Statistical Thermodynamics of RNA Secondary Structures

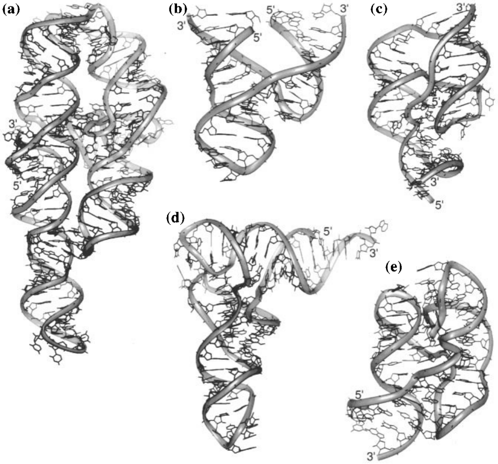
Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science & Interdisciplinary Center for Bioinformatics, **University of Leipzig**

Max-Planck-Institute for Mathematics, Leipzig
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
The Santa Fe Institute (external faculty)

Oldenburg, 02 Nov 2017

RNA Structure

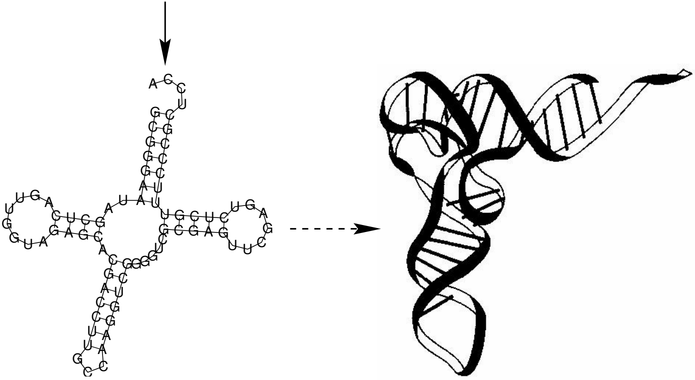


(a) Group I intron P4-P6 domain
(b) Hammerhead ribozyme
(c) HDV ribozyme
(d) Yeast tRNA$^{\mathrm{phe}}$
(e) L1 domain of 23S rRNA

Hermann & Patel, JMB 294, 1999

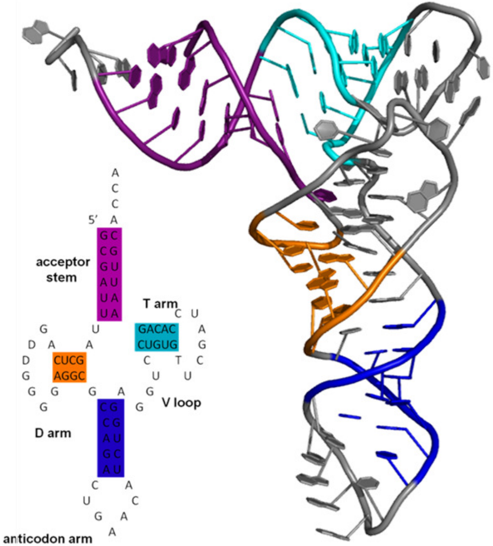
The RNA Secondary Structure Model

GCGGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUUCCCGCUCCA



1. Secondary structures are folding intermediates
2. Secondary structures capture most of the energy of folding

RNA Structure



RNA Folding as Matching Problem

- Vertex set $V = \{1, \dots n\}$ labeled by the nucleotides $\in \{A, U, G, C\}$
- Legal base pairs $\{i, j\} \in E$ iff $\{\ell(i), \ell(j)\} = GC, AU, GU$
- Secondary structure = circular matching on $G(V, E)$.
- circular ⇔ non-crossing rule $\{i, j\}, \{k, l\} \in M$ and $i < k < j$ then $i < l < j$. excludes pseudoknots
- steric constraint: $\{i, j\} \in M$ implies $|i - j| > 3$.
- energy function defined on edges or certain cycles

Dot-Bracket Notation

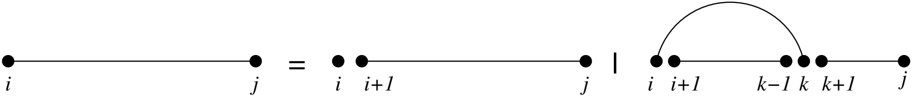
- Base pairs do not cross each other, i.e., every pair is either contained, containing, or on the side of another base pair: like matching parentheses.
- Each base is either unpaired, or opening or closing a base pair use symbols "." "(" ")"
- hairpin `((((....))))`
- a clover leaf `(((.(((...))).(((...))).(((...))).)))`

Let's count the structures . . .

Counting secondary structures. Given a sequence of length n. $\Pi_{kl} = 1$ if sequence positions k, l **can** form a pair **GC, CG, AU, UA, GU, UG** and $\Pi_{kl} = 0$ otherwise.

N_{ij} = number of structures of the *subsequence* from i to j.

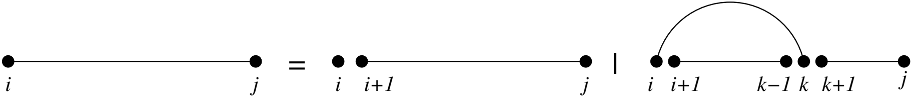
Basic recursion:



$$N_{ij} = N_{i+1,j} + \sum_{k=i+m}^{j} \Pi_{ik} N_{i+1,k-1} N_{k+1,j}$$

RNA Folding in a nutshell

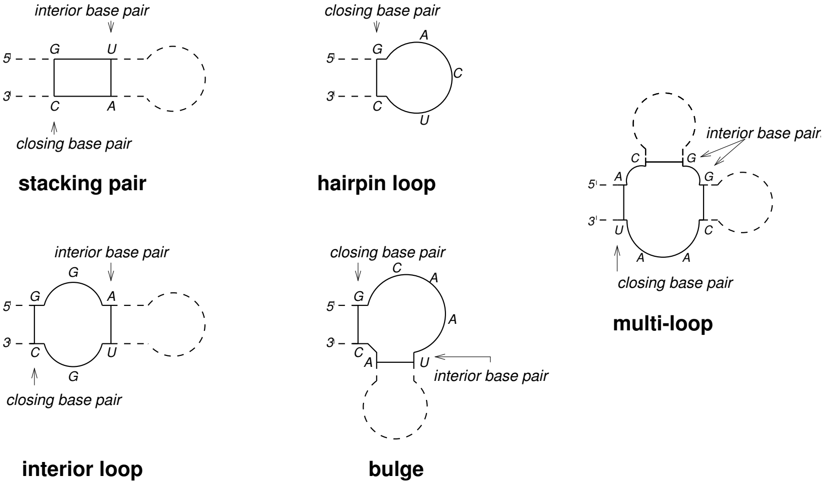
Other quantities can be obtained analogously:



$$N_{ij} = N_{i+1,j} + \sum_{\substack{k \\ (i,k)\text{pair}}} N_{i+1,k-1} N_{k+1,j}$$

$$E_{ij} = \min\left\{ E_{i+1,j} + \min_{\substack{k \\ (i,k)\text{pair}}} \left(E_{i+1,k-1} + E_{k+1,j} + \varepsilon_{ik} \right) \right\}$$

Realistic Energy Model



stacking pair

hairpin loop

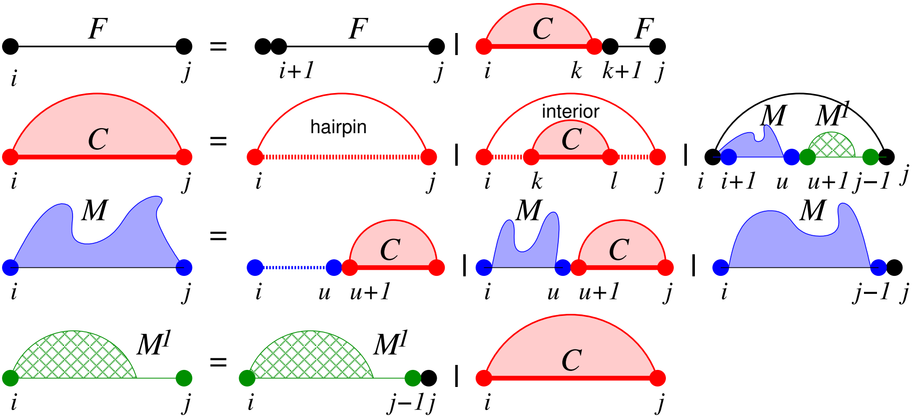
multi-loop

interior loop

bulge

Parameters from large number of melting experiments by Douglas Turner, David Matthews, John Santa Lucia, and others

Forward Recursion

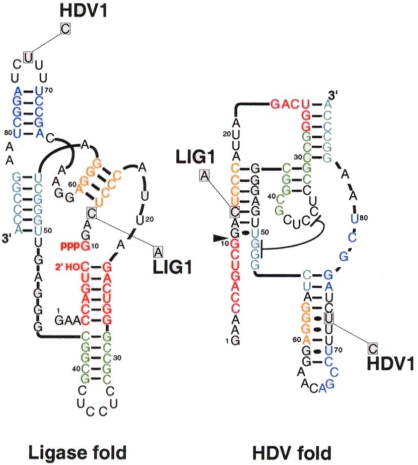


Backward Recursion: Minimum Energy Structure

Idea: use a stack on which the sub-sequences are stored that still need to be investigated

1. $\mathfrak{S} \rightarrow [1, n]$
2. General Recursion while $\mathfrak{S} \neq \emptyset$
    1. take interval from $\mathfrak{S} \rightarrow [i, j]$.
    2. if $E_{i,j} = E_{i+1,j}$:
       Position i unpaired
       $\mathfrak{S} \rightarrow [i+1, j]$
    3. else: find k so that $E_{i,j} = E_{i+1,k-1} + E_{k+1,j} + \beta_{ik}$
       Basepair (i, k)
       $\mathfrak{S} \rightarrow [i+1, k-1] \qquad \mathfrak{S} \rightarrow [k+1, j]$

One Sequence, Two target Structures



Schultes, EA & Bartel, DP; Science (2000), **289**:448-452

Suboptimal structure

with a little bit of more thinking we can also get all structures within a certain energy range above the ground state.

```
UCAGUGAUUUCAGCUCUUUUAGUAUUUGUCCAGCAGGUUUCCCGCCCCGCGGGAAGCCCCACUGU
.(((((......(((................))).((((((((((...)))))))))).))))). -22.00
.(((((......(((...(........)...))).((((((((((...)))))))))).))))). -21.50
.(((((....................(((...)))((((((((((...)))))))))).))))). -21.20
.(((((......(((....((.....))...))).((((((((((...)))))))))).))))). -21.20
.(((((......(((....(((...)))...))).((((((((((...)))))))))).))))). -21.10
.(((((......(((.....)))............((((((((((...)))))))))).))))). -21.00
```

... many structures with often very small energy ranges

Boltzmann Ensembles of Secondary Structures

- Probability of a structure: $Prob(s) \propto \exp(-E(s)/RT)$
- Normalization constant = partition function

$$Z = \sum_{s} \exp(-E(s)/RT)$$

- Link to thermodynamics: Free energy

$$\Delta G = -RT \ln Z$$

- Probability that we observe a structure from a set Ψ?

