Coarse grained RNA folding kinetics

Ronny Lorenz
ronny@tbi.univie.ac.at

Institute for Theoretical Chemistry
University of Vienna

Bled, Slovenia, February 19, 2010

Why are we interested in this?

- RNAs with (long term stable) metastable structure states
- different functions coupled by change in conformation
- examples: RNA switches (thermometers, riboswitches, ...)

Questions arise:

- How does the structure (state) population density looks like in equilibrium?
- Starting from an initial population density, does the system reach its equilibrium directly (traps)?
- ...

RNA folding process in terms of a Markov process

- state space of allowed conformations (secondary structures)
- move-set defining elementary transitions between states (insert/deletion of base pairs)
- transition rates of allowed transitions (Metropolis/Kawasaki rule)

The master equation

$$\frac{d}{dt}\vec{p}(t) = \mathbf{R}\vec{p}(t) \quad \text{with formal solution} \quad \vec{p}(t) = e^{t\cdot\mathbf{R}} \cdot \vec{p}(0).$$

What we need is

- the initial population density $\vec{p}(0)$
- the transition rate matrix $\mathbf{R} = (r_{xy})$
- somebody who actually computes $\vec{p}(t)$

RNA folding process in terms of a Markov process

- state space of allowed conformations (secondary structures)
- move-set defining elementary transitions between states (insert/deletion of base pairs)
- transition rates of allowed transitions (Metropolis/Kawasaki rule)

The master equation

$$\frac{d}{dt}\vec{p}(t) = \mathbf{R}\vec{p}(t) \quad \text{with formal solution} \quad \vec{p}(t) = e^{t\cdot\mathbf{R}} \cdot \vec{p}(0).$$

What we need is

- the initial population density $\vec{p}(0)$
- the transition rate matrix $\mathbf{R} = (r_{xy})$
- something that actually computes $\vec{p}(t)$ (`treekin`)

But nature spoils things for us:

- number of states grows exponentially with sequence length
- direct computation of master equation becomes infeasible

But nature spoils things for us:

- number of states grows exponentially with sequence length
- direct computation of master equation becomes infeasible

Solution: Coarse graining of the state space!

But nature spoils things for us:

- number of states grows exponentially with sequence length
- direct computation of master equation becomes infeasible

Solution: Coarse graining of the state space!

- Partition the state space into macrostates
- compute effective transition rates between the partitions
- diagonalize the rate matrix
- compute eigenvalues and eigenvectors
- solve master equation

But nature spoils things for us:

- number of states grows exponentially with sequence length
- direct computation of master equation becomes infeasible

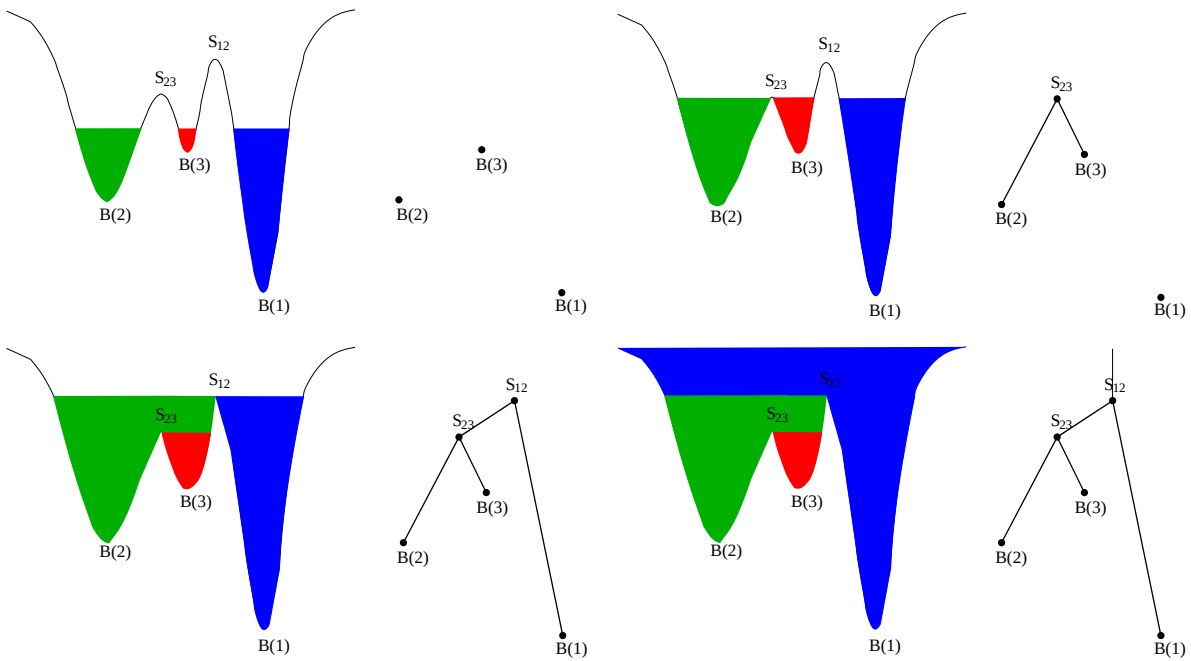
Solution: Coarse graining of the state space!

- Partition the state space into macrostates
- compute effective transition rates between the partitions
- diagonalize the rate matrix
- compute eigenvalues and eigenvectors
- solve master equation

How to partition the state space anyway?

The flooding algorithm (Flamm et al. 2002)

- energy sorted list of structure states
- identification of all local minima
- identification of minimal saddle points connecting them
- assigning each structure to its respective gradient basin



How to obtain the rate matrix $\mathbf{R} = (r_{xy})$?

