Energy Landscapes and Dynamics of Biopolymers

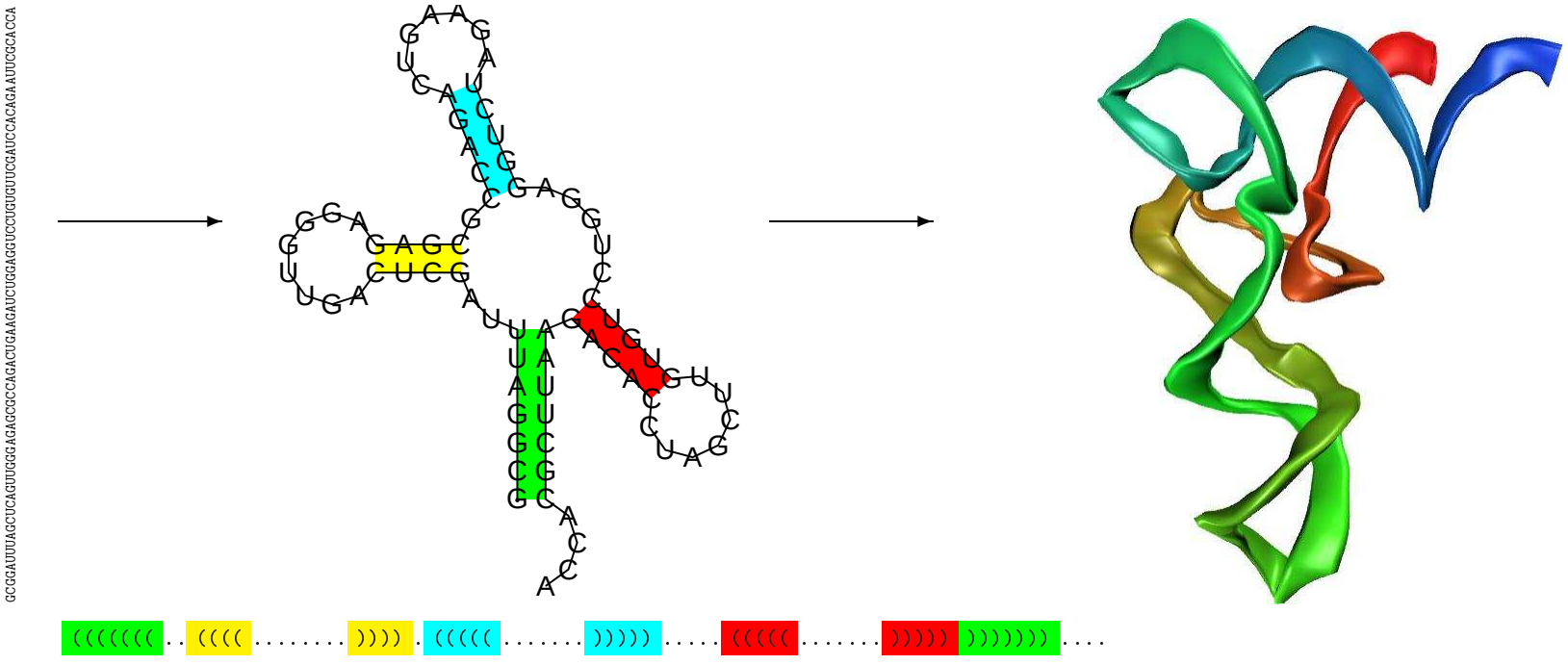
Michael Thomas Wolfinger

University of Vienna

5 March 2012

Outline

The RNA model



A secondary structure is a list of base pairs that fulfills two constraints:

- A base may participate in at most one base pair.
- Base pairs must not cross, i.e., no two pairs (i,j) and (k,l) may have $i < k < j < l$. (no pseudo-knots)

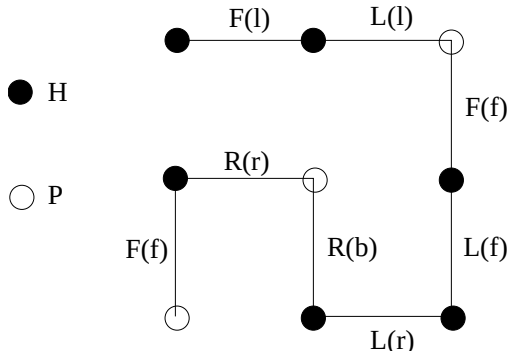
The optimal as well as the suboptimal structures can be computed recursively.

The HP-model

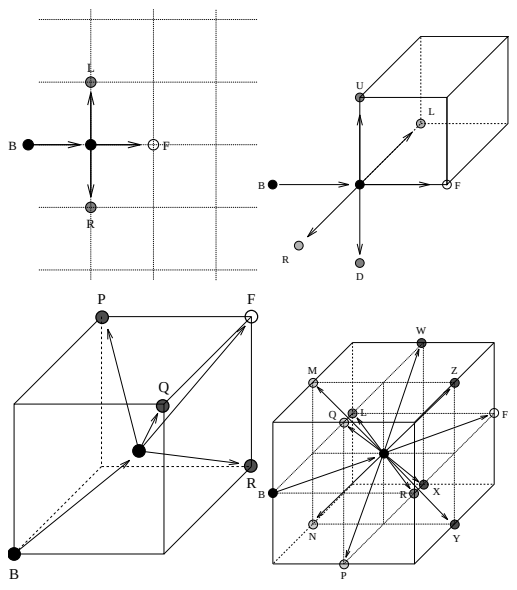
In this *simplified model*, a conformation is a *self-avoiding walk (SAW)* on a given lattice in 2 or 3 dimensions. Each bond is a straight line, bond angles have a few discrete values. The 20 letter alphabet of amino acids (monomers) is reduced to a two letter alphabet, namely **H** and **P**. H represents hydrophobic monomers, P represents hydrophilic or *polar* monomers.

