RNAiFold:
Complete Inverse Folding
for Synthetic Design
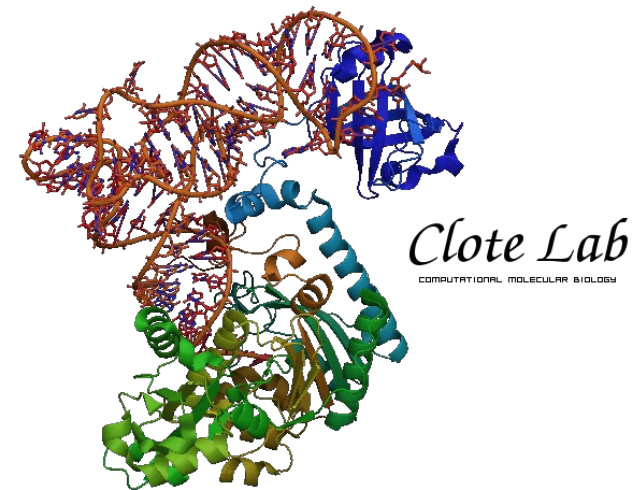
Juan Antonio García Martín
23/07/2015

Benasque 2015: Computational Analysis of RNA Structure and Function

RNA folding

Sequence + **Structure** ----→ Function

Folding

AAGCGCAAAGCGAGACGCAAGCGC

MFE

Inverse folding

Synthetic RNA design approaches

Computational

Energy model based

Experimental

SELEX

AD-HOC

Collaborative

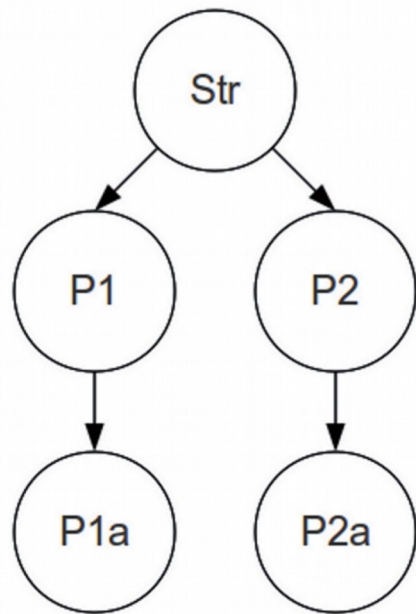Crowd science

Weighted sampling/Similarity
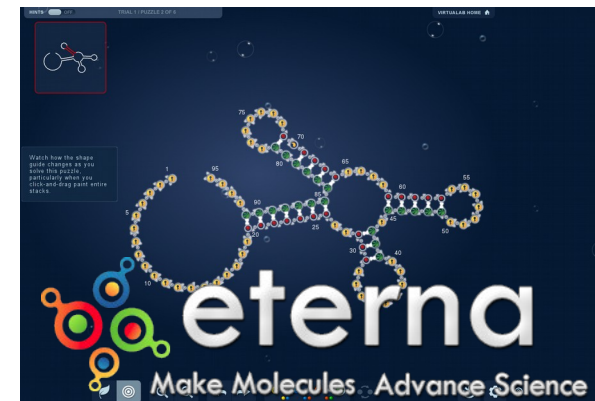
Objective

Novel algorithm design and implementation of tools to design synthetic RNA

Approach

Generate a large number of sequences whose minimum free energy structure is identical to the target design structure, and subsequently filter with respect to different criteria in order to select the most promising candidates for biochemical validation

Complete modular RNA inverse folding

- Design:
  - Constraint programming
    - Model: Variables with a discrete domain
    - Constraints: Defines the scope of the search
    - Heuristics: Increases search speed
    - Objective: Stop condition (# solutions, optimization)

- Requirements:
  - Able to generate all possible solutions for small structures and millions for longer structures.
  - Fast when completeness is not required
  - Allows design flexibility

RNAiFold: Insights

CP Model

Tree-like decomposition

Tree reduction

Search heuristics

MFE helix check

MFE helix check

Garcia-Martin JA, Clote,P , Dotu, I . RNAiFold: a constraint programming algorithm for RNA inverse folding and molecular design. J. Bioinform. Comput. Biol. 2013;11:1350001.

