From Sequence to 2D
Part II

Ivo Hofacker

Institute for Theoretical Chemistry, University of Vienna

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

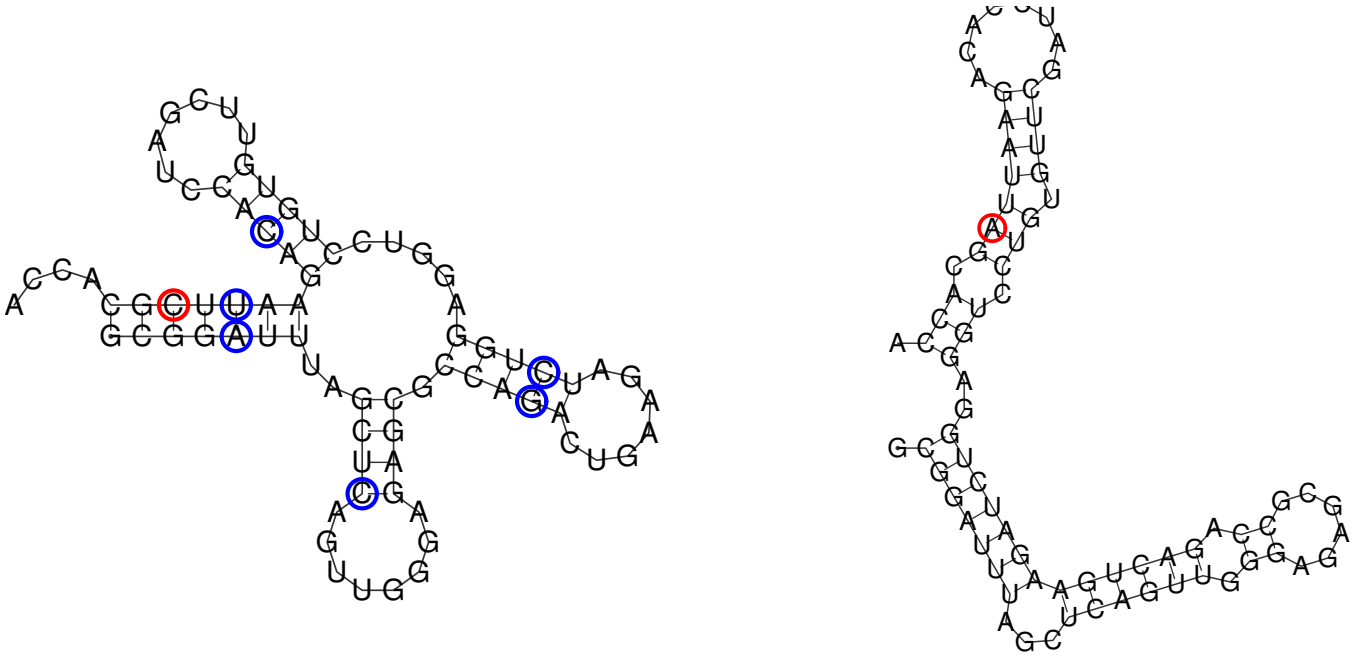
AlgoSB 2025
Marseille, December 2025

Can external Information help Structure Prediction?

Structure prediction from a single sequence is inaccurate.
How to improve this?

- Include experimental data (see tomorrow)
- Include more sequences!
- If several sequences form the same structure, prediction should get easier!

The Effect of Mutations



- Consistent and compensatory mutations often conserve the structure (blue)
- A single mutation (red) can radically change the structure
- Accumulating mutations quickly randomize any structure

Consensus Structure Prediction

- RNA families usually exhibit a well conserved consensus structure
- much stronger conserved than sequences
- if many sequences are known to have the same structure, base pairs can be deduced purely from sequence co-variations
- even a single homologous sequence can improve structure prediction
- conversely, conservation of structure implicates function

Strategies for Predicting Consensus Structures

Whenever possible, don't rely on a single sequence!

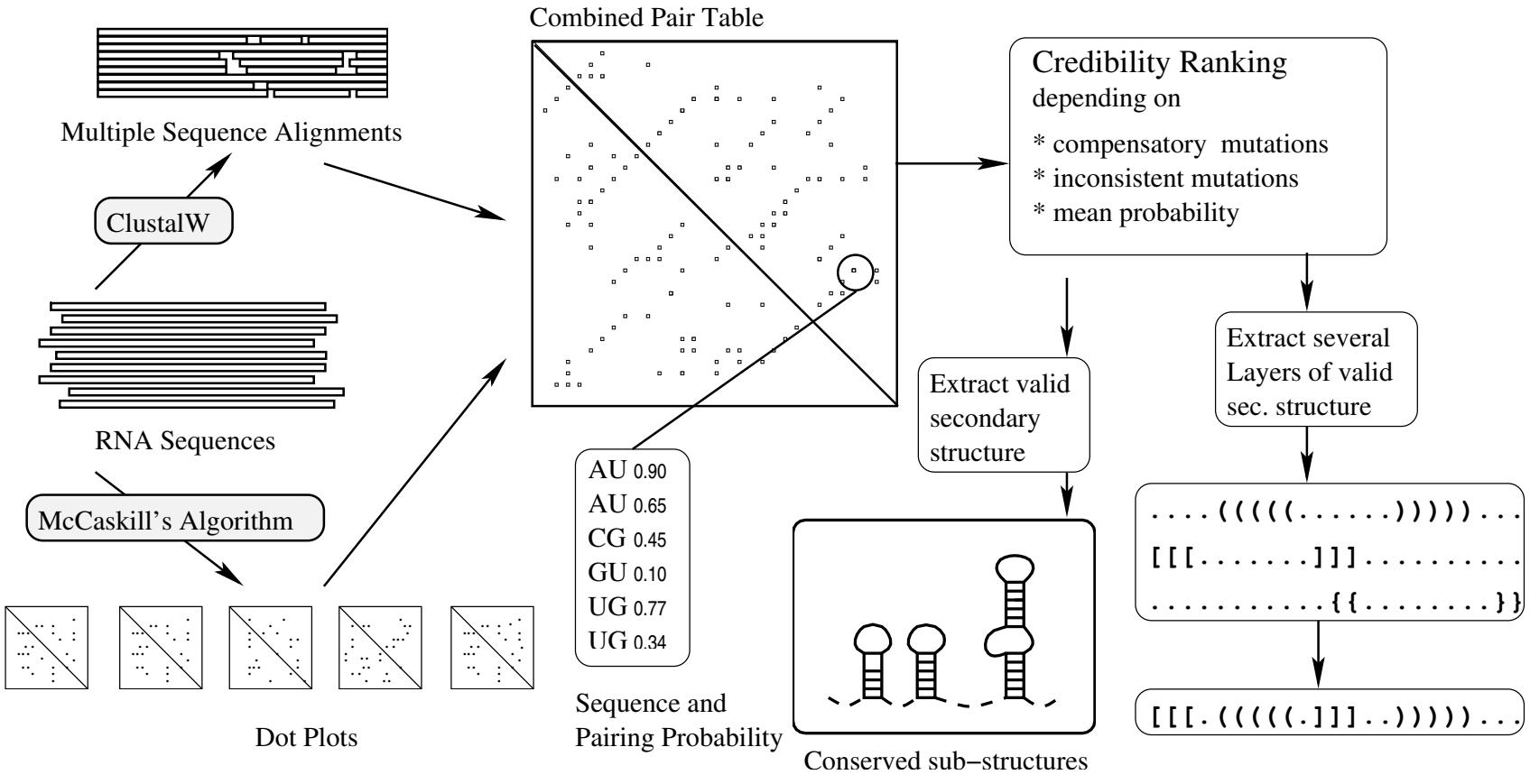
- Align Sequences, predict structure from alignment
  `RNAalifold`, `pfold`; `alidot`, `ConStruct`
  Sensitive to alignment errors
- Predict structures, then align structures
  `RNAforester`, `MARNA`
  Possibly sensitive to prediction errors
- Combine structure prediction and alignment
  The "Sankoff algorithm" `FoldAlign`, `DynAlign`, `stemloc`, `PMcomp`, `LocARNA`
- Alignment-free: Predict near-optimal coarse grained structures
  look for shapes common to all sequences `RNAcast`

Searching for conserved structure motifs

The **Alidot** approach:

1. Predict secondary structures for each sequence individually
   Use a) predicted mfe structures b) pair probabilities c) locally optimal structures from RNA**L**fold.
2. Use a standard *multiple alignment package* (Clustal W).
   (Sequences should be similar enough to allow accurate alignment)
3. Combine structure prediction and sequence alignment to get a list of candidate base pairs for conserved structures.
4. Sort list by *credibility* using compensatory mutations, inconsistent mutations, and predicted probability as criteria.
5. Extract predicted secondary structure motifs.

Alidot Flowchart



Alignment Folding

Alternative to `alidot`: Combine covariance analysis and folding into one dynamic programming algorithm.

- computes optimal consensus structure for a given alignment.
- Use a average energy over all sequences :

$$E_c(A, \Psi) = \frac{1}{N} \sum_{s \in A} E(s, \Psi)$$

- Optimize average energy E_c over all sequences in alignment
- Usual variants, mfe, partition function, sampling ...
- Efficient: $\mathcal{O}(N \cdot n^2 + n^3)$ CPU and $\mathcal{O}(n^2)$ memory, for alignment length n and N sequences.

Implemented in `RNAalifold` in the Vienna RNA Package.

Alignment Folding

Alternative to `alidot`: Combine covariance analysis and folding into one dynamic programming algorithm.

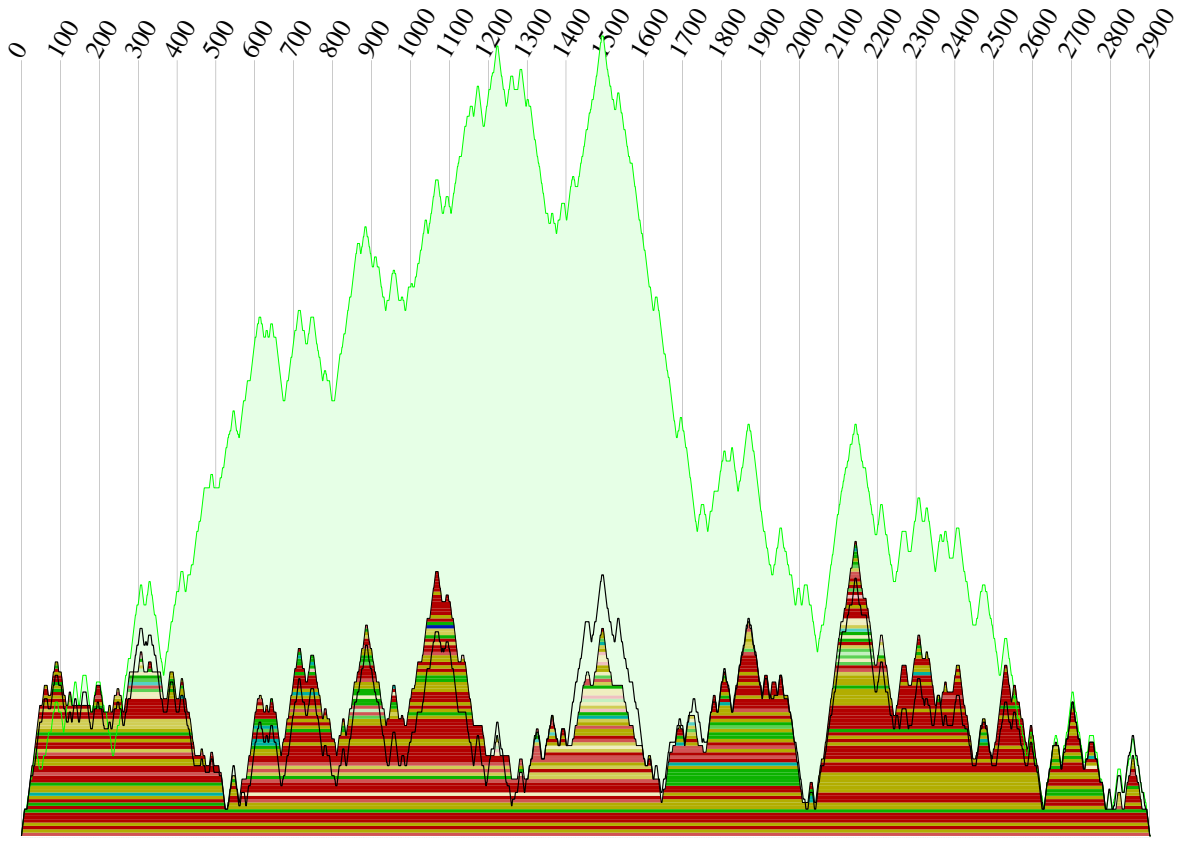
- computes optimal consensus structure for a given alignment.
- Use a average energy over all sequences plus a covariance score:

$$E_c(A, \Psi) = \frac{1}{N} \sum_{s \in A} E(s, \Psi) + cv \cdot \sum_{(i,j) \in \Psi} B_{ij}$$

- Optimize average energy E_c over all sequences in alignment
- Usual variants, mfe, partition function, sampling ...
- Efficient: $\mathcal{O}(N \cdot n^2 + n^3)$ CPU and $\mathcal{O}(n^2)$ memory, for alignment length n and N sequences.