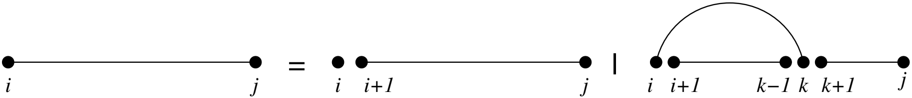
$$Prob(\Psi) = \sum_{s \in \Psi} \frac{1}{Z} \exp(-E(s)/RT) = Z(\Psi)/Z$$

... too many structure to enumerate in practise.

- Most important example: Compute the probabilities of all base pairs, i.e. Ψ = set of structure with a given base pair i, j

RNA Folding in a nutshell

Compute partitial paritition function:



$$N_{ij} = N_{i+1,j} + \sum_{\substack{k \\ (i,k)\text{pair}}} N_{i+1,k-1} N_{k+1,j}$$

$$E_{ij} = \min\left\{ E_{i+1,j} + \min_{\substack{k \\ (i,k)\text{pair}}} \left(E_{i+1,k-1} + E_{k+1,j} + \varepsilon_{ik}\right) \right\}$$

$$Z_{ij} = Z_{i+1,j} + \sum_{\substack{k \\ (i,k)\text{pair}}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\varepsilon_{ik}/RT)$$

Backward Recursion: Base Pairing Probabilities

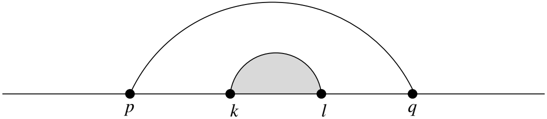
$$p_{ij} = \frac{Z_{1,i-1}\widehat{Z}_{i,j}Z_{j+1,n}}{Z_{1,n}} + \sum_{k<i}\sum_{l>j} p_{kl}\Xi_{ij,kl}\,.$$

$\Xi_{ij,kl}$ is a ratio of the two partition functions:
$\widehat{Z}_{ij,kl}$ ... both i,j and k,l pair
$\widehat{Z}_{kl}$ ... k,l pair.
Simplest case:
$\widehat{Z}_{ij,kl} = Z_{k+1,i-1}\widehat{Z}_{ij}Z_{j+1,l-1}\zeta_{kl}$ where $\zeta_{kl} = \exp(-\beta_{kl}/RT)$ is the Boltzman factor of the pairing energy

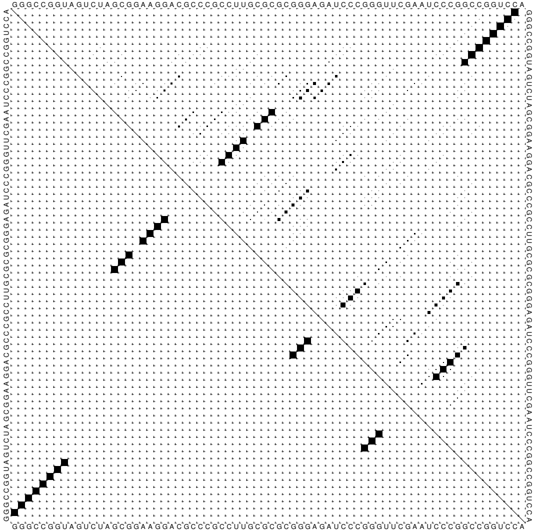
Backward recursion: full model

Backward recursion:

$$
\begin{aligned}
P_{kl} = P^{\circ}_{kl} + \sum_{p<k;q>l} P_{pq} \frac{Z^B_{k,l}}{Z^B_{p,q}} \Bigg\{ & e^{-\mathcal{I}(p,q,k,l)} \\
& + \left(\sum_{p<u<k} Z^M_{p+1,u} Z^{M1}_{u+1,k-1}) \right) e^{-(a+(q-l-1)c)} \\
& + \left(\sum_{l<u<q} Z^M_{l+1,u} Z^{M1}_{v+1,q-1}) \right) e^{-(a+(k-p-1)c)} \\
& + Z^M_{p+1,k-1} Z^M_{l+1,q-1} \Bigg\}
\end{aligned}
$$



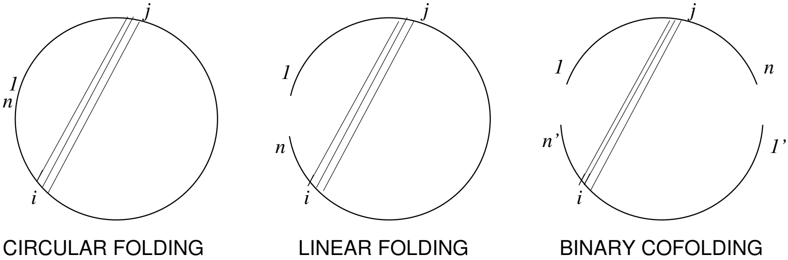
Base Pairing Probability Matrices



Circular, Linear, and Interacting RNAs

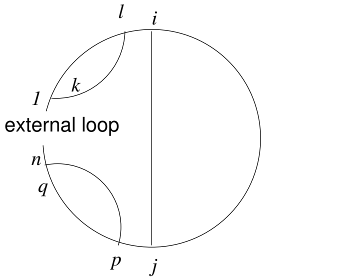
In the maximum matching case

$\Longrightarrow$ same algorithm for all three cases

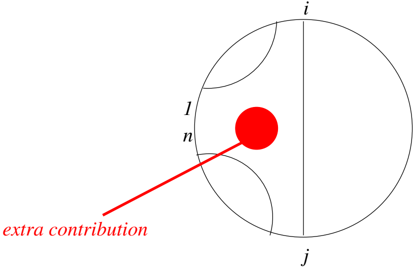


Linear versus Circular Folding

Linear folding: energy contributions *inside* a pair (i,j) *only*.
Co-folding: additional contribution for loop spanning $[n,1]$.



no energy contribution for external loop



no external loop

Local structures

Idea: Restrict Recursion to base pairs (i,j) with $j-i<L$.

Special interest in robust structures:

$Z_{ij}^{u,L}$ …partition function of sub-sequence $[i,j]$ when sequence window $[u,u+L]$ is folded

$p_{ij}^{u,L}$ …probability that i and j form a base pair when window $[u,u+L]$ is folded.

$$Z_{ij}^{u,L} = \begin{cases} Z_{ij} & \text{if} \quad [i,j]\subseteq[u,u+L] \\ 0 & \qquad\qquad \text{otherwise} \end{cases}$$

$$\begin{aligned} p_{ij}^{u,L} &= \frac{Z_{1,i-1}^{u,L}\widehat{Z}_{i,j}^{u,L}Z_{j+1,n}^{u,L}}{Z_{u,u+L}^{u,L}} + \sum_{k<i}\sum_{l>j} p_{kl}^{u,L}\Xi_{ij,kl}^{u,L} \\ &= \frac{Z_{u,i-1}\widehat{Z}_{i,j}Z_{j+1,u+L}}{Z_{u,u+L}} + \sum_{k<i}\sum_{l>j} p_{kl}^{u,L}\Xi_{ij,kl}\,. \end{aligned}$$

Robust local structures

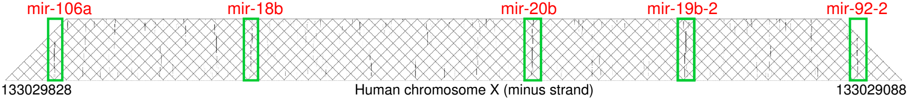
Average probability of an (i,j) pair over all folding windows containing the sequence interval $[i,j]$

$$\pi^L_{ij} = \frac{1}{L-(j-i)+1}\sum_{u=j-L}^{i} p^{u,L}_{ij} \,.$$

Direct Recursion:

$$\pi^L_{ij} = \underbrace{\frac{1}{L-(j-i)+1}\sum_{u=j-L}^{i}\frac{Z^{u,L}_{1,i-1}\widehat{Z}^{u,L}_{i,j}Z^{u,L}_{j+1,n}}{Z^{u,L}_{1,n}}}_{\pi^{*L}_{ij}} + \frac{1}{L-(j-i)+1}\sum_{u=j-L}^{i}\sum_{k<i}\sum_{l>j} p^{u,L}_{kl}\Xi_{ij,kl}$$

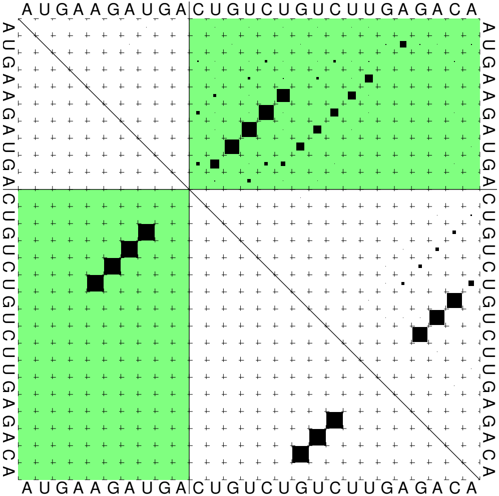
$$= \pi^{*L}_{ij} + \sum_{k=j-L}^{i-1}\sum_{l=j+1}^{i+L}\sum_{u=l-L}^{k}\frac{p^{u,L}_{kl}\Xi_{ij,kl}}{L-(j-i)+1} = \pi^{*L}_{ij} + \sum_{k=j-L}^{i-1}\sum_{l=j+1}^{i+L}\frac{L-(k-l)+1}{L-(j-i)+1}\pi^L_{kl}\Xi_{ij,kl} \,. \tag{1}$$

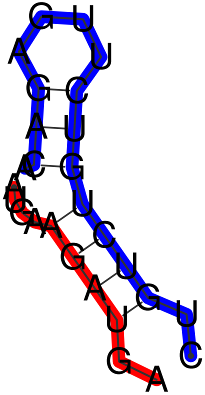


Cofold: How to deal with Concentration?