RNAiFold
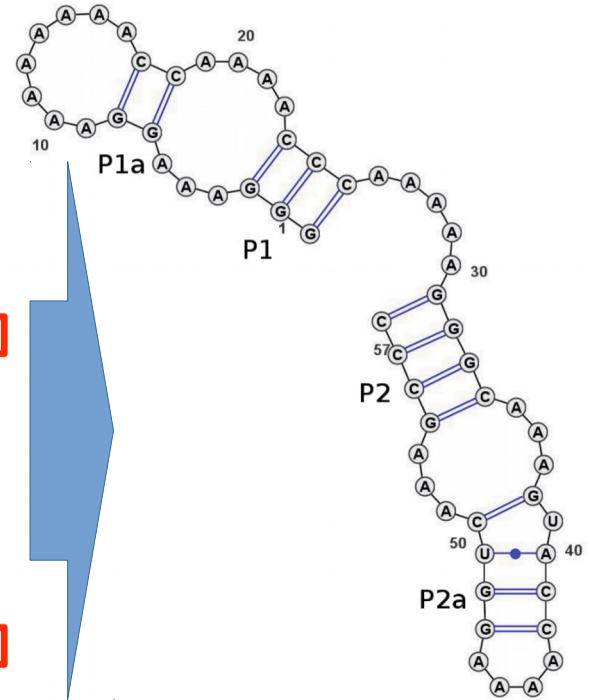
Additional features

Cofolding
Template sequence
GC%
Limit #base pairs per type
Consecutive bases
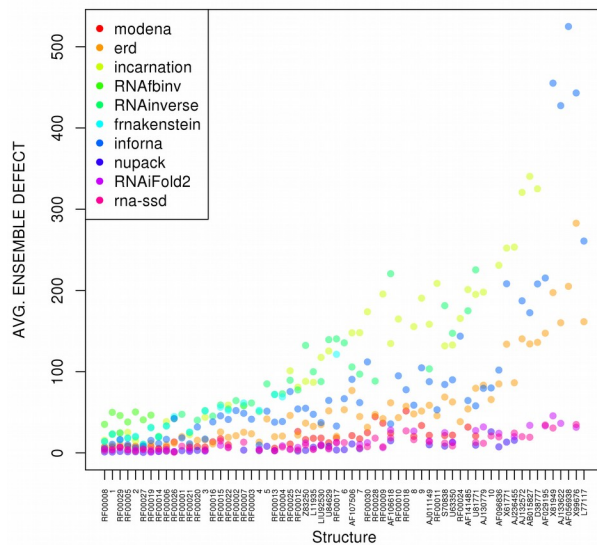Incompatible base pairs

First complete inverse folding

LNS

Garcia-Martin JA, Clote,P , Dotu, I . RNAiFold: a constraint programming algorithm for RNA inverse folding and molecular design. J. Bioinform. Comput. Biol. 2013;11:1350001.

Performance

Table 1. **Rfam CP Results.**

Parameters		CP		INFO-RNA		MODENA		RNA-SSD		RNAinverse	
RF id	n	sol	time	sol	time	sol	time	sol	time	sol	time
RF00001.121	117	38	21.5	**50**	**0.0**	6	36.8	22	1.0	41	233.1
RF00002.2	151	**44**	29.5	4	62.6	20	39.4	6	**12.2**	0	-
RF00003.94	161	0	-	1	72.1	**29**	**70.2**	0	-	0	-
RF00004.126	193	**50**	1.5	**50**	**0.1**	34	52.9	**50**	2.0	**50**	48.3
RF00005.1	74	**50**	0.2	**50**	**0.0**	33	12.4	**50**	0.1	**50**	0.1
RF00006.1	89	**50**	0.3	**50**	**0.0**	37	15.1	**50**	0.6	**50**	4.3
RF00007.20	154	**50**	5.6	**50**	**0.0**	34	44.4	**50**	1.1	**50**	12.4
RF00008.11	54	**50**	0.1	**50**	**0.0**	26	8.7	**50**	**0.0**	**50**	**0.0**
RF00009.115	348	**48**	**20.8**	0	-	29	214.1	26	48.2	0	-
RF00010.253	357	**0**	-	**0**	-	**0**	-	**0**	-	**0**	-
RF00011.18	382	**0**	-	**0**	-	**0**	-	**0**	-	**0**	-
RF00012.15	215	**50**	**2.7**	15	25.0	27	64.5	28	28.8	1	139.4
RF00013.139	185	**50**	1.6	**50**	**0.8**	12	51.5	49	2.8	**50**	19.8
RF00014.2	87	**50**	0.3	**50**	**0.0**	33	17.5	49	0.1	**50**	**0.0**
RF00015.101	140	49	1.3	**50**	**0.2**	38	29.1	40	0.6	**50**	52.4
RF00016.15	129	**0**	-	**0**	-	**0**	-	**0**	-	**0**	-
RF00017.90	301	**50**	19.3	**50**	**0.0**	28	208.1	**50**	7.0	**50**	10.0
RF00018.2	360	**47**	**12.1**	1	697.0	28	331.5	0	-	0	-
RF00019.115	83	**50**	0.2	**50**	**0.0**	32	14.9	**50**	0.2	**50**	0.3
RF00020.107	119	**0**	-	**0**	-	**0**	-	**0**	-	**0**	-
RF00021.10	118	**50**	0.3	**50**	**0.0**	37	27.8	49	0.2	**50**	0.2
RF00022.1	148	**50**	0.7	**50**	**0.0**	38	32.6	24	0.9	35	225.5
RF00024.16	451	**0**	-	**0**	-	**0**	-	**0**	-	**0**	-
RF00025.12	210	**50**	1.4	9	47.9	33	54.2	29	2.9	0	-
RF00026.1	102	**50**	**0.4**	33	5.5	38	15.2	**50**	1.4	44	173.2
RF00027.7	79	**50**	0.1	**50**	**0.0**	32	17.4	**50**	0.1	**50**	0.4
RF00028.1	344	**39**	**6.2**	0	-	0	-	4	71.2	0	-
RF00029.107	73	**50**	0.3	**50**	**0.0**	37	10.4	**50**	0.2	**50**	0.3
RF00030.30	340	**46**	**6.8**	1	57.3	22	186.8	34	39.3	0	-
sum	-	**1111**	**133.2**	813	271.5	683	1555.5	860	220.9	771	919.7
avg	-	**38.3**	**5.7**	28.0	12.9	23.6	67.6	29.7	10.0	26.6	54.1

