Sampling Energy landscapes and Modeling co-transcriptional RNA folding

Ivo L. Hofacker

Institute for Theoretical Chemistry
and
Research Group Bioinformatics and Computational Biology
University of Vienna

Benasque, August 2022

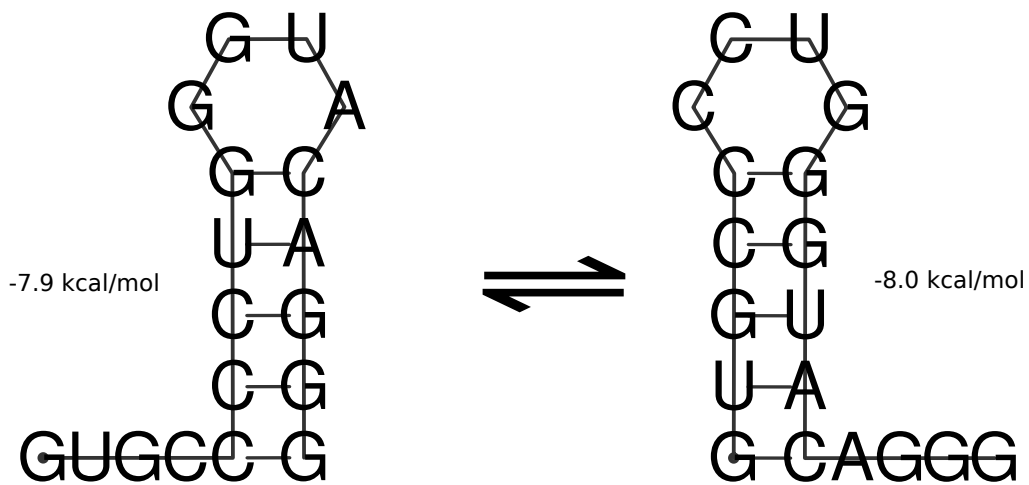
tbi

Thermodynamic vs. Kinetic Folding

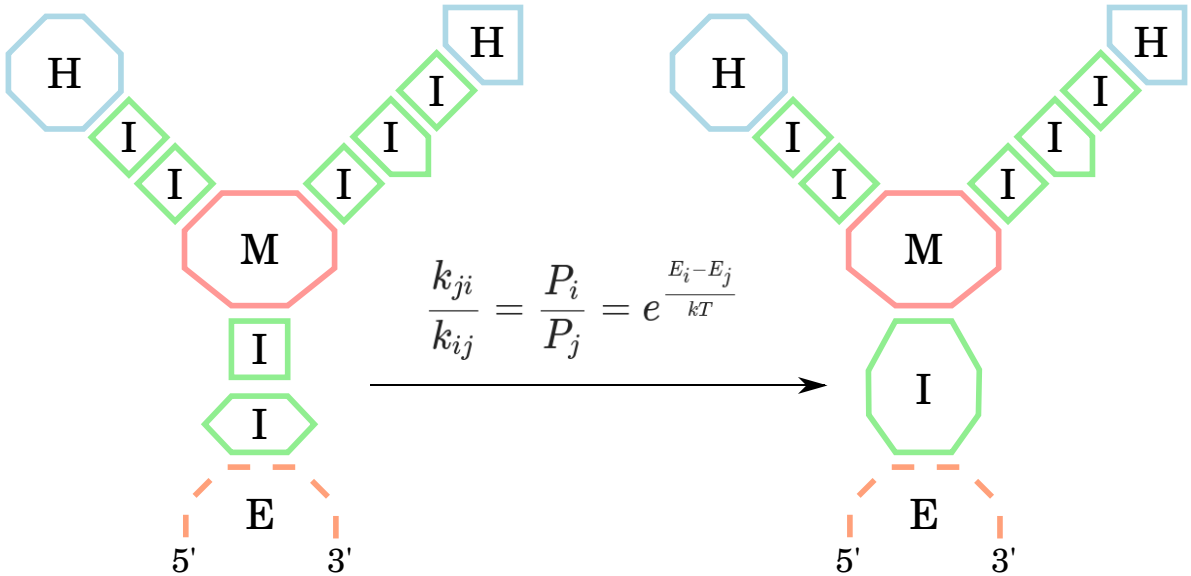
Equilibrium properties for RNA secondary strutcures can be calculated efficiently

But what about dynamics?

- On what time scale is equilibrium reached?
- How fast/slow is re-folding between dissimilar structures?
- What structures are populated initially?



Folding Dynamics as Markov Process



- Time evolution determined by master equation:

$$\frac{dP_i(t)}{dt} = \sum_{j \neq i} [P_j(t)k_{ji} - P_i(t)k_{ij}].$$

- Rate model must satisfy detailed balance
  energies from the Turner model
- guarantees correct convergence to equilibrium