Estimation from rates between micro states k_{yx}:

$$\begin{aligned} r_{\beta\alpha} &= \sum_{x\in\alpha}\sum_{y\in\beta} k_{yx}\mathrm{Prob}[x|\alpha] \\ &\approx \sum_{x\in\alpha}\sum_{y\in\beta} k_{yx}\cdot\frac{e^{-\frac{E(x)}{kT}}}{Q_\alpha} \end{aligned}$$

with:

$$k_{yx} = \begin{cases} e^{-\frac{E(y)-E(x)}{kT}} & \text{if } E(x) < E(y) \\ 1 & \text{otherwise.} \end{cases}$$

How to obtain the rate matrix $\mathbf{R} = (r_{xy})$?

Estimation from rates between micro states k_{yx}:

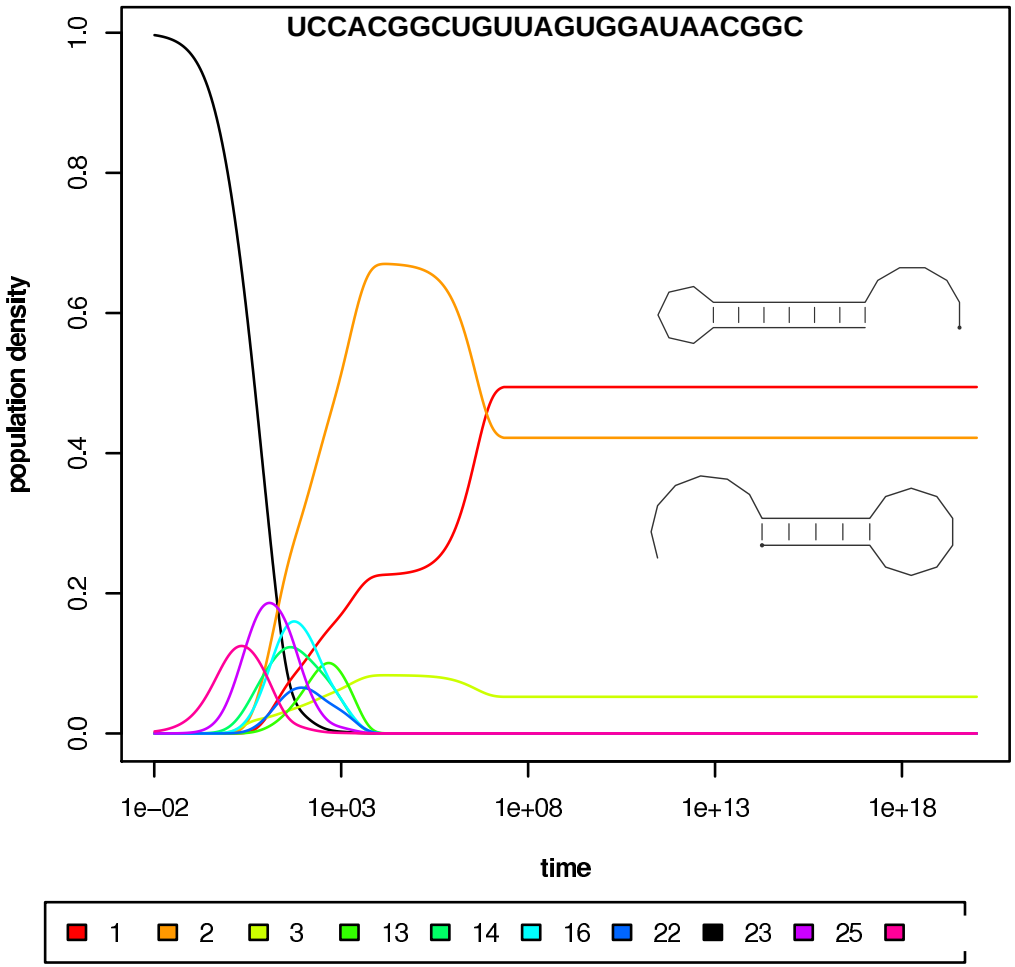
$$\begin{aligned} r_{\beta\alpha} &= \sum_{x\in\alpha}\sum_{y\in\beta} k_{yx}\mathrm{Prob}[x|\alpha] \\ &\approx \sum_{x\in\alpha}\sum_{y\in\beta} k_{yx}\cdot\frac{e^{-\frac{E(x)}{kT}}}{Q_\alpha} \end{aligned}$$

with:

$$k_{yx} = \begin{cases} e^{-\frac{E(y)-E(x)}{kT}} & \text{if } E(x) < E(y) \\ 1 & \text{otherwise.} \end{cases}$$

Limited to RNA molecules no longer than some 100 nt

small example with length = 25 nt, initial state = unfolded chain



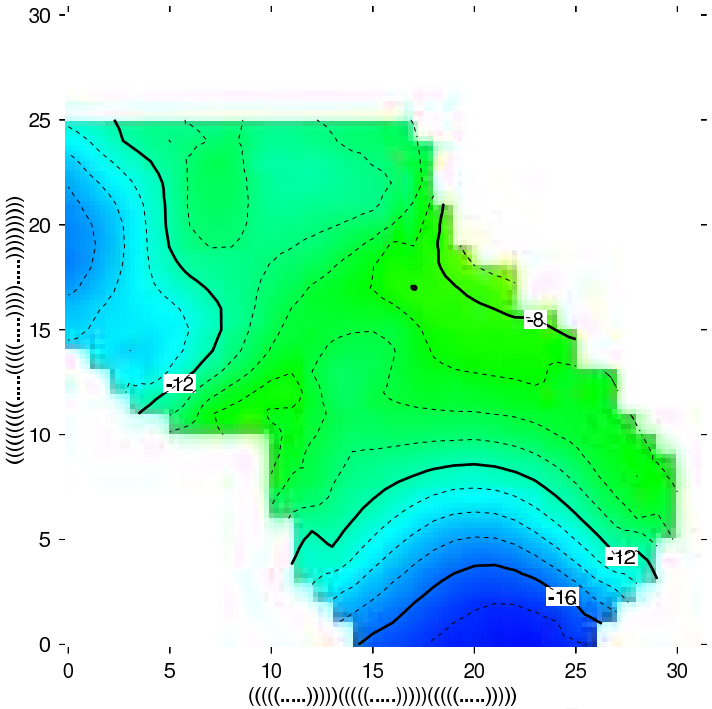
What about sampling secondary structures from the state space according their Boltzmann probability to estimate partition functions and transition rates between macrostates?