Garcia-Martin JA, Clote,P , Dotu, I . RNAiFold: a constraint programming algorithm for RNA inverse folding and molecular design. J. Bioinform. Comput. Biol. 2013;11:1350001.
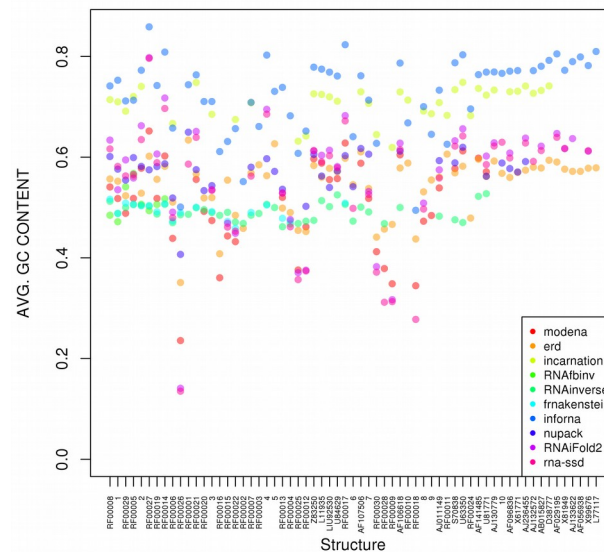
Output sequences

Method	ERD	FRNA	Incarnation	Info-RNA	MODENA	Nupack	RNA-SSD	RNAfbinv	RNAiFold2	RNAinverse
Output (%)	100%	30%	60%	95%	60%	57%	90%	13%	65%	65%
Target (%)	85%	38%	0%	57%	45%	70%	82%	0%	100%	18%
Avg str len	397	122	352	393	234	256	400	74	363	208
Avg output	117	325	41,535	195	50	22	1	2	55,476	935
P(S)	3.32%	1.70%	0.06%	3.17%	11.30%	30.01%	2.24%	0.36%	23.21%	0.78%
Native cont. (%)	85±9	61±15	63±13	76±12	89±9	98±1	85	32±6	93±2	57±12
Avg E	-0.41	-0.24	-0.46	-0.63	-0.46	-0.44	-0.30	-0.14	-0.56	-0.23
Pos entropy	0.33	0.71	0.41	0.44	0.15	0.07	0.36	0.88	0.12	0.80
MH diversity	0.16	0.35	0.21	0.22	0.07	0.03	0.18	0.45	0.06	0.38
Vienna diversity	0.11	0.23	0.15	0.16	0.05	0.02	0.11	0.30	0.05	0.26
Exp bp dist	0.09	0.21	0.27	0.16	0.06	0.01	0.08	0.38	0.03	0.24
Ens def	0.14	0.32	0.39	0.22	0.08	0.02	0.14	0.56	0.04	0.37
Exp num bp	0.28	0.29	0.34	0.30	0.26	0.29	0.28	0.28	0.27	0.28
GC-content (%)	55%	49%	71%	72%	50%	57%	36%	51%	57%	49%

AVG. ENSEMBLE DEFECT   AVG. GC CONTENT   AVG. EXPECTED PROPORTION OF NATIVE CONTACTS
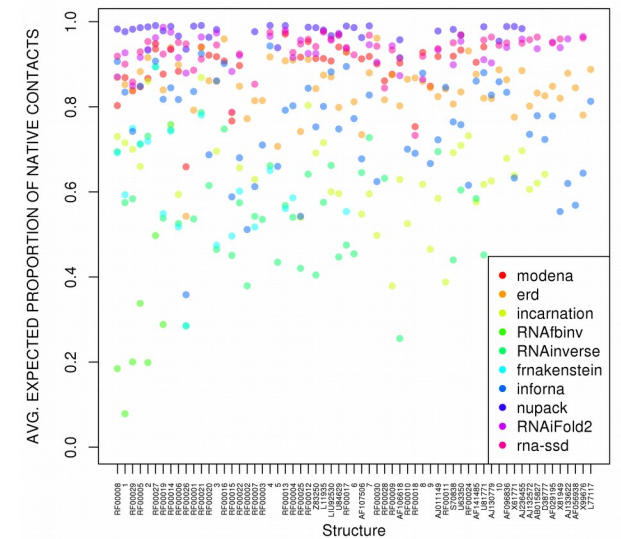
RNAiFold 1.0: Limitations

- Memory usage

  - 32 bit software limits memory to 4GB

- Sequence length

  - Complete search is unpractical

    for very long structures

- COMET availability

  - Available only via webserver

RNAiFold 2.0

- Implemented in C++ using OR-Tools (open source).
- Modular and scalable

Software	⇓	WS	PK	H	MT	PT	T	EM	D	SeqC	StrC	AaC	O	Num
RNAiFold 2.0	✓	✓	—	✓	—	✓	✓	'99,'04	0,1,2,3	✓	✓	✓	mfe	MAX
RNAinverse	✓	✓	—	—	—	—	✓	'99,'04	0,1,2,3	IUPAC⋆	—	—	mfe, prob	100
RNA-SSD	—	✓	—	—	—	—	✓	'99	1	IUPAC⋆	—	—	mfe	10
Info-RNA	✓	✓	—	—	—	—	—	'04	1	IUPAC	—	—	mfe, prob	50
NUPACK	✓	✓	—	✓⋆	—	—	✓	'99,'04	0,1,2	✓	—	—	ens def	10
MODENA	✓	—	✓	—	—	—	—	I	def	—	—	—	mfe, prob	?
Frnakenstein	✓	—	—	—	✓	—	✓	I	def	—	—	—	various	?
IncaRNAtion	✓	—	—	—	—	—	✓	'04⋆	—	IUPAC	—	—	pf sampling	—
ERD	✓	✓	—	—	—	—	✓	I	def	IUPAC⋆	—	—	mfe	MAX⋆
RNAdesign	✓	—	—	—	✓	—	✓	'04	def	—	—	—	various	—
RNAfbinv	✓	—	—	—	—	✓	—	'99, I	def	local A,C,G,U	—	—	mfe	—

