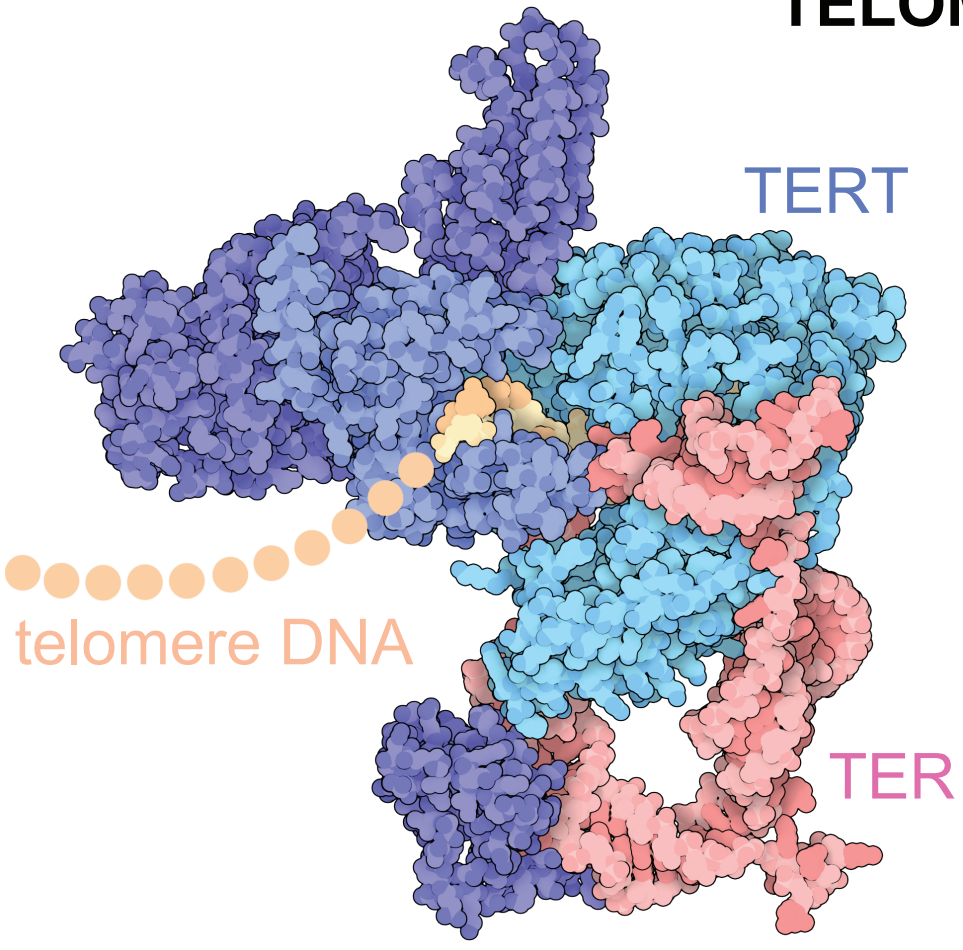
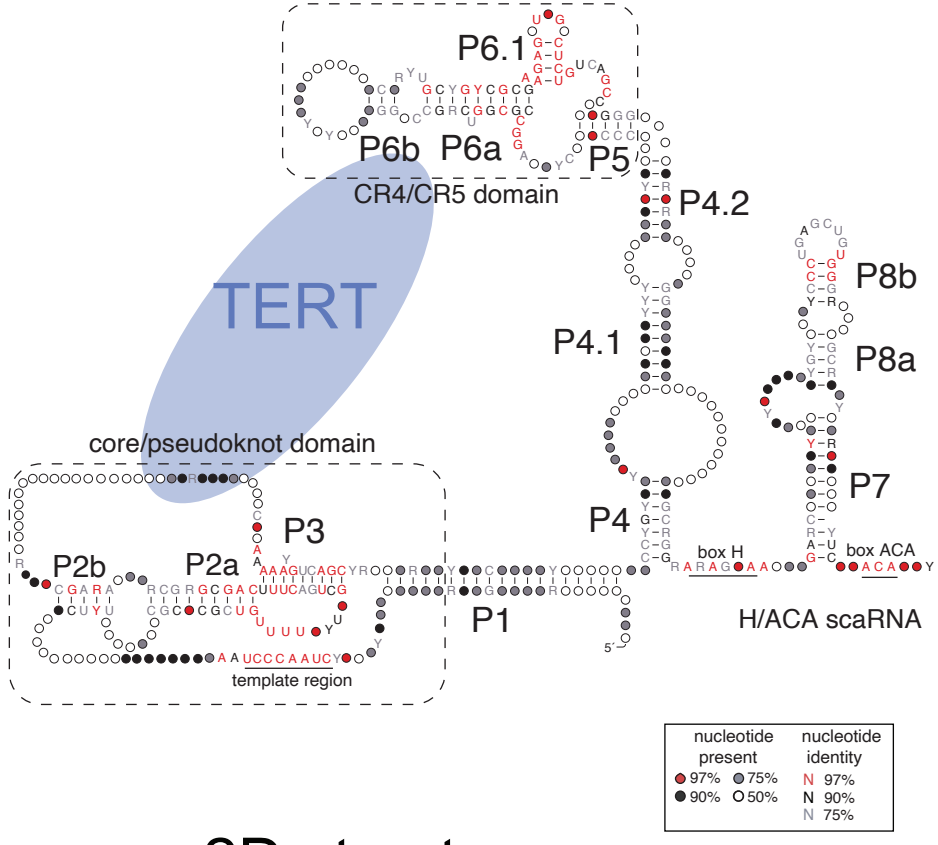
RNA structure prediction using positive and negative evolutionary information

E Rivas, Harvard University
rivaslab.org

Many functional RNAs have conserved structures

TELOMERASE





3D structure (4.8 Å)
Tetrahymena Telomerase

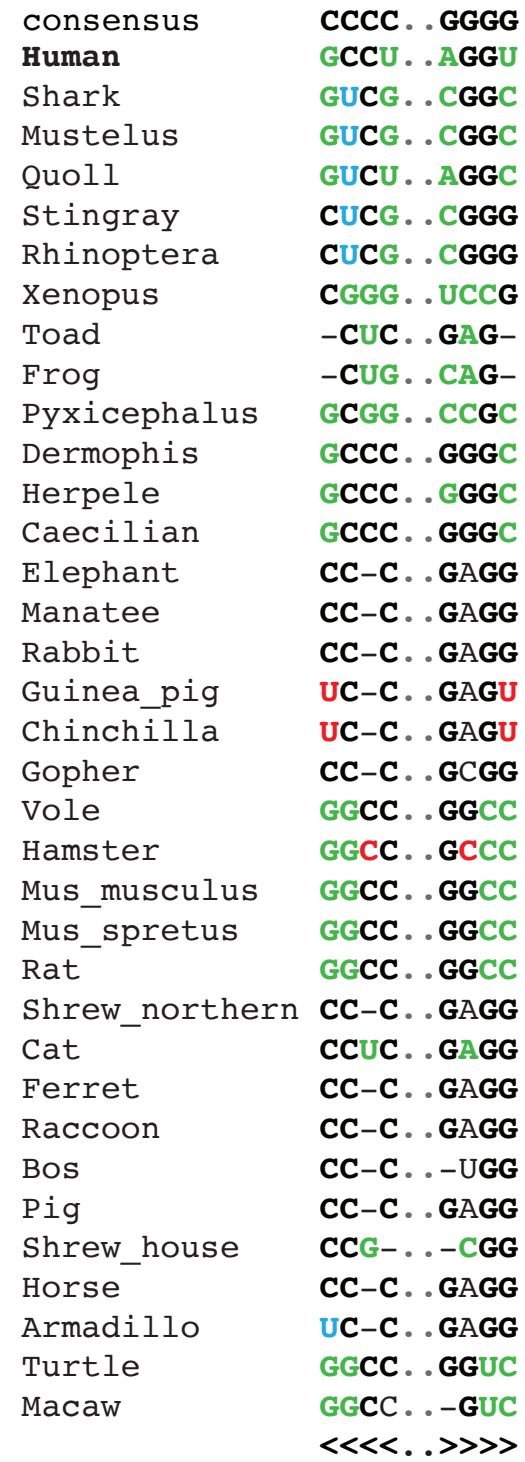
Jiang *et al.*, Cell, 2018

2D structure
vertebrate telomerase RNA (TER)

vTER P4.1

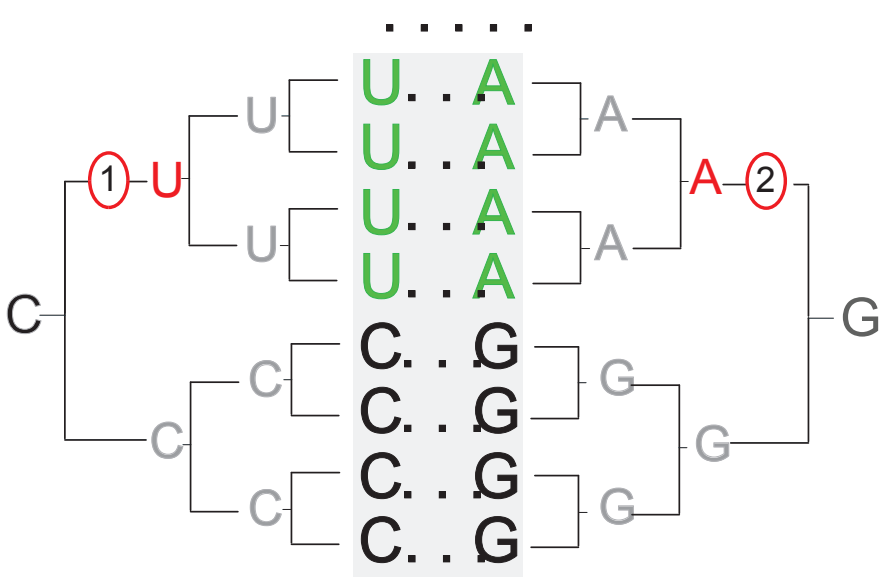
compensatory pair
half-compensatory pair
broken pair

