Computational approaches to RNA Folding Kinetics

Ivo L. Hofacker

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

AlgoSB
Marseille, Janvier 2019

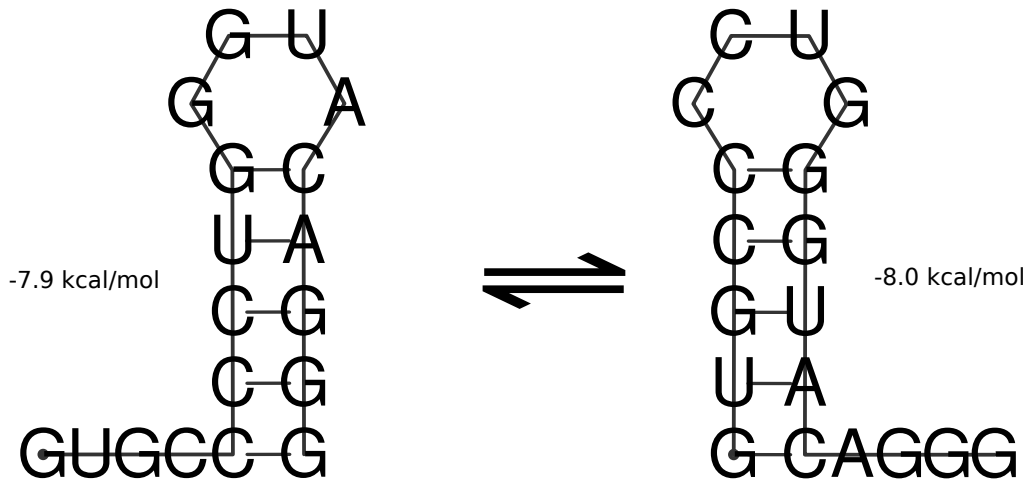
tbi

Thermodynamic vs. Kinetic Folding

Equilibrium properties for RNA secondary strutcures can be calculated efficiently

But what about dynamics?

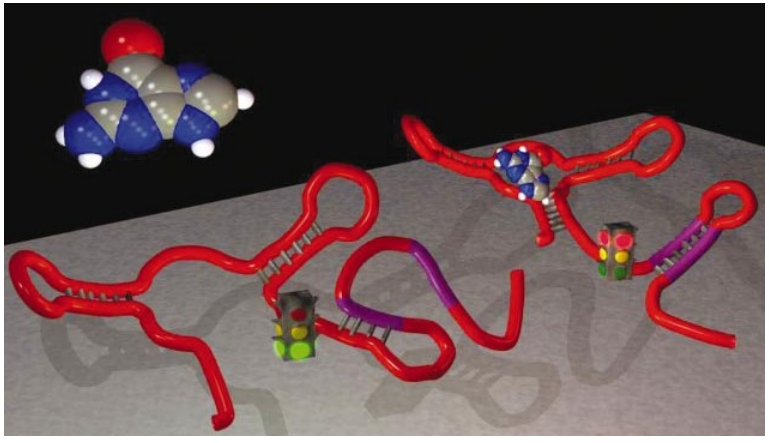
- On what time scale is equilibrium reached?
- How fast/slow is re-folding between dissimilar structures?
- What structures are populated initially?

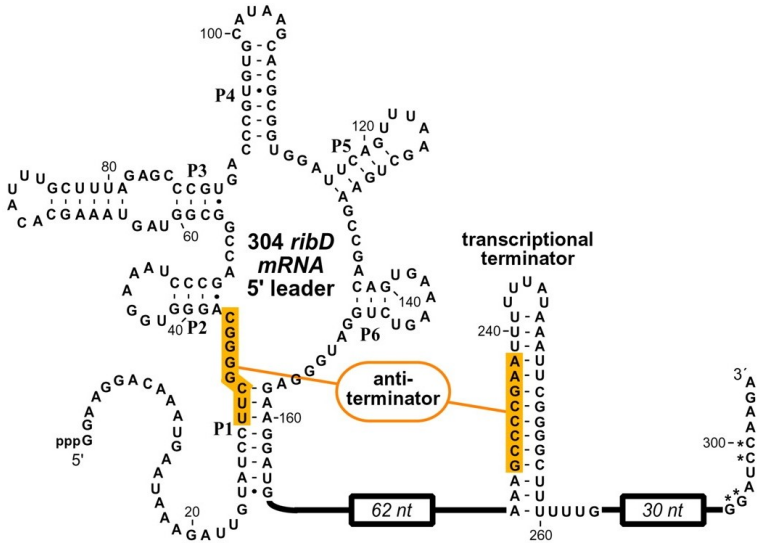


Structural changes are common in functional RNA

RNA switches toggle between active and inactive states by changing conformation.
Used especially to control mRNA translations; triggered by:

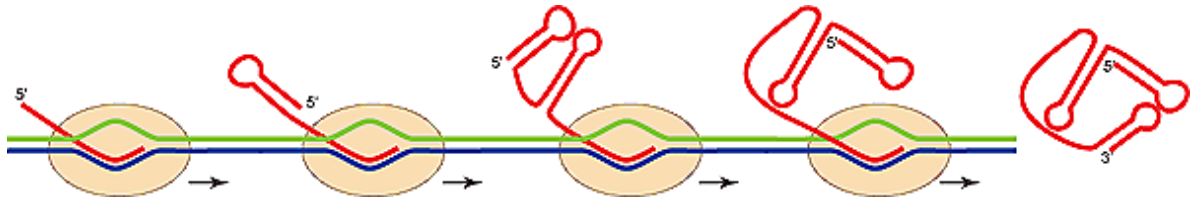
- binding of proteins or small ligands
- chemical modification, e.g. tRNA
- temperature dependent switches
- timed mRNA switches, e.g. HOK





Folding during Transcription

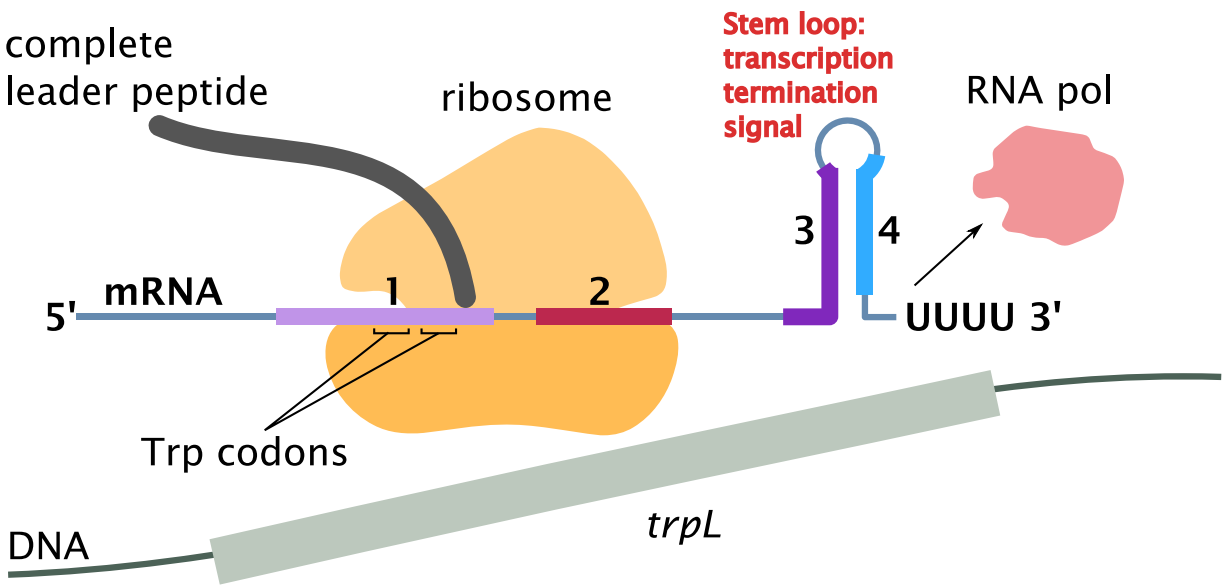
Almost all RNA structures may be affected by co-transcriptional folding:



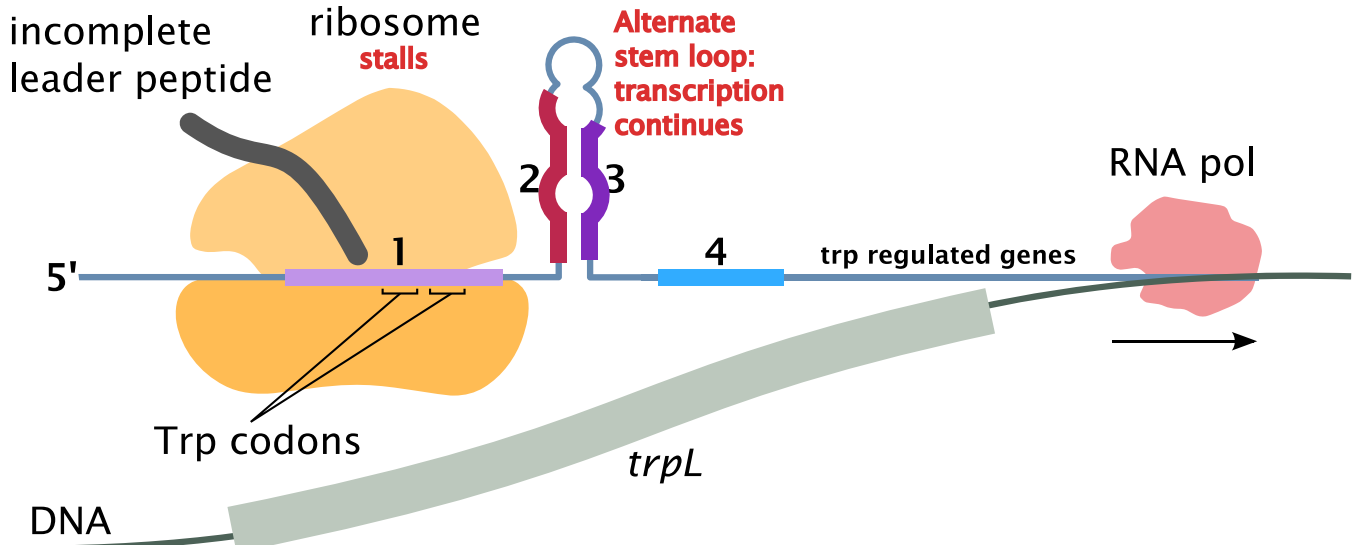
- RNA is transcribed at a rate of only 25–50 nucleotides per second
- The nascent chain starts folding as soon as its leaves the ribosome
- Stems formed by the incomplete chain may be too stable to refold later on
- Co-transcriptional folding may drive the folding process to a well-defined folded state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

Regulation of the Trp Operon

High level of tryptophan

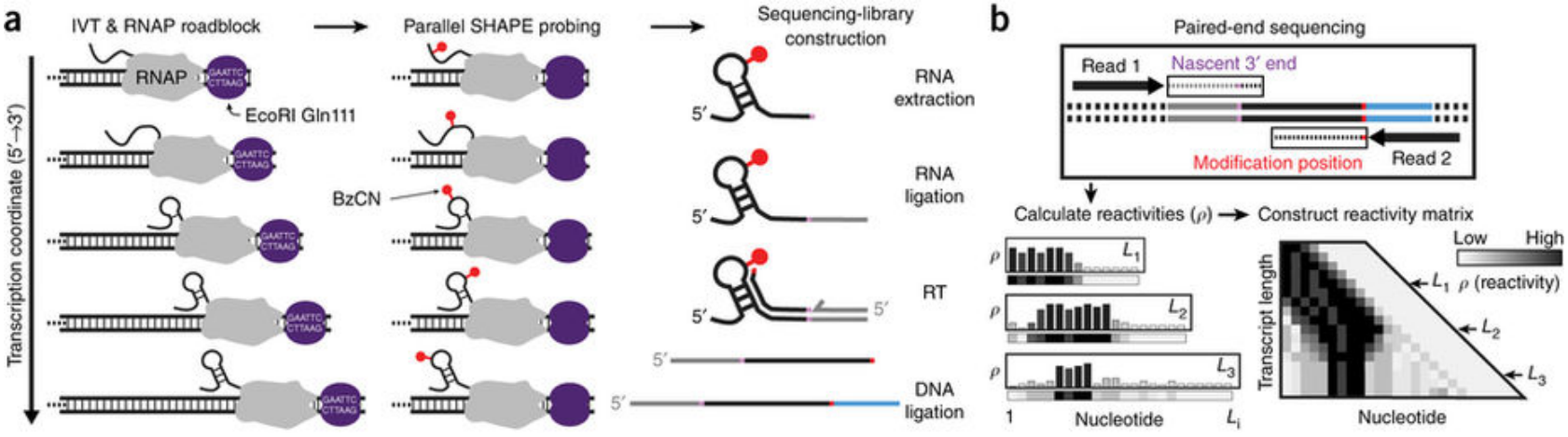


Low level of tryptophan



Co-Transcriptional Structure Probing

Co-transcriptional is becoming experimentally accesssible



Watters et al, Nat. Struct. Biol. 2016

Folding Dynamics as Markov Process

Let's compute prob. $P_x(t)$ of observing structure x at time t. Given transition rates k_{xy}, this gives rise to a *Markov process* with master equation

$$\frac{dP_x(t)}{dt} = \sum_{y \neq x} [P_y(t) k_{xy} - P_x(t) k_{yx}].$$

or in matrix form, with $k_{xx} = -\sum_{x \neq y} k_{yx}$:

$$\frac{d}{dt} P(t) = \mathbf{K} P(t).$$

A *formal* solution can be written simply

$$P(t) = e^{t \cdot \mathbf{K}} P(0)$$

Way too many states to solve directly (10^{17} for a tRNA)

Three 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 grained dynamics via Barriers / Treekin / Barmap
  - Identify local minima, assign macro-states
  - Energy barriers and transition rates (`barriers`)
  - Solve $P_x(t)$ on coarse grained landscape (`treeekin`)
  - Extend sequence and transfer population to next landscape (`barmap`)
- Heuristic landscape construction
  - Model landscape by small set of representative structures
  - Estimate energy barriers and rates
  - Can be nicely combined with co-transcriptional folding `DrTransformer`

Folding Dynamics as Markov Process

But, for a tRNA the dimension of K is about $10^{17} \times 10^{17}$
The formal solution is therefore of limited use.
We can:

- Solve toy models by integration of the master equation
- Perform stochastic folding simulations.
  Needs many trajectories.
- Reduce the number of conformations by coarse graining
  i.e. lump structures together into *macro states*
- Just try to compute a single best folding pathway.