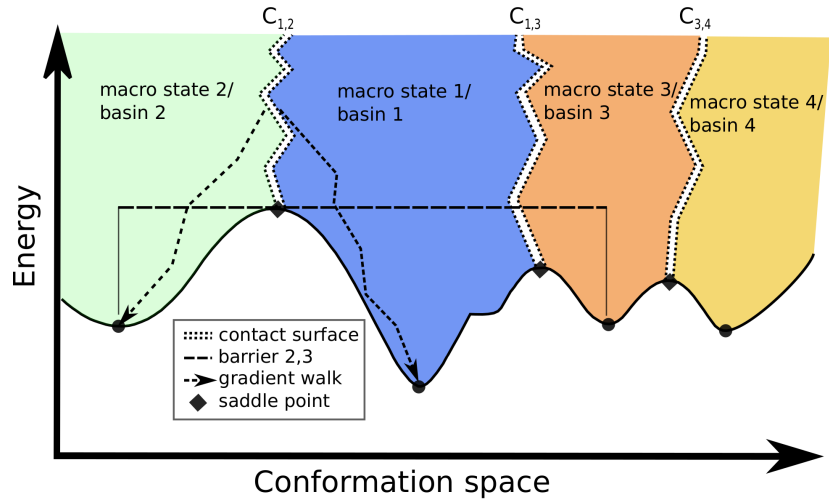
Strategies for Predicting Folding Kinetics

- Folding trajectories via Monte-Carlo simulation
  - Time-consuming
  - Need statistics over many trajectories
  - Non-trivial to analyze and interpret
  - `kinfold`, `KineFold`
- Coarse graining the energy landscape
  - Identify representative structures (local minima)
  - Assign transition rates
  - Either exact enumeration (`barriers`)
    using heuristics (`RNAlocmin`, `RNAxplorer`)
  - Solve $P_x(t)$ on coarse grained landscape (`treeekin`)

Coarse Graining the folding dynamics

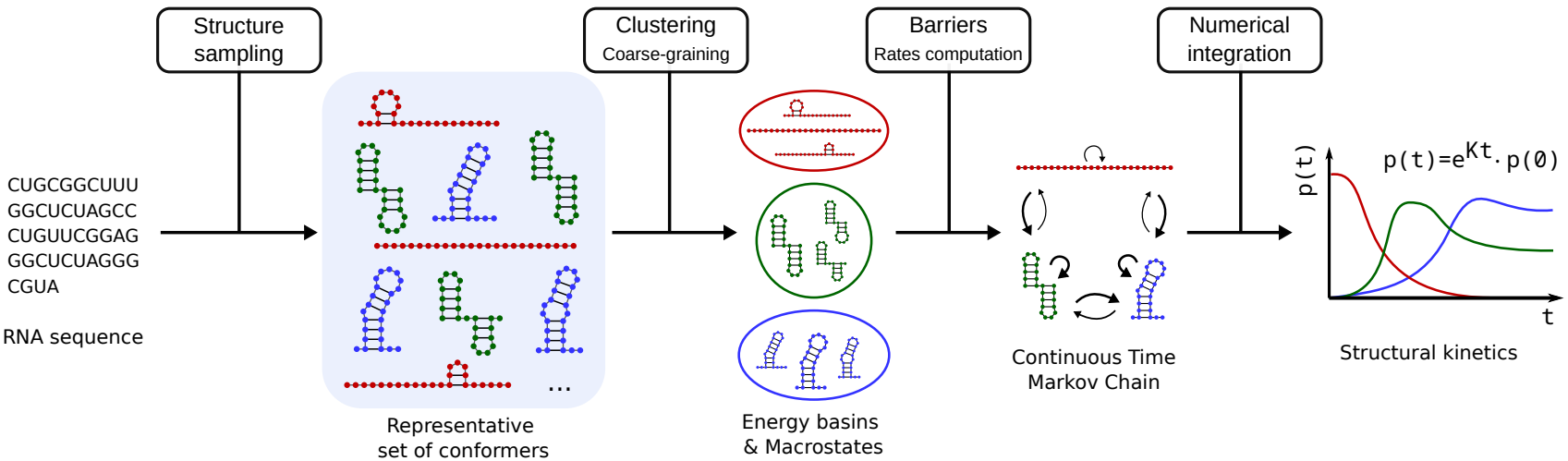
For a reduced description we need

- macro-states form a partition of full configuration space
- macro-states defined via gradient walks
- effective transition rates between macro states



Transition rates e.g. from Arrhenius rule $r_{\beta\alpha} = \exp\left(-(E^*_{\beta\alpha} - G_\alpha)/RT\right)$.

Workflow for Folding Kinetics



Sampling Representative Structures

- Exhaustive enumeration
  short sequences only
- Boltzmann sampling
  lot's of samples close to MFE but poor diversity
  - nonredundant sampling (`RNAsubopt`)
  - sampling only local minima (`RNAlocopt`)
  - sampling with temperature control (`RNAlocmin`)
  - sampling with guiding potentials (`RNAxplorer`)

RNAxplorer Sampling

Boltzmann sampling produces structure s with $p(s) = \frac{1}{Z} \exp\left(-E(s)/RT\right)$

1. Set energy function $E = E_{\text{Turner}}$
2. Boltzmann sampling from E
3. Identify most overrepresented structure $\hat{s}$
4. Set $E_{\text{new}} = E_{\text{old}} + E_{\text{guide}}(\hat{s})$
5. Goto 2

E_{guide} penalizes any structures similar to $\hat{s}$

RNAxplorer Guiding Potentials

Two choices of guiding potential:

- penalize all base pairs present in $\hat{s}$

$$E_{\text{guide}}(s, \hat{s}) = \alpha \cdot \frac{|s \cap \hat{s}|}{|\hat{s}|}$$