Pattern of sequence changes in a conserved RNA structure

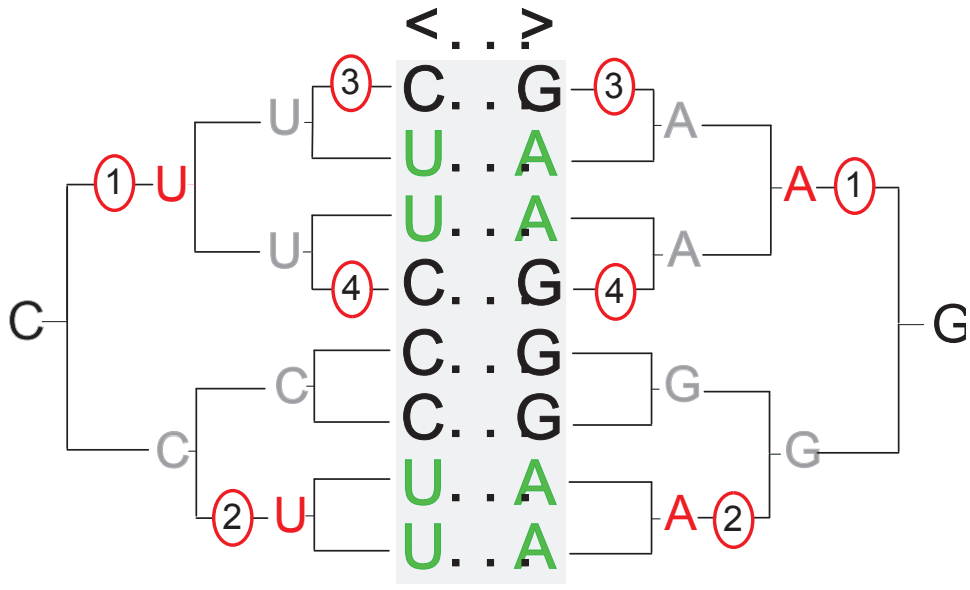


Spurious pairwise covariations can appear from uncorrelated substitutions on a phylogenetic tree

two independent positions

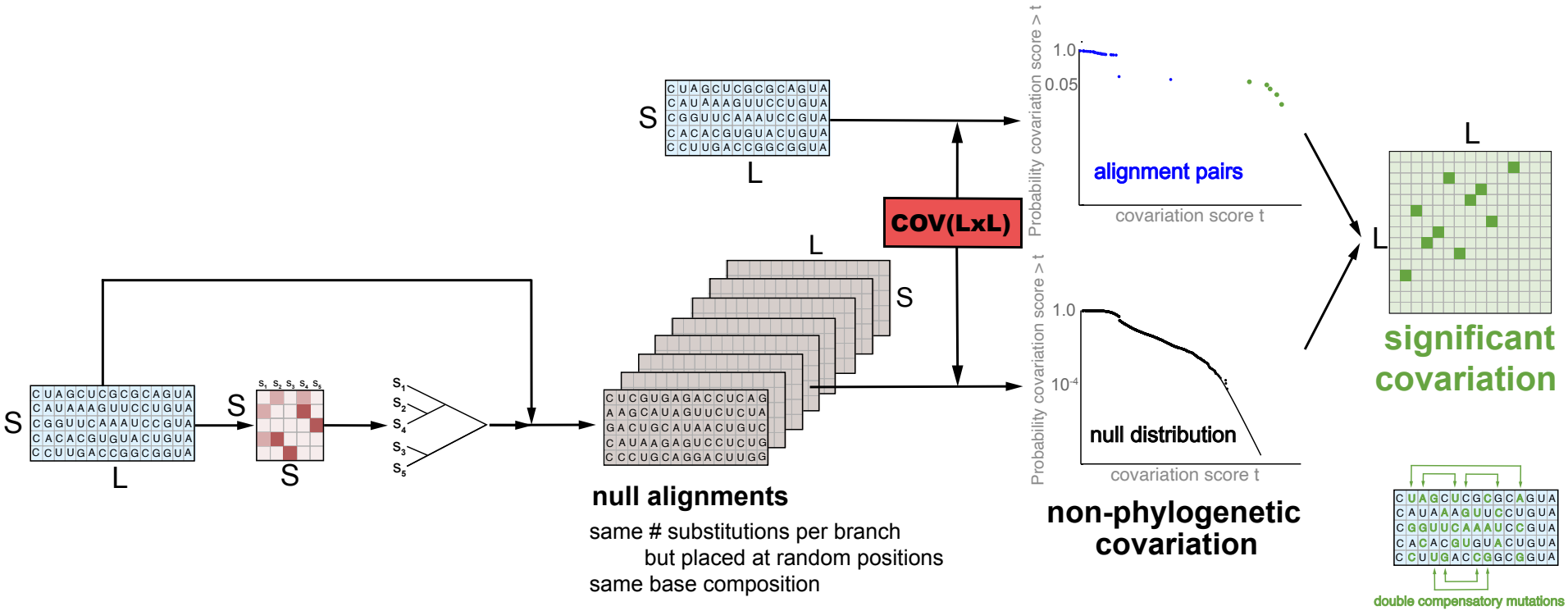


two base-paired positions



A statistical test to simulate phylogenetic covariations

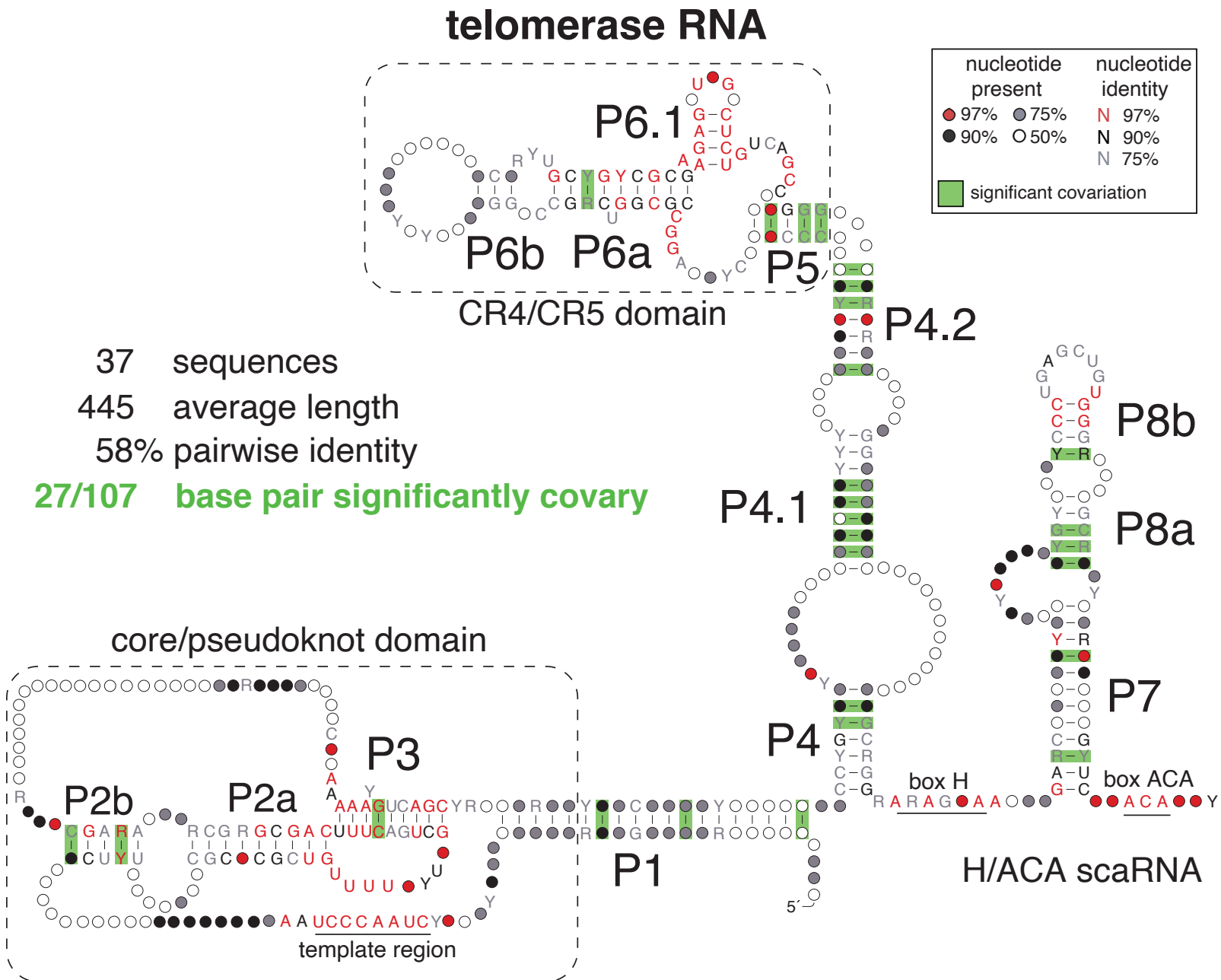
RNA structural covariation above phylogenetic expectation



eddylab.org/rscape

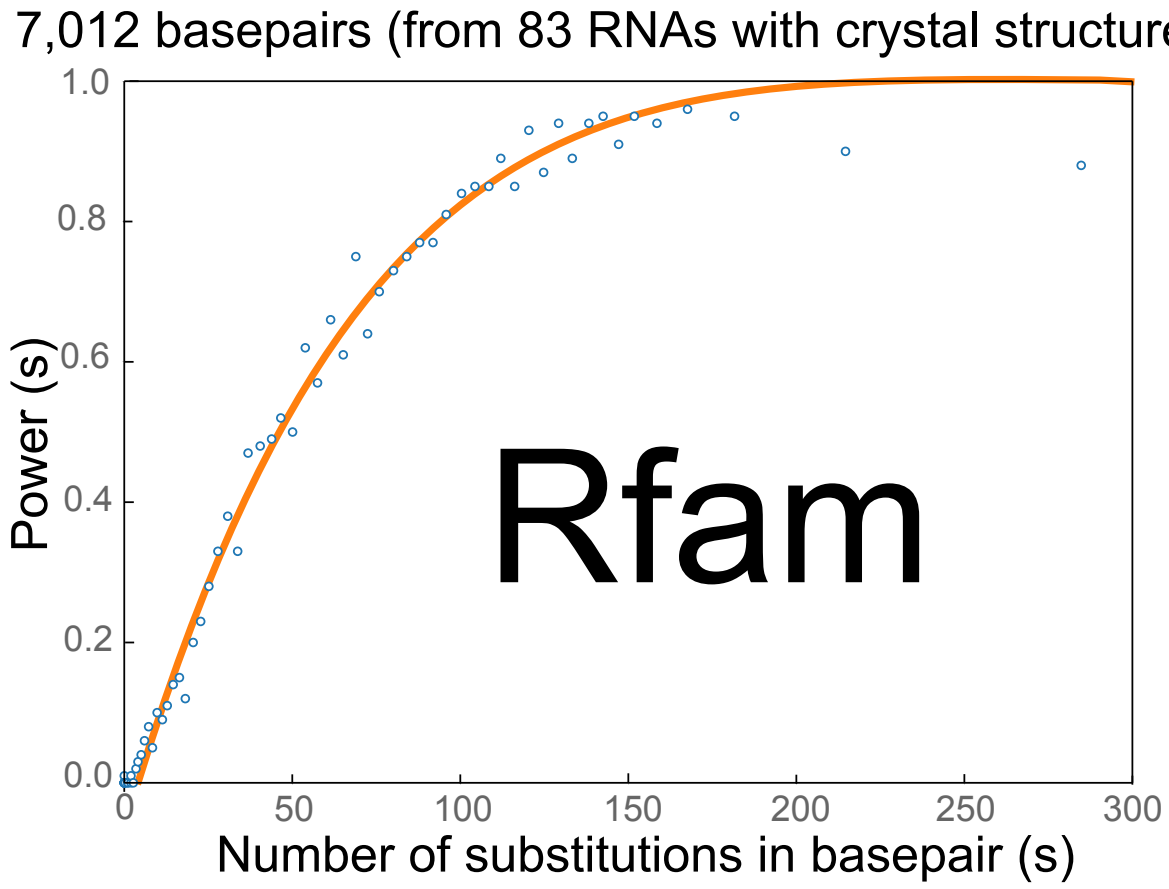
Lack of evidence for conserved secondary structure in long noncoding RNAs
Rivas, Clements & Eddy, *Nat Methods*, 2017.