Implemented in `RNAalifold` in the Vienna RNA Package.

Example: E.coli 23S RNA from 5 sequences



Accuracy for ribosomal RNAs

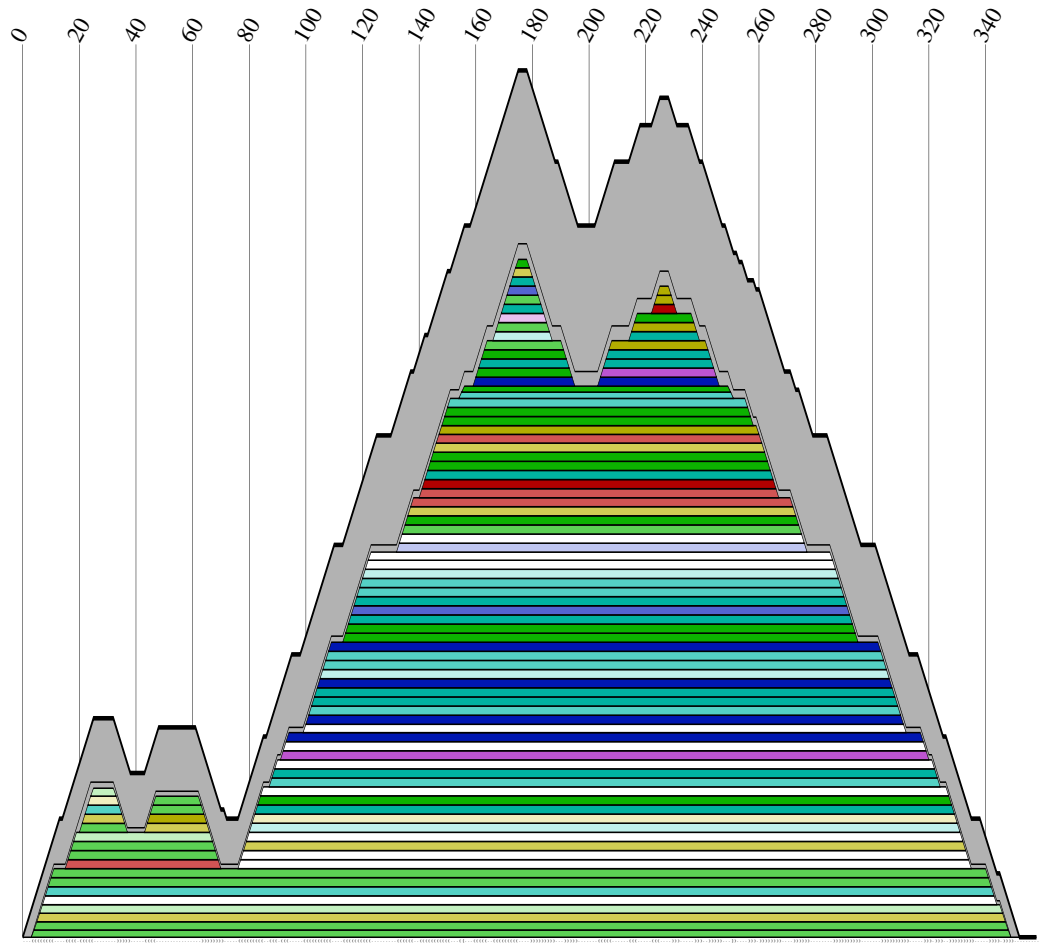
Correctly predicted base pairs 16S and 23S rRNA from E. Coli.

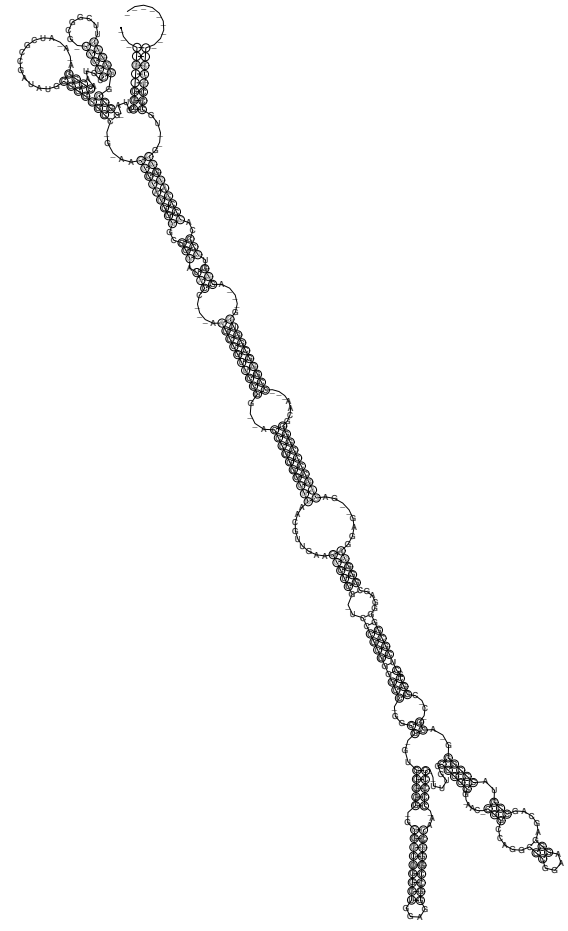
(relative to R. Gutell's structure)

	Clustal W		RDB		Clustal W		RDB	
N	raw	filled	raw	filled	raw	filled	raw	filled
	E.coli 16sRNA				23sRNA			
1	47.2	N/A	47.2	N/A	52.2	N/A	52.2	N/A
2	64.7	67.1	73.8	73.4	71.0	69.4	83.7	82.6
3	74.1	77.2	78.1	79.9	71.2	73.7	85.3	84.9
5	74.5	81.2	85.2	86.6	76.2	82.4	86.6	86.8
9	74.1	82.1	85.9	88.6	74.6	82.6	86.1	86.2

RDB Alignment: *Ribosomal Database Project* [Maidak et al., NAR (2000)]

Consensus structure of 14 SRP RNA





How to score covariances

Mutual Information

Entropy of a random variable X with probability distribution $p(x)$ is

$$H(X) = -\sum_x p(x) \log_2 p(x)$$

The mutual information of two distribution is given by

$$\begin{aligned} M(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

Obviously we have

$$M(X;Y) = M(Y;X) \text{ and } M(X;Y) \geq 0$$
$$\text{with } M(X;Y) = 0 \iff p(x,y) = p(x)p(y)$$

Mutual Information II

Easily computed directly from frequencies in column i and j of alignment:

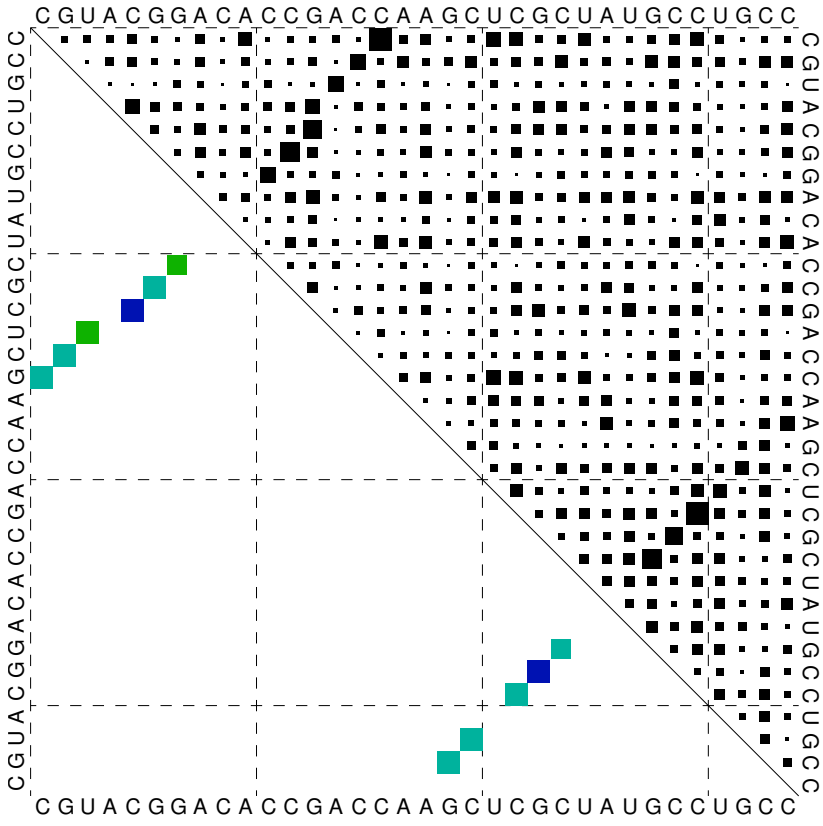
$$M_{i,j} = \sum_{x,y} f_{ij}(xy) \log_2 \frac{f_{ij}(xy)}{f_i(x) f_j(y)}$$

For the 4 letter alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$, $0 \leq M_{ij} \leq 2$ bits.

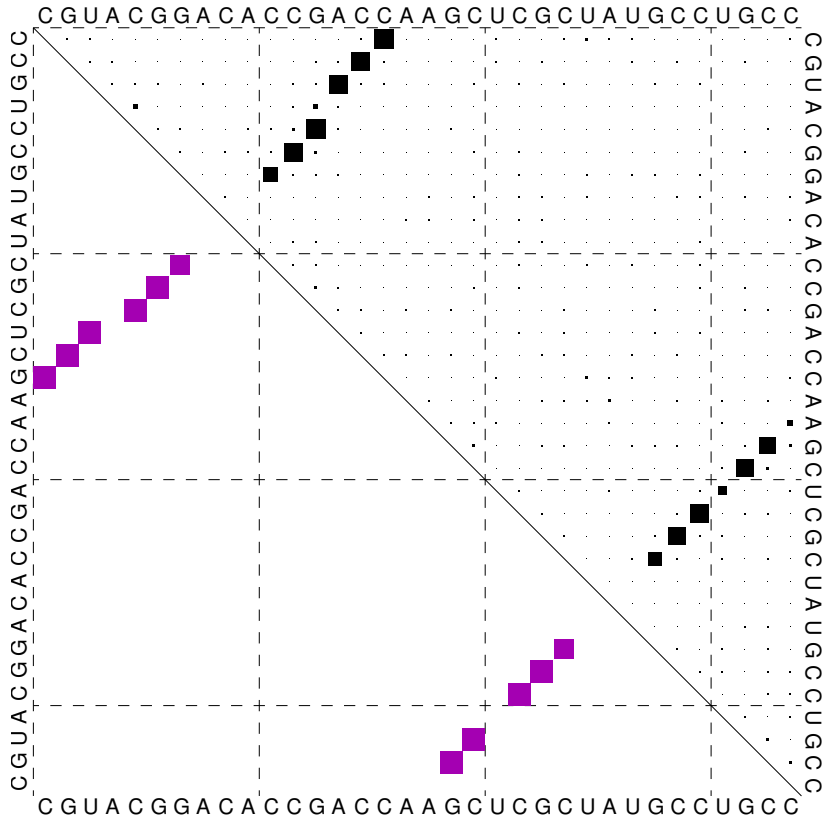
+ Completely parameter free
+ No model of sequence evolution or phylogenetic tree needed
± Uses no prior knowledge about secondary structures
+ can detect tertiary contacts and functional constraints
- poor signal to noise for small data sets
- only compensatory mutations contribute, consistent mutations (GC $\rightarrow$ GU) are neglected

Artificial Test Case

Generate sequences folding into the structure `(((.(((...)))))).((.(((...))).)).` using `RNAinverse`.