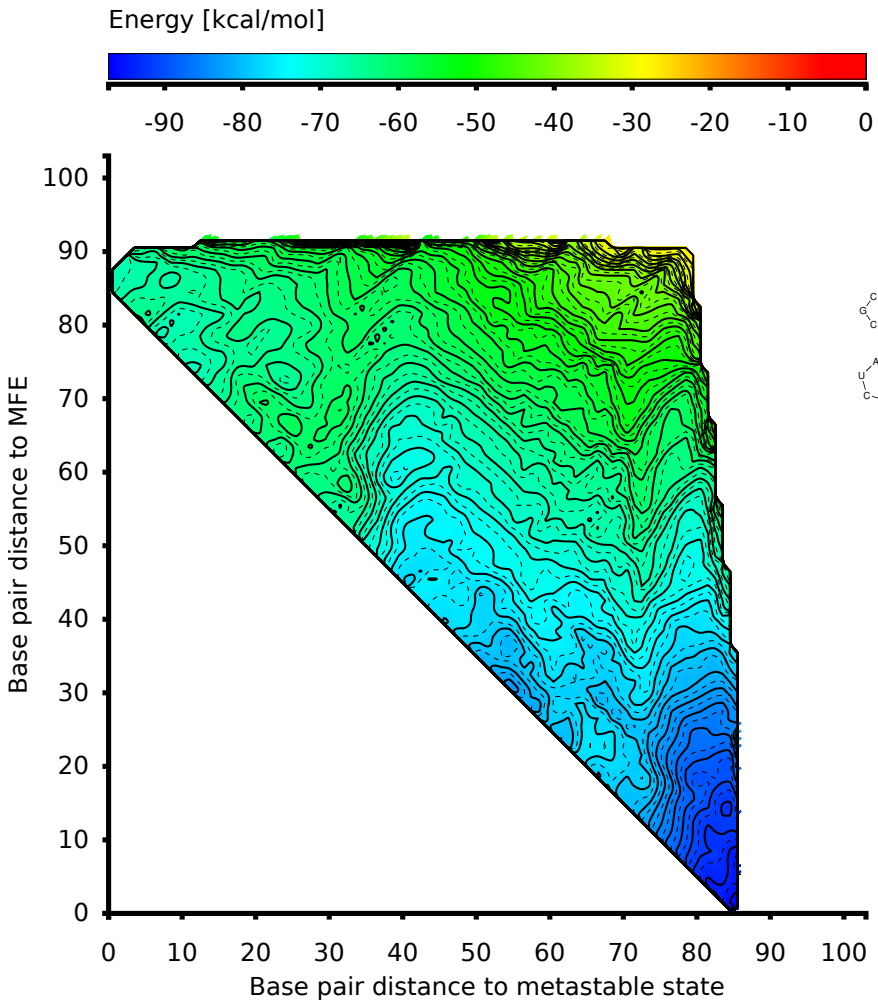
- penalize by distance from $\hat{s}$

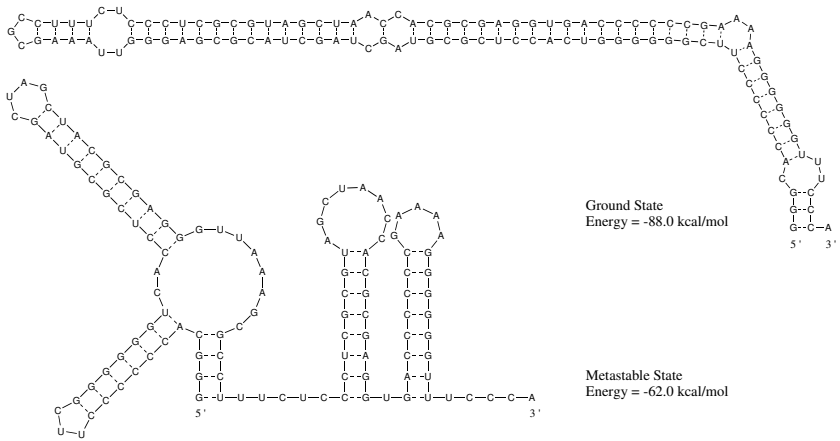
$$E_{\text{guide}}(s, \hat{s}) = \alpha \cdot d_{\text{BP}}(s, \hat{s})$$

Both guiding potentials implemented as soft constraints in ViennaRNA.