- Algorithmically that same as linear folding
  special energy contribution for “loop with the cut”
- Additional energy contribution for forming duplex
- At least 5 molecular species need to be taken into account (Dmitrov & Zuker, 2005): A, B, A_2, B_2, AB.
- Their folding energies and partition functions are easily computed

Cofold





Dot plot (left) and mfe structure representation (right) of the cofolding structure of the two RNA molecules AUGAAGAUGA (red) and CUGUCUGUCUUGAGACA.

Cofold: Concentration dependencies

$$\mathcal{Q} = V^n \frac{a!b! \times (Z'^A)^{n_A}(Z'^{AA})^{n_{AA}}(Z'^{AB})^{n_{AB}}(Z'^{BB})^{n_{BB}}(Z'^B)^{n_B}}{n_A! n_B! \, 2\, n_{AA}! \, 2\, n_{BB}! n_{AB}!}$$

where $a = n_A + 2n_{AA} + n_{AB}$. The system minimizes the free energy $-kT \ln \mathcal{Q}$.

Equilibria:

$[AA] = K_{AA}\,[A]^2\,, \qquad [BB] = K_{BB}\,[B]^2\,. \qquad [AB] = K_{AB}\,[A]\,[B].$

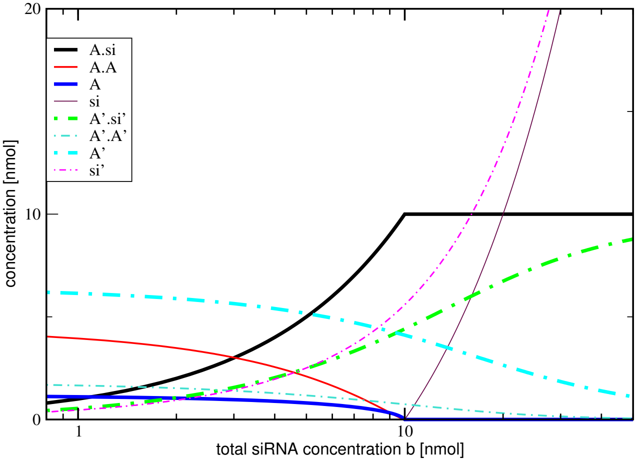
with

$$K_{AA} = \frac{Z'^{AA}}{(Z^A)^2} = \frac{(Z^{AA} - (Z^A)^2)e^{-\Theta_I/RT}/2}{(Z_A)^2} = \frac{1}{2}\,e^{-\Theta_I/RT}\left(\frac{Z^{AA}}{(Z^A)^2} - 1\right)$$

$$K_{BB} = \frac{1}{2}\,e^{-\Theta_I/RT}\left(\frac{Z^{BB}}{(Z^B)^2} - 1\right)$$

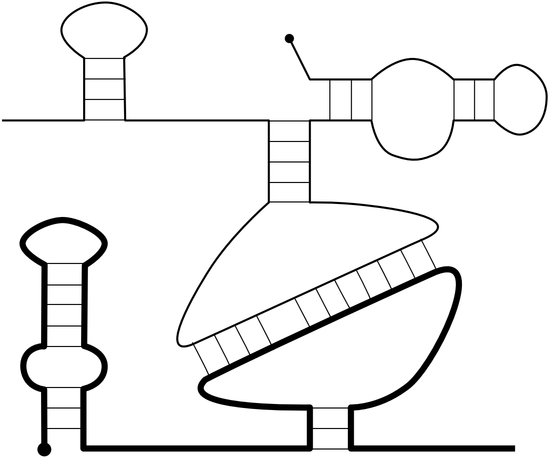
$$K_{AB} = e^{-\Theta_I/RT}\left(\frac{Z^{AB}}{Z^A Z^B} - 1\right)$$

Concentration Dependence



Example for the concentration dependency for two mRNA-siRNA binding experiments.

RNAup: Small RNAs Binding to Large Ones



`RNAup`: Small RNAs Binding to Large Ones

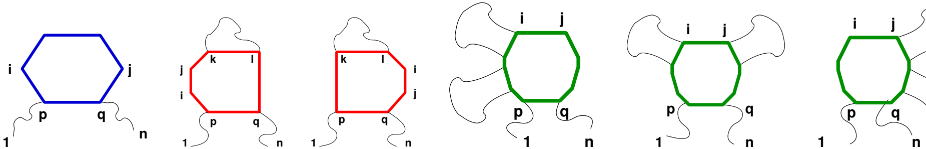
- RNA folding excludes pseudoknots, i.e., non outerplanar graphs
- `cofold` thus does not allow small RNA binding into loop regions of large ones
- ... but this happens in reality

Remedy: Compute energy/partition function

$$P_u[i,j] = \underbrace{\frac{Z[1,i-1] \times 1 \times Z[j+1,N]}{Z}}_{exterior} + \sum_{\substack{p,q \\ p<i\leq j<q}} \underbrace{P_{pq} \times \frac{Z_{pq}[i,j]}{Z^b[p,q]}}_{enclosed}$$

that subsequence $[i,j]$ is unpaired and the energy of binding a short molecule in this location

RNAup



(a) (b') (b") (c) (d) (e)

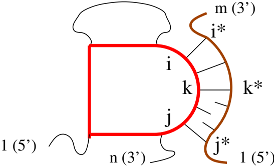
$$
\begin{aligned}
Z_{pq}[i,j] = {} & \underbrace{\exp(-\beta H(p,q))}_{(a)} \\
& + \sum_{\substack{p\,<\,i\,\le\,j\,<\,k \text{ or} \\ l\,<\,i\,\le\,j\,<\,q}} \underbrace{Z^b[k,l]e^{-\beta I(p,q;k,l)}}_{(b)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^{m2}[p+1,i-1]e^{-\beta c(q-i)}}_{(c)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^m[p+1,i-1]Z^m[j+1,q-1]e^{-\beta c(j-i+1)}}_{(d)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^{m2}[j+1,q-1]e^{-\beta c(j-p)}}_{(e)}
\end{aligned}
$$

RNAup: Interaction part

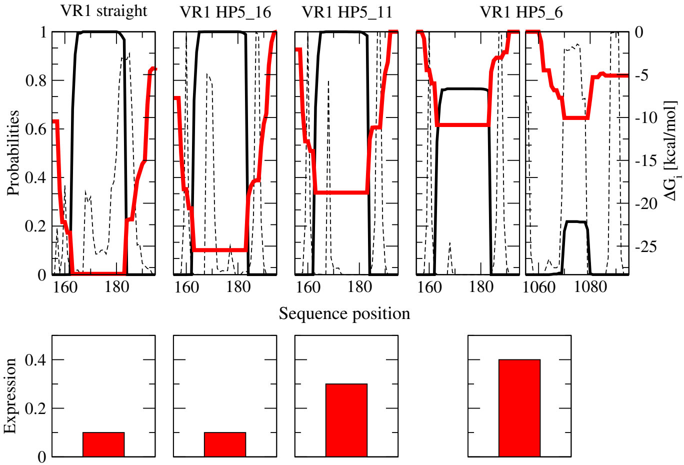
$$Z^I[i,j,i^*,j^*] = \sum_{\substack{i<k<j \\ i^*>k^*>j^*}} Z^I[i,k,i^*,k^*] e^{-\beta I(k,k^*;j,j^*)}$$

$$Z^*[i,j] = P_u[i,j] \sum_{i^*>j^*} Z^I[i,j,i^*,j^*];$$

$$P^*[i,j] = Z^*[i,j] / \sum_{k<l} Z^*[k,l]$$



RNAup: Application



Binding of siRNAs to VR mRNA.

$P_u[i,i]$ (dashed line), P_i^* (thick black line), ΔG_i (thick red line). Below: activity of siRNA

Application: Bacterial regulation

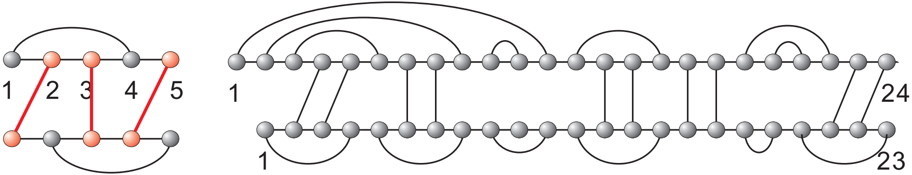
mRNA	sRNA	regulation	$\Delta\Delta G$	Position	Pos.lit.	cite
RyhB	sodB	-	-11.50	-18,+4	-4,+5	Geissmann:04
DsrA	hns	-	-14.60	-10,+11	+7,+19	Lease:98
MicA	ompA	-	-13.60	-21,-6	-21,-6	Rasmussen:05
MicC	ompC	-	-15.80	-30,-15	-30,-15	Chen:04
MicF	ompF	-	-17.80	-11,+9	-11,+10	Chen:04
Spot42	galK	-	-17.00	-18,+30	-19,+21	Moeller:02
SgrS	ptsG	-	-17.33	-28,-10	-28,+4	Kawamoto:06
GcvB	dppA	-	-17.30	-30,-7	-31,-14	Sharma:07[a]
DsrA	rpoS	+	-14.52	-126,-97	-119,-97	Majdalani:02
RprA	rpoS	+	-15.90	-134,-94	-117,-94	Majdalani:02

[a] GcvB/dppA interaction was studied in `Salmonella enterica` serovar Typhimurium not in E.coli.