MI from 10 sequences



MI from 100 sequences

Alifold Covariance Score

Let $\Pi_{ij}^{\alpha} = 1$ if sequence α can pair positions i, j;
$d_{ij}^{\alpha,\beta}$ hamming distance of α and β at positions i and j (e.g. 0,1, or 2).

$$C_{ij} = \frac{1}{N(N-1)} \sum_{\alpha,\beta} d_{ij}^{\alpha,\beta} \Pi_{ij}^{\alpha} \Pi_{ij}^{\beta}$$

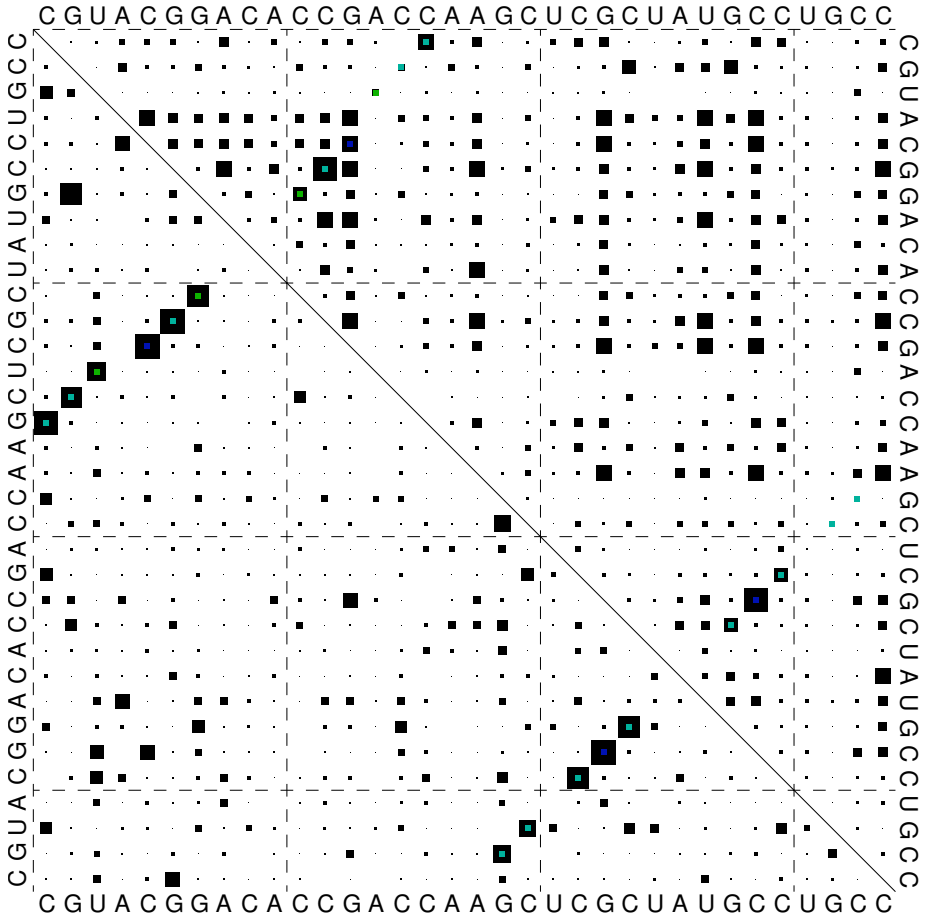
$$= \sum_{\mathrm{xy},\mathrm{x'y'}} f_{ij}(\mathrm{xy}) \mathbf{D}_{\mathrm{xy},\mathrm{x'y'}} f_{ij}(\mathrm{x'y'})$$

where $\mathbf{D}_{\mathrm{xy},\mathrm{x'y'}}$ contains $d_H(xy, x'y')$ if xy and $x'y'$ are allowed pairs , else 0.
Including a penalty for non-standard pairs set

$$B_{ij} = C_{ij} - \varphi \left(1 - \frac{1}{N} \sum_{\alpha} \Pi_{ij}^{\alpha}\right)$$

MI vs. Covariance score

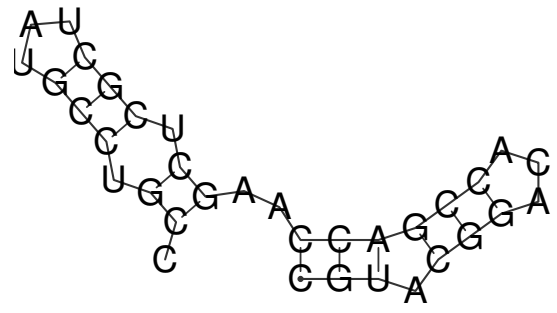
Comparing mutual information and covariance score



upper left: MI; lower right: covariance

Generate 5 sequences that all fold into the same structure using `RNAinverse`.

Compare mutual information (upper right) and covariance (lower left) score



(((.(((...)))))).((.(((...))).)).

Ribosum Scores

Classical substitution score from sequence alignment:

$$s(a,b) = \log \frac{f(a,b)}{f(a)f(b)}$$

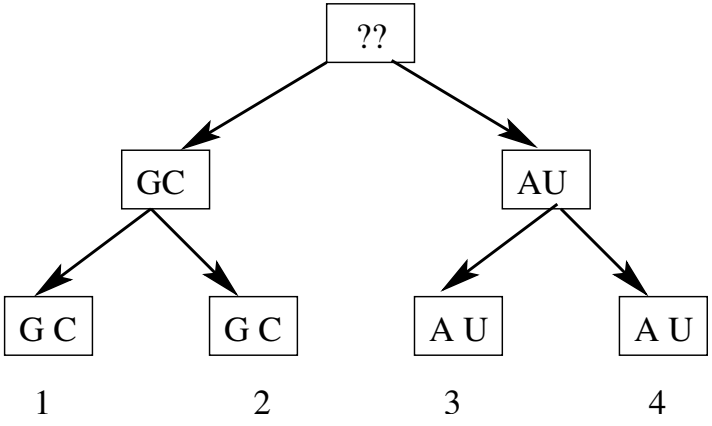
Equivalent scores for pairs of columns:

$$R(ab,cd) = \log \frac{f(ab;cd)}{f(ac)f(bd)}$$

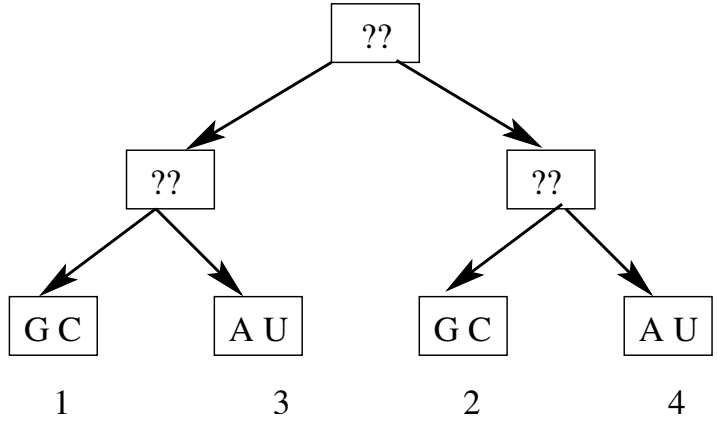
$f(ab;cd)$:	Probability of observing a pair (a,b) in one sequence and (c,d) in another – frequencies taken from rRNA structures
$f(ac)$:	Probability of observing a and c in one column

Scoring with Phylogenetic Tree

Same data pair frequencies may be produced by different histories



Single Mutation



Multiple Mutations

Given a tree T compute probability of data given two models for a) conserved pair b) independent positions. Use the log-odds score:

$$score = \log \frac{P(d|T, \wedge pair)}{P(d|T \wedge nopair)}$$

To calculate $P(d|model)$ need to sum over all possible histories. Luckily, this can be done recursively (DP).

Variations

Consensus structure vs. predicting one structure with help from other sequences

TurboFold (Harmanci et al 2011)

1. Probabilistic pairwise Alignment to compute match probabilities
2. Compute pair probabilies
3. Compute "extrinsic information" for pair (i, j) in Sequence m

$$\pi_{ij}^{m} = \alpha \sum_{s \neq m} (1 - ID(s, m)) p_{kl}^{s} P^{m,s}(i \sim k) P^{m,s}(j \sim l)$$

4. Compute pair probabilities with modified energy function

$$E(\Psi) = E^{0}(\Psi) - \gamma \sum_{(i,j) \in \Psi} \log(\pi_{ij})$$

5. Goto step 3

Pfold

Probabilistic SCFG based consensus structure prediction.
Compute the most probably structure σ given Alignment A, Phylogenetic tree T, and a model of Evolution M.

$$P(\sigma|A, T, M) \propto P(A|\sigma, T, M)P(\sigma)$$

$P(A|\sigma, T, M)$ can be computed as in maximum likelihood phylogeny reconstruction.
$P(\sigma)$ computed using a simple SCFG:

$$\begin{aligned} S &\rightarrow LS\,(86.9\%) \mid L\,(13.1\%) \\ L &\rightarrow s\,(89.5\%) \mid dFd\,(10.5\%) \\ F &\rightarrow dFd\,(78.8\%) \mid LS\,(21.2\%) \end{aligned}$$

Structural alignment

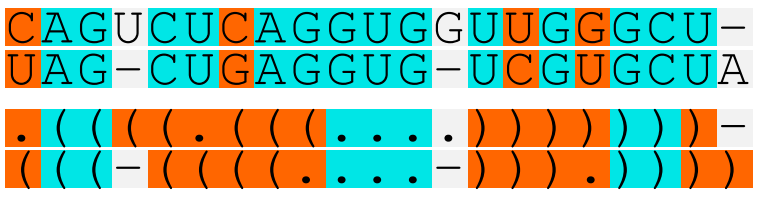
Sequence alignment is often not appropriate for structural RNAs

The "correct" sequence alignment need not be the correct structural alignment

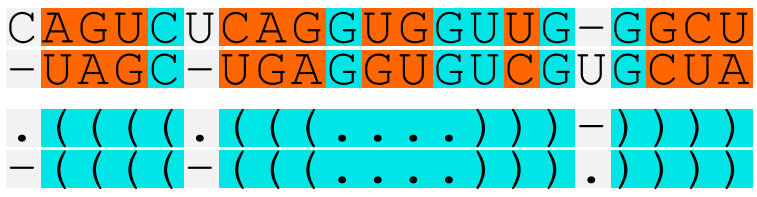
```
CAGUCUCAGGUGGUUGGGCU
.((((.(((....)))))))
```

```
UAGCUGAGGUGUCGUGCUA
((((((((....)))).))))
```

Sequence alignment



Structure alignment



Using pure sequence alignments is still the common approach.
Structural alignment important when sequence identity $< 60\%$

Conclusions (this part)

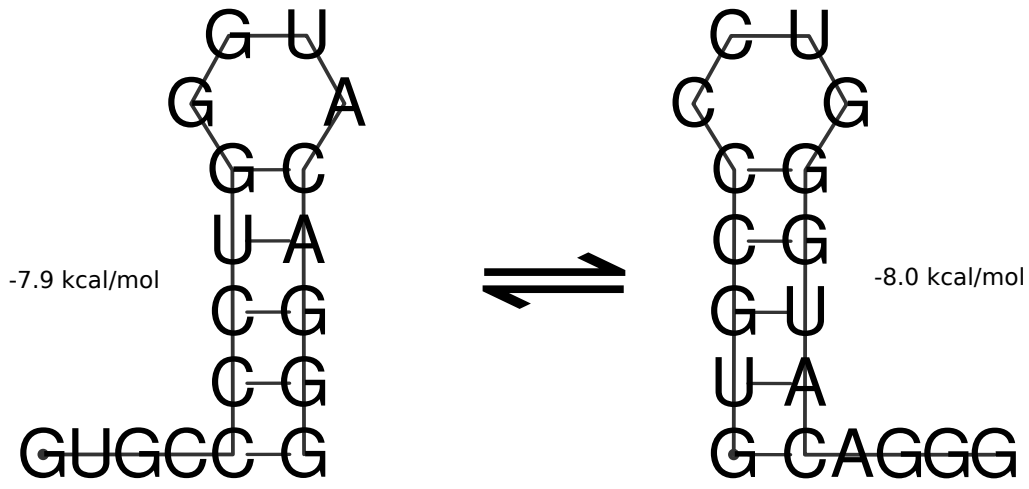
- If homologous sequences are available, use them!
- Structural conservation implies function
- Pure sequence alignment is not always sufficient

Thermodynamic vs. Kinetic Folding

Equilibrium properties for RNA secondary strutcures can be calculated efficiently

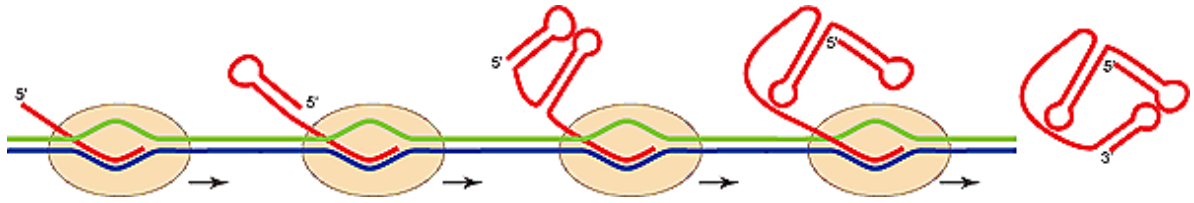
But what about dynamics?

