Coarse grained RNA folding kinetics

Ronny Lorenz
ronny@tbi.univie.ac.at

Institute for Theoretical Chemistry
University of Vienna

Vienna, Austria, May 5, 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, ...)

Arising questions:

- Population of conformations towards equilibrium given initial population density
  (fast, slow, via longterm stable intermediates, ...)
- Influence of cotranscriptional folding
- Influence of temperature
- ...

RNA folding as a Markov process

- State space

$$\mathcal{S} = \{s | s \text{ is secondary structure for the sequence}\}$$

- Neighborhood relation

$$\mathcal{N}(s_i, s_j) = \begin{cases} \text{true} & \text{if } d_{BP}(s_i, s_j) == 1 \\ \text{false} & \text{otherwise.} \end{cases}$$

- Transition rates $\mathbf{R} = (r_{ij})$

$$r_{ij} = \begin{cases} f(s_i, s_j) & \text{if } \mathcal{N}(s_i, s_j) \\ 0 & \text{otherwise.} \end{cases}$$

- $\vec{p}(0) \ldots$ population density of all states at time 0

The master equation

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

But nature spoils things for us:

- number of states grows exponentially with sequence length
- matrix exponential exceeds computability
- direct computation of master equation becomes infeasible even for small RNA sequences

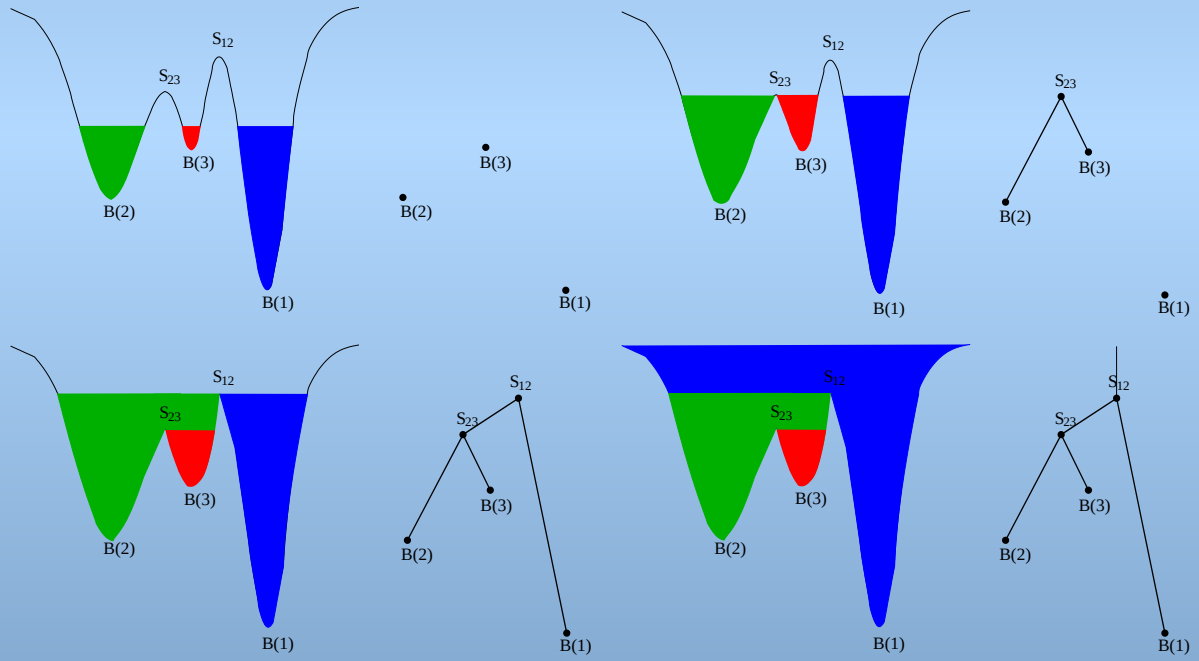
Solution: Coarse graining of the state space!

- Partition the state space into macrostates
- compute effective transition rates between the partitions
- solve master equation for the smaller problem

How to construct the macrostates
...and compute their transition rates?