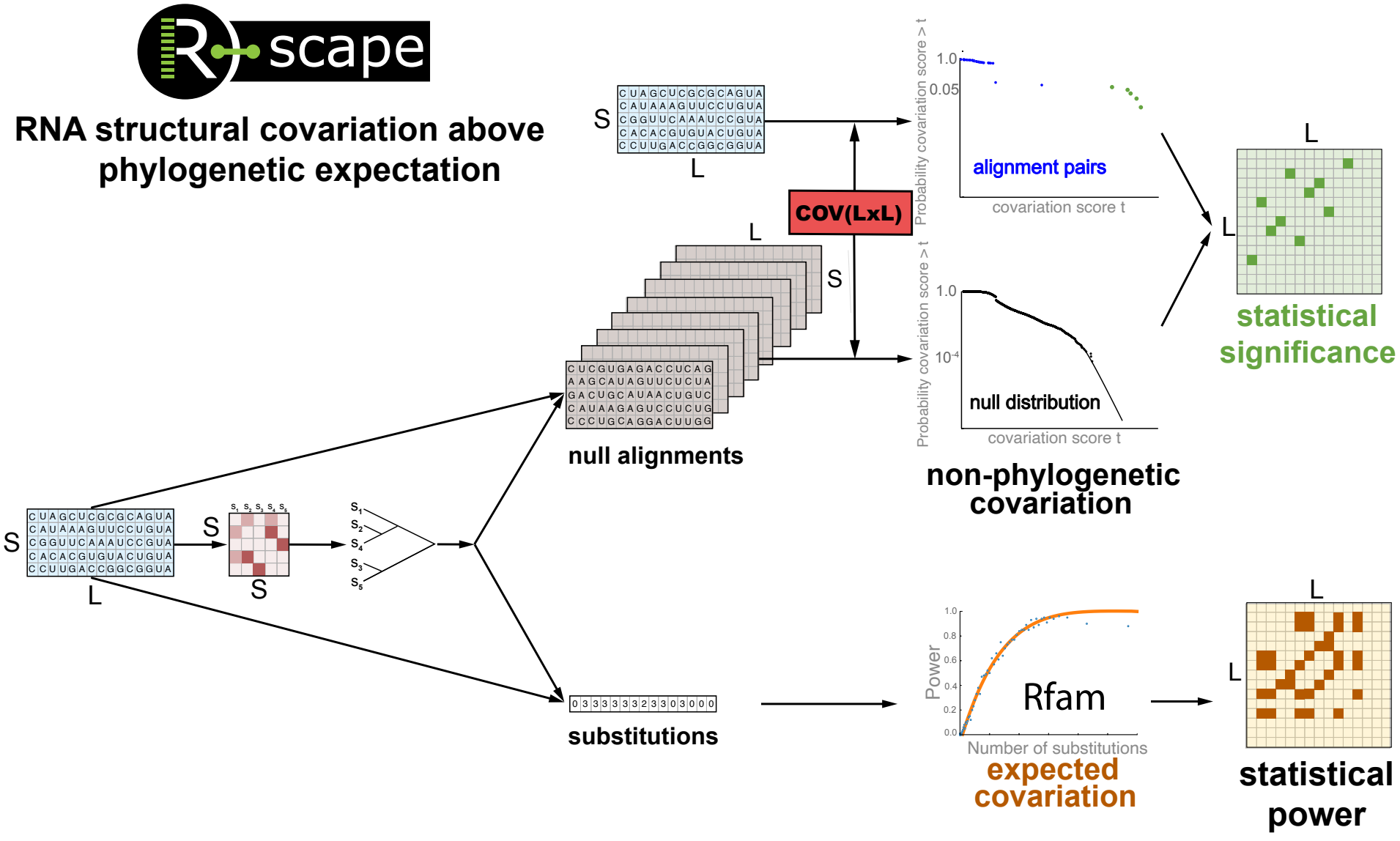
R-scape finds RNA structural covariation above phylogenetic expectation



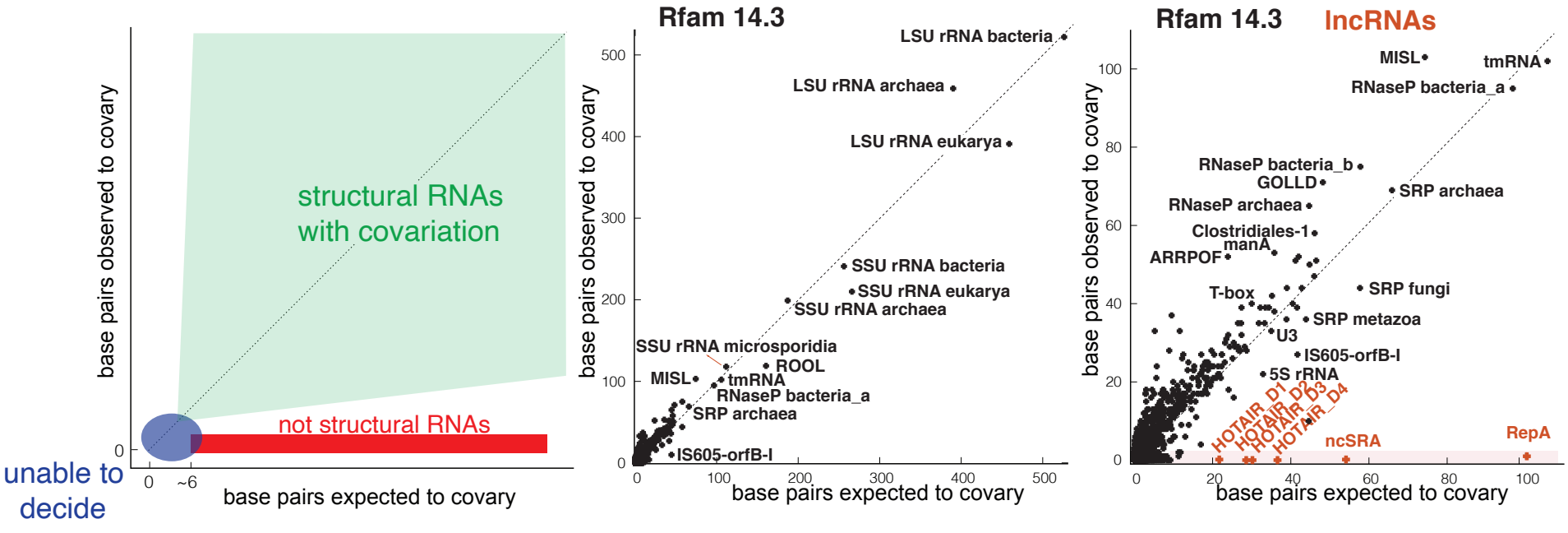
Statistical Power = expected covariation given the variation observed

$$\text{power}(s) = P\,(\text{basepair with s substitutions has an E-value} < 0.05)$$

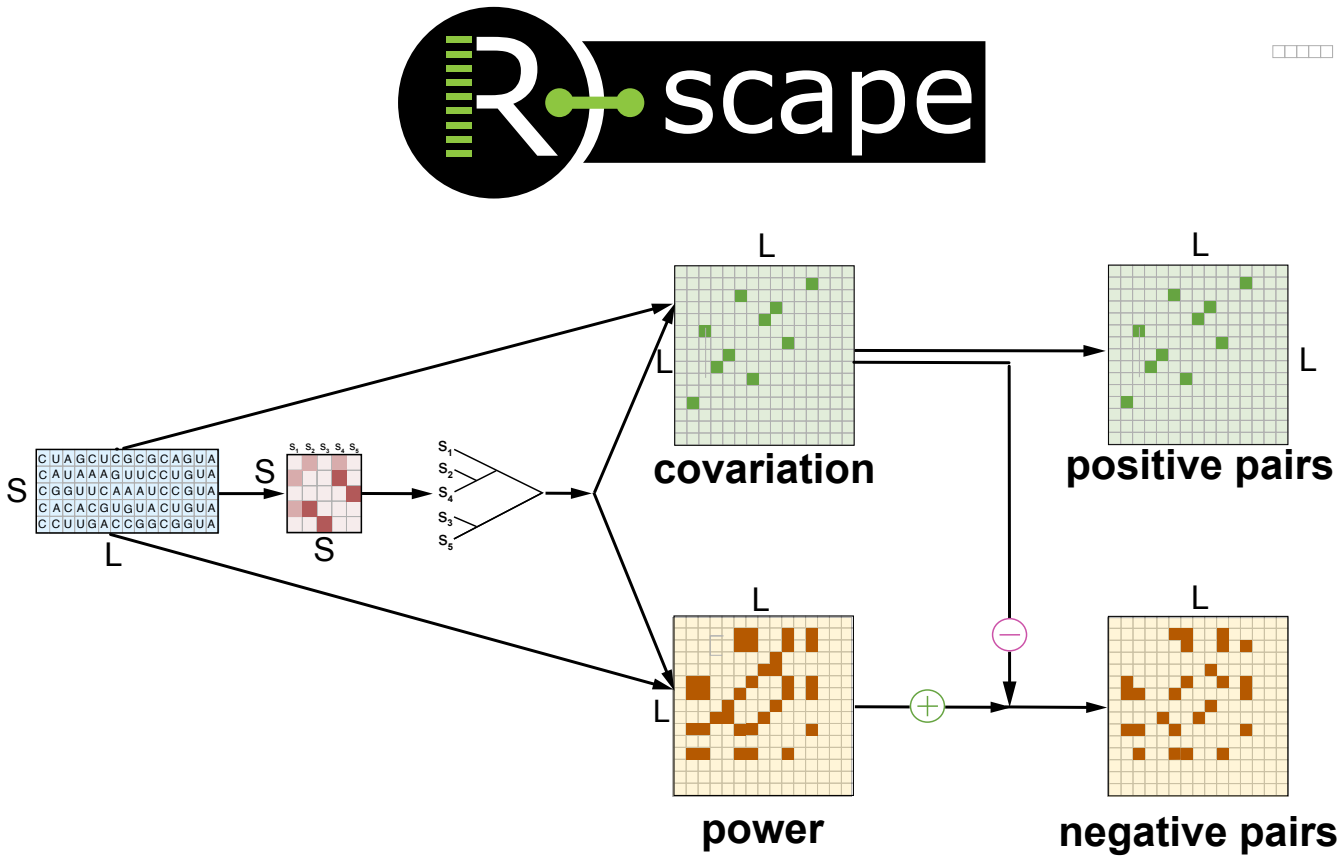



RNA structural covariation above phylogenetic expectation
S
L
COV(LxL)
null alignments
alignment pairs
covariation score t
Probability covariation score > t
1.0
0.05
10-4
null distribution
non-phylogenetic covariation
statistical significance
substitutions
Power
Rfam
Number of substitutions
expected covariation
statistical power

How to tell when an RNA has an evolutionarily conserved structure ... or not?



Estimating the power of sequence covariation for detecting conserved RNA structure
Rivas *et al.*, *Bioinformatics*, 2020.