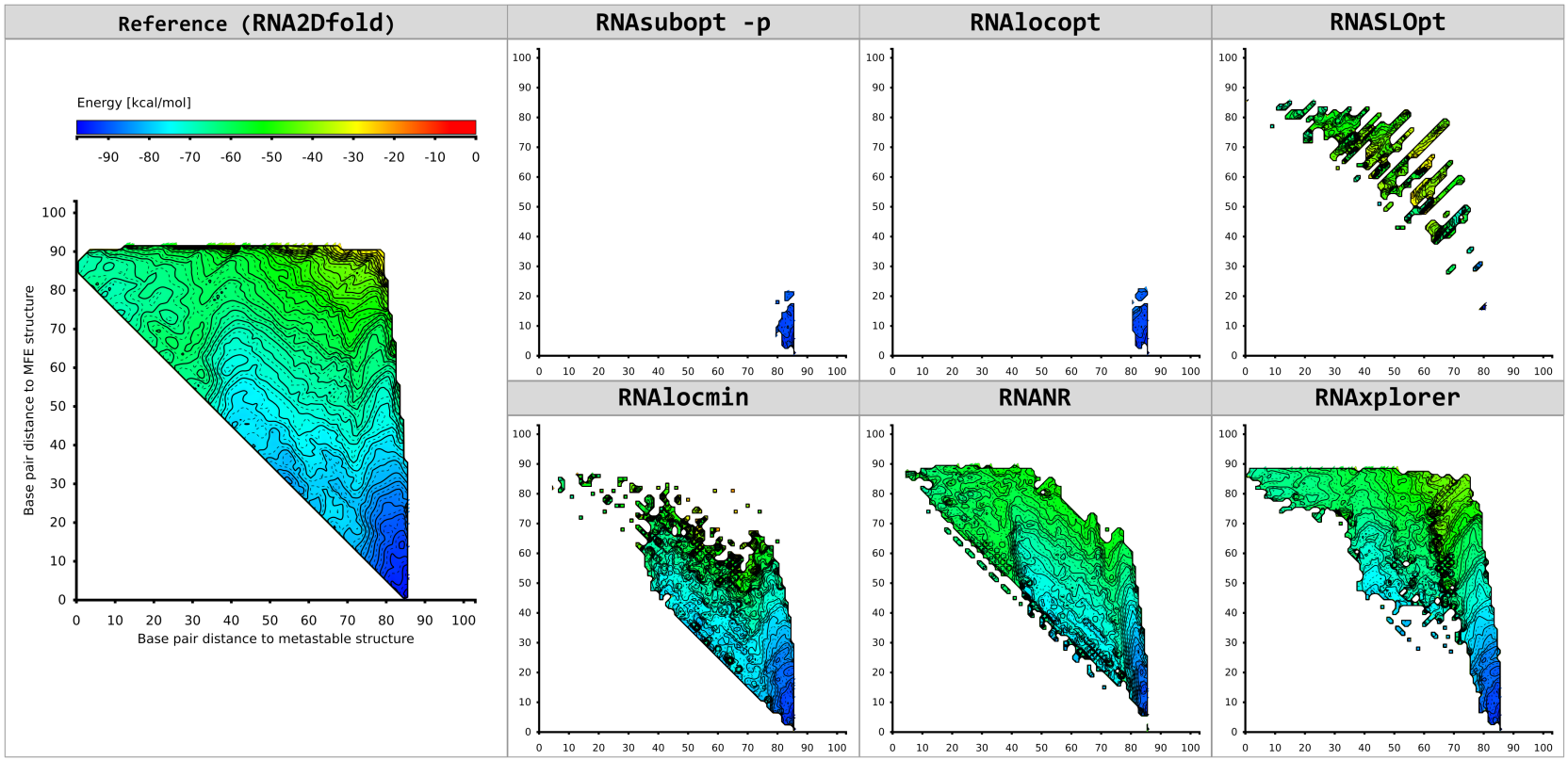
Comparison with RNA2DFold

2D projection of the SV11 energy landscape



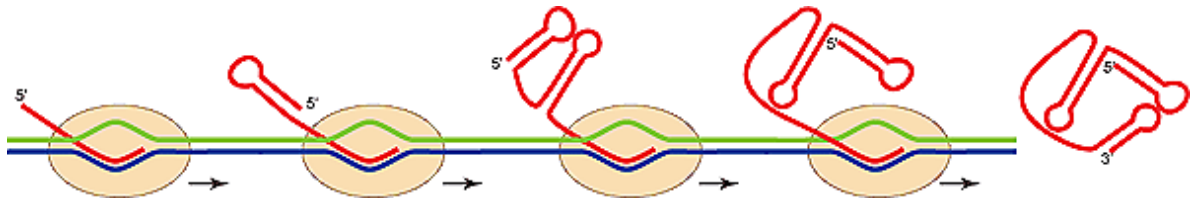


Comparison with RNA2DFold



Folding during Transcription

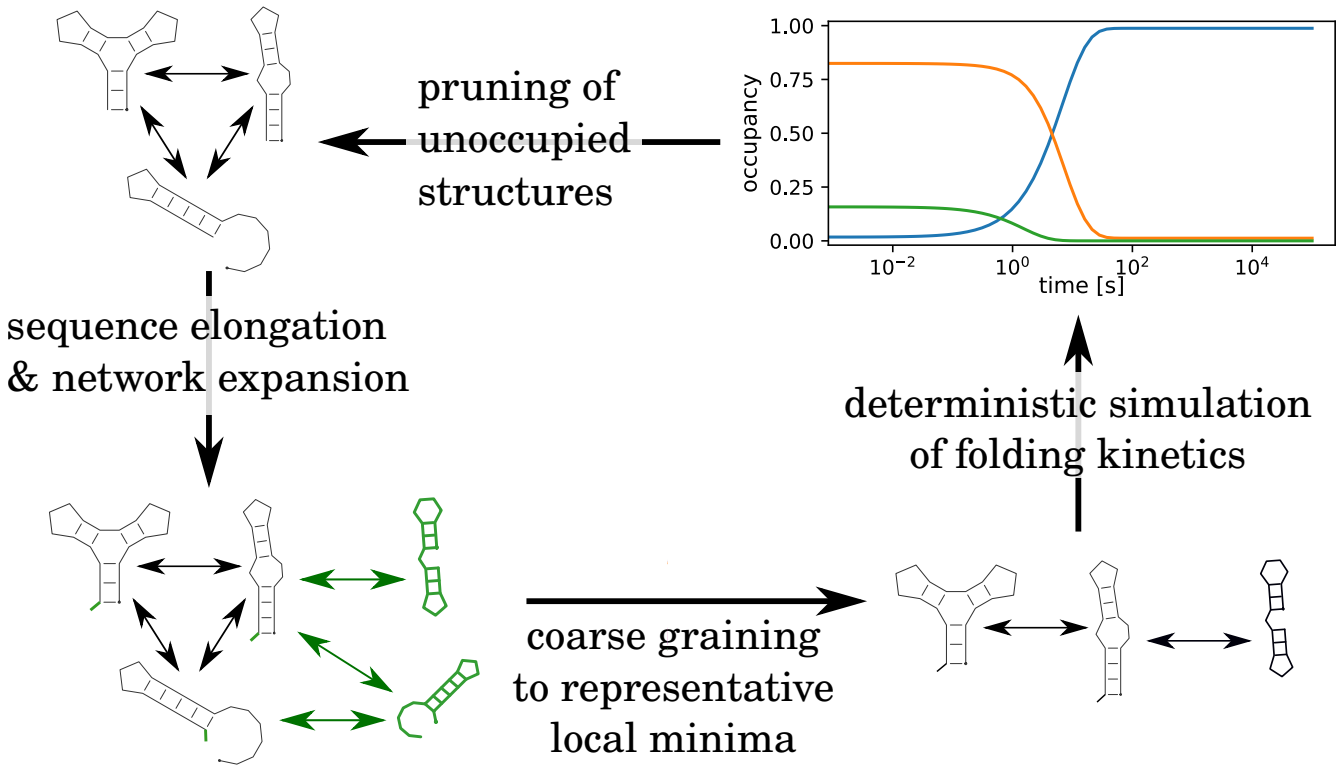
All RNA structures are affected by co-transcriptional folding:



- RNA is transcribed at a rate of only 25–50 nucleotides per second
- The nascent chain starts folding as soon as its leaves the polymerase
- Stems formed early on may be too stable to refold later
- Co-transcriptional folding may drive the folding process to a well-defined folded state (possibly different from the MFE)
- An energy barrier of 5kcal/mol is sufficient to prevent refolding during extension