Advantages:

- lattice-independent folding algorithms
- simple energy function
- hydrophobicity can be reasonably modeled



FRRLLFLF

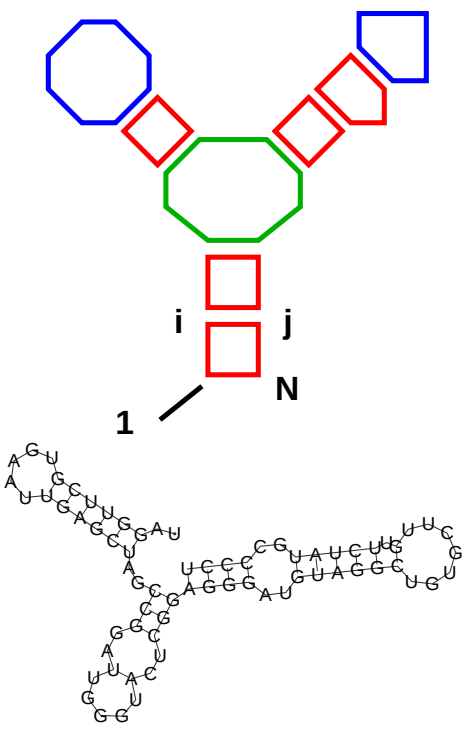


Energy functions

RNA

The standard energy model expresses the free energy of a secondary structure S as the sum of the energies of its loops l

$$E(S) = \sum_{l \in S} E(l)$$



$E = -17.5 kcal/mol$

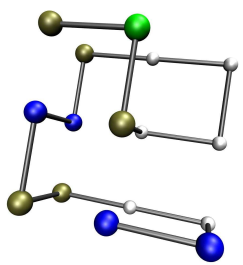
Lattice Proteins

The energy function for a sequence with n residues $\mathfrak{S} = \mathfrak{s}_1\mathfrak{s}_2\ldots\mathfrak{s}_n$ with $\mathfrak{s}_i \in \mathscr{A} = \{a_1, a_2, \ldots, a_b\}$, the alphabet of b residues, and an overall configuration $x = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ on a lattice $\mathscr{L}$ can be written as the sum of pair potentials

$$E(\mathfrak{S}, x) = \sum_{\substack{i < j-1 \\ |\mathbf{x}_i - \mathbf{x}_j| = 1}} \Psi[\mathfrak{s}_i, \mathfrak{s}_j].$$

	H	P
H	-1	0
P	0	0

	H	P	N	X
H	-4	0	0	0
P	0	1	-1	0
N	0	-1	1	0
X	0	0	0	0



$E = -16$

Folding landscape - energy landscape

The energy landscape of a biopolymer molecule is a complex surface of the (free) energy versus the conformational degrees of freedom.

Number of RNA secondary structures

$c_n \sim 1.86^n \cdot n^{-\frac{3}{2}}$

dynamic programming algorithms available

Number of LP structures

$c_n \sim \mu^n \cdot n^{\gamma-1}$

problem is NP-hard

dim	Lattice Type	μ	γ
2	SQ	2.63820	1.34275
	TRI	4.15076	1.343
	HEX	1.84777	1.345
3	SC	4.68391	1.161
	BCC	6.53036	1.161
	FCC	10.0364	1.162

Formally, three things are needed to construct an energy landscape:

- A set X of configurations
- an energy function $f : X \rightarrow \mathbf{R}$
- a symmetric neighborhood relation $\mathfrak{N} : X \times X$

Folding landscape - energy landscape

The energy landscape of a biopolymer molecule is a complex surface of the (free) energy versus the conformational degrees of freedom.

Number of RNA secondary structures

$$c_n \sim 1.86^n \cdot n^{-\frac{3}{2}}$$

dynamic programming algorithms available

Number of LP structures

$$c_n \sim \mu^n \cdot n^{\gamma-1}$$

problem is NP-hard

dim	Lattice Type	μ	γ
2	SQ	2.63820	1.34275
	TRI	4.15076	1.343
	HEX	1.84777	1.345
3	SC	4.68391	1.161
	BCC	6.53036	1.161
	FCC	10.0364	1.162

Formally, three things are needed to construct an energy landscape:

- A set X of configurations
- an energy function $f : X \to \mathbf{R}$
- a symmetric neighborhood relation $\mathfrak{N} : X \times X$

Folding landscape - energy landscape

The energy landscape of a biopolymer molecule is a complex surface of the (free) energy versus the conformational degrees of freedom.

Number of RNA secondary structures

$c_n \sim 1.86^n \cdot n^{-\frac{3}{2}}$

dynamic programming algorithms available

Number of LP structures

$c_n \sim \mu^n \cdot n^{\gamma-1}$

problem is NP-hard

dim	Lattice Type	μ	γ
2	SQ	2.63820	1.34275
	TRI	4.15076	1.343
	HEX	1.84777	1.345
3	SC	4.68391	1.161
	BCC	6.53036	1.161
	FCC	10.0364	1.162

Formally, three things are needed to construct an energy landscape:

- A set X of configurations
- an energy function $f : X \rightarrow \mathbf{R}$
- a symmetric neighborhood relation $\mathfrak{N} : X \times X$

Folding landscape - energy landscape

The energy landscape of a biopolymer molecule is a complex surface of the (free) energy versus the conformational degrees of freedom.

Number of RNA secondary structures

$c_n \sim 1.86^n \cdot n^{-\frac{3}{2}}$

dynamic programming algorithms available

Number of LP structures

$c_n \sim \mu^n \cdot n^{\gamma-1}$

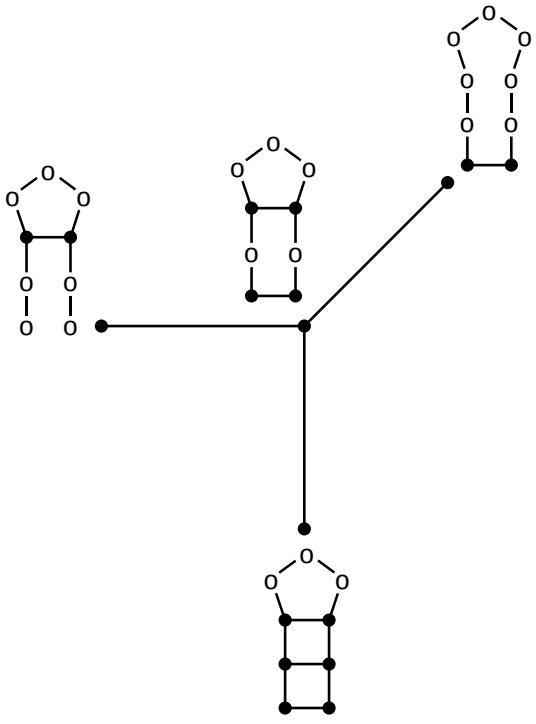
problem is NP-hard

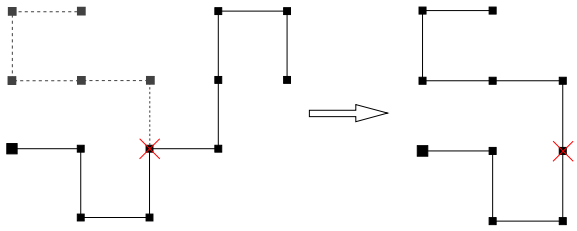
dim	Lattice Type	μ	γ
2	SQ	2.63820	1.34275
	TRI	4.15076	1.343
	HEX	1.84777	1.345
3	SC	4.68391	1.161
	BCC	6.53036	1.161
	FCC	10.0364	1.162