R-scape
S
L
S
S
covariation
positive pairs
power
negative pairs

Cascade covariation/variation Constrained RNA Folding

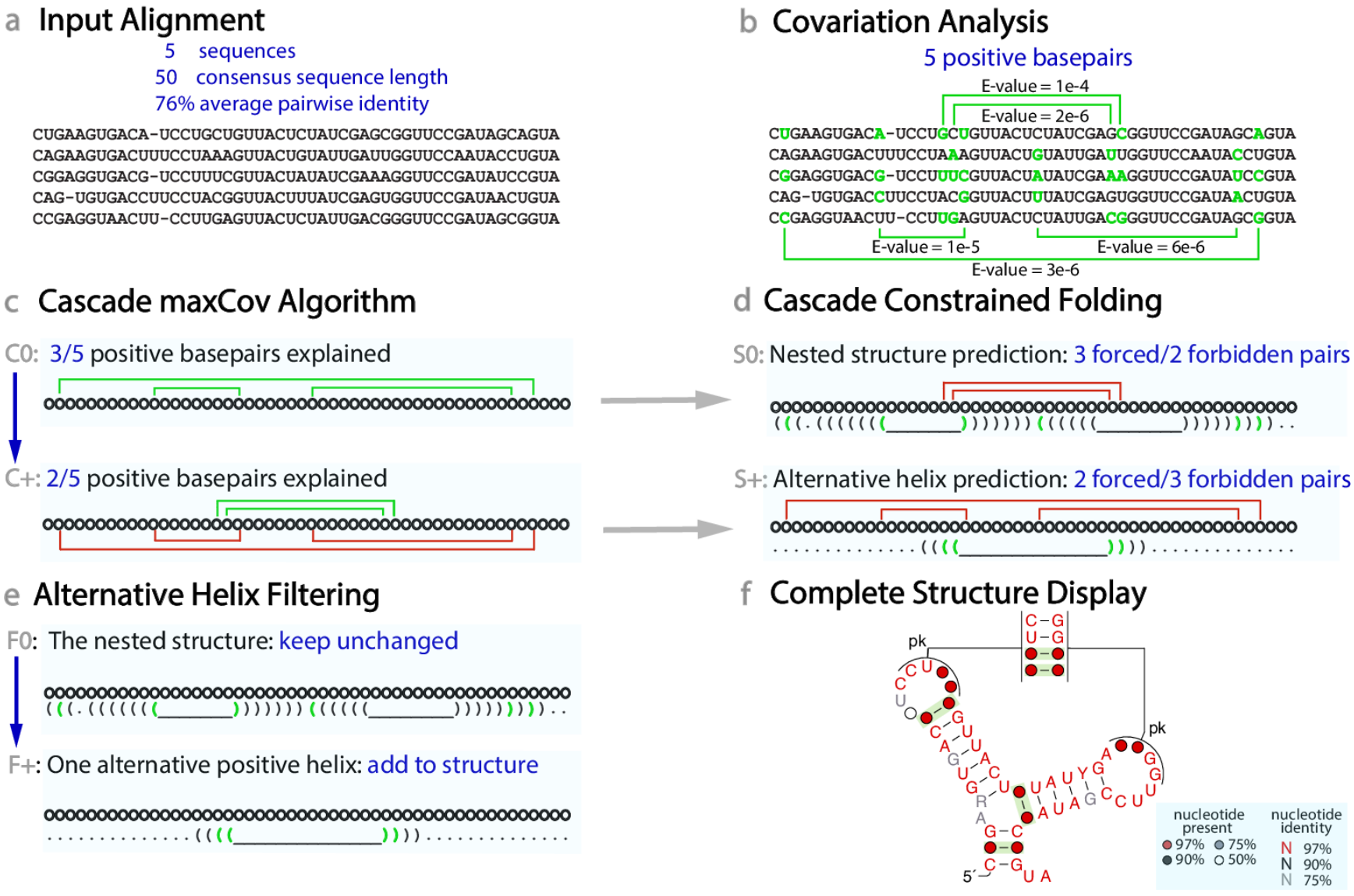
- Produces an RNA structure that incorporates **all** positive basepairs
- Produces an RNA structure that forbides negative basepairs
- Uses a battery of probabilistic folding algorithms (computationally efficient)
- Visualization to critically analyze the structure
- In vivo structures

CaCoFold can incorporate pseudoknots, base triplets, other 3D interactions, even non RNA-structure related covariation

RNA structure prediction using positive and negative evolutionary information,
E Rivas, PLoS Comp Bio, 2020.

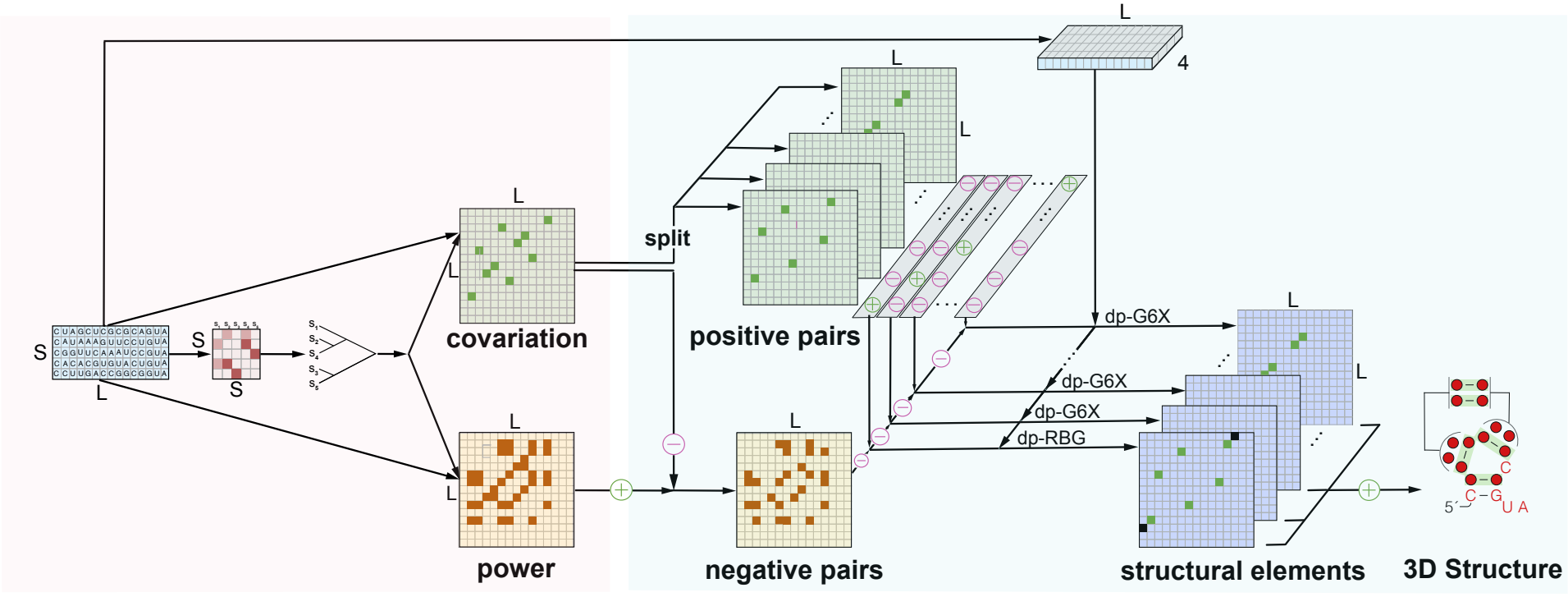
RNA structure prediction using positive and negative evolutionary information

CaCoFold



R-scape

CaCoFold



Rivas, *PLOS Comp Biol*, 2020.

a **Model used by the maxCov algorithm**

Nussinov Grammar

```
S -> o S          any non-covarying residue
S -> o S o S      a covarying basepair
S -> S S
S -> end
```

b **Model used by the folding algorithm (first layer)**

RNA Basic Grammar (RBG)

```
S -> o S      a free unpaired residue
S -> L S
S -> end

L ->  o F o   a helix starts
L ->  o P o   a one-basepair helix ends

F ->  o F o   a helix adds one more basepair
F ->  o P o   a helix ends
```

what can happen at the end of a helix...

```
P ->      o...o          a hairpin      loop
P -> o...o L             a left  bulge loop
P ->       L o...o       a right bulge loop
P -> o...o L o...o       an internal    loop
P -> M1 M                a multiloop starts

M -> M1 M   multiloop adds one more branch
M -> R      multiloop about to add right residues

R -> R  o   a right-unpaired residue in multiloop
R -> M1     multiloop about to add left residues

M1 -> o M1  a left-unpaired residue in multiloop
M1 -> L     multiloop starts another helix
```

c **Model used by the folding algorithm (additional layers)**

G6X Grammar

```
S -> L
S -> L S
S -> end

L -> o F o    a helix starts
L ->  o o     a basepair of contiguous residues
L ->   o      an unpaired residue

F -> o F o    a helix adds one more basepair
F ->  o o     a helix ends without a hairpin
F ->  L S     a helix ends, more stuff to come
```

```
o        a non-covarying RNA residue
o o      a covarying RNA basepair
o        an RNA residue, not forming any basepairing
o...o    a set of contiguous unpaired RNA residues
o o      an RNA basepair; bases could be at arbitrary distance in the RNA backbone
```

`S,L,F,P,M,M1,R` non-terminals that have to be transformed following one of the allowed rules

transfer-messenger RNA (tmRNA)

a Input Alignment

Rfam RF00023 seed alignment

477 sequences
354 consensus sequence length
357 average sequence length
42% average pairwise identity

b Covariation Analysis

All possible pairs analyzed equally

119 annotated basepairs in alignment
(not used in analysis)

414 columns analyzed:
121 positive basepairs (significantly covary)
109 positive basepairs expected by power
31,027 negative basepairs

c Cascade maxCov Algorithm

121 positive basepairs explained in 6 layers

layer 1: 69 layer 2: 41
layer 3: 5 layer 4: 3
layer 5: 2 layer 6: 1

d Cascade Constrained Folding

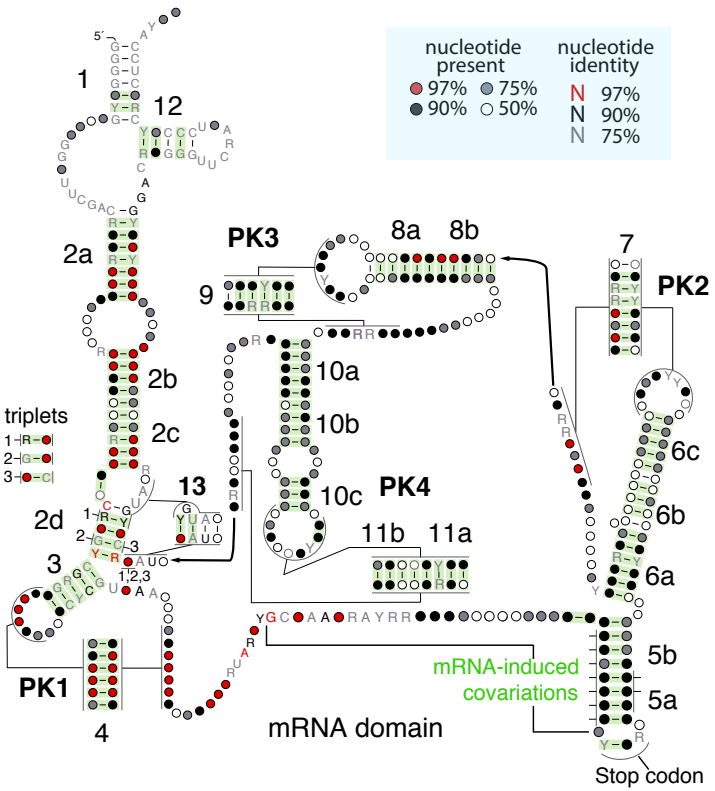
139 annotated pairwise interactions
121/139 positive basepairs
74 pairs not in final ss due to forbidden negative basepairs

e Alternative Helix Filtering

18 alternative helices

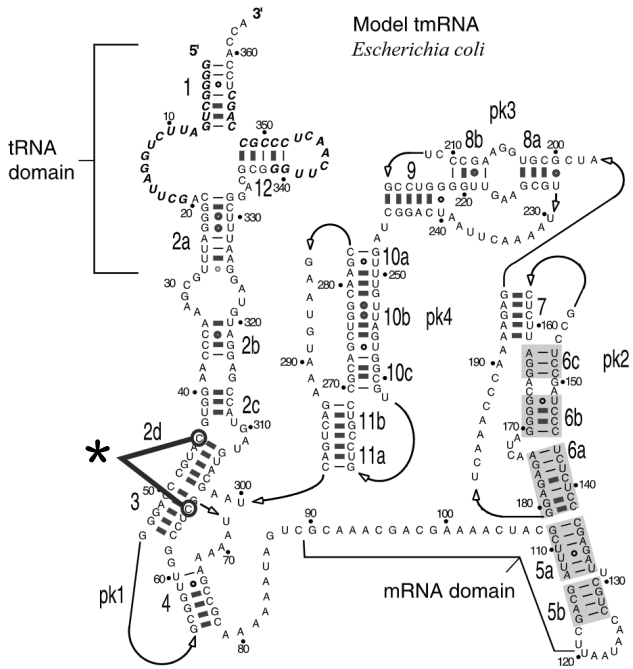
- 5 pseudoknots
- 3 triplets
- 10 mRNA-induced covariations