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



Folding during Transcription

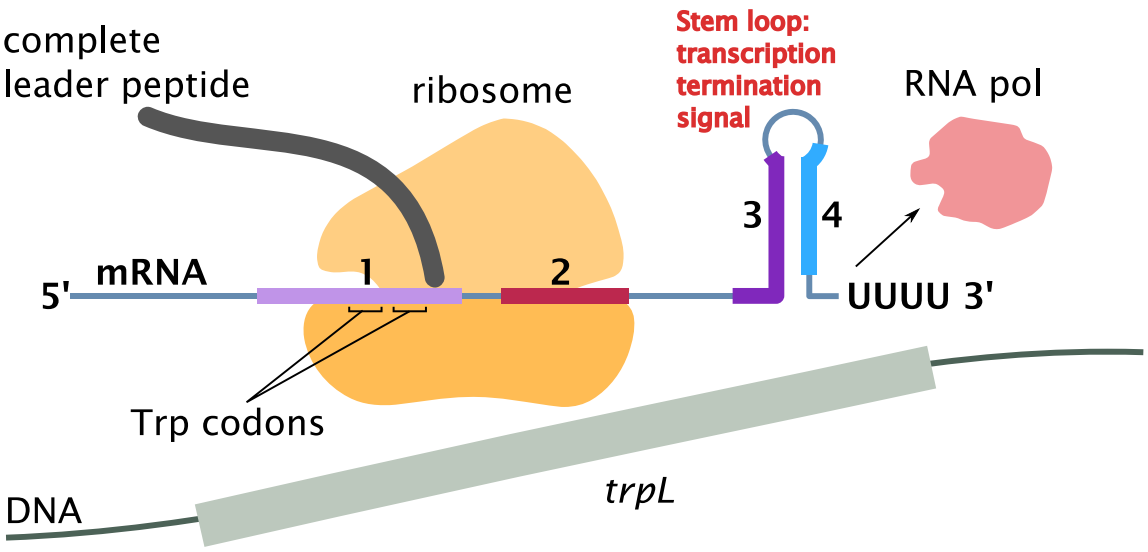
Almost all RNA structures may be affected by co-transcriptional folding:



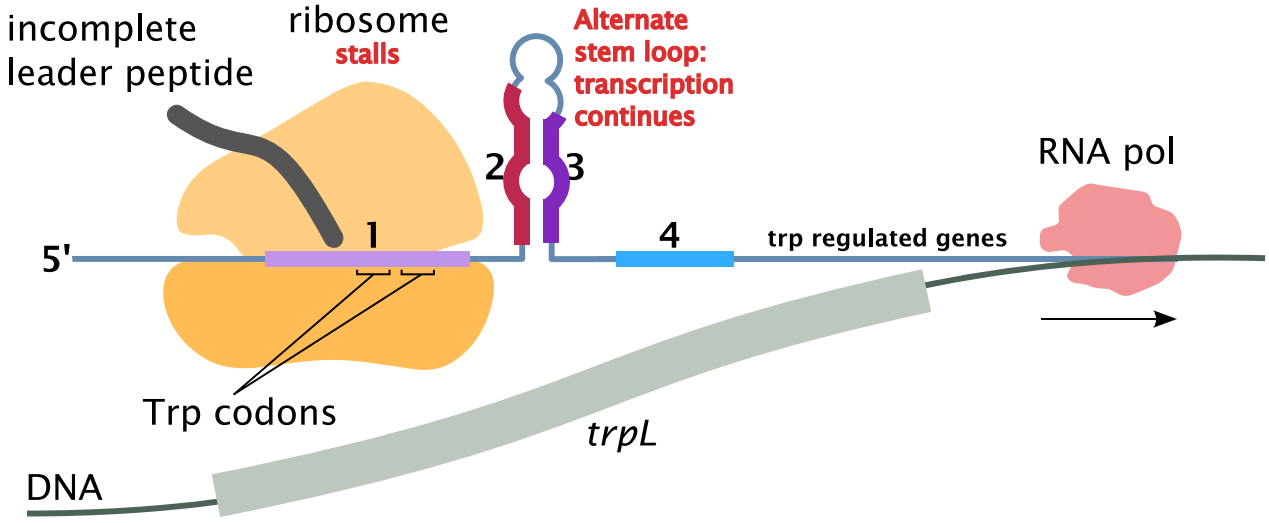
- RNA is transcribed at a rate of only 25–50 nucleotides per second
- The nascent chain starts folding as soon as its leaves the ribosome
- Stems formed by the incomplete chain may be too stable to refold later on
- Co-transcriptional folding may drive the folding process to a well-defined folded state (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

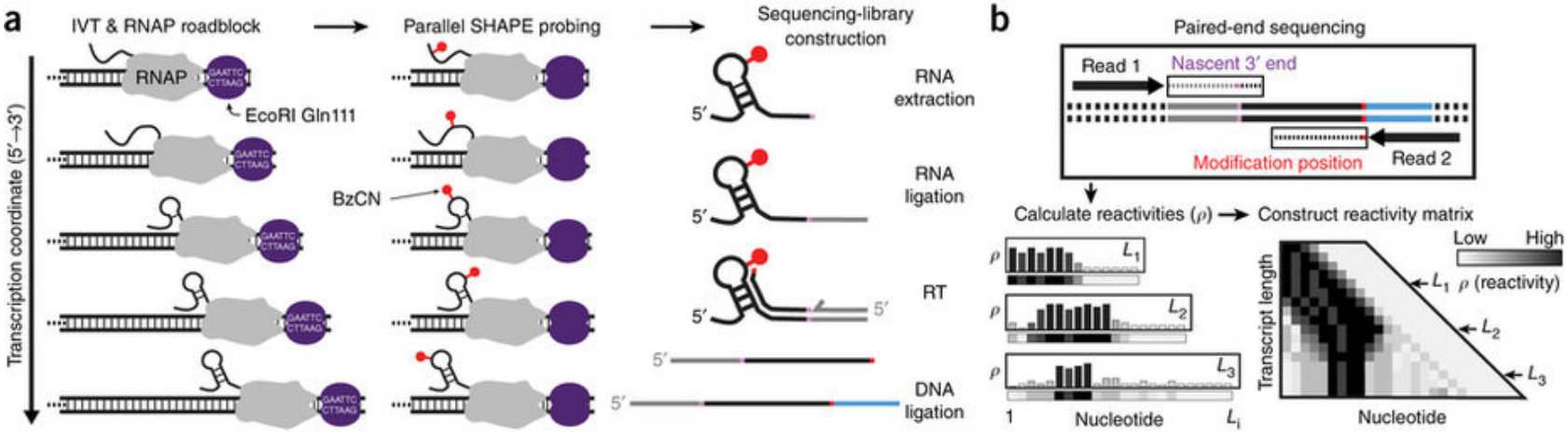


Low level of tryptophan



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

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

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)

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.

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`

Stochastic Simulations

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

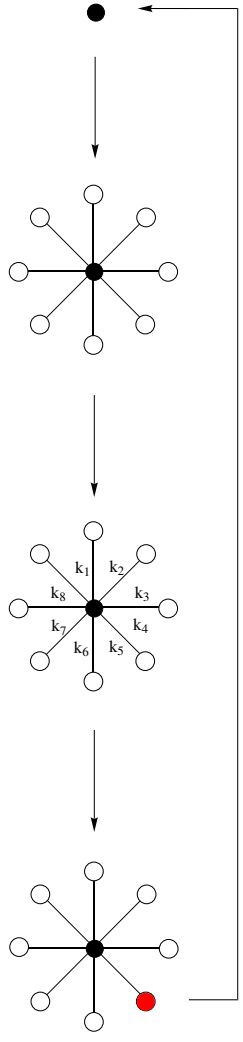
Generate all neighbors using a move-set:
close single base-pair and *open* single base-pair

Assign rates to each move, e.g.:

$$k_i = k_0 \cdot \min\left\{1, \exp\left(-\frac{\Delta E}{kT}\right)\right\}$$

Select a move i with probability $\propto k_i$

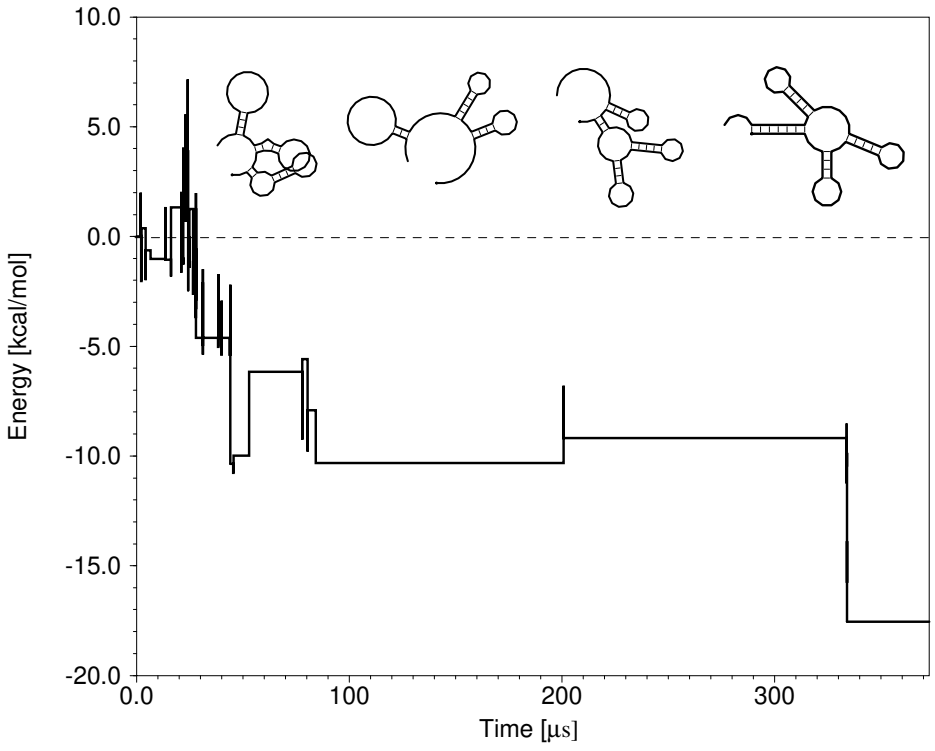
Advance clock by $1/\sum_i k_i$ (on average).



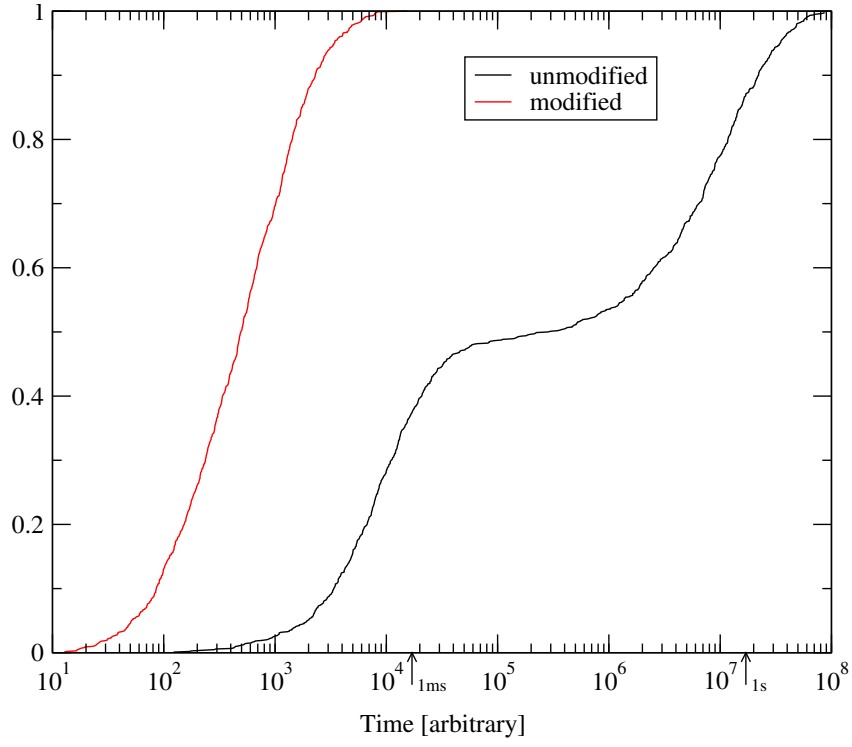
- computationally somewhat expensive
- need to analyze many trajectories
+ easy to include co-transcriptional folding

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

Kinetic Rate Models

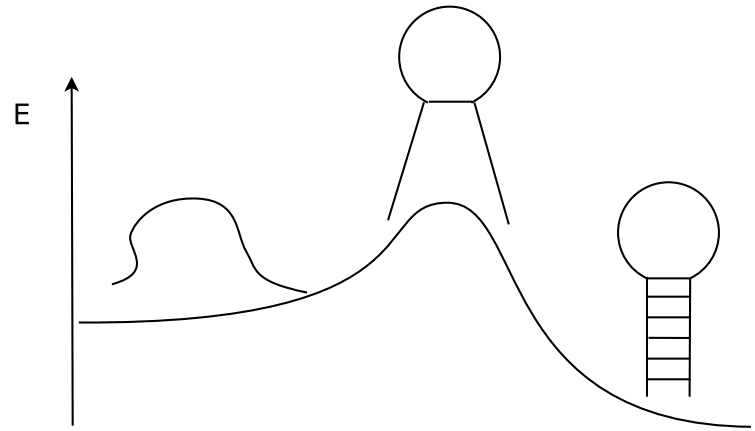
The simplest rate model satisfying detailed balance is the Metropolis rule

$$k_{x\to y} = k_0 \cdot \min\left(1,\ e^{-(\Delta G(y)-\Delta G(x))/RT}\right)$$