Alkan's RIP Model

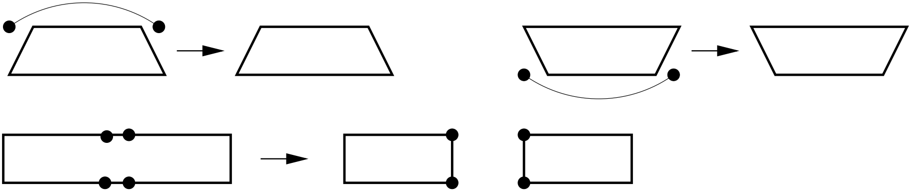
Two arbitrary secondary structures and non-crossing intermolecular base-pairs

Forbidden configuration: the "zigzag"



Solvable by dynamic programing in the absence of "zigzags":
previous work by several groups:
Alkan, Pervouchine, Mneimneh, Backofen & Sahinalp

RIPing it appart

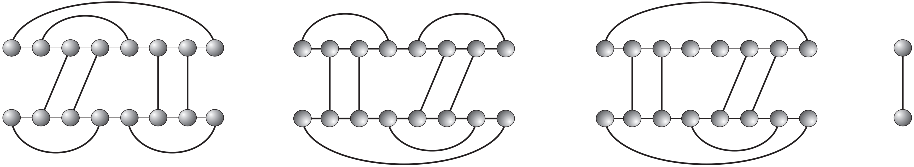


1. one of the partners is enclosed by a base pair:
   → "remove" this pair to reduce to a smaller problem.
2. neither of the partners is enclosed by a base pair:
   Then there are breakpoints p and q in the two sequences such that no pairs connect the block structure $x[1, p] : y[q + 1, n]$ with $x[p + 1, n] : y[1, q]$.
   → cut at p and q and treat the two blocks separately.

Tight Structures

Problem: decomposition is not unique. We therefore cannot use this to count structures or to compute a partition function.

We need an unambiguous decomposition



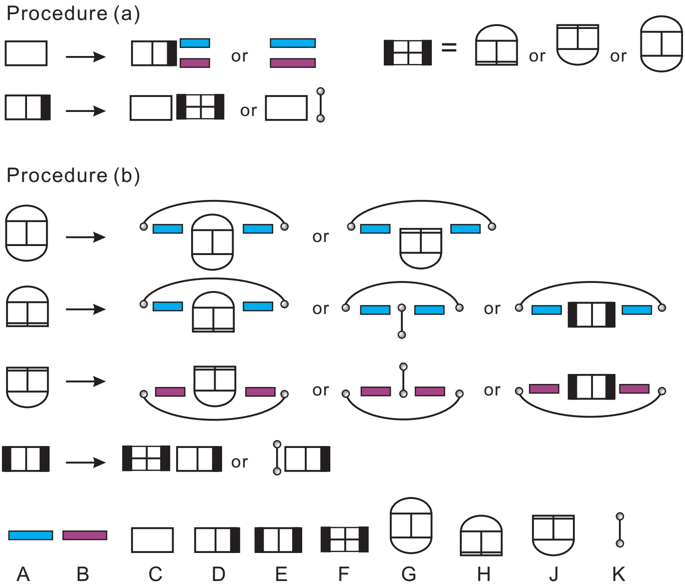
enclosed by a pair in one or both sequences

can be reduced by "arc removal"

We need to think about case 3: which of the two arcs?

→ define preference for the upper arc

An unambiguous grammar

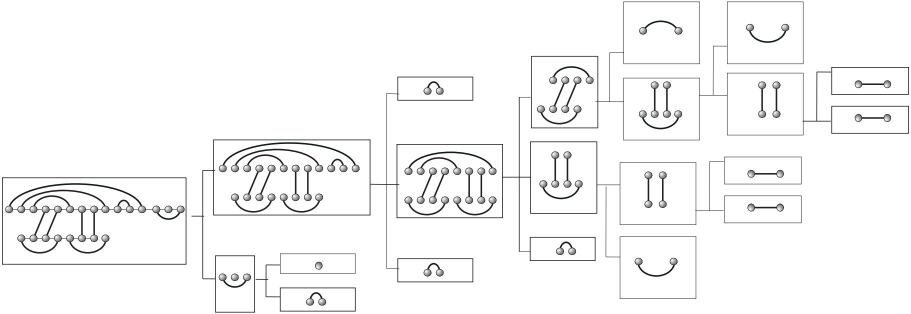


Procedure (a)

Procedure (b)

Decomposition Tree

Example for a parse tree:

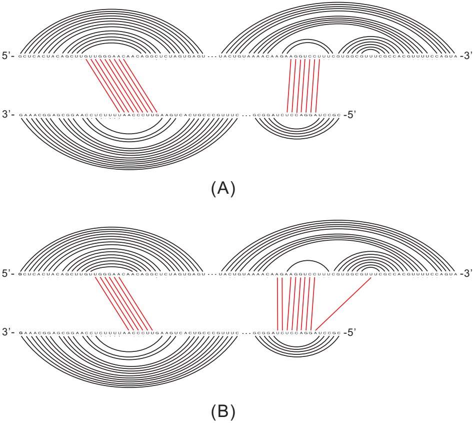


Implementation

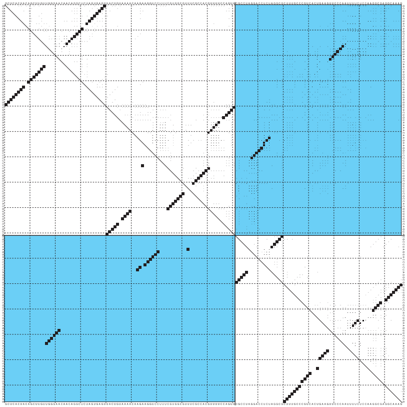
- Ugly but doable:
  $16 + 24 + 18 + 15 = 73$ fourdimensional arrays
- $\mathcal{O}(n^6)$ time and $\mathcal{O}(n^4)$ memory
- most of the effort is necessary to determine WHERE the likely interactions are. Much cheaper to compute the interaction energy only.

A similar approach has been taken by Rolf Backenofen, Cenk Sahinalp and their collaborators.

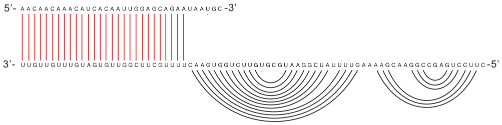
OxyS/fhlA



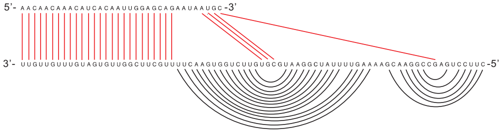
5'- GCUCACUACAGCUUGUUGGGAACAACAGGCUCUAGUGAGU ... UACUGUAAAACAAGAAGGUCCUUUCGUGGCGUUUCGCCACGUUUUCCAGUA -3'
3'- GAAACGGAGCGGAACCUCUUUUAACCCUUGAAGUCACUGCCCGUUUC ... GCGGAUCUCCAGGAUCCGC -5'
(A)
5'- GCUCACUACAGCUUGUUGGGAACAACAGGCUCUAGUGAGU ... UACUGUAAAACAAGAAGGUCCUUUCGUGGCGUUUCGCCACGUUUUCCAGUA -3'
3'- GAAACGGAGCGGAACCUCUUUUAACCCUUGAAGUCACUGCCCGUUUC ... GCGGAUCUCCAGGAUCCGC -5'
(B)



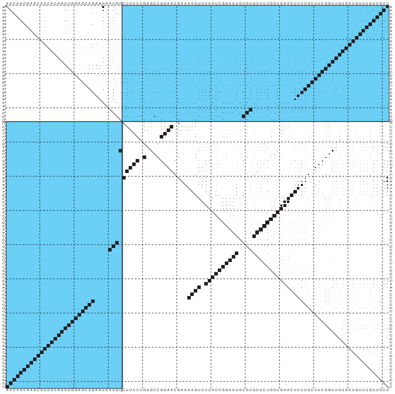
gcvB/dppA


5'- AACAACAAACAUCACAAUUGGAGCAGAAUAAUGC -3'
3'- UUGUUGUUUGUAGUGUUGGCUUCGUUUUCAAGUGGUCUUGUGCGUAAGGCUAUUUUGAAAAGCAAGGCCGAGUCCUUC -5'

(A)

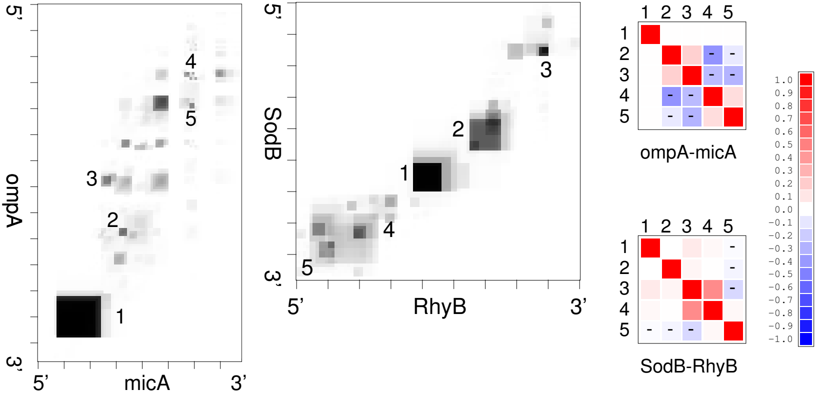

5'- AACAACAAACAUCACAAUUGGAGCAGAAUAAUGC -3'
3'- UUGUUGUUUGUAGUGUUGGCUUCGUUUUCAAGUGGUCUUGUGCGUAAGGCUAUUUUGAAAAGCAAGGCCGAGUCCUUC -5'

(B)