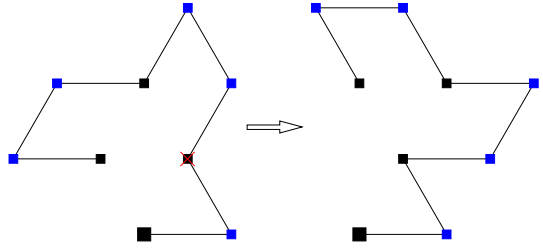
Formally, three things are needed to construct an energy landscape:

- A set X of configurations
- an energy function $f : X \rightarrow \mathbf{R}$
- a symmetric neighborhood relation $\mathfrak{N} : X \times X$

The move set







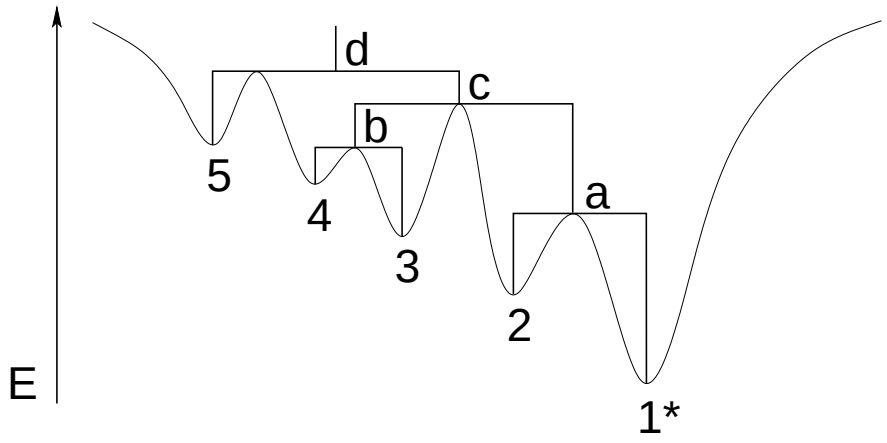
- For each move there must be an inverse move
- Resulting structure must be in X
- Move set must be *ergodic*

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 thar can be reached by a downhill walk starting at this point



We further define

- a walk between two conformations x and y as a list of conformations $x = x_1 \dots x_{m+1} = y$ such that $\forall 1 \leq i \leq m : \mathfrak{N}(x_i, x_{i+1})$
- the lower part of the energy landscape (written as $X^{\leq \eta}$) as *all* conformations x such that $E(\mathfrak{S}, x) \leq \eta$ (with a predefined threshold η).

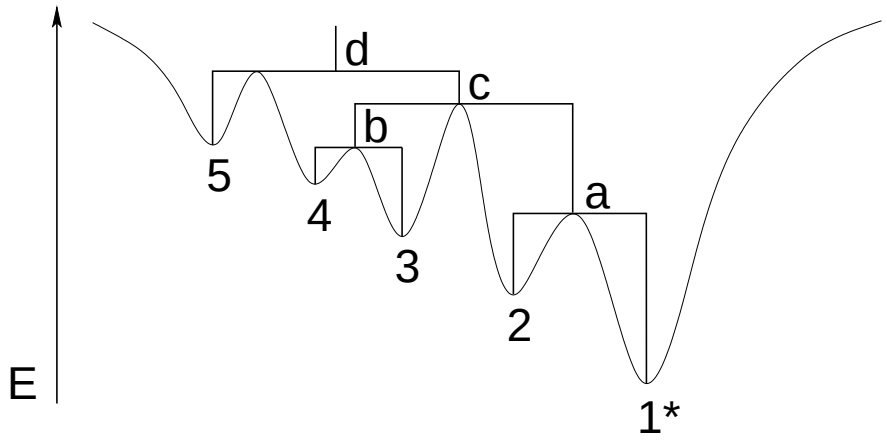
C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.
Barrier trees of degenerate landscapes.
Z. Phys. Chem., 216:155–173, 2002.

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 thar can be reached by a downhill walk starting at this point



We further define

- a walk between two conformations x and y as a list of conformations $x = x_1 \dots x_{m+1} = y$ such that $\forall 1 \leq i \leq m : \mathfrak{N}(x_i, x_{i+1})$
- the lower part of the energy landscape (written as $X^{\leq \eta}$) as *all* conformations x such that $E(\mathfrak{S}, x) \leq \eta$ (with a predefined threshold η).

C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.
Barrier trees of degenerate landscapes.
Z. Phys. Chem., 216:155–173, 2002.

The lower part of the energy landscape

Two conformations x and y are mutually accessible at the level η (written as $x \leftrightarrow\!\!\!\!\!\!\underline{\quad}^{\eta}\!\!\!\!\!\!\rightarrow y$) if there is a walk from x to y such that all conformations z in the walk satisfy $E(\mathfrak{S}, z) \leq \eta$. The saddle height $\hat{f}(x, y)$ of x and y is defined by

$$\hat{f}(x, y) = \min\{\eta \mid x \leftrightarrow\!\!\!\!\!\!\underline{\quad}^{\eta}\!\!\!\!\!\!\rightarrow y\}$$

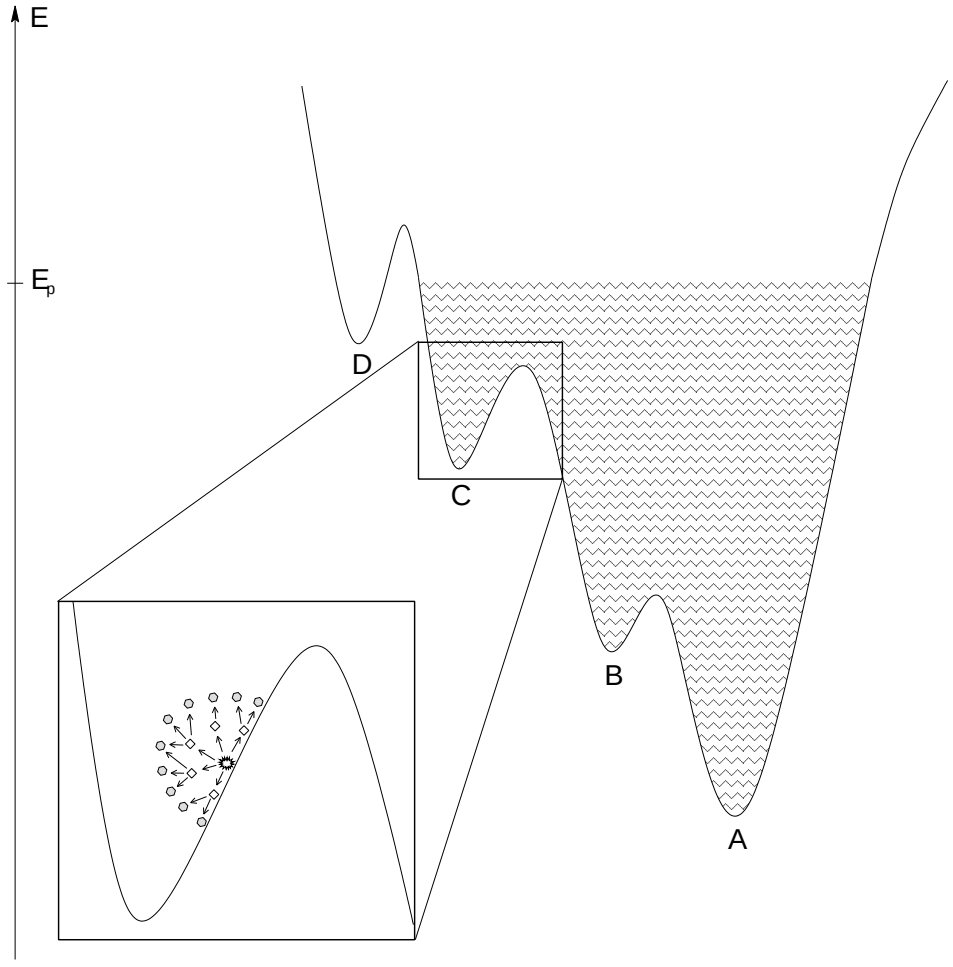
Given the set of all local minima $X_{\text{min}}^{\leq \eta}$ below threshold η, the lower energy part $X^{\leq \eta}$ of the energy landscape can alternatively be written as