f Complete structure display

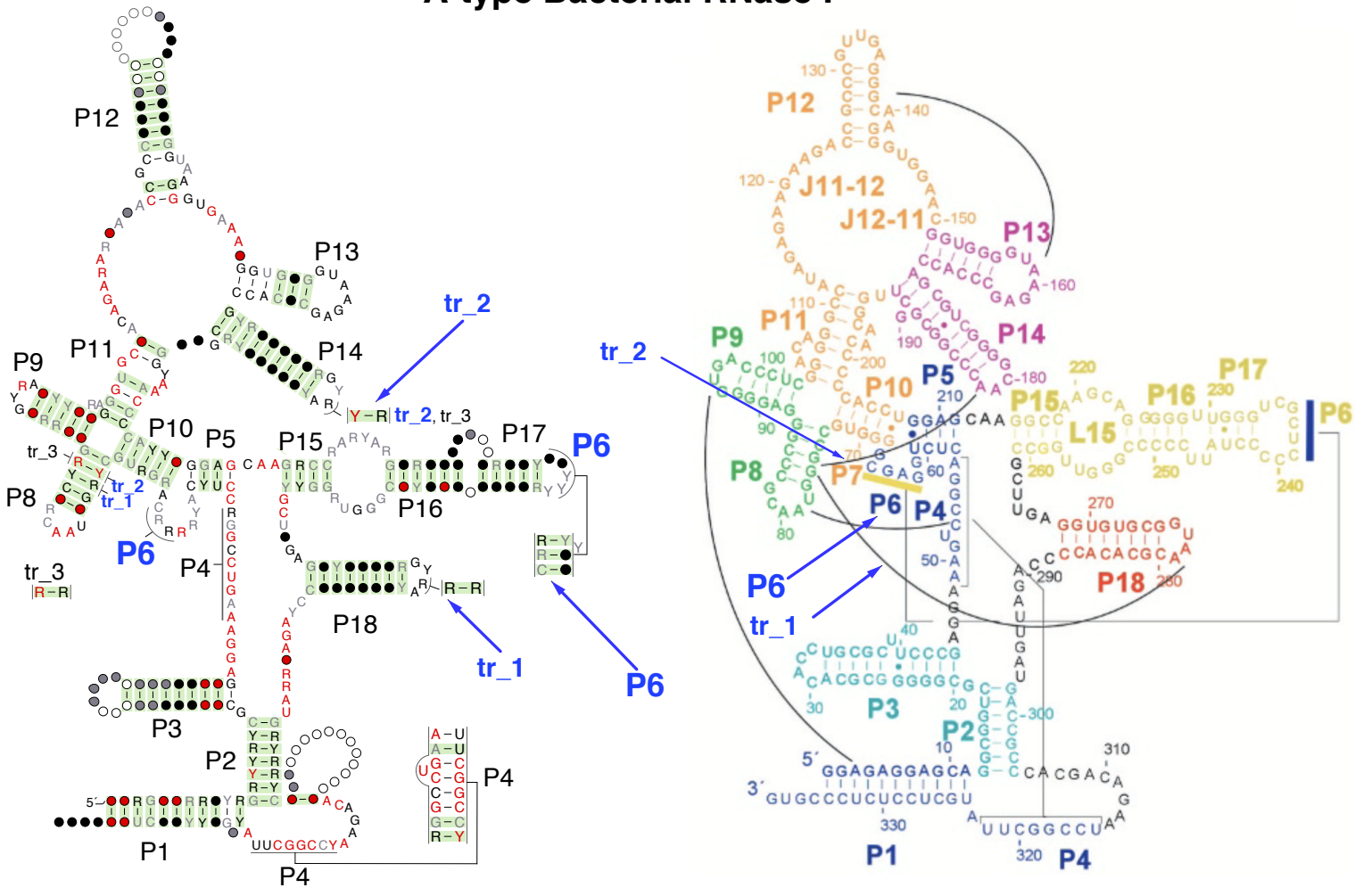


g Structure comparison

Kelley *et al.*, RNA 2001, Fig 4



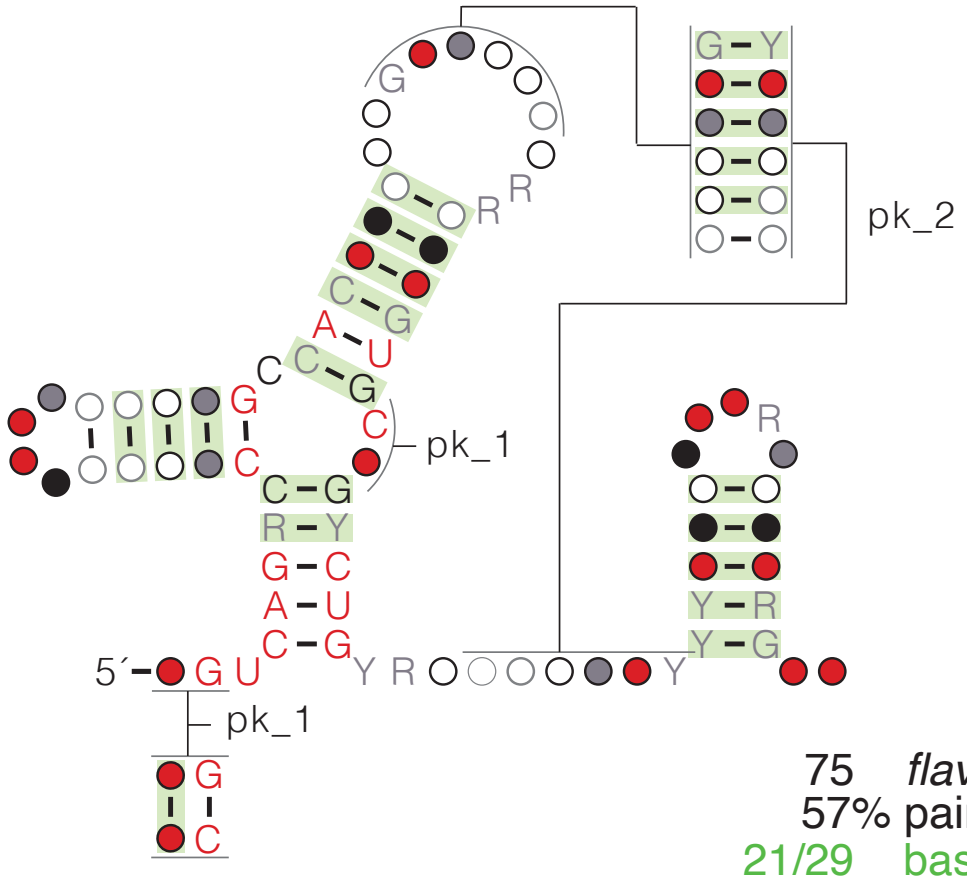
A-type Bacterial RNase P

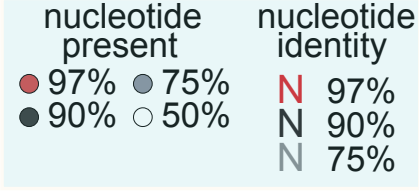


CaCoFold

Torres-Larios *et al.*, Nature 2005, Fig 2c

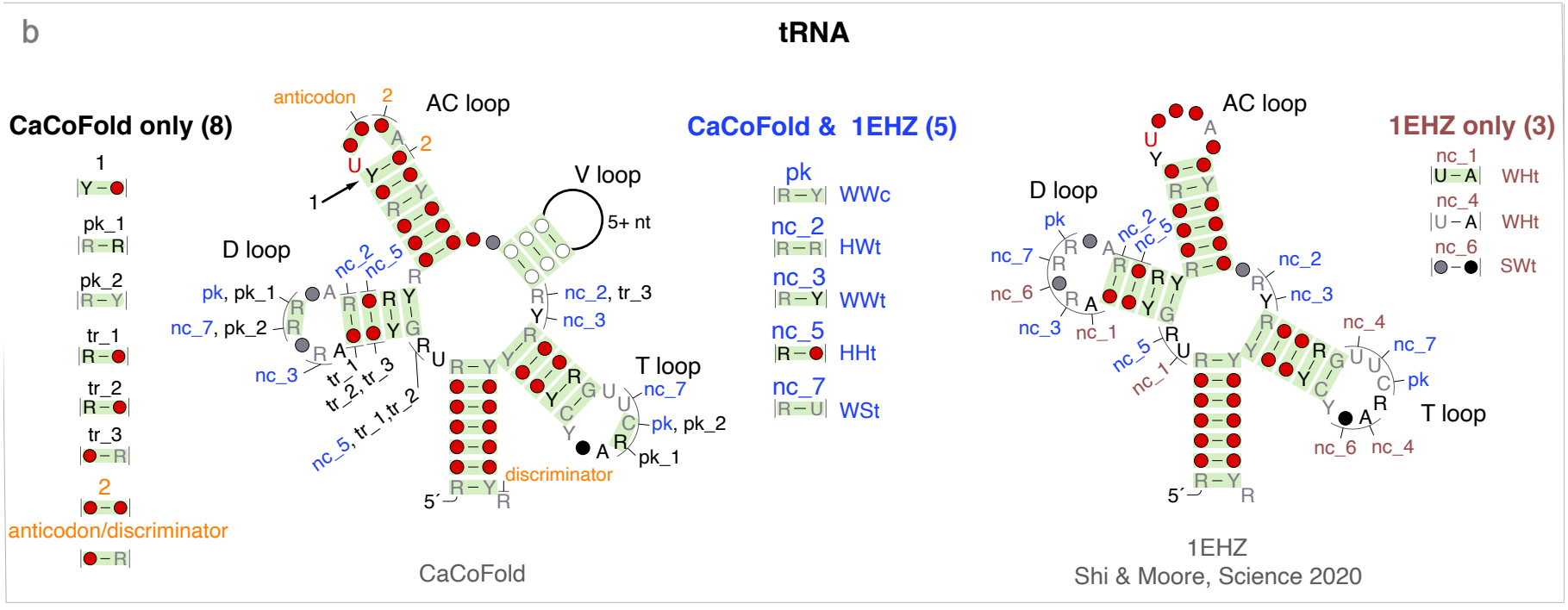
exoribonuclease resistant RNA





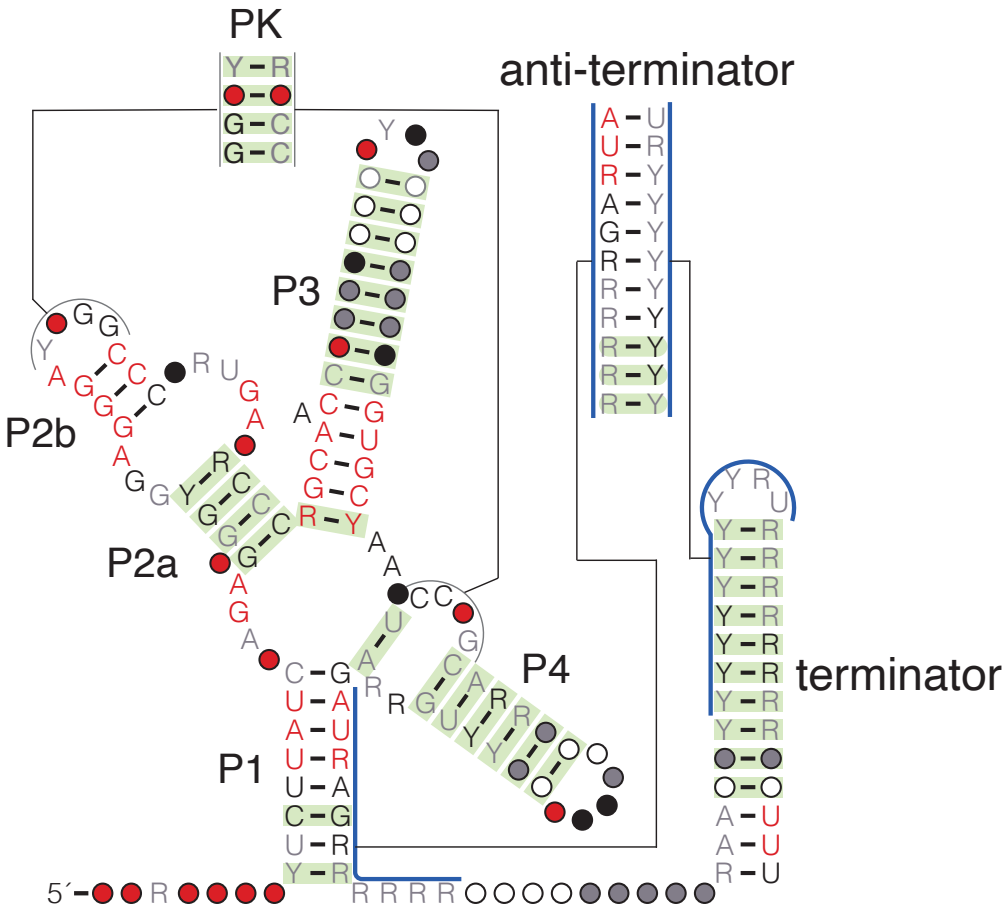
75 *flaviviridae* sequences
57% pairwise identity
21/29 base pair significantly covary