DrTransformer: Heuristic Co-transcriptional Folding

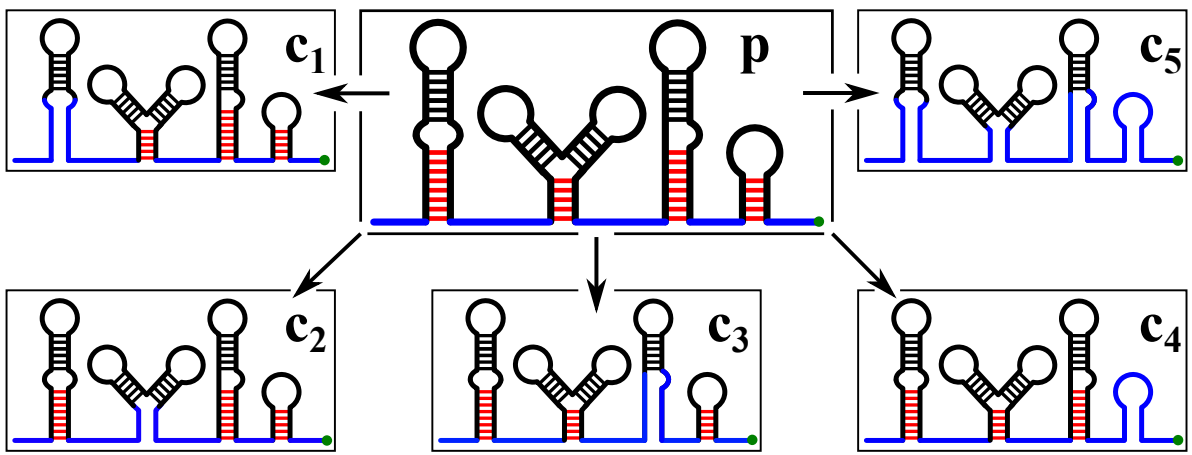
- Simulate a **small** network consisting only of the most relevant structural states
- Evolve network as RNA grows



DrTransformer: "Breathing" neighbors

Which new structures should be added after an elongation step?

- Elongation can only effect the surroundings of the exterior loop
- Partially unfold all helices that protrude from exterior loop
- Use constrained folding to re-fold exterior loop surroundings



DrTransformer: Connecting States

- Connect structures based on distance.
  Add an edge between x and y, if there is no i such that