- Partial target structure(**)
- Sequence constraints
  - GC-content (*)
  - Number of base pairs(***)
  - Maximum, minimum and consecutive nucleotides by region(***)
  - Structural compatibility and incompatibility constraints (***)
- Amino acid coding constraints (***)
  - Partial and overlapping regions
  - Restrict search to amino acids similar in BLOSUM62 similarity score
  - Maximization of BLOSUM62 similarity score

Supported by:
(*) Some other methods
(**) Only one other software
(***) RNAiFold only

JA Garcia-Martin, I Dotu, P Clote, (2015) RNAiFold 2.0 (Synthetic RNA Design): A web server to design RNA molecules (submitted).
Available at http://bioinformatics.bc.edu/clotelab/RNAiFold2.0

Applications (I): IRES discovery

Dotu I, Lozano G, Clote P, Martinez-Salas E. *Using RNA inverse folding to identify IRES-like structural subdomains. RNA Biol. 2013 Dec 1;10(12):1842-52.*

Applications(I): IRES discovery

Dotu I, Lozano G, Clote P, Martinez-Salas E. Using RNA inverse folding to identify IRES-like structural subdomains. RNA Biol. 2013 Dec 1;10(12):1842-52.

Applications(II): SECIS insertion

Model: formate dehydrogenase F (fdhF)
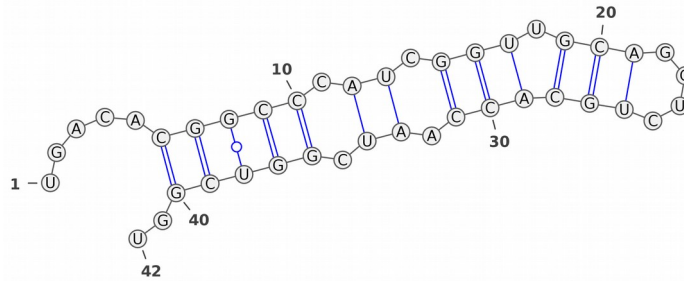
S. enterica (CDS70432.2)
UHGPSVAGLHQSVG
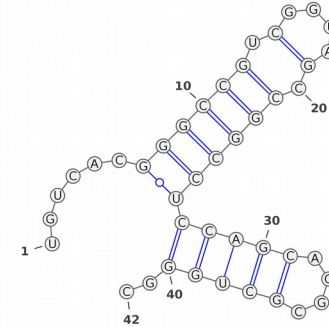
R. ornithinolytica (AJF73661.1)
CHGPSVAGLQQALG

RnaiFold 2.0 Constraints
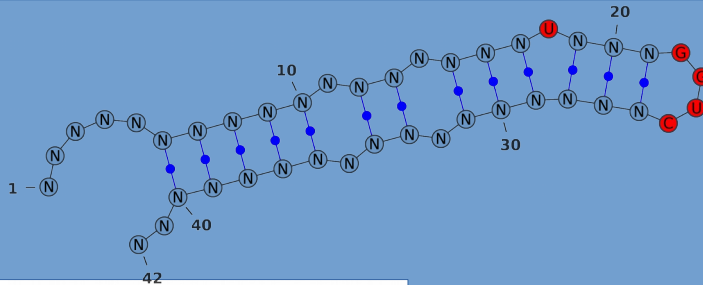
Amino acid target:
UHGPSVAGLQQALG
BLOSUM62 similarity → -1
Maximize BLOSUM62 score

Solution
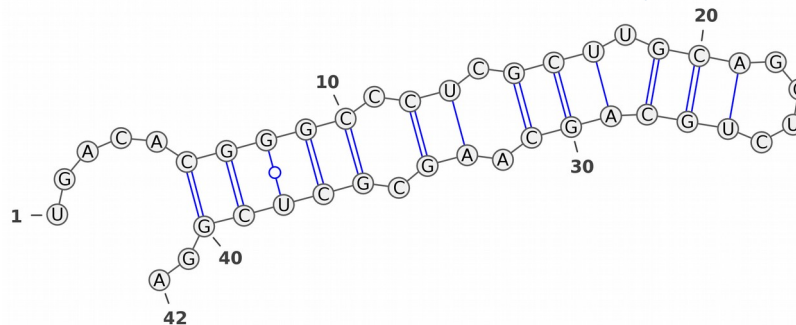
UHGPS**L**AGLQQALG
Score: 68 out of 71

