RNA profiling: Extracting structural signals from noisy distributions

Emily Rogers and Christine Heitsch
School of Mathematics
Georgia Institute of Technology

Discrete mathematical biology

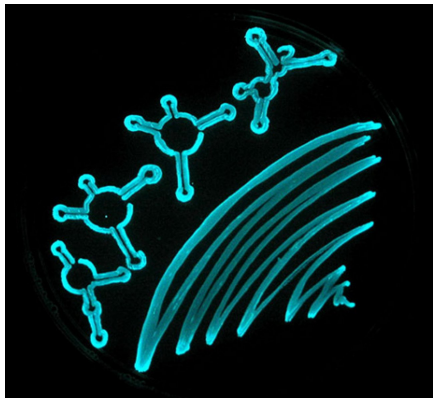
Combinatorics. . .

"as motivated by"

1, 1, 2, 5, 14, 42, 132, 429, 1430,
4862, 16796, 58786, 208012,
742900, 2674440, 9694845,
35357670, 129644790,...
(OEIS A000108)

and

1, 1, 2, 8, 42, 262, 1828, 13820,
110954, 933458, 8152860,
73424650, 678390116,
6405031050, 61606881612,...
(OEIS A005315)



Henke & Bassler, Princeton.

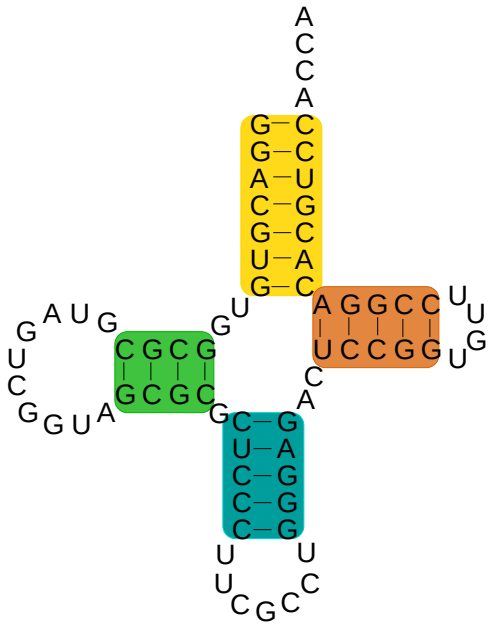
"with applications to"

. . . molecular biology.

RNA sequences fold via base pairings

Primary sequence $\longrightarrow$ secondary structure $\longrightarrow$ 3D molecule

GGGCGUAUGGCG
CGUAGUCGGUAG
CGCGCUCCCUUC
GCCUGGGAGACU
CCGGUGUUCCGG
ACACGUCCACCA



RNA secondary structures balance energetically favorable helices (consecutive base pairs) against destabilizing loops (single-stranded bases).

Predicting MFE secondary structures

Dynamic programming solves discrete optimization efficiently, but quality of free energy approximation by the NNTM (nearest neighbor thermodynamic model) objective function varies widely.

Abbreviation	Sequence	Length (nt)	MFE accuracy
T1	*H. sapiens* (AC004932_g)	72	0.00
T2	*S. tokodaii* (BA00002_e)	74	0.26
T3	*S. tokodaii* (BA000023_b)	74	0.45
T4	*L. delbrueckii* (CP000412_o)	72	0.75
T5	*O. nivara* (AP006728_af)	73	1.00
S1	*E. coli* (V00336)	120	0.26
S2	*G. arboreum* (U31855)	120	0.47
S3	*A. tabira* (AB015591)	120	0.59
S4	*S. cerevisiae* (X67579)	118	0.71
S5	*D. mobilis* (X07545)	135	0.88

Ill-conditioning of the NNTM optimization

". . . raises some questions for biologists and biochemists as well as for mathematicians." M. Zuker, **1986**, *RNA folding prediction: the continued need for interaction between biologists and mathematicians.*

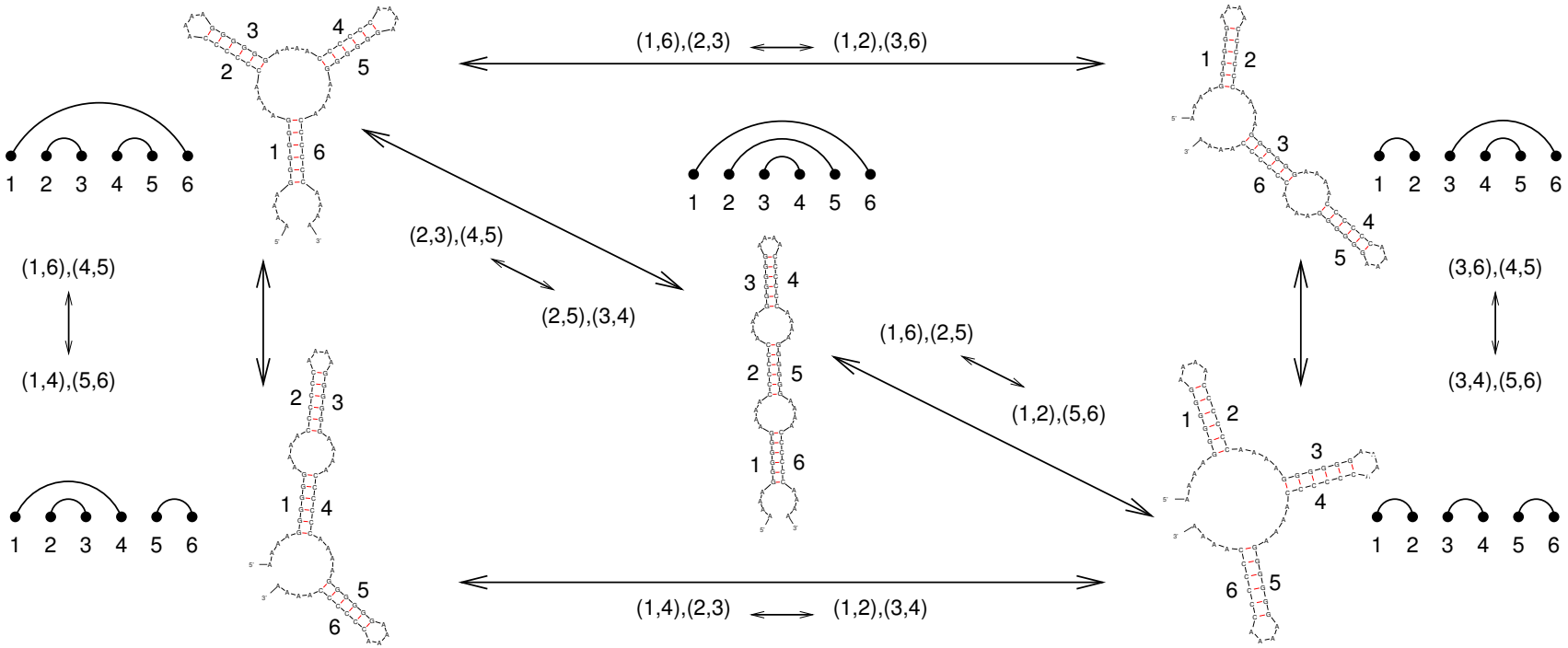
1. "Tradeoff between biochemical reality and quick and efficient folding algorithms."

2. "**Accept multiple solutions** or else be prepared to supply auxiliary information."

3. "Hope(d) that more sophisticated energy rules will become available in the future."

(2a) Accept multiple solutions (bio $\rightarrow$ math)

Combinatorial understanding of RNA configuration landscapes?

