winterseminar
Bled, Slovenia
February 2016

RTH
CENTER FOR NON-CODING RNA
IN TECHNOLOGY AND HEALTH

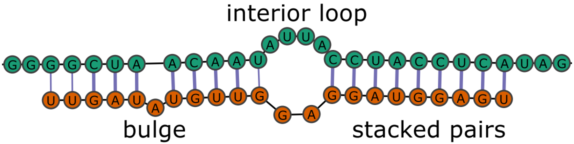
UNIVERSITY OF COPENHAGEN
MEDICAL SCIENCES

Motivation

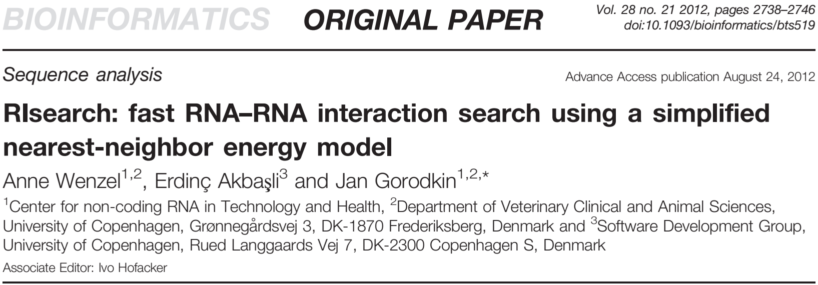
Regulatory, non-coding RNAs

- often function by forming a duplex with other RNAs
- many identified but unknown targets
- their interactome provides insight to function

Goal: Predict RNA–RNA duplexes *in silico* on genome-wide scale



Recap: RIsearch



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 28 no. 21 2012, pages 2738–2746
doi:10.1093/bioinformatics/bts519

Sequence analysis

Advance Access publication August 24, 2012

RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model

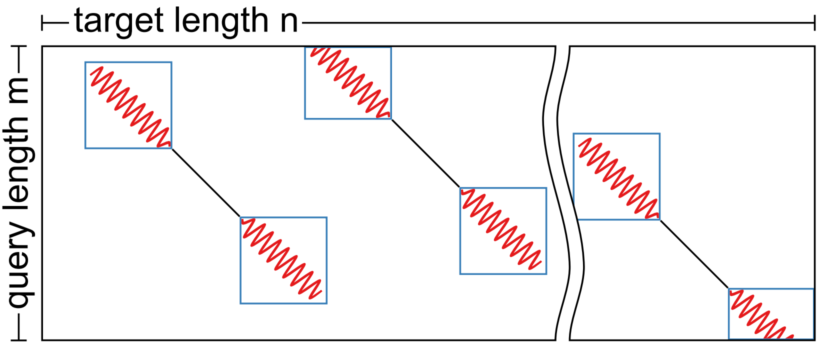
Anne Wenzel[1,2], Erdinç Akbaşli[3] and Jan Gorodkin[1,2,*]

[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 and [3]Software Development Group, University of Copenhagen, Rued Langgaards Vej 7, DK-2300 Copenhagen S, Denmark

Associate Editor: Ivo Hofacker

- dinucleotide scoring matrix in Smith–Waterman-like algorithm
- approximates Nearest Neighbor energy model
- computed energies deviate from full model, BUT
- candidates are ranked similar
- fast method for predicting near-complementary duplexes

Search space



- RIsearch: DP over $\boxed{m \times n}$
- idea: seed-and-extend (DP on either end over $\boxed{l^2}$)
- GUUGle + RIsearch performs very well for miRNAs
- Now: one stop shop → RIsearch2

RIsearch2: outline

Suffix array enables very fast seed detection

- build generalized suffix array of the target (entire human genome stored in 47 GiB)
- build partial suffix array of the query (according to seed settings)
- match suffix arrays (allowing wobble pairs)