Stochastic Simulations

Simulate folding kinetics by Gillespie (rejectionless Monte Carlo) algorithm :

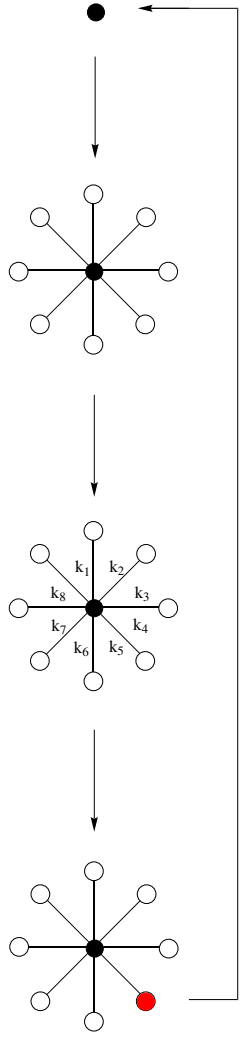
Generate all neighbors using a move-set
Close base pair – Open base pair

Assign rates to each move, e.g.:

$$k_i = \Gamma \cdot \min\left\{1, \exp\left(-\frac{\Delta E}{kT}\right)\right\}$$

Select a move i with probability $\propto k_i$

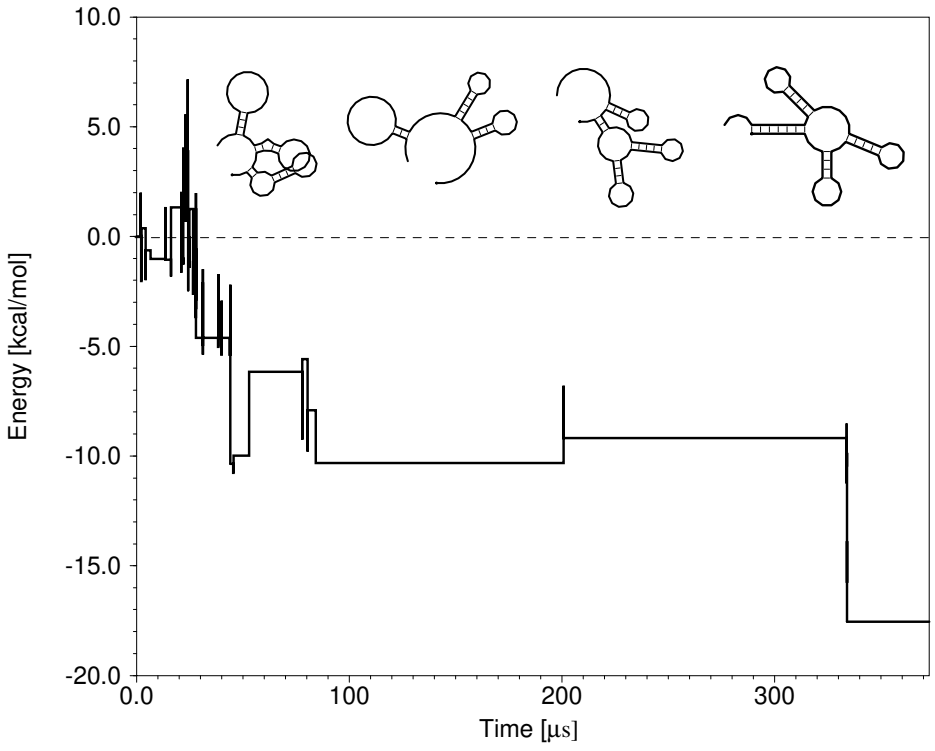
Advance clock by $1/\sum_i k_i$ (on average).



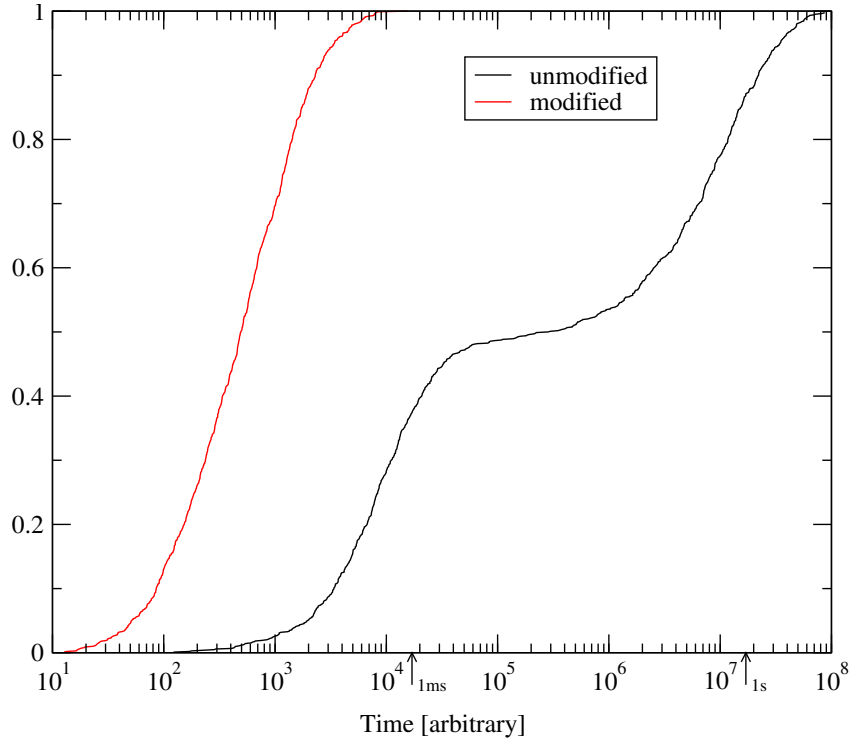
- computationally somewhat expensive
- need to analyze many trajectories
- easy to include co-transcriptional folding

Simulated folding of $tRNA^{phe}$

Many trajectories have to be collected in order to do statistics.



energy profile of a single trajectory

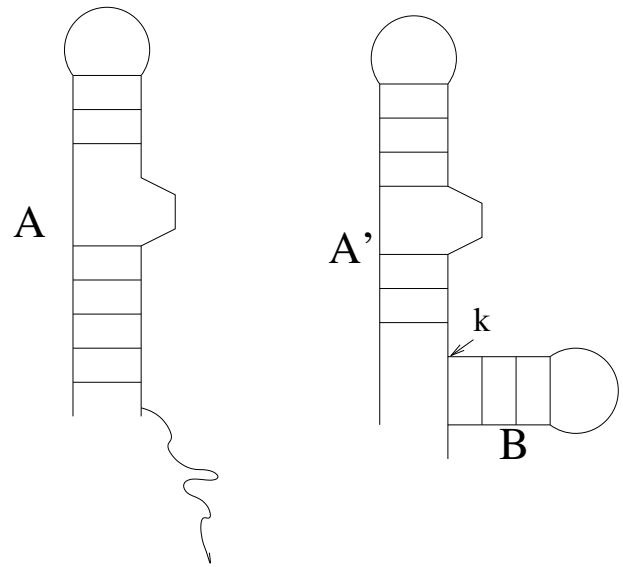


distribution of first passage times

Folding Simulation using Isambert's `Kinefold`

- Use opening/closing of entire helices as move set
- Allows pseudo-knots,
- Suitable for RNAs up to several hundred nt.

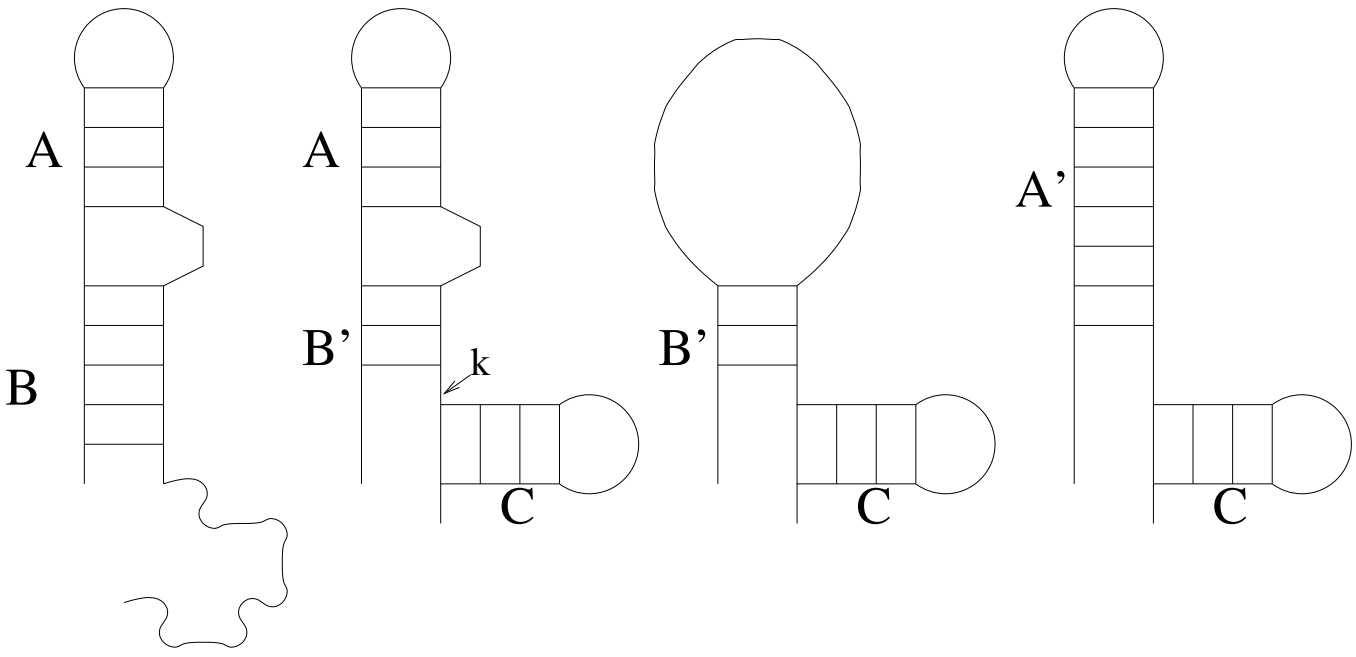
Helix moves require a local conflict resolution after each step



Web service available at `http://kinefold.curie.fr`

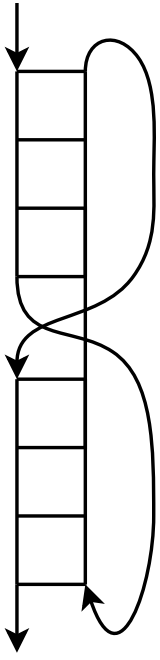
Potential problems with Conflict Resolution

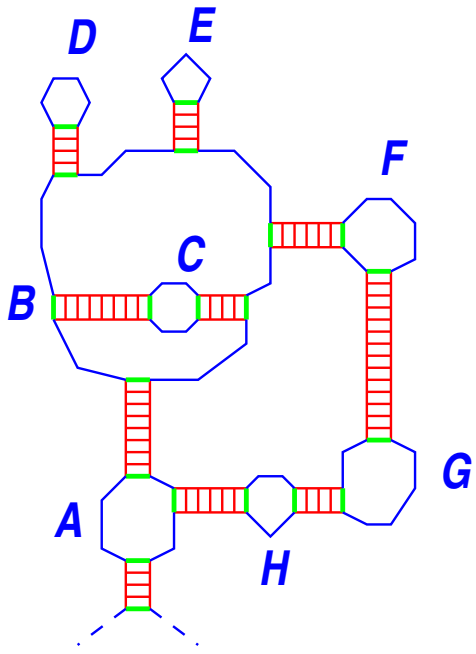
Maintaining detailed balance with helix moves is non-trivial:

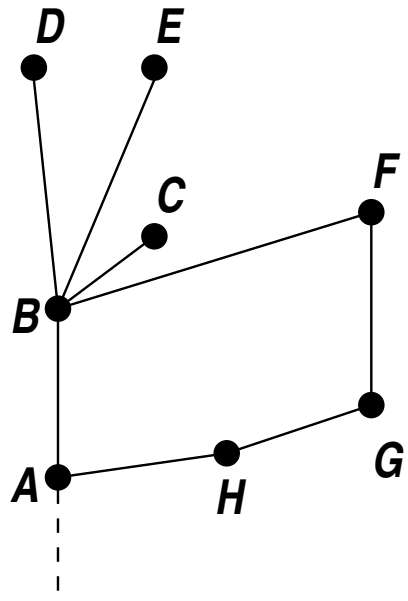


Pseudo-knots

- Pseudo-knots do not pose a problem for folding simulations.
- Still requires accurate pseudo-knot energies







- Frequently only H-type knots are considered.
- Kinefold allows complex pseudo-knots whose entropy is approximated by a cross-linked "Gaussian gel"

Kinetic Rate Models

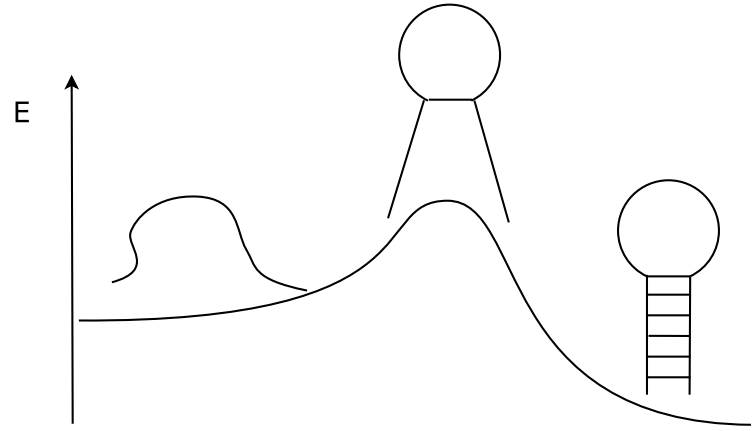
The simplest rate model satisfying detailed balance is the Metropolis rule

$$k_{xy} = \Gamma \cdot \max\left(1,\ e^{(\Delta G(x) - \Delta G(y))/RT}\right)$$

More accurate models define a transition state with free energy $\Delta G^{\dagger}$ and Arrhenius rates:

$$k_{xy} = \Gamma \exp\left(-(\Delta G^{\dagger}_{xy} - \Delta G(x))/RT\right)$$

This is essential for large moves (e.g. helix moves).



Abstract Definition of Landscapes

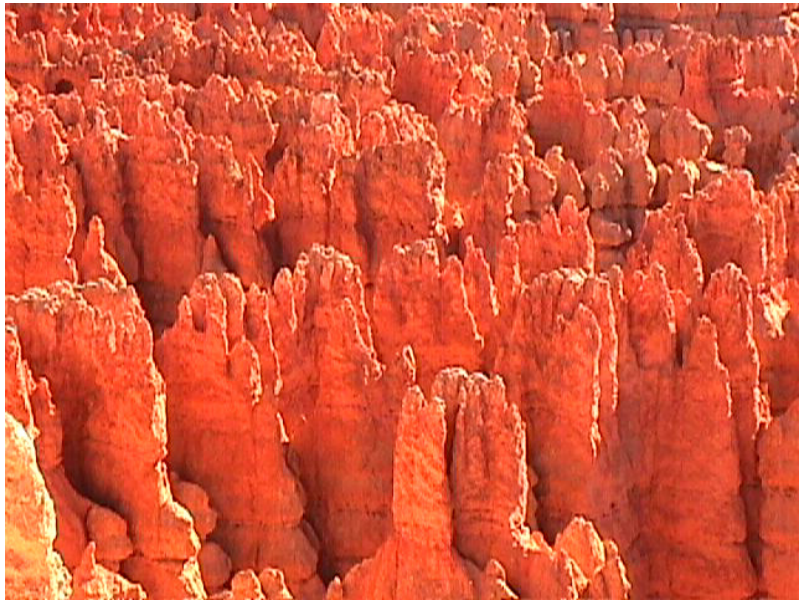
A **landscape** is a triple $(V, \mathcal{X}, f)$ where

- V is a set of *configurations*.
  E.g.: RNA sequences, tours of a travelling salesman, spin configurations,

  secondary structures of given RNA molecule;
- f is a *cost* or *fitness function* $f : V \to \mathbb{R}$;
- $\mathcal{X}$ is a way of defining "nearness", "closeness", "dissimilarity", or "accessibility" among the configurations.

  E.g. an *adjacency relation* (thus a graph), *transition matrix* (defining a Markov chain), or a *(pre)topology* on V.

Ruggedness



Rugged: Bryce Canyon UT

Smooth: Capulin Volcano NM

Measures of Ruggedness:

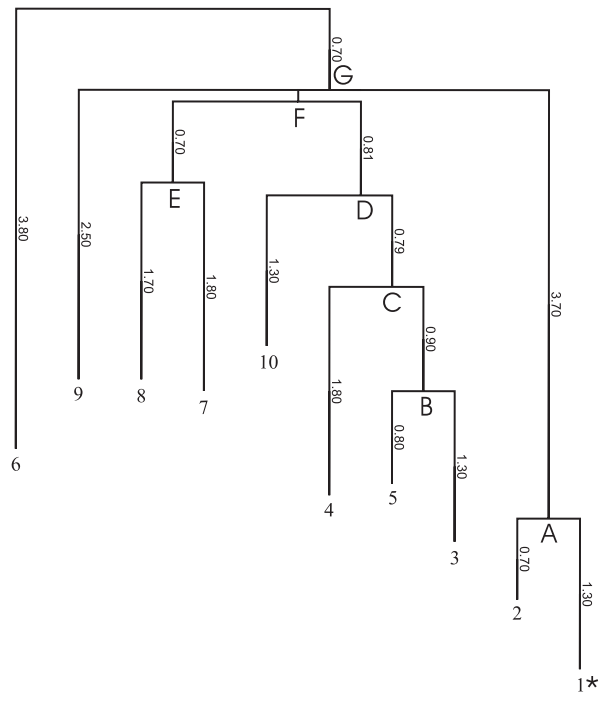
- Number of Local Minima and Maxima
- Correlation length
- Basin sizes

- Length of Adaptive Walks

RNA Landscape Analysis

Barrier trees

- Contains all local minima as leafs
- Barrier heights and saddles between minima
- Groups structures into *macro states*
- Transition rates between macro states $\rightarrow$ coarse grained dynamics
- Time and space proportional to the size of the landscape
  Limited to RNA < 100nt
- Sampling based heuristics for longer RNAs



Flamm et al, Z. Phys. Chem. (2002), Kucharík et al, Bioinformatics (2014, 2016)

Calculating barrier trees

The flooding algorithm:

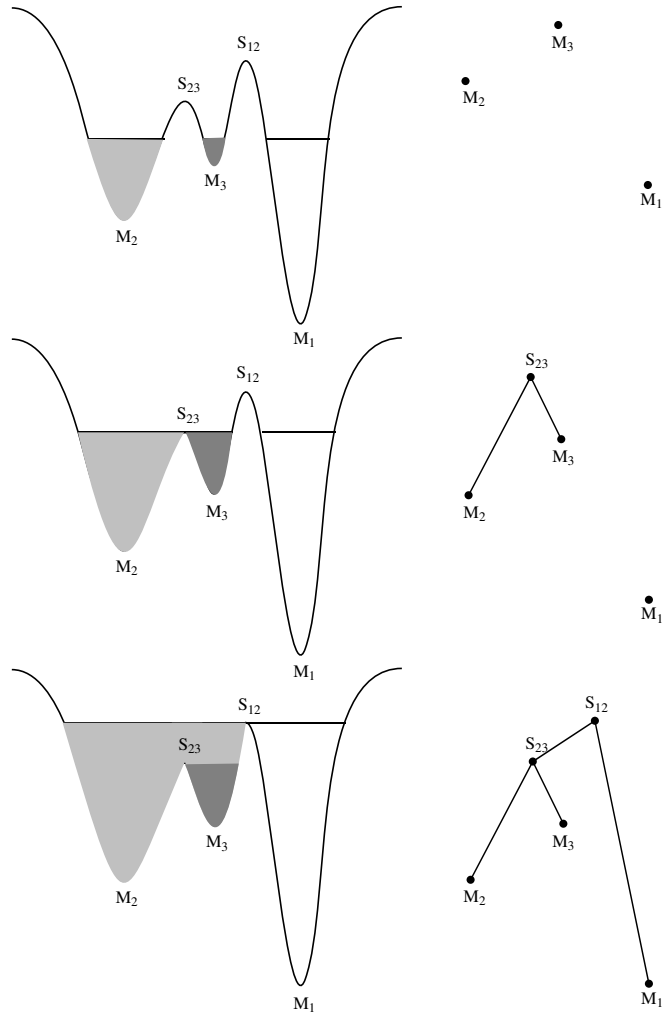
Read conformations in energy sorted order.

For each confirmation x we have three cases:

(a) x is a *local minimum* if it has no neighbors we've already seen

(b) x belongs to basin $B(s)$, if all known neighbors belong to $B(s)$

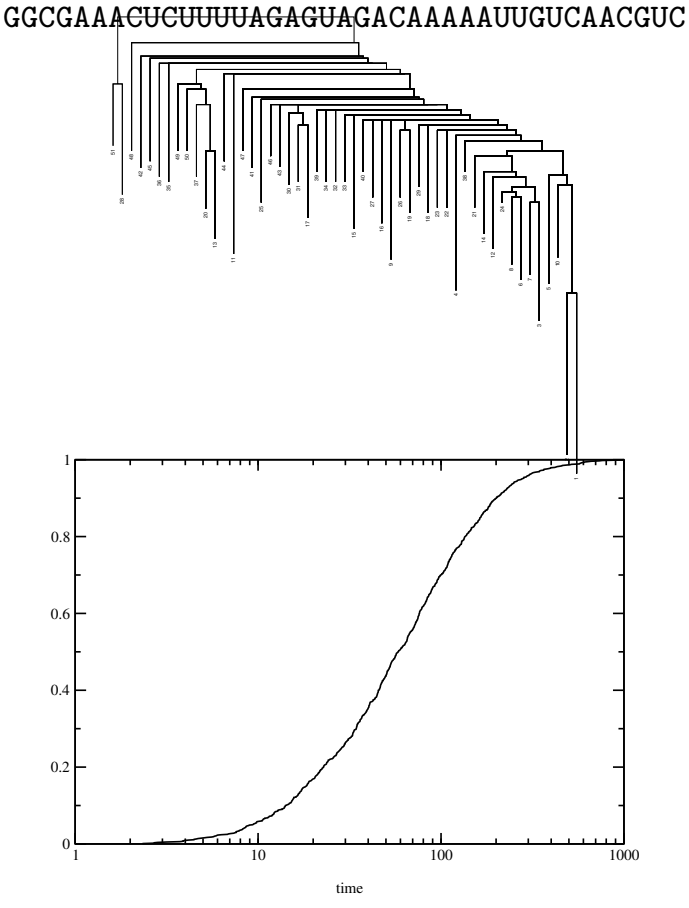
(c) if x has neighbors in several basins $B(s_1) \ldots B(s_k)$ then it's a *saddle point* that *merges* these basins.

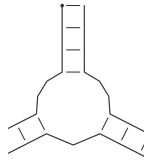


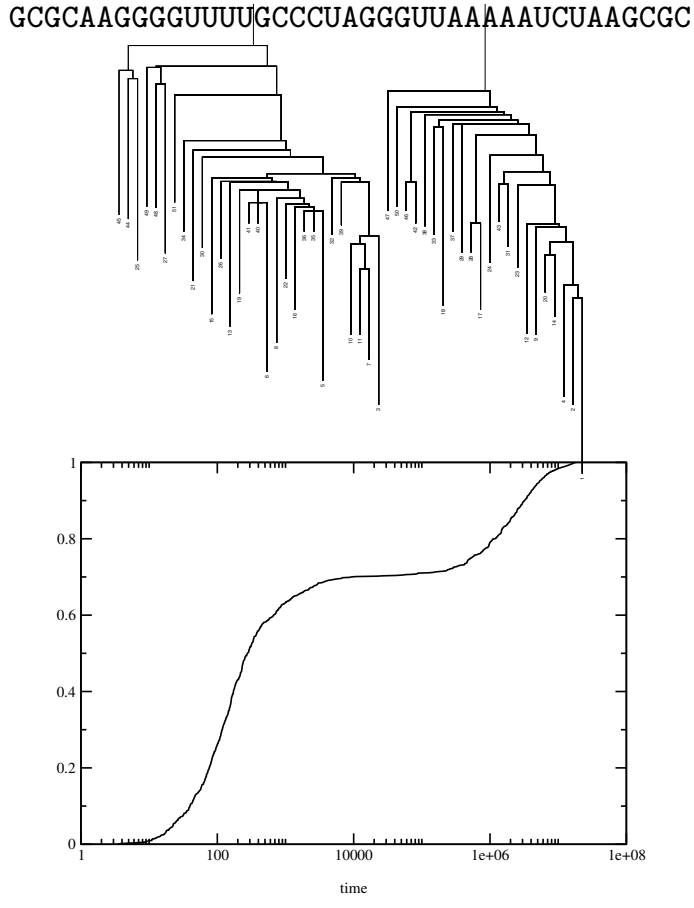
The barriers program

- Computes all local minima
- Barrier heights and saddle points between minima
- Optimal refolding paths between any two minima
- Groups structures into *macro states* connected to each minimum
- Computes effective transition rates between macro states $\rightarrow$ coarse grained dynamics can be computed without simulation
- Time and space $\mathcal{O}(N \cdot n)$ for an RNA of length n with N structures. However, N grows exponentially

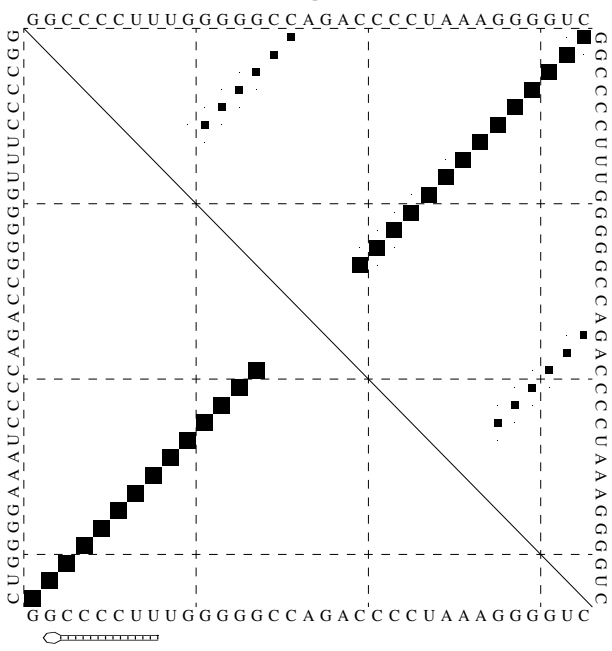
Fast Folder vs. Slow Folder

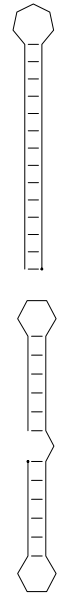






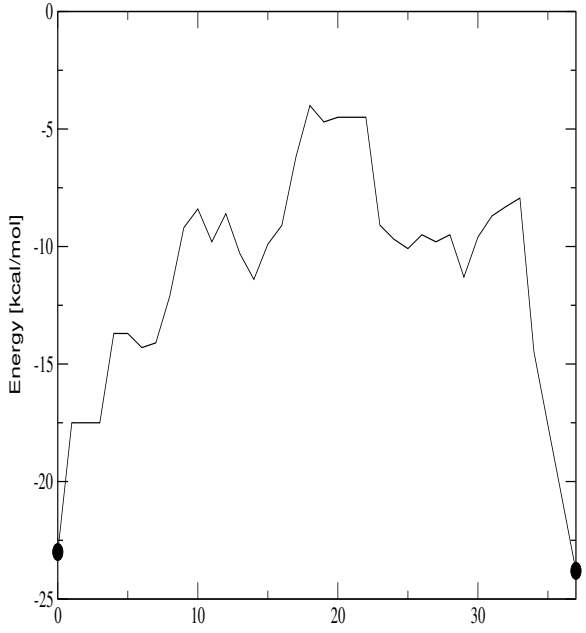
A designed bi-stable Sequence

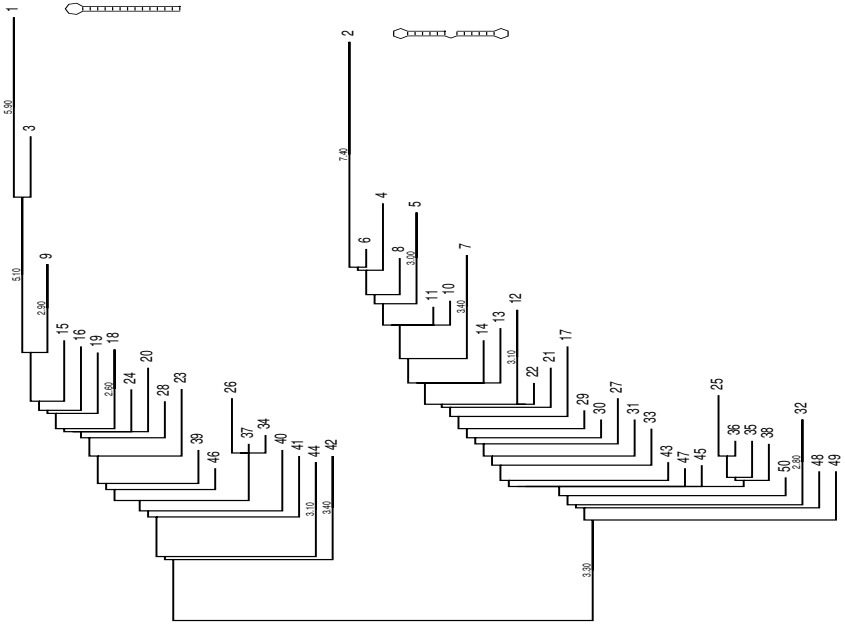




-23.8 kcal

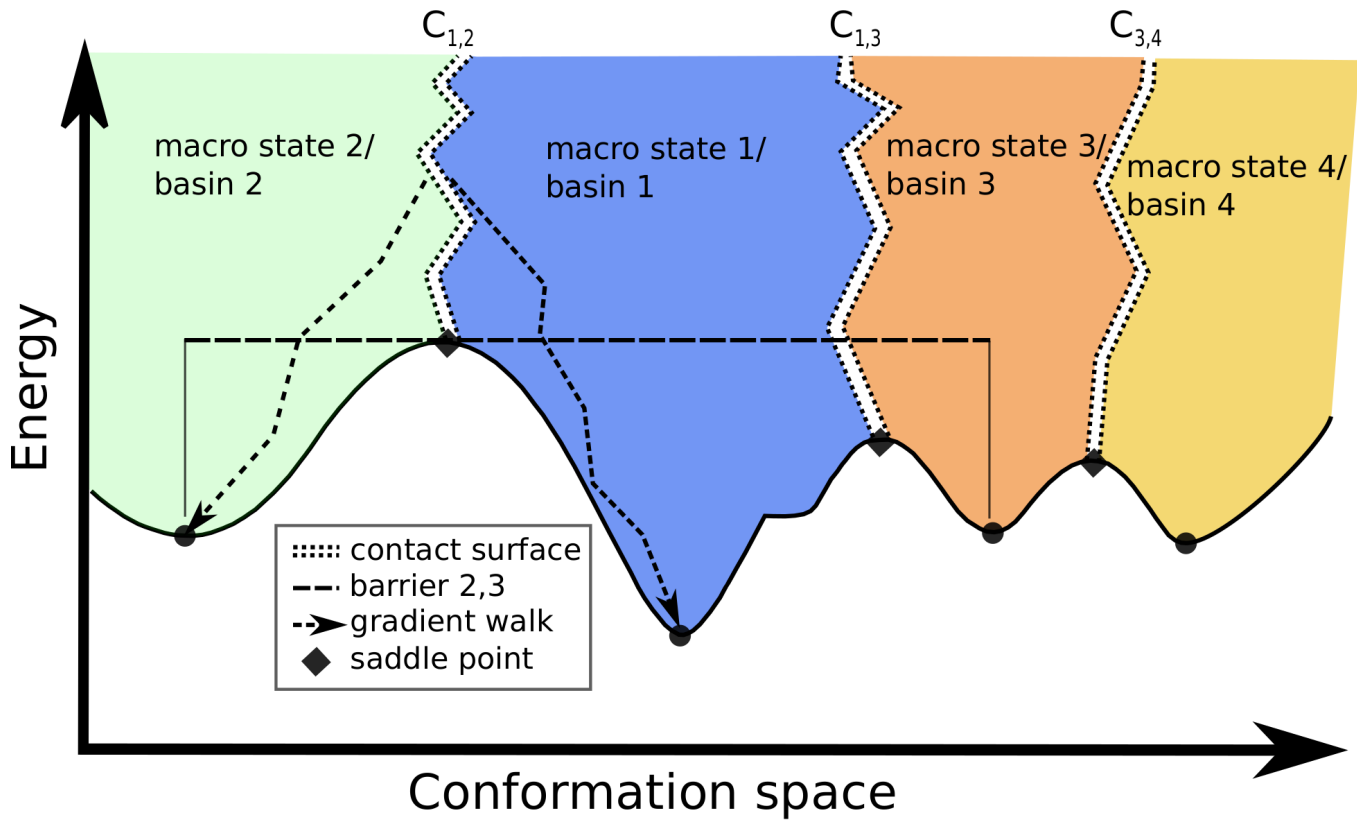
-23.0 kcal





```
((((((....)))))).((((((....))))))  -23.00
(((((.(...)))))).((((((....))))))  -17.50
((((.((...)))))).((((((....))))))  -17.50
(((.(((...)))))).((((((....))))))  -17.50
.((.(((...))))).((((((....))))))  -13.70
.(.(((...))))).((((((....))))))  -13.70
.(.(((...))))).).((((((....))))))  -14.30
...(((...))))...((((((....))))))  -14.10
....(((...))).....((((((....))))))  -12.10
.....((...)).....((((((....))))))  -09.20
.....(.....).....((((((....))))))  -08.40
...................((((((....))))))  -09.80
..........(.....).((((((....))))))  -08.60
..........((...)).((((((....))))))  -10.30
........(((...))))((((((....))))))  -11.40
.......(((...)))).(((((....)))))..  -09.90
......(((((...)))))((((....))))..  -09.10
.(...)(((((...)))))((((....))))..  -06.20
.(....((((((.....))))))((((....))))..  -10.00
((....(((((...)))))((((....))))))  -04.70
(((...(((((...))))).(((....)))))  -04.50
((((..(((((...)))))..((....))))))  -04.50
(((((.(((((...)))))...(....)))))  -04.50
((((((((((...))))).........))))))  -09.09
((((((((((...))))).........))))))  -09.69
((((((((((...))))).........))))))  -10.09
((((((((((...)))(.....).))))))  -09.50
((((((((((...)).(.....)))))))))  -09.80
((((((((((...))(.....)))))))))  -09.50
((((((((((...))((....))))))))))  -11.30
((((((((((.....)((....))))))))))  -09.60
(((((((((.(....)((....))))))))))  -08.70
((((((((((....).(....))))))))))  -08.30
((((((((((....).......))))))))))  -07.94
((((((((((..........))))))))))  -14.48
((((((((((.........))))))))))  -17.60
((((((((((.......))))))))))  -20.70
((((((((((.....))))))))))  -23.80
```

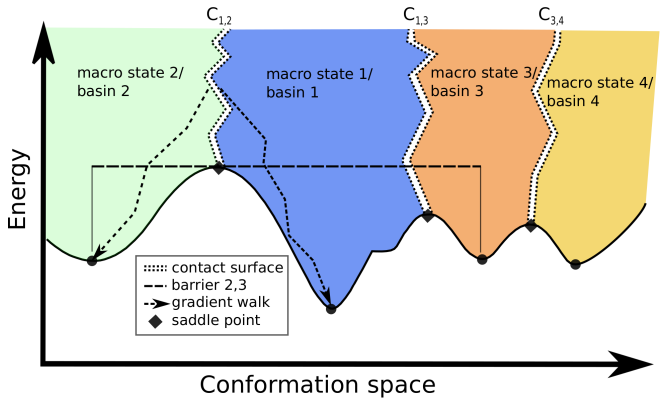
Coarse Graining the Landscape



Coarse Graining the folding dynamics

For a reduced description we need

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



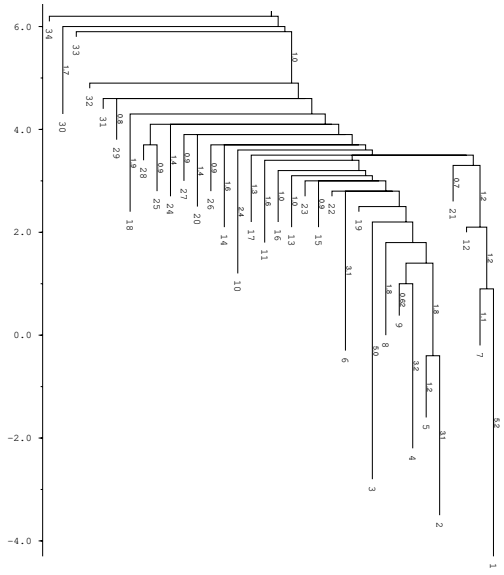
Transition rates could follow an Arrhenius rule
$r_{\beta\alpha} = \exp\left(-(E^*_{\beta\alpha} - G_\alpha)/RT\right)$.

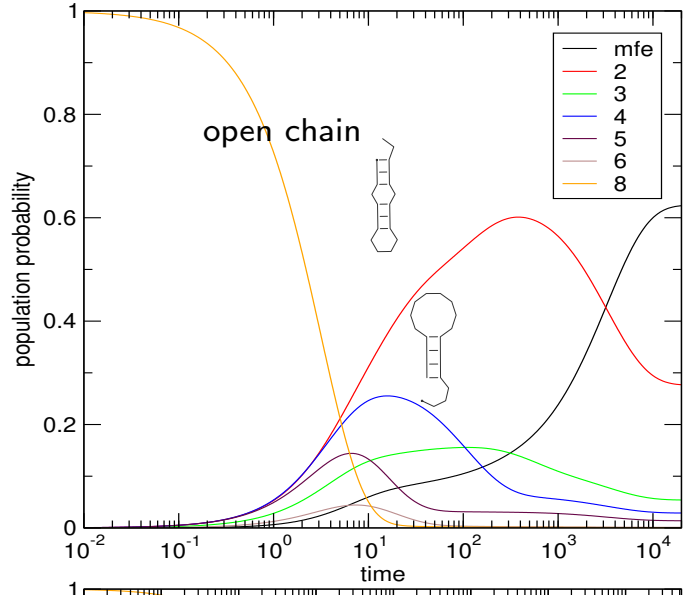
Better: include *all* transition states

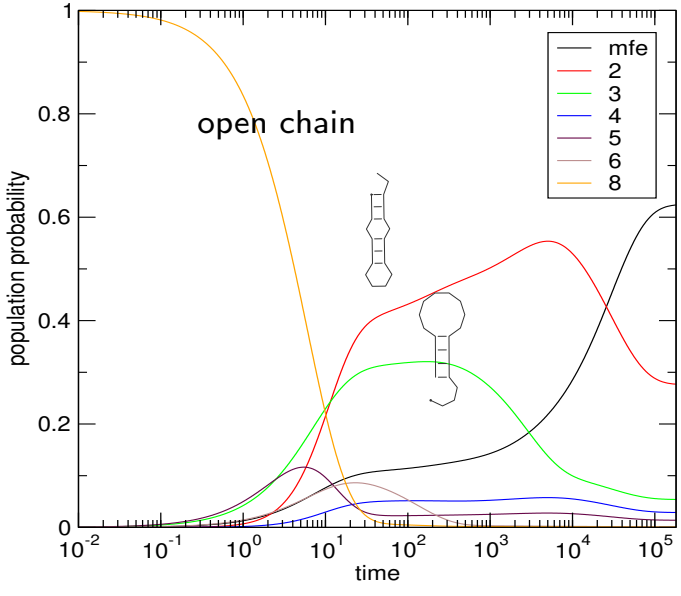
$$r_{\beta\alpha} = \sum_{y\in\beta}\sum_{x\in\alpha} r_{yx}\mathrm{Prob}[x|\alpha] \approx \frac{1}{Z_\alpha}\sum_{y\in\beta}\sum_{x\in\alpha} r_{yx}e^{-E(x)/RT}$$

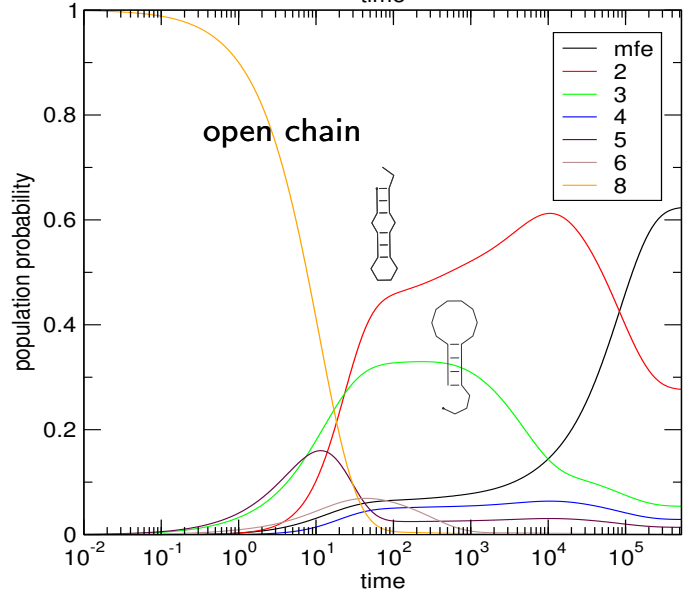
assuming local equilibrium.

Coarse grained dynamics vs. full dynamics





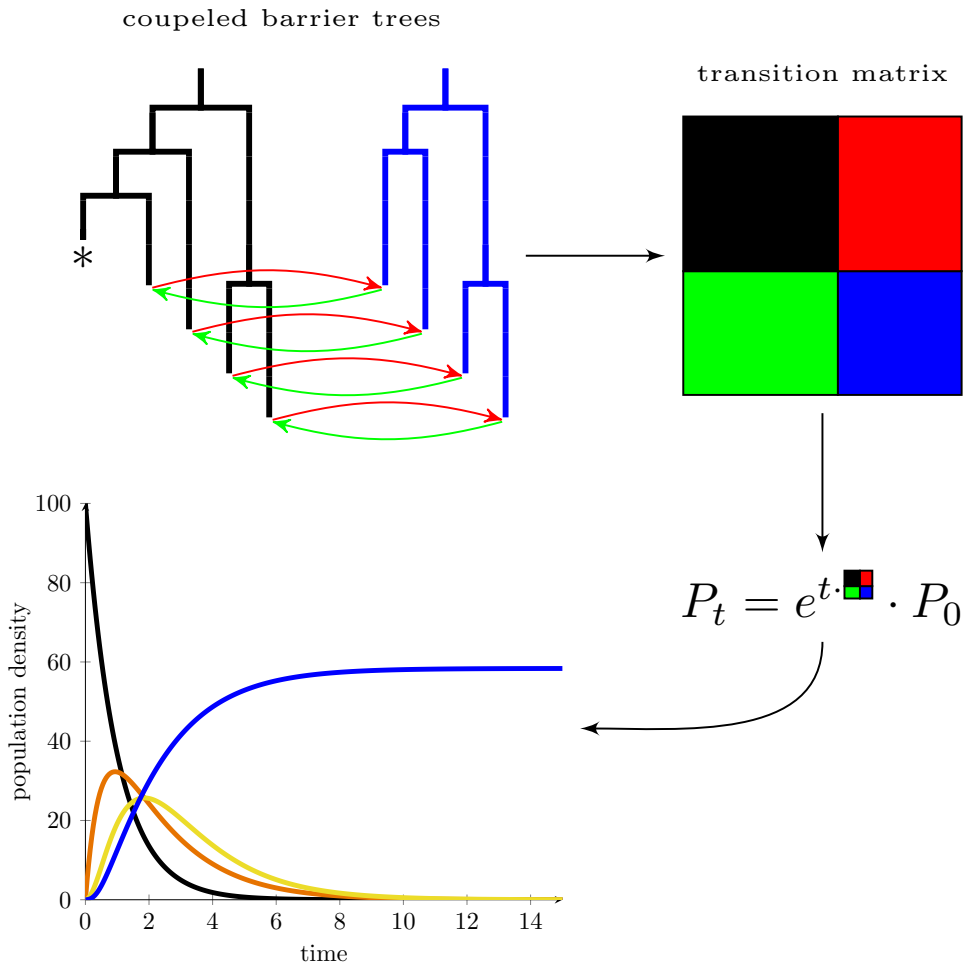




How to include Ligand Binding ?

- Need to know binding motif and binding rates from experiment
- Simple strategy:
  - Add binding energy $\theta = RT \ln \frac{K_d}{c^{\ominus}}$ to every binding competent structure
  - Assumes infinite ligand concentration and infinitely fast binding
- Treat binding / unbinding events explicitly
  - Barrier trees for bound and unbound states
  - Usual rates within bound / unbound structures
  - Concentration dependent rate of complex formation
    $k_{\text{off}} = k_{\text{on}} e^{-\theta/RT}, \quad r = k_{\text{on}} \cdot C$

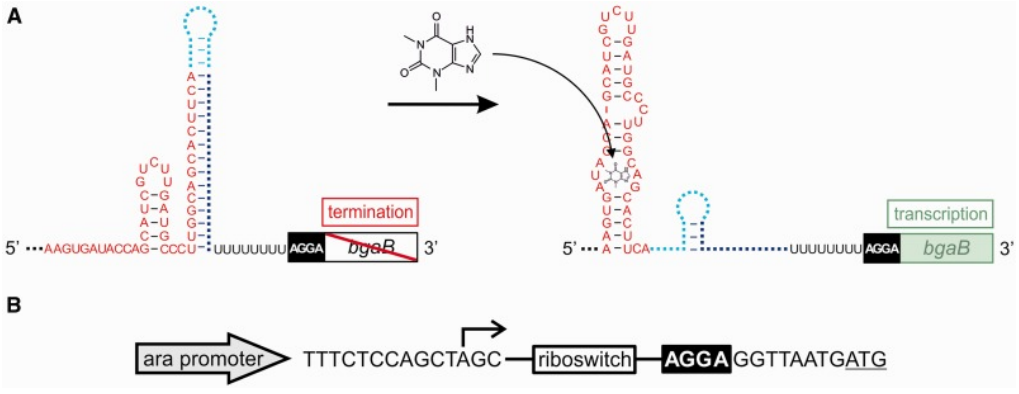
How to include Ligand Binding ?



Kühnl et al, BMC Bioinf. (2017), Wolfinger et al. Methods (2018)

An Artifical Riboswitches

A designed *transcriptional* switch

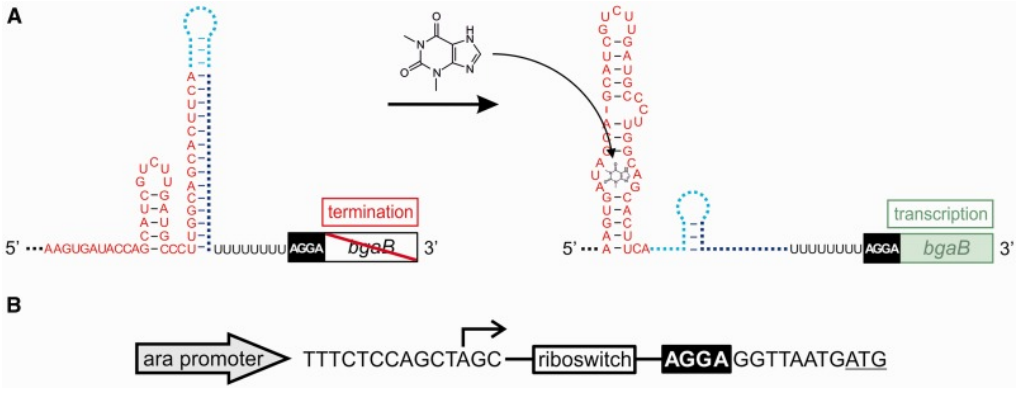


Wachsmuth et al, NAR (2013)

- Theophylline binding to the aptamer inhibits terminator hairpin
- How to model the effect of the ligand?
- *Co-transcriptional* folding
  Terminator can act only if it is formed fast enough

An Artifical Riboswitches

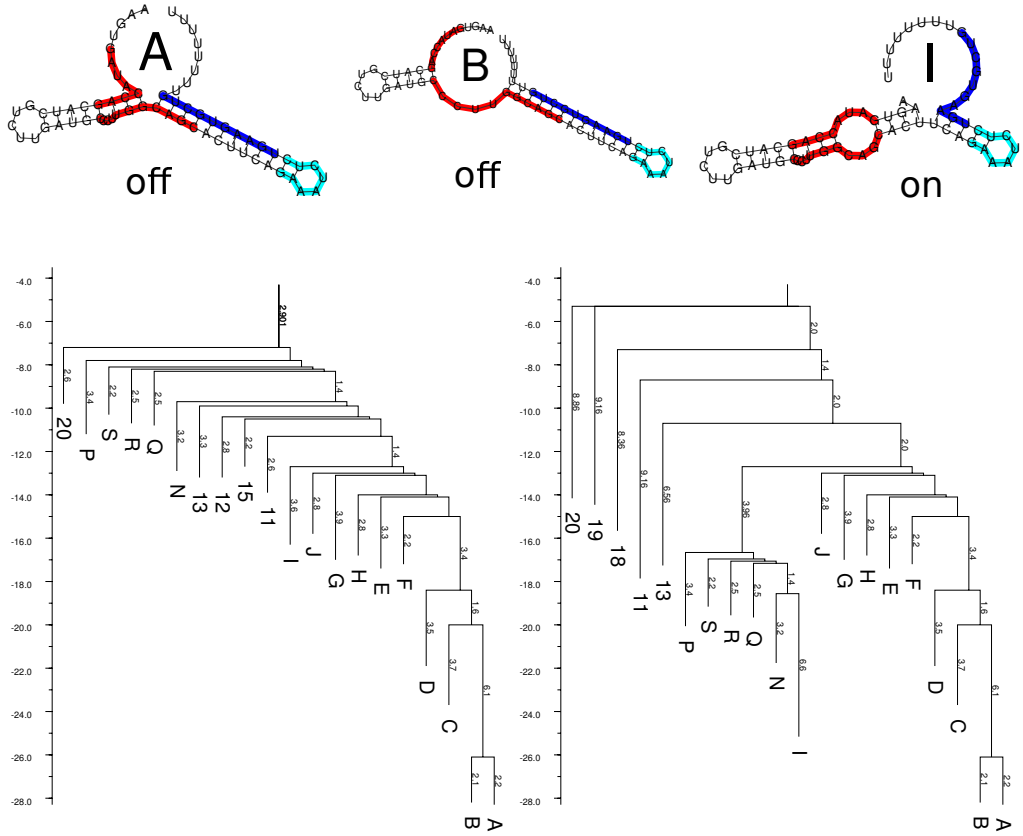
A designed *transcriptional* switch



Wachsmuth et al, NAR (2013)

- Theophylline binding to the aptamer inhibits terminator hairpin
- How to model the effect of the ligand?
- *Co-transcriptional* folding
  Terminator can act only if it is formed fast enough

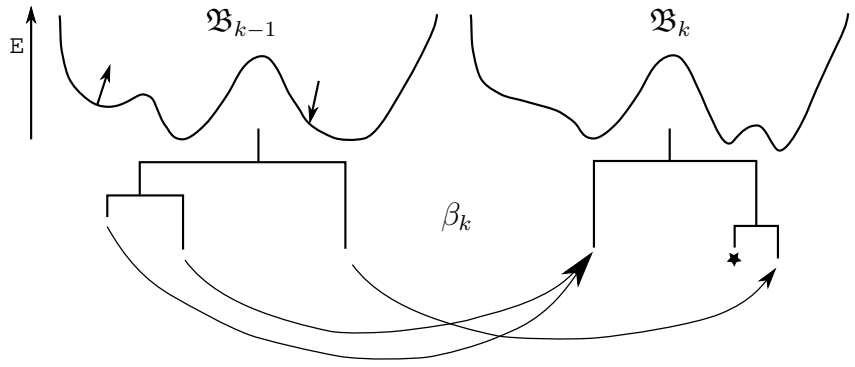
Barrier Tree for RS10 with and without Theophylline



- Binding motif and K_d measurements
- Binding-competent structures are stabilized by about 8.9kcal/mol
- $\Rightarrow$ Distortion of the folding landscape by ligand

Co-transcriptional with BarMap

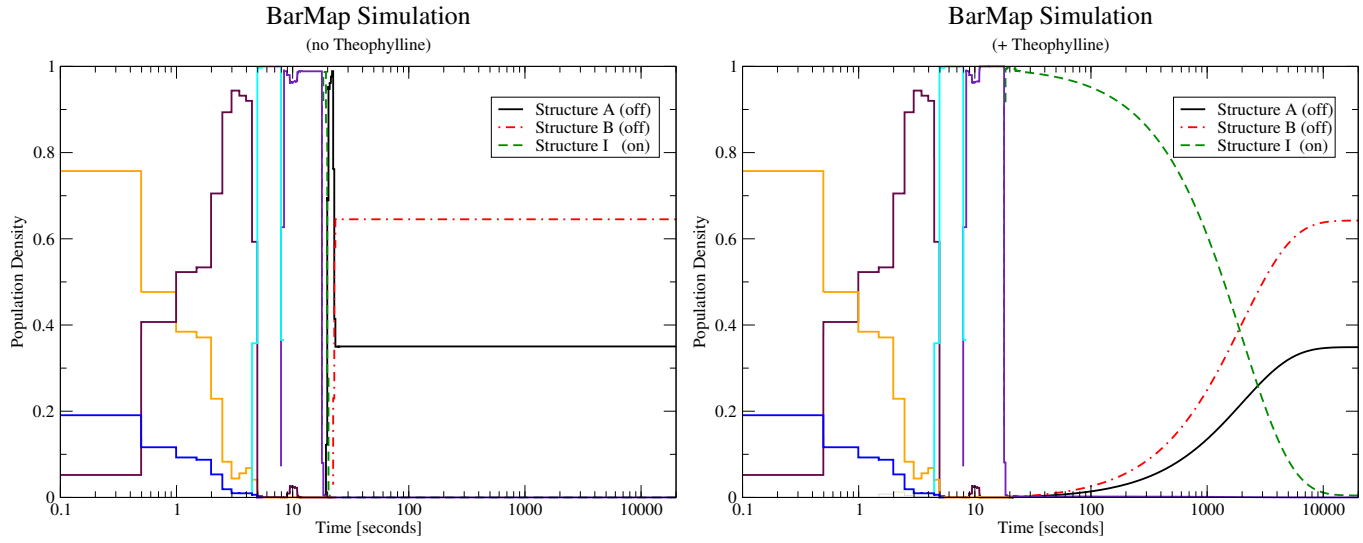
Each extension of the RNA structure modifies the landscape:



- Compute barrier trees for each sequence length $1 \ldots n$
- Compute a mapping between the minima of subsequent landscapes
- Compute dynamics piece-wise:
  - Compute dynamics on landscape for length k
  - Transfer population to landscape of length $k + 1$

Hofacker et al., RNA (2010)

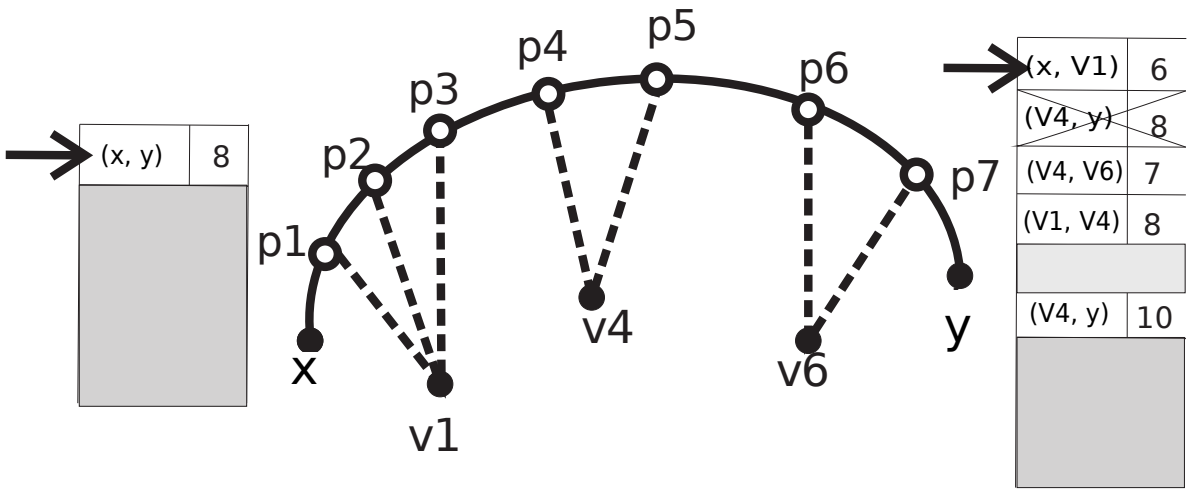
Co-transcriptional of the RS10 Riboswitch



- Without theophylline, the RNA is in equilibrium at the end of transcription
  Terminator is formed, transcription terminates
- With theophylline, almost 100% in state I (on-state)
- Only few of the initial designs show switching behavior

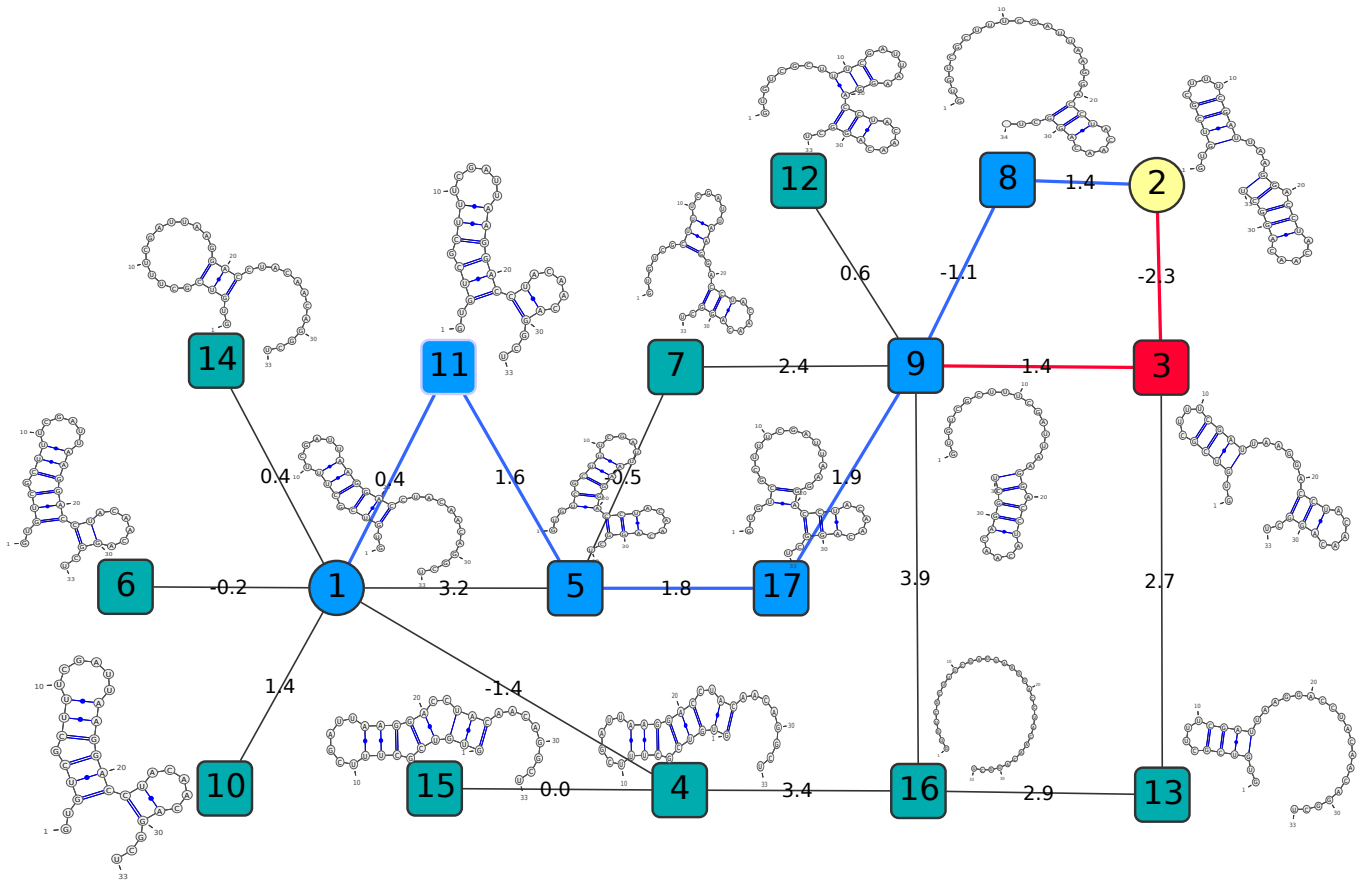
Approximation of Basins and Barriers

- Idea: sample local minima and connect them by **direct paths**
- Sampling:
  - sample secondary structures from a Boltzmann ensemble
  - use adaptive or gradient walk to find the corresponding minimum
- Construct connecting paths recursively: subdivide estimates at intermediate minima



$\Longrightarrow$ Basin hopping graph of the landscape

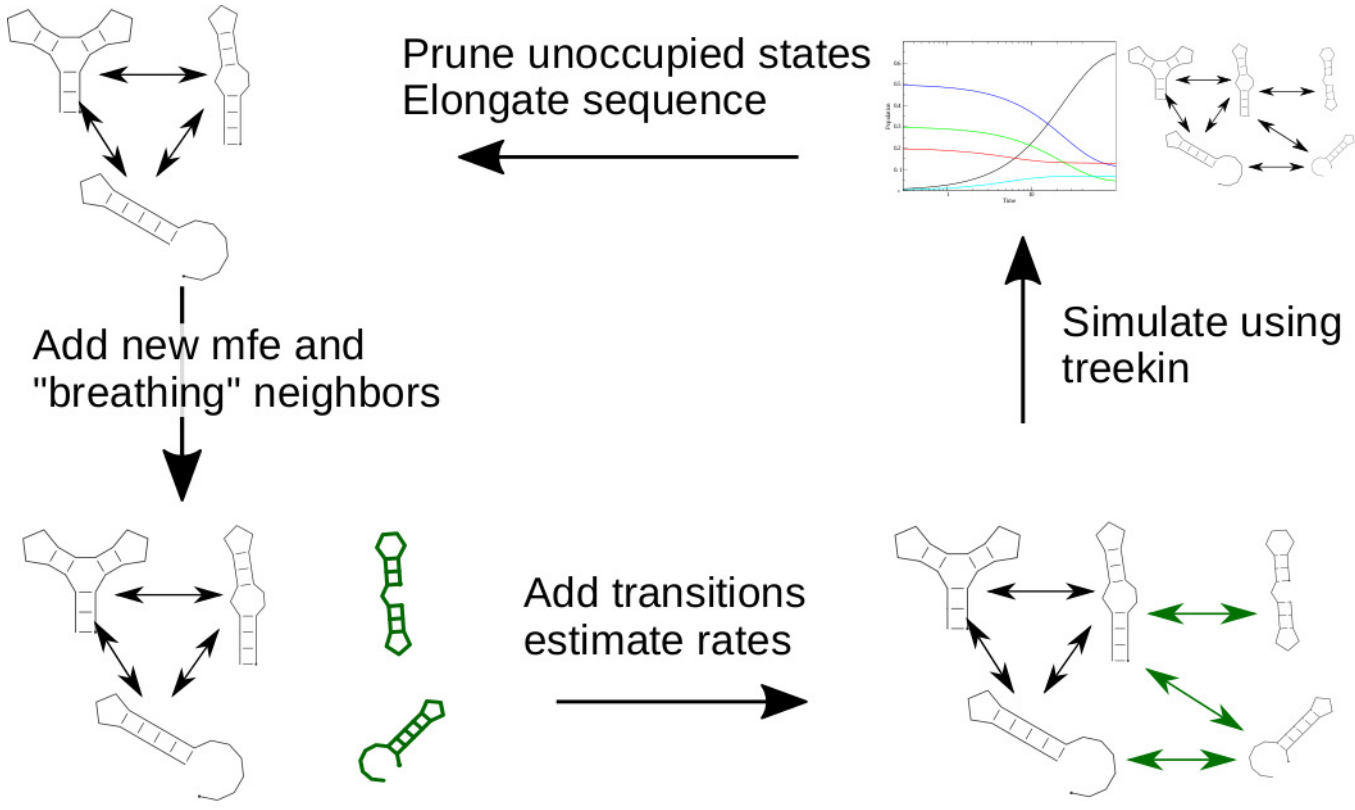
Basin Hopping Graph



Kucharík et al, Bioinformatics (2014, 2016)

DrTransformer: Ultrafast co-transcriptional Folding

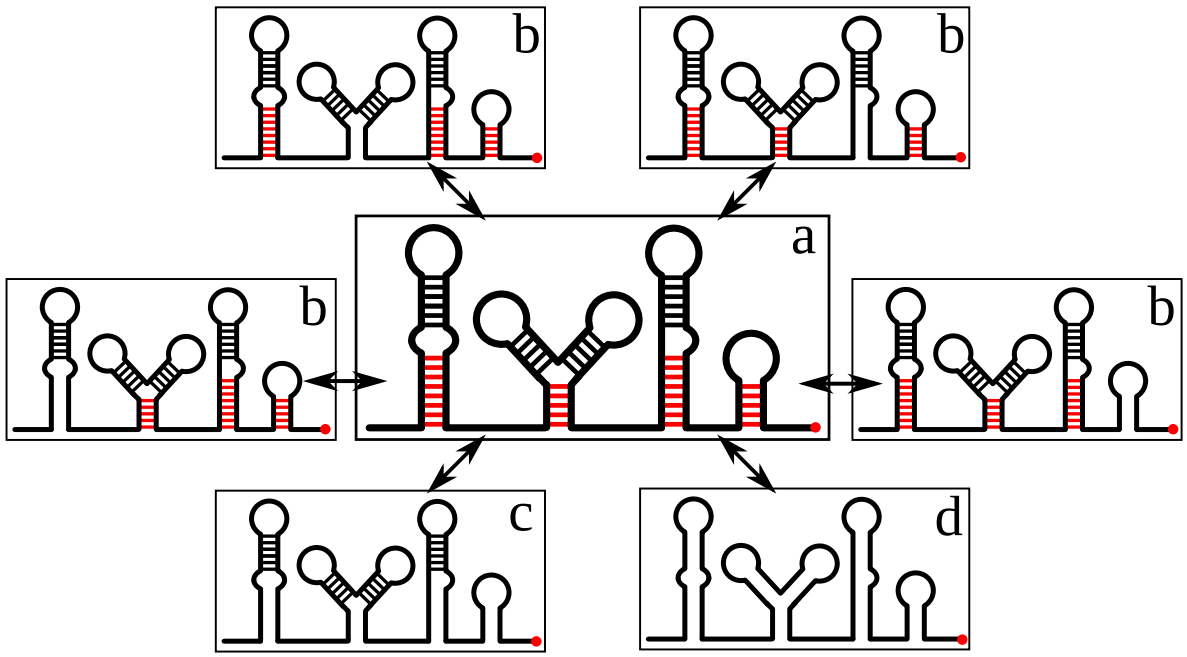
- Simulate a **small** network consisting only of the most relevant structural states
- Evolve network as RNA grows



DrTransformer: "Breathing" neighbors

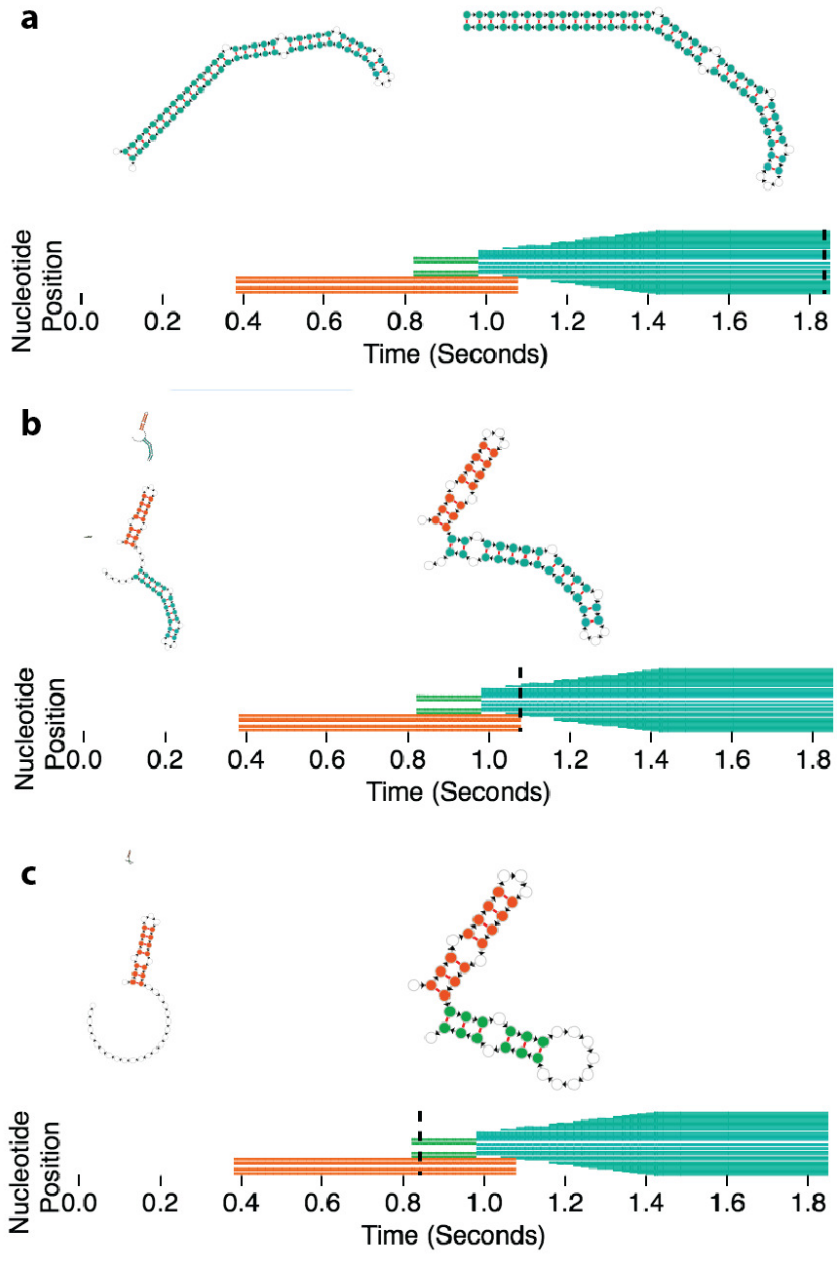
Which new structures should be added after an elongation step?

- 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



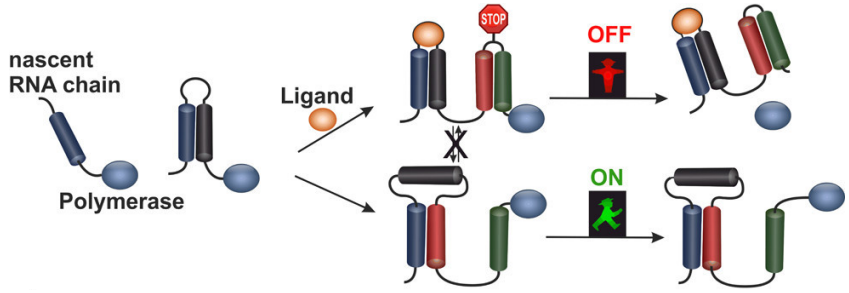
DrTransformer Visualization

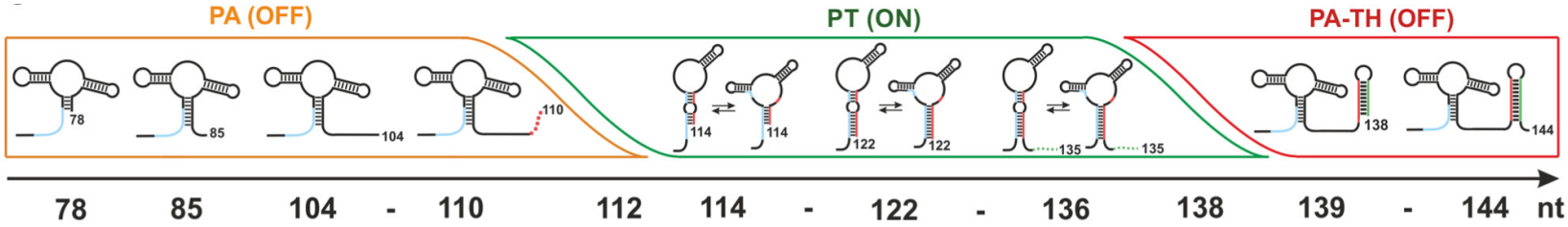
- Simple webinterface
- Interactive visualization Javascript and SVG
- Structure ensemble as function of time



Example: The dG-Riboswitch

- Aptamer for 2'deoxyguanosin
- Binding leads to transcription termination
- NMR analysis (Schwalbe lab): Ground state structure contains terminator even without ligand

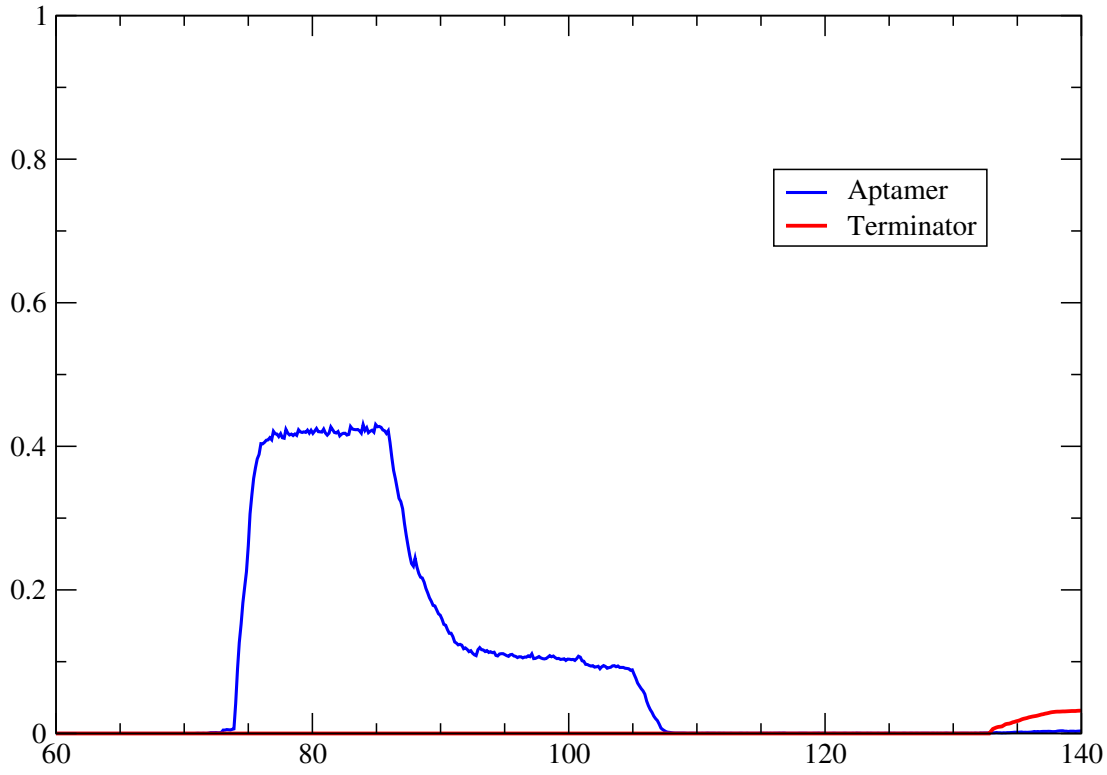




Helmling et al, JACS (2017)

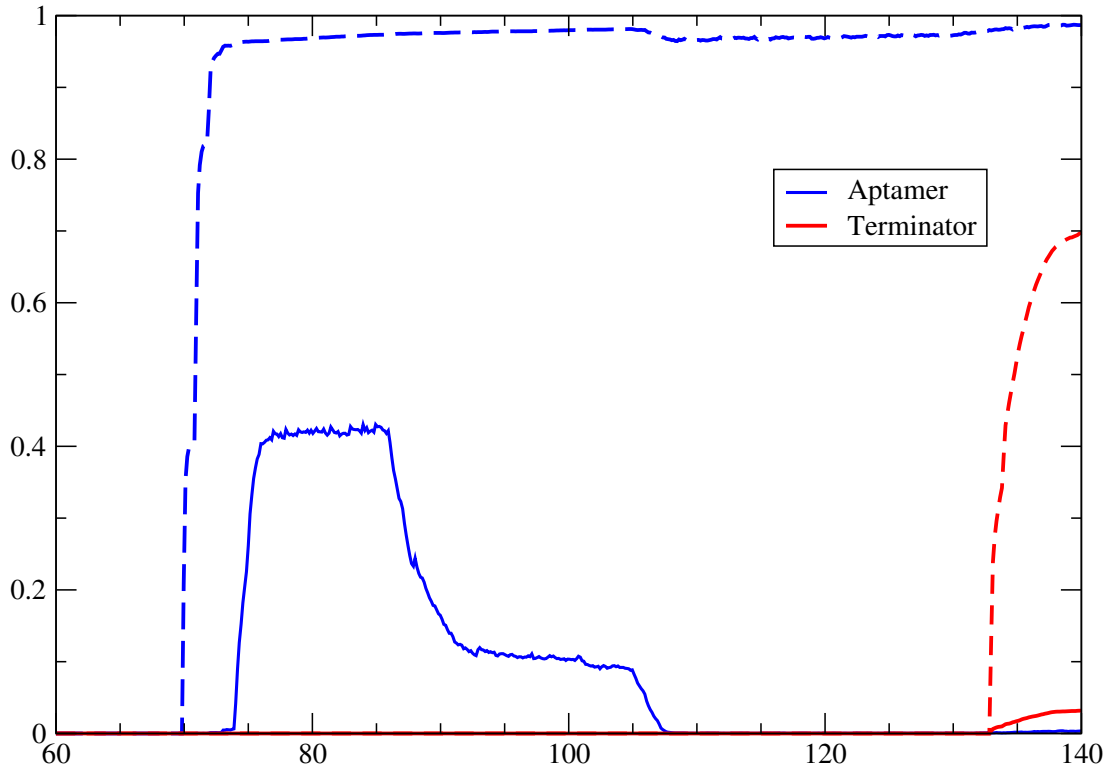
Kinfold simulation of the dG Riboswitch

- 10000 Kinfold trajectories (186 cpu hours)
- Classify each structure as aptamer and/or terminator
- Simulation with ligand: Add a bonus of 8kcal/mol for each binding competent structure



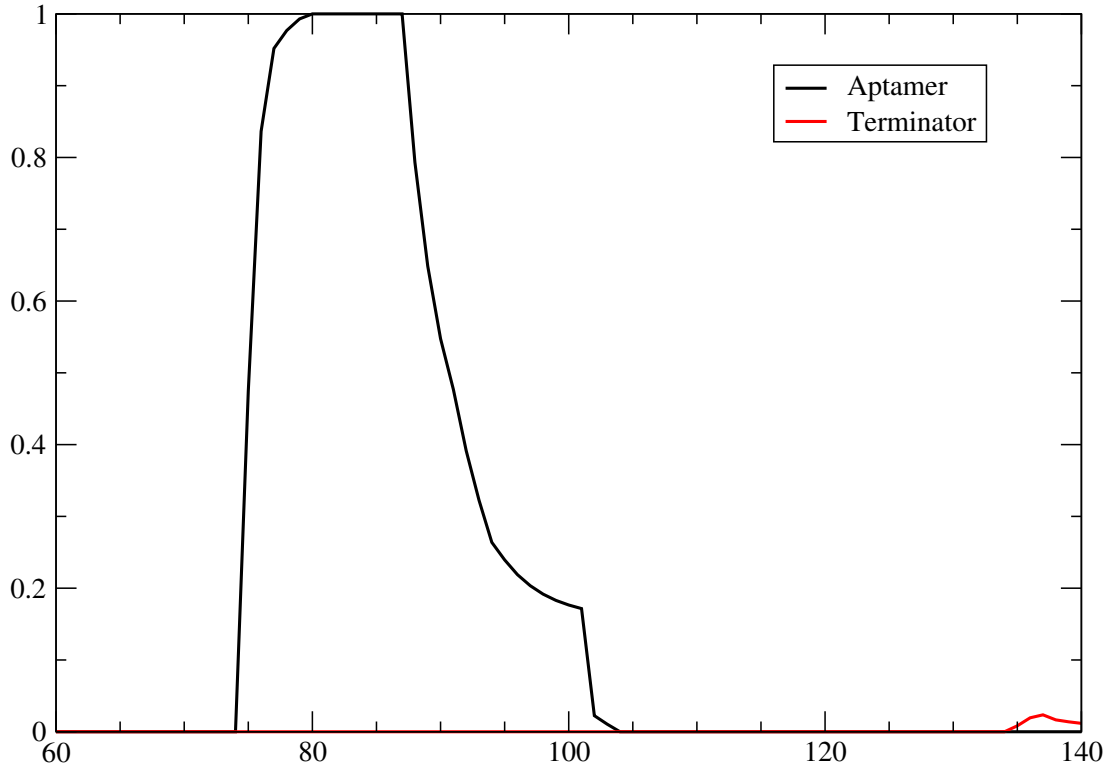
Kinfold simulation of the dG Riboswitch

- 10000 Kinfold trajectories (186 cpu hours)
- Classify each structure as aptamer and/or terminator
- Simulation with ligand: Add a bonus of 8kcal/mol for each binding competent structure

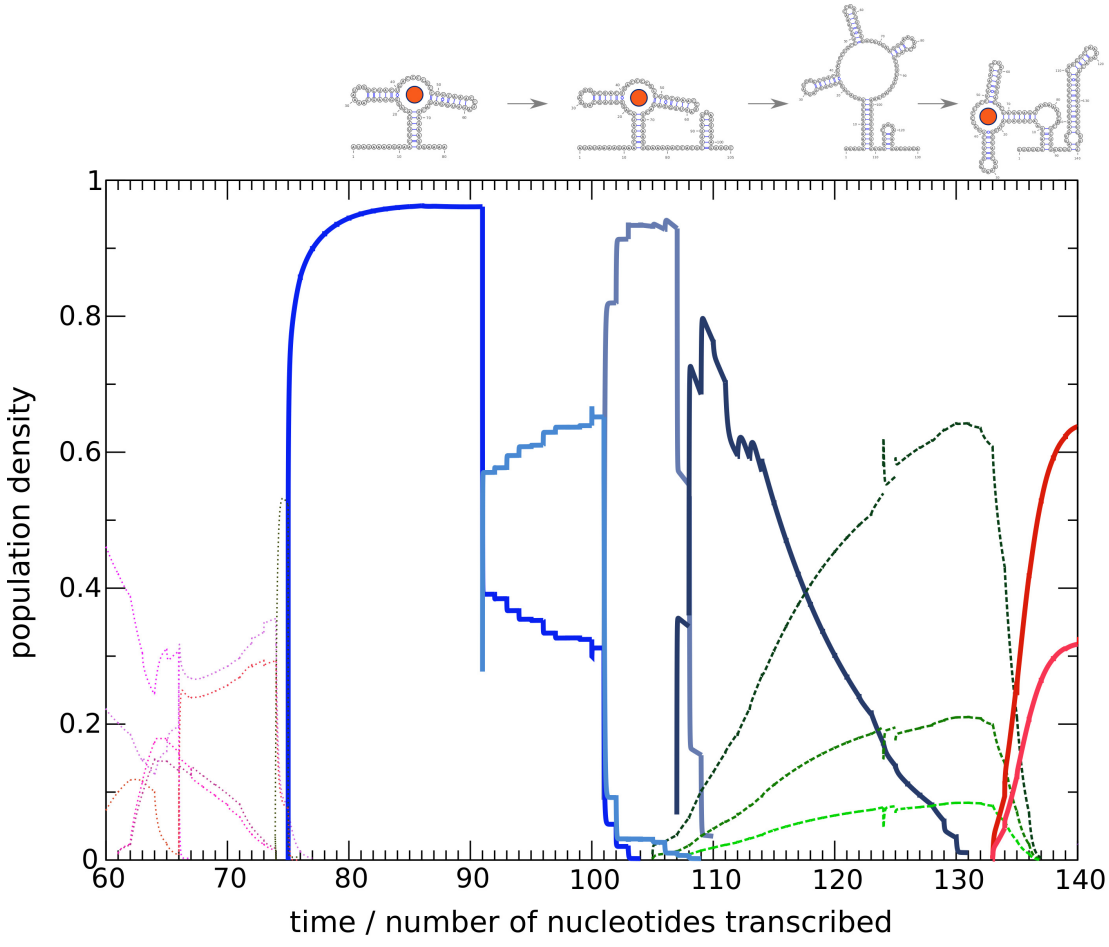


DrTrafo simulation of the dG Riboswitch

- Only 1 run needed (3 cpu sec)
- Classify each structure as aptamer and/or terminator
- Final state 1% population in terminator
- Simulation with ligand not yet possible



BarMap simulation of the dG Riboswitch



Simulation at 25C, transcription speed 25 nt/sec, ligand concentration of 1mM

Take home messages

- RNAs don't always reach their MFE or equilibrium state in reasonable time.
- Co-transcriptional folding essential to regulatory elements such as riboswitches
- Predicting kinetics is much harder than predicting equilibrium
- Previous methods too slow too cumbersome
- Faster, easy to interpret methods, now available

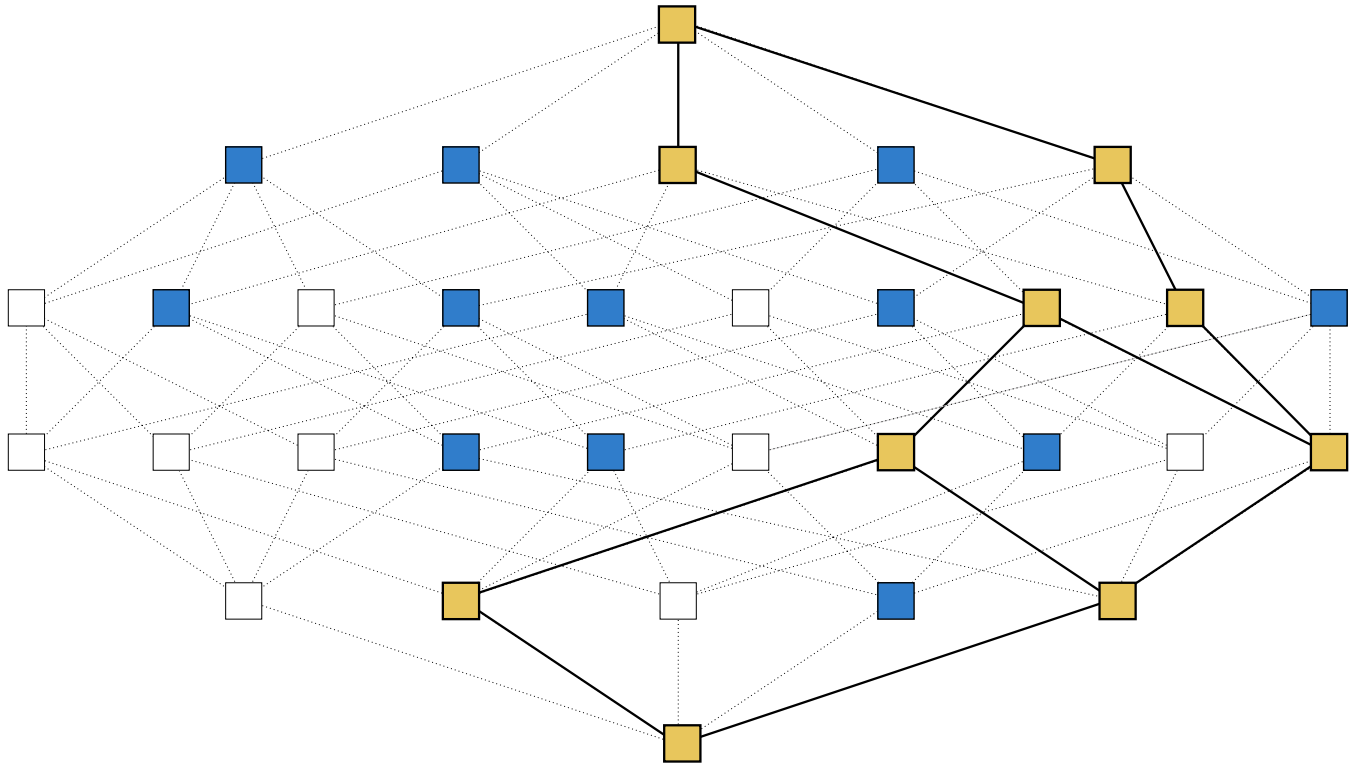
Acknowledgments

Christoph Flamm
Peter Stadler (Leipzig)
Ronny Lorenz (ViennaRNA)
Michael **Wolfinger** (`barriers`, `treekin`)
Marcel **Kucharik**, Jing Qin (now SDU Odense) (`BGH`)
Stefan **Badelt** (now CalTech, `DrTransformer`)
Peter Kerpedjiev (now Harvard, DrTransformer visualization)
Stefan **Hammer** (visualization)

H. Schwalbe, B. Fürtig, Ch. Helmling (Frankfurt, (dG-ribsowitch)

The findpath re-folding path heuristic

Perform a bounded breadth first search of direct paths.



- Only consider **direct** paths, i.e. where distance decreases with each step.
- Up to $D(x, y)!$ direct paths.
- Bound the search by keeping only m best candidates from each distance class.