  $$\max\{d(x,i), d(i,y)\} < d(x,y)$$

  If edges exist between x and i, and i and y, add a shortcut, if

  $$d(x,i) + d(i,y) > d(x,y)\,.$$

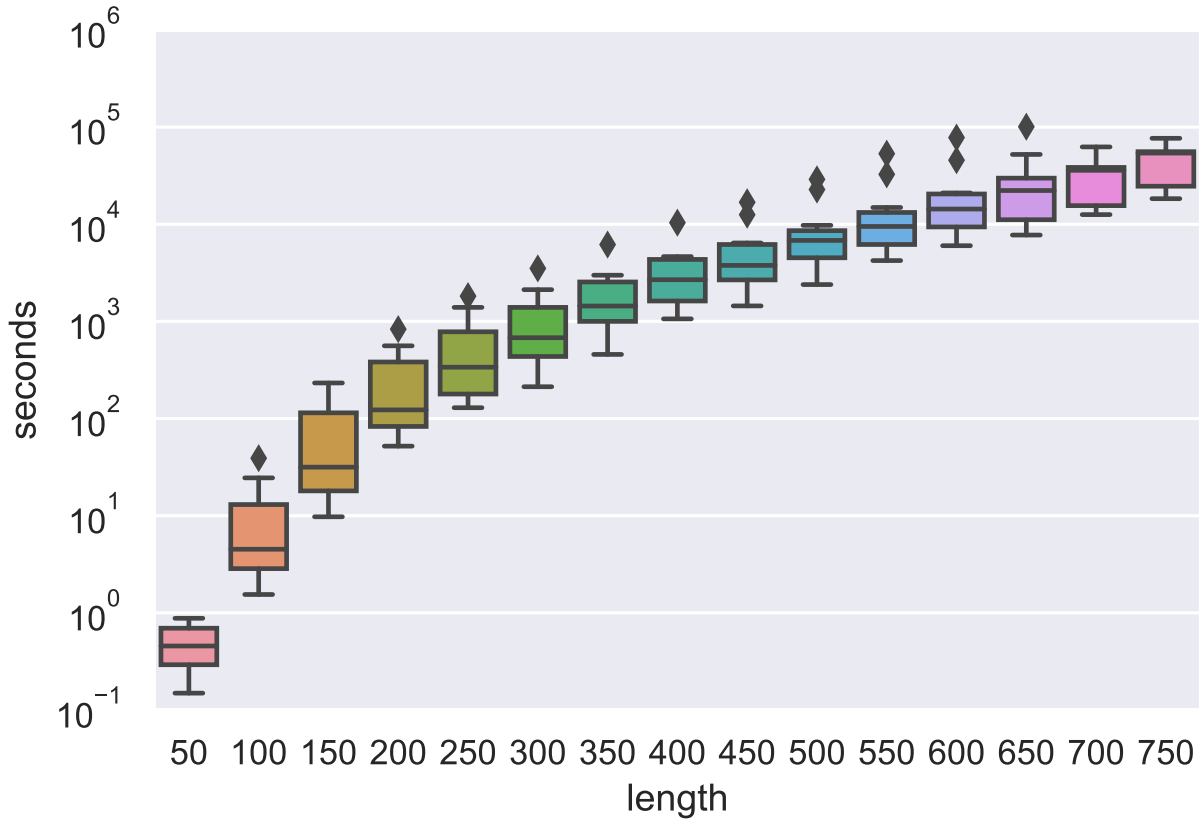
- Estimate saddle points using `findpath`
- Add addditional minima encountered along paths

DrTransformer: Coarse Graining and Pruning

Runtime depends critically on the number of structure states
Keep the number of active structure states small by

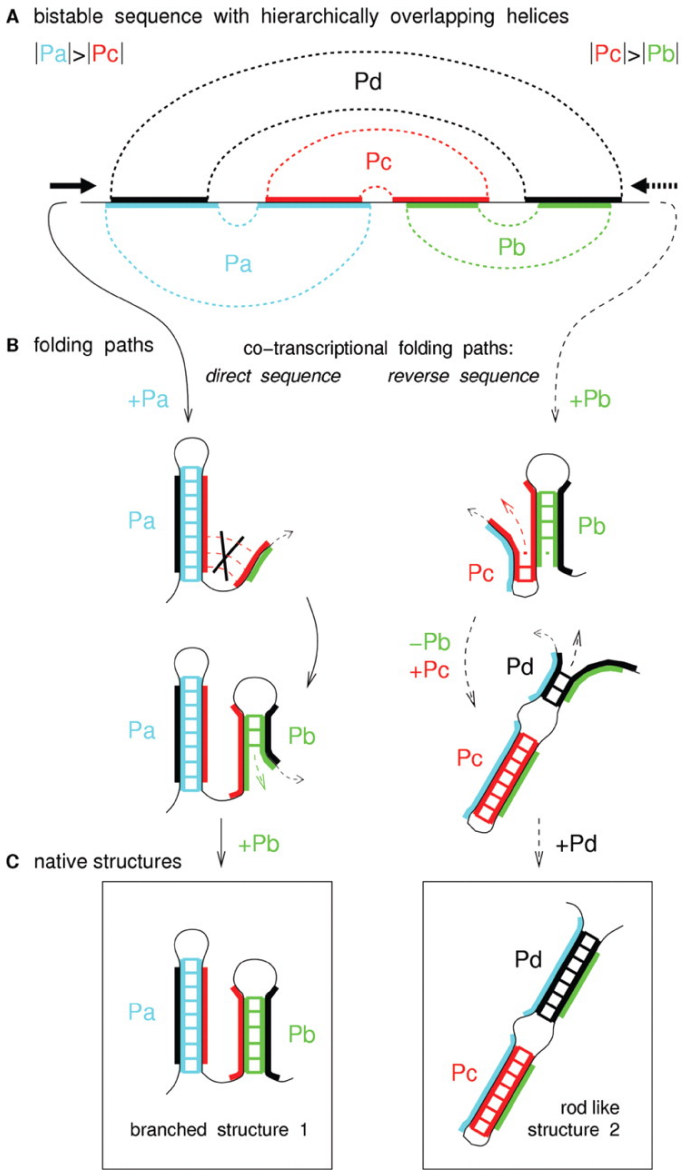
- Coarse graining:
  remove all shallow local minima with, i.e. that can reach other minima via a barrier $< \delta \approx 1\,\text{kcal/mol}$
- Pruning:
  After each simulation step, remove unoccupied states