The flooding algorithm, gradient basins and barrier trees[1]

- 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



Limited to RNA molecules no longer than some 100 nt

[1] Flamm et al. 2002

Gradient basin transition rates[2]

Estimation of gradient basin rates along the barrier tree:

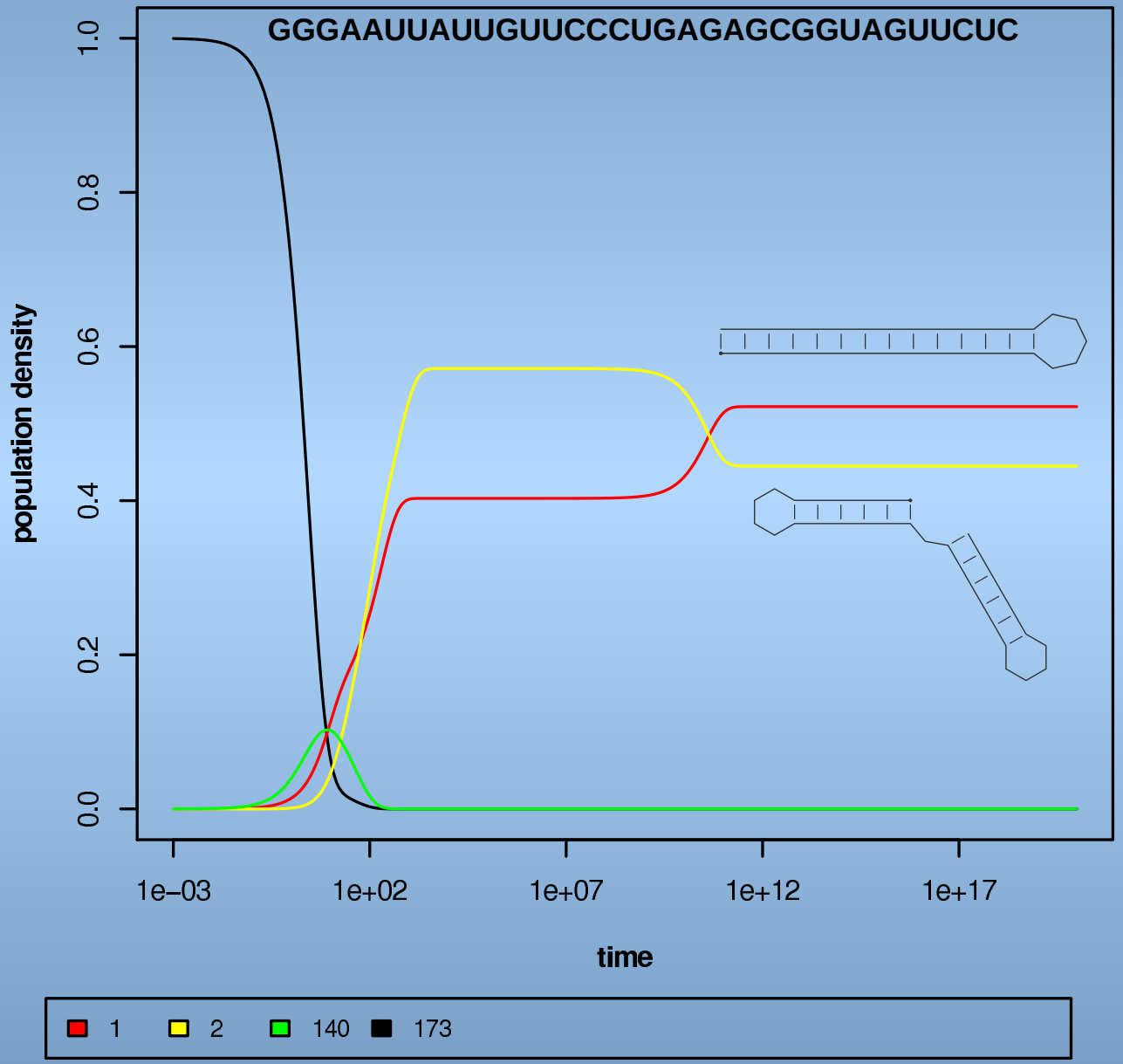
$$r_{\beta\alpha} = e^{-\frac{E_{\alpha\beta}-G_\alpha}{kT}}$$

with:

$$\begin{aligned} E_{\alpha\beta} &\ldots \text{ energy of saddle connecting state } \alpha \text{ and } \beta \\ G_\alpha &= -kT \cdot \ln Q_\alpha \\ Q_\alpha &= \sum_{i \in \alpha} e^{-\frac{E_i}{kT}} \end{aligned}$$