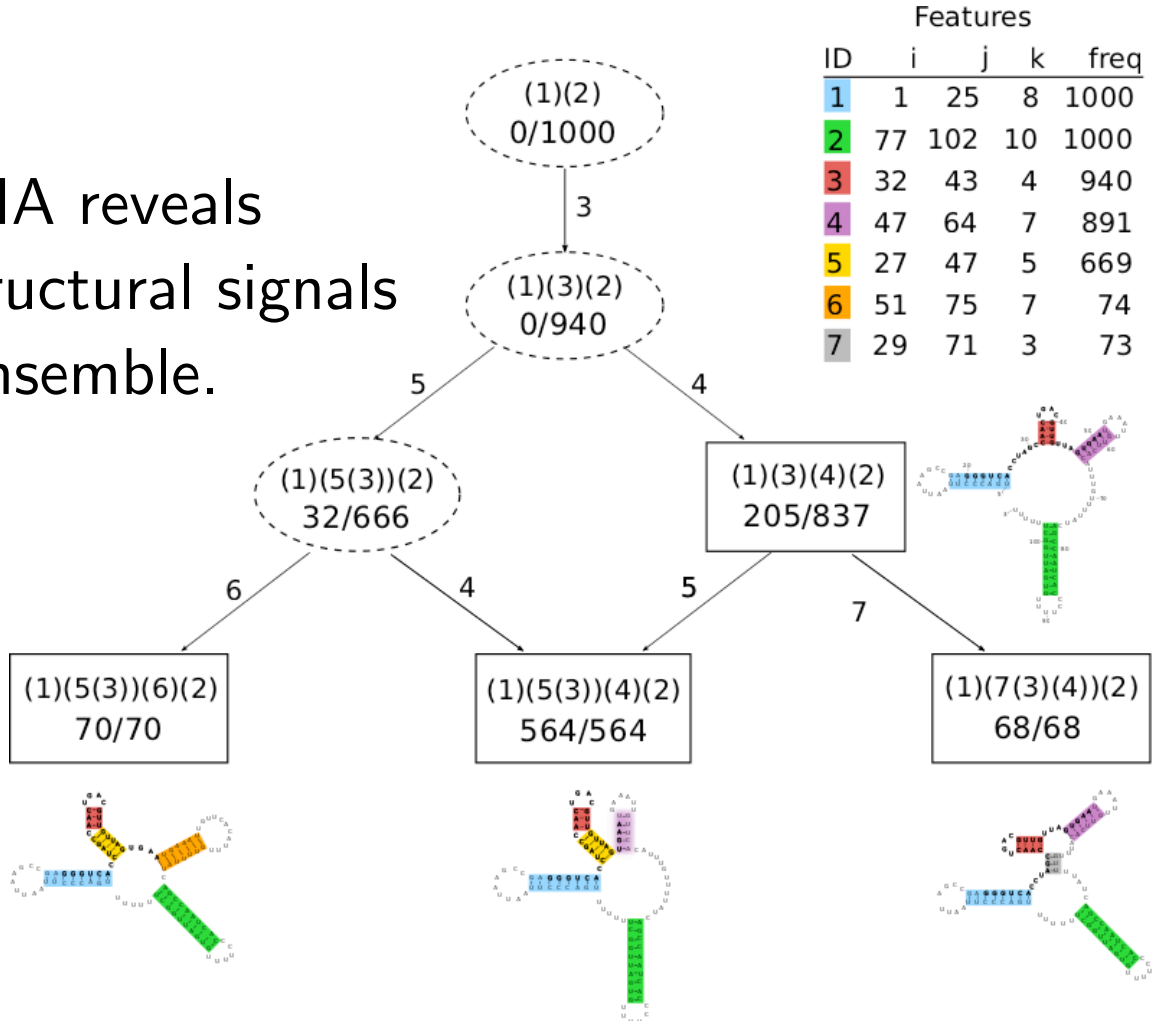


Saturated secondary structures for the RNA sequence $R = ({\rm A}^4{\rm G}^6{\rm A}^4{\rm C}^6)^4{\rm A}^4$.

$\Longrightarrow$ Results on Kreweras complementation and meander graphs.

(2a) Accept multiple solutions (math → bio)

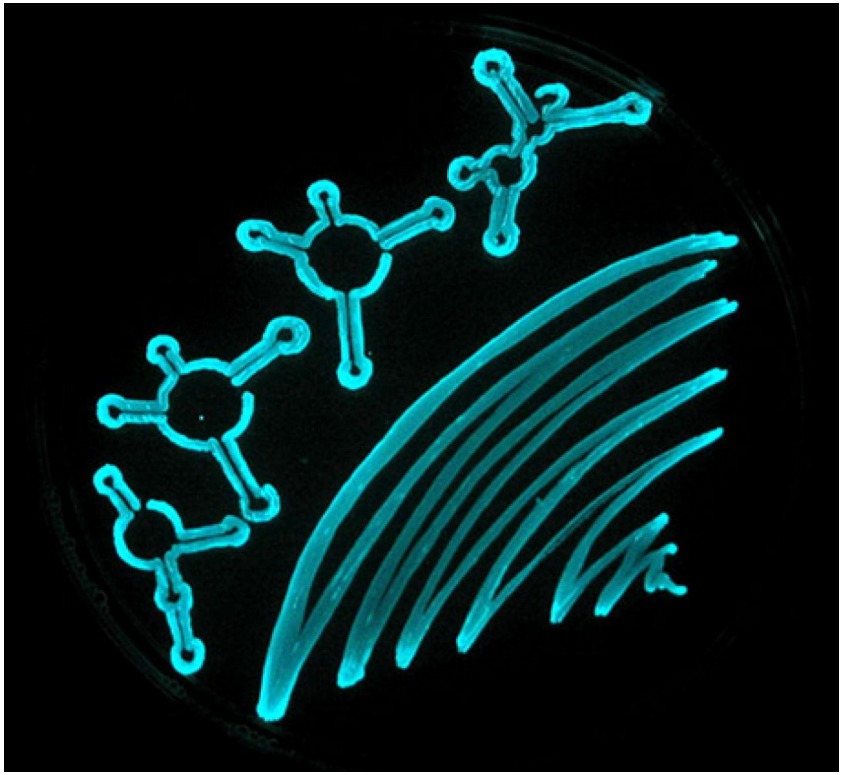
Profiling small RNA reveals multimodal substructural signals in a Boltzmann ensemble.



Rogers & Heitsch, Nucleic Acids Res, 2014.

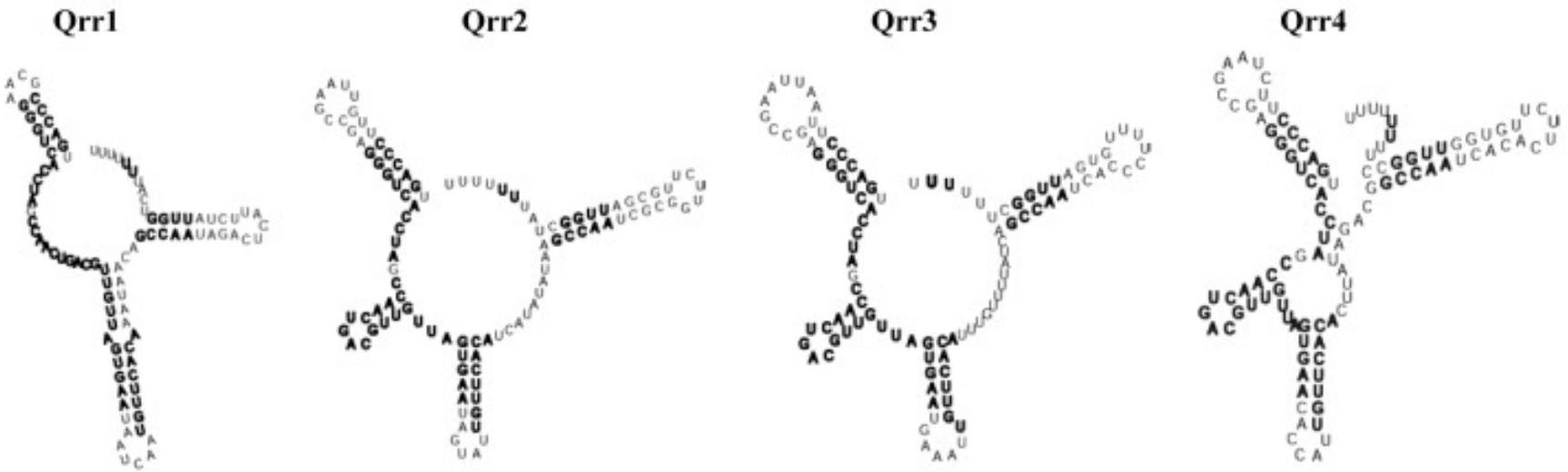
It's a small RNA world after all?