What about sampling secondary structures from the state space according their Boltzmann probability to estimate partition functions and transition rates between macrostates?

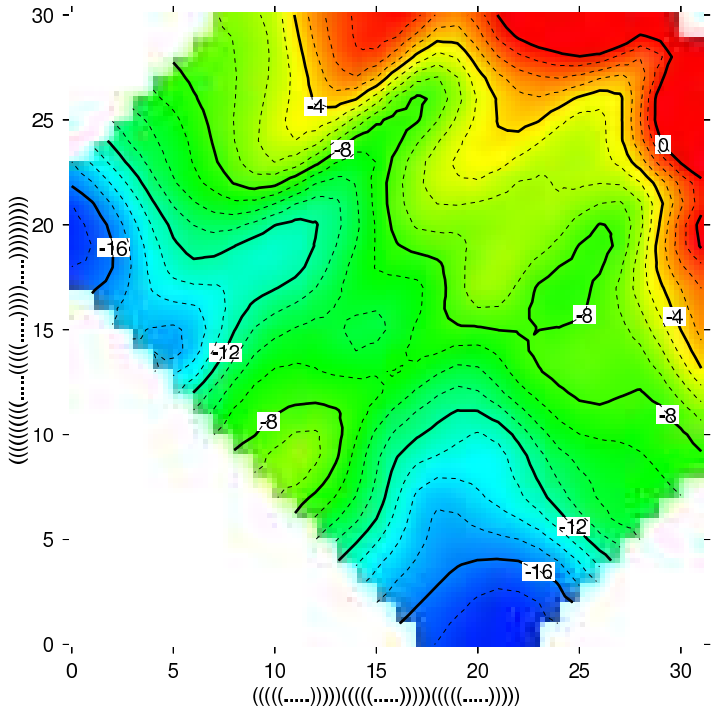
Sampling may not explore the landscape sufficiently!

MFE representatives with respect to two reference structures

```
GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAUUGUGUCU
(((((.....)))))(((((.....)))))(((((.....)))))
((((((((((.....(((((.....)))))).....)))))))))
```



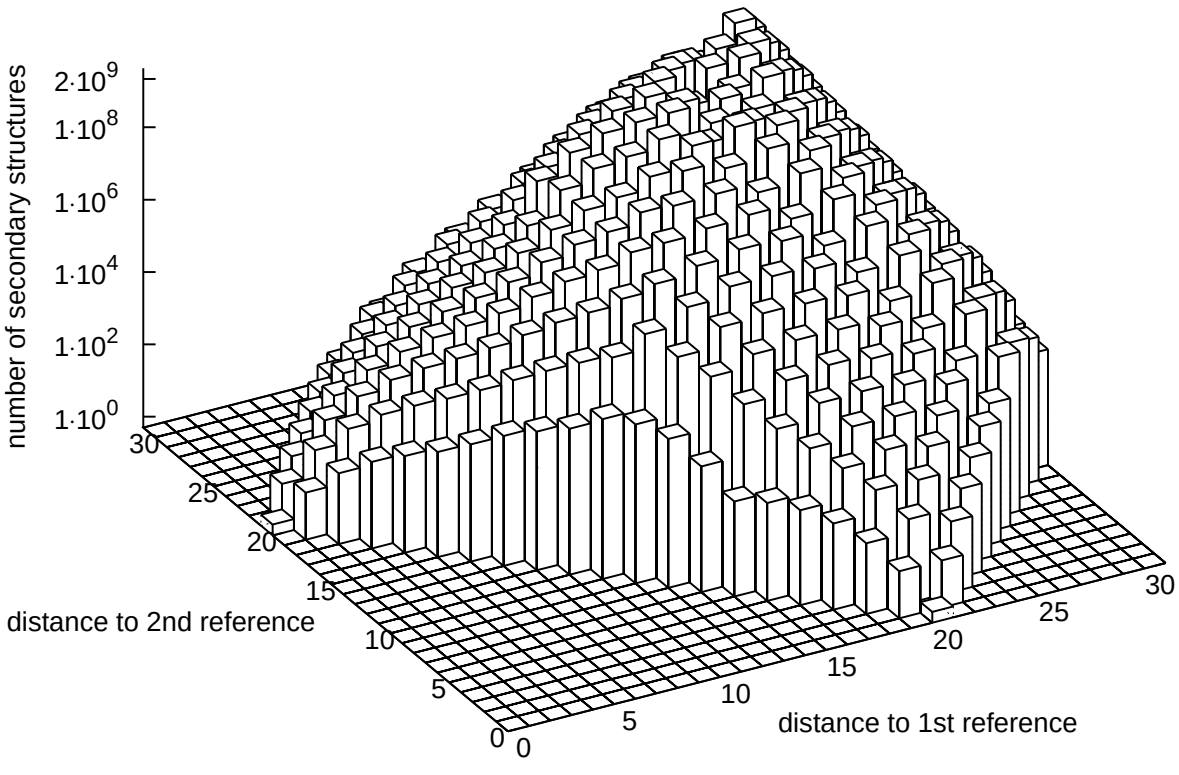
Landscape projection obtained by sampling 10^7 structures from the ensemble