Applications: Computational design of Hammerhead Ribozymes
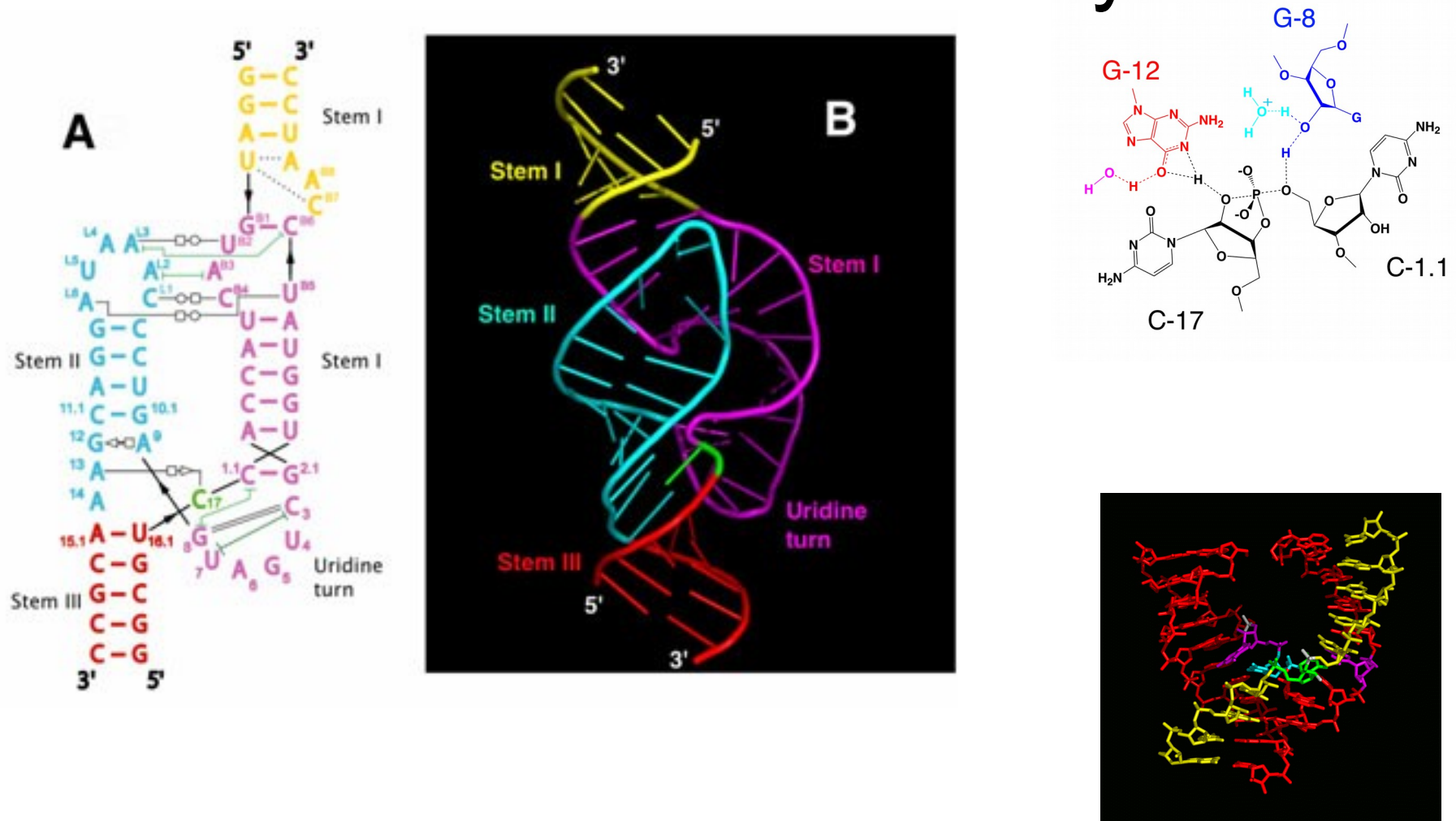
Figures: Monika Martick et al (2008),

1. Sequence conservation

Peach Latent Mosaic Viroid
(PLMVd) AJ005312.1/282-335
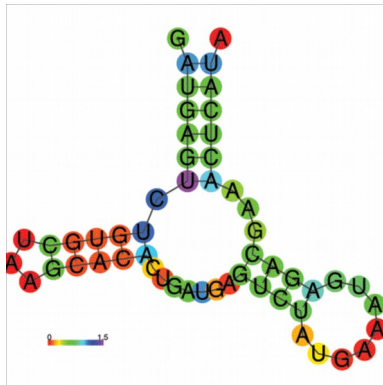
Turner '99   Turner '04
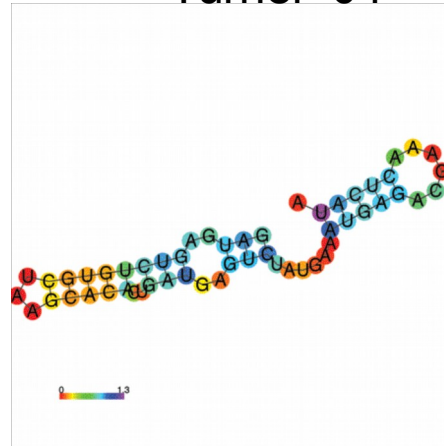
Rank	Position (1-54)	Nucleotide	Frequency	Percentage
1	7	U	1	100%
2	23	U	1	100%
3	27	G	1	100%
4	22	C	1	100%
5	48	A	1	100%
6	47	A	1	100%
7	28	A	1	100%
8	25	A	1	100%
9	24	G	0.988095	99%
10	46	A	0.988095	99%
11	6	G	0.987342	99%
12	45	G	0.97619	98%
13	49	C	0.97561	98%
14	29	G	0.964286	96%
15	44	C	0.964286	96%
16	38	A	0.957746	96%
17	8	C	0.949367	95%
18	35	G	0.942857	94%
19	42	G	0.928571	93%
20	31	C	0.892857	89%