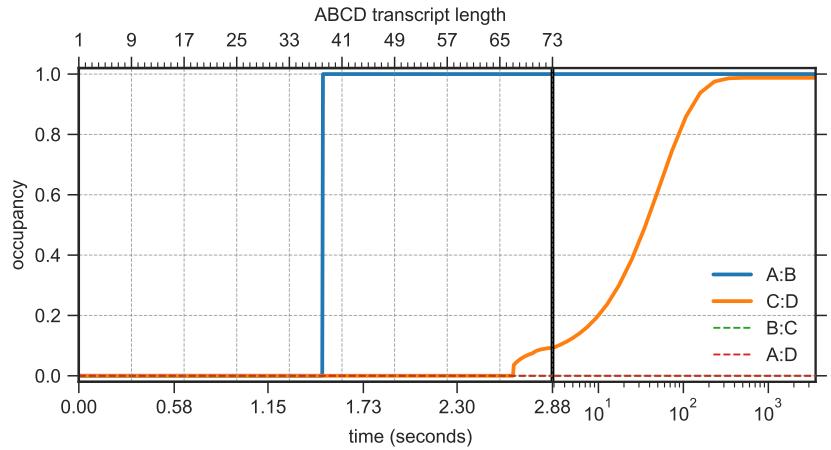
DrTransformer Runtime

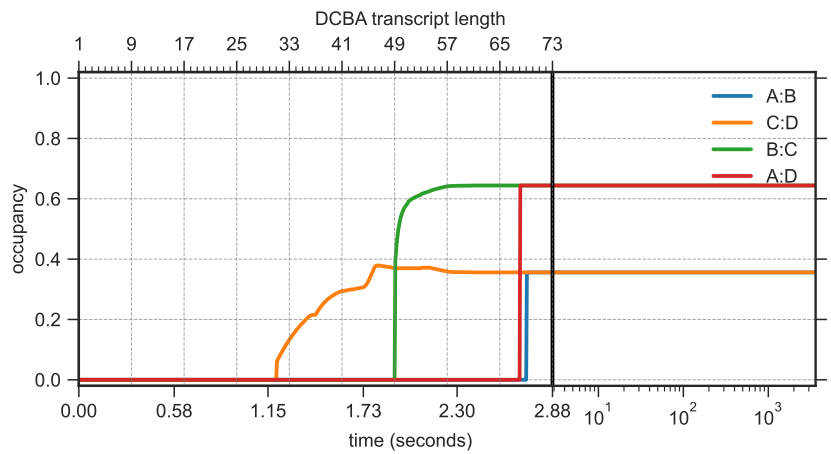


Runtime for group II intron sequences.