Landscape projection obtained by `RNA2Dfold`

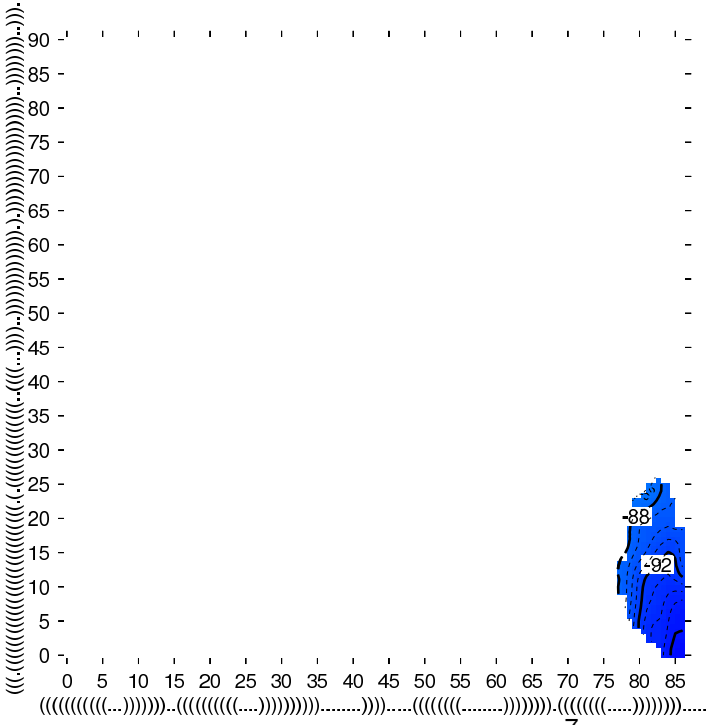
MFE representatives with respect to two reference structures

```
GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAUUGUGUCU
(((((.....)))))(((((.....)))))(((((.....)))))
((((((((((.....(((((.....)))))).....)))))))))
```

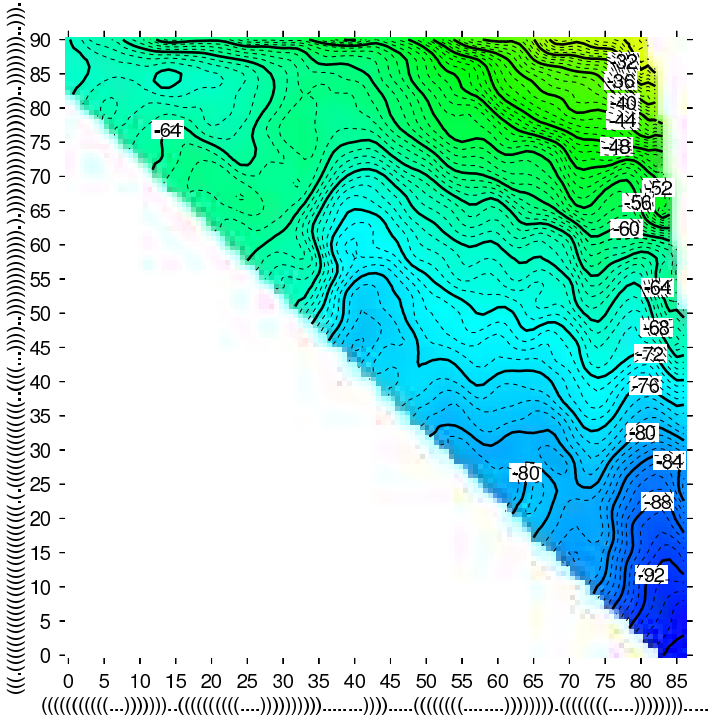


MFE representatives with respect to two reference structures

```
GGGCACCCCCCUUCGGGGGGUCACCUCGCGUAGCUAGCUACGCGAGGGUUAAAGCGCCUUUCUCCCUCGCGUAGCUAACCACGCGAGGUGACCCCCCGAAAAGGGGGGUUUCCCA
(((((((((((...)))))))..(((((((((((....)))))))))).......))))).....(((((((.......)))))))..(((((((.....)))))))).....
(((.(((((((((((((((((((((((((.(.(((((((((((((..((((...)))).))))))))))))))).).)))))))))))))))..)))))))..))).
```



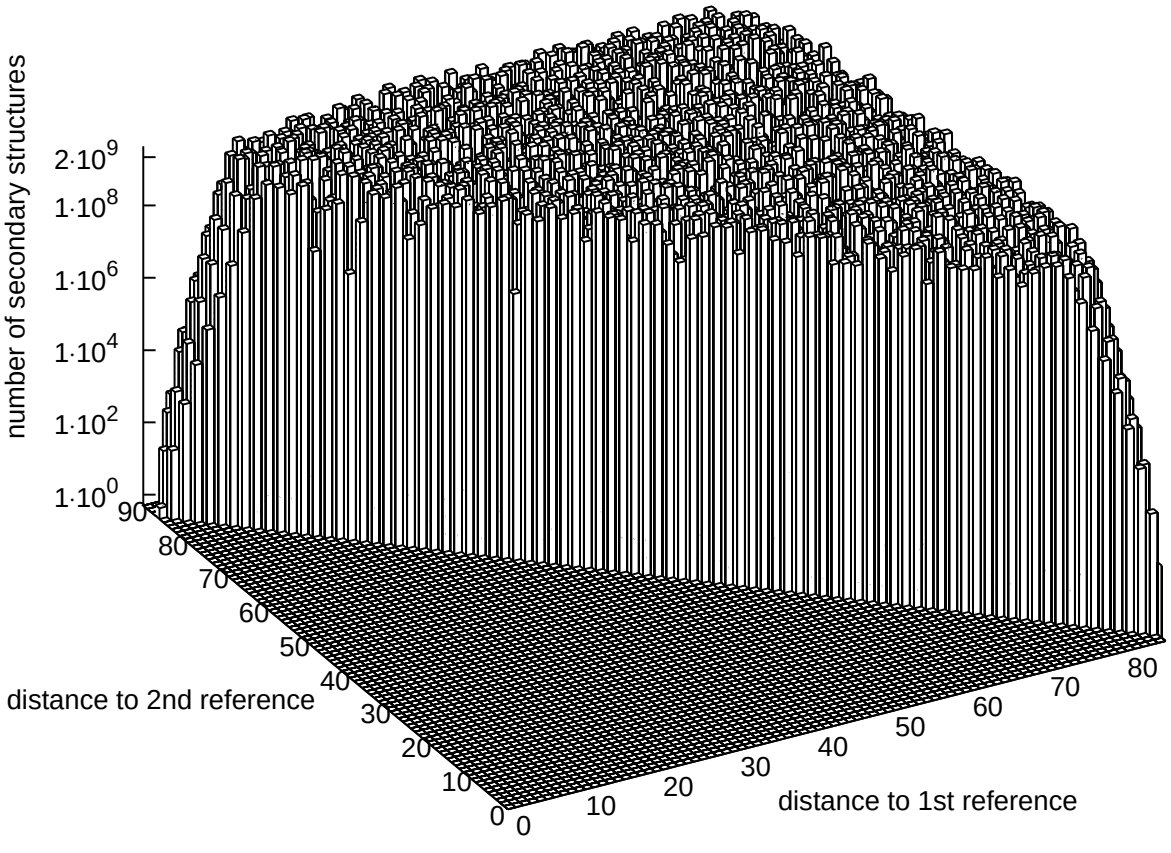
Landscape projection obtained by sampling 10^7 structures from the ensemble