The four *Vibrio cholerae* quorum regulating RNA (VcQrr 1–4) predicted minimum free energy (MFE) secondary structures drawn on a petri dish.
J Henke and B Bassler, Princeton.



Many bacteria regulate behavior (like marine bioluminescense or cholera virulence) via quorum sensing using small RNA molecules.

Sequence alignment reveals functional signal

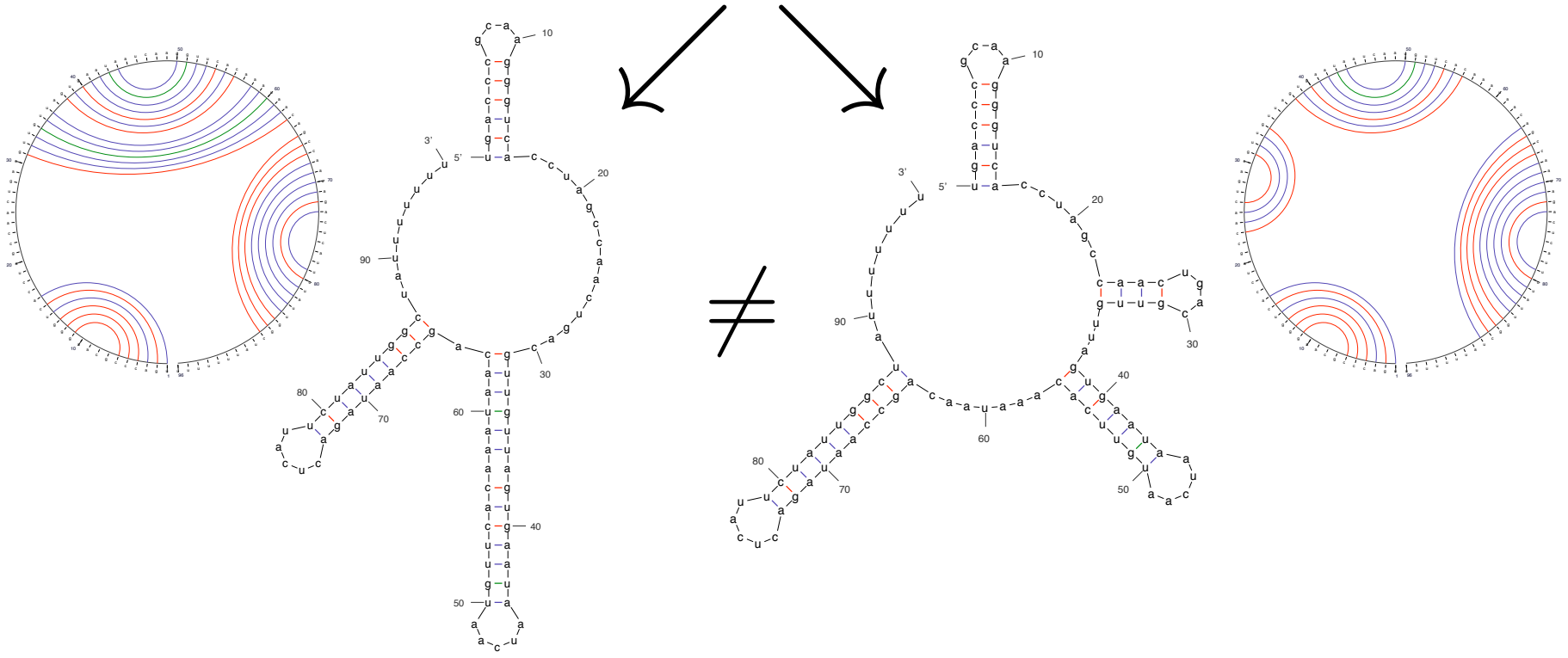


`RNAfold` MFE structures. Lenz, Mok, Lilley, Kulkarni, Wingreen & Bassler, 2004.

Conservation (bolded) strongly implies functional significance, yet MFE predictions vary and native structure(s) are not yet known.

Challenge: too many suboptimal folds

Vibrio cholerae Qrr 1 = **ugaccc gcaa ggguca ccuagc caac ugac guug uua gugaaua aucaa uguucac aaauaac agccaauaga cucau ucuauuggcu auuuuuuu**



$\Delta G(S_1) = -37.5$ kcal/mol $\Delta G(S_2) = -35.2$ kcal/mol

Ensemble analysis improves over MFE

Gibbs/Boltzmann sampling: Stochastically generate suboptimal structures s (foldings with ΔG close to MFE) according to

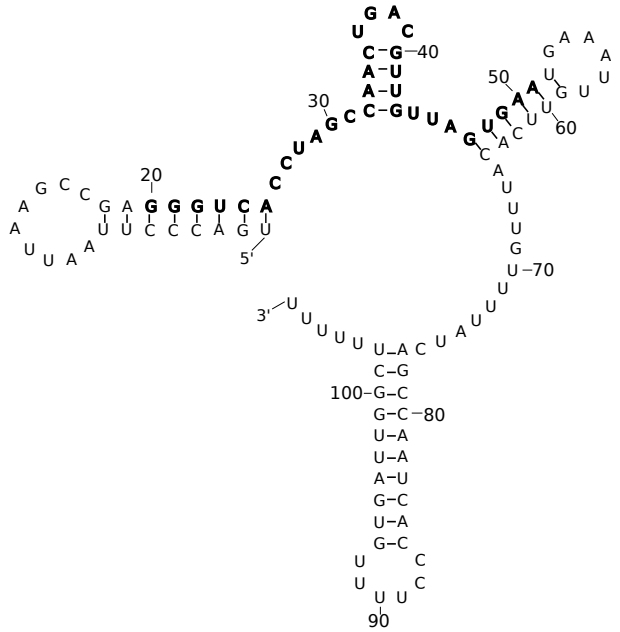
$$P(s) = \frac{e^{-\Delta G(s)/kT}}{Z}$$

McCaskill, 1990: Partition function $Z = \sum_s e^{-\Delta G(s)/kT}$ and BP probabilities can be calculated efficiently by dynamic programming.

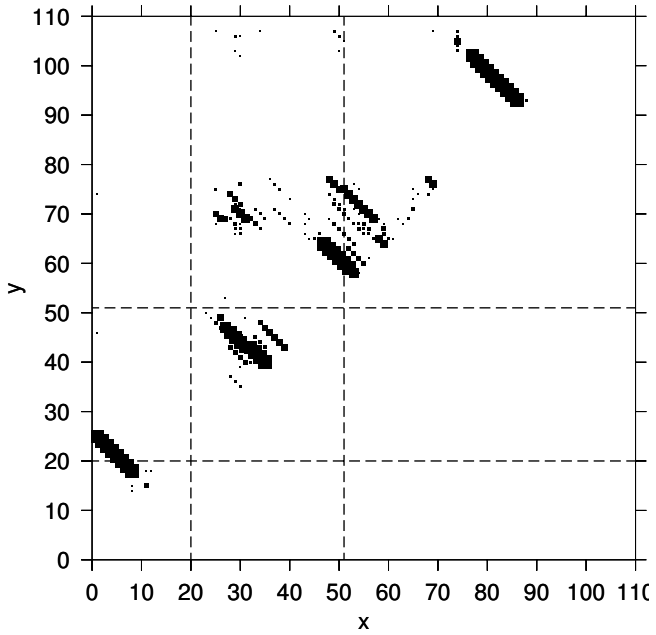
Ding & Lawrence, 2003: Sampling algorithm given and implemented in `Sfold`. Cluster results and report centroids.

Voß, Giegerich, and Rehmsmeier, 2006: Sampling implemented in `RNAshapes`. Report shapes and their shrep.