AACAACAAACAUCACAAUUGGAGCAGAAUAAUGCCUUCCUGAGCCGGAACGAAAAGUUUUAUCGGAAUGCGUGUUCUGGUGAACUUUUGCUUCGGUUGUGAUGUUUGUUGUU

Interaction Regions

Probability $\pi_{i,j}$ that the basepair i, j is contained in an interacting region



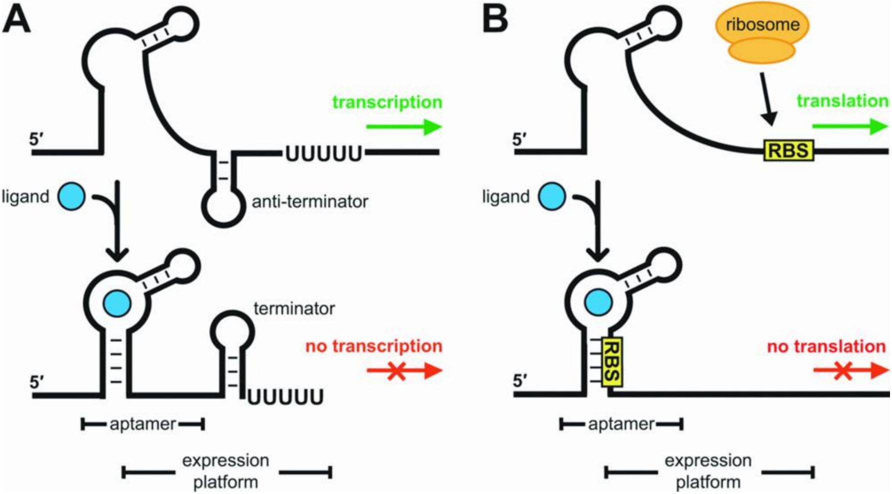
... and correlations between them

Design of Artificial Riboswitches

- Riboswitches are a convenient gadget in synthetic biology
- Task: combine ligand-specific sensor with an effector (i.e., some form of a regulatory element)
- Question: to what extent is this really modular?
- Idea: use RNA structure prediction to model the interplay of sensor and effector

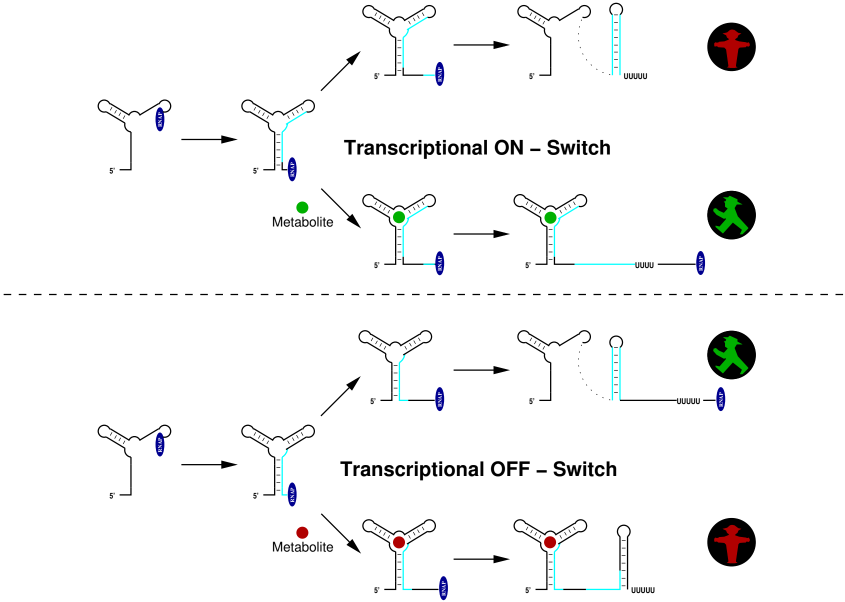
Riboswitches: Regulators of Gene Expression

Transcriptional *versus* translational riboswitch



Kim & Breaker, *Biol. Cell* (2008)

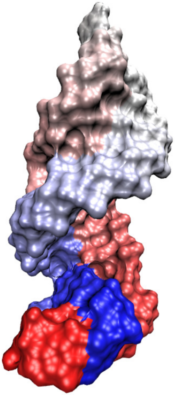
Ribo-Switching of Transcription



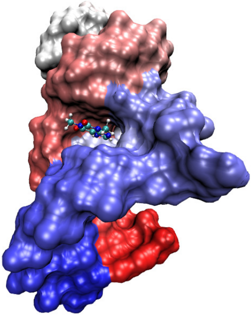
Transcriptional ON – Switch

Transcriptional OFF – Switch

Theophylline Aptamer

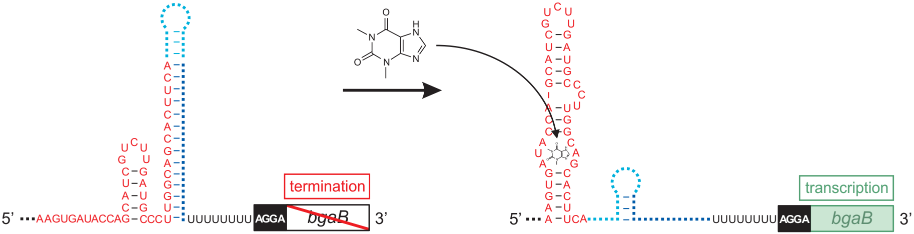


Unbound aptamer
Model predicted using Rosetta



Theophylline bound aptamer
Crystal Structure
(PDB-ID 1O15)

Design Idea


5' AAGUGAUACCAG CCCU UUUUUUUU AGGA bgaB 3'
termination
5' A UCA UUUUUUUU AGGA bgaB 3'
transcription

Goal: a theophylline triggered on-switch

Essence of the Multistable Design Problem

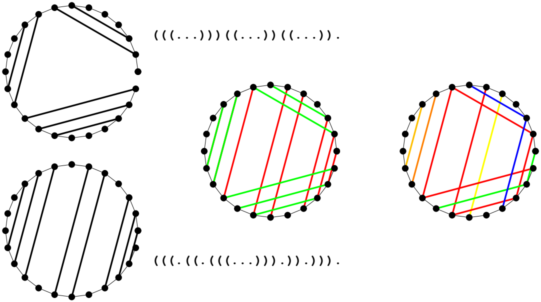
- Design a sequence that *compatible* with not just one but *several* target structures
- Each target should be almost a ground state
- **Questions:**
  - When can this be solved?
  - How can we include ligang specificity
- First step: generate sequences that are compatible with all design goals.
- 2nd step: optimize the sequences toward the design goal(s)

Bi-Stable Structures

Given two structures $\mathcal{S}_1$ $\mathcal{S}_2$, are there sequences compatible to both?
intersection theorem:

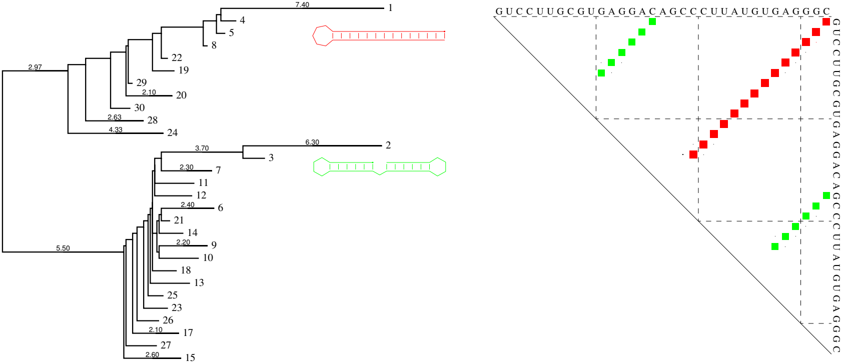
$$\mathbf{C}[\mathcal{S}_1] \cap \mathbf{C}[\mathcal{S}_2] \neq \emptyset$$

Proof: Dependency graph decomposes into paths and cycles of even length



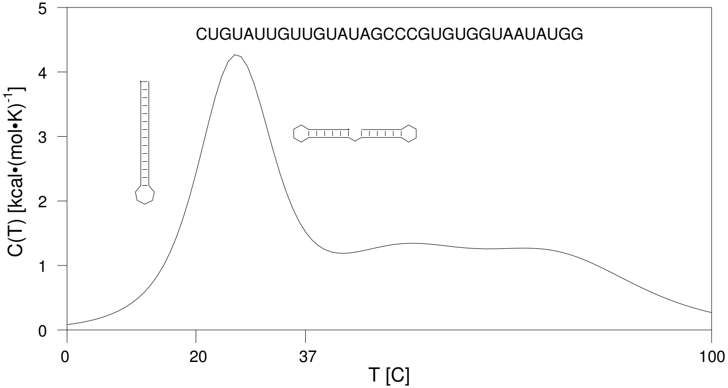
the alternating sequence `AUAUAU...` is compatible with each path and cycle.

Examples of bistable structures


7.40
1
4
5
8
22
19
29
2.10
20
30
2.97
2.63
28
4.33
24
3.70
6.30
2
3
2.30
7
11
12
2.40
6
21
14
2.20
9
10
18
13
25
23
26
2.10
17
27
5.50
2.60
15
G U C C U U G C G U G A G G A C A G C C C U U A U G U G A G G G C
G U C C U U G C G U G A G G A C A G C C C U U A U G U G A G G G C

$$\Xi(x) = E(x,\Omega_1) + E(x,\Omega_2) - 2G(x) + \xi\left(E(x,\Omega_1) - E(x,\Omega_2)\right)^2$$

A thermometer-like structure



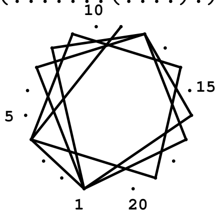
$$\Xi(x) = \left(E_{T_1}(x,\Omega_1) - G_{T_1}(x)\right) + \left(E_{T_2}(x,\Omega_2) - G_{T_2}(x)\right) \\ + \xi\left\{\left(E_{T_1}(x,\Omega_1) - E_{T_1}(x,\Omega_2)\right) + \left(E_{T_2}(x,\Omega_2) - E_{T_2}(x,\Omega_1)\right)\right\}$$

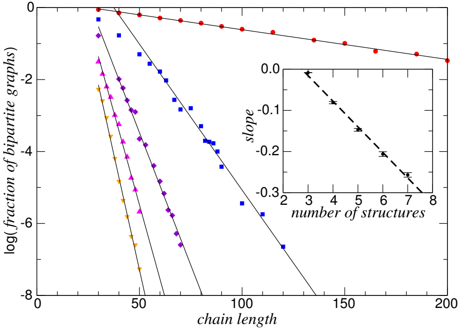
Multi-Stable Structures

Generalization to multiple Targets:

Theorem. There is a sequence satisfying each secondary structure constraints S_1, S_2, …, S_M if and only if the overlap graph $S_1 \cup S_2 \cup \cdots \cup S_M$ is bipartite.