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

$$k_{x\to y} = k_0 \exp\left(-(\Delta G^{\dagger}_{xy} - \Delta G(x))/RT\right)$$

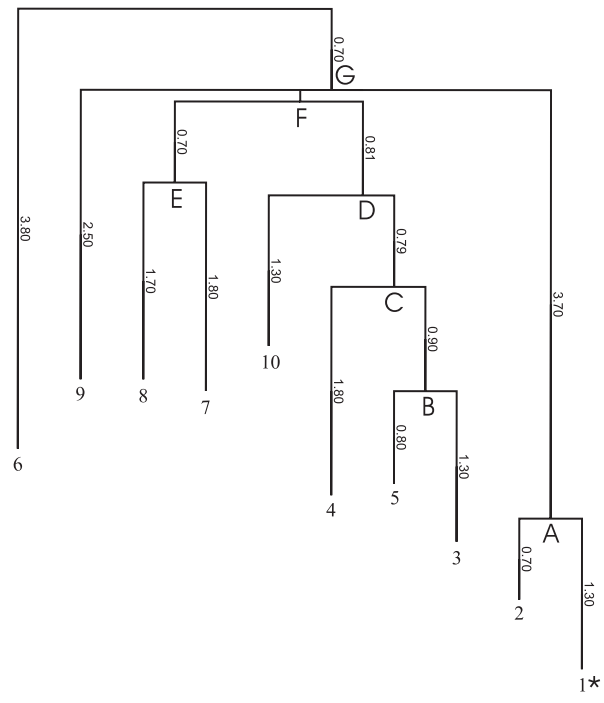
This is essential for large moves (e.g. helix moves).



RNA Landscape Analysis

Barrier trees

- Contains all local minima as leaves
- 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:

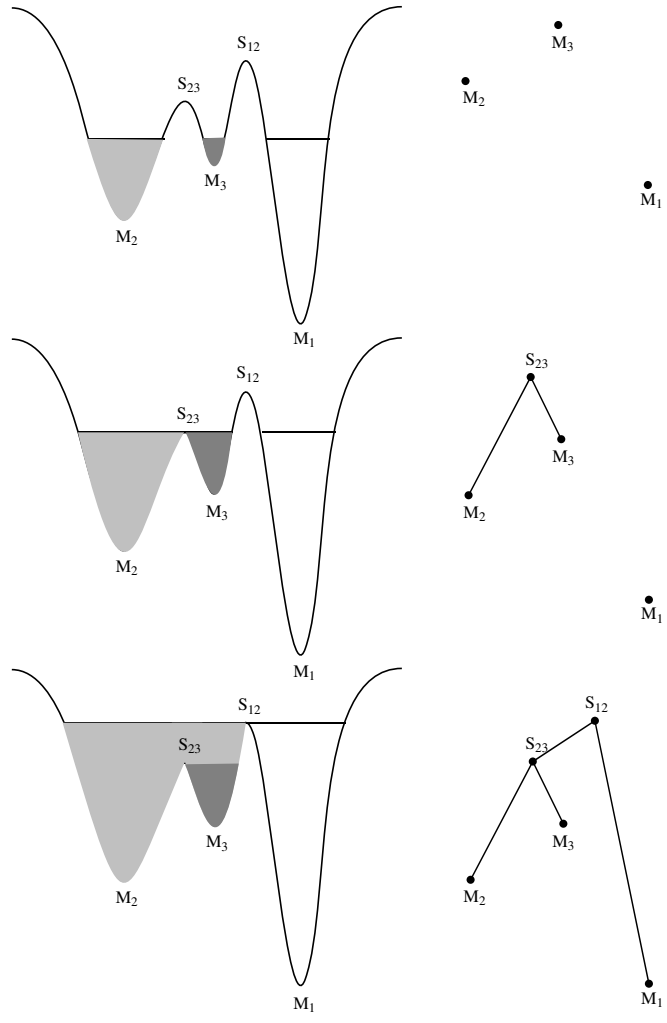
Read conformations in energy sorted order.

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.