BP probabilities reveal well-determined regions



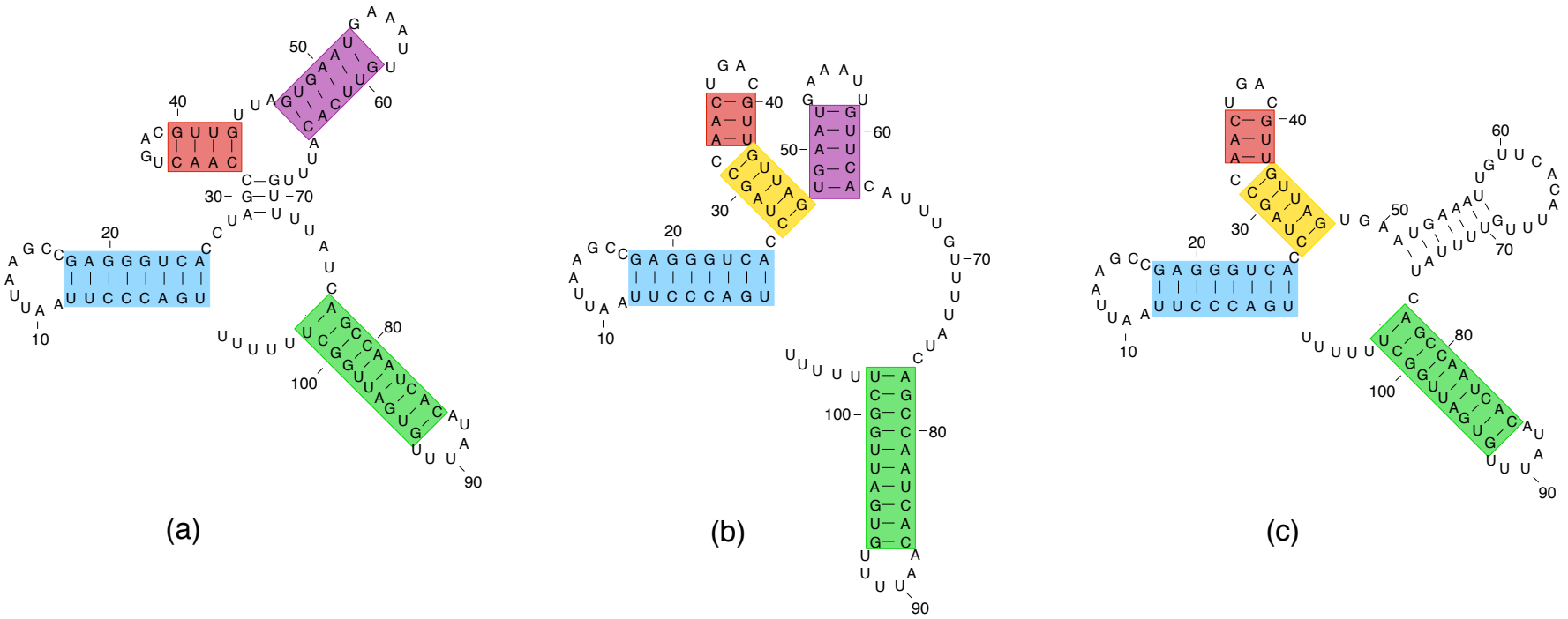
VcQrr3 MFE stucture



Probability dot plot

High probability helices correlate with native pairings,
but conserved region has competing structural alternatives.
How to find the (sub)structural signal?

Making sense of 1000 VcQrr3 structures



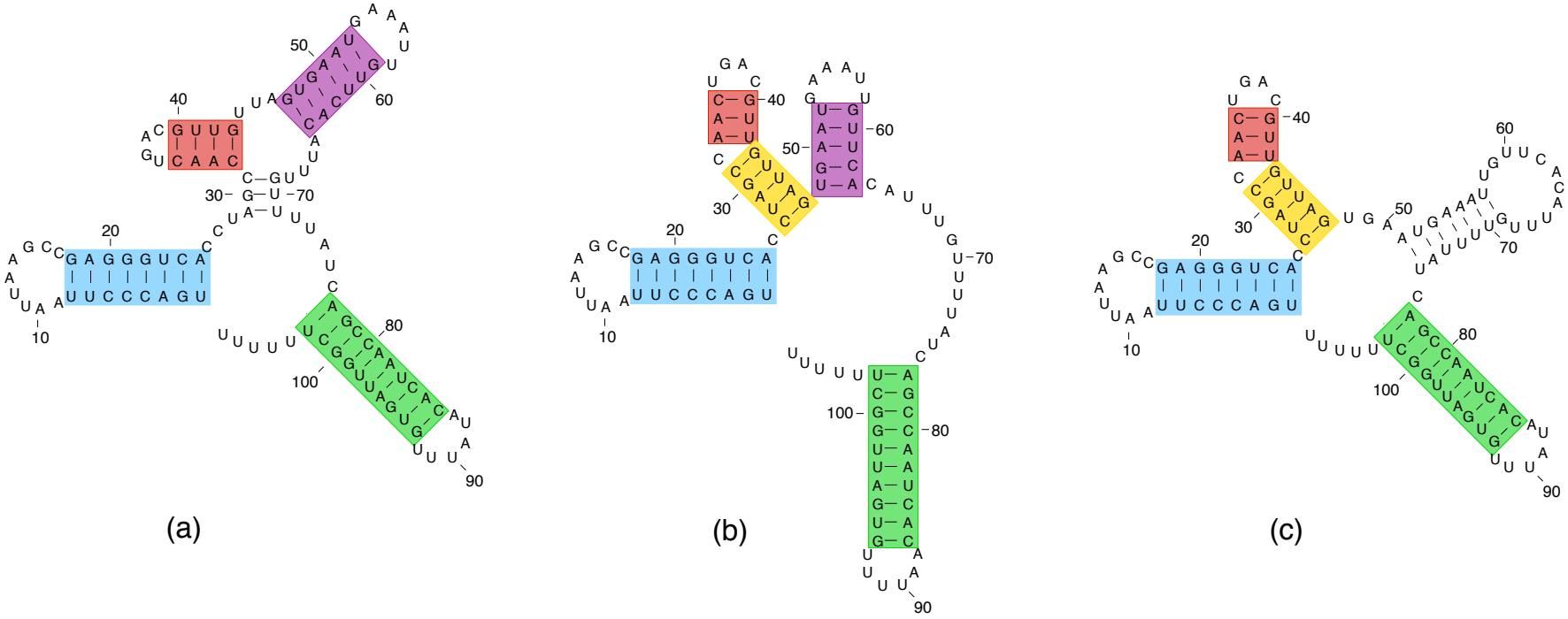
`Sfold` clusters b with a but not c. (BP metric, DIANA, max CH index.)

`RNAshapes` groups b with c but not a. ([][[]][][] versus [][[][]][].)

Our new method recognizes each structure as belonging to a distinct profile, while clearly indicating common base pairings.

Helix class as basic structural unit

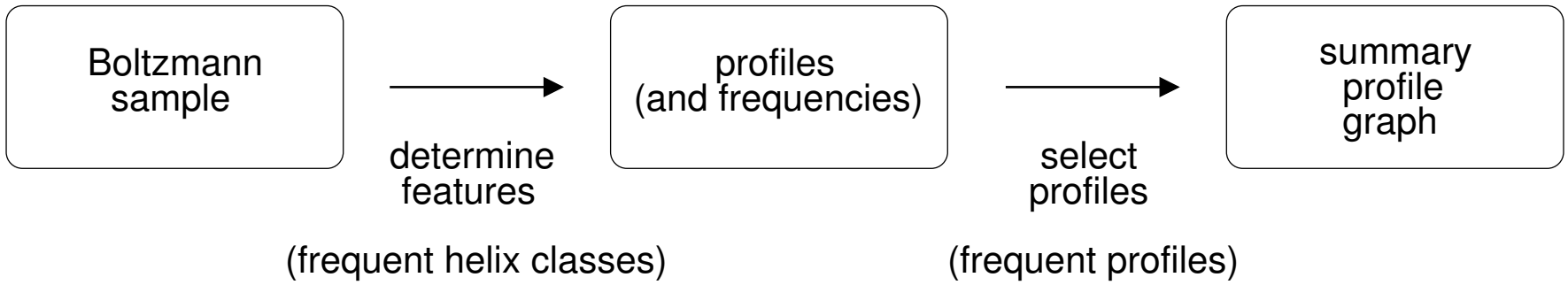
A helix $h = (i, j, k)$ is *maximal* if it is non-extendable (if either $(i-1, j+1)$ or $(i+k, j-k)$ are non-canonical base pairs or $j - i - 2k < 2$).