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





Solving Multi-Constraint Design Problems

- one possibility: constraint programming [Dotu's work]
- stochastic heutristics
  - Complex search space. Only $\mathbf{C} := \bigcap_{i=1}^{M} \mathbf{C}(\Omega_i)$ allowed
  - How to choose a good (fair) starting position?
    simple for $M = 2$: constraints are path and cycles. Simple recursions to sample uniformly from $\mathbf{C}$
  - Difficult for $M > 2$: need more complex descompositions of graphs

Fibonacci numbers and bistable RNAs

$p_P(k;\mathrm{X}|\mathrm{Y})$ ... probability of sampling X after a path of length k if the other end is Y

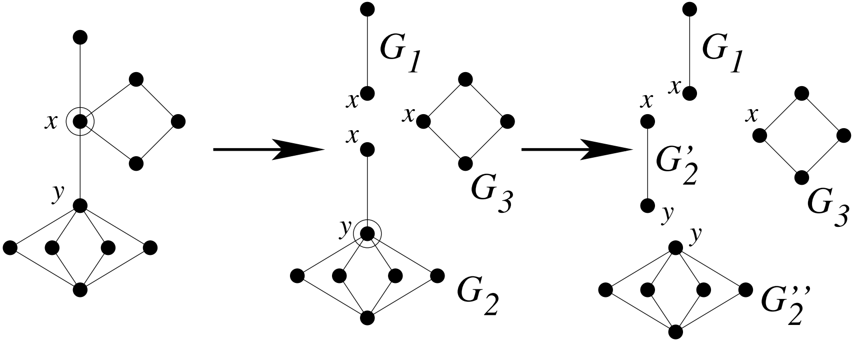
$$\begin{aligned} p_P(k;\mathrm{G}|\mathrm{U}) = p_P(k;\mathrm{U}|\mathrm{G}) &= \frac{\mathrm{Fib}(n-k+1)}{\mathrm{Fib}(n-k+2)} \\ p_P(k;\mathrm{A}|\mathrm{U}) = p_P(k;\mathrm{C}|\mathrm{G}) &= \frac{\mathrm{Fib}(n-k)}{\mathrm{Fib}(n-k+2)} \end{aligned} \tag{2}$$

Cyles are like path that are 1 vertex shorted.

Flamm et al, RNA 2001

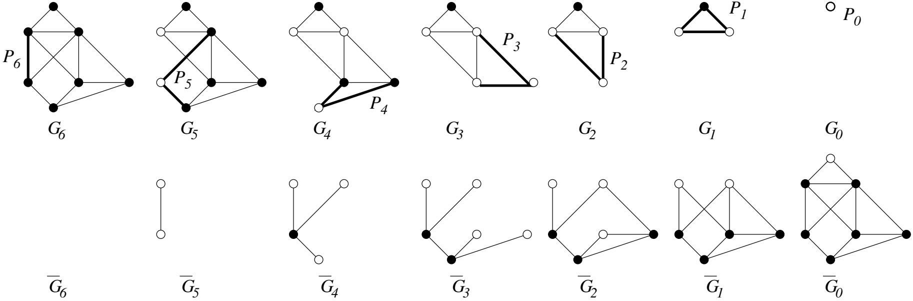
The General case

- First Step: block decomposition of the overlap graph.
  Color every block separately with fixed colors at the cut points



Coloring dense blocks: Ear decomposition

ear decomposition

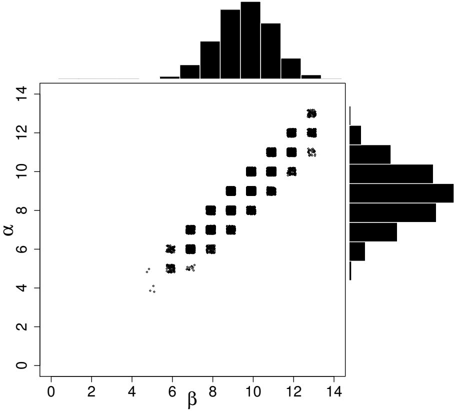


complement graph with attachment vertices

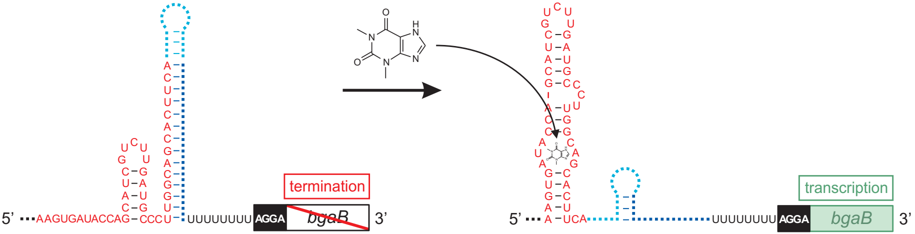
- dynamic programming approach to count colorings with given color combinations at the attachment vertices.
- memory exponential in the maximum number of attachment vertices α, CPU time in the maximum size of the union of attachment vertices in consecutive steps β

Coloring dense blocks: DP

computational effort depends strongly on the ear decomposition



Back to the Theophylline Switch



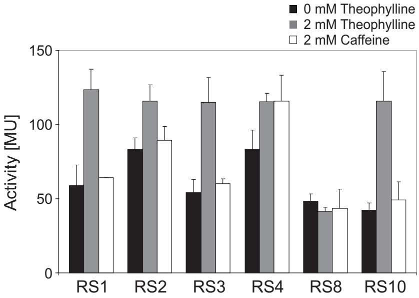
Goal: a theophylline triggered on-switch

Designed Theophylline Switches

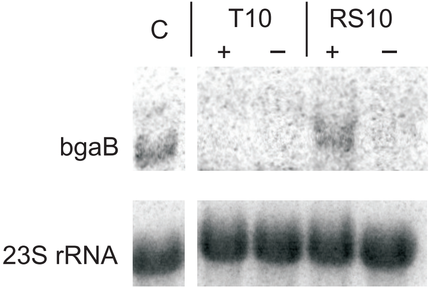
	sensor spacer 3'-part terminator U stretch	Energy RS (kcal/mol)	Energy T (kcal/mol)
RS1	AAGUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCACUUCAUUACAUCUGAAGUGCUGCCUUUUUUUU ...........(((((.....)))))....((((((((((((.......))))))))))))........ (((((...((((((((.....)))))...)))...)))))((((........)))).............	-27.4 -13.1	-21.0
RS2	AAGUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCACUUCAUGAUCUCGCUUGAAGUGCUGCUUUUUUUU ...........(((((.....)))))....((((((((((((..........))))))))))))....... (((((...((((((((.....)))))...)))...)))))........((((...))))............	-26.0 -14.1	-19.7
RS3	AAGUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCACUUCAUUUACAUACUCGGUAAACUGAAGUGCUGCCAUUUUUUUU ...........(((((.....)))))...((((((((((((((((((.......))))).)))))))))))))........ (((((...((((((((.....)))))...)))...)))))......(((((((((....))).)))).))...........	-32.5 -16.7	-25.8
RS4	AAGUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCACUUCAAACCGAAAUUUGCGCUUGAAGUGCUGCUUUUUUUU ...........(((((.....)))))....(((((((((((((..((.......)).))))))))))))))....... (((((...((((((((.....)))))...)))...)))))......((((...(((((...)))))...))))....	-26.9 -17.3	-20.6
RS8	AAGUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCACUUCACUCCUAGUGGAGUGAAGUGCUGUUUUUUUU ........((((((((.....)))))...)))((((((((((((((....))))))))))))))........ (((((...((((((((.....)))))...)))...)))))((((((....)))))).................	-35.4 -22.2	-29.0
RS10	AAGUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCACUUCAGAAAUCUCUGAAGUGCUGUUUUUUUU ........((((((((.....)))))...)))((((((((((((....))))))))))))........ (((((...((((((((.....)))))...)))...)))))((((....))))................	-28.3 -15.1	-21.9

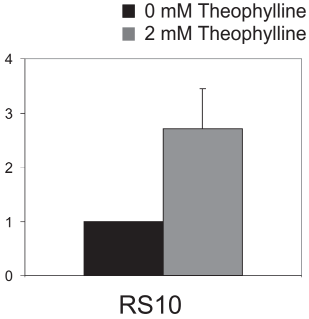
Construct Expression

```
       sensor        spacer             3'-part terminator  poly(U)
RS1:   aptamer-UUACAUC------------UGAAGUGCUGCC--UUUUUUUU

RS2:   aptamer-UGAUCUCGCU---------UGAAGUGCUGC---UUUUUUUU

RS3:   aptamer-UUUACAUACUCGGUAAAC-UGAAGUGCUGCCA-UUUUUUUU

RS4:   aptamer-AACCGAAAUUUGCGCU---UGAAGUGCUGC---UUUUUUUU

RS8:   aptamer-CUCCUAGUGGAG-------UGAAGUGCUG----UUUUUUUU

RS10:  aptamer-GAAAUCUC-----------UGAAGUGCUG----UUUUUUUU
```



Transcriptional Switching





Northern blot of RS10 and terminator T10

A more principled way to include ligand binding

- Known/measured binding energy $-\varepsilon$ of the ligand to a particular structural motif Ψ necessary for binding.
- without ligand: we want structural feature Ω
- with ligand: we want structural feature Ψ
- Compute the partition function $Z[\Psi]$ over all structures with feature Ψ.