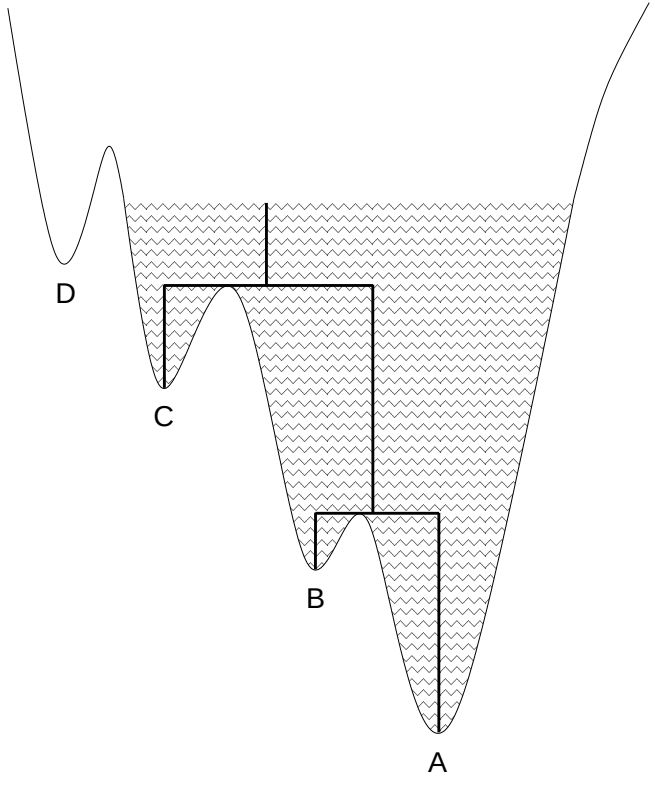
$$X^{\leq \eta} = \{y \mid \exists x \in X_{\text{min}}^{\leq \eta} : \hat{f}(x, y) \leq \eta\}$$

Given a restricted set of low-energy conformations, X_{init}, and a reasonable value for η, the lower part of the energy landscape can be calculated.

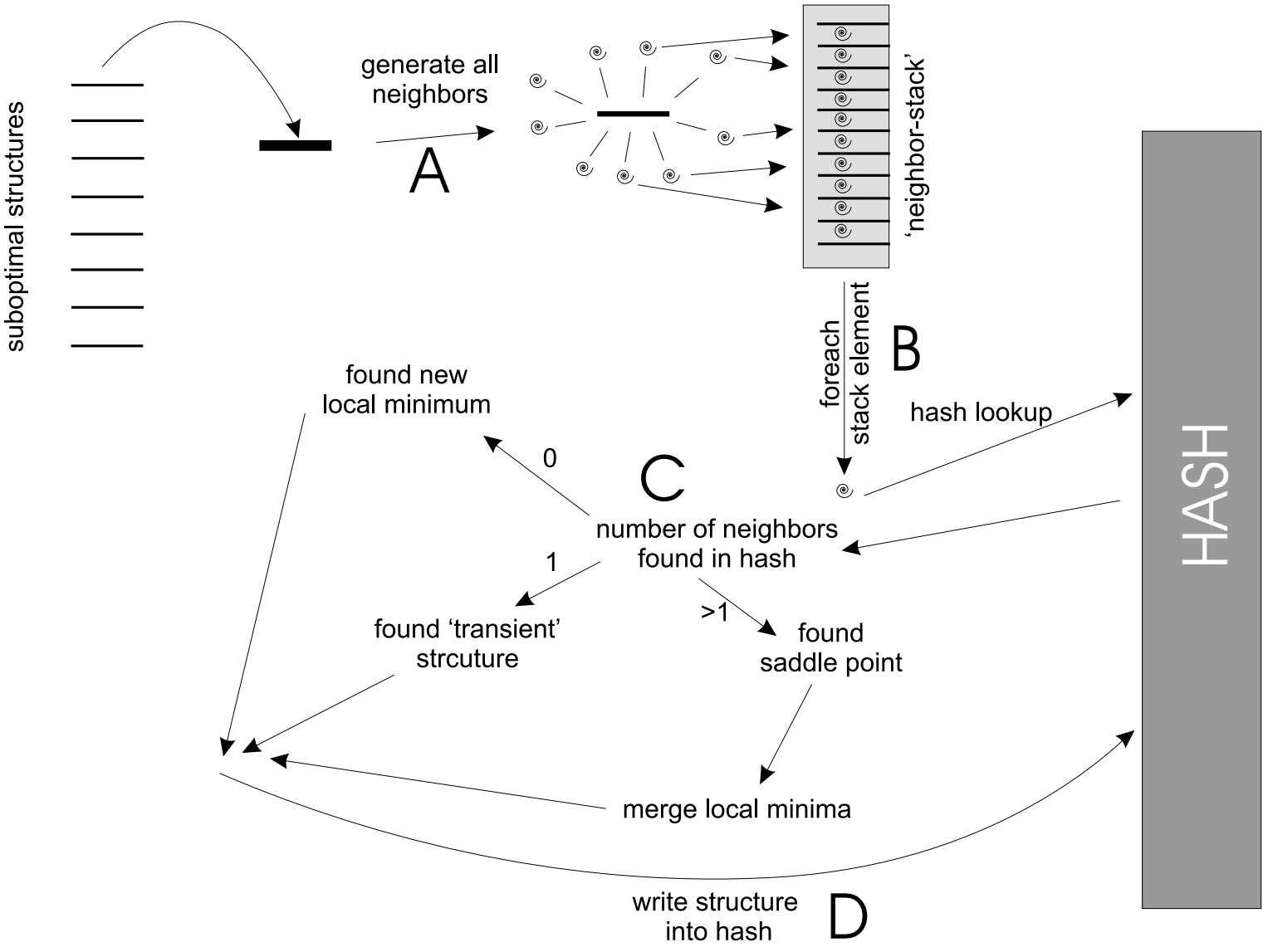
M. T. Wolfinger, S. Will, I. L. Hofacker, R. Backofen, and P. F. Stadler.
Exploring the lower part of discrete polymer model energy landscapes.
Europhys. Lett., 2006.