Steckelberg, Vicens, Kieft, mBio, 2018
Szucs, Nichols, Jones, Vicens, Kieft, mBio, 2020


b
tRNA
CaCoFold only (8)
1
pk_1
pk_2
tr_1
tr_2
tr_3
2
anticodon/discriminator
anticodon
AC loop
V loop
5+ nt
D loop
T loop
discriminator
CaCoFold & 1EHZ (5)
pk
nc_2
nc_3
nc_5
nc_7
WWc
HWt
WWt
HHt
WSt
1EHZ only (3)
nc_1
nc_4
nc_6
WHt
WHt
SWt
CaCoFold
1EHZ
Shi & Moore, Science 2020

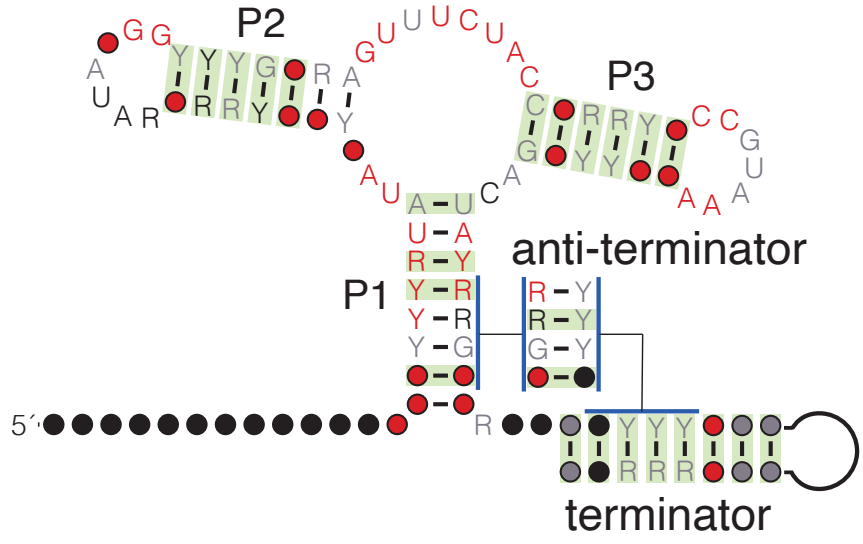
Alternative/overlapping structures

SAM-I riboswitch

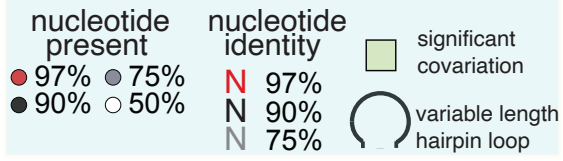


Zhu & Meyer, *RNA Biology*, 2015

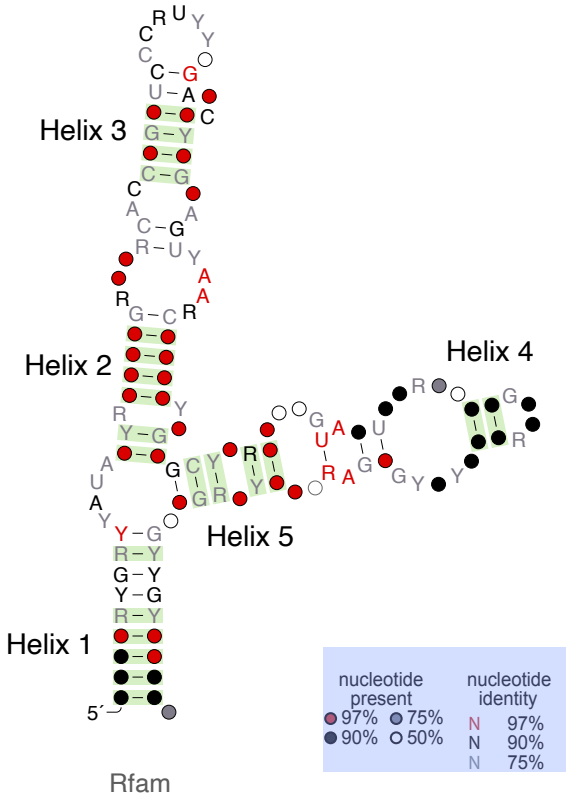