Extend seeds with DP as before

- DP matrix anchored at first/last position of seed

RIsearch2: building the target index

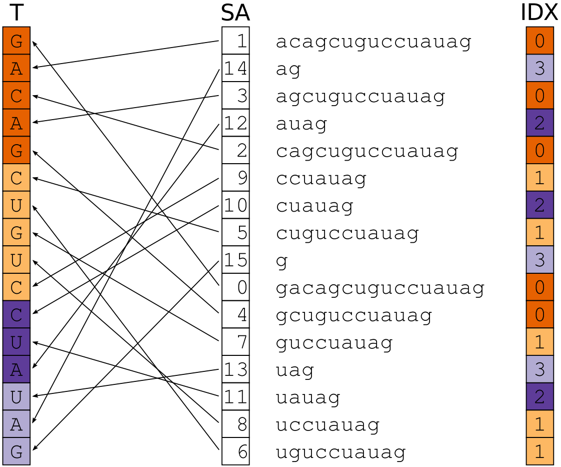
```
>targetSeqA
gacag                gacagcuguccuauag   T
>targetSeqB          0000011111222333  IDX
cua
```

T	SA		IDX
G	0	gacagcuguccuauag	0
A	1	acagcuguccuauag	0
C	2	cagcuguccuauag	0
A	3	agcuguccuauag	0
G	4	gcuguccuauag	0
C	5	cuguccuauag	1
U	6	uguccuauag	1
G	7	guccuauag	1
U	8	uccuauag	1
C	9	ccuauag	1
C	10	cuauag	2
U	11	uauag	2
A	12	auag	2
U	13	uag	3
A	14	ag	3
G	15	g	3

RIsearch2: building the target index

```
>targetSeqA
gacag
>targetSeqB
cua
```

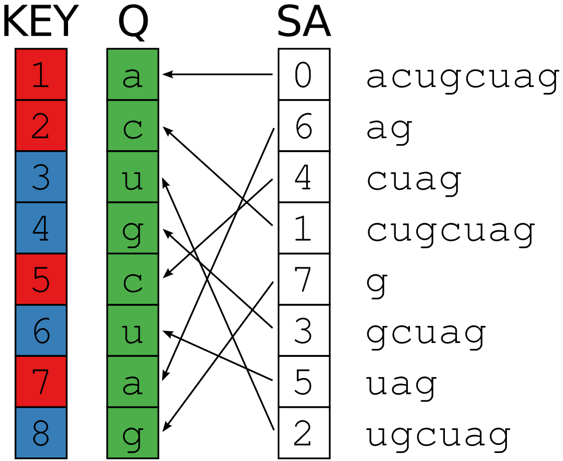
```
gacagcuguccuauag    T
0000011111222333    IDX
```



T	SA		IDX
G	1	acagcuguccuauag	0
A	14	ag	3
C	3	agcuguccuauag	0
A	12	auag	2
G	2	cagcuguccuauag	0
C	9	ccuauag	1
U	10	cuauag	2
G	5	cuguccuauag	1
U	15	g	3
C	0	gacagcuguccuauag	0
C	4	gcuguccuauag	0
U	7	guccuauag	1
A	13	uag	3
U	11	uauag	2
A	8	uccuauag	1
G	6	uguccuauag	1

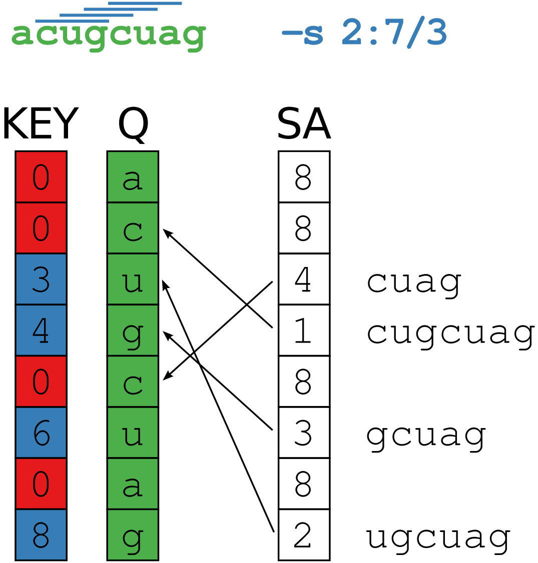
RIsearch2: building (partial) query SA

acugcuag   -s 2:7/3



KEY	Q	SA	
1	a	0	acugcuag
2	c	6	ag
3	u	4	cuag
4	g	1	cugcuag
5	c	7	g
6	u	3	gcuag
7	a	5	uag
8	g	2	ugcuag