The Flooder approach





The concept of BARRIERS



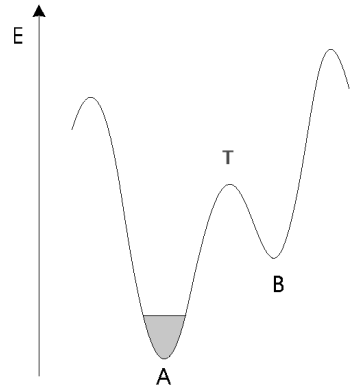
The algorithm of Barriers

Barriers

Require: all suboptimal secondary structures within a certain energy range from mfe

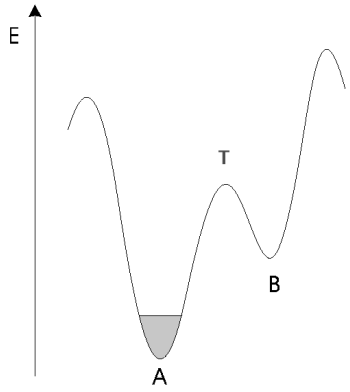
```
 1: ℬ ⇐ ∅
 2: for all x ∈ subopt do
 3:     𝒦 ⇐ ∅
 4:     𝒩 ⇐ generate_neighbors(x)
 5:     for all y ∈ 𝒩 do
 6:         if b ⇐ lookup_hash(y) then
 7:             𝒦 ⇐ 𝒦 ∪ b
 8:         end if
 9:     end for
10:     if 𝒦 = ∅ then
11:         ℬ ⇐ ℬ ∪ {x}
12:     end if
13:     if |𝒦| ≥ 2 then
14:         merge_basins(𝒦)
15:     end if
16:     write_hash(x)
17: end for
```