Landscape projection obtained by `RNA2Dfold`

MFE representatives with respect to two reference structures

```
GGGCACCCCCCUUCGGGGGGUCACCUCGCGUAGCUAGCUACGCGAGGGUUAAAGCGCCUUUCUCCCUCGCGUAGCUAACCACGCGAGGUGACCCCCCGAAAAGGGGGGUUUCCCA
((((((((((...))))))).((((((((((....))))))))))........)))).....(((((((........)))))))..(((((((.....)))))))).....
(((.(((((((((((((((((((((((.(.(((((((((((((..((((...))))..))))))))))))).).)))))))))))))))))..)))))))..))).
```



Simulating folding dynamics becomes easier with prior knowledge

- MFE structure is most probable in equilibrium (1^{st} reference)
- sometimes a metastable state is known (2^{nd} reference)
- partitioning into distance classes (κ, λ-neighborhoods) wrt. two reference structures
- MFEs and partition functions can be computed in $\mathcal{O}(n^7)$
- computable for sequence up to 500 nt on modern machines
- Boltzmann sampling from each κ, λ-neighborhood

How to obtain the rate matrix $\mathbf{R} = (r_{xy})$?

Approximation of the macro rates by Boltzmann sampling from each distance class S_α:

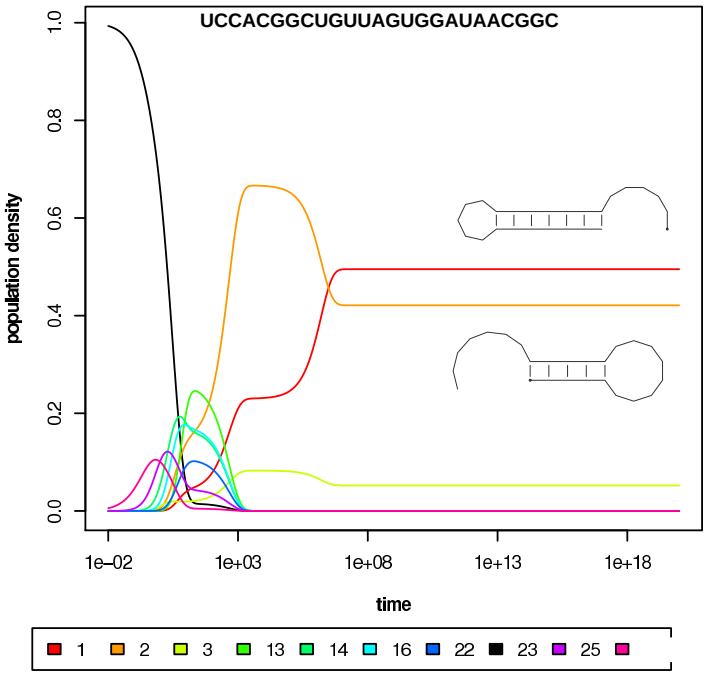
$$r_{\beta\alpha} \approx \frac{1}{|S_\alpha|} \sum_{x \in S_\alpha} \sum_{y \in \beta \cap \mathcal{N}(x)} k_{yx}$$

with:

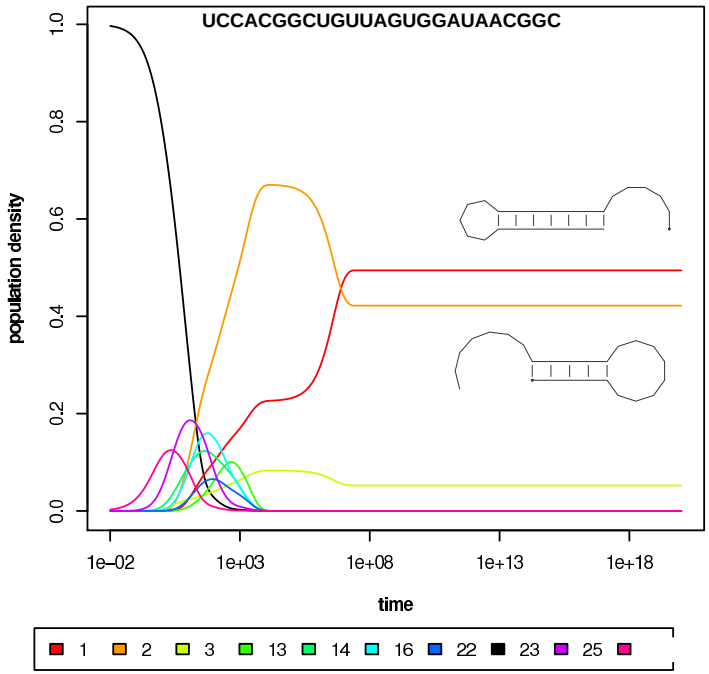
$$k_{yx} = \begin{cases} e^{-\frac{E(y)-E(x)}{kT}} & \text{if } E(x) < E(y) \\ 1 & \text{otherwise.} \end{cases}$$