A helix class $[h]$ is the set of all helices within a maximal one h;
$(33, 42, 3) \in [(32, 43, 4)] =$ ▮ and $(48, 63, 5), (47, 64, 6) \in [(47, 64, 7)] =$ ▮.

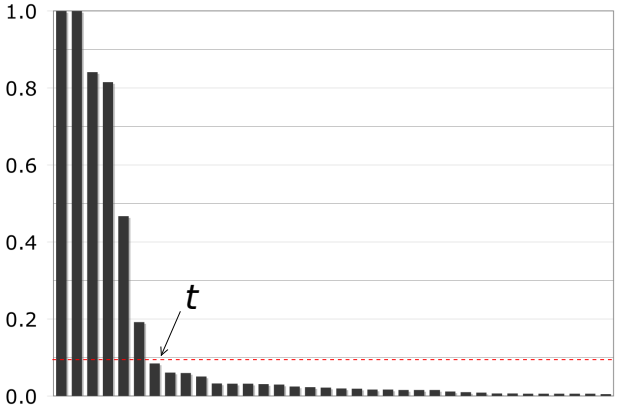
Profiling based on helix classes

- Given Boltzmann sample (typically 1000 structures).
- Generate helix classes in sample.
- Determine *featured* helix classes based on frequency.
- Generate *profiles* = maximal combinations of features.
- Select profiles based on frequency.
- Generate summary profile graph.



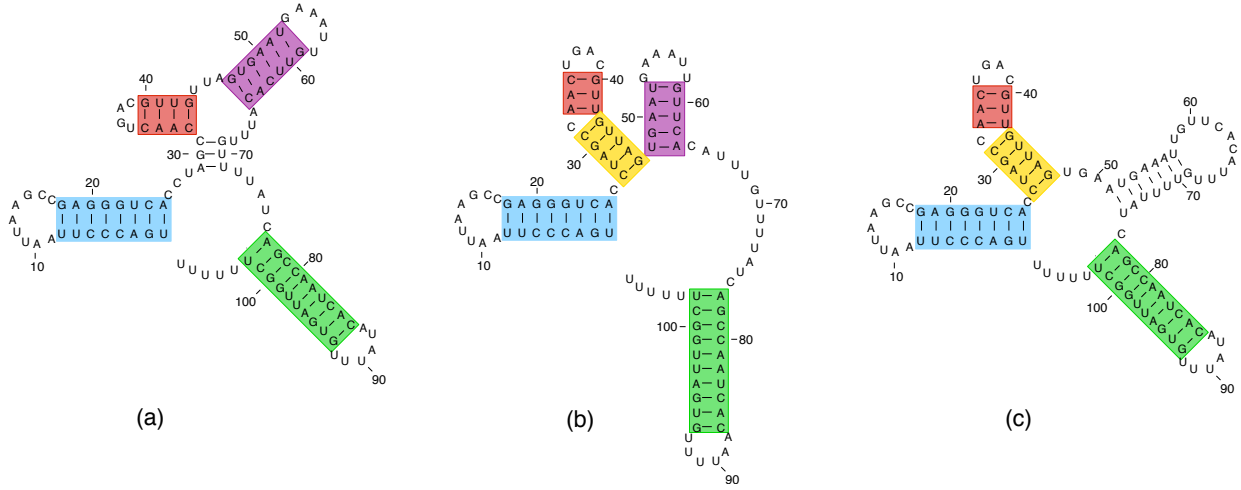
Determining VcQrr3 features & profiles

Observed $|[h]| \geq 0.05$



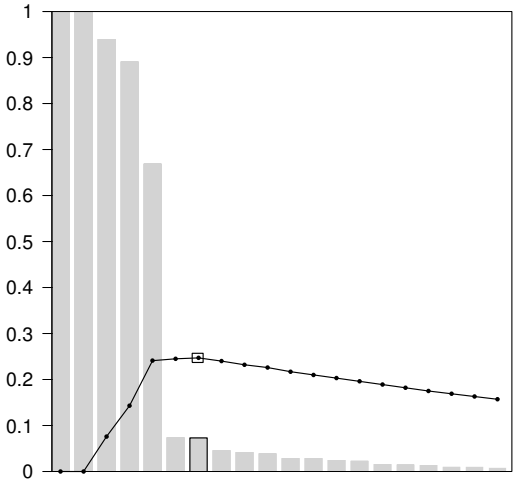
HC	ID	(i, j, k)	Freq
	1	$1, 25, 8$	1000
	2	$77, 102, 10$	1000
	3	$32, 43, 4$	940
	4	$47, 64, 7$	891
	5	$27, 47, 5$	699
	6	$51, 75, 7$	74
	7	$29, 71, 3$	73

Order $[h]$ by frequency $f(h) = |[h]|$; truncate long, noisy tail.



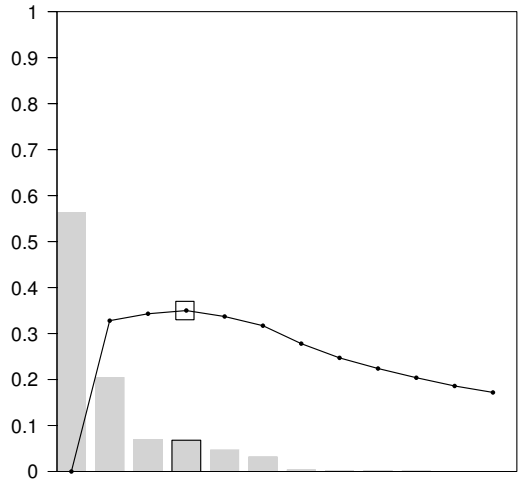
profiles: a = (1)(7(3)(4))(2); b = (1)(5(3))(4)(2); c = (1)(5(3))(6)(2);

Information theoretic threshold selection



Helix classes

Probability histograms. 194 helices in 1000 sampled structures. 20 of 88 helix classes shown (prob of 20th is 0.8%). All 13 profiles (last 7 have frequency < 5). Threshold at 7th helix class and 4th profile.



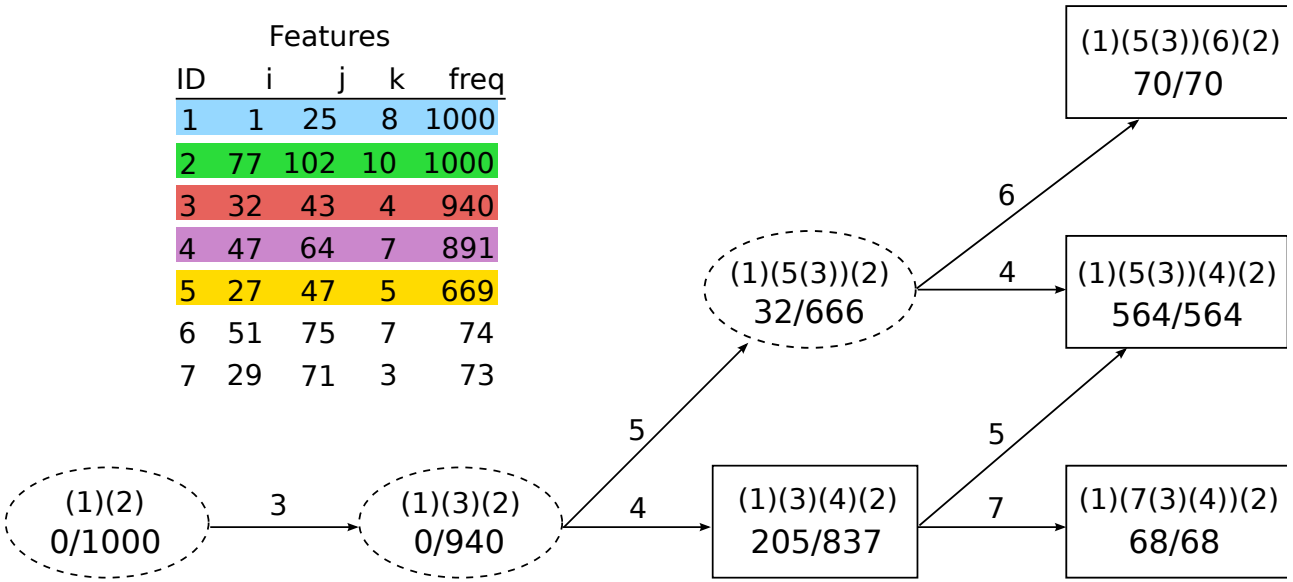
Profiles

Order structural units u by decreasing frequency $f(u)$.
Remove noise of low probability pairings by thresholding.
Use maximum average entropy, normalized by most probable $f(u_1)$;
$p(X_u) = f(u)/f(u_1)$ if $X_u = 1$ and $1 - f(u)/f(u_1)$ if $X_u = 0$.
$H(X_u) = -\sum_{x=0,1} p(x) \log\, p(x)$, $t = \mathrm{argmax}_k \frac{1}{k} \sum_{i=1}^{k} H(X_{c_i})$.