The flooding algorithm

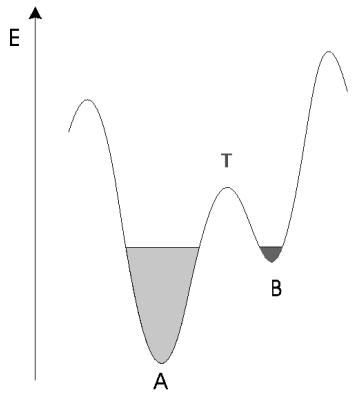


1

The flooding algorithm

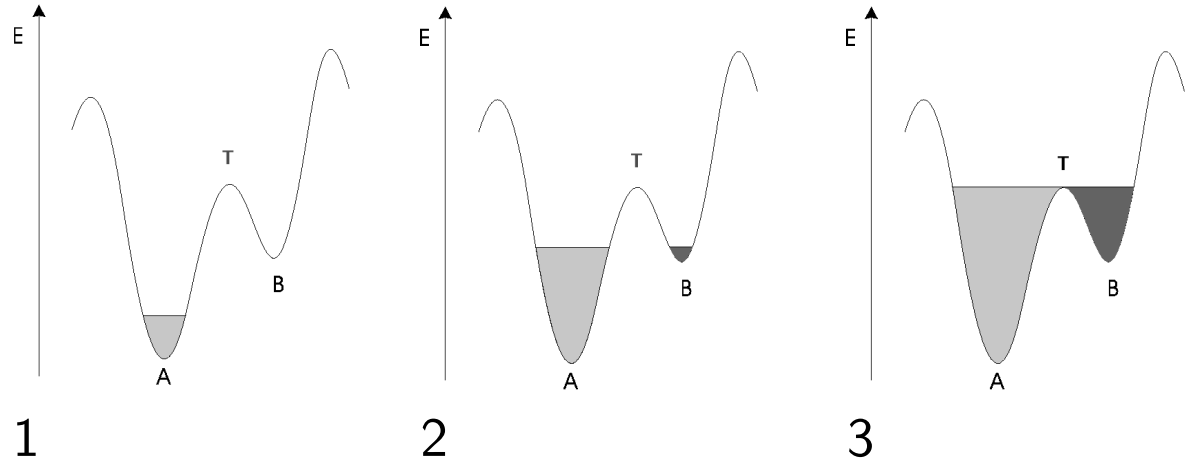


1

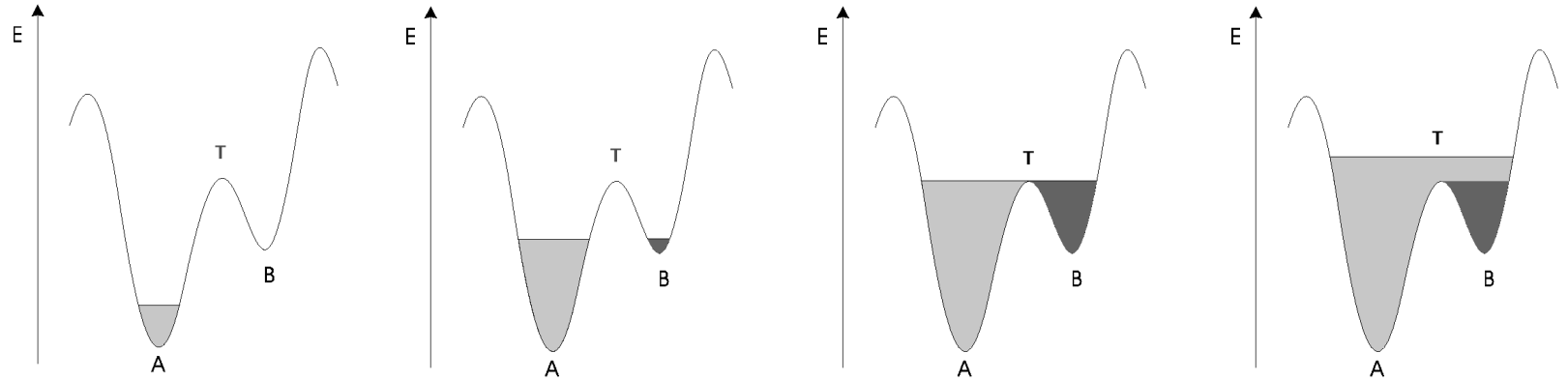


2

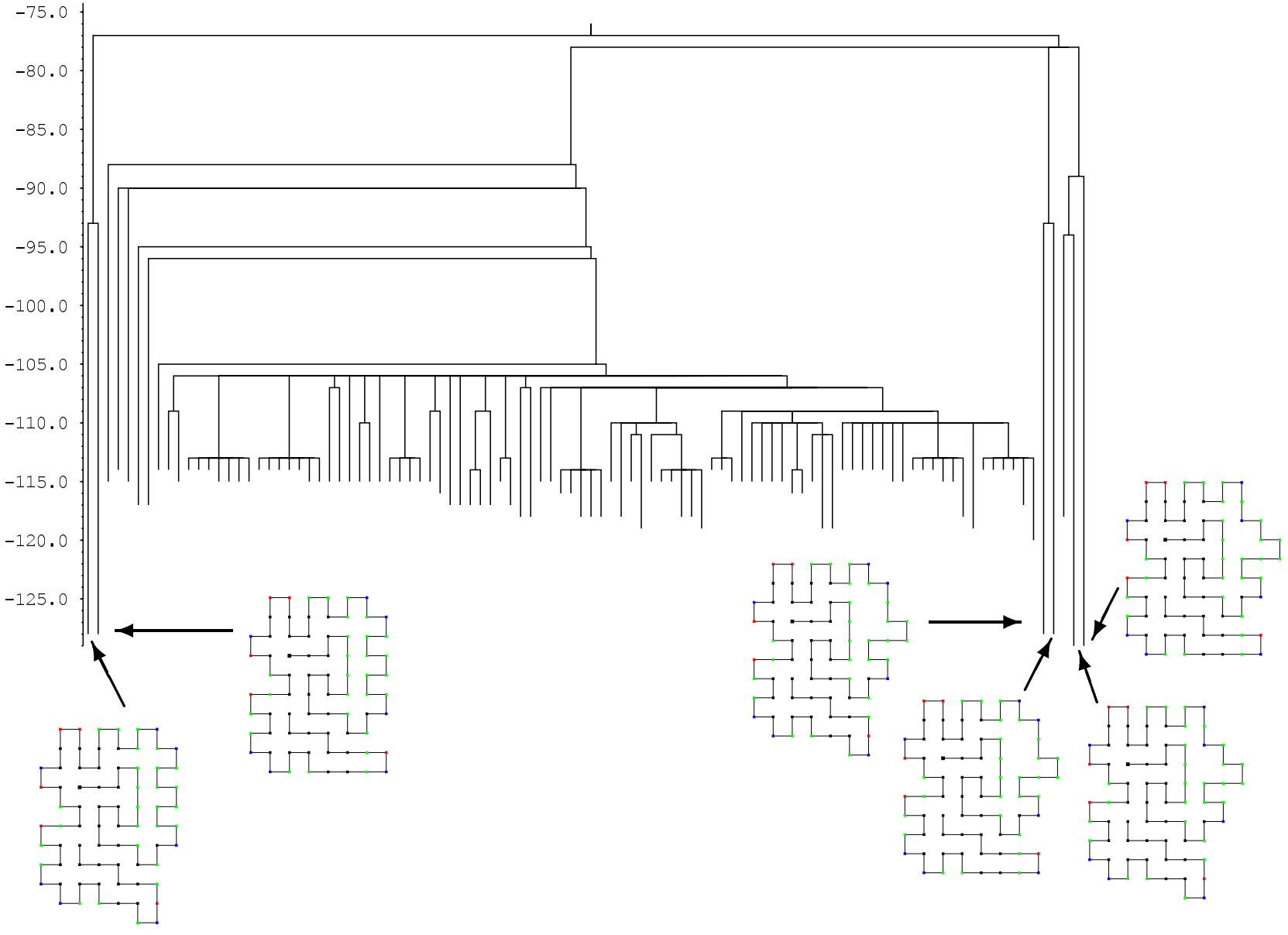
The flooding algorithm



The flooding algorithm



Barrier tree example



Information from the barrier trees

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

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

Folding kinetics

Biomolecules 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 (predict folding pathways)
- Predicting structures for growing fragments of the sequence
- Analysis of the energy landscape based on complete suboptimal folding

Biopolymer dynamics

The probability distribution P of structures as a function of time is ruled by a set of forward equations, also known as the master equation

$$\frac{dP_t(x)}{dt} = \sum_{y \neq x} [P_t(y)k_{xy} - P_t(x)k_{yx}]$$

Given an initial population distribution, how does the system evolve in time? (What is the population distribution after n time-steps?)

$$\frac{d}{dt}P_t = \mathbf{U}P_t \quad \Longrightarrow \quad P_t = e^{t\mathbf{U}}P_0$$

KINFOLD: A stochastic kinetic foling algorithm

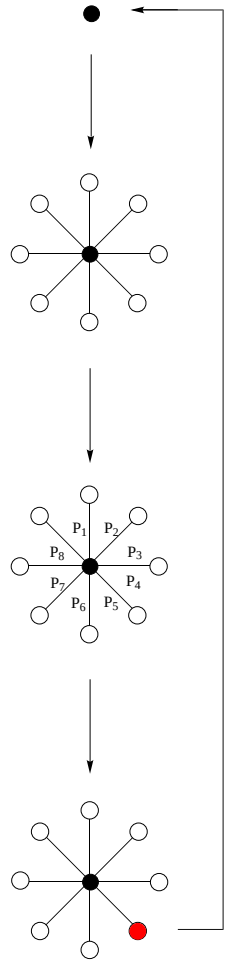
Simulate folding kinetics by a rejection-less Monte-Carlo type algorithm:

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 $1/\sum_i P_i$.