[2] Wolfinger et al. 2004

Small example with unfolded chain as initial state



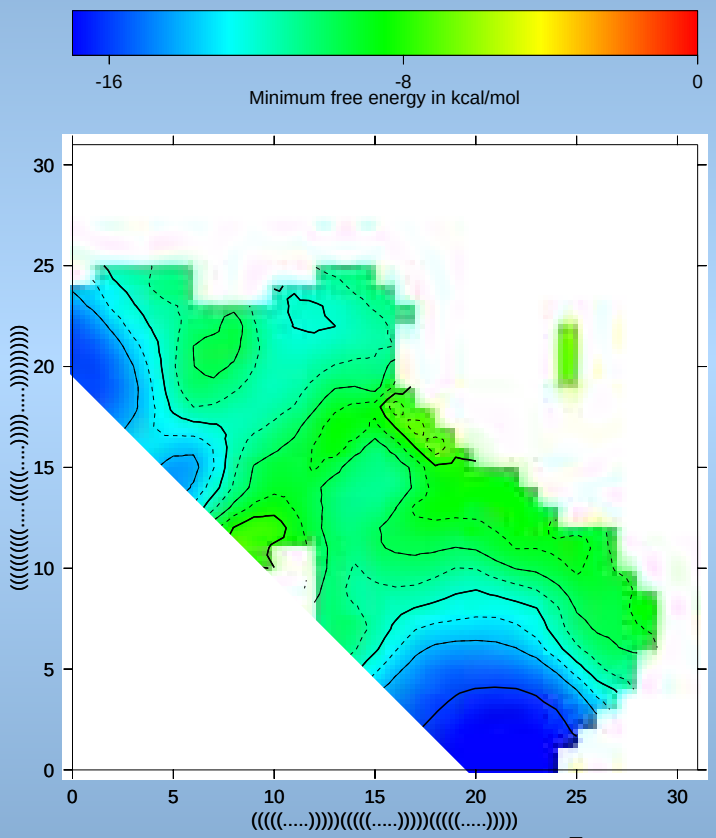
How to circumvent exhaustive enumerations?

- Sampling of secondary structures according their Boltzmann probability
- Sort samples into the macro states
- Estimate partition functions from samples
- Estimate transition rates

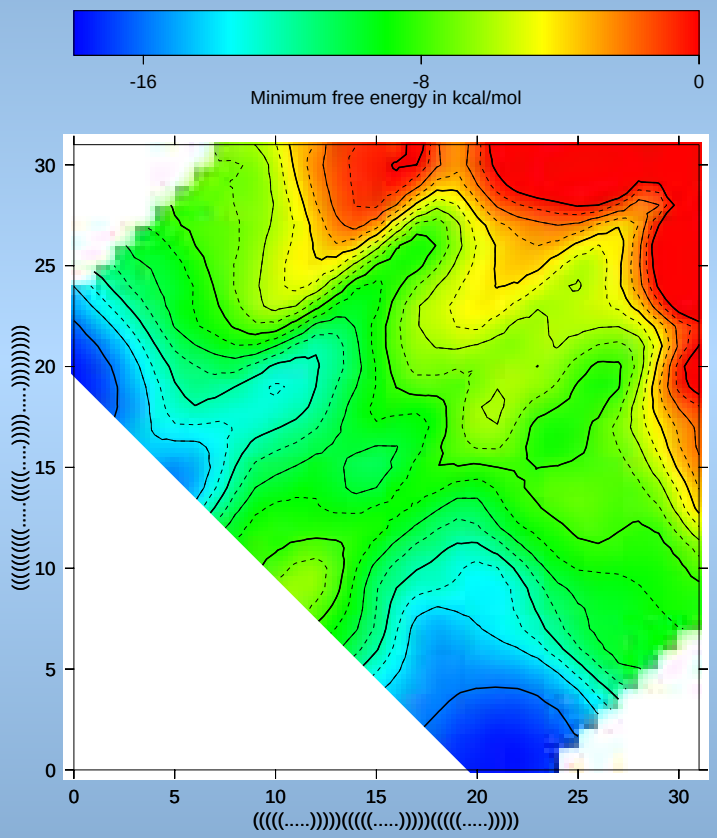
Sampling may not explore the state space sufficiently!