VcQrr3 summary profile graph

Features

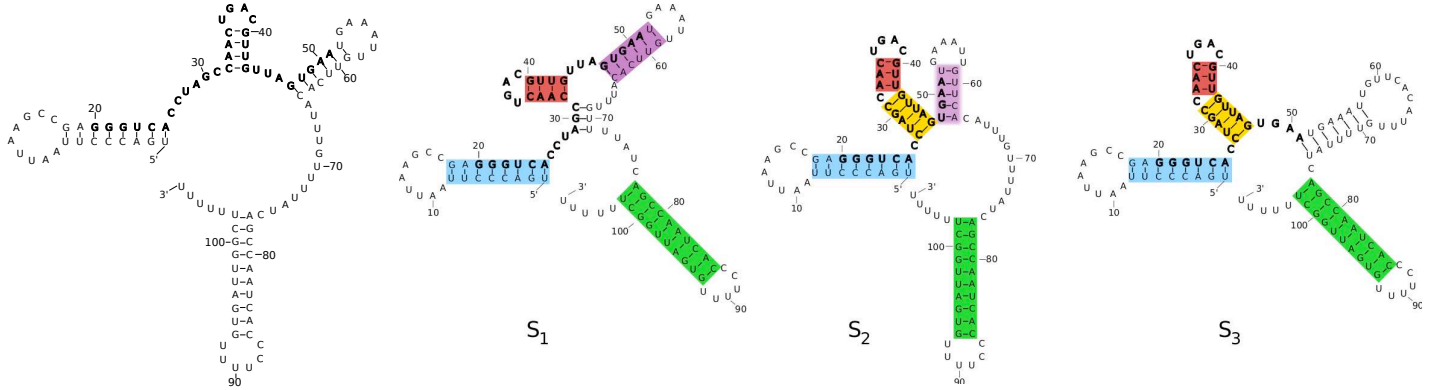
ID	i	j	k	freq
1	1	25	8	1000
2	77	102	10	1000
3	32	43	4	940
4	47	64	7	891
5	27	47	5	669
6	51	75	7	74
7	29	71	3	73



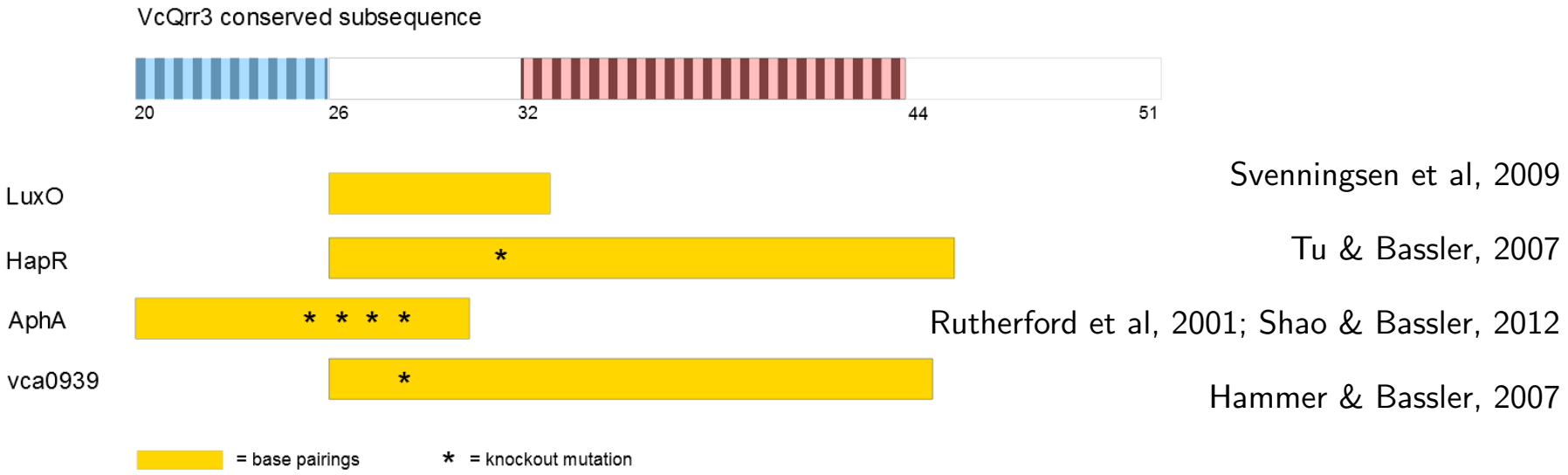
Hasse diagram of selected profiles; dashed ovals are "intersection" profiles added for completeness.

Ratios give specfic frequency (size of profile) over general frequencey (number of all structures in sample containing this set of features).

A multimodal substructural signal revealed



Key variable region overlaps with four known VcQrr targets!



Profiling extracts signal from noisy samples

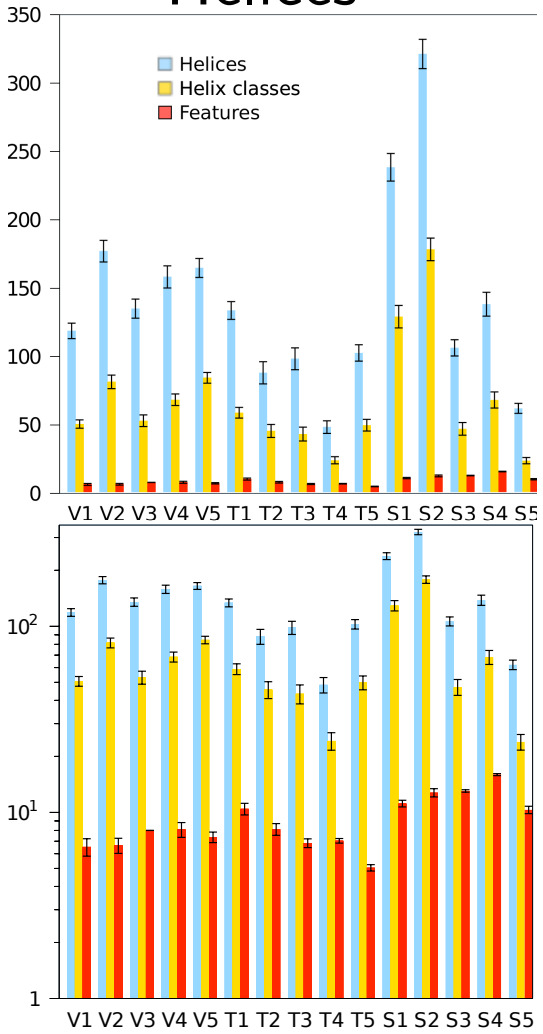
Abbr	Organism	Len	Acc
V1	*V. cholerae*	96	–
V2	*V. cholerae*	107	–
V3	*V. cholerae*	107	–
V4	*V. harveyi*	95	–
V5	*V. harveyi*	107	–
T1	*H. sapiens*	72	0.00
T2	*C. diphtheriae*	74	0.28
T3	*S. tokodaii*	74	0.45
T4	*L. delbrueckii*	72	0.75
T5	*O. nivara*	73	1.00
S1	*E. coli*	120	0.26
S2	*G. arboreum*	120	0.47
S3	*A. tabira*	120	0.59
S4	*R. norvegicu*	121	0.76
S5	*D. mobilis*	133	0.88