TREEKIN: Barrier tree kinetics

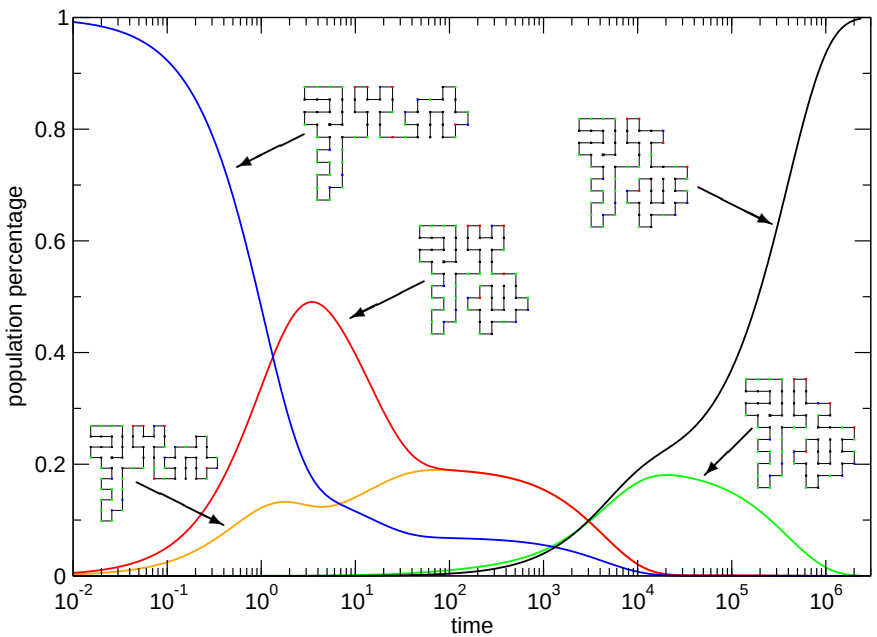
- Local minima
- Saddle points
- Barrier heights
- Gradient basins
- Partition functions
- Free energies of (gradient) basins

With this information, a reduced dynamics can be formulated as a Markov process by means of macrostates (i.e. basins in the barrier tree) and Arrhenius-like transition rates between them.

$$\frac{d}{dt}P_t = \mathbf{U}P_t \quad \Longrightarrow \quad P_t = e^{t\mathbf{U}}P_0$$

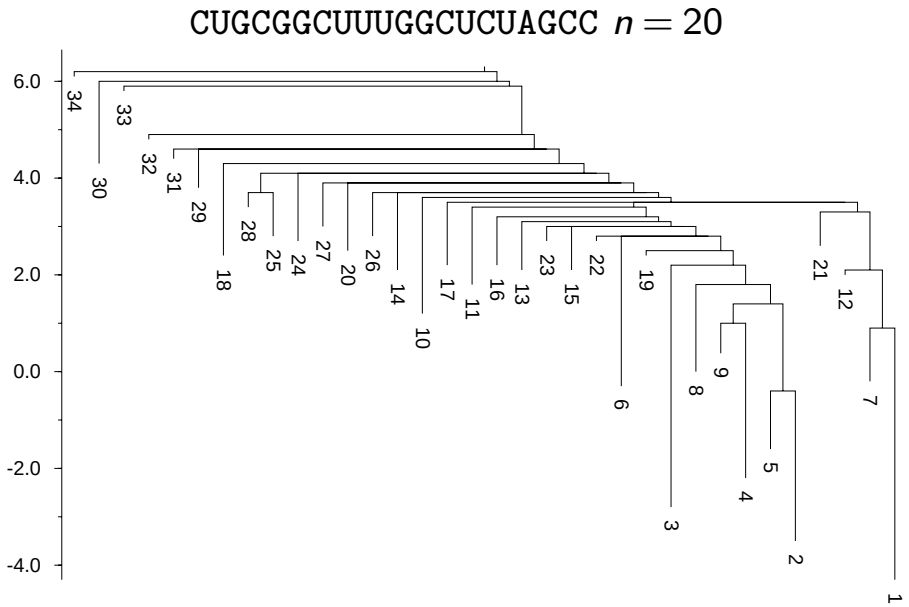
- macro-states form a partition of the full configuration space
- transition rates between macro-states

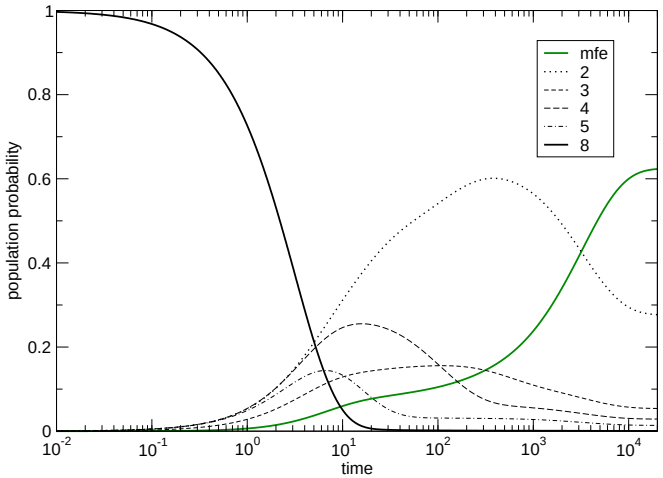
$$r_{\beta\alpha} = \Gamma_{\beta\alpha} \exp\left(-(E^*_{\beta\alpha} - G_\alpha)/kT\right)$$



M. T. Wolfinger, W. A. Svrcek-Seiler, C. Flamm, I. L. Hofacker, and P. F. Stadler.
Efficient computation of RNA folding dynamics.
J. Phys. A: Math. Gen., 37(17):4731–4741, 2004.

