Folding Dynamics of RNA Secondary Structures

Ivo L. Hofacker

Institute for Theoretical Chemistry, University Vienna

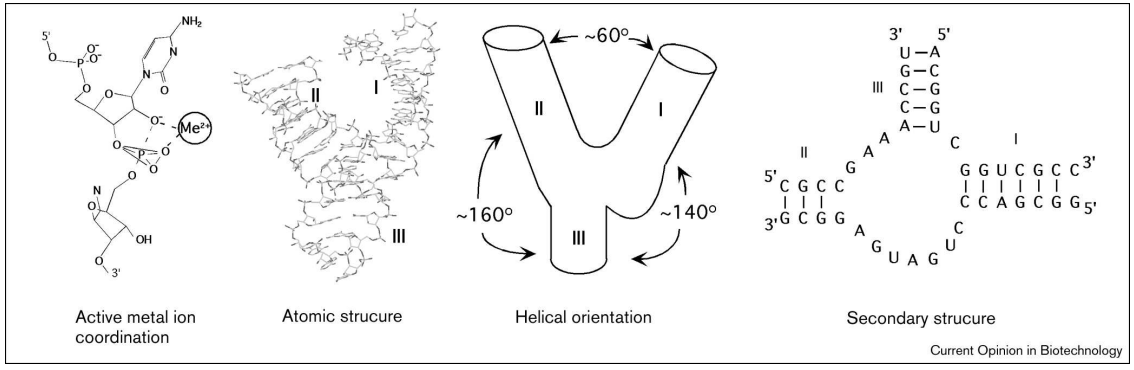
http://www.tbi.univie.ac.at/~ivo/

Cambridge, July 2005

The RNA secondary structure model

RNA secondary structure provide an ideal model to study biopolymer folding

- provide a biochemically useful structure description
- mathematically and computationally easy to handle
- energy model based on carefully measured parameters
- efficient algorithms for structure prediction



Computing RNA secondary structures

Most *equilibrium* properties can be computed exactly and efficiently by dynamic programming

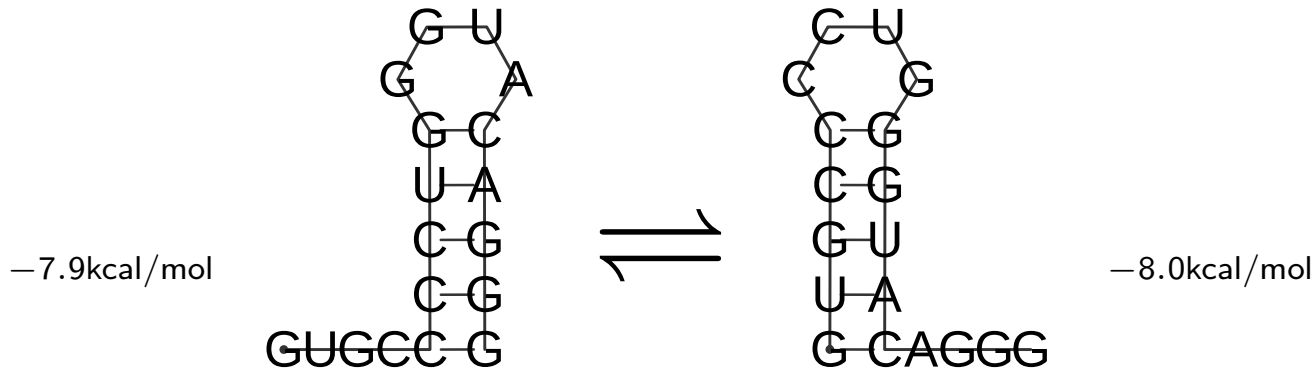
- Minimum free energy structure (Zuker & Stiegler '81)
- Suboptimal structures
  - representative suboptimal structures (Zuker '89)
  - all structures within an energy range (Wuchty et.al. '99)
  - Boltzmann weighted samples (Ding & Lawrence '03)
- Partition function and base pair probabilities (McCaskill '90)
- Density of states (Cupal '96)
- Minimum free energy with pseudoknots (Rivas & Eddy '99)

Free software available in the `Vienna RNA Package` at
`http://www.tbi.univie.ac.at/~ivo/RNA/`

Thermodynamic vs. Kinetic Folding

Equilibrium properties can be calculated efficiently
But what about dynamics?

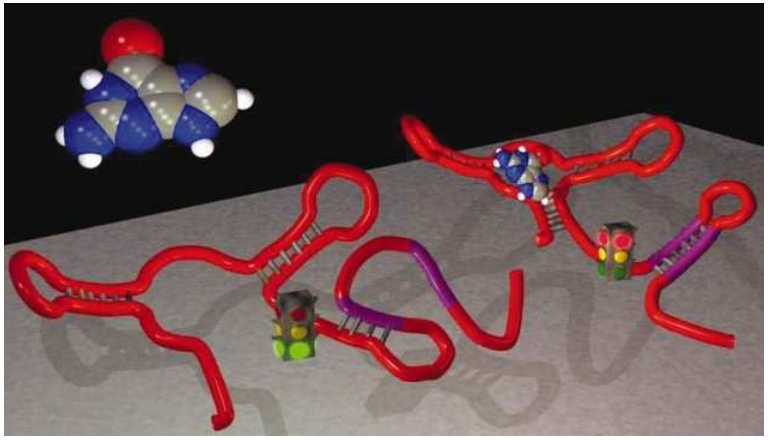
- On what time scale is equilibrium reached?
- How fast/slow is re-folding between dissimilar structures?
- What structures are populated initially?

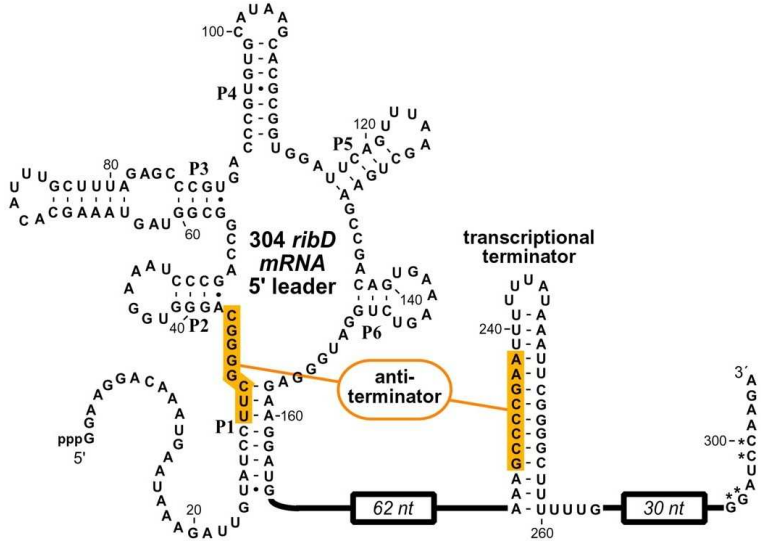


Structural changes are common in functional RNA

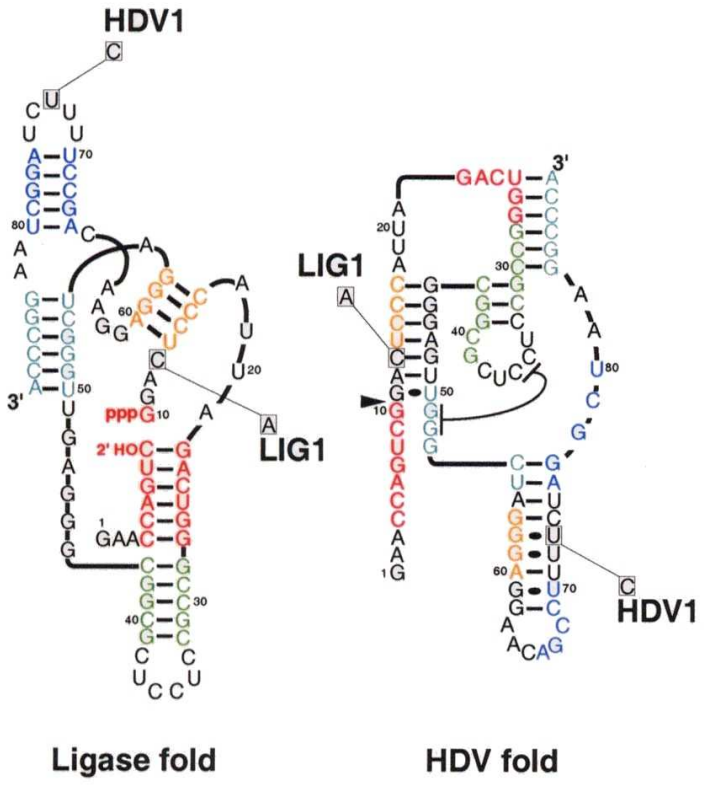
RNA switches toggle between active and inactive states by changing conformation.
Used especially to control mRNA translations; triggered by:

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

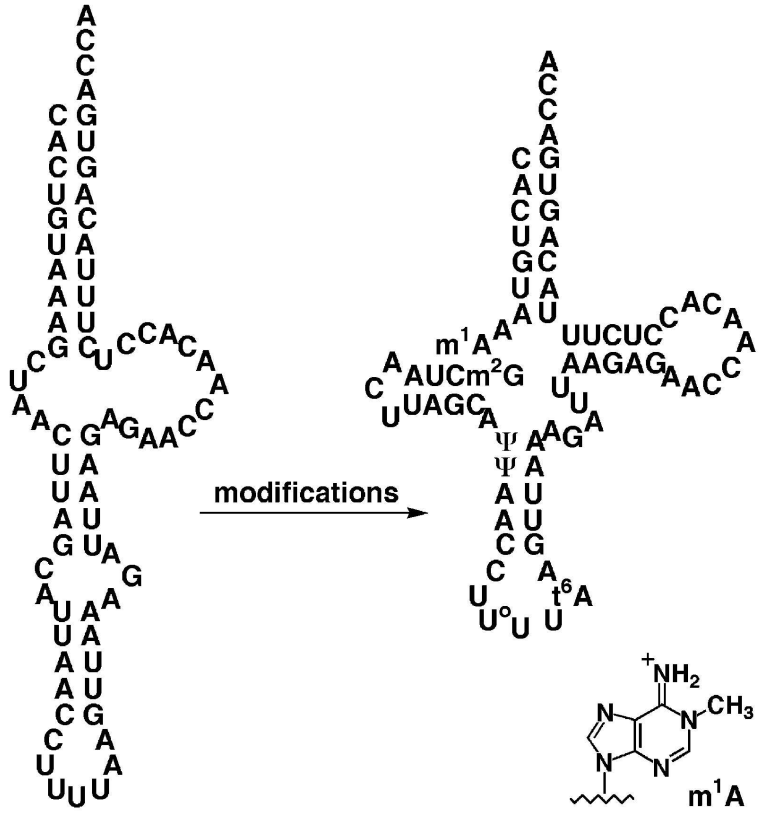




Examples of RNA switches



A Ribozyme with two functions (Schultes & Bartel 2000)



Chemical modification triggers the cloverleaf fold of a tRNA (Helm & Giegé 1999)

Predicting dynamics of RNA folding

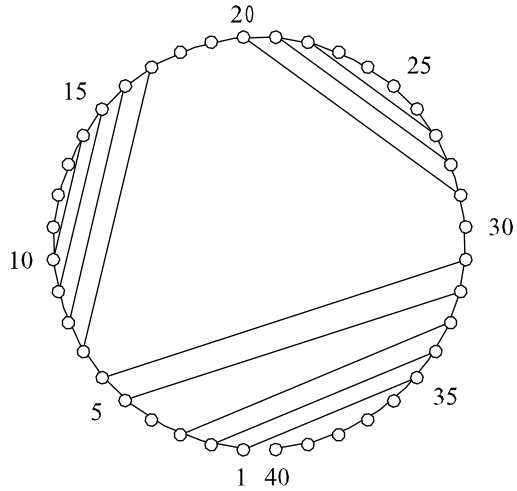
Folding dynamics described by a Morkov process with master equation

$$\frac{\mathrm{d}p_x}{\mathrm{d}t} = \sum_{y \in X} r_{xy} p_y(t)\,, \qquad \text{with } r_{xx} = -\sum_{y \neq x} r_{yx}.$$

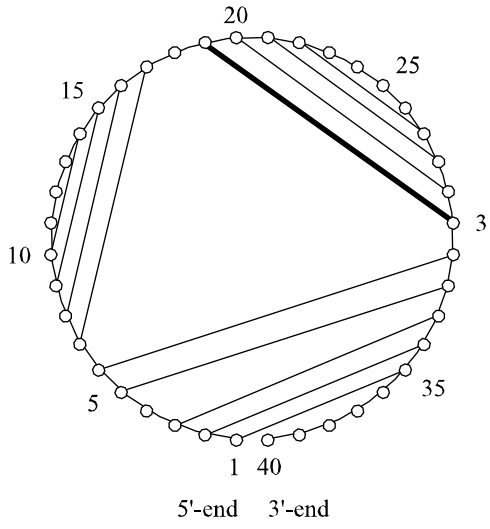
- Integration of the master equation (toy models only).
- Stochastic folding simulations. Needs many trajectories.
- Qualitative analysis of the energy landscape to identify possible traps (local minima). $\longrightarrow$ coarse grained versions of the Markov process

Need to model the rate r_{xy}. For small moves Metropolis rule is sufficient.

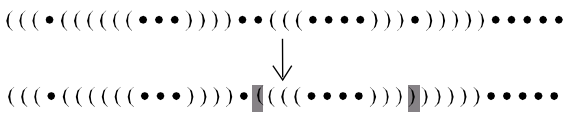
Elementary move set for RNA secondary structures



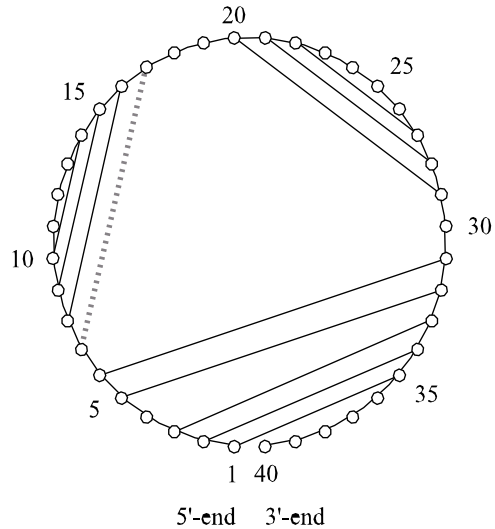
Add base pair

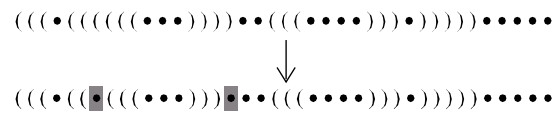


(((•((((((•••))))))••(((••••)))•)))))•••••



Remove base pair





Kinetic Folding Algorithm

Simulate folding kinetics by a Monte-Carlo type algorithm:

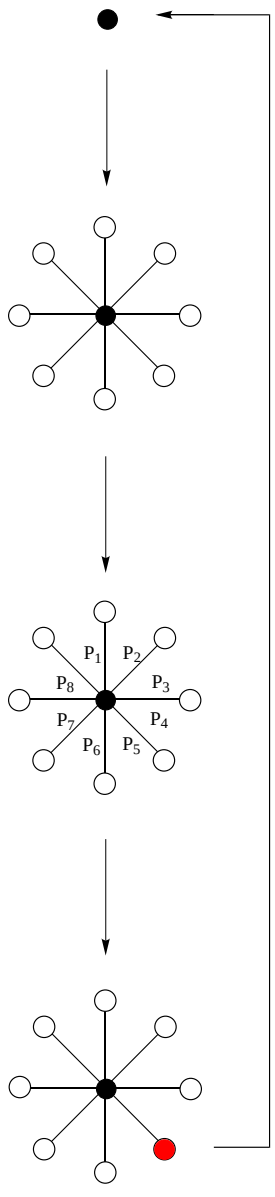
Generate all neighbors using the move-set

- Basepair Insertion
- Basepair Deletion

Assign rates to each move, e.g.

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

Advance clock $1/\sum_i P_i$.
select a move with probability proportional to its rate



Characterization of Landscapes

A landscape consists of a configuration space V, a move set within that configuration space and an energy function $f : V \to \mathbb{R}$.

Simplest move set for secondary structures: opening and closing of pairs.

Speed of optimization depends on the *roughness* of the Landscape.
Measures of roughness suggested in the literature:

- Number of local optima
- Correlation lengths (e.g. along a random walk)
- Lengths of adaptive walks
- Folding temperature vs. glass temperature T_f / T_g
- Energy barriers between the local optima. Especially, the maximum barrier height ("depth" in SA literature)

Energy barriers

$$E[s,w] = \min\left\{ \max\left[f(z) \middle| z \in \mathbf{p} \right] \;\middle|\; \mathbf{p} : \text{path from } s \text{ to } w \right\},$$

$$B(s) = \min\left\{ E[s,w] - f(s) \middle| w : f(w) < f(s) \right\}$$

Depth and Difficulty
(borrowed from simulated annealing theory)

$$\mathrm{D} = \max\left\{ B(s) \middle| s \text{ is not a global minimum } \right\}$$

$$\psi = \max\left\{ \frac{B(s)}{f(s) - f(\min)} \middle| s \text{ is not a global minimum} \right\}$$

Calculating barrier trees

The flooding algorithm:

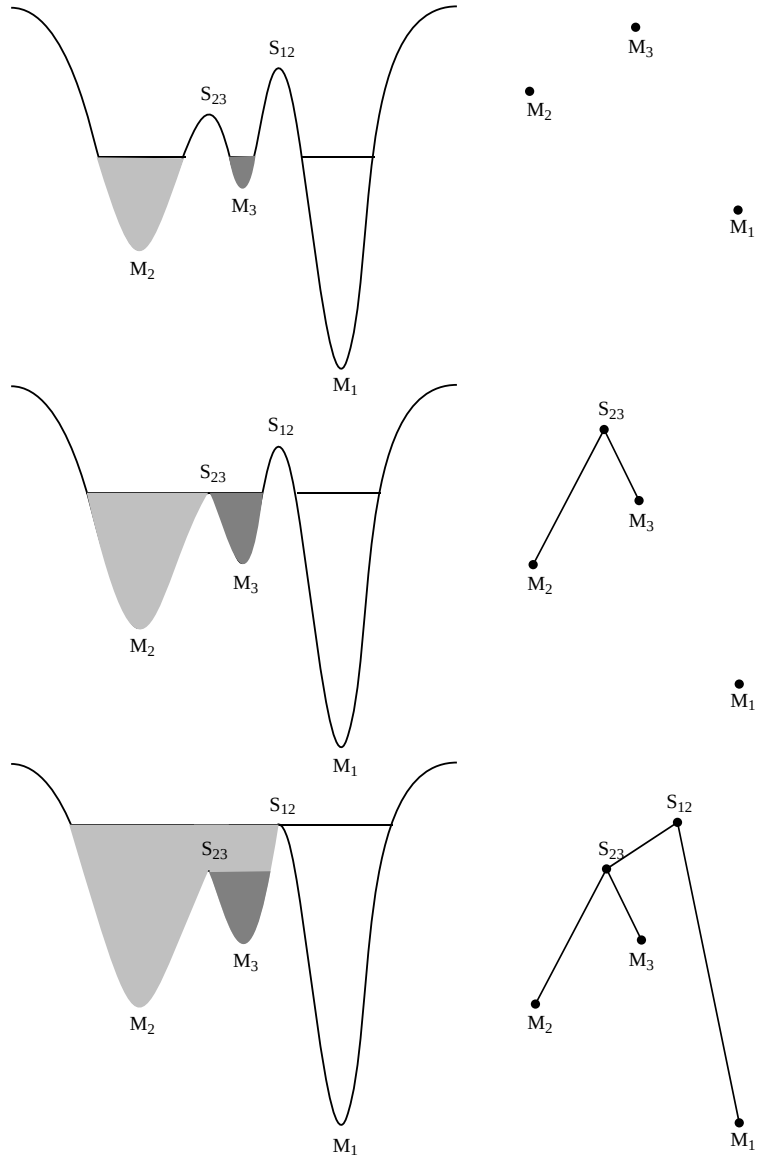
Read conformations in energy sorted order.

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. Basins $B(s_1), \ldots, B(s_k)$ are then united and are assigned to the deepest of local minimum.

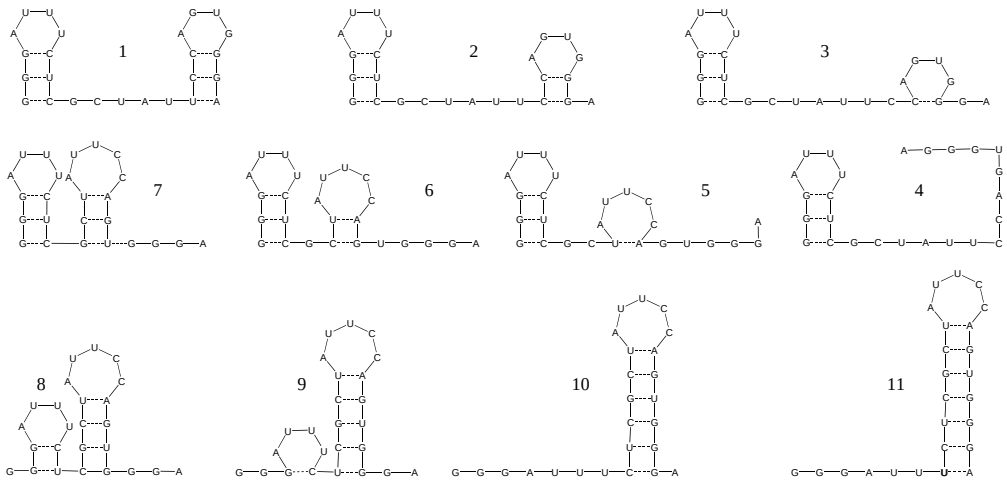


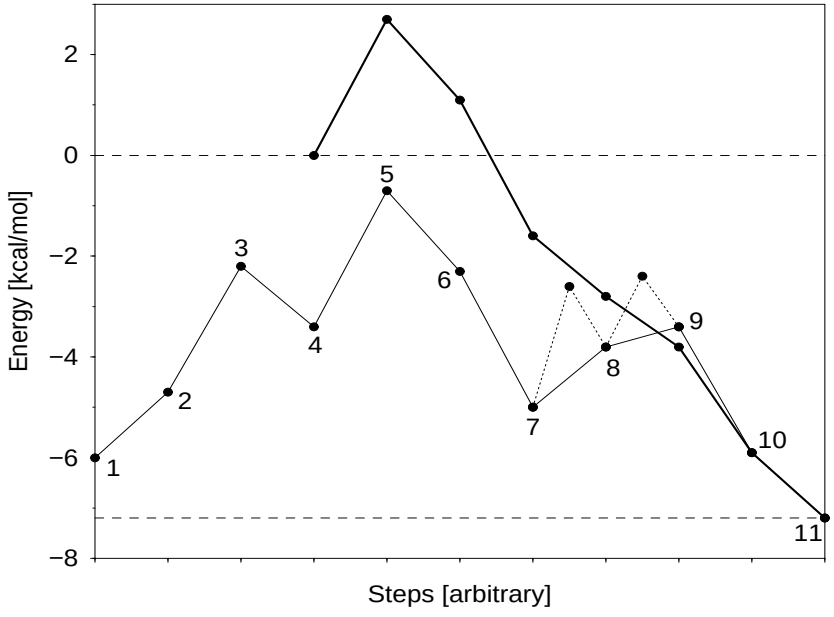
Information from the Barrier Trees

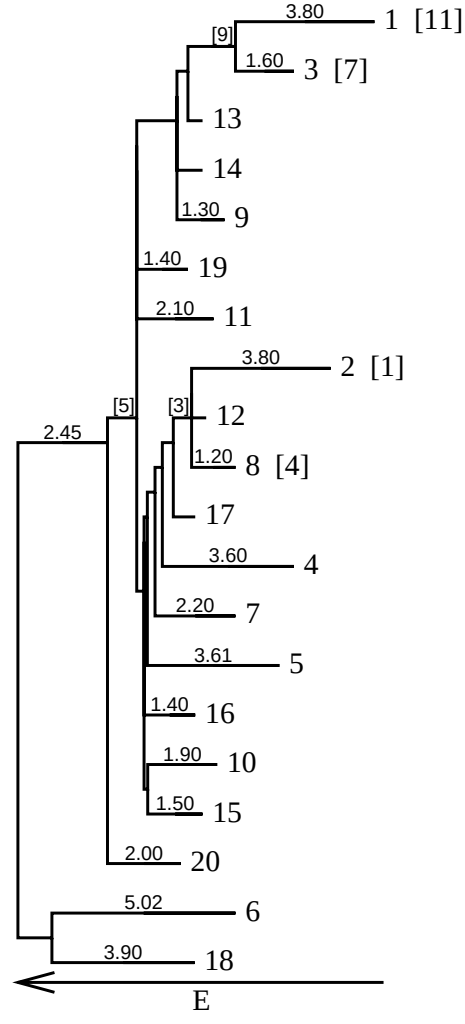
- Local minima
- Saddle points
- Barrier heights
- Gradient basins
- Partition functions and free energies of (gradient) basins
- Effective refolding rates between gradient basins
- Optimal refolding paths
- Depth and Difficulty of the landscape

A *gradient basin* is the set of all initial points from which a gradient walk (steepest descent) ends in the same local minimum.

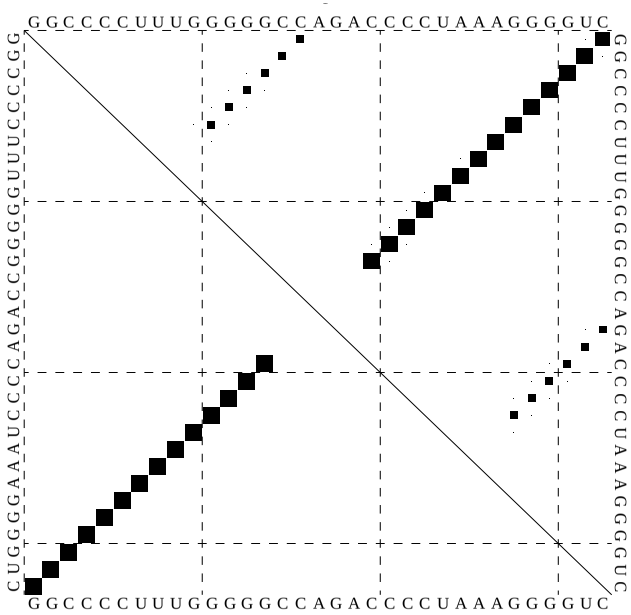
Energy Landscape of a Toy Sequence

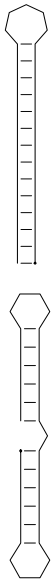






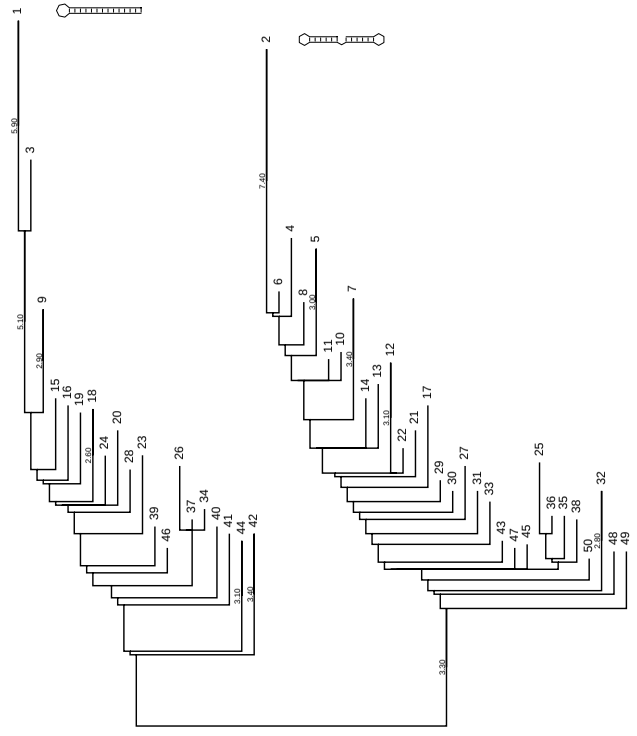
A Designed Bi-stable Sequence



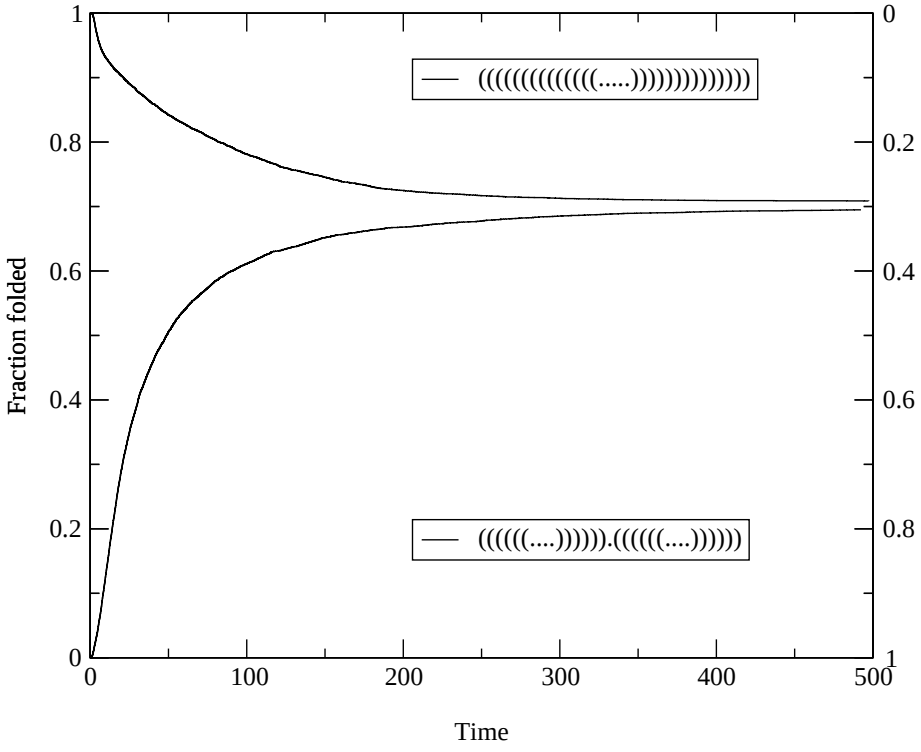


-23.8 kcal

-23.0 kcal



Barrier Tree and refolding Path




```
((((((....)))))).((((((....)))))) -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
.(....(((((...)))).((((((....)))))). -04.00
((....(((((...)))).((((((....)))))) -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
```

- The two component structure is kinetically prefered, because both hairpins act as nucleation centers
- For the full length chain 75% of trajectories reach the two component stucture first
- Much stronger effect for co-transcriptional folding:
  only 1 in 1000 trajectories ends in the one component structure

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

How can we optimally choose the macro-states?
Use the gradient basins around each local minimum.

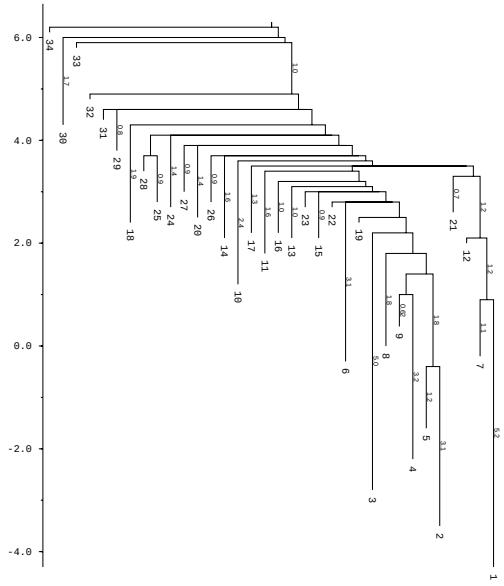
Transition rates could follow an Arrhenius rule
$r_{\beta\alpha} = \exp\left(-(E^*_{\beta\alpha} - G_\alpha)/RT\right)$.

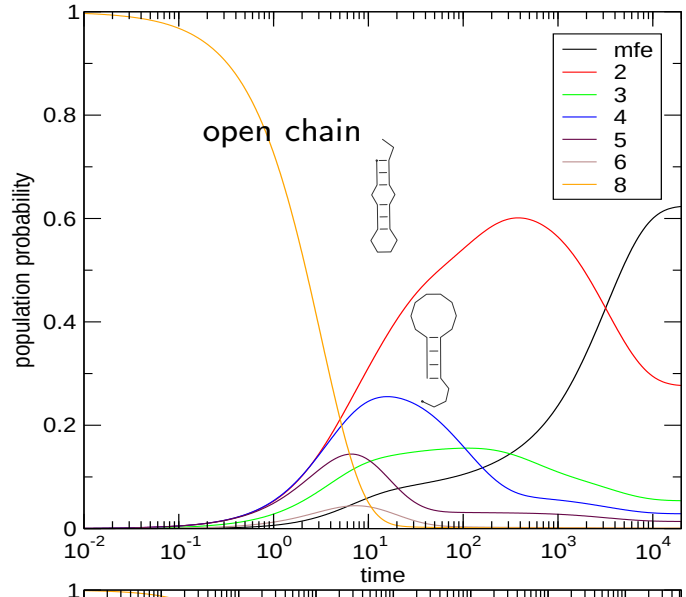
Or compute macro state rates from microscopic ones

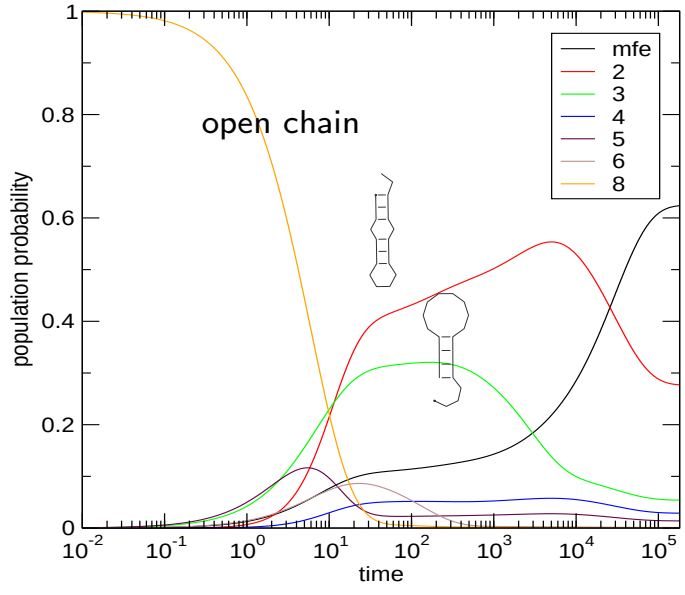
$$r_{\beta\alpha} = \sum_{y\in\beta}\sum_{x\in\alpha} r_{yx}\mathrm{Prob}[x|\alpha] = \frac{1}{Z_\alpha}\sum_{y\in\beta}\sum_{x\in\alpha} r_{yx}e^{-E(x)/RT}$$

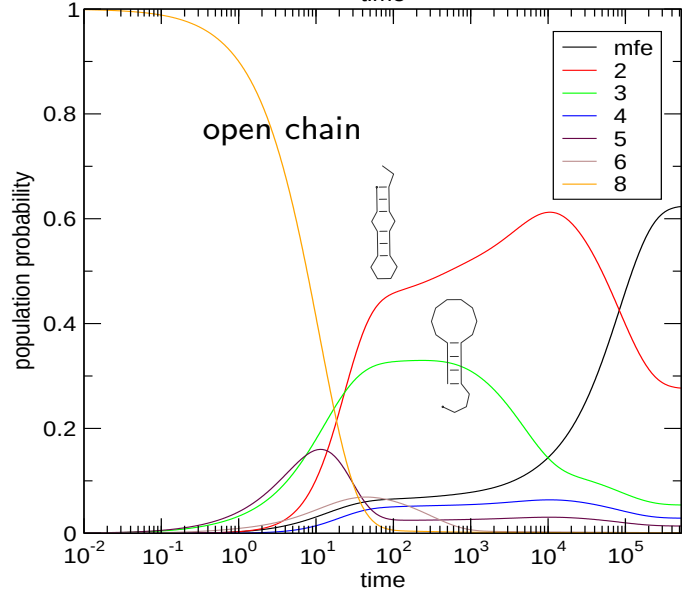
assuming local equilibrium.

Coarse grained dynamics vs. full dynamics









Folding during Transcription

- RNA is transcribed at a rate of only 30–40 nucleotides per second
- The nascent chain starts folding as soon as its leaves the ribosome
- Stem 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

Kinetic Folding Algorithm

Simulate folding kinetics by a Monte-Carlo type algorithm:

Generate all neighbors using the move-set

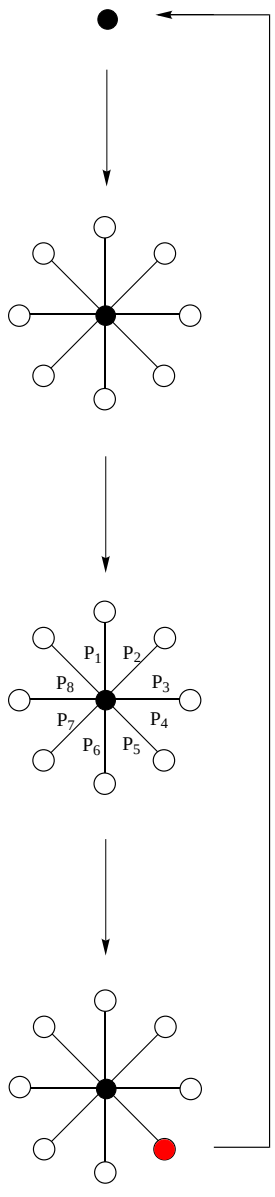
- Basepair Insertion
- Basepair Deletion

Assign rates to each move, e.g.

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

Advance clock $1/\sum_i P_i$.

select a move with probability proportional to its rate



Kinetic Folding Algorithm

Simulate folding kinetics by a Monte-Carlo type algorithm:

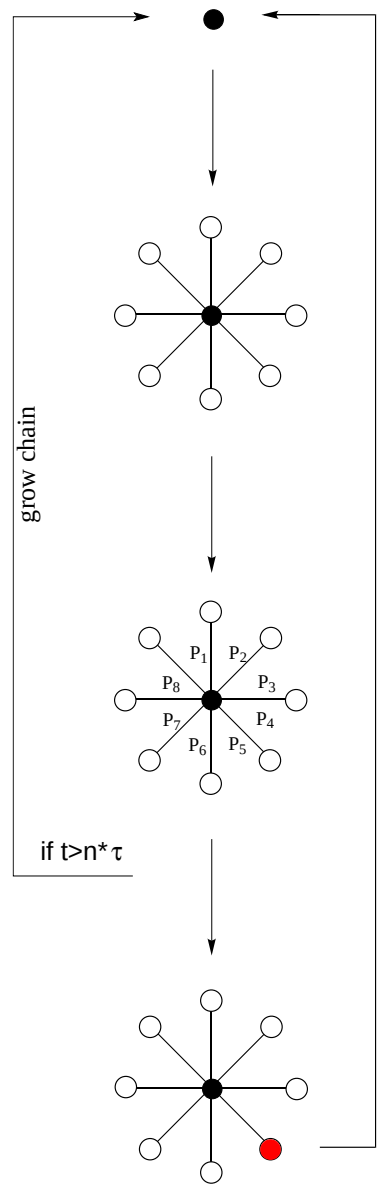
Generate all neighbors using the move-set

- Basepair Insertion
- Basepair Deletion

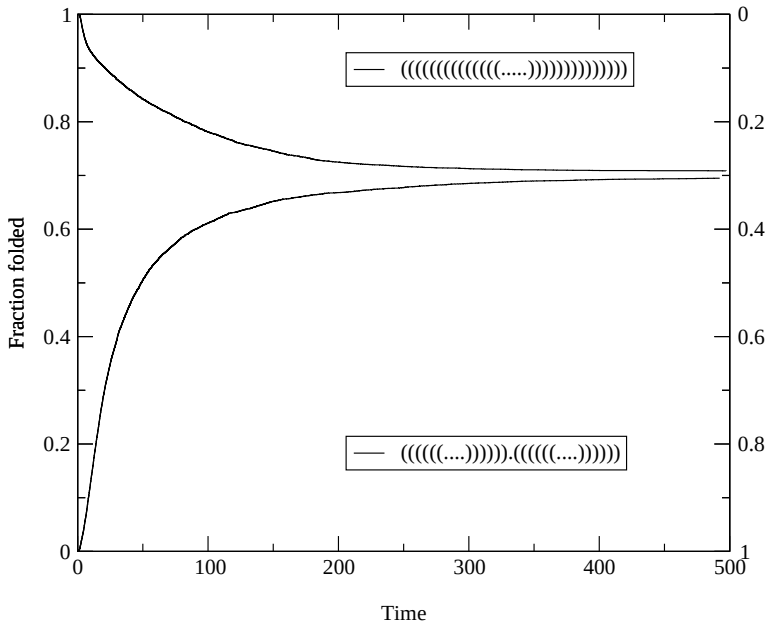
Assign rates to each move, e.g.

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

Advance clock $1/\sum_i P_i$.
extend chain by one if $t > n \cdot \tau$ else
select a move with probability proportional to its rate



Barrier Tree and refolding Path




```
((((((....)))))).((((((....)))))) -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
.(....(((((...)))))((((....)))))). -04.00
((....(((((...)))))((((....)))))) -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
```

- The two component structure is kinetically prefered
- From the open chain 75% of trajectories end in the two component stucture
- Much stronger effect for co-transcriptional folding:
  only 1 in 1000 trajectories ends in the one component structure

Some Examples

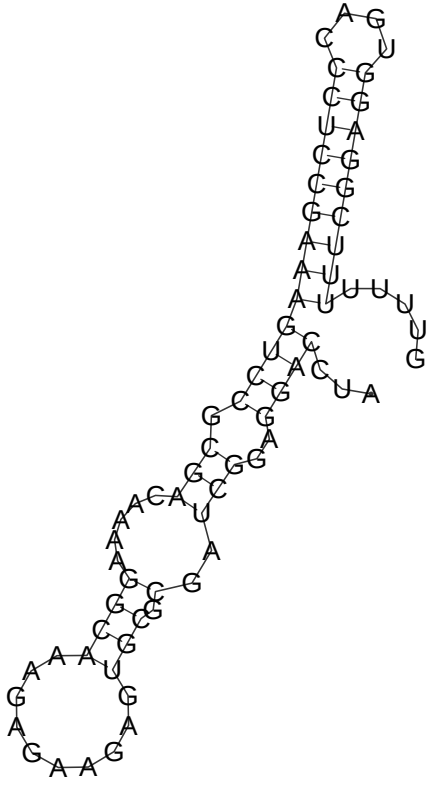
Effect of co-transcriptional folding for some bi-stable structures taken from the PARNAS web site.

name	full seq	slow	fast	very fast	equil.	maxB [1]
MS2	69/31	99.6/0.4	59/41	76/24	99.9/0.1	8.1
S15	60/40	99.7/0.3	99.5/0.5	60/40	99/1	6.24
dsrA	32/68	63/37	42/58	65/35	62/38	7.8
attenuator	85/15	99.9/0.1	25/75	69/31	94/6	13.7

With realistically slow transcription rate, co-transcriptional folding often leads to equilibrium.

[1] kcal/mol

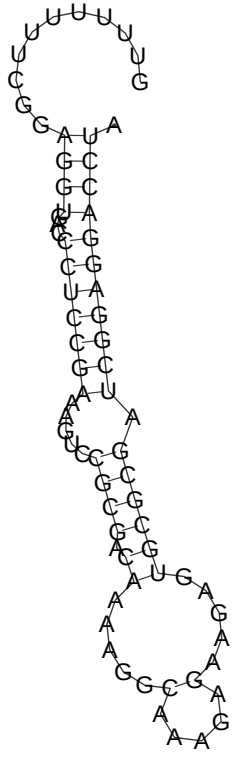
Attenuator example



−22.9kcal/mol

13.7kcal

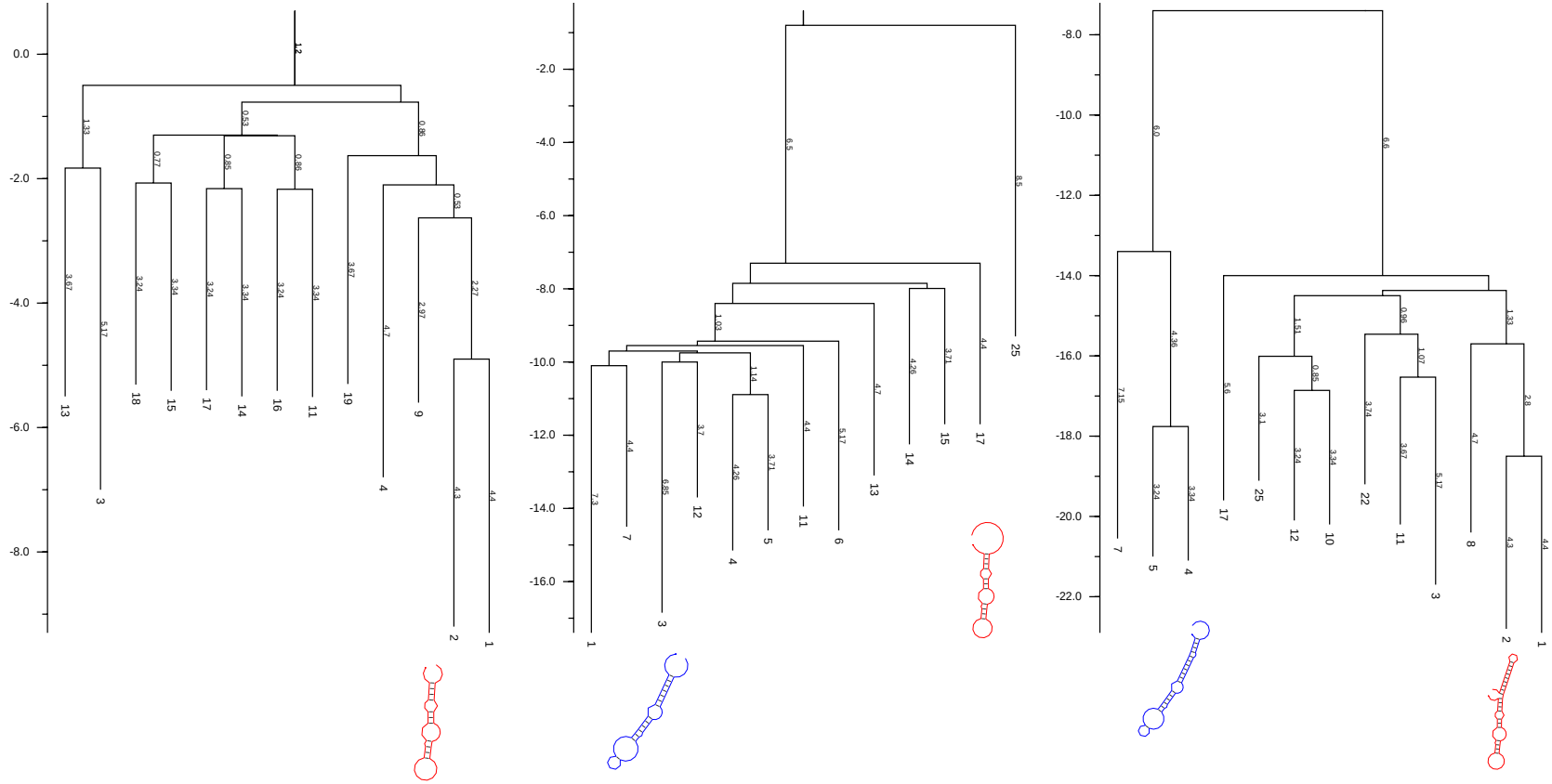
$\rightleftharpoons$



−21.1kcal/mol

Barrier Trees of Growing Sequence

$n = 50$ $n = 60$ $n = 70$



AUCCAGGAGGCUAGCGCGUGAGAAGAGAAACGGAAAACAGCGCCUGAAAGCCUCCCAGUGGAGGCUUUUUUU

...((((..(((..(.(((.........)))).....))).)))).... (-9.3)

.....((((((.((((((.....(.....).....)).))))).....)))))).... (-17.4)

...((((..(((..(.(((.........)))).....))).)))).............. (-9.3)

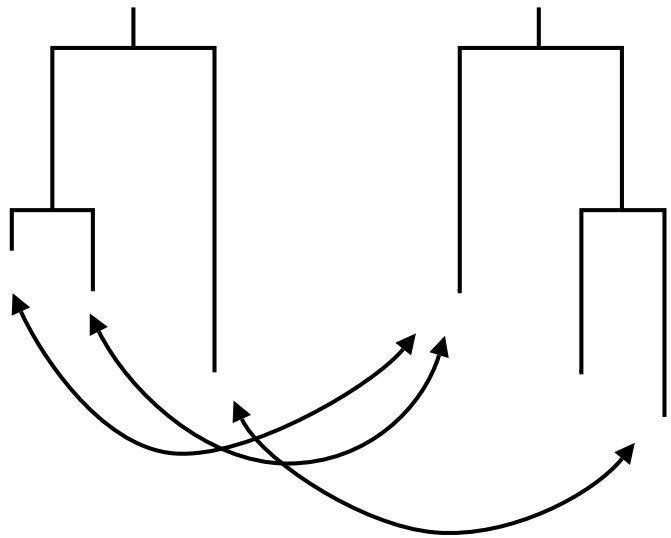
...((((..(((..(.(((.........)))).....))).))))((((((((...)))))))).. (-22.9)

.((((((((((.((((((.....(.....).....)).))))).....))))))...)))).... (-21.1)

Mapping between Barrier Trees

Each structure x at length n corresponds to an extended structure $x\bullet$ at length $n+1$.
For a minimum m, the correponding minimum m' can be found by a gradient walk starting with $m\bullet$.

- Two minima may be mapped to the same minimum in the $n+1$ landscape.
- In addition new minima may appear.



Mapping between Barrier Trees Example

`bar_map.pl` computes the mapping between a sequence of bar files

```
> bar_map.pl attenuator_*.bar
```

```
 44    46    48    50    52    54    56    58    60    62    64    65    66    67    68   70
  6 ->  1 ->  1 ->  1 ->  1 -> 17 -> 16 -> 25 -> 25 ~> 26 ~> 27 ->  7 ->  3 ->  1 ~>  1 ~> 1
  7 ~>  2 ~>  2 ~>  2 ~>  2 ~> 18 ~> 18 ~> 25 -> 25 ~> 26 ~> 27 ->  7 ->  3 ->  1 ~>  1 ~> 1
    ->  5 ->  4 ->  4 -> 12 ~> 18 ~> 18 ~> 25 -> 25 ~> 26 ~> 27 ->  7 ->  3 ->  1 ~>  1 ~> 1
    ->  8 ->  5 ->  5 -> 16 ~> 18 ~> 18 ~> 25 -> 25 ~> 26 ~> 27 ->  7 ->  3 ->  1 ~>  1 ~> 1
        ....
                                                                  ->  8 ~>  4 ~>  2 ~>  2 ~> 2
  1 ->  7 ~> 14 -> 14 ~>  9 -> 20 -> 20 ~>  1 ->  1 ~>  1 ->  1 ->  1 ->  1 ->  3 ->  3 -> 4
  4 -> 14 ~> 14 -> 14 ~>  9 -> 20 -> 20 ~>  1 ->  1 ~>  1 ->  1 ->  1 ->  1 ->  3 ->  3 -> 4
    -> 19 -> 11 -> 11 ~>  9 -> 20 -> 20 ~>  1 ->  1 ~>  1 ->  1 ->  1 ->  1 ->  3 ->  3 -> 4
          -> 15 -> 15 ~>  9 -> 20 -> 20 ~>  1 ->  1 ~>  1 ->  1 ->  1 ->  1 ->  3 ->  3 -> 4
                     ->  3 ->  1 ->  1 ->  1 ->  1 ~>  1 ->  1 ->  1 ->  1 ->  3 ->  3 -> 4
                     ->  6 ->  4 ~>  1 ->  1 ->  1 ~>  1 ->  1 ->  1 ->  1 ->  3 ->  3 -> 4
                     ->  7 ->  5 ~>  1 ->  1 ->  1 ~>  1 ->  1 ->  1 ->  1 ->  3 ->  3 -> 4
        ....
  2 ~>  9 ~> 17 ~> 17 ~> 10 ~> 21 ~> 21 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  4 ~>  4 ~> 5
          -> 16 ~> 16 ~> 10 ~> 21 ~> 21 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  4 ~>  4 ~> 5
                     ->  4 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  4 ~>  4 ~> 5
                     ->  8 ~>  6 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  4 ~>  4 ~> 5
                     -> 21 -> 11 ->  6 ->  6 ->  6 ~>  5 -> 11 -> 13 -> 15 -> 19 ~>  4 ~> 5
                           -> 19 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  2 ~>  4 ~>  4 ~> 5
```

Coarse grained Simulation with Chain Growth

How to generalize the coarse grained simulations for co-transcriptional folding

1. Simulate folding on barrier tree of size n for time τ
2. map final population to size barrier tree of size $n + 1$
3. use mapped population as initial condition for next simulation

Not yet implemented...

Summary

- Folding dynamics can be simulated through either explicit MC simulation or coarse grained computation on the barrier tree.
- Both approaches can be generalized to co-transcriptional folding
- Co-transcriptional folding can focus the outcome on just one structure
- Results can depend strongly on transcription speed
- Need to fix our time-scale by comparison with experiment