  Partition function over structures without feature Ψ is $Z[\neg\Psi] := (Z - Z[\Psi])/Z$.
- Binding distorts the ensembl of structure when the ligand is present: $Z_L = Z[\not\Psi] + Z[\Psi]\exp(-\varepsilon)$
- Objectives
  - without ligand: $p_0(\Omega) := Z[\Omega]/Z \to max$ and $p_0(\Psi)$ should be small
  - with ligand: $p_L(\Psi) := Z[\Psi]\exp(-\varepsilon)/Z_L \to \max$ and $p_L(\Omega)$ should be small.

    ...easy if Ψ and Ω are mutually exclusive, otherwise we also need the partition function $Z(\Omega \wedge \Psi)$.

Using Constraints ...

Modified folding algorithms that scores certain structures differently
$Z\{\psi; e\}$ scores loop(s) with bonus energies e
Key relationship:

$$\frac{[RNA \cdot L]}{[RNA]\,[L]} = K = \frac{Z\{\psi; e\}}{Z z_L}$$

Set $z_L = 1$ for a small molecular ligand and gauge the binding energies accordingly.

$$Z\{\psi; e\} \approx Z[\not\Psi] + Z[\Psi]\exp(-\varepsilon)$$

in the more general model with (soft) constraints we may include a more elaborate parametrization that includes e.g. a set of variant binding site structures ...
Details of the theory still need to be developed ...

Folding Kinetics

RNA molecules may have kinetic traps which prevent them from reaching equilibrium within the lifetime of the molecule. Long molecules are often trapped in such meta-stable states during transcription.
Possible solutions are

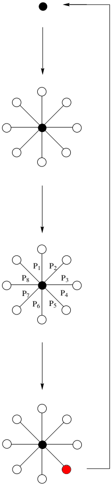
- Stochastic folding simulations can predict folding pathways and final structures. Computationally expensive, few programs available.
- Predicting structures for growing fragments of the sequence can show whether large scale re-folding will occur during transcription. Cheap but inaccurate.
- Analysis of the energy landscape based on complete suboptimal folding can identify possible traps (local minima).

Kinetic Folding Algorithm

Simulate folding kinetics by a Monte-Carlo type algorithm:
Generate all neighbors using the move-set

Assign rates to each move, e.g.

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



Select a move with probability proportional to its rate

Advance clock $1/\sum_i P_i$.

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 structure: opening and closing of base 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)

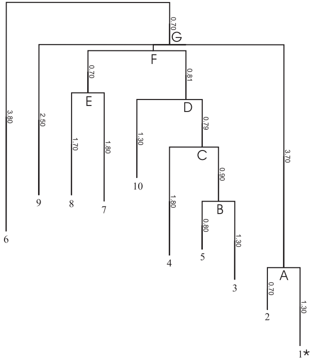
$$\mathrm{D} = \max\left\{ B(s) \middle| s \text{ is not a global minimum } \right\}$$

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

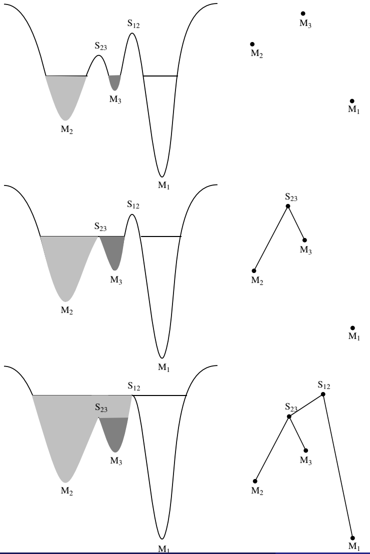
Energy Barriers and Barrier Trees

Some topological definitions:
A structure is a

- *local minimum* if its energy is lower than the energy of **all** neighbors
- *local maximum* if its energy is higher than the energy of **all** neighbors
- *saddle point* if there are at least two local minima that can be reached by a downhill walk starting at this point



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 $B(s)$

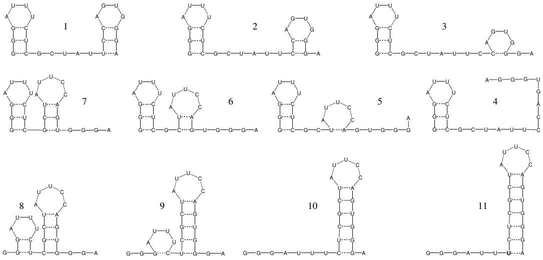
(c) if x has neighbors in several basins $B(s_1) \dots B(s_k)$ then it's a saddle point that *merges* these basins. Basins $B(s_1), \dots, B(s_k)$ are then united and are assigned to the deepest of local minimum.

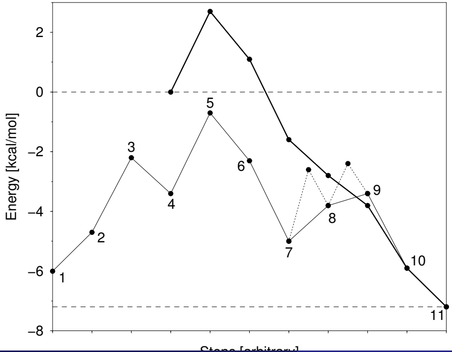
Information from the Barrier Trees

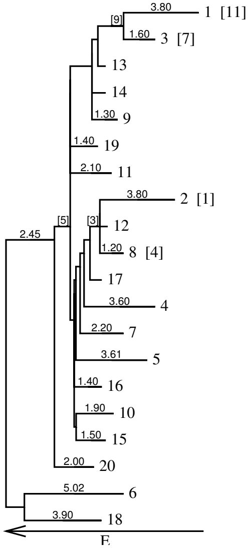
- Local minima
- Saddle points
- Barrier heights
- Gradient basins
- Partition functions and free energies of (gradient) basins
- Depth and Difficulty of the landscape

N.B.: 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







Folding Kinetics

Transition rates from x to y:

$$\begin{aligned} r_{yx} &= r_0 e^{-\frac{E^{\neq}_{yx} - E(x)}{RT}} \quad \text{for } x \neq y \\ r_{xx} &= -\sum_{y \neq x} r_{yx} \end{aligned}$$

Kinetics as a Markov process:

$$\frac{\mathrm{d}p_x}{\mathrm{d}t} = \sum_{y \in X} r_{xy} p_y(t)\,.$$

Transition states:

$$E^{\neq}_{yx} = \max\{E(x), E(y)\}$$

or more complex models (Tacker et al 1994, Schmitz et al 1996)

Reduced Description of the Folding Dynamics

Macrostates = Classes of a partition of the state space.
Partition function for a macro state:

$$Z_\alpha = \sum_{x\in\alpha} \exp(-E(x)/RT)$$

Free energy of a macro state:

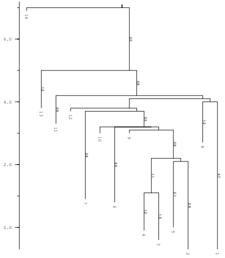
$$G(\alpha) = -RT \ln Z_\alpha$$

$$\begin{aligned} r_{\beta\alpha} &= \sum_{y\in\beta}\sum_{x\in\alpha} r_{yx}\mathrm{Prob}[x|\alpha] \quad \text{for } \alpha \neq \beta \\ &= \frac{1}{Z_\alpha}\sum_{y\in\beta}\sum_{x\in\alpha} r_{yx} e^{-E(x)/RT} \end{aligned}$$

$r_{\beta\alpha}$ "on flight" while executing the `barriers` program.
Transition state free energy:

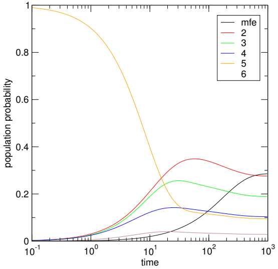
$$G^{\neq}_{\beta\alpha} = -RT \ln \sum_{y\in\beta}\sum_{x\in\alpha} e^{-\frac{E^{\neq}_{xy}}{RT}}$$

Different Approximations

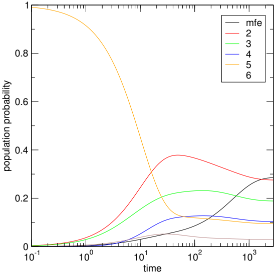


lilly

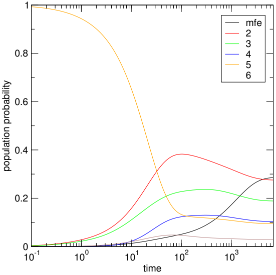
A simple model sequence



tree

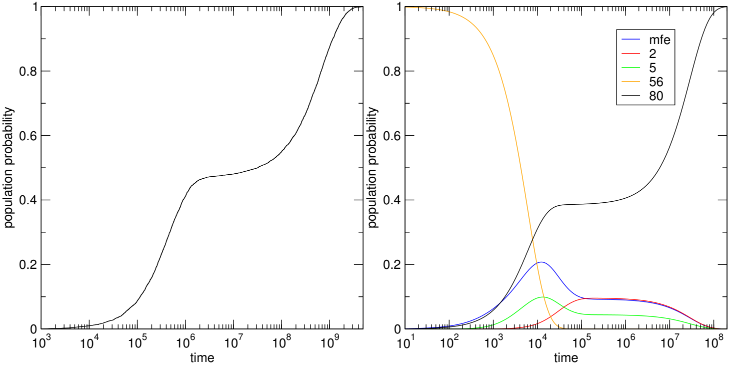


macro-states



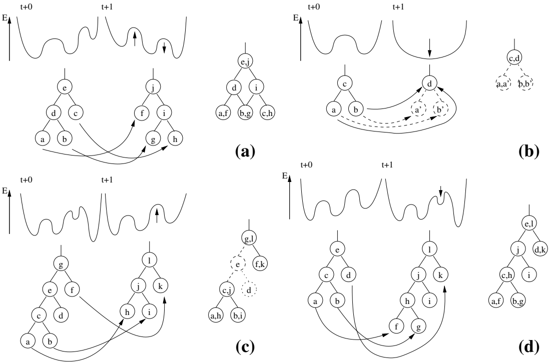
complete

Refolding of a tRNA



BarMaps: A static approximation

Idea: couple RNA folding with transcriptional chain elongation.
More general problem in the backgroud: perturbation of landscapes
Pragmatic approach: map local minima to local minima