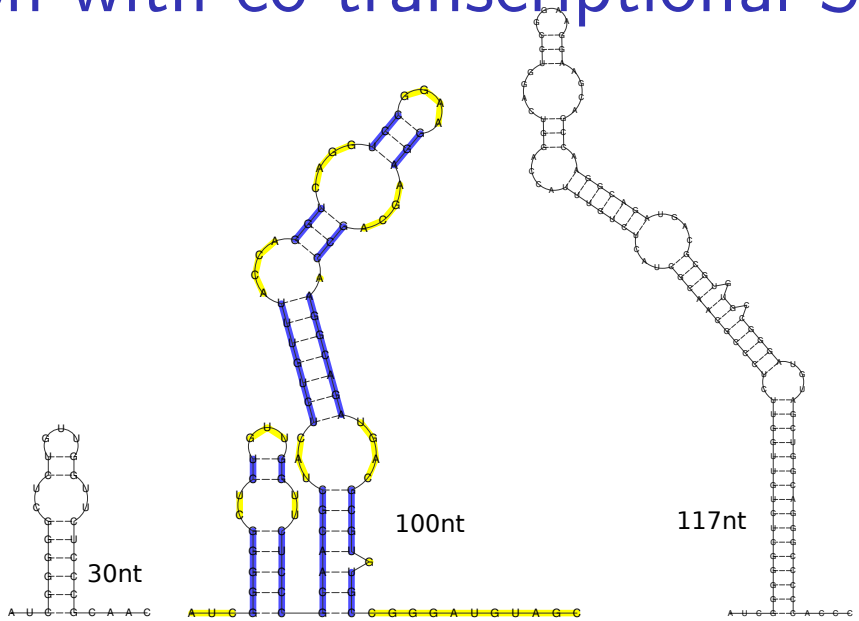
Test on an artificial sequence

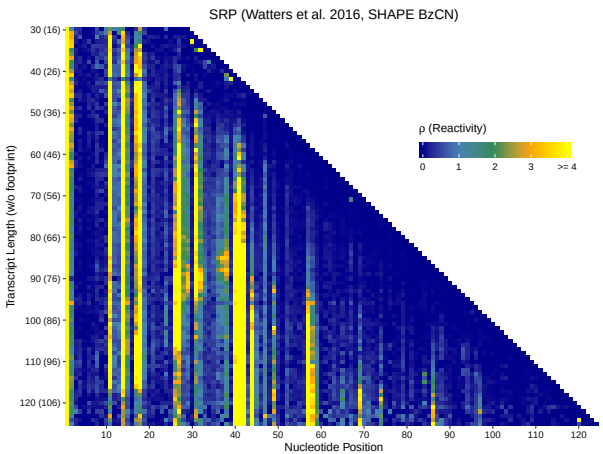


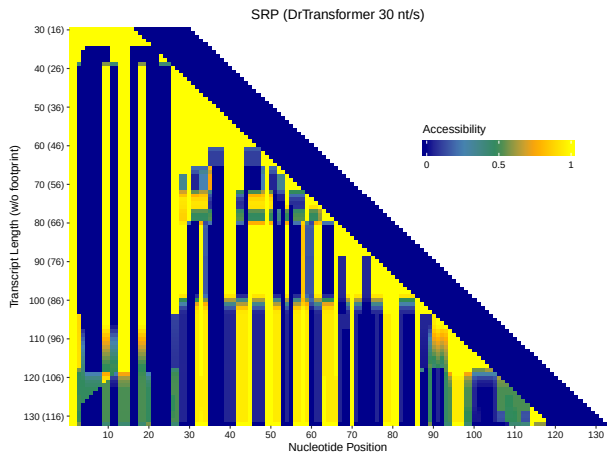




Comparison with co-transcriptional SHAPE-seq



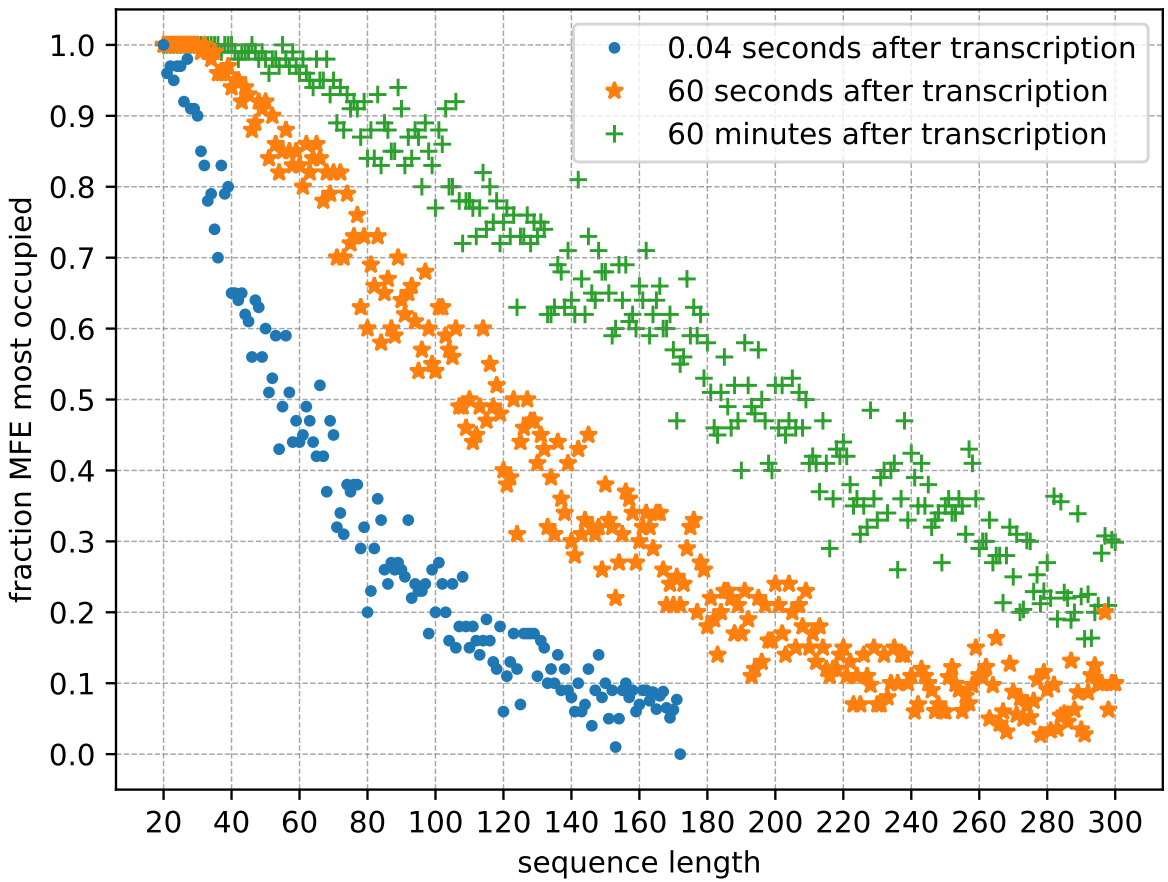




Watters et al, Nat. Struct. Biol. 2016

Is RNA structure determined by co-transcriptional folding?

- Some RNAs will be trapped during co-transcriptional folding can't reach their MFE within reasonable time
- How often does this happen?



Acknowledgments

RNAxplorer:

- Gregor Entzian
- Andrea Tanzer
- Yann Ponty
- Ronny Lorenz

DrTransformer:

- **Stefan Badelt**

Links:

Entzian et al, "RNAxplorer: ..." , Bioinformatics, 2021

`https://github.com/ViennaRNA/RNAxplorer`

`https://github.com/ViennaRNA/drtransformer`