RIsearch2: building (partial) query SA



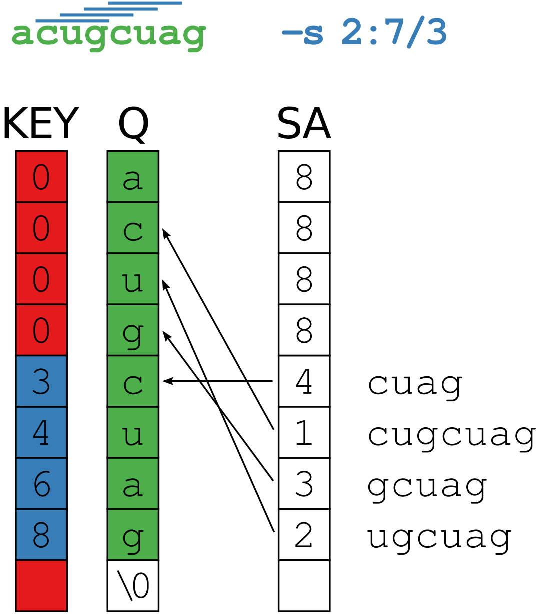
acugcuag

-s 2:7/3

KEY	Q	SA	
0	a	8	
0	c	8	
3	u	4	cuag
4	g	1	cugcuag
0	c	8	
6	u	3	gcuag
0	a	8	
8	g	2	ugcuag

RIsearch2: building (partial) query SA

acugcuag   -s 2:7/3

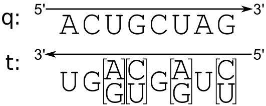


RIsearch2: parallel matching SAs

KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		

SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

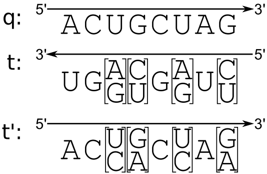
RIsearch2: parallel matching SAs



KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		

SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

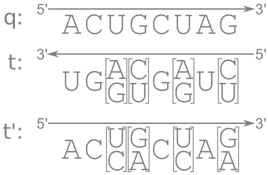
RIsearch2: parallel matching SAs



KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		

SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

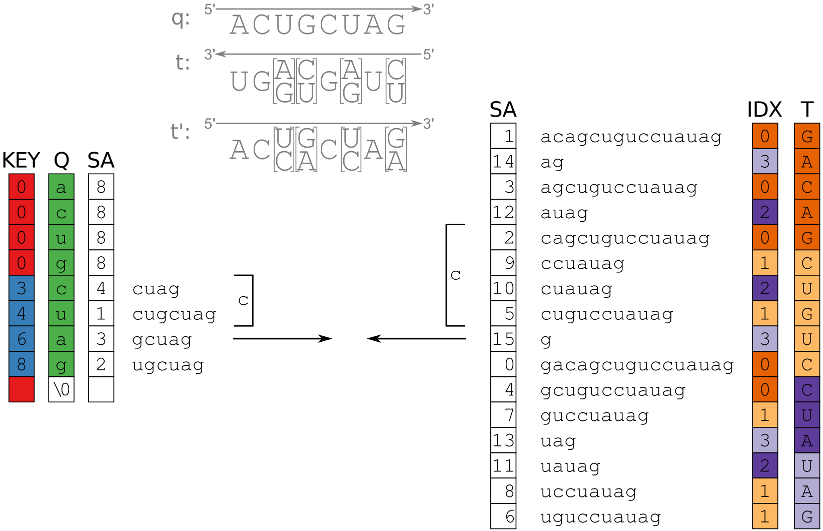
RIsearch2: parallel matching SAs



KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		

SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

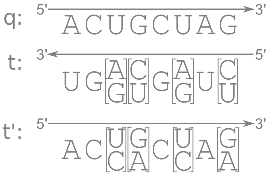
RIsearch2: parallel matching SAs



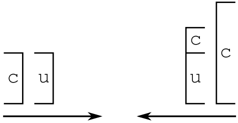
KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		

SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

RIsearch2: parallel matching SAs

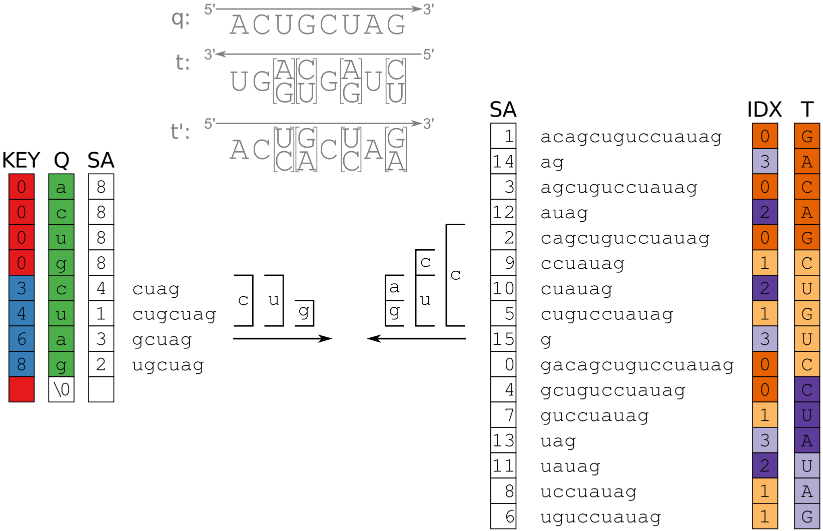

q: 5' ACUGCUAG 3'
t: 3' UG A/G C/U G A/G U C/U 5'
t': 5' AC U/C G/A C U/C A G/A 3'

KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		



SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

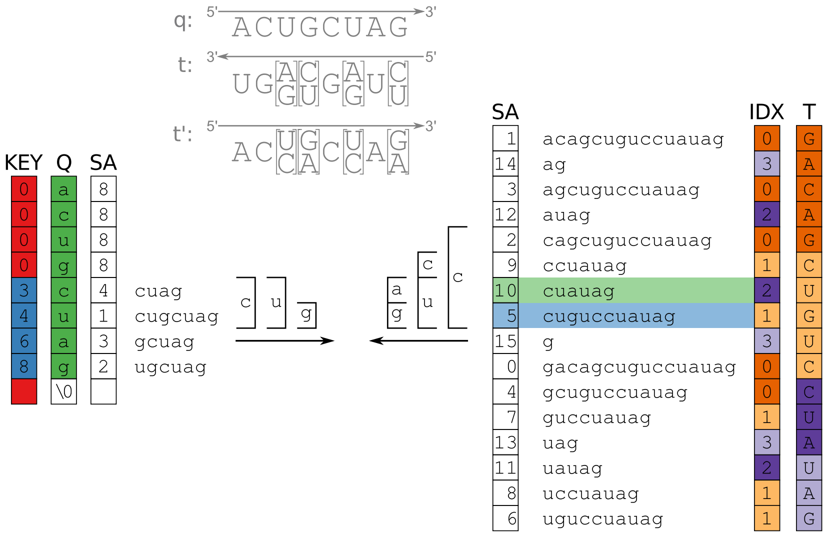
RIsearch2: parallel matching SAs


q: 5' ACUGCUAG 3'
t: 3' UG A/G C/U G A/G U C/U 5'
t': 5' AC U/C G/A C U/C A G/A 3'
c
u
g
a
g
c
u
c

KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		

SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

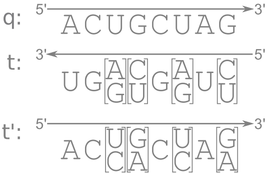
RIsearch2: parallel matching SAs


q: 5' ACUGCUAG 3'
t: 3' UG A/G C/U G A/G U C/U 5'
t': 5' AC U/C G/A C U/C A G/A 3'
c
u
g
a
g
c
u
c

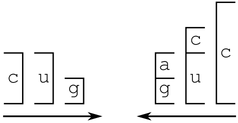
KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		

SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

RIsearch2: parallel matching SAs



KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		

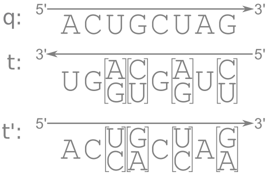


SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G

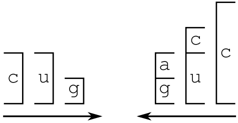
T gacagcuguccuauag

IDX 0000011111222333

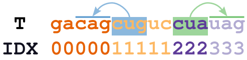
RIsearch2: parallel matching SAs


q: 5' ACUGCUAG 3'
t: 3' UG A G C U G A G U C U 5'
t': 5' AC U C G A C U C A G A 3'

KEY	Q	SA	
0	a	8	
0	c	8	
0	u	8	
0	g	8	
3	c	4	cuag
4	u	1	cugcuag
6	a	3	gcuag
8	g	2	ugcuag
	\0		