Kinetic Folding of LONG Molecules

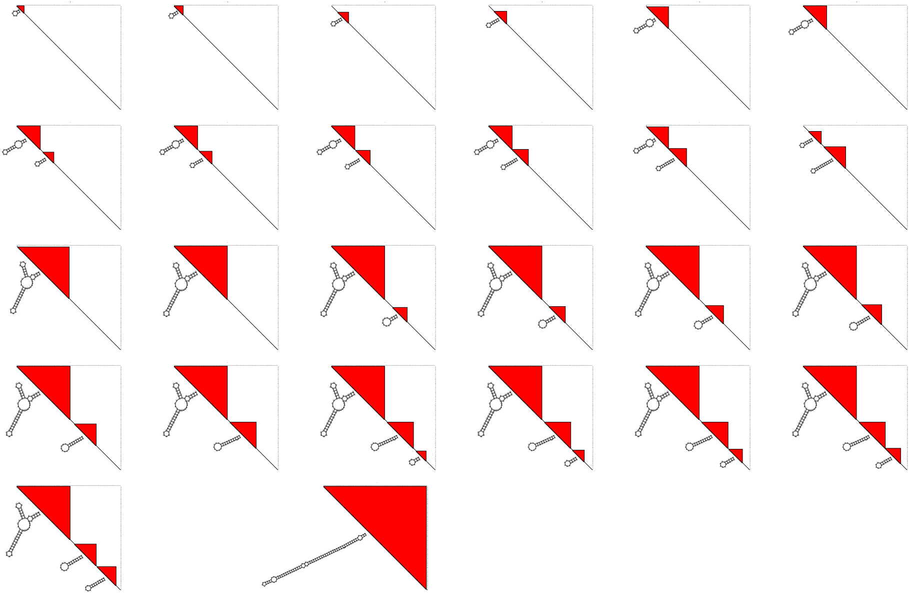
- For large RNAs ($N \gg 100$), direct Monte Carlo simulations become very slow
- Barrier Trees cannot be computed for $\gg 10^8$ low energy conformations
- Folding should be dominated by thermodynamically determined "domains"

Idea: "folding front" that progresses from local to more and more global interactions.

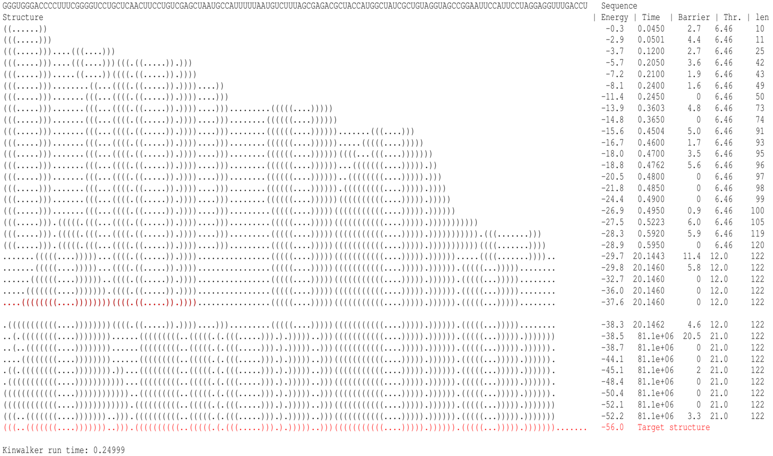
$\Longrightarrow$ `kinwalker`

Nice side effect: growing chains can easily be included as well

kinwalker at work



kinwalker: an example



```
GGGUGGGACCCCUUUCGGGGUCCUGCUCAACUUCCUGUCGAGCUAAUGCCAUUUUUAAUGUCUUUAGCGAGACGCUACCAUGGCUAUCGCUGUAGGUAGCCGGAAUUCCAUUCCUAGGAGGUUUGACCU   Sequence
Structure                                                                                                                        | Energy | Time   | Barrier | Thr. | len
((......))                                                                                                                          -0.3   0.0450      2.7   6.46     10
(((.....)))                                                                                                                         -2.9   0.0501      4.4   6.46     11
(((.....)))....(((....)))                                                                                                           -3.7   0.1200      2.7   6.46     25
(((.....)))....(((....)))(((.((.....)).)))                                                                                          -5.7   0.2050      3.6   6.46     42
(((.....))).....((....))((((.((.....)).))))                                                                                         -7.2   0.2100      1.9   6.46     43
(((.....)))........((...((((.((.....)).))))....))                                                                                   -8.1   0.2400      1.6   6.46     49
(((.....))).......(((...((((.((.....)).))))....)))                                                                                 -11.4   0.2450        0   6.46     50
(((.....))).......(((...((((.((.....)).))))....))).........(((((....)))))                                                          -13.9   0.3603      4.8   6.46     73
(((.....))).......(((...((((.((.....)).))))....)))........((((((....))))))                                                         -14.8   0.3650        0   6.46     74
(((.....))).......(((...((((.((.....)).))))....)))........((((((....)))))).......(((....)))                                        -15.6   0.4504      5.0   6.46     91
(((.....))).......(((...((((.((.....)).))))....)))........((((((....)))))).....(((((....)))))                                      -16.7   0.4600      1.7   6.46     93
(((.....))).......(((...((((.((.....)).))))....)))........((((((....))))))((((...(((....)))))))                                    -18.0   0.4700      3.5   6.46     95
(((.....))).......(((...((((.((.....)).))))....)))........((((((....))))))...(((((((....))))).))                                   -18.8   0.4762      5.6   6.46     96
(((.....))).......(((...((((.((.....)).))))....)))........((((((....))))))..((((((((....))))).)))                                  -20.5   0.4800        0   6.46     97
(((.....))).......(((...((((.((.....)).))))....)))........((((((....)))))).(((((((((....))))).))))                                 -21.8   0.4850        0   6.46     98
(((.....))).......(((...((((.((.....)).))))....)))........((((((....))))))((((((((((....))))).)))))                                -24.4   0.4900        0   6.46     99
(((.....))).......(((...((((.((.....)).))))....))).........(((((....)))))(((((((((((....))))).))))))                               -26.9   0.4950      0.9   6.46    100
(((.....))).(((((.(((...((((.((.....)).))))....))).........(((((....)))))(((((((((((....))))).)))))))))))                          -27.5   0.5223      6.0   6.46    105
(((.....))).(((((.(((...((((.((.....)).))))....))).........(((((....)))))(((((((((((....))))).))))))))))).(((.......)))            -28.3   0.5920      5.9   6.46    119
(((.....))).(((((.(((...((((.((.....)).))))....))).........(((((....)))))(((((((((((....))))).)))))))))))((((.......))))           -28.9   0.5950        0   6.46    120
.......(((((....)))))...((((.((.....)).))))................(((((....)))))(((((((((((....))))).)))))).....((((.......))))..         -29.7  20.1443     11.4  12.0     122
.......(((((....)))))...((((.((.....)).))))................(((((....)))))(((((((((((....))))).)))))).(((((...))))).........         -29.8  20.1460      5.8  12.0     122
......((((((....))))))..((((.((.....)).))))................(((((....)))))(((((((((((....))))).)))))).(((((...))))).........         -32.7  20.1460        0  12.0     122
.....(((((((....))))))).((((.((.....)).))))................(((((....)))))(((((((((((....))))).)))))).(((((...))))).........         -36.0  20.1460        0  12.0     122
....((((((((....))))))))((((.((.....)).))))................(((((....)))))(((((((((((....))))).)))))).(((((...))))).........         -37.6  20.1460        0  12.0     122

.(((((((((((....))))))))((((.((.....)).))))....))).........(((((....)))))(((((((((((....))))).)))))).(((((...))))).........         -38.3  20.1462      4.6  12.0     122
..(.((((((((....))))))))......(((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))))         -38.5   81.1e+06  20.5  21.0     122
..(..(((((((....))))))))......(((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))).         -38.7   81.1e+06     0  21.0     122
....((((((((....))))))))......(((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))).         -44.1   81.1e+06     0  21.0     122
.((.((((((((....)))))))).))...(((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))).         -45.1   81.1e+06     2  21.0     122
.(((((((((((....)))))))))))...(((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))).         -48.4   81.1e+06     0  21.0     122
((((((((((((....))))))))))))..(((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))).         -50.4   81.1e+06     0  21.0     122
((((((((((((....)))))))))))).((((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))))         -52.1   81.1e+06     0  21.0     122
(((..(((((((....)))))))..))).((((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))))         -52.2   81.1e+06   3.3  21.0     122
(((..(((((((....)))))))..))).((((((((((..(((((.(.(((.....))).).)))))..)))(((((((((((....))))).)))))).(((((...))))).)))))))....... -56.0   Target structure

Kinwalker run time: 0.24999
```

Folding pathway of MS2 A-protein 5'UTR. Kinwalker correctly identifies the “trap structure” described by Meerten *et al.* (red color)

Summary

- RNA structures can be computed efficiently by means of dynamic programming
- Computations are based on a set of carefully measures energy parameters and an additive energy model
- Algorithms exist for ground state energy and structure, full partition functions, density of states, interacting structures, . . .
- The folding kinetics of a given RNA Sequence can also be investigated as the level of secondary structures
- VIENNA RNA PACKAGE

Co-Conspirators

- Peter Schuster, Walter Fontana, Ivo L. Hofacker, Christoph Flamm
- Christian Höner zu Siederdissen, Felix Kühnl, Sebastian Will, Jör Fallmann, and many others in my lab
- Ronny Lorenz, the Master of Disaster and the Vienna RNA Package, and many others in Ivo's Group in Vienna
- Rolf Backofen and others in Freiburg
- Christian Reidys, Jing Qin, Fenix Huang at Virginia Tech
- Jan Gorodkin and his vikings @ RTH in København
- Daniel Gerighaus & Dirk Zeckzer (visualization)