Progress toward Predicting Viral RNA Structure from Sequence:

How Parallel Computing can Help Solve the RNA Folding Problem

Susan J. Schroeder
University of Oklahoma

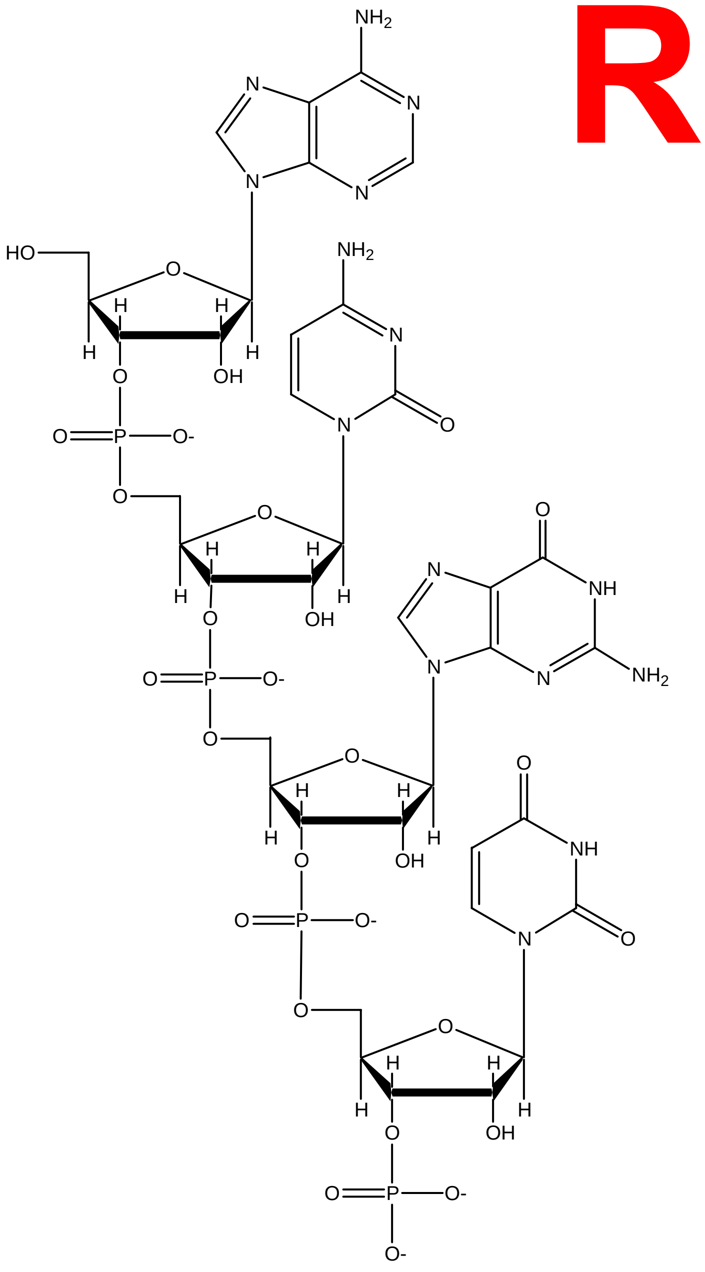
October 7, 2008

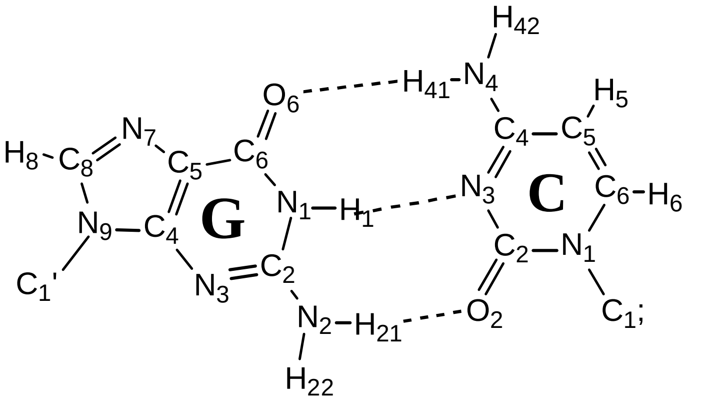
We finished the genome map,

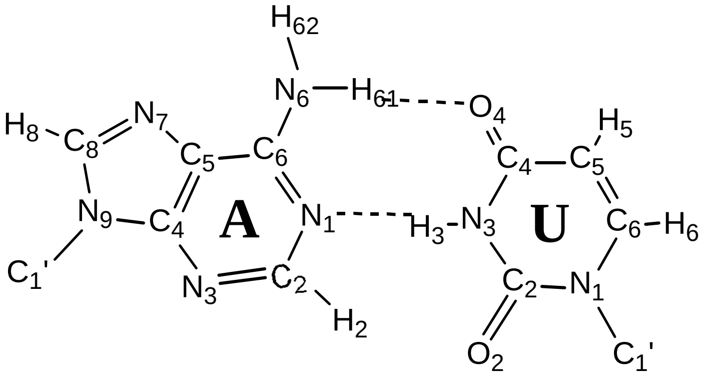
now we can’t figure out how to fold it!

Science (1989) **243,** p.786

RNA


NH2
N
N
N
N
HO
O
H
H
H
H
OH
NH2
N
N
O
O
O=P
O-
O
O
H
H
H
H
OH
O
N
NH
N
N
NH2
O=P
O-
O
O
H
H
H
H
OH
O
NH
N
O
O=P
O-
O
O
H
H
H
H
OH
O
O=P
O-
O-

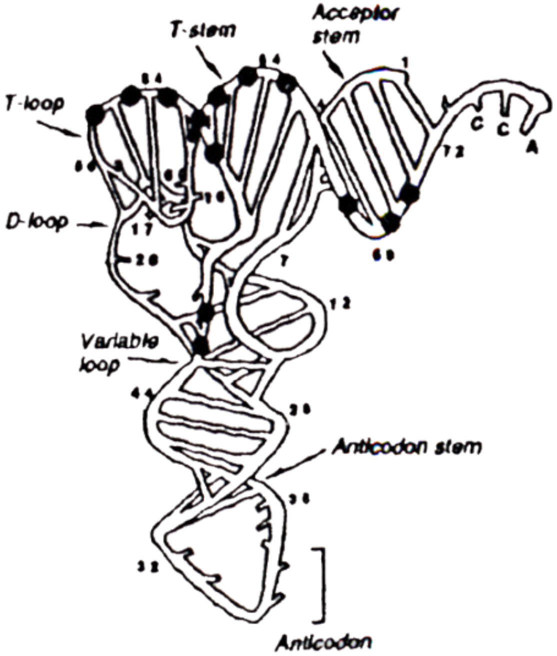

H42
O6
H41-N4
H5
H8-C8
N7
C5
C6
C4
C5
G
C
N1-H1
N3
C6-H6
N9
C4
C2
N1
C1'
N3
C2
O2
C1;
N2-H21
H22


H62
N6
H61
O4
H8
C8
N7
H5
C5
C6
C4
C5
A
U
N9
C4
N1
H3-N3
C6-H6
C1'
N3
C2
C2
N1
H2
O2
C1'

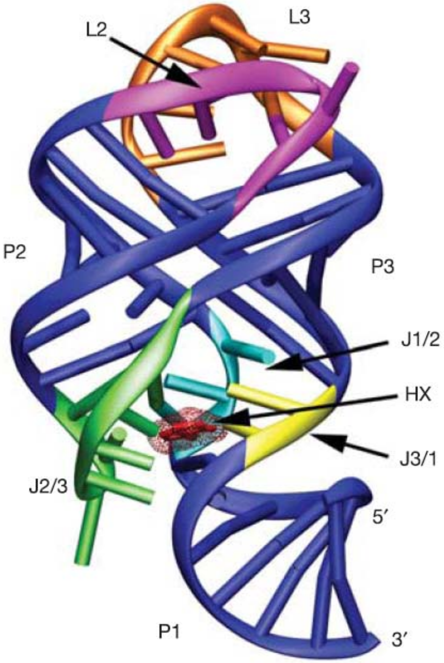
Sequence→ Structure→ Function

5’GCGGAUUUAG^{2M}
CUCAGU^{DH}U^{DH}GGG
AGAGCG^{M2}CCAGA
C^{0M}UG^{OM}AAG^{Y}AU^{PS}
C^{5M}UGGAGG^{7M}UC
C^{5M}UGUGU^{5M}U^{PS}C
GA^{1M}UCCACAGAA
UUCGACCA

