Secondary structure prediction for circular RNAs

Ronny Lorenz
ronny@tbi.univie.ac.at

Institute for Theoretical Chemistry
University of Vienna

Bled, Slovenia, February 23, 2008

There are circular RNAs?

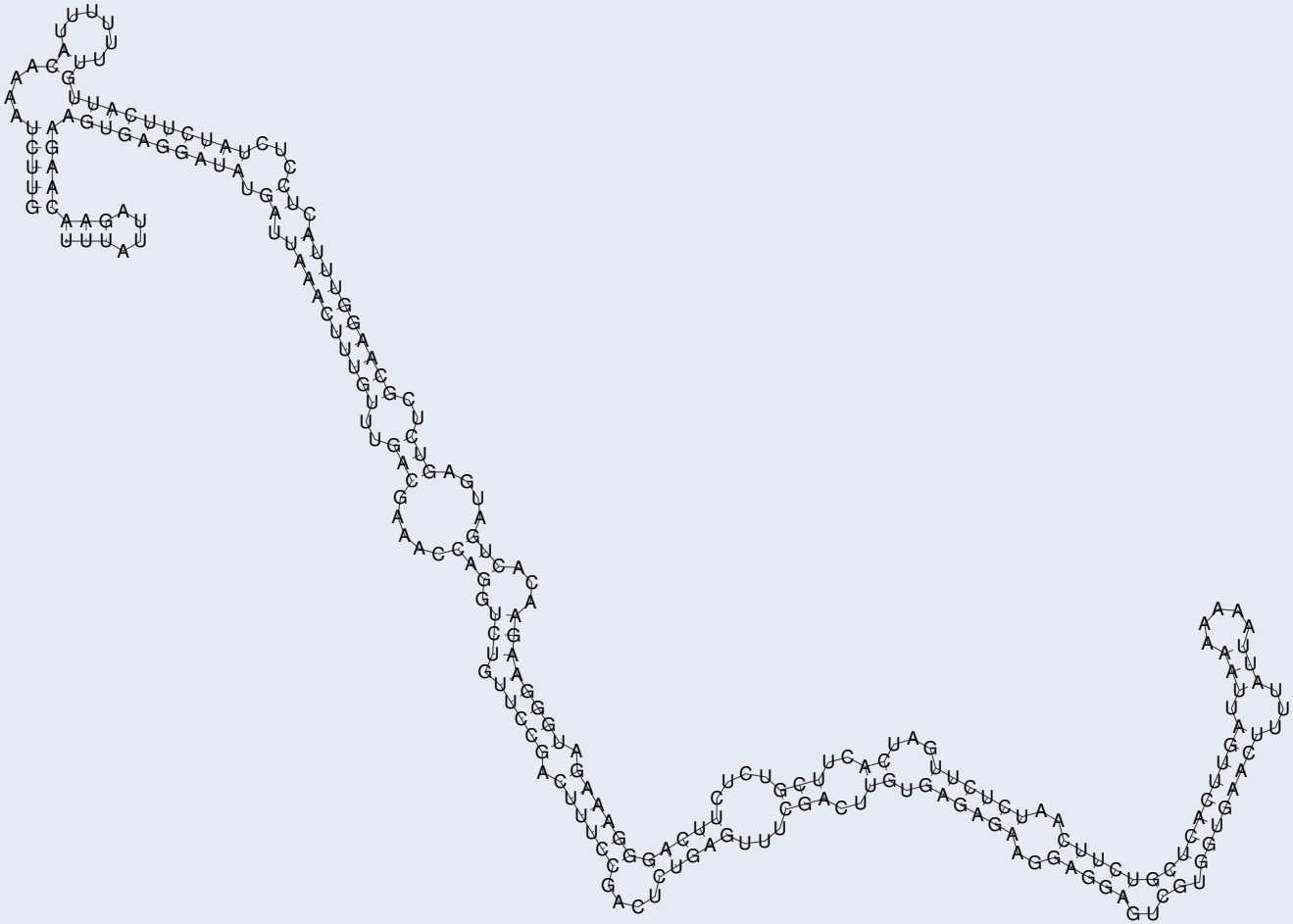
There are circular RNAs?

- Spliced group I introns form circles
- Some tRNA splicing products in archaea
- Box C/D RNAs in Pyrococcus furiosus
- Hepatitis δ Virus (HDV)
- Viroids
- Satellite RNAs
- RNA aptameres

More than 1300 circular viroid RNA genomes and related objects in the `Subviral RNA Database`

In almost all cases, circular RNAs code not for any mRNA but for their own structure

Predicted MFE secondary structure of *ASBVd*

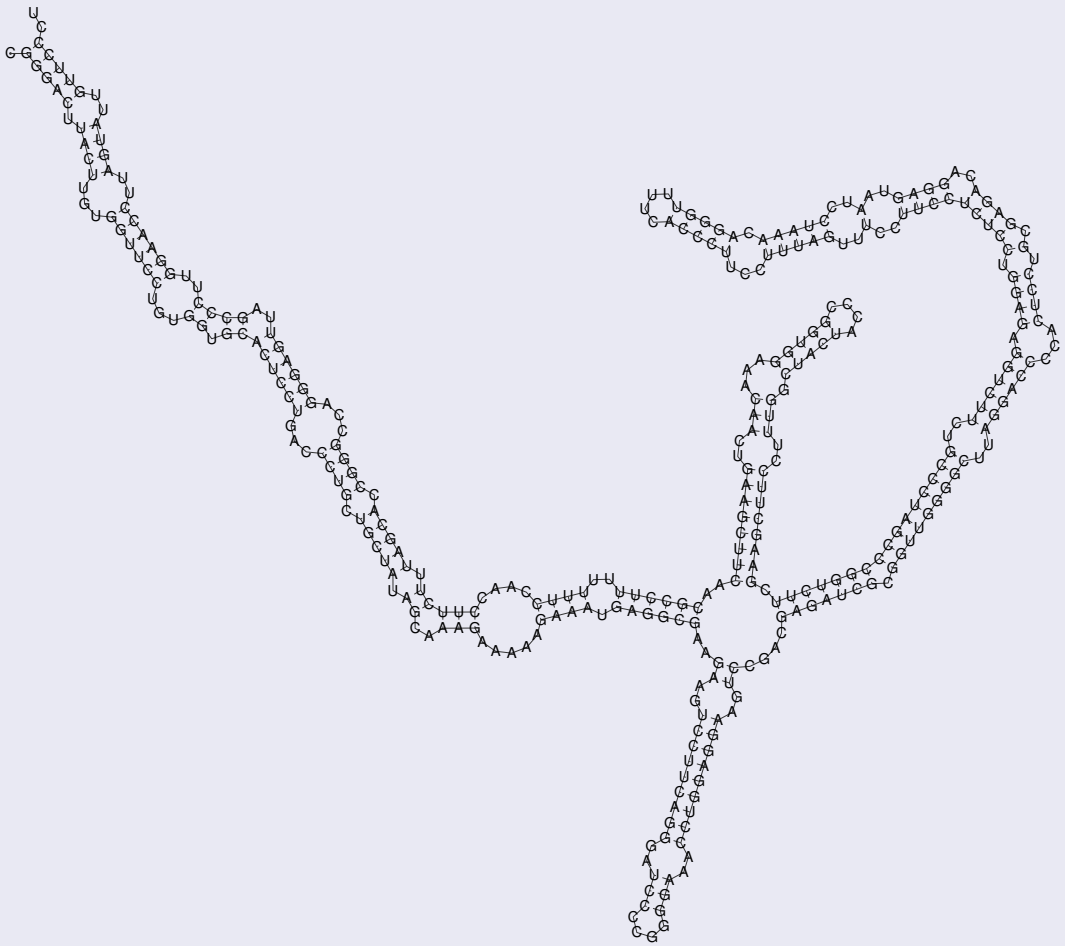


(Avocado Sunblotch Viroid)

Symptoms of *ASBVd* on fruit

(Image taken from "2006 Florida Plant Disease Management Guide: Avocado (Persea americana)", Palmateer, A.J. and Ploetz, R.C. and Harmon, P.F.)

Predicted MFE secondary structure of *CSVd*

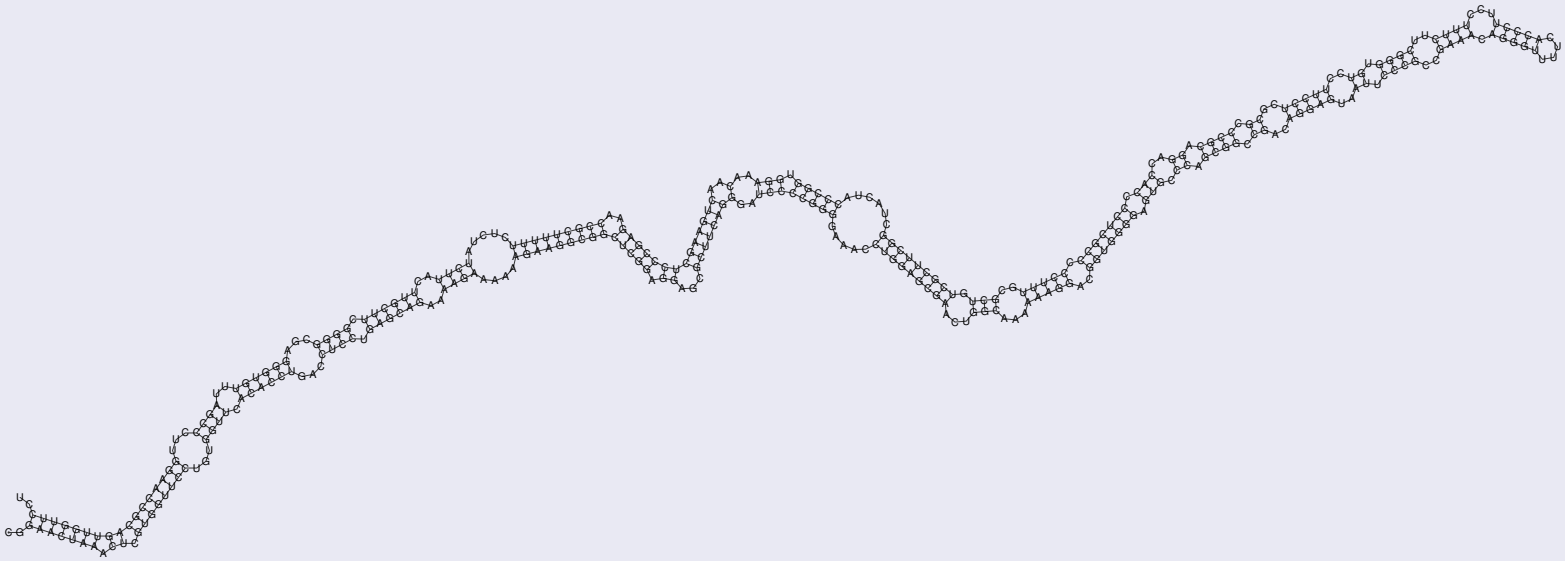


(Chrysanthemum Stunt Viroid)

Symptoms of *CSVd* infection on flowers of chrysanthemum cv. Gillglow

(Image taken by J. Dunez, France, Bugwood.org)

Predicted MFE secondary structure of *PSTVd*



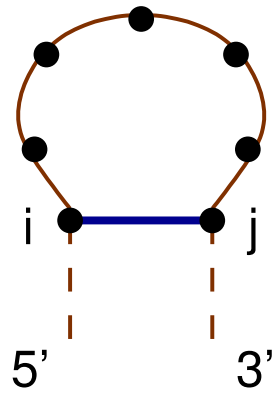
(Potato Spindle Tuber Viroid)

PSTVd on potato

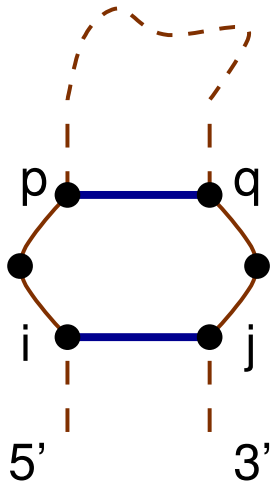
(Image taken by USDA ARS Archive, USDA Agricultural Research Service, USA, Bugwood.org)

Loop-based energy model:

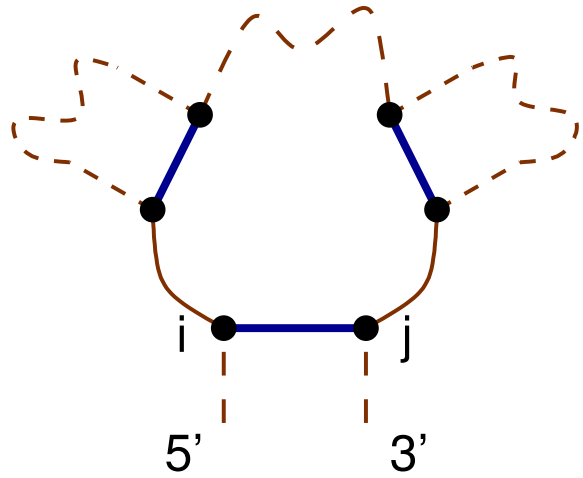
- free energy of a secondary structure is additively composed by free energy of its loops



Hairpin loop



Interior loop



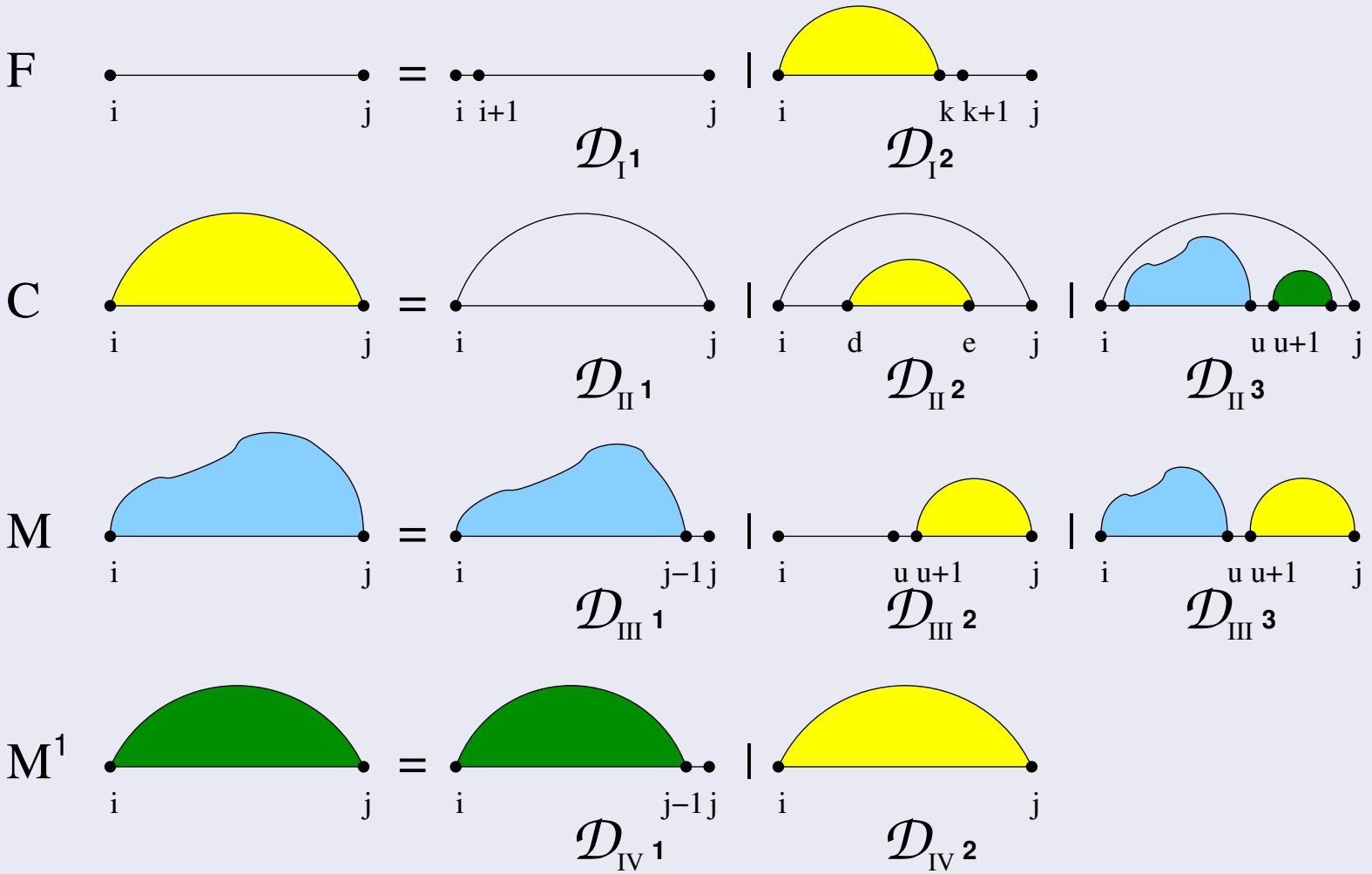
Multi loop

Dynamic programming algorithms:

Dynamic programming algorithms:

- optimal secondary structure according to free energy (MFE)
- suboptimal secondary structures with free energy in an interval arround MFE
- partition function and base pairing probabilities
- statiscally representative samples of the Boltzmann ensemble (stochastic backtracking)

Loop decomposition for linear RNAs



Whats the difference when predicting secondary structures for circular RNAs?

Whats the difference when predicting secondary structures for circular RNAs?

- consecutive numbering: circles have no beginning

Whats the difference when predicting secondary structures for circular RNAs?

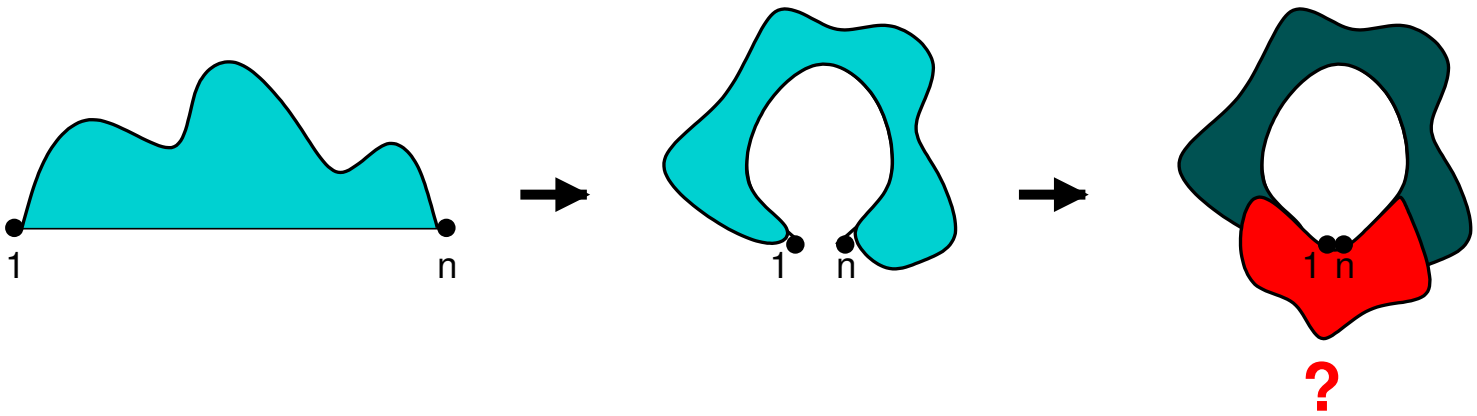
- consecutive numbering: circles have no beginning
  Cut the circle at an arbitrary point to obtain a linear sequence of nucleotides

Whats the difference when predicting secondary structures for circular RNAs?

- consecutive numbering: circles have no beginning
  Cut the circle at an arbitrary point to obtain a linear sequence of nucleotides
  And now treat it as if it is a linear RNA?

Whats the difference when predicting secondary structures for circular RNAs?

- consecutive numbering: circles have no beginning
  Cut the circle at an arbitrary point to obtain a linear sequence of nucleotides
  And now treat it as if it is a linear RNA?



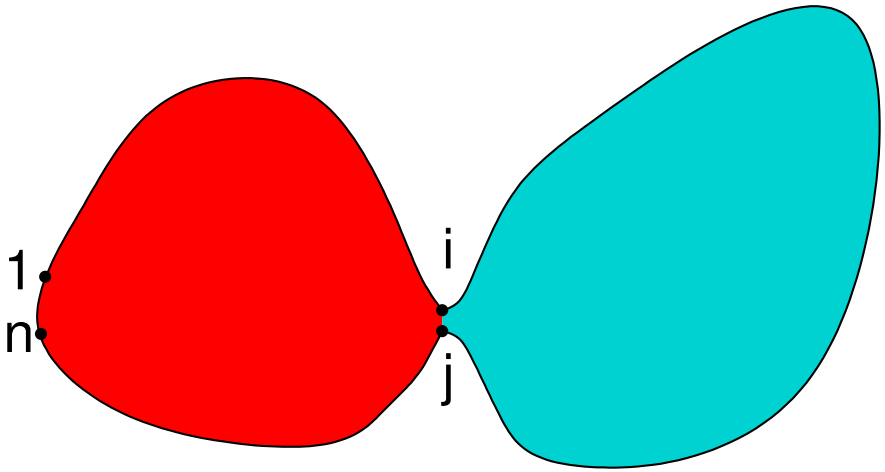
How to extend linear recursions for circular RNAs?

How to extend linear recursions for circular RNAs?

If at least one basepair exists:

(Hofmann et al., J. Biomol. Struct. Dyn., 1984)

- Take free energy of exterior loop and free energy of interior loop



$$\min_{i<j} \{C(j,i) + C(i,j)\}$$

- doubles computation time and memory requirements!

How to extend linear recursions for circular RNAs?
If at least one basepair exists:
(Zuker et al., Bulletin of Mathematical Biology, 1984)

- Concatenate sequence [1,n] on itself so that nucleotides $n+1, \ldots, 2n$ are the same as $1, \ldots, n$
- fill C matrix with linear algorithm and new condition:

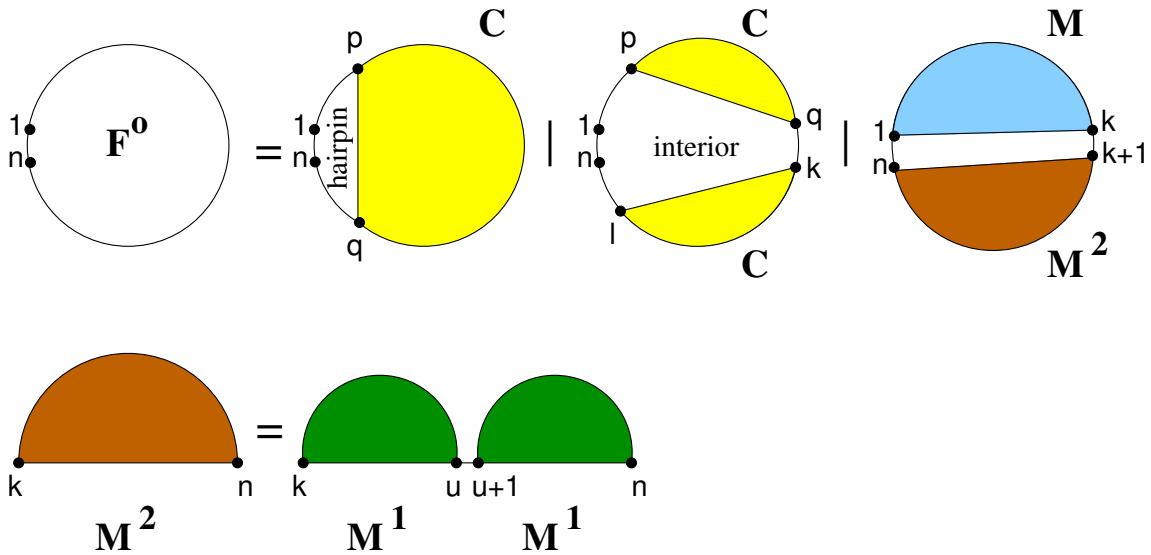
$$C(i,j) = \infty, \text{ if } j - i > n - 2$$

$$\min_{i<j} \{C(i,j) + C(j, i+n)\}$$

- quadruples the memory requirements and roughly triples computation time

How to extend linear recursions for circular RNAs?
If at least one basepair exists:
(Hofacker et al., Bioinformatics, 2005)

- Compute energy of the exterior loop as a kind of post-processing step



- $\mathcal{O}(n)$ additional memory, roughly doubling of computation time

Memory efficient algorithm was applied to:

Memory efficient algorithm was applied to:

- partition function and base pairing prob. for single sequences
- partition function and base pairing prob. for aligned sequences
- algorithms for computing suboptimal secondary structures (structures within an energy range δ arround the MFE and stochastic backtracking)

Memory efficient algorithm was applied to:

- partition function and base pairing prob. for single sequences
- partition function and base pairing prob. for aligned sequences
- algorithms for computing suboptimal secondary structures (structures within an energy range δ arround the MFE and stochastic backtracking)

These are now implemented in the `ViennaRNAPackage` and available in the programs: `RNAfold`, `RNAalifold` and `RNAsubopt`

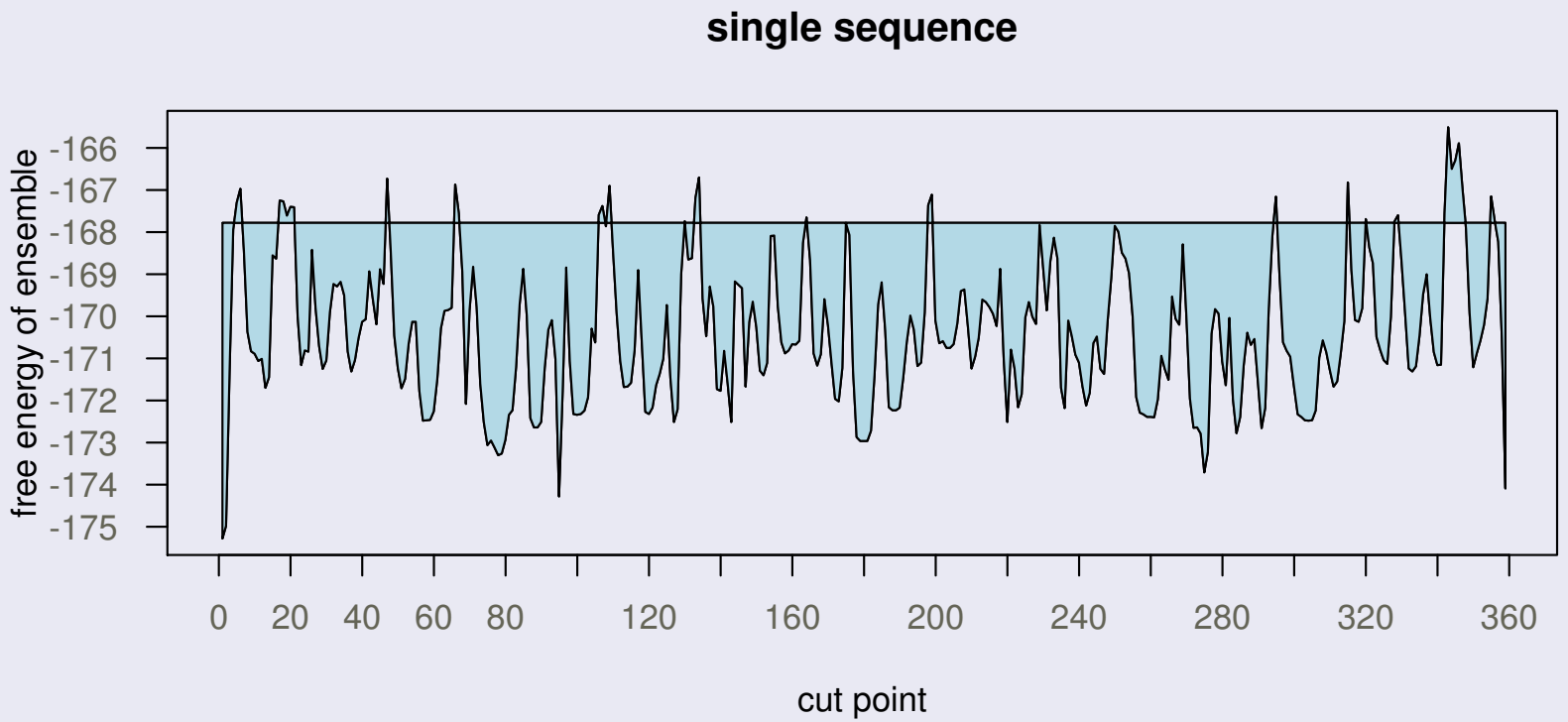
What can one do now?

What can one do now?

- predict secondary structures for even large circular RNAs
- compute partition function and base pairing probabilities
- generate samples of suboptimal secondary structures
- handle circular AND linear RNAs with the same tools
- ...

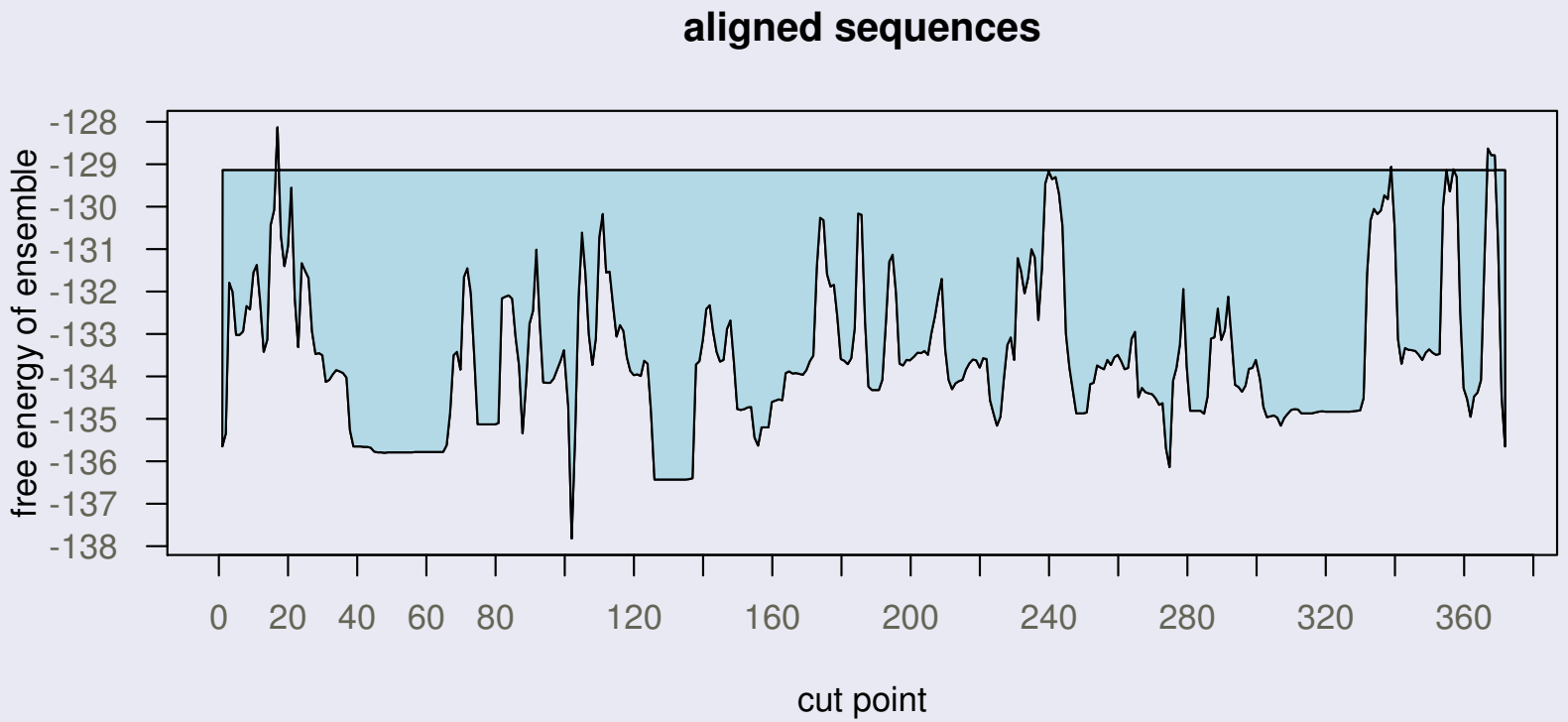
- *investigate if theres a difference between folding linear and circular (cut point specificity)*

Free energy of ensemble of an RNA sequence



Predicted Gibbs free energy $G = -kT \cdot ln(\mathcal{Q})$ compared to
$G^{\circ} = -167.78\ kcal/mol$
(*PSTVd, Acc. No. M16826*)

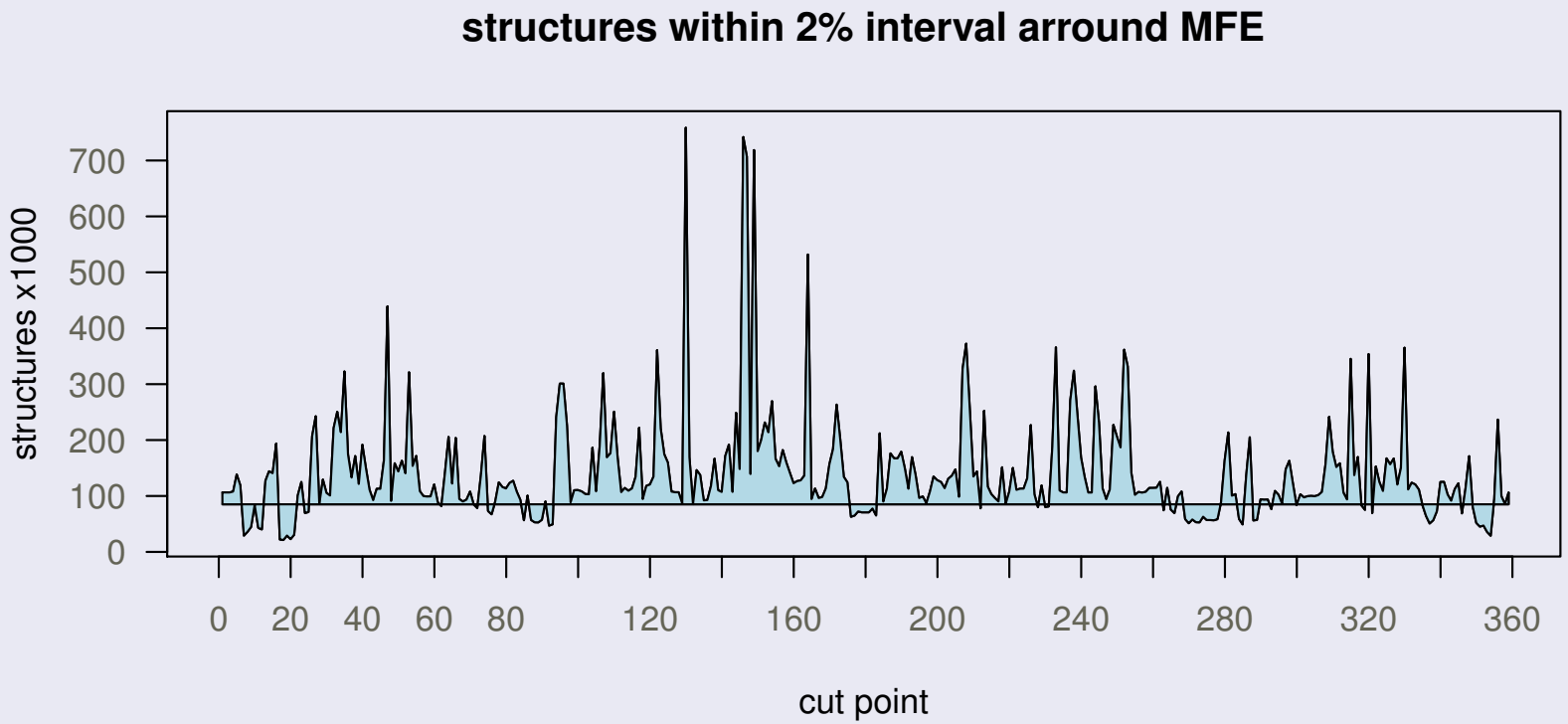
Free energy of ensemble of an alignment of RNA sequences



Predicted Gibbs free energy $G^{\mathcal{A}} = -kT \cdot ln(\mathcal{Q}^{\mathcal{A}})$ of an alignment compared to $G^{\mathcal{A}\circ} = -129.14\ kcal/mol$

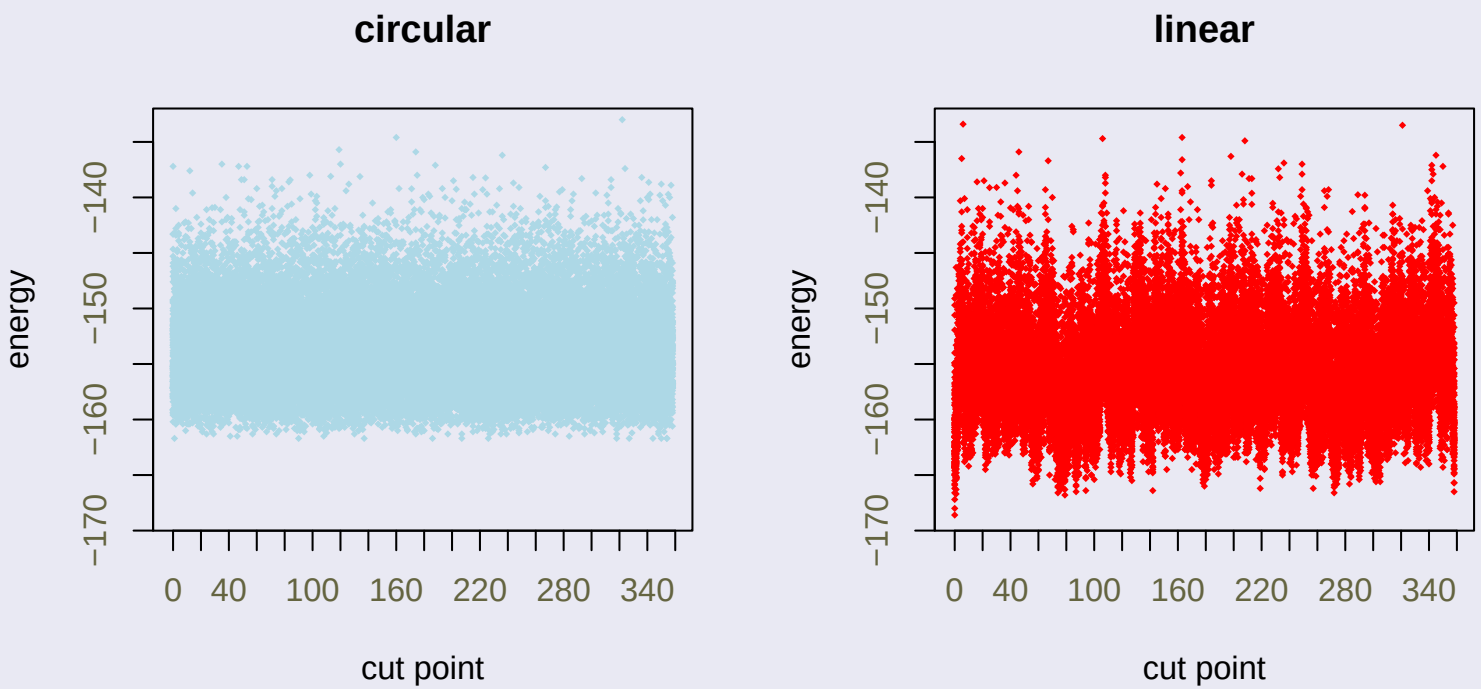
(134 *PSTVd* seq. from `Subviral RNA database`)

Suboptimal secondary structures



Number of secondary structures when folding linear compared to 85085 structures obtained by folding with circular extension (*PSTVd, Acc. No. M16826*)

Stochastic backtracking



Free energies in *kcal/mol* of stochastically sampled suboptimal structures (*PSTVd, Acc. No. M16826*)

Searching for multistable secondary structure states, folding paths and folding kinetics

- Viroids are known to exihibit different secondary structures, e.g. during replication and while acting pathogenic
- sequence too long for an analysis!
  e.g. PSTVd: 360nt, MFE: -160.3 kcal/mol

  Subopt Wuchty:
  24.000.000 within interval of 5.7 kcal/mol

  Stochastic backtracking:
  17.900.000 sampled
  11.000.000 after removing duplicates

Searching for multistable secondary structure states, folding paths and folding kinetics

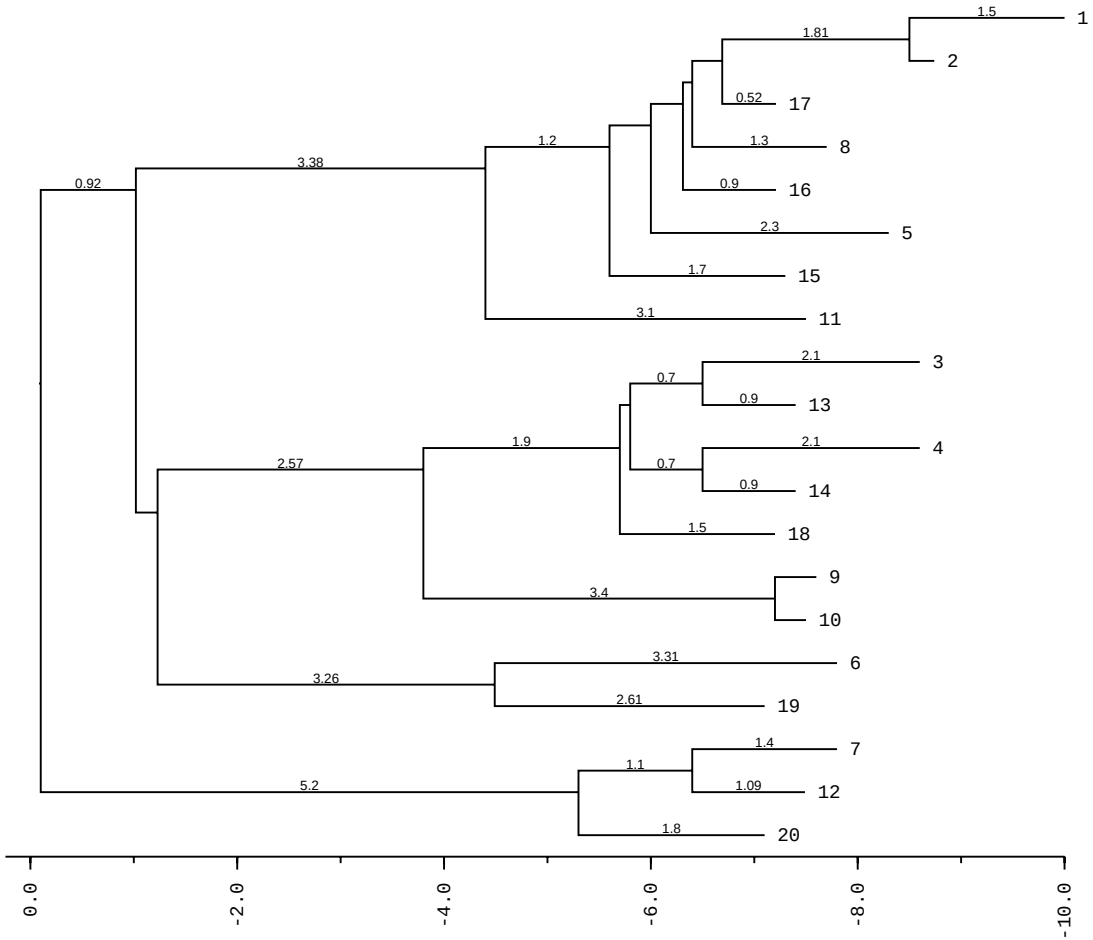
- Viroids are known to exihibit different secondary structures, e.g. during replication and while acting pathogenic
- sequence too long for an analysis!
  e.g. PSTVd: 360nt, MFE: -160.3 kcal/mol

  Subopt Wuchty:
  24.000.000 within interval of 5.7 kcal/mol

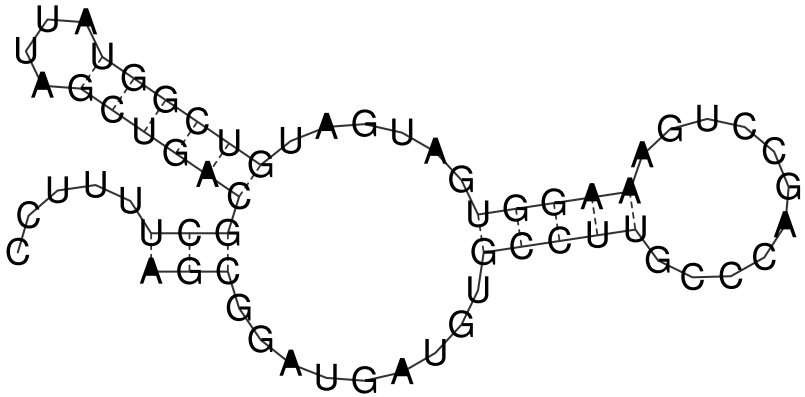
  Stochastic backtracking:
  17.900.000 sampled
  11.000.000 after removing duplicates
  this gets even worse with bigger samples ($>$ 50% duplicates)
  energy landscape is not well connected

Searching for multistable secondary structure states and folding paths and folding kinetics

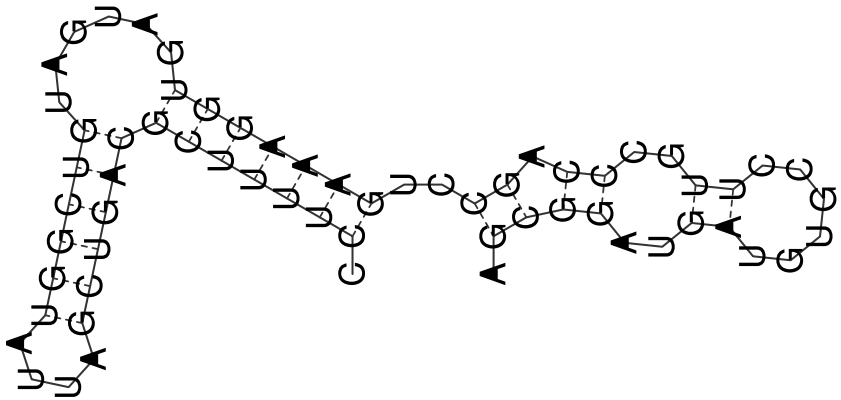
- small circular sRNA (sR29, 63nt) of *P. furiosus*
- `RNAsubopt` generated secondary structure space (6.5M structs)
- use these structures as input for barrier tree generation with `barriers` (20 local minima)
- select two stable structures (local minima) and compute the folding path


1.5
1
1.81
2
0.52
17
1.3
8
1.2
3.38
0.92
0.9
16
2.3
5
1.7
15
3.1
11
2.1
3
0.7
0.9
13
1.9
2.1
4
2.57
0.7
0.9
14
1.5
18
9
3.4
10
3.31
6
3.26
2.61
19
1.4
7
1.1
5.2
1.09
12
1.8
20
0.0
-2.0
-4.0
-6.0
-8.0
-10.0

select for example structure at local minimum 4

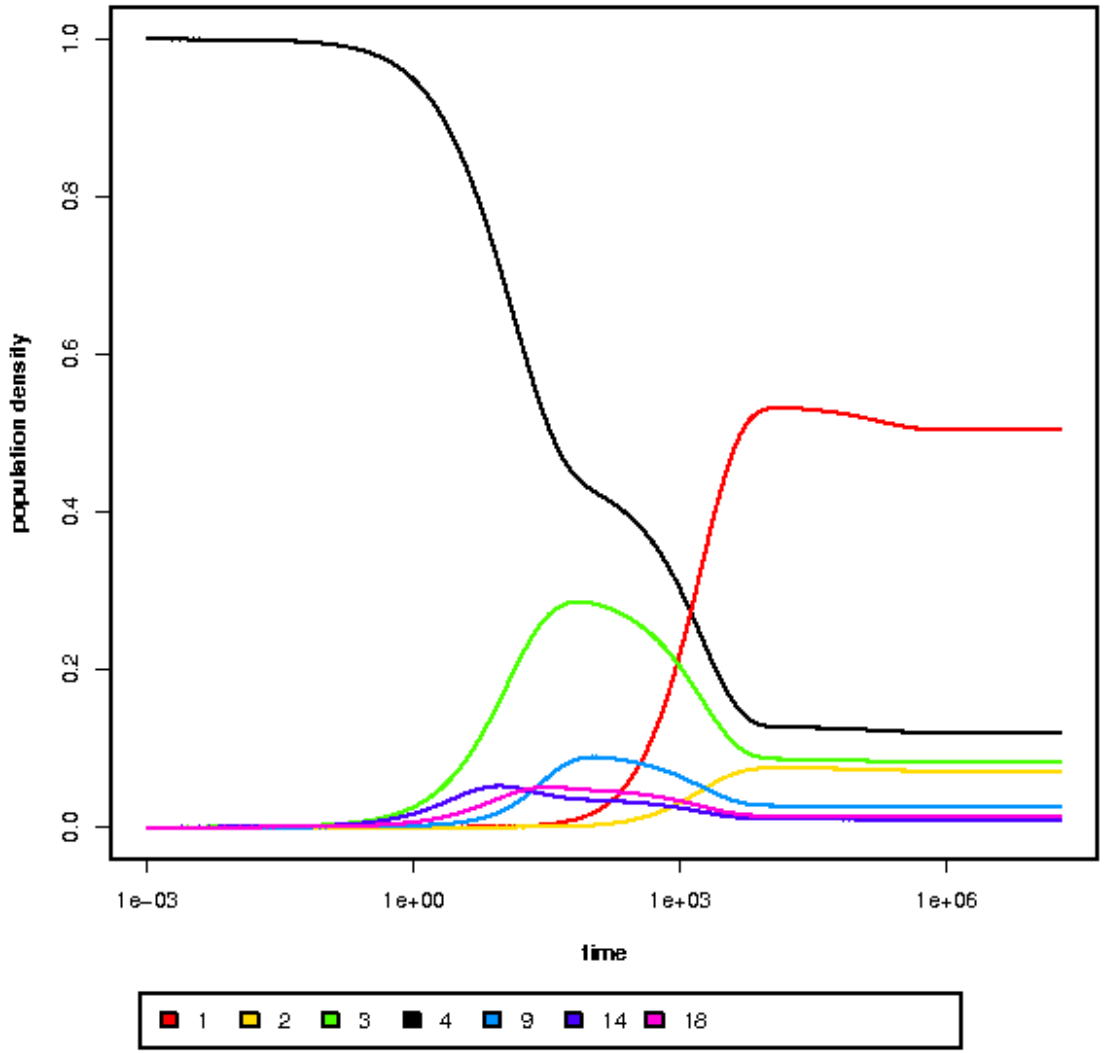


and investigate folding path to MFE structure 1



show visualization

Folding kinetics starting in macro state 4



More work to do:

More work to do:

- Find efficient folding paths between two given strucures without the need for the entire sec. structure space
- find more small circular RNAs to investigate
- apply circular extensions to the design tools for Multistable RNAs

Thank you for your attention