SA		IDX	T
1	acagcuguccuauag	0	G
14	ag	3	A
3	agcuguccuauag	0	C
12	auag	2	A
2	cagcuguccuauag	0	G
9	ccuauag	1	C
10	cuauag	2	U
5	cuguccuauag	1	G
15	g	3	U
0	gacagcuguccuauag	0	C
4	gcuguccuauag	0	C
7	guccuauag	1	U
13	uag	3	A
11	uauag	2	U
8	uccuauag	1	A
6	uguccuauag	1	G



RIsearch2: Benchmark

100 miRNAs vs. repeat-masked human genome

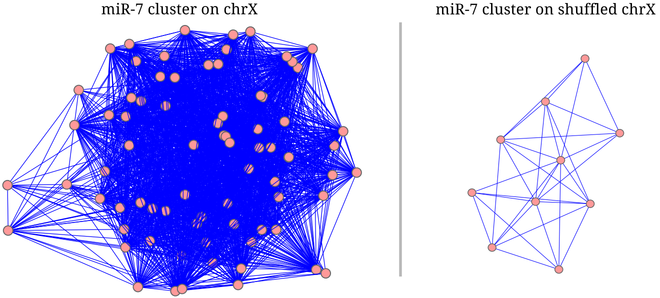
variant	run time [h:mm:ss]
RIsearch	
1 optimal per query–target pair (miR–chr-strand)	65:58:21
suboptimals $\leq -10\,\mathrm{kcal/mol}$	184:47:47
RIsearch2 with different seed sizes anywhere in the query	
-s 8	2:28:39
-s 7	6:36:24
-s 6	19:10:45
RIsearch2 with seeds position-constrained in the query	
-s 1:8/6 (1–6 / 2–7 / 3–8)	4:05:43
-s 2:7 (2–7)	1:29:08
-s 2:7/5 (2–6 / 3–7)	8:49:09

This is single-core time, additionally RIsearch2 is multi-threaded.

Application 1: miRNA sponge discovery

`RIsearch2` is especially efficient when constrained to seed region

- screen all human mature miRNAs against the human genome
- identify genomic regions with high binding site density
- select candidates based on difference real vs. shuffled



miR-7 cluster on chrX

miR-7 cluster on shuffled chrX

Pan *et al.*, in prep.

Application 2: search for off-targets

Off-targeting is a problem in probe/siRNA design

siRNAs can effect transcripts other than the intended target

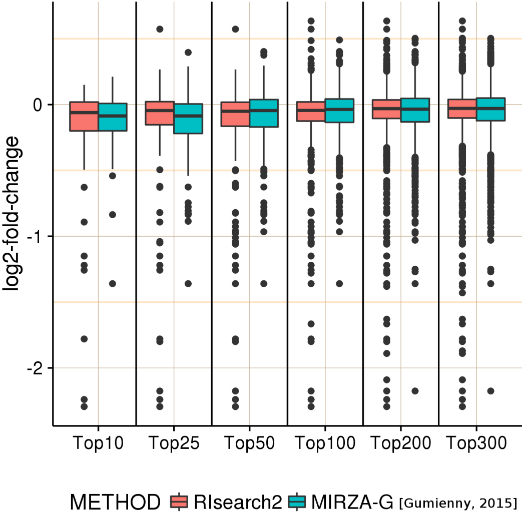
- (near-)perfect complementarity : silencing
- imperfect binding to 3' UTR : miRNA-like effect

Approach:

- use `RIsearch2` to screen for potential off-targets
- combine with accessibility profiles and transcript abundance
- partition function to quantify off-target volume (per siRNA)
- compute off-targeting probabilities (per transcript)

Application 2: search for off-targets (cont.)

Predicted off-targets are down-regulated upon siRNA transfection (combined results for 6 siRNAs [Burchard, 2009])



[Gumienny, 2015]

Acknowledgements

`RIsearch2` coders (in descending order of recentness of their contributions):

Ferhat Alkan
Oana Palasca
Peter Kerpedjiev
Anders F Rudebeck

Peter F. Stadler
Ivo L. Hofacker

Jan Gorodkin and our whole group at RTH

DET FRIE FORSKNINGSRÅD
DANISH COUNCIL FOR
INDEPENDENT RESEARCH

Innovation Fund Denmark