MFE representatives wrt. two reference structures[3]

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



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

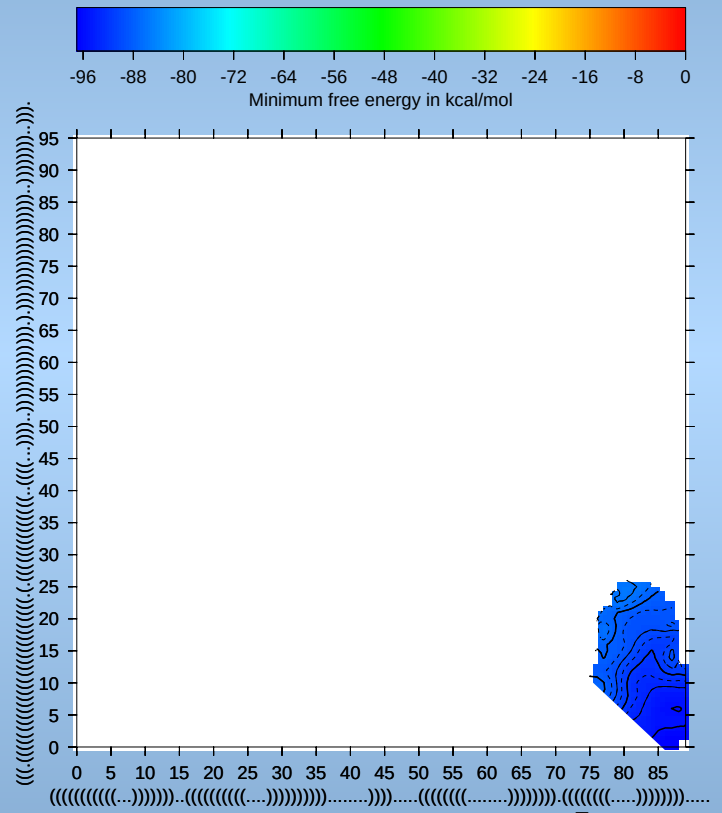


Landscape projection obtained by `RNA2Dfold`

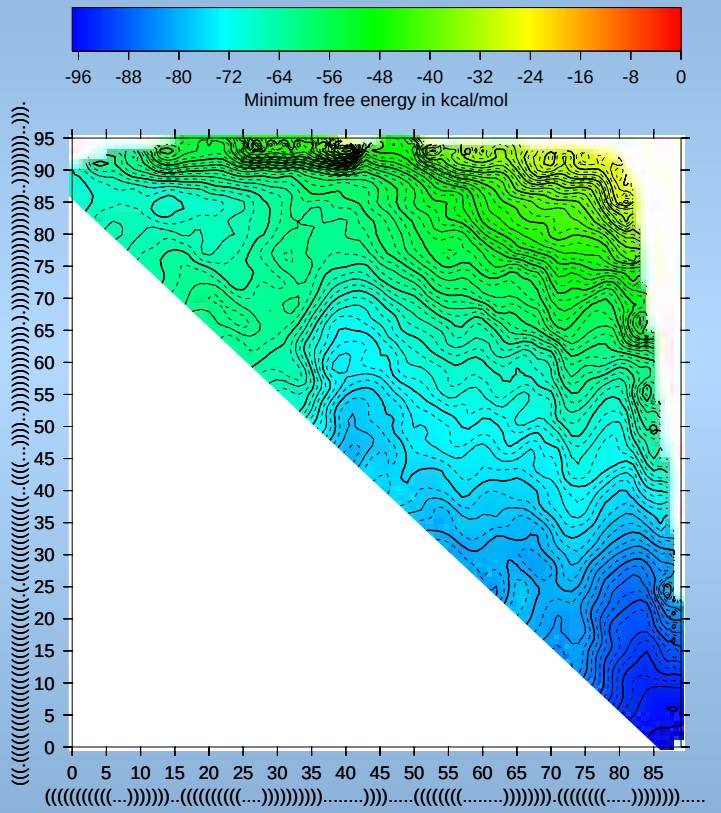
[3]Lorenz et al. 2009

MFE representatives wrt. two reference structures

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



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



Landscape projection obtained by `RNA2Dfold`

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:

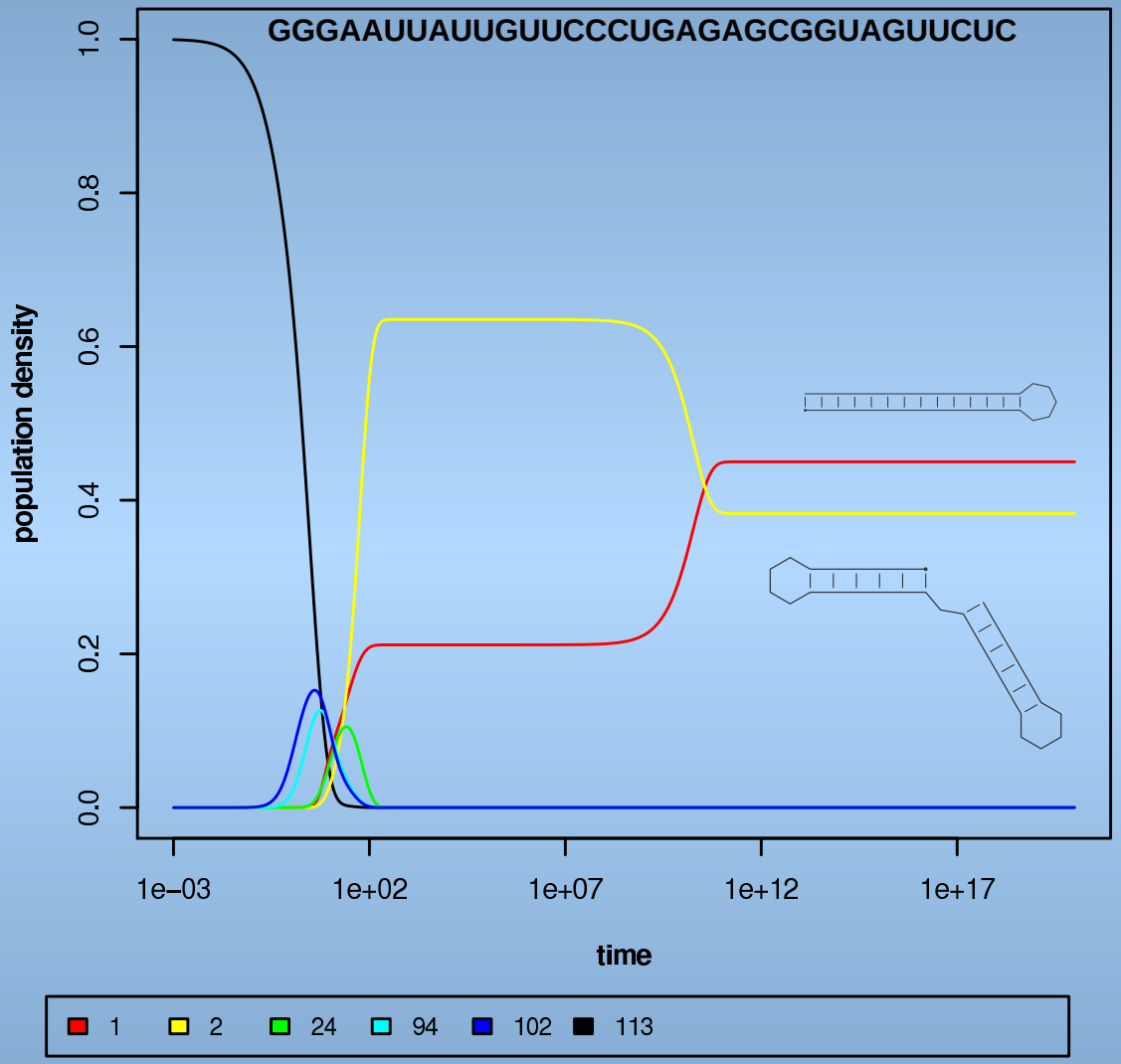
$$r_{\beta\alpha} \approx \frac{1}{|S_\alpha|} \sum_{x \in S_\alpha} \sum_{y \in \beta, \mathcal{N}(x,y)} k_{yx}$$

with:

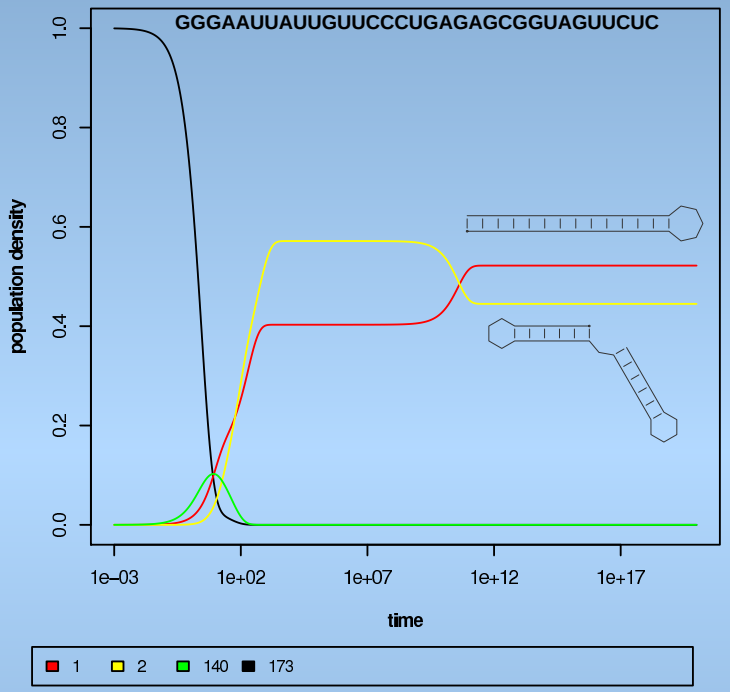
$$k_{yx} = \begin{cases} e^{-\frac{E(y)-E(x)}{kT}} & \text{if } E(x) < E(y) \\ 1 & \text{otherwise.} \end{cases}$$

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

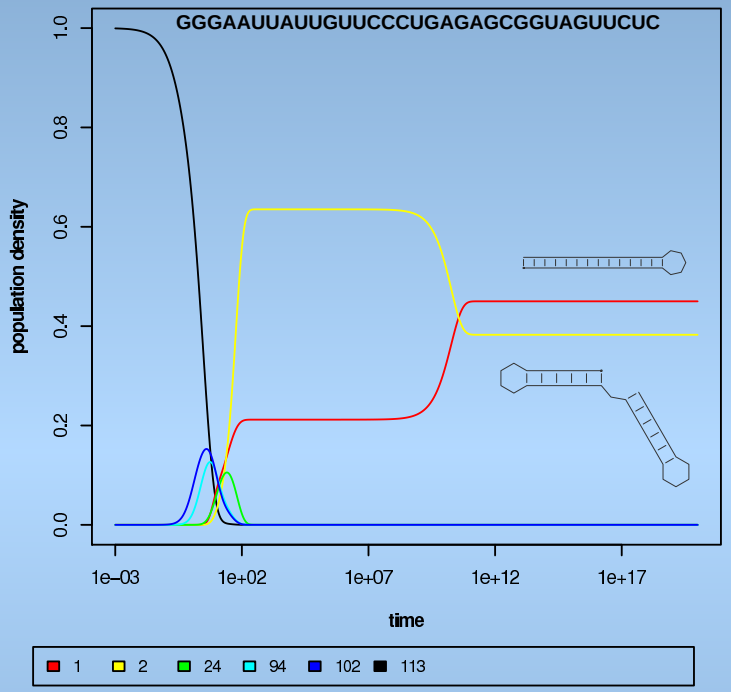
Small example with unfolded chain as initial state



Small example with unfolded chain as initial state



Kinetic with gradient basin macro states



Kinetic with distance class macro states

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 structural diversity
- significantly longer RNAs can be analyzed
- method used may also work for other partitionings (*RNAshapes*, etc.)

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”.