5’GGACAUAUAAU
CGCGUGGAUAUG
GCACGCAAGUUU
CUACCGGGCACC
GUAAAUGUCCGA
CUAUGUCCA



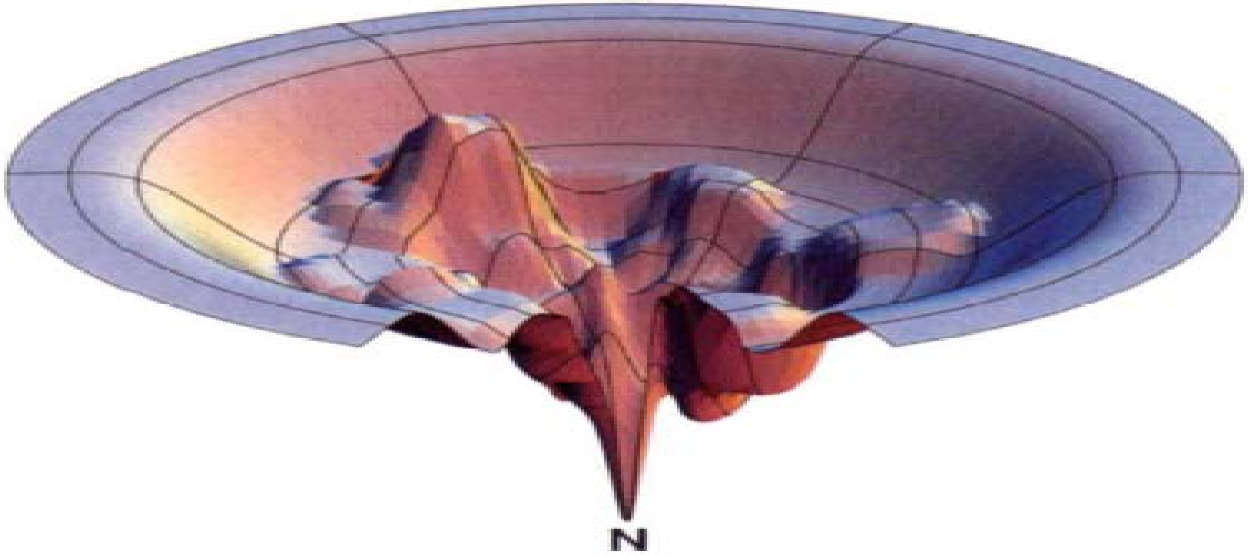
tRNA



Guanine Riboswitch

Batey, R. et al, 2004 Nature vol. 432, p. 412

RNA Folding Problem

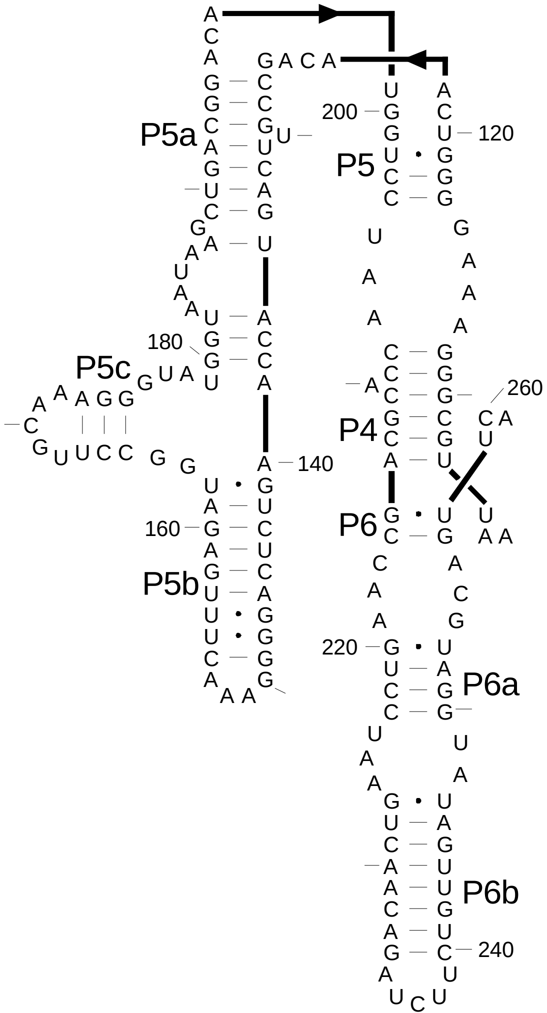


- Folding a polymer with negative charge
- Watson - Crick base pairing
- Hierarchical folding 1° → 2° → 3°

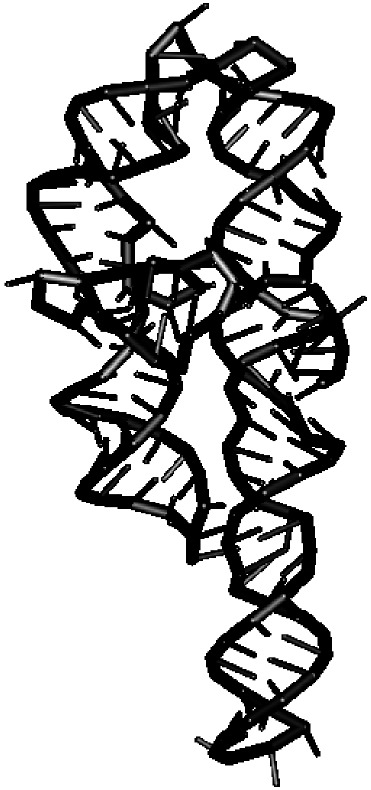
Figure from Dill &Chan (1997) Nat. Struct. Biol. Vol. 4, pp. 10-19

1° → 2° → 3°

AAUUGCGGGAAAGGGGUCAA
CAGCCGUUCAGUACCAAGUC
UCAGGGGAAACUUUGAGAUG
GCCUUGCAAAGGGUAUGGUA
AUAAGCUGACGGACAUGGUC
CUAACCACGCAGCCAAGUCC
UAAGUCAACAGAUCUUCUGU
UGAUAUGGAUGCAGUUCA

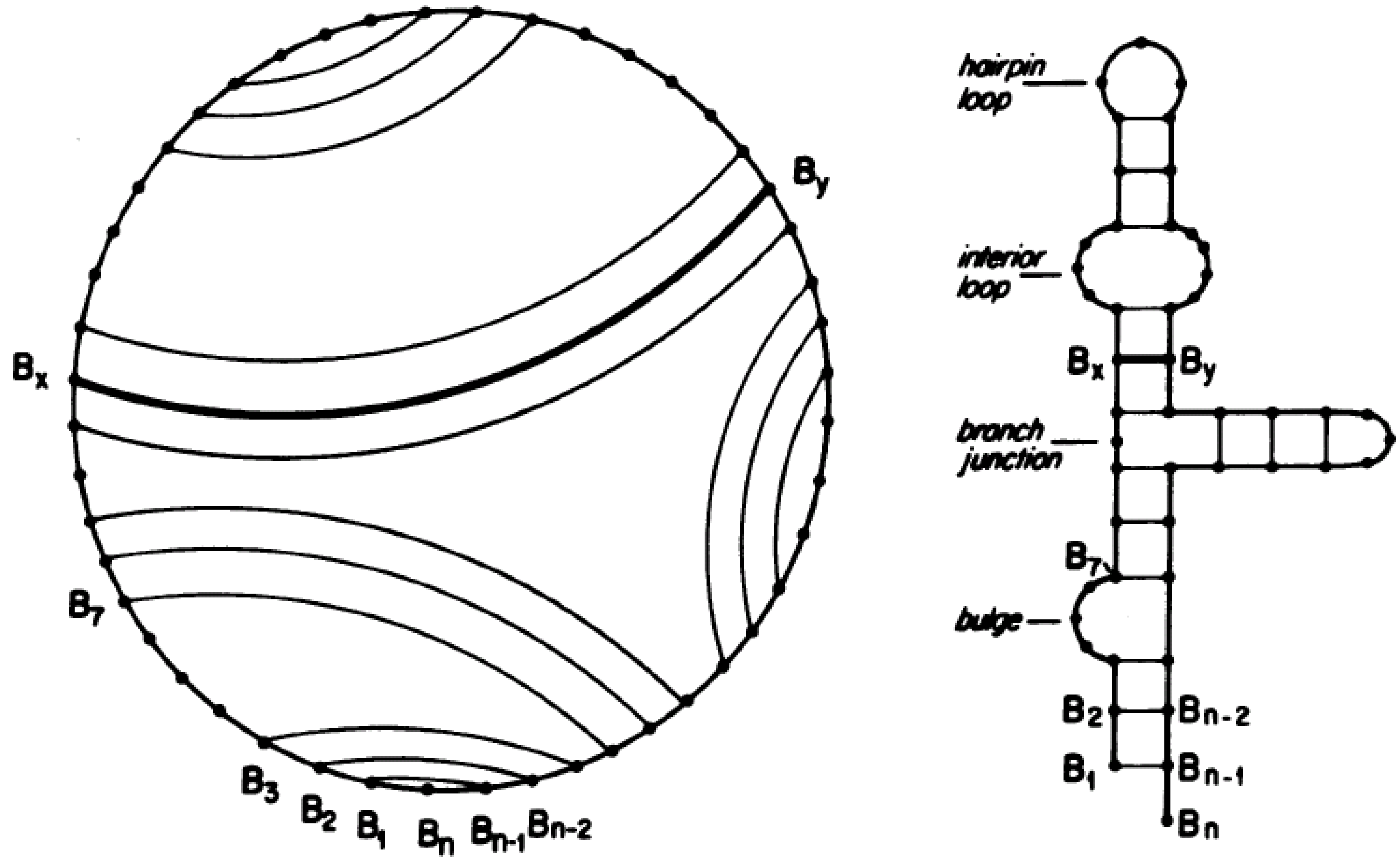


Waring &Davies 1984
Gene 28:277



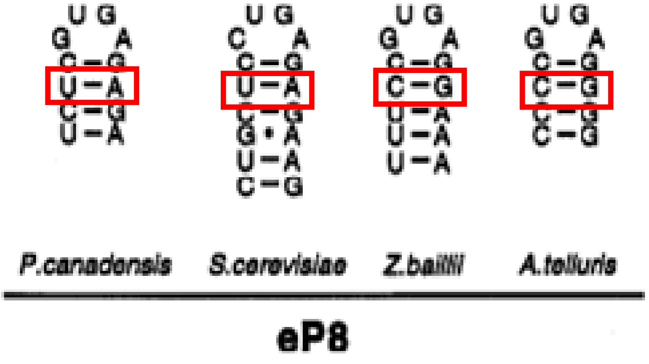
Cate et al. 1996
Science 273:1678

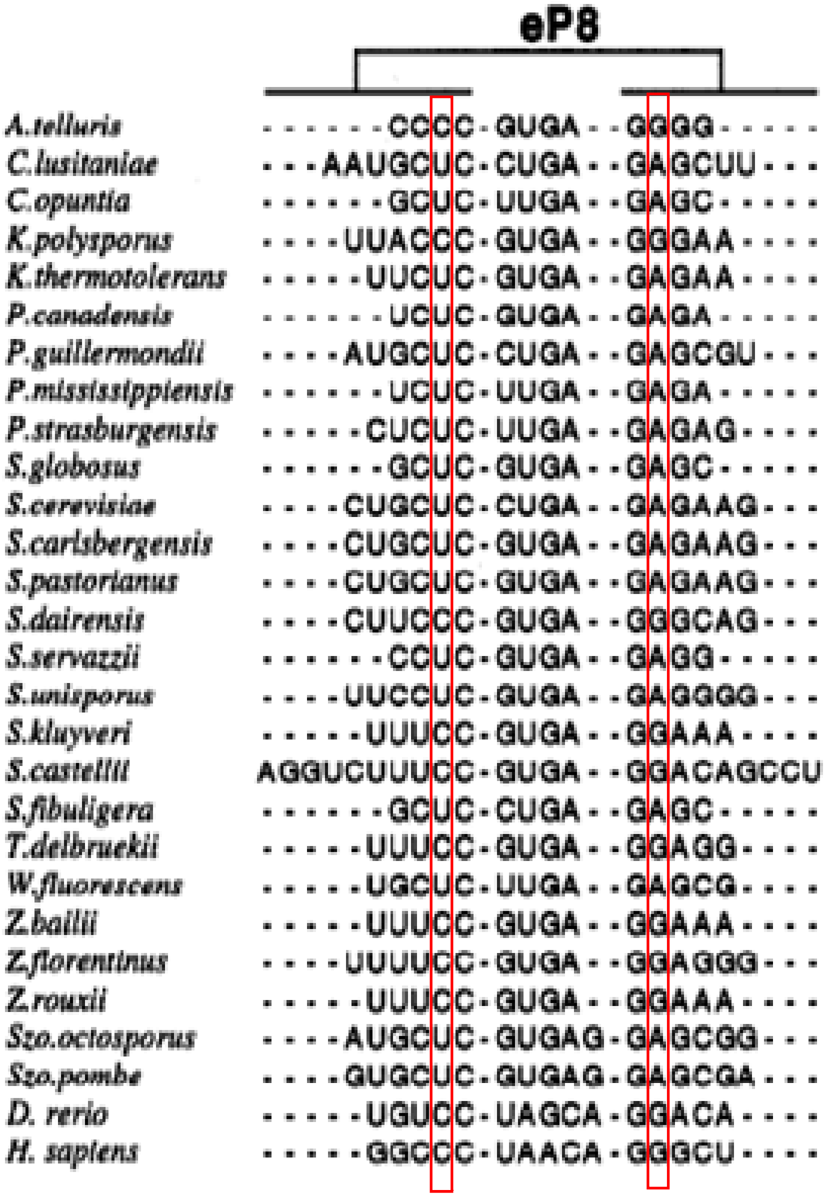
RNA Secondary Structure has Helices and Loops



Nussinov & Jacobson 1980 PNAS v 11 p 6310 Fig.1

An example of phylogenetic alignment and structure prediction for RNAse P



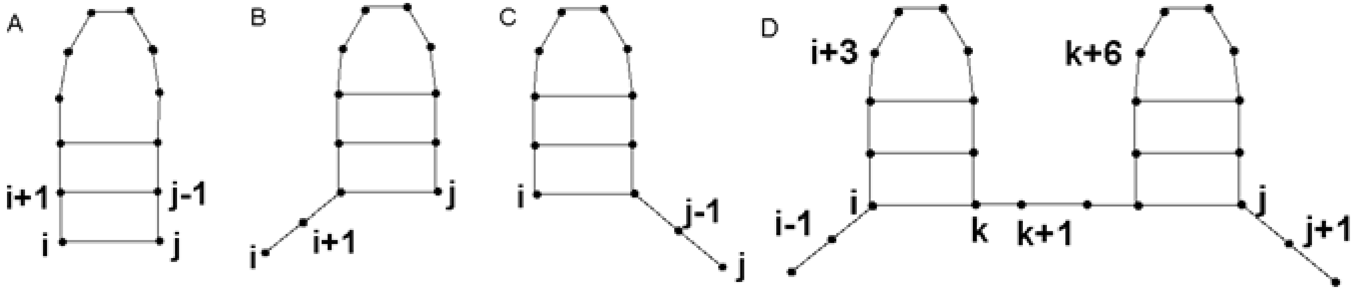


	eP8
A.telluris	------CCCC-GUGA--GGGG-----
C.lusitaniae	---AAUGCUC-CUGA--GAGCUU---
C.opuntia	------GCUC-UUGA--GAGC-----
K.polysporus	----UUACCC-GUGA--GGGAA----
K.thermotolerans	-----UUCUC-GUGA--GAGAA----
P.canadensis	------UCUC-GUGA--GAGA-----
P.guillermondii	----AUGCUC-CUGA--GAGCGU---
P.mississippiensis	------UCUC-UUGA--GAGA-----
P.strasburgensis	-----CUCUC-UUGA--GAGAG----
S.globosus	------GCUC-GUGA--GAGC-----
S.cerevisiae	----CUGCUC-CUGA--GAGAAG---
S.carlsbergensis	----CUGCUC-GUGA--GAGAAG---
S.pastorianus	----CUGCUC-GUGA--GAGAAG---
S.dairensis	----CUUCCC-GUGA--GGGCAG---
S.servazzii	------CCUC-GUGA--GAGG-----
S.unisporus	----UUCCUC-GUGA--GAGGGG---
S.kluyveri	-----UUUCC-GUGA--GGAAA----
S.castellii	AGGUCUUUCC-GUGA--GGACAGCCU
S.fibuligera	------GCUC-CUGA--GAGC-----
T.delbruekii	-----UUUCC-GUGA--GGAGG----
W.fluorescens	-----UGCUC-UUGA--GAGCG----
Z.bailii	-----UUUCC-GUGA--GGAAA----
Z.florentinus	----UUUUCC-GUGA--GGAGGG---
Z.rouxii	-----UUUCC-GUGA--GGAAA----
Szo.octosporus	----AUGCUC-GUGAG-GAGCGG---
Szo.pombe	----GUGCUC-GUGAG-GAGCGA---
D. rerio	-----UGUCC-UAGCA-GGACA----
H. sapiens	-----GGCCC-UAACA-GGGCU----

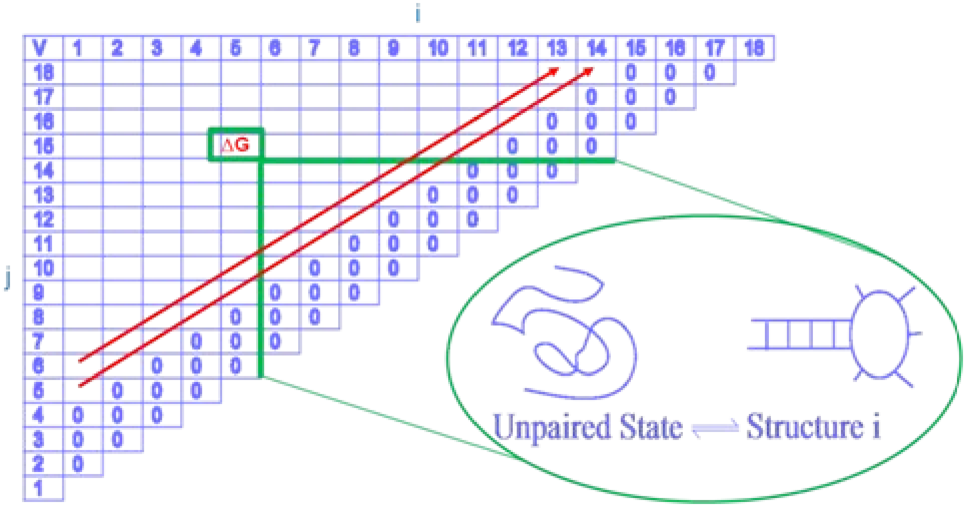
Frank et al. 2000 RNA v 6 p1895

How Dynamic Programming Algorithms Calculate RNA Secondary Structure

- Stochastic context-free grammar defines possible base pairs as 1 of 4 possible cases



- Recursion statement finds maximum value for each small subset of RNA sequence
- Fill an array with scores for each substructure
- Traceback through the array to find the lowest free energy structure
- O(N^2) memory storage
- O(N^3) runtime



V	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
18															0	0	0	
17														0	0	0		
16													0	0	0			
15					ΔG							0	0	0				
14											0	0	0					
13										0	0	0						
12									0	0	0							
11								0	0	0								
10							0	0	0									
9						0	0	0										
8					0	0	0											
7				0	0	0												
6			0	0	0													
5		0	0	0														
4	0	0	0															
3	0	0																
2	0																	
1																		

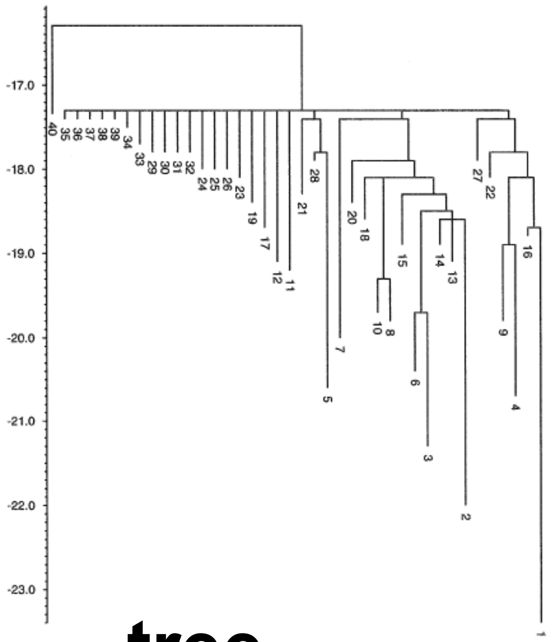
Websites for Folding Algorithms to Predict RNA 2°

MFOLD	http:// www.bioinfo.rpi.edu/applications/mfold
Vienna package	http:// www.tbi/univie.ac.at/~ivo/RNA
RNAStructure	http://rna.urmc.rochester.edu
SFOLD	http://sfold.wadsworth.org
PKNOTS	http:// selab.wustl.edu
STAR4.4	http://biology.leidenuniv.nl/~batenburg/STRAbout.html

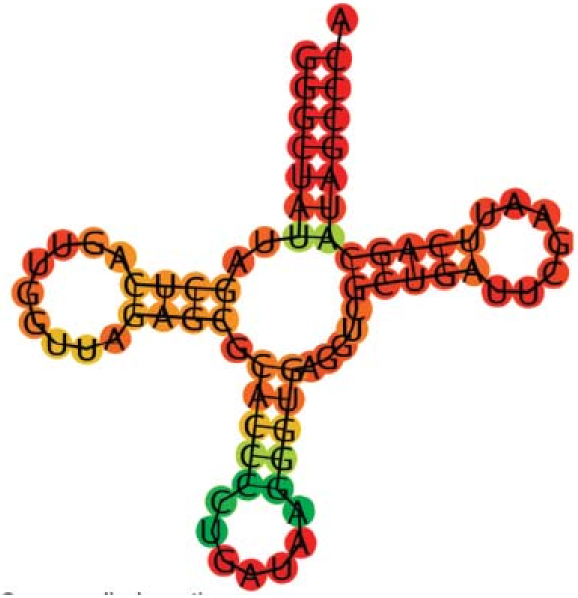
Representations of RNA Secondary Structure

```
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGCUGAUUCGAAUUCAGCAUAGCCCA
(((((((..((((.........)))).(((((.......))))).....(((((.......))))))))))))).
```

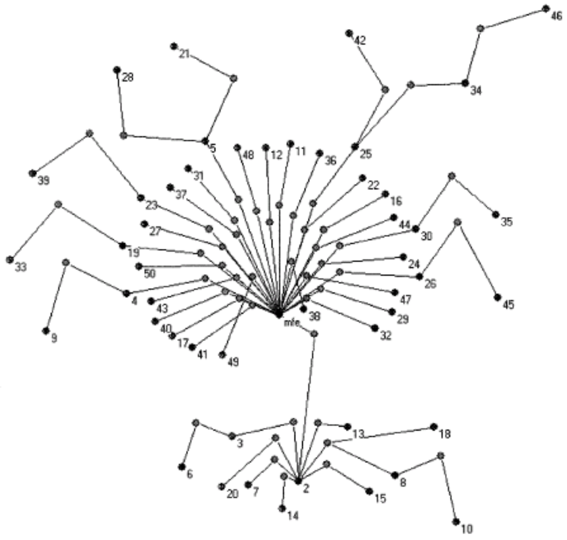
dots & parentheses



tree



graph



merged landscape

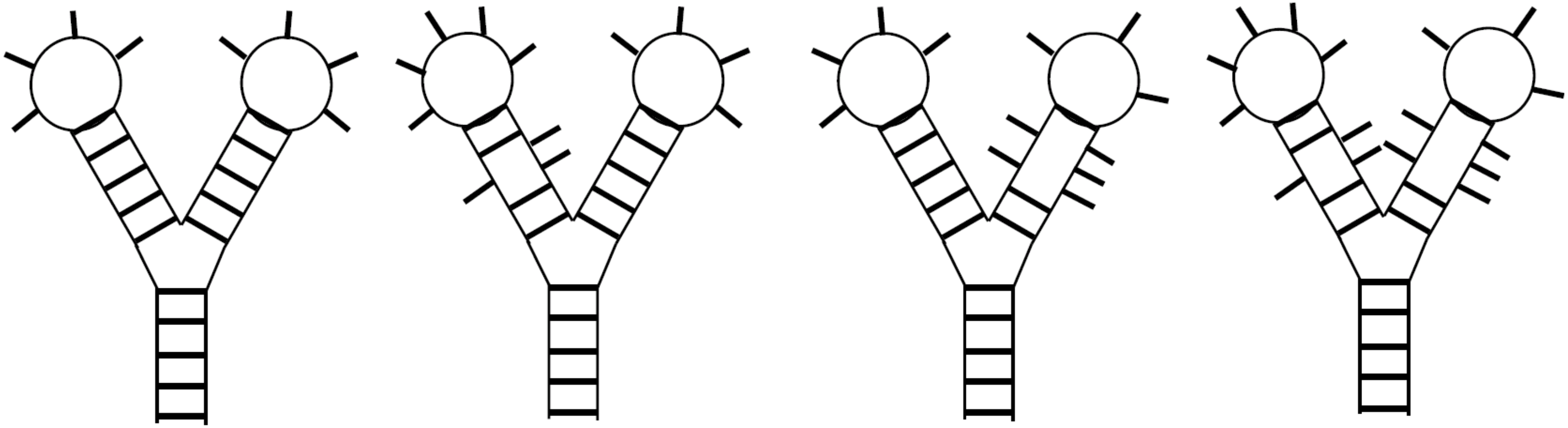
Wuchty 2003, Nucl. Acids Res. v 31, p 1115 Fig. 7
Gruber et al. 2008, Nucl. Acids Res. v 36, p. W73 Fig. 1

RNAStructure Predicts Secondary Structure Well

RNA	Lowest ΔG^{o} Structure	Best Suboptimal Structure
average	73%	87%
Group II introns	88%	94%
tRNA	87%	97%
5 S rRNA	74%	96%
Group I introns	69%	84%
SRP RNA	66%	88%
Rnase P	63%	76%
23 S rRNA (as domains)	55% (74%)	61% (88%)
16 S rRNA (as domains)	44% (61%)	54% (76%)

Mathews et al. (2004) *Proc. Natl. Acad. Sci.* vol. 101, pp. 7287-7292

Mfold and RNAStructure Sample Suboptimal Structures



Lowest Free Energy Structure

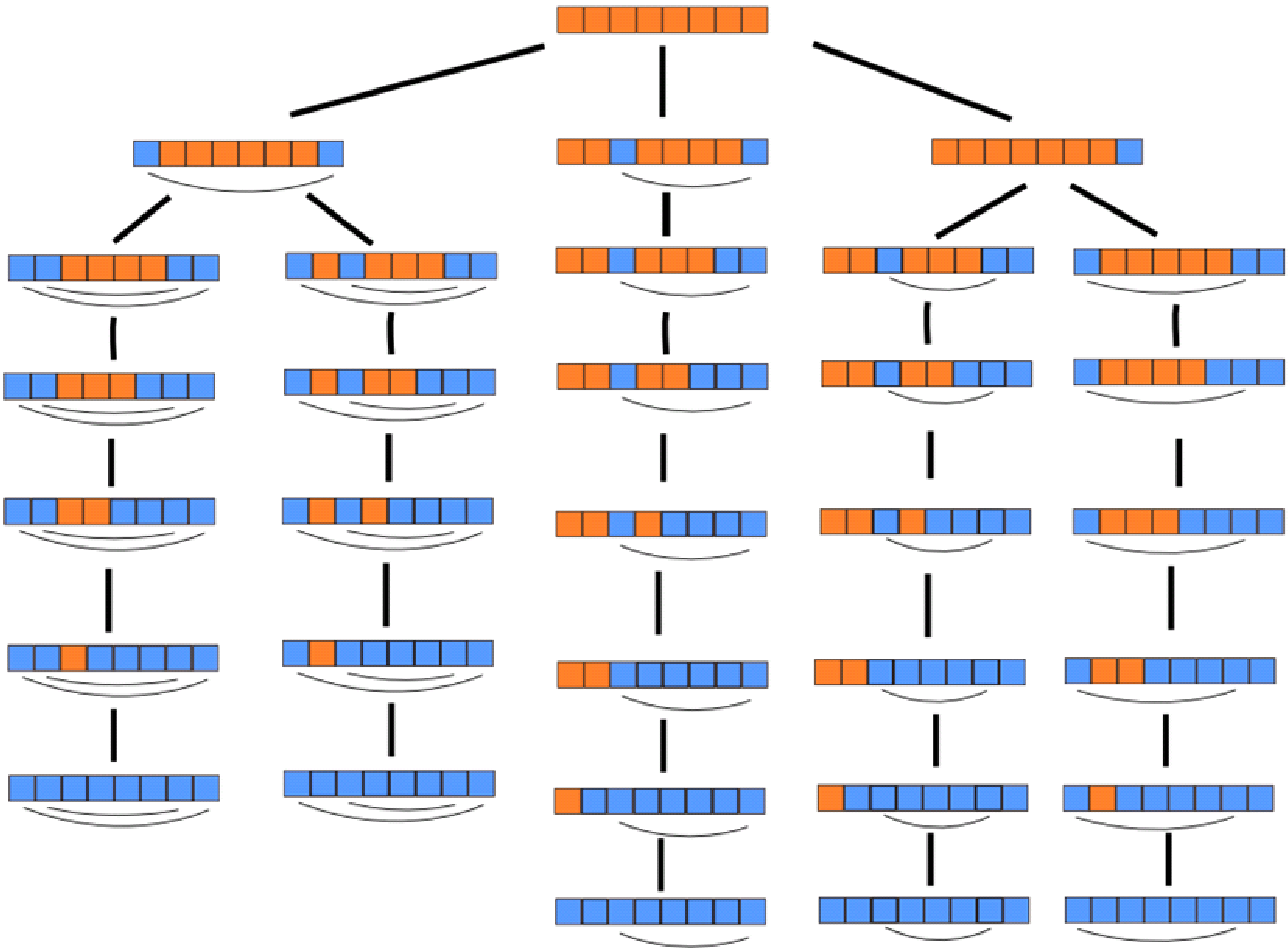
Suboptimal Structure 1

Suboptimal Structure 2

Structure that Cannot be Predicted

Mathews, 2006, J. Mol. Biol. V359 p. 528 Fig. 2

How Wuchty's Algorithm is like a Tree



STMV RNA folding problem

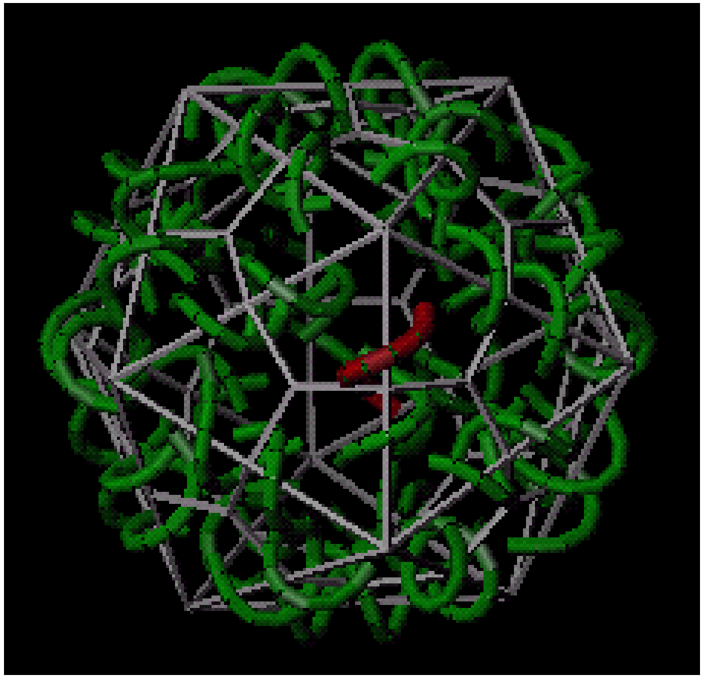
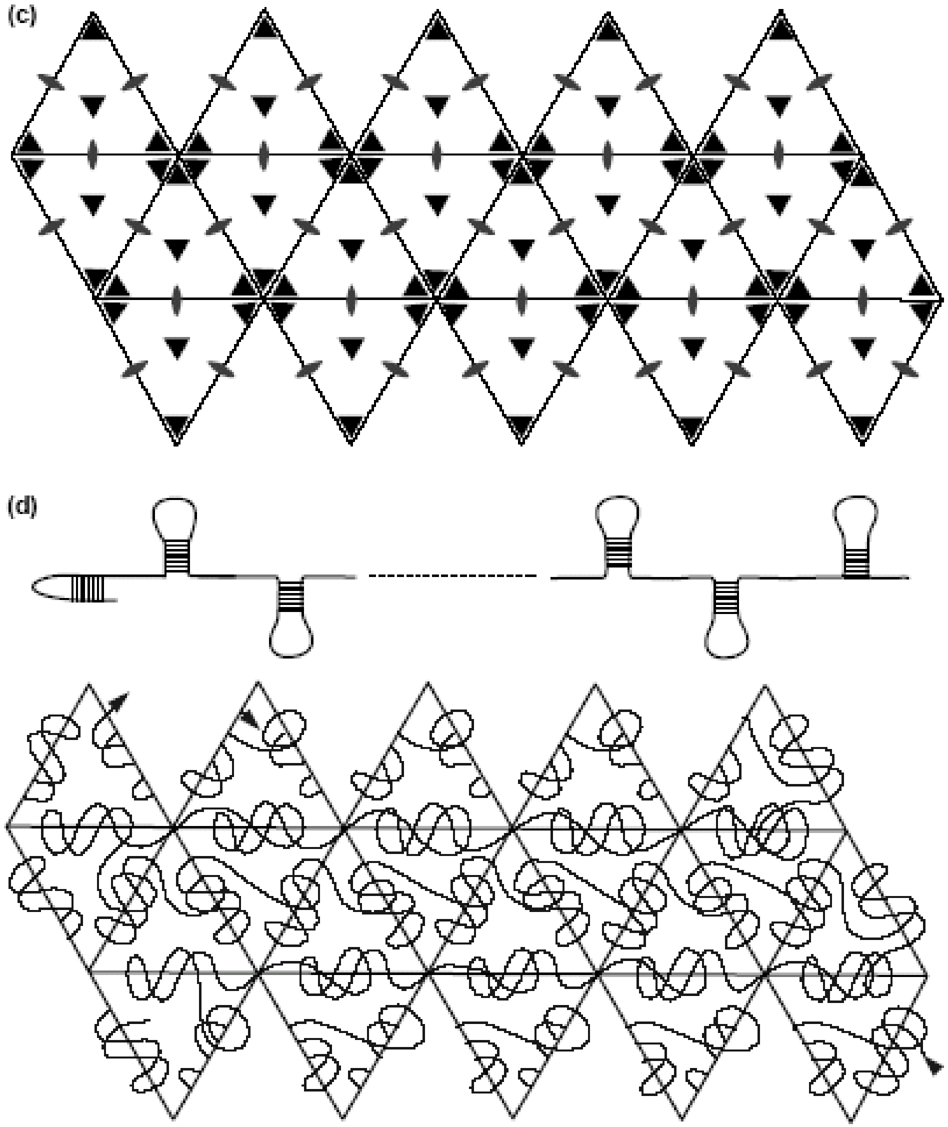


Figure reproduced from VIPER website
Reddy et al. 2001

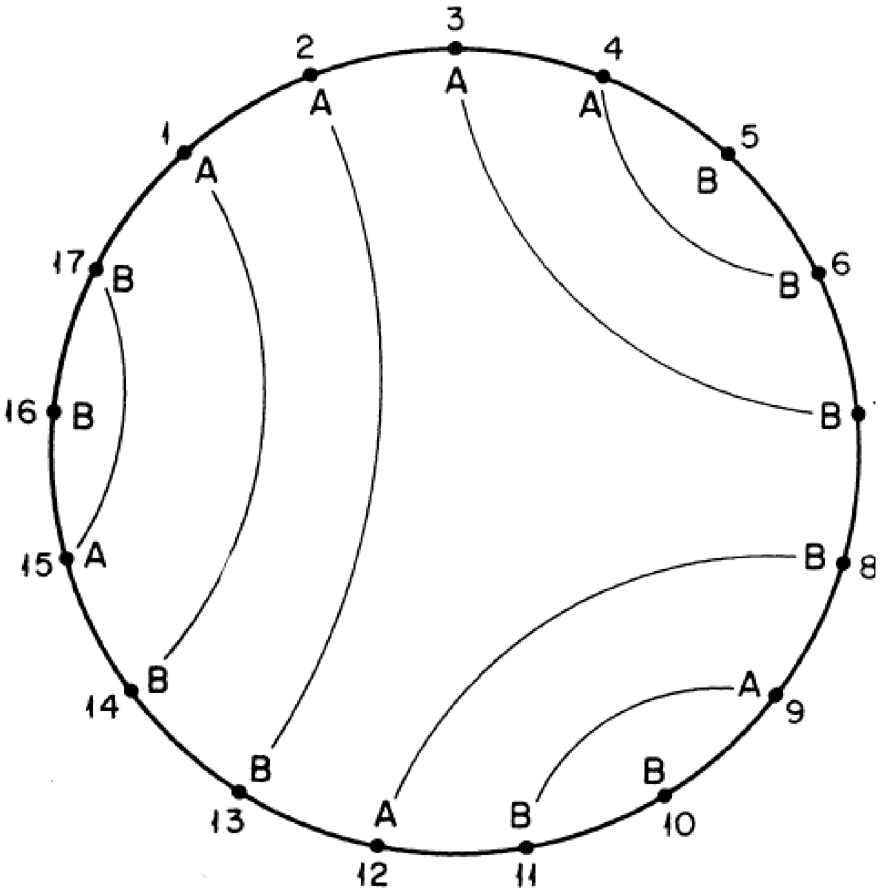
- Crystal structure to 1.8 Å resolution (Larson et al., 1998)
- 59% of the 1,058 RNA nucleotides are in helices
- RNA is icosahedrally averaged
- Identity of nucleotides in helices remains obscure
- Structure of 41% of the RNA remains unknown

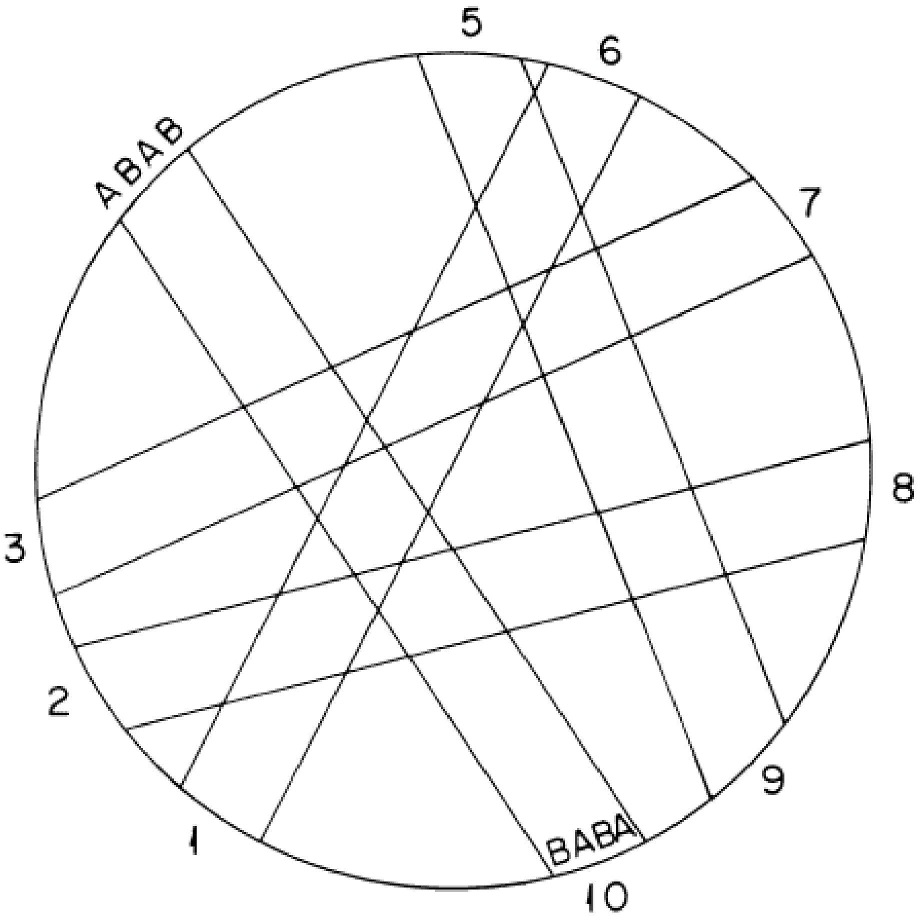
Current model for STMV RNA



Larson, S. & McPherson, A. Current Opinions in Structural Biology, vol. 11, p. 61.

Nussinov algorithms for maximizing matches and blocks

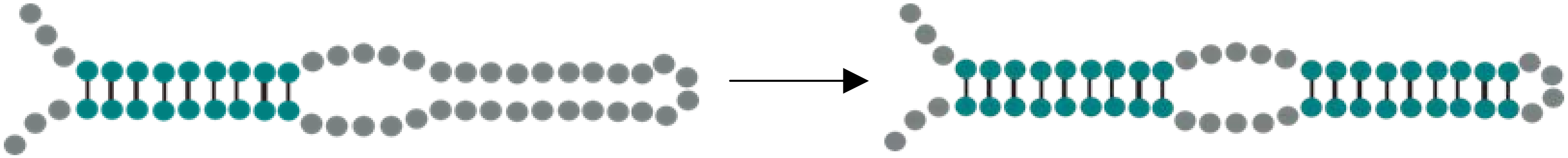




Nussinov et al. 1978 SIAM v35, p. 71,78 Fig. 1, 5

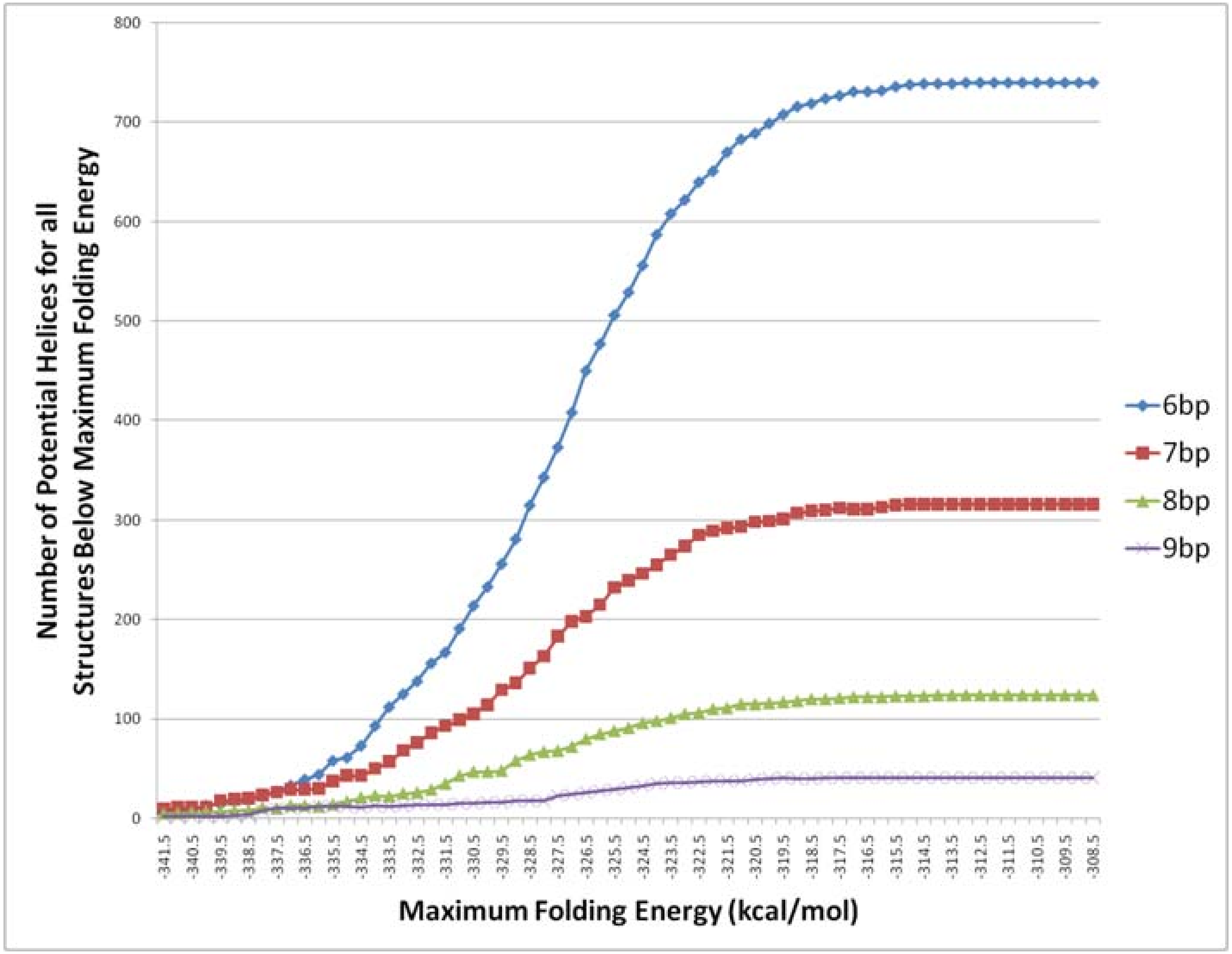
Combinatorial Search of STMV RNA

- **Locate potential helical structures between pairing bases *i* and *j***
- **Assemble non-overlapping potential helices *(i,j)* with *(p>j,q)* or *(k>i+l, k+l<q<j)***
- **Nested searching identifies "helices within helices"**

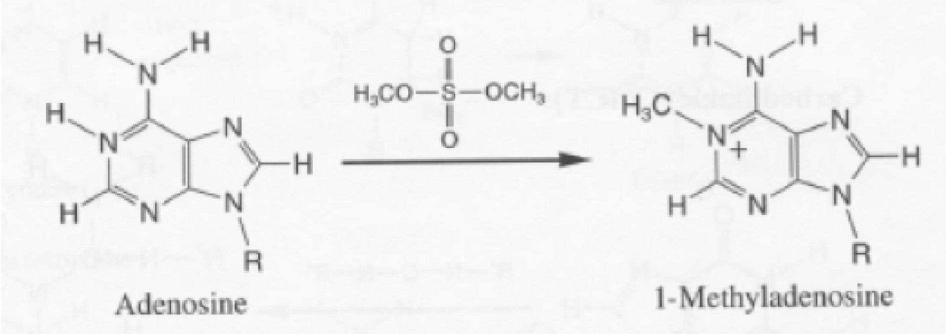


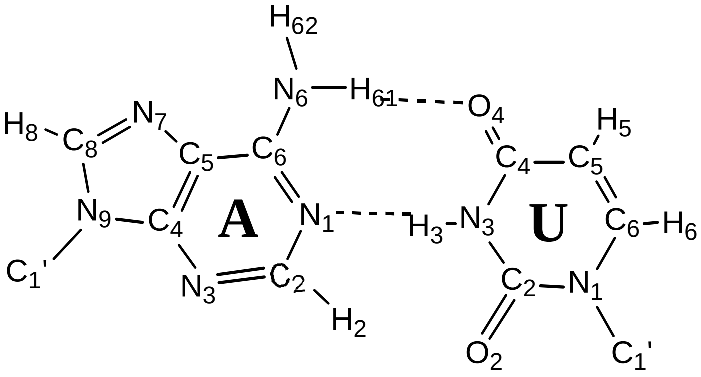
- **Over 144,000 perfect 6-pair helices, but no possible simultaneous combination of 30 helices in STMV RNA**

Many more possible structures contain 30 imperfect helices in the STMV sequence



Chemical modification data restrains possible base pairing





5’AAGAUGU AAA CCAGGA3’
3’CUGCA AA GGUCCU5’

5’AAGAUGU AAA CCAGGA3’
3’CUGCA AA GGUCCU5’

5’AAGAUGU AAA CCAGGA3’
3’CUGCA AA GGUCCU5’

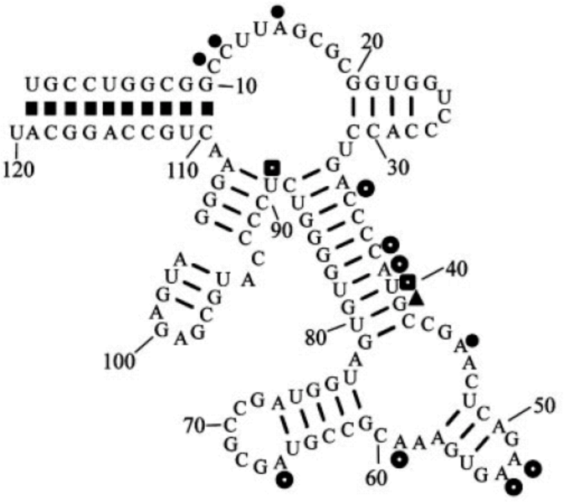
5’AAGAUGU AAA CCAGGA
3’CUGCA AA GGUCCU

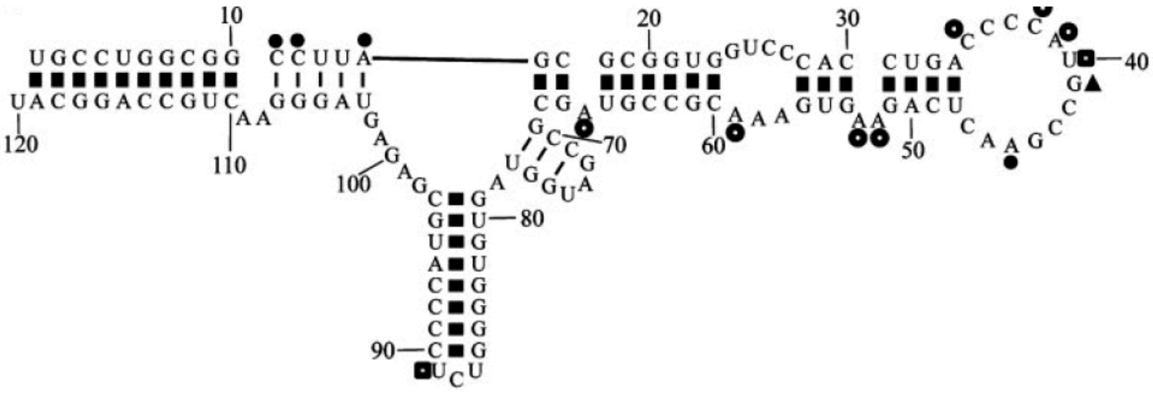
5’AAGAUGU AAA CCAGGA
3’CUGCA AA GGUCCU

✖ 5’AAGAUGU AAA CCAGGA
3’CUGCA AA GGUCCU

Including Results from Chemical Probing Improves Secondary Structure Prediction

E.coli 5S rRNA



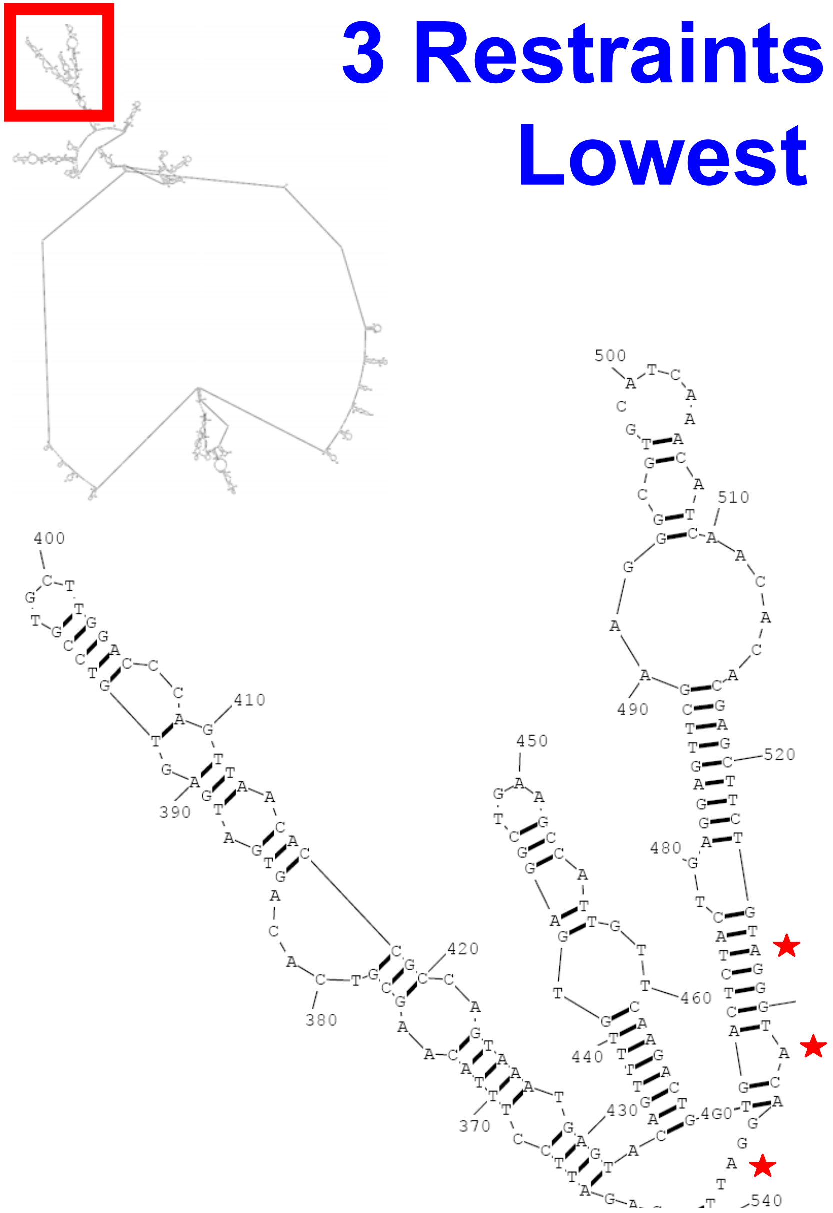


LOWEST 26.3 %
BEST 86.8 %

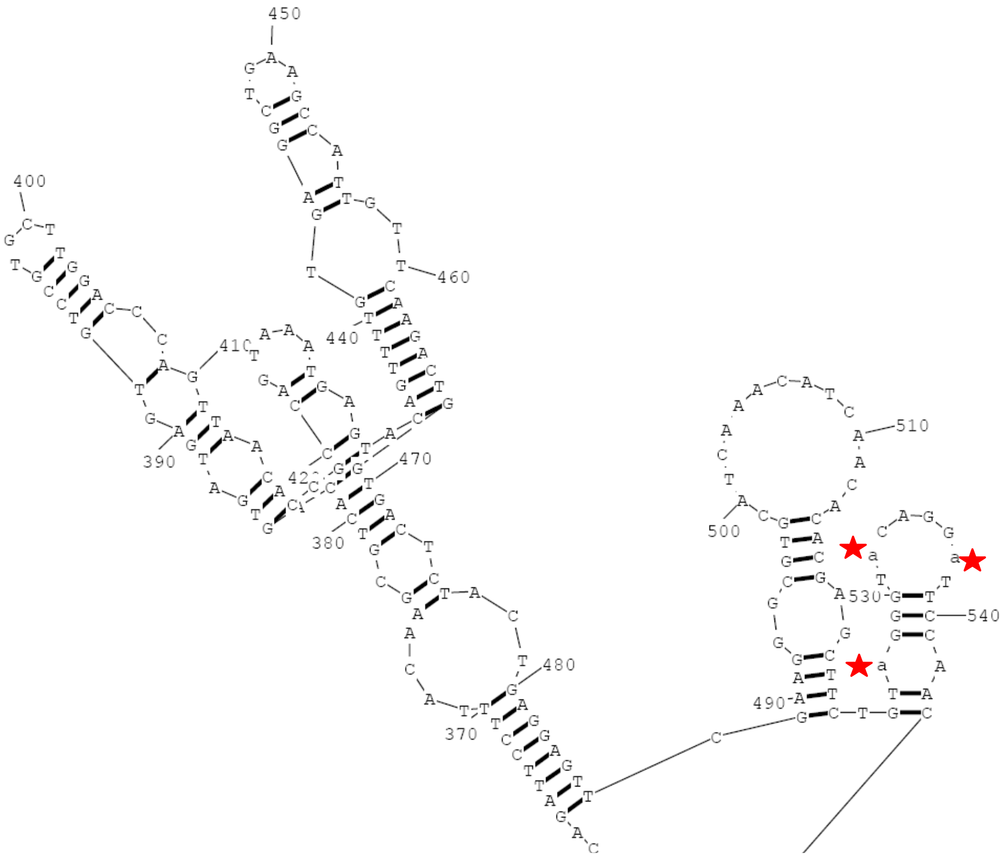
Folded with constraints from *in vivo* chemical modification
LOWEST 86.8 %
BEST 97.4 %

Mathews et al. (2004) *Proc. Natl. Acad. Sci.* **101**, 7290 Figure 1

3 Restraints Can Change the Lowest Energy Fold

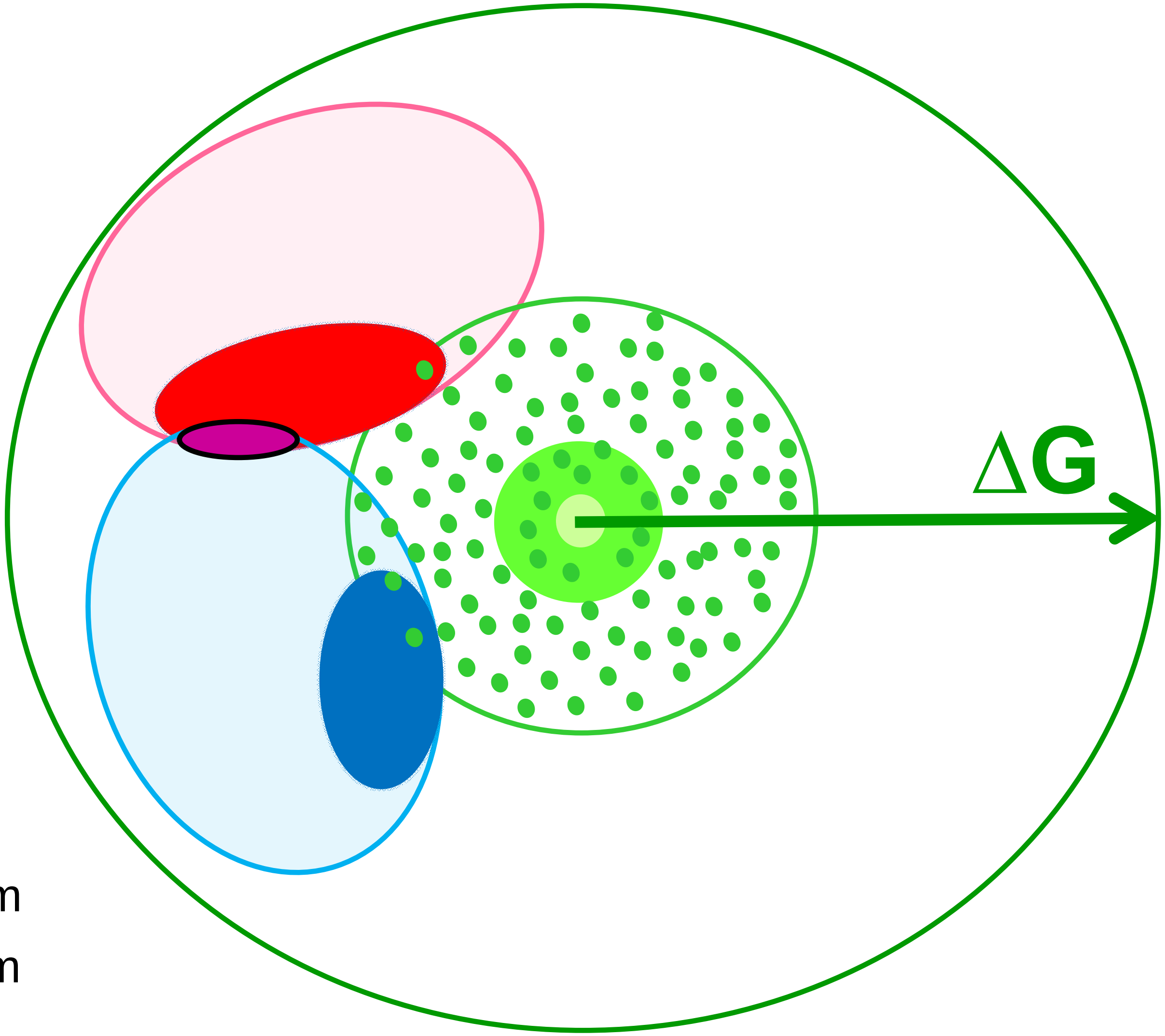


Native -341.1 kcal/mol



A527, A532, A537
restrained to be single stranded
-341.0.kcal/mol

Free Energy Landscape of STMV RNA



- 🟩 MFE
- 🟩 Wuchty
- ▨ Zuker
- 🟥 Chemical data
- 🟥 Future data
- 🟦 # Helices>1mm
- 🟦 # Helices,1 mm
- 🟪 Goal

How can Parallel Computing Help Solve the RNA Folding Problem?

- Utilize tree structure of RNA secondary structure prediction
- Expand range of free energies that can be computed for an RNA free energy landscape
- Explore more possible RNA structures

Acknowledgements

- Deb Mathews, University California Riverside
- David Mathews, University of Rochester
- Jeanmarie Verchot-Lubicz, Oklahoma State University
- Cal Lemke, Oklahoma University greenhouses

- Lab Members:

  Xiaobo Gu, Steven Harris, Koree Clanton-Arrowood, Brina Gendhar, Shelly Sedberry, Brian Doherty, Ted Gibbons, Sean Lavelle, John McGurk, Becky Myers, Mai-Thao Nguyen, Samantha Seaton, Jon Stone
- Funding

The University of Oklahoma 1890

NSF

PhRMA Foundation

OCAST

OU Cancer Institute