Rfam logo of positions aligned to PLMVd
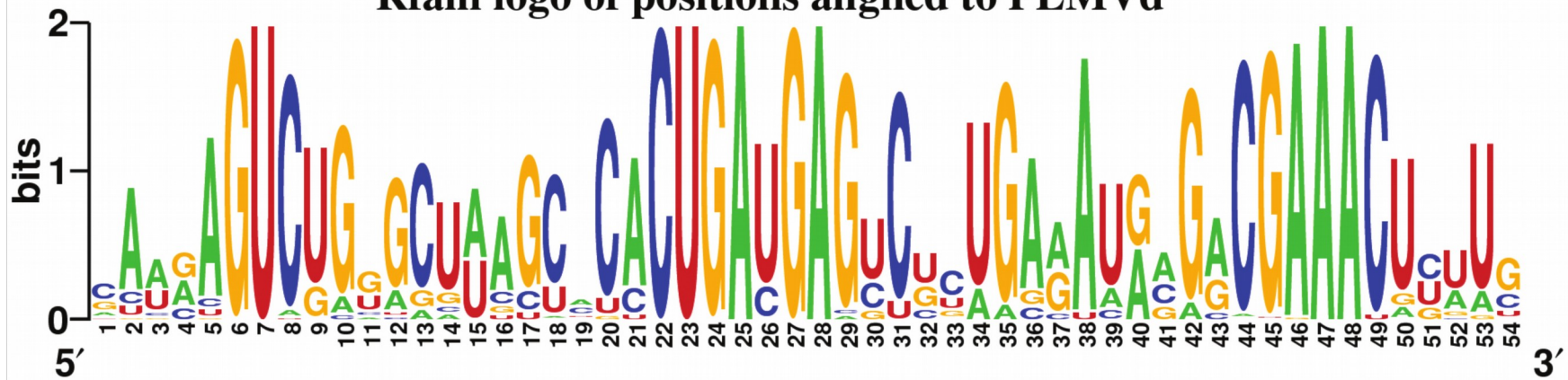
Dotu I, Garcia-Martin JA, Slinger BL, Mechery V, Meyer MM, Clote P.
Complete RNA inverse folding: computational design of functional hammerhead ribozymes. Nucleic Acids Res. 2015;42(18):11752-62.

2. Generating sequences
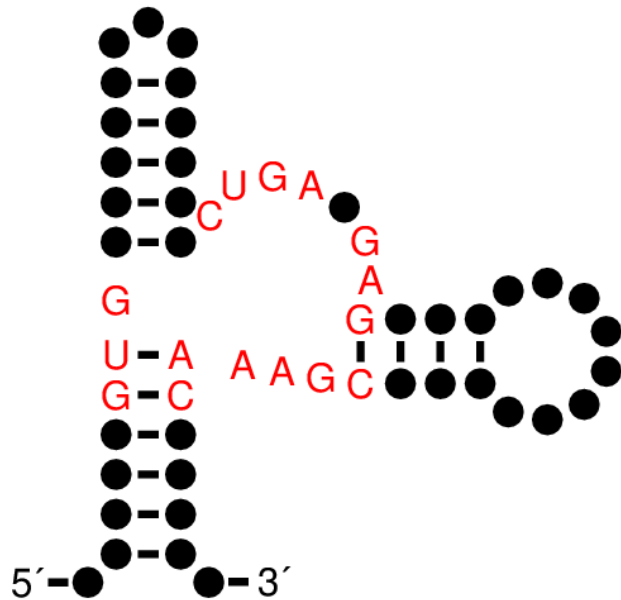
RNAiFold

No possible solution
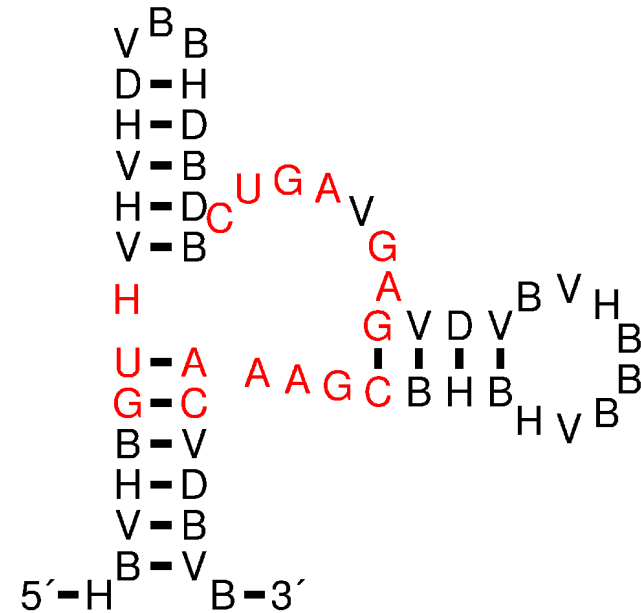
RNAiFold

~1.4 million sequences

Dotu I, Garcia-Martin JA, Slinger BL, Mechery V, Meyer MM, Clote P.
Complete RNA inverse folding: computational design of functional hammerhead ribozymes. Nucleic Acids Res. 2015;42(18):11752-62.