- detailed balance must not be effected by sampling errors
- sample size of 1000 per macro state proved sufficient for the examples tested

small example with length = 25 nt, initial state = unfolded chain



1000 samples per macrostate



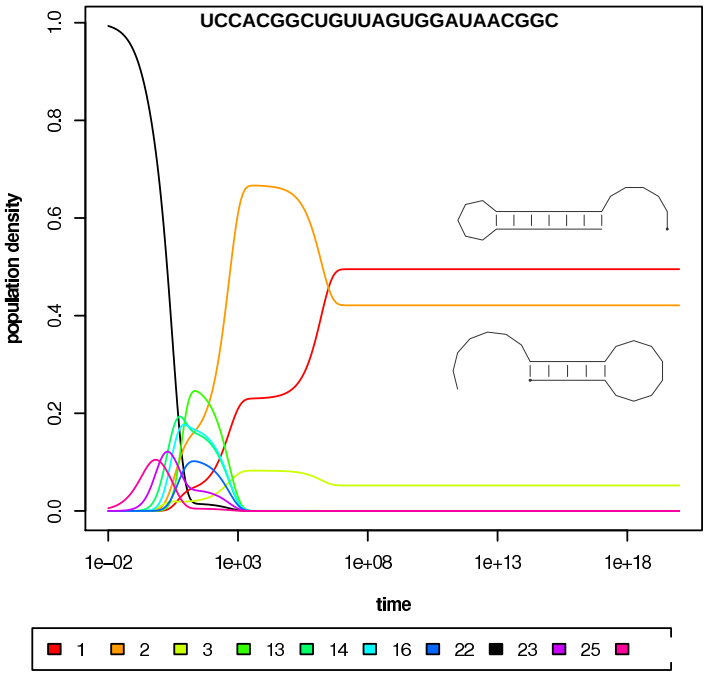
Full kinetic with 10000 microstates sorted into same distance classes

This method may also work for other partitionings of the state space

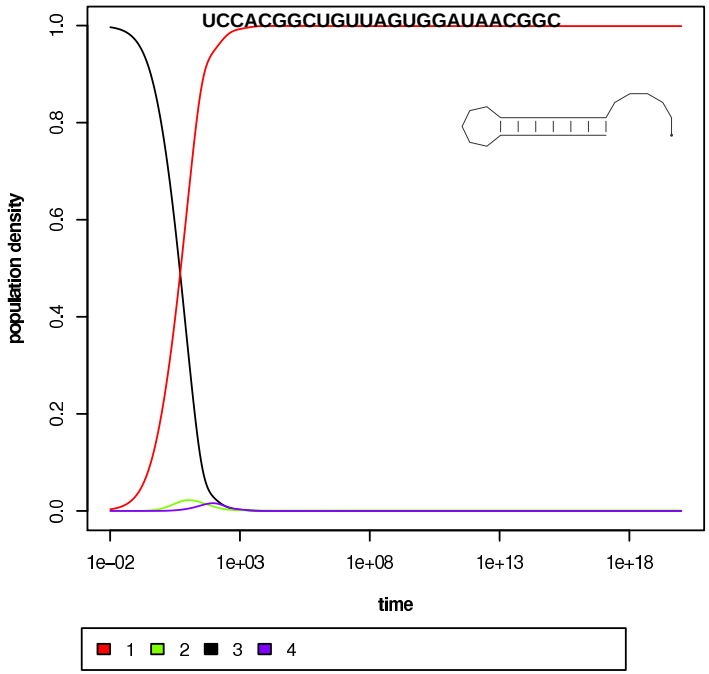
RNA shapes

- no RNAshapes stochastic backtracking available
- expected behavior can be computed for previous example

small example with length = 25 nt, initial state = unfolded chain

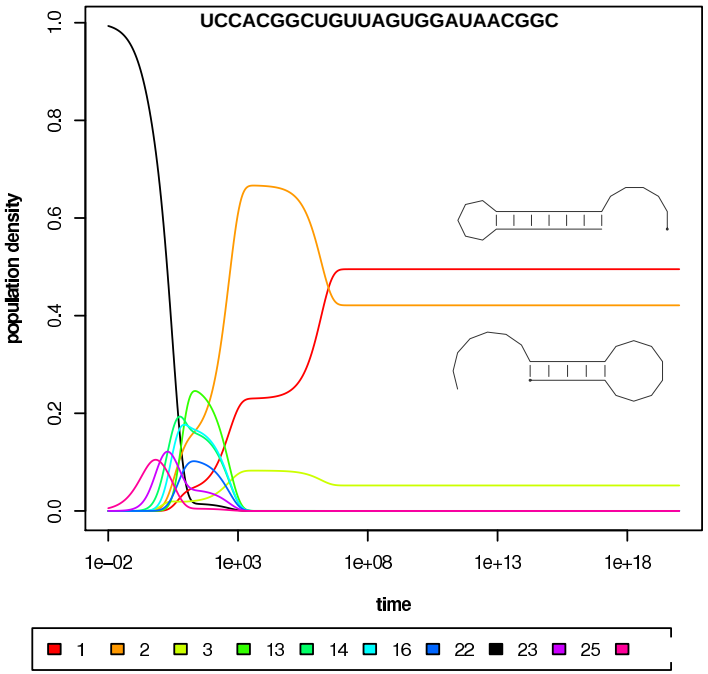


1000 samples per macrostate (κ, λ-approach)