(Minimal) webserver:

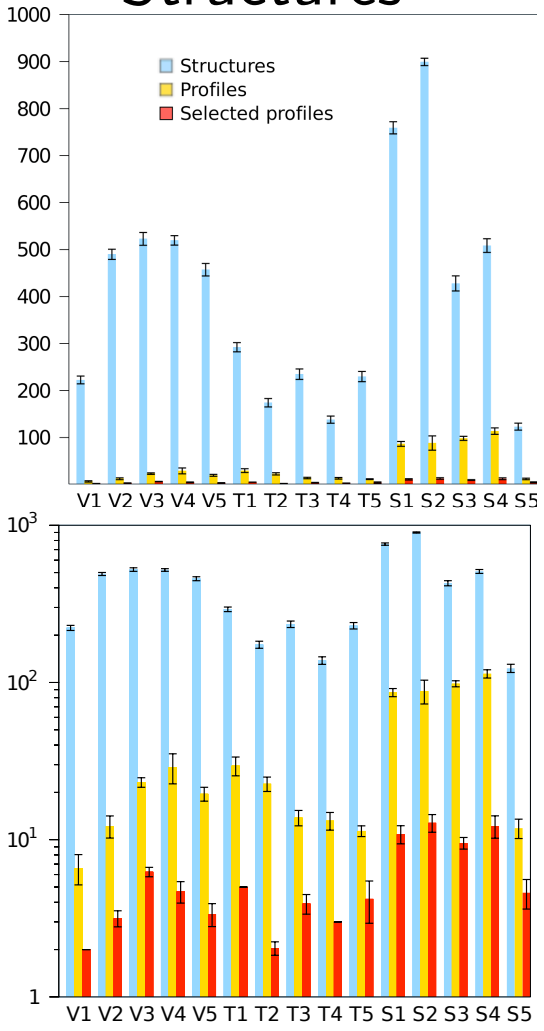
`rnaprofiling.gatech.edu`

`RNAStructProfiling` on GitHub/GTfold

Helices



Structures

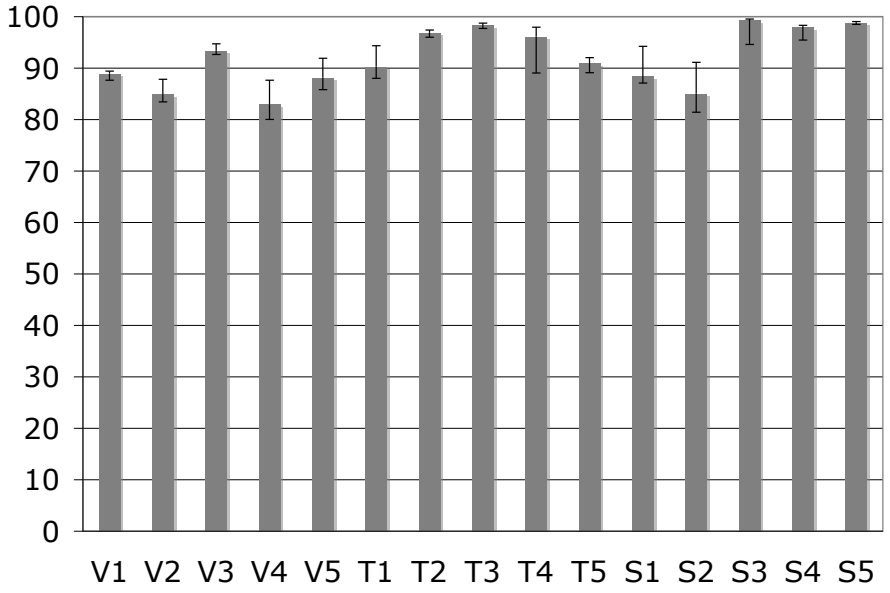


Coverage is not sacrificed for stability

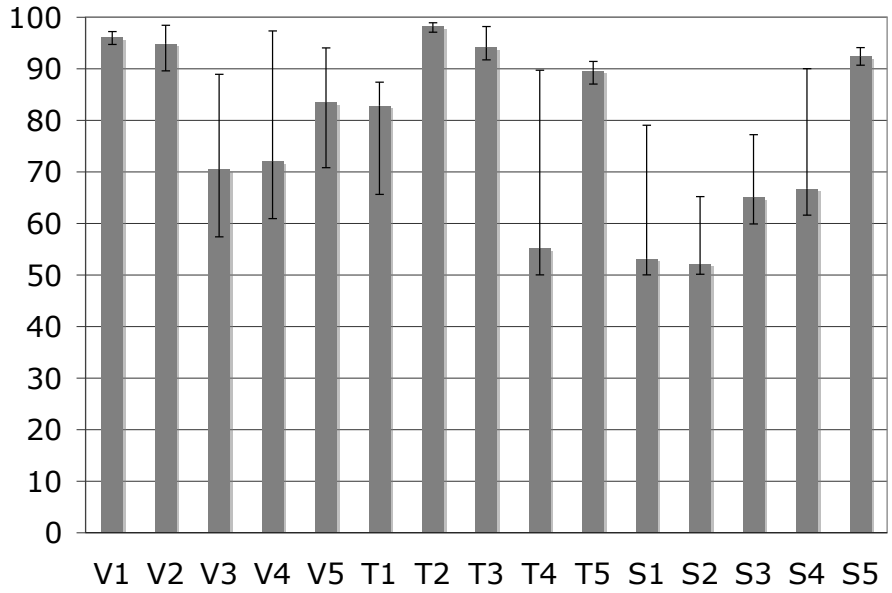
Profiling separates "signal" of high probability combinations of base pairs from "noise" of low probability helices.

Evaluate coverage by percentage of helices/structures in sample with multiplicity (!) included in features/selected profiles.

% Helices in features



% Structures in selected profiles



Multiple evaluation criteria

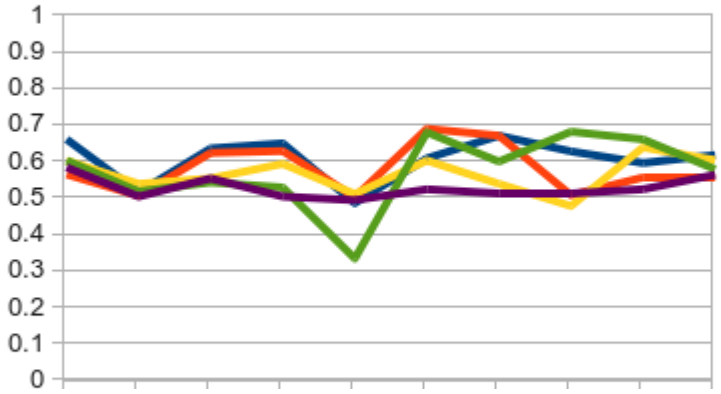
- Informative — highlight similarities/differences between groups
- Transparent — easily assign new structure to appropriate group
- Coherent — within group similarity higher than between group

- Reproducible — low variation in groupings between samples
- Complete — good coverage by summarizing majority of sample

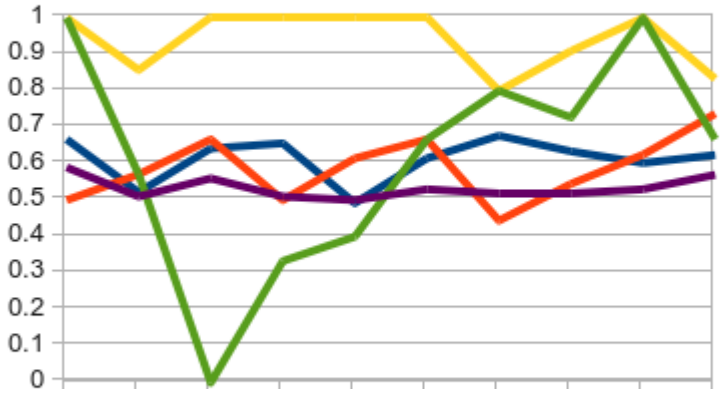
- Efficient — profiling faster than clustering (x3) and shapes (x6)
- Accurate — all three comparable, correlated with MFE accuracy

Accuracy comparison (Rogers & Heitsch, WIRE, 2016)