3. Filtering and selection

ID	Selection criteria
HH1	GC-content $30 - 39\%$, $P(S_0, s) \geq 40\%$, smallest (binary) entropy distance for active site
HH2	GC-content $40 - 49\%$, $P(S_0, s) \geq 40\%$, smallest (binary) entropy distance for active site
HH3	GC-content $40 - 49\%$, $P(S_0, s) \geq 40\%$, second smallest (binary) entropy distance for active site
HH4	GC-content $50 - 59\%$, $P(S_0, s) \geq 40\%$, smallest (binary) entropy distance for active site
HH5	GC-content $60 - 69\%$, $P(S_0, s) \geq 40\%$, smallest (binary) entropy distance for active site
HH6	GC-content $30 - 39\%$, $P(S_0, s) \geq 40\%$, largest (binary) entropy distance for active site
HH7	GUC in positions 6-8, smallest ensemble defect
HH8	smallest ensemble defect, entropy and expected bp distance and highest Boltzmann probability
HH9	$P(S_0, s) \leq 20\%$, smallest (binary) entropy distance for active site
HH10	smallest (binary) entropy distance for active site

Dotu I, Garcia-Martin JA, Slinger BL, Mechery V, Meyer MM, Clote P.
Complete RNA inverse folding: computational design of functional hammerhead ribozymes. Nucleic Acids Res. 2015;42(18):11752-62.

4. Experimental validation

Dotu I, Garcia-Martin JA, Slinger BL, Mechery V, Meyer MM, Clote P.
Complete RNA inverse folding: computational design of functional hammerhead ribozymes. Nucleic Acids Res. 2015;42(18):11752-62.
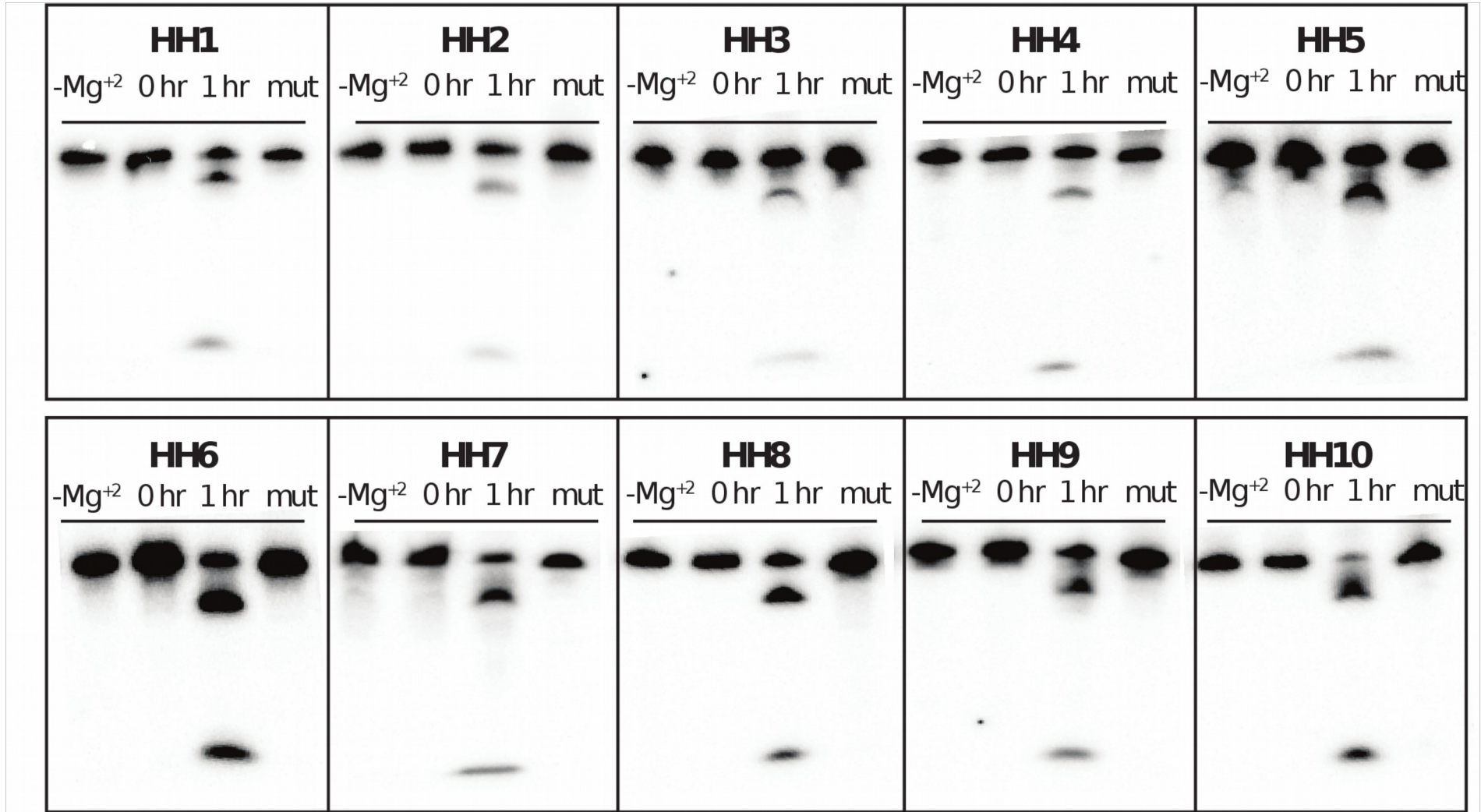
Kinetics
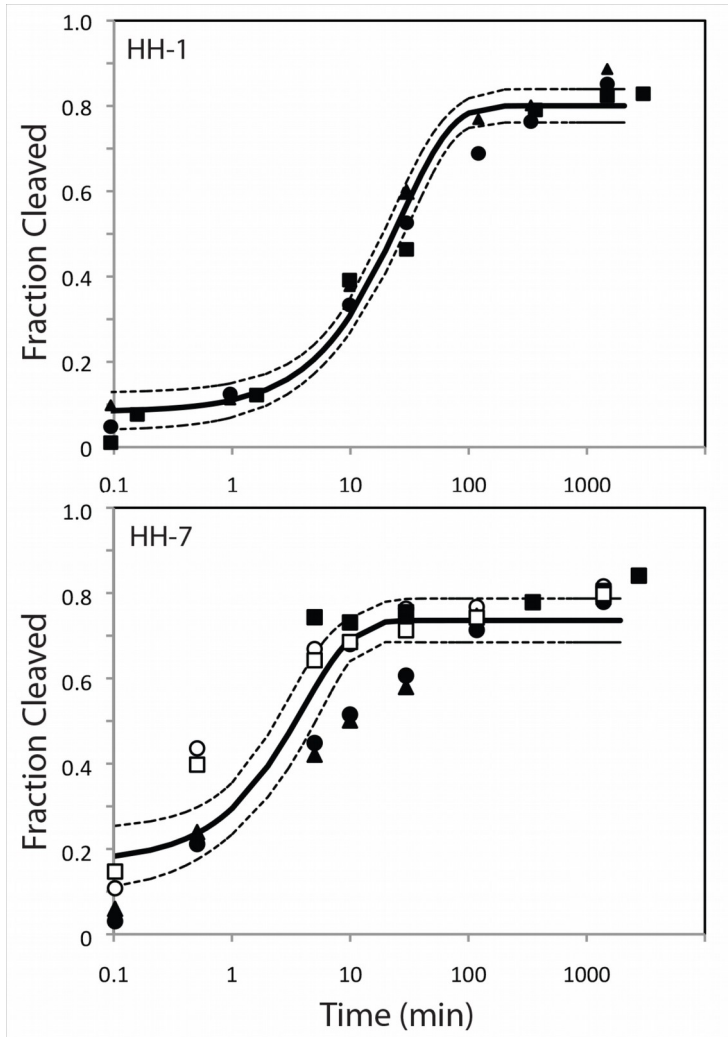
ID	K_{obs}	F_{max}	MSE	Pos Ent	Ens Def	EBPD Dis Act
HH1	0.037	0.79	0.0029	0.27	4.17	0.0501207
HH2	0.0057	0.74	0.003	0.29	4.55	0.0386253
HH3	0.0027	0.65	0.0039	0.26	4.12	0.0410984
HH4	0.0127	0.55	0.0048	0.40	6.76	0.0354213
HH5	0.0085	0.52	0.0066	0.38	6.24	0.033132
HH6	0.102	0.73	0.0047	0.41	8.14	0.059864
HH7	0.25	0.74	0.0107	0.12	2.38	0.0406728
HH8	0.02	0.68	0.0124	0.08	1.45	0.0662421
HH9	0.025	0.76	0.0015	0.25	4.53	0.0328018
HH10	0.14	0.77	0.01	0.29	4.98	0.0269354