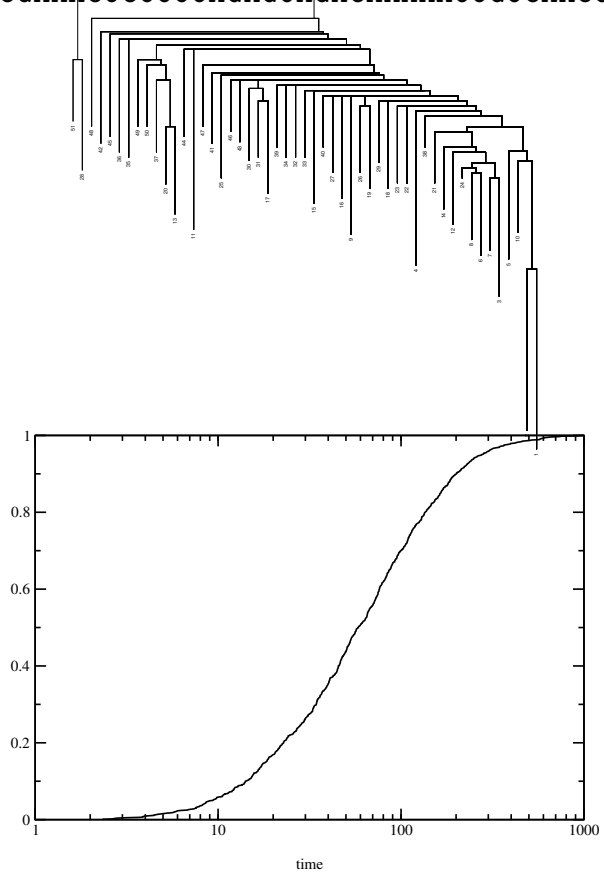


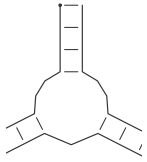
The barriers program

- 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

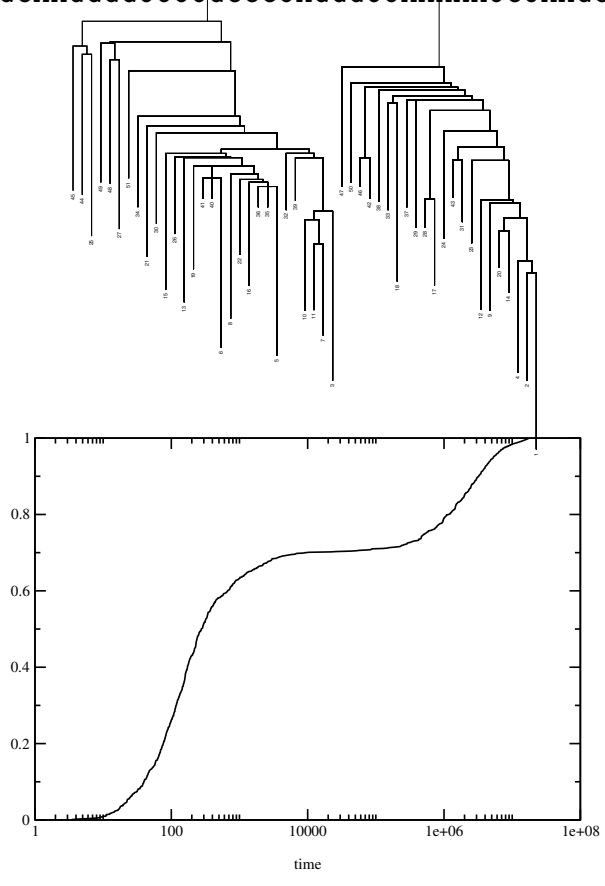
Fast Folder vs. Slow Folder

GGCGAAACUCUUUUAGAGUAGACAAAAAUUGUCAACGUC

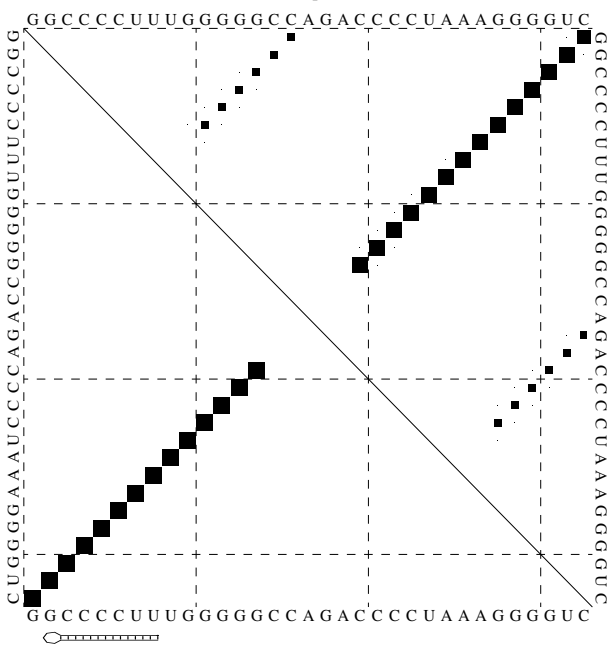


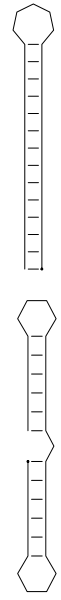


GCGCAAGGGGUUUUGCCCUAGGGUUAAAAAUCUAAGCGC



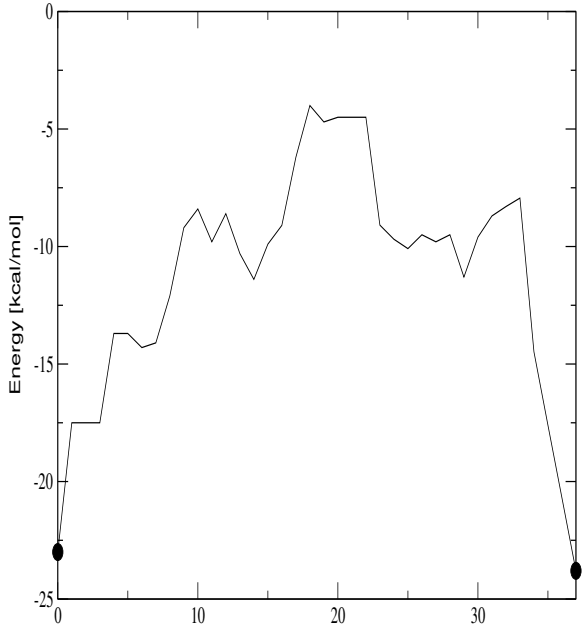
A designed bi-stable Sequence

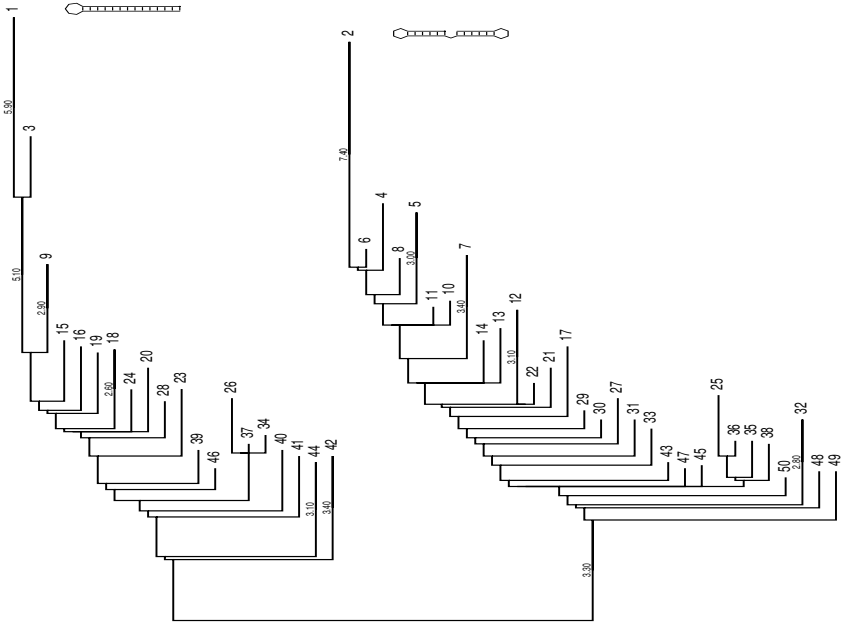




-23.8 kcal

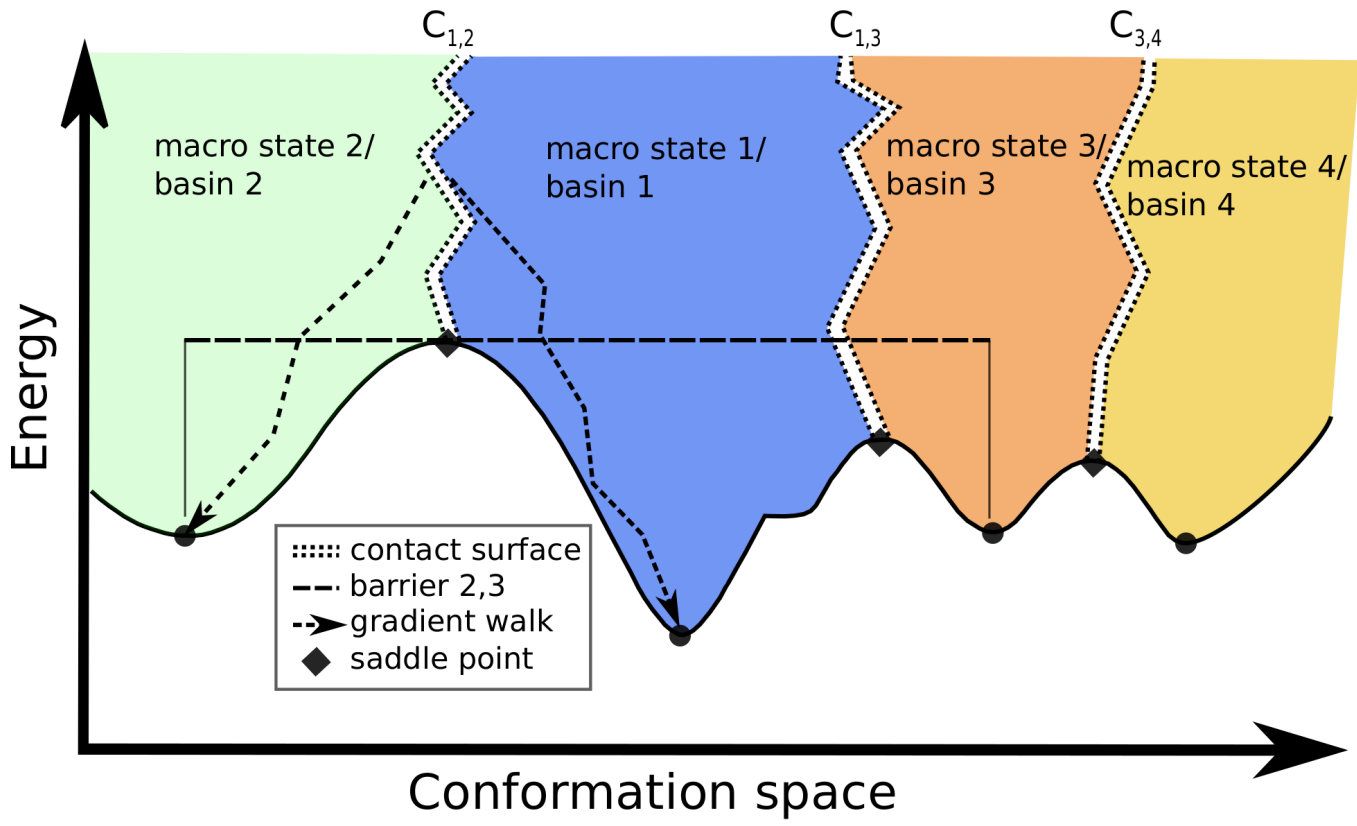
-23.0 kcal





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

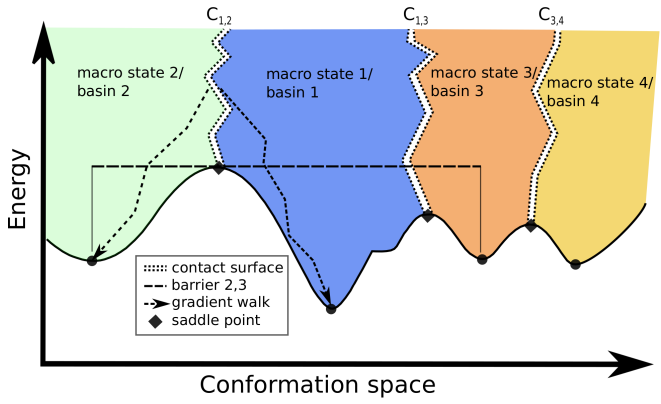
Coarse Graining the Landscape



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



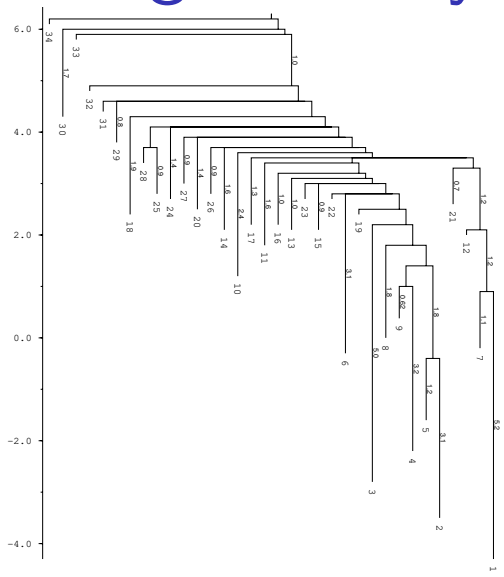
Transition rates could follow an Arrhenius rule
$r_{\beta\alpha} = \exp\left(-(E^*_{\beta\alpha} - G_\alpha)/RT\right)$.

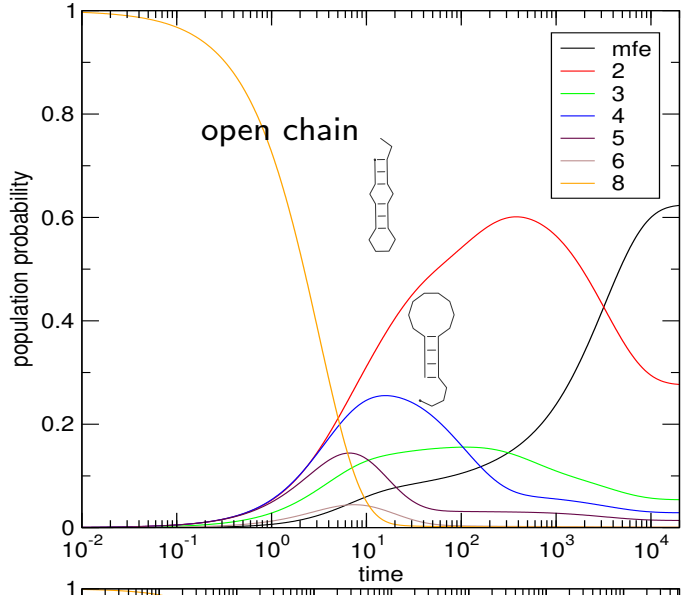
Better: include *all* transition states

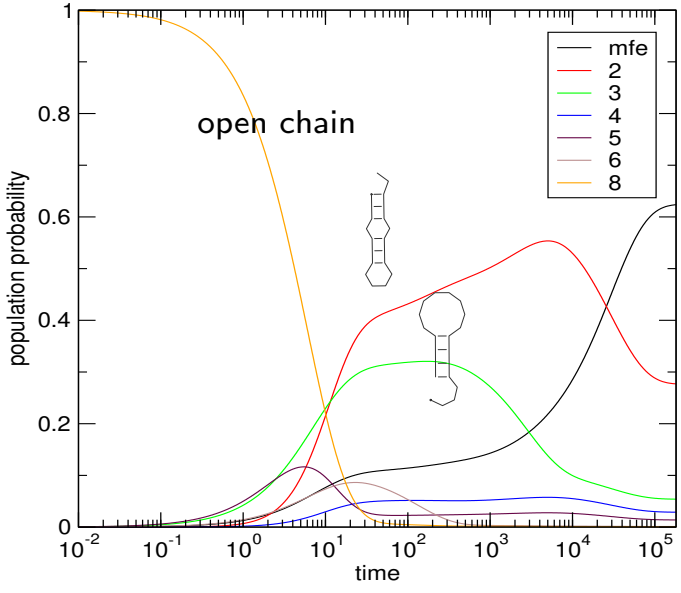
$$r_{\beta\alpha} = \sum_{y\in\beta}\sum_{x\in\alpha} r_{yx}\mathrm{Prob}[x|\alpha] \approx \frac{1}{Z_\alpha}\sum_{y\in\beta}\sum_{x\in\alpha} r_{yx}e^{-E(x)/RT}$$

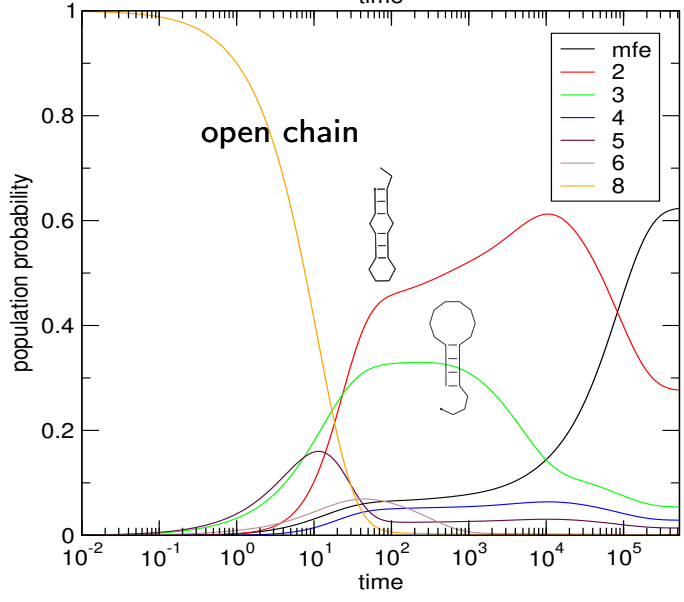
assuming local equilibrium.

Coarse grained dynamics vs. full dynamics



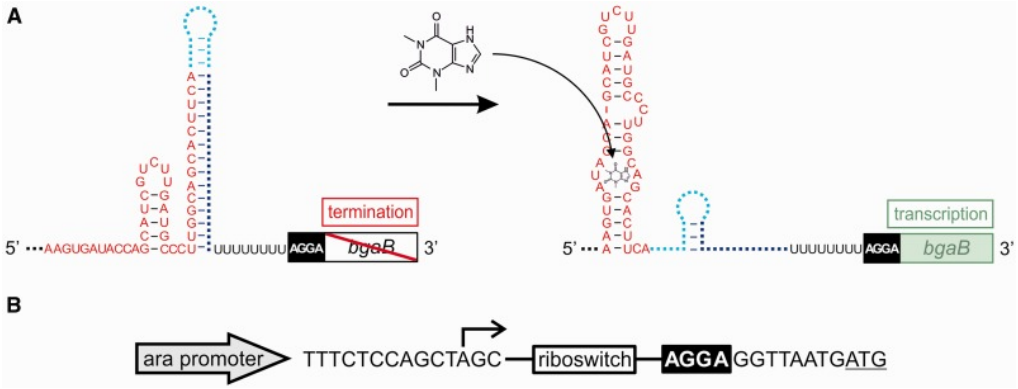






An Artificial Riboswitch

A designed *transcriptional* switch

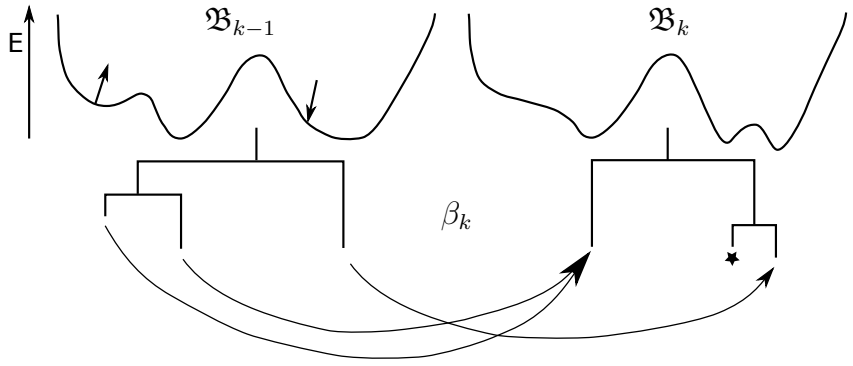


Wachsmuth et al, NAR (2013)

- 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

Co-transcriptional with BarMap

Each extension of the RNA structure modifies the landscape:



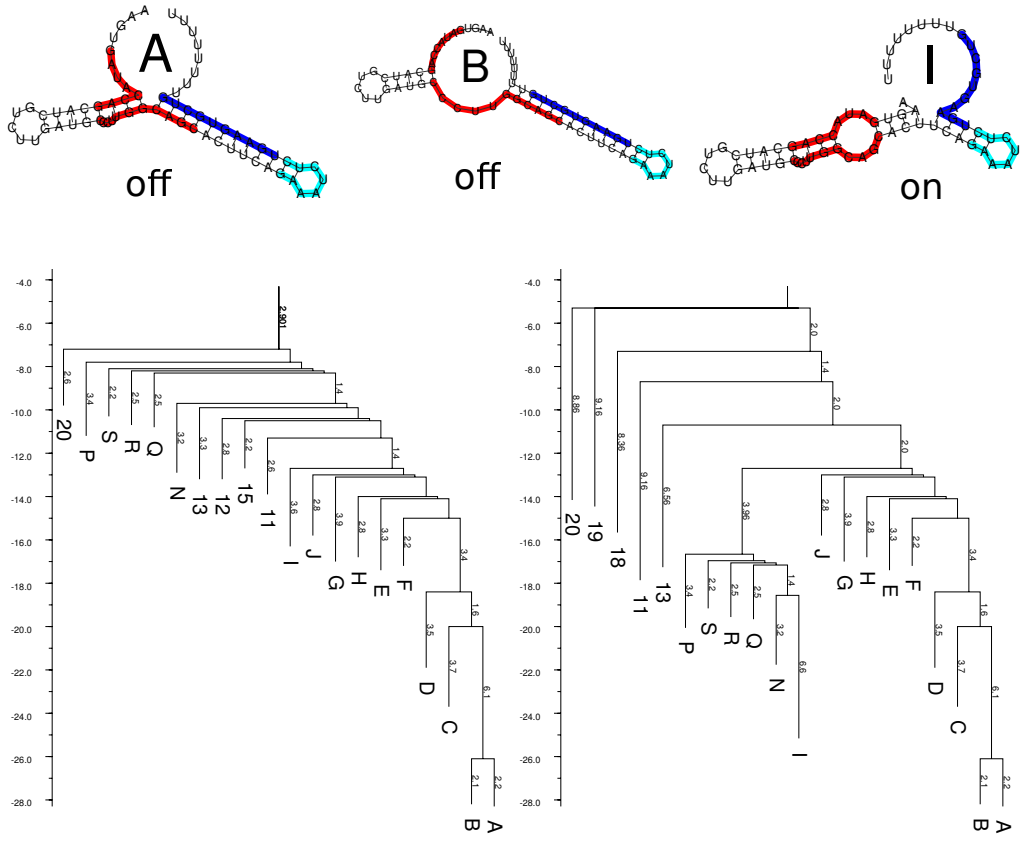
- Compute barrier trees for each sequence length $1 \ldots n$
- Compute a mapping between the minima of subsequent landscapes
- Compute dynamics piece-wise:
  - Compute dynamics on landscape for length k
  - Transfer population to landscape of length $k + 1$

Hofacker et al., RNA (2010)

How to include Ligand Binding ?

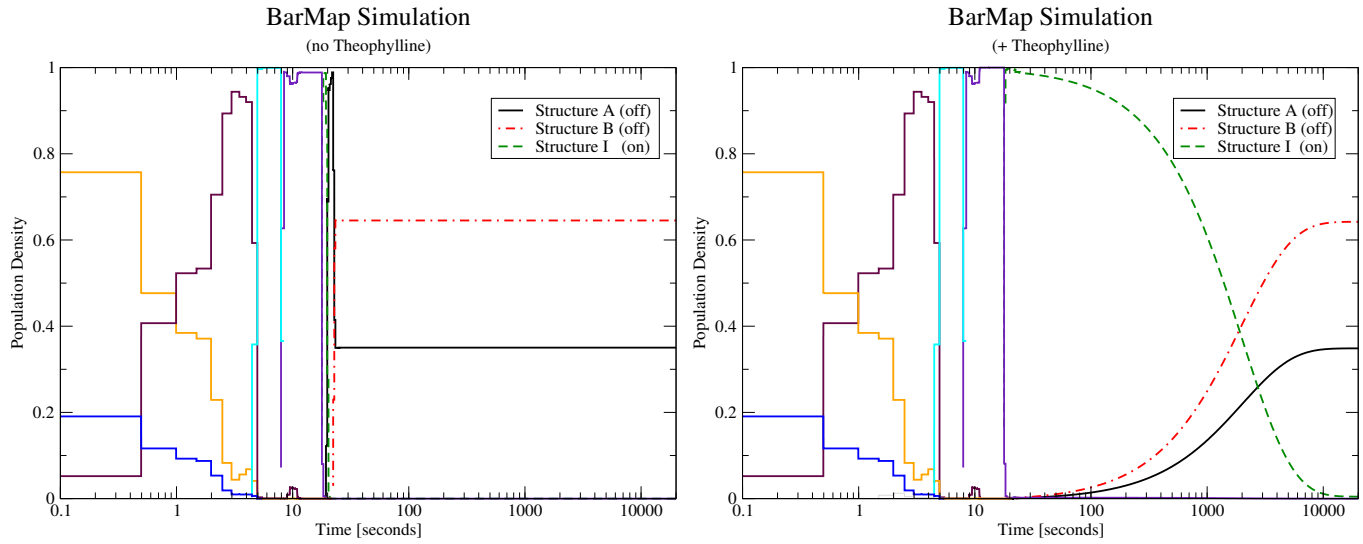
- Need to know binding motif and binding rates from experiment
- Simple strategy:
  - Add binding energy $\theta = RT \ln \frac{K_d}{c^{\ominus}}$ to every binding competent structure
  - Assumes infinite ligand concentration and infinitely fast binding
- Treat binding / unbinding events explicitly
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

Barrier Tree for RS10 with and without Theophylline



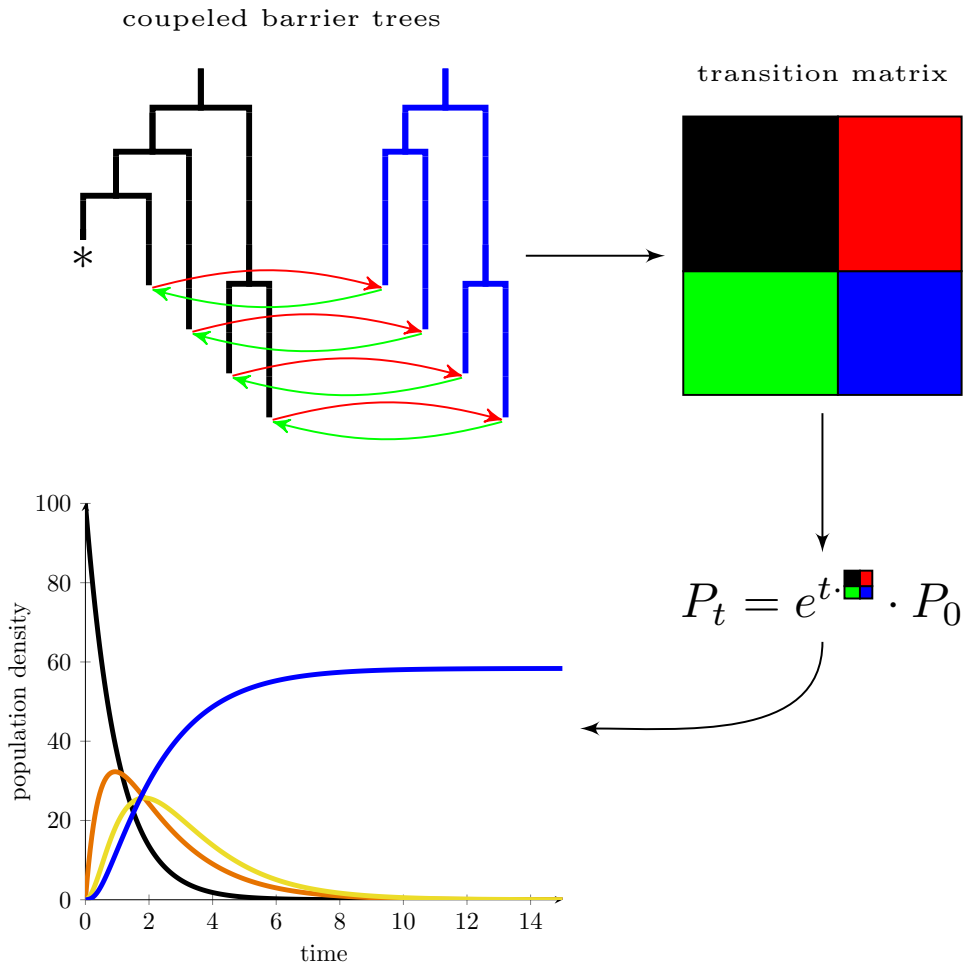
- Binding motif and K_d measurements
- Binding-competent structures are stabilized by about 8.9kcal/mol
- $\Rightarrow$ Distortion of the folding landscape by ligand

Co-transcriptional 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

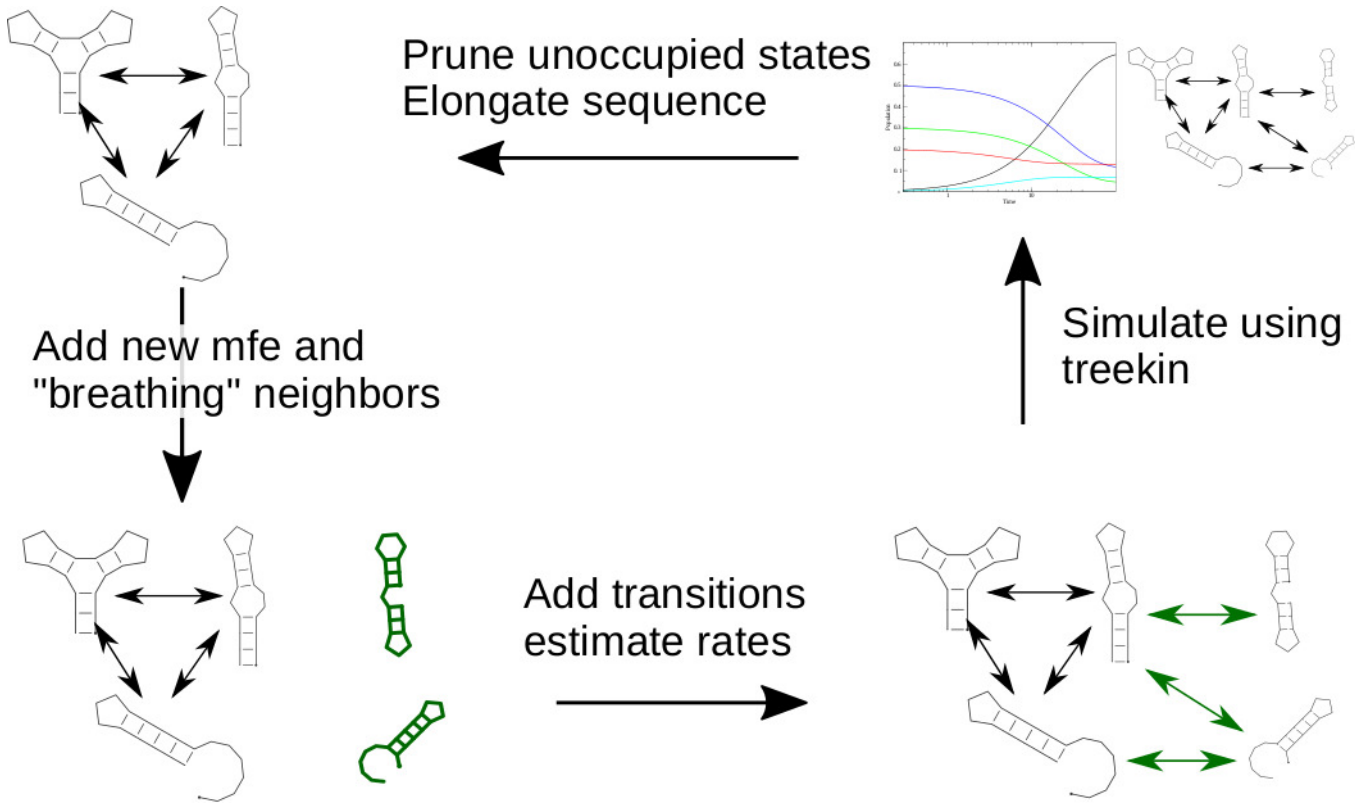
With explicit binding events...