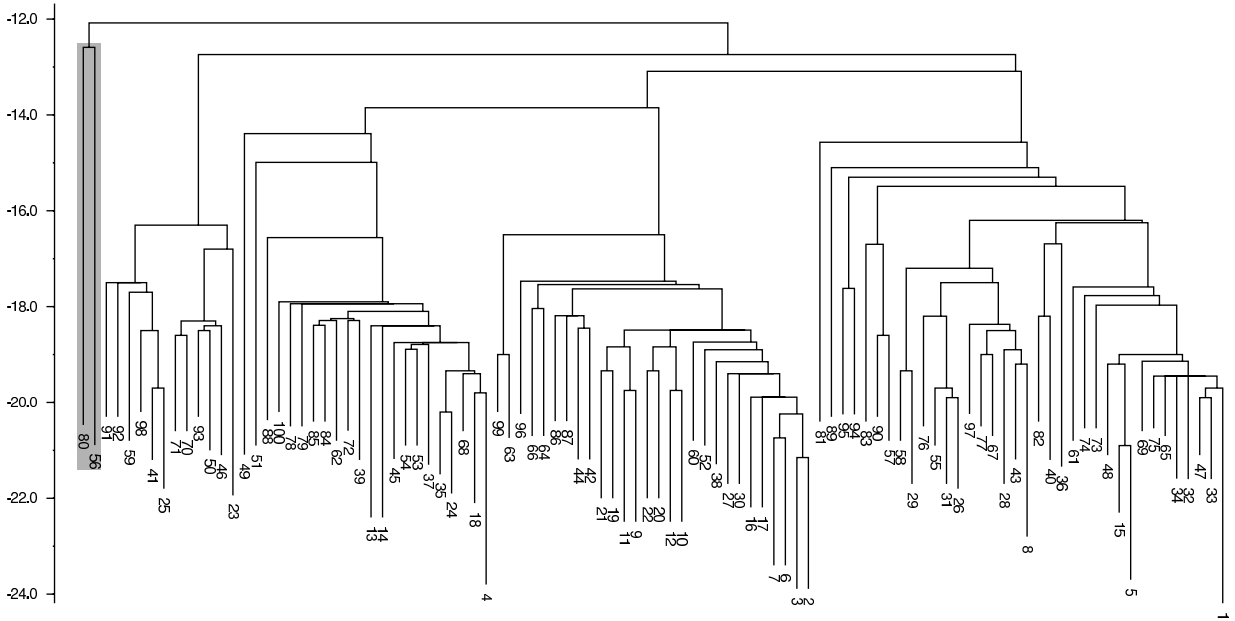
Dynamics of a short artificial RNA

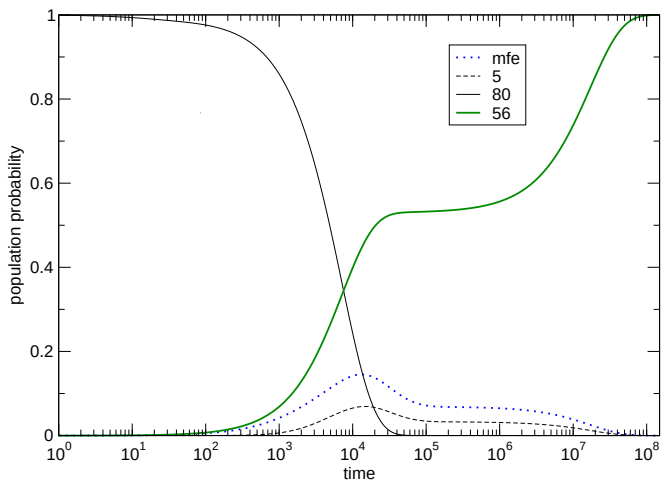


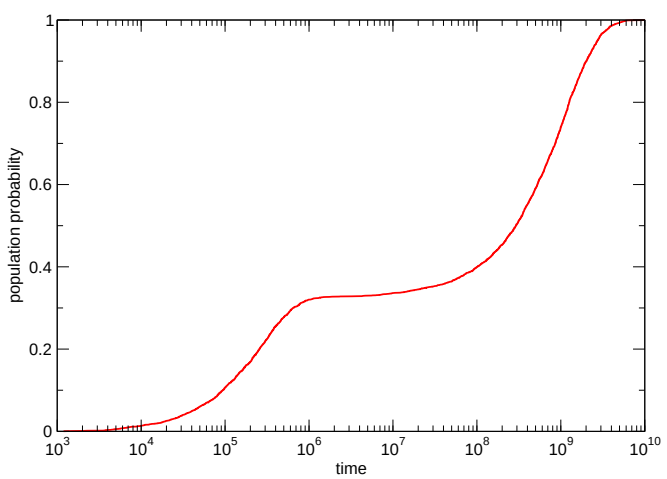


Dynamics of tRNA

GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAUCUGGAGGUCCUGUGTPCGAUCCACAGAAUUCGCACCA

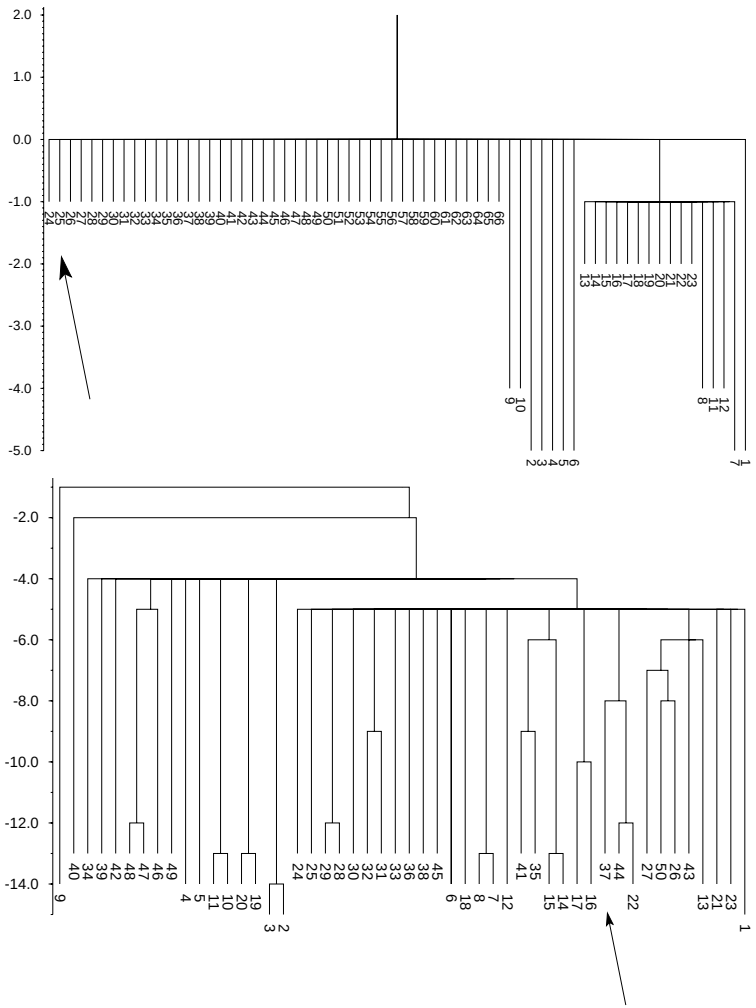


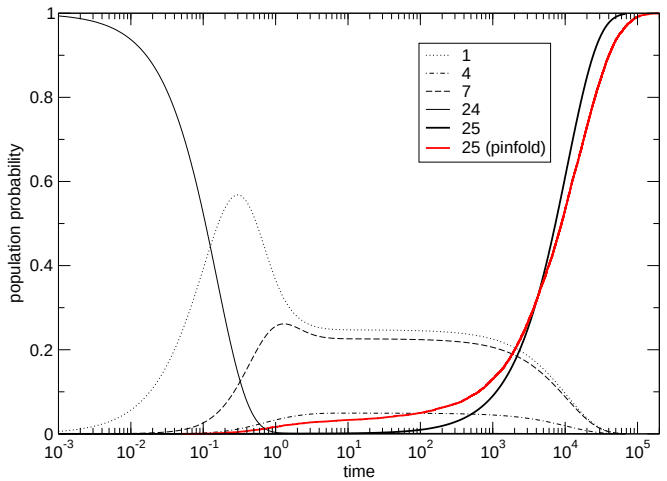