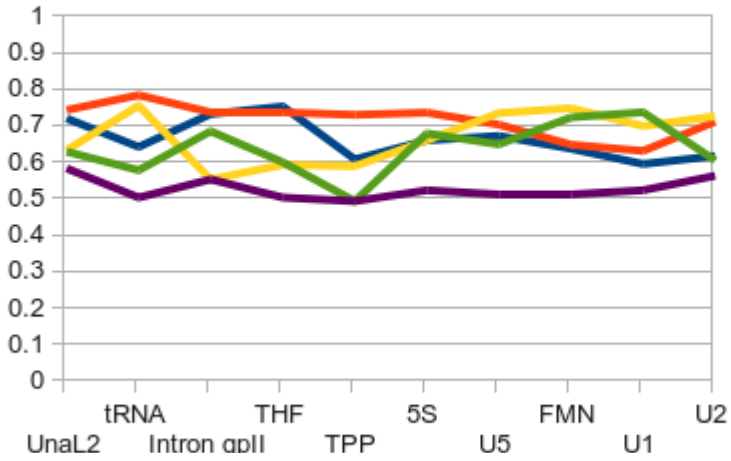
Most probable rep. structure



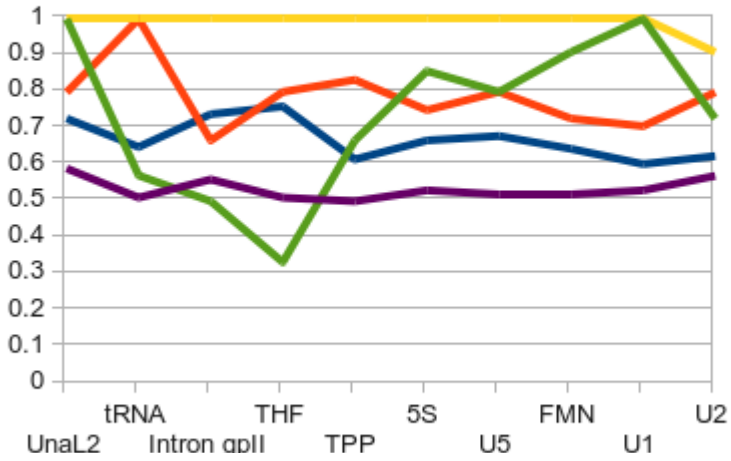
Most probable signature



Best rep. structure



Best signature

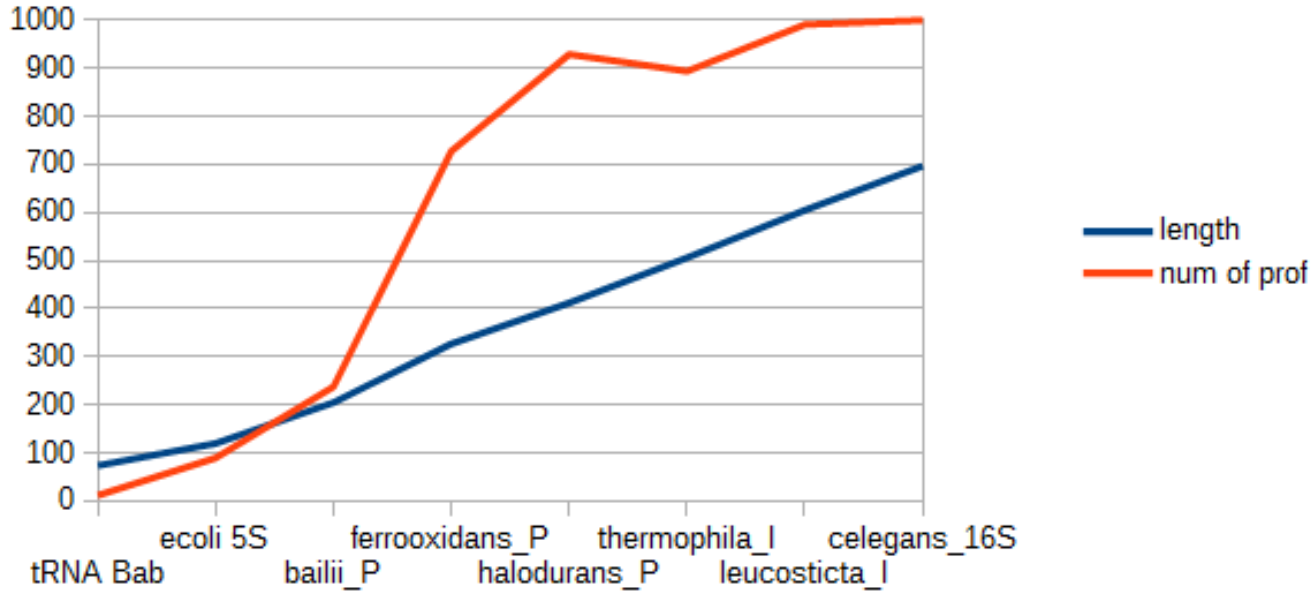


Sfold (blue); Profiling (red); Shape (yellow); Hishape (green); MFE (plum)

Punchline and caveat

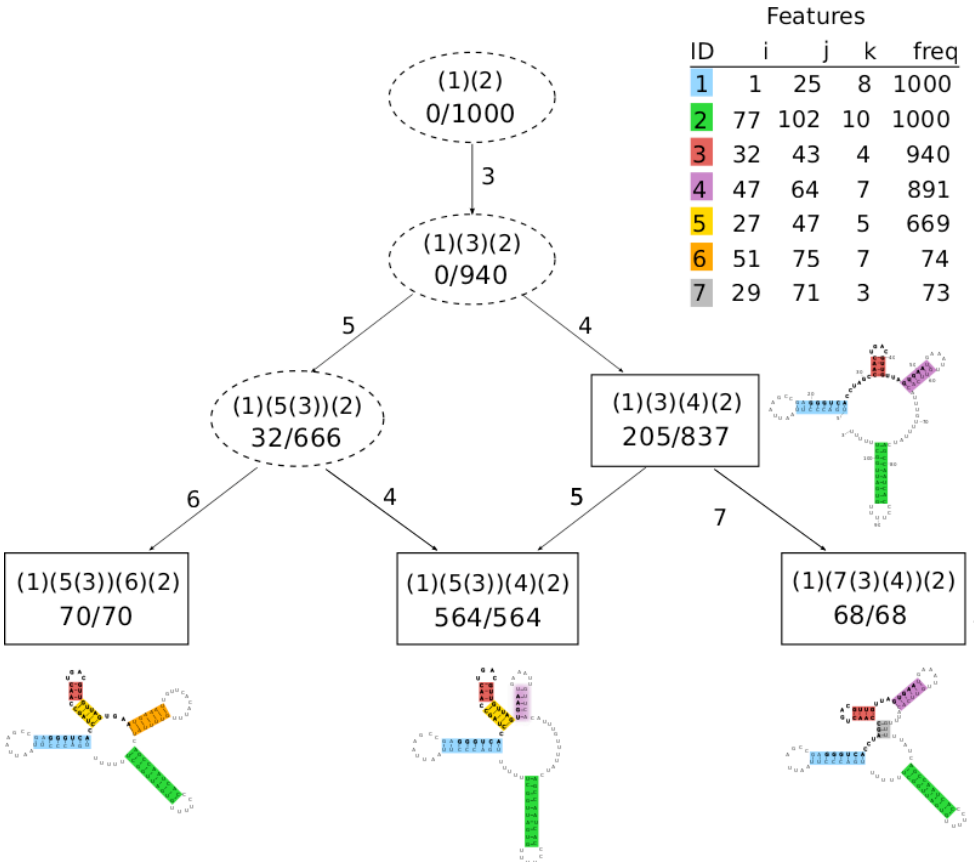
Structural signal from Boltzmann ensemble mined at different granularity (Sfold/bp; profiling/helices; shape/branching).

Abstraction reduces errors (systematic & random). But. . .



. . . utility limited to $\sim$ 200 nucleotides by NNTM ill-conditioning!

In the meantime . . .



Minimal webserver: `rnaprofiling.gatech.edu`

`RNAStructProfiling` on Sourceforge or GitHub