Correlations: Pos. entropy (-0.46) Ens. defect (-0.37)

RNAiFold web server

J. A. Garcia-Martin, P. Clote, and I. Dotu. RNAiFold: a web server for RNA inverse folding and molecular design. Nucleic. Acids. Res., 41(Web):W465–W470, July 2013. Available at http://bioinformatics.bc.edu/clotelab/RNAiFold

RNAiFold 2.0 web server

JA Garcia-Martin, I Dotu, P Clote, (2015) RNAiFold 2.0 (Synthetic RNA Design): A web
server to design RNA molecules (submitted). Available at http://bioinformatics.bc.edu/clotelab/RNAiFold2.0

RNAiFold 2.0
Some ongoing work

- Optimize inverse folding
  - New search heuristics and structure decomposition strategies
  - Constraint evaluation order

- Add new functionalities
  - Design constraints
  - Incorporate target functions and energy models
  - Evaluate new measures of structural diversity

- Expand objective functions (e.g. ensemble defect)
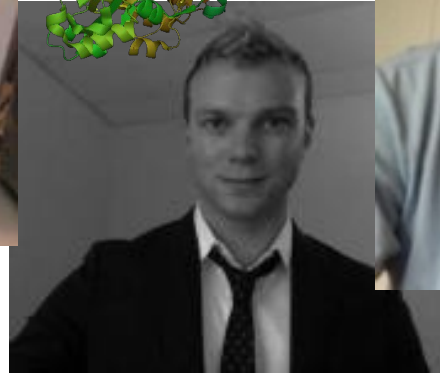
ONGOING WORKS

Conclusions

- Advantages of using a CP approach
  - Scalability
  - Flexibility for constrained design
  - Completeness (can also determine that NO solution exists)
- Applications of RNAiFold
  - Properties of the sequence ensemble
  - Discovery of functional RNAs
  - Synthetic design

Thanks!!

Peter Clote

Ivan Dotú

Amir Bayegan

Evan Senter

NSF DBI-1262439

BOSTON COLLEGE

The Meyer Lab

Betty Slinger Michelle Meyer