Purine riboswitch



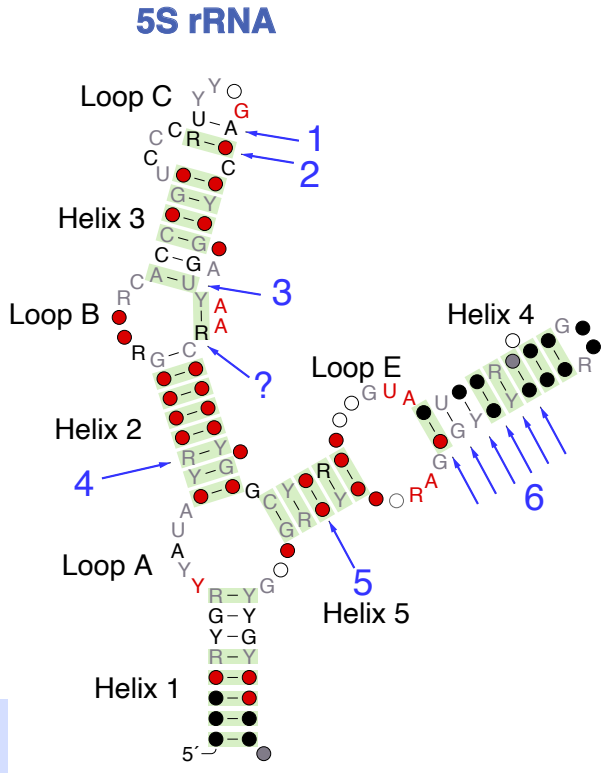
Ritz, Martin, Laederach, *PLCB*, 2013



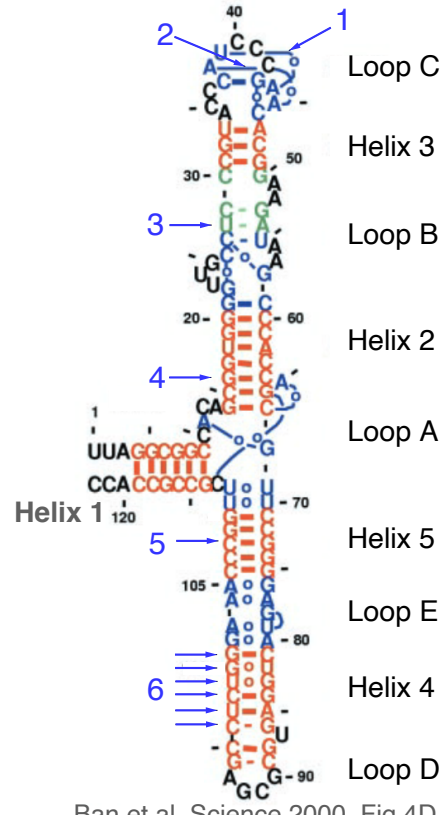
CaCoFold helps improve Rfam structures



Rfam



CaCoFold

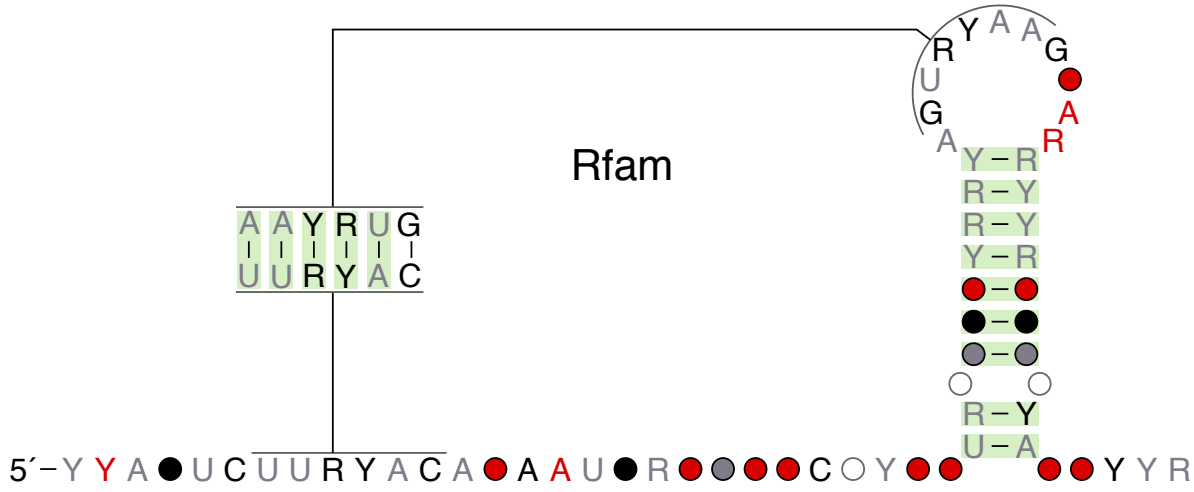


Ban et al, Science 2000, Fig 4D

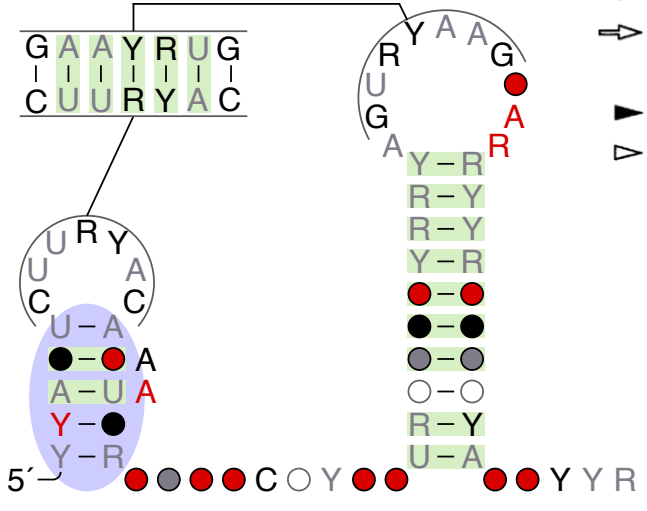
Type 2

Coronavirus 3'UTR pseudoknot

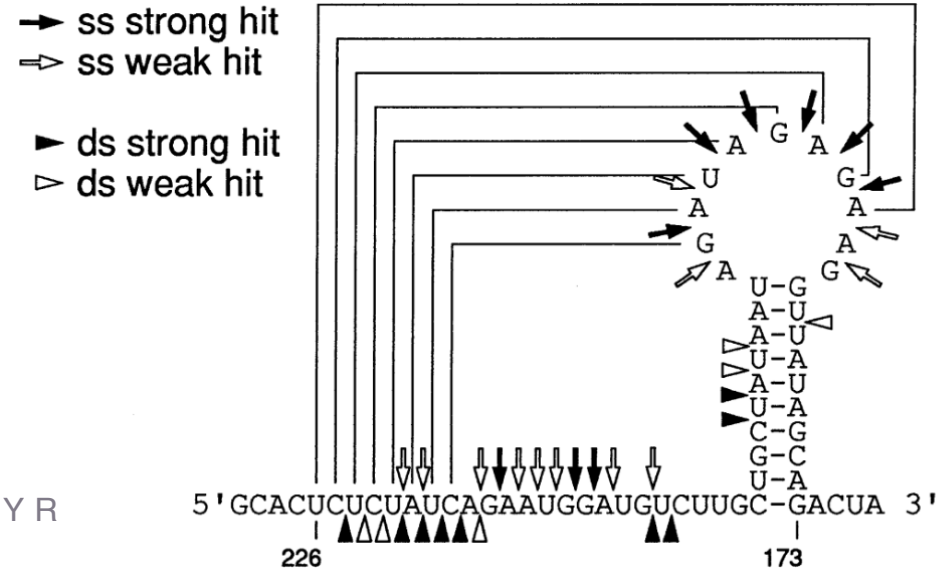
Rfam



CaCoFold

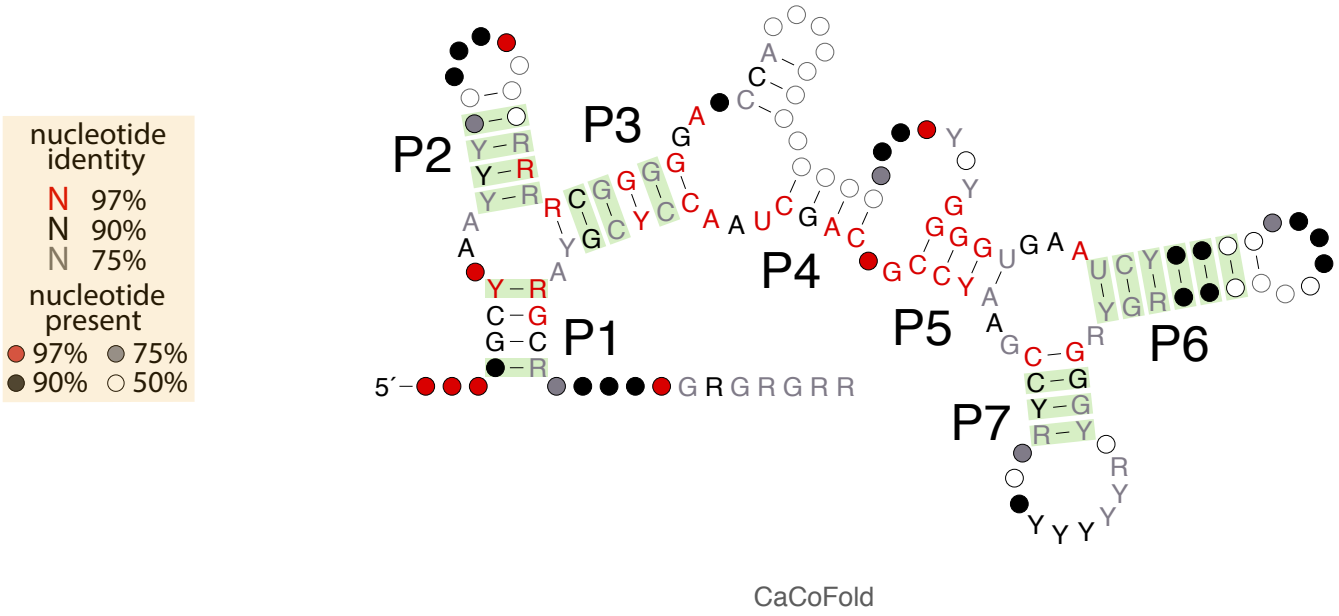


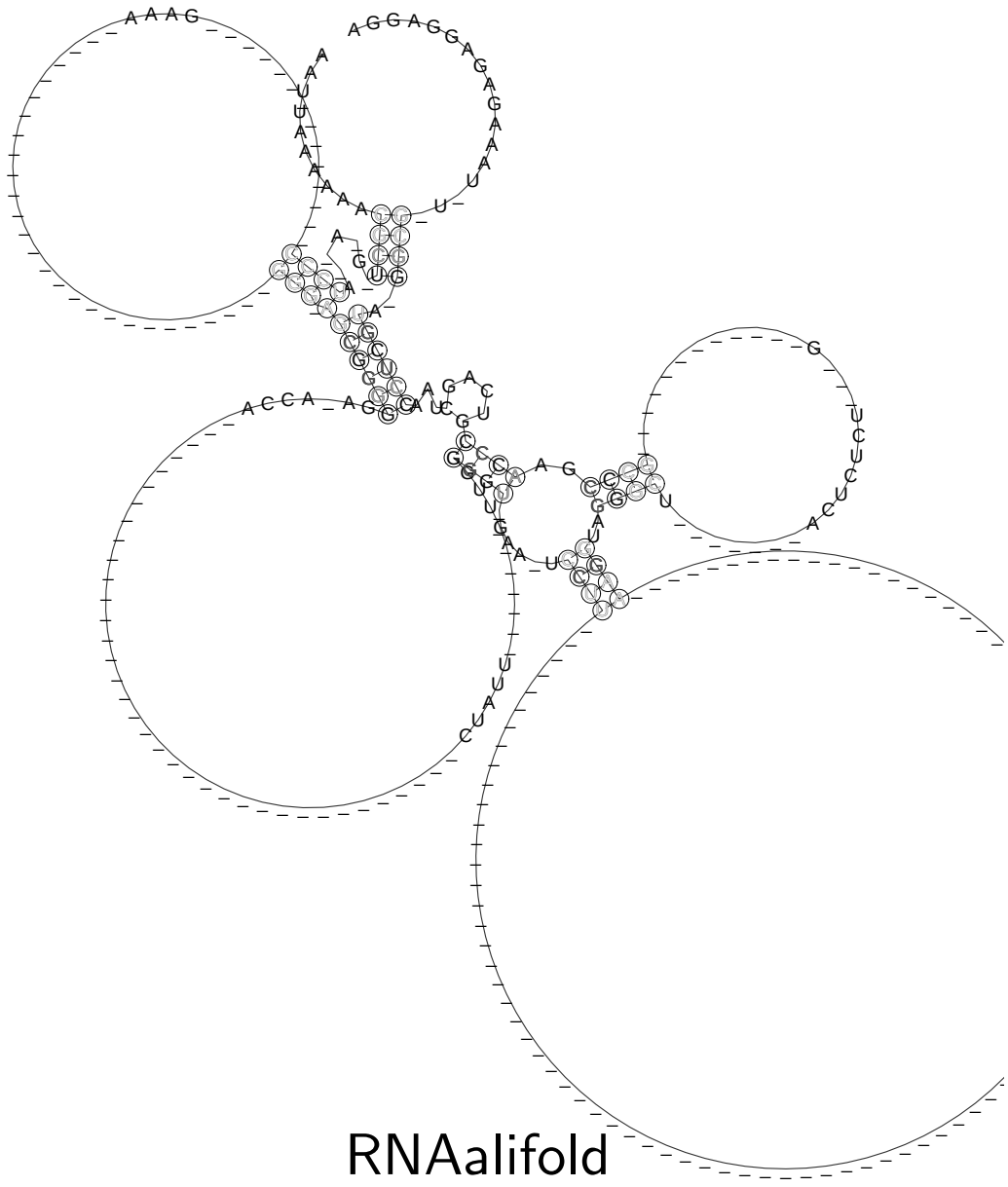
Williams *et al.*, J. Virol. 1999, Fig 4B



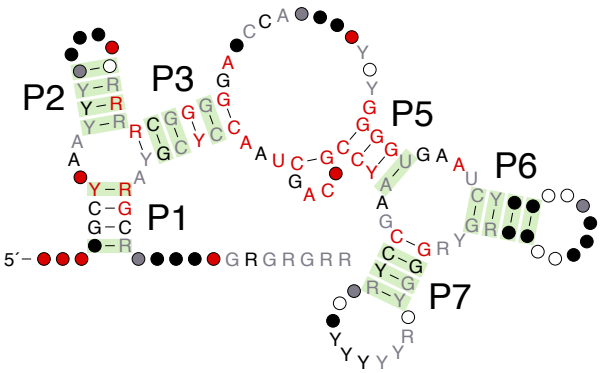
new helix with covariation support

Cyclic di-AMP riboswitch



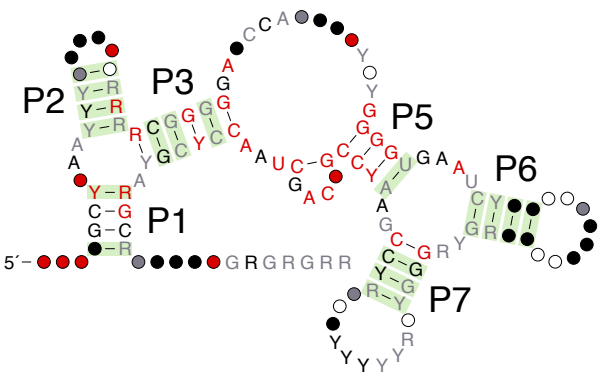

RNAalifold

Cyclic di-AMP riboswitch

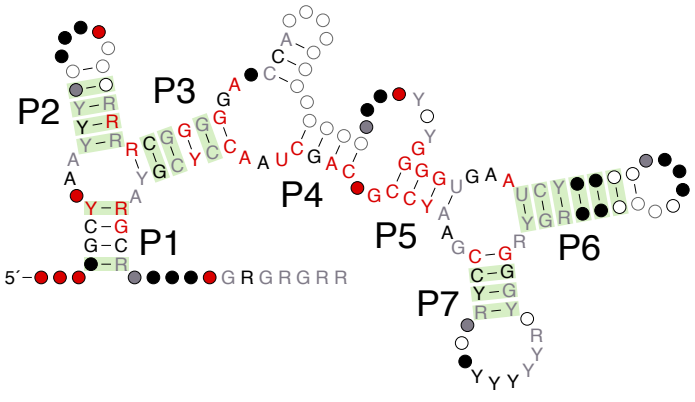


RNAalifold-R-scape

Cyclic di-AMP riboswitch

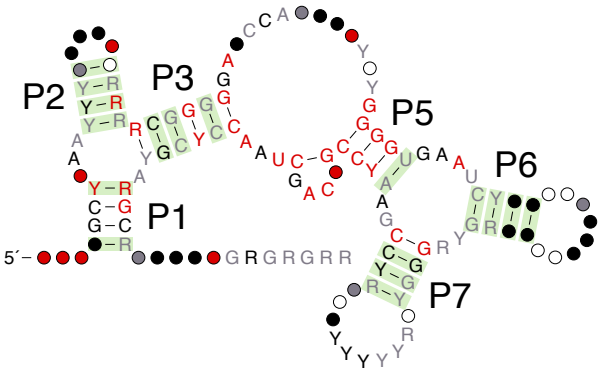


RNAalifold-R-scape

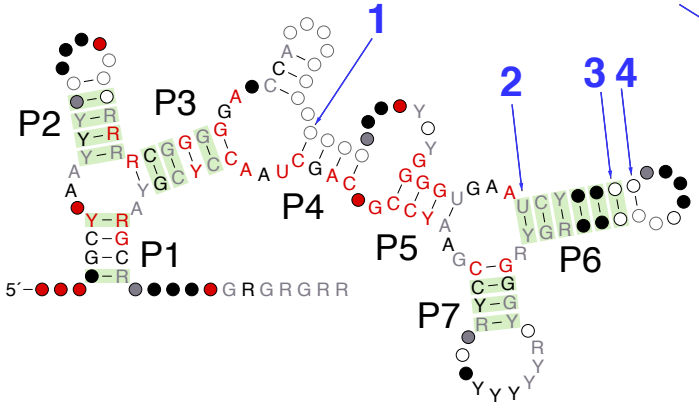


CaCoFold

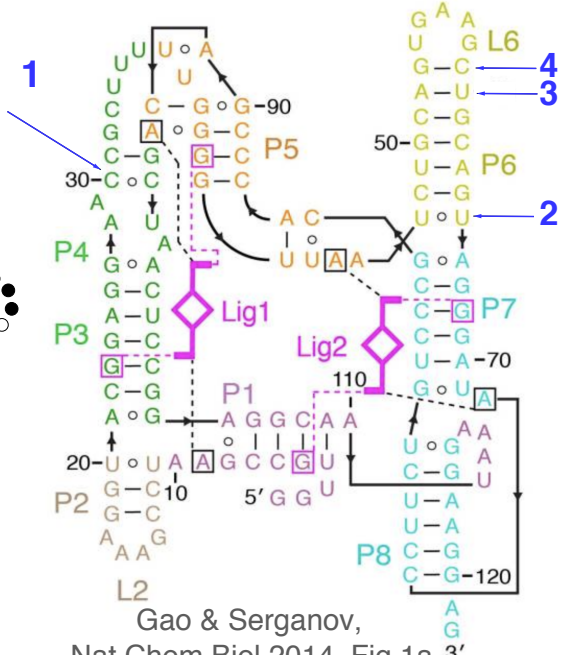
Cyclic di-AMP riboswitch



RNAalifold-R-scape



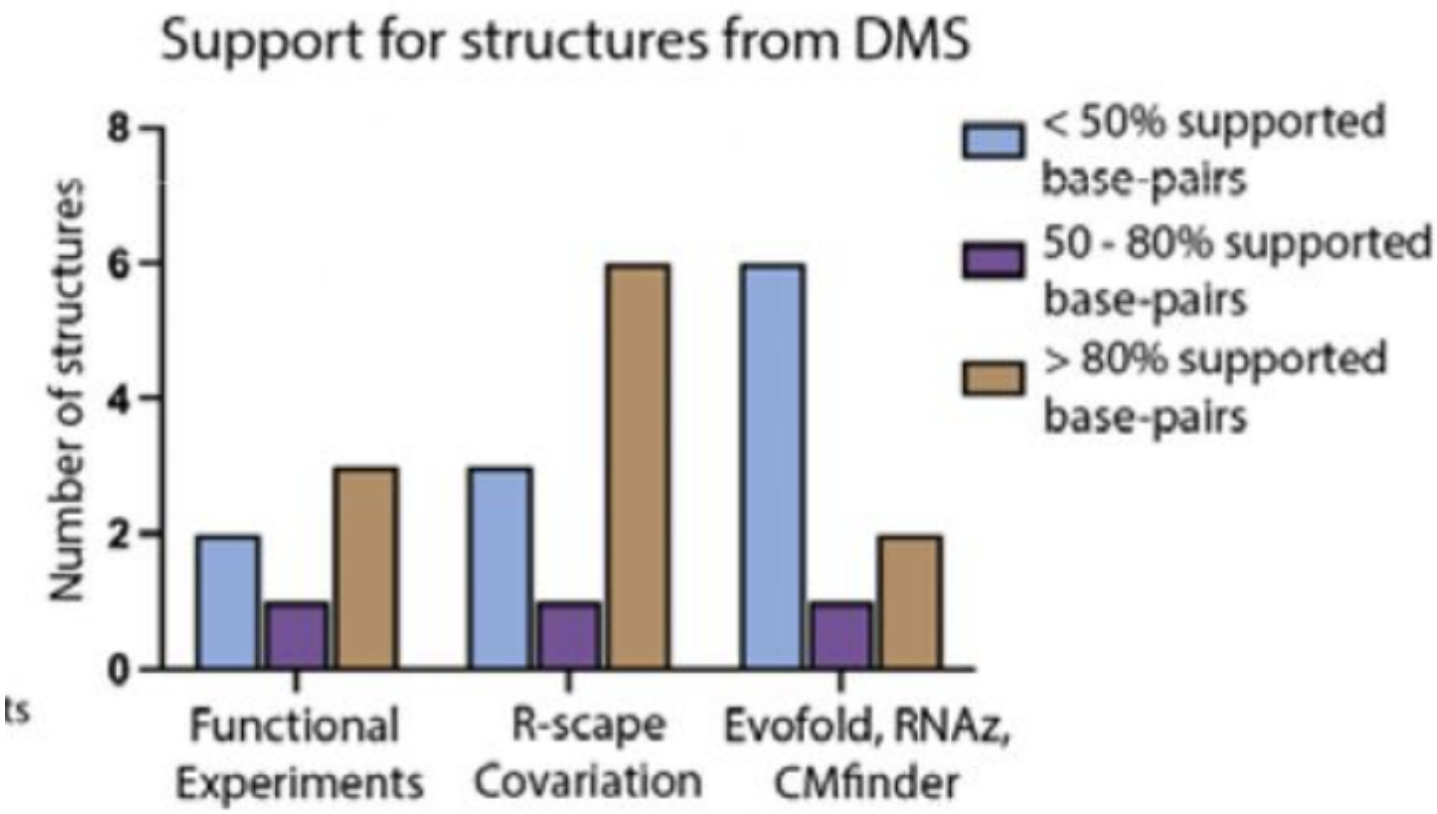
CaCoFold



Gao & Serganov,
Nat Chem Biol 2014, Fig 1a

CaCoFold find in vivo structures

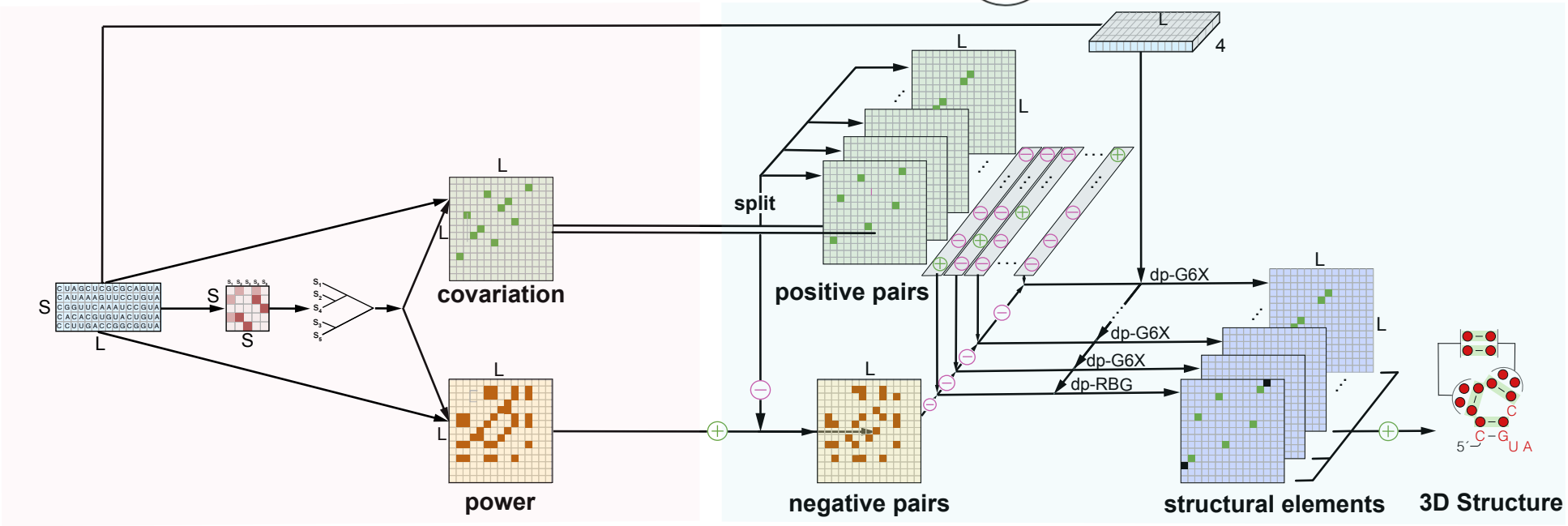
structures with support from DMS reactivity for in vivo formation

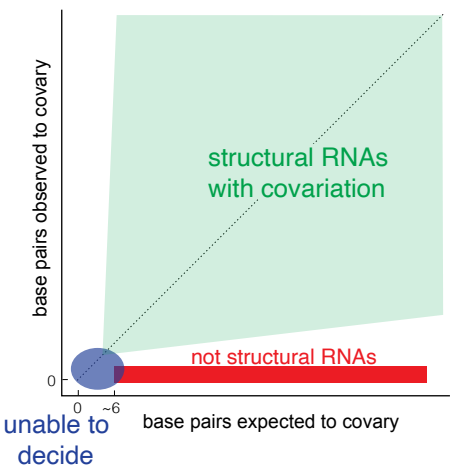


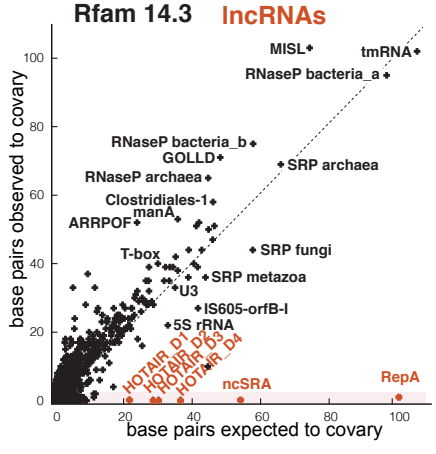
"RNA structure landscape of S. cerevisiae intron"
Rangan, Hunter, Pham, Ares Jr., Das

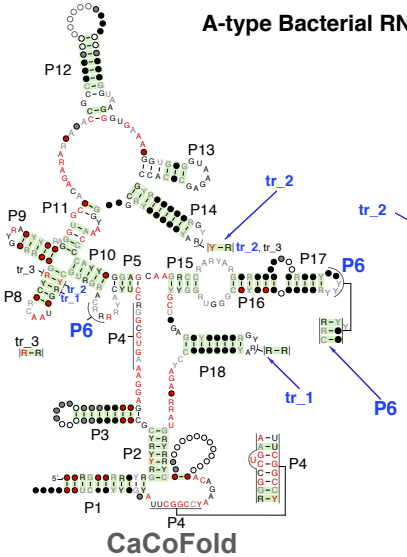
R-scape

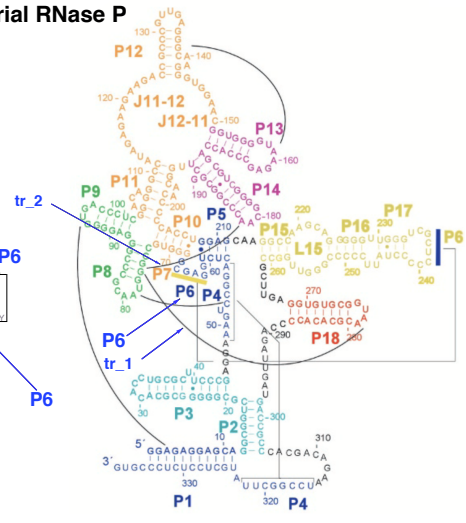
CaCoFold











Torres-Larios et al., Nature 2005, Fig 2c