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

DrTransformer: Fast co-transcriptional Folding

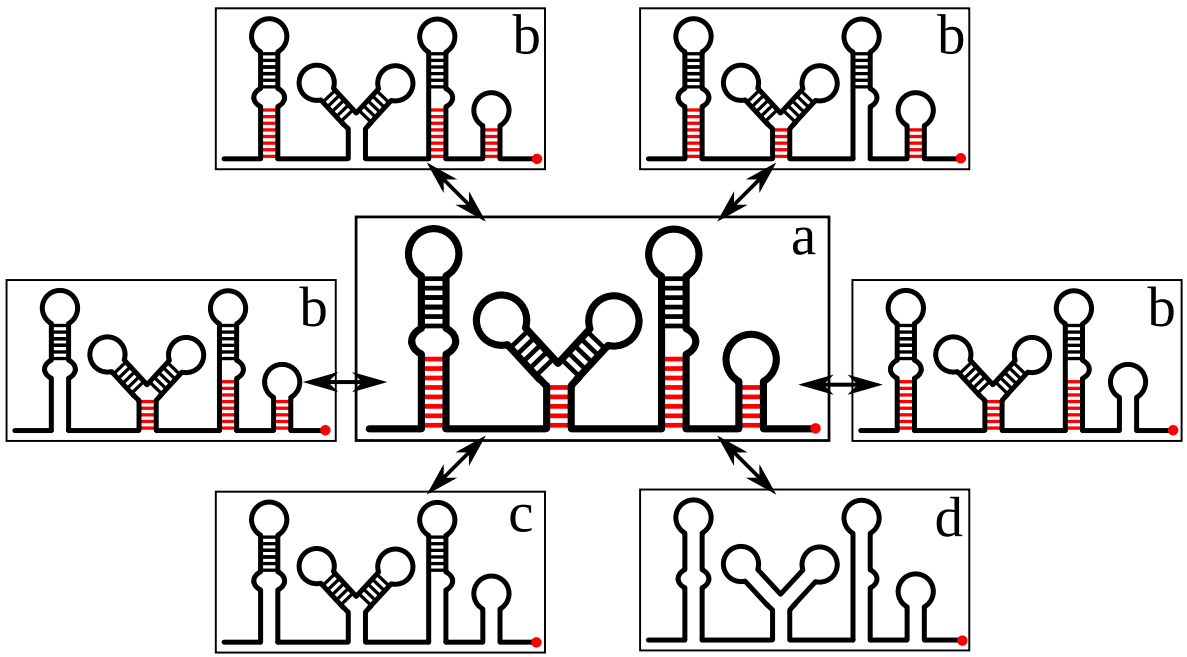
- Simulate a **small** network consisting only of the most relevant structural states
- Evolve network as RNA grows



DrTransformer: "Breathing" neighbors

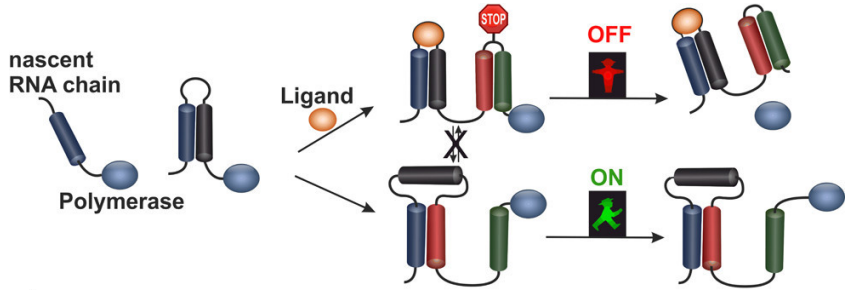
Which new structures should be added after an elongation step?

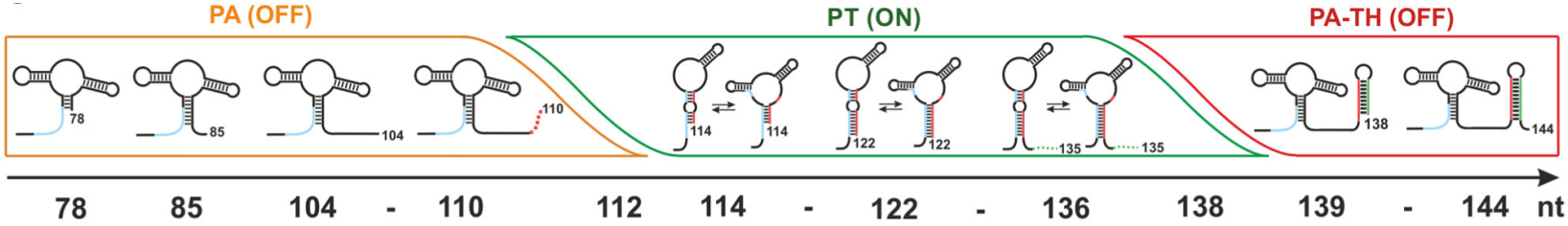
- Elongation can only effect the surroundings of the exterior loop
- Partially unfold all helices that protrude from exterior loop
- Use constrained folding to fold exterior loop surroundings



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

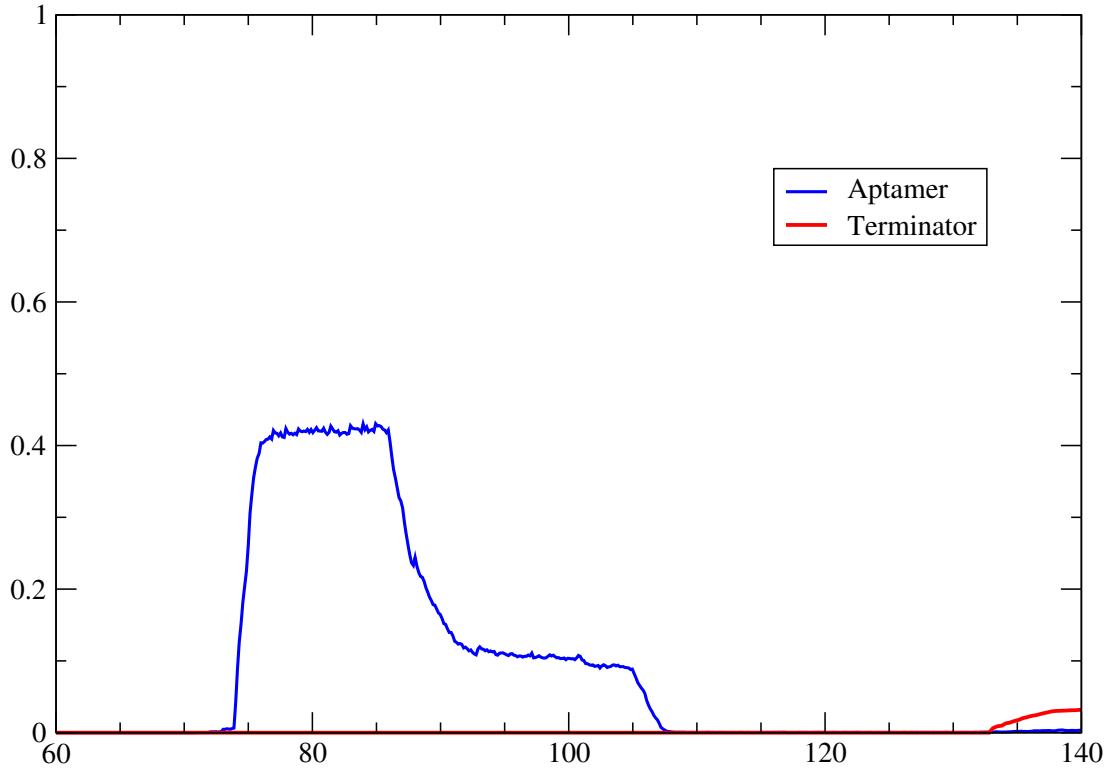




Helmling et al, JACS (2017)

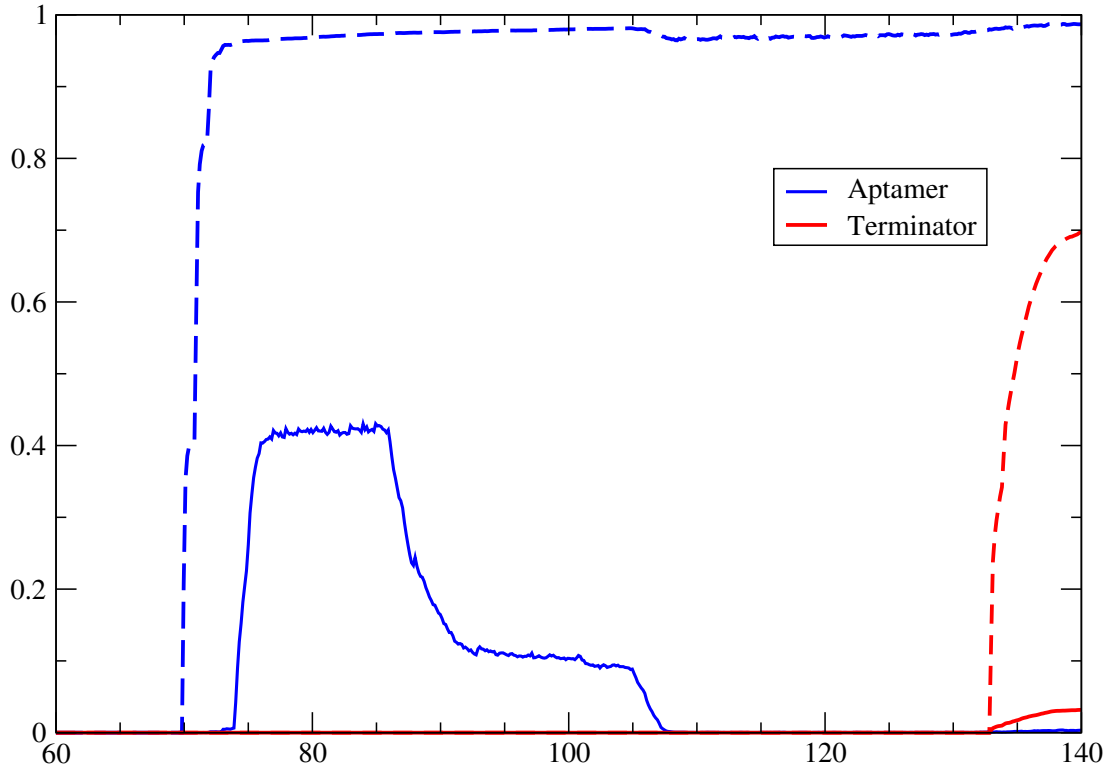
Kinfold simulation of the dG Riboswitch

- 10000 Kinfold trajectories (186 cpu hours)
- Classify each structure as aptamer and/or terminator



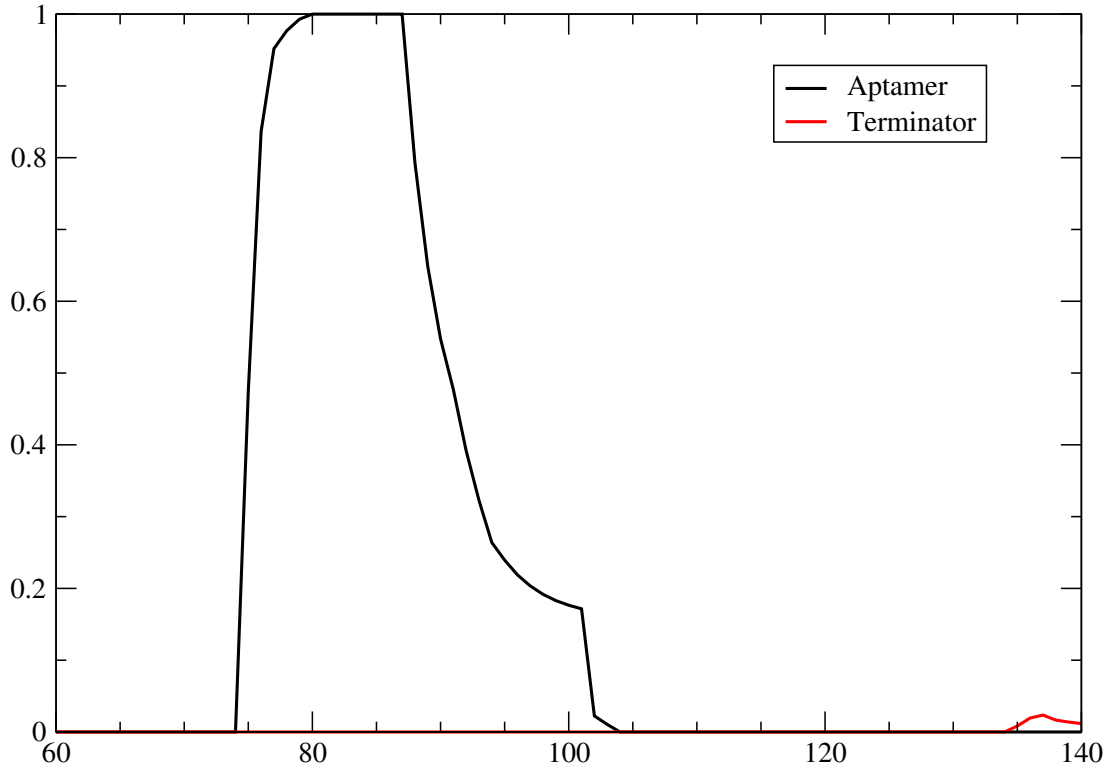
Kinfold simulation of the dG Riboswitch

- 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



Take home messages (this part)

- 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