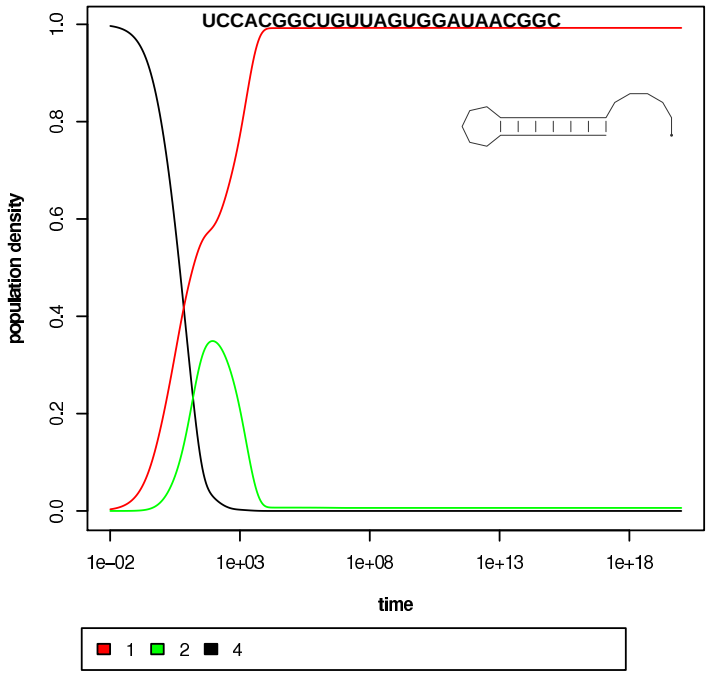


Full kinetic sorted into shapes (abstraction level 5)

small example with length = 25 nt, initial state = unfolded chain

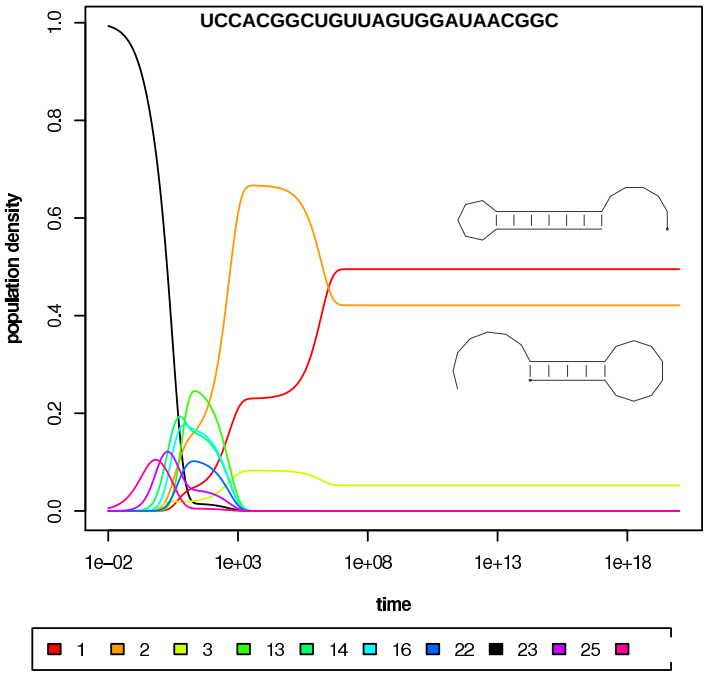


1000 samples per macrostate (κ, λ-approach)

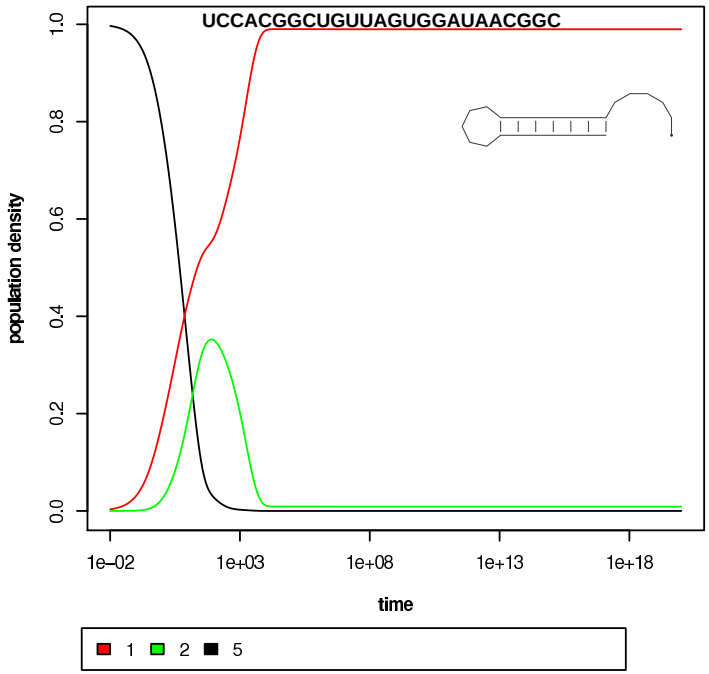


Full kinetic sorted into shapes (abstraction level 4)

small example with length = 25 nt, initial state = unfolded chain

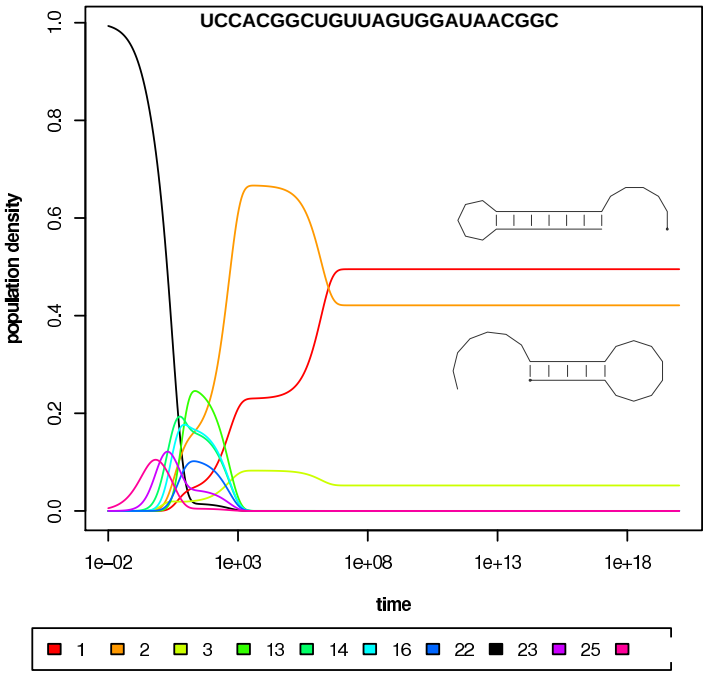


1000 samples per macrostate (κ, λ-approach)

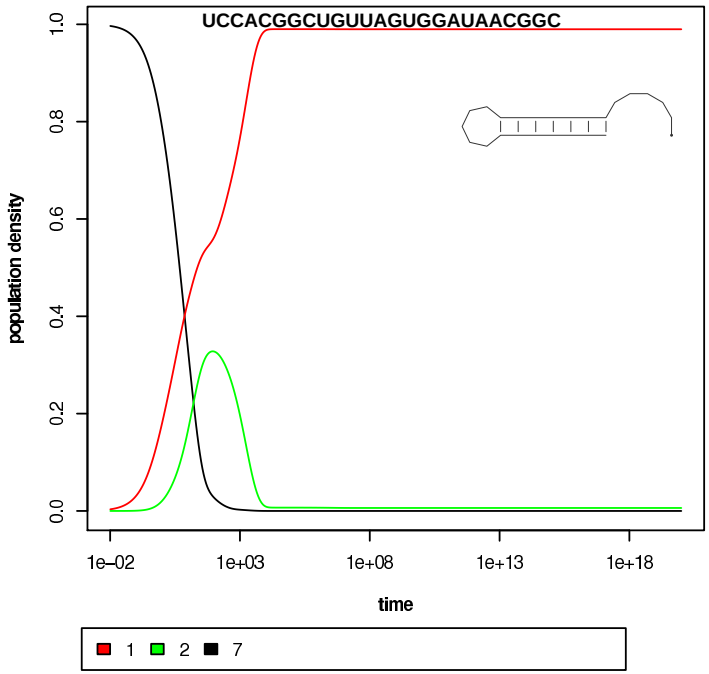


Full kinetic sorted into shapes (abstraction level 3)

small example with length = 25 nt, initial state = unfolded chain

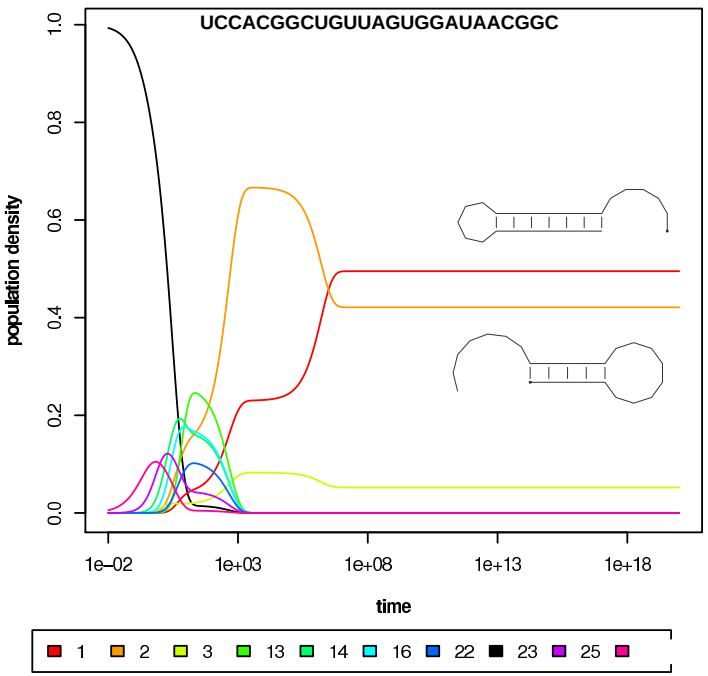


1000 samples per macrostate (κ, λ-approach)

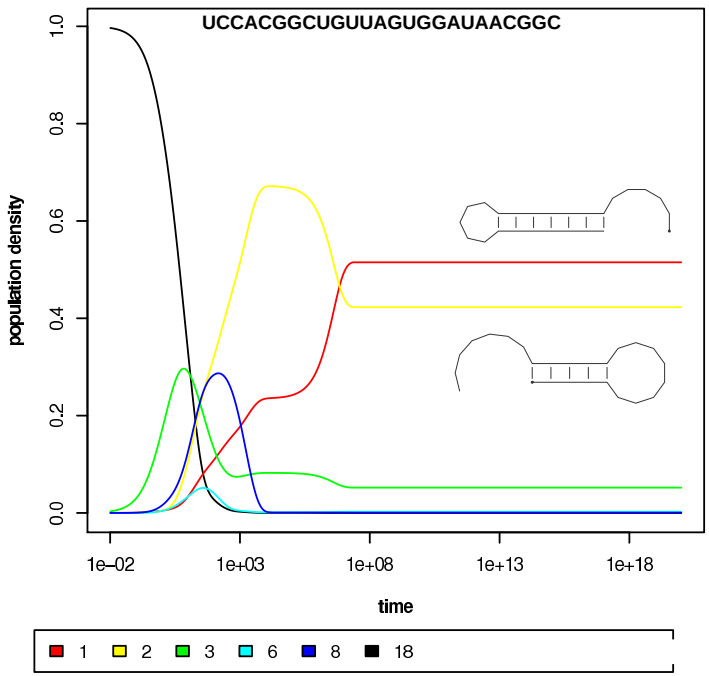


Full kinetic sorted into shapes (abstraction level 2)

small example with length = 25 nt, initial state = unfolded chain



1000 samples per macrostate (κ, λ-approach)



Full kinetic sorted into shapes (abstraction level 1)

To summarize

- prior knowledge can ease computational effort
- Boltzmann sampling may not explore important parts of the structure space
- sampling from distance classes implicitely explores more strucutral diversity
- significantly longer RNAs can be analyzed
- transition rate sampling may also work for RNAshapes partitioning

Thanks to:

Christoph Flamm
Christian Höner zu Siederdissen
Ivo Hofacker

...and You!

This work has been funded, in part, by the Austrian GEN-AU projects “bioinformatics integration network III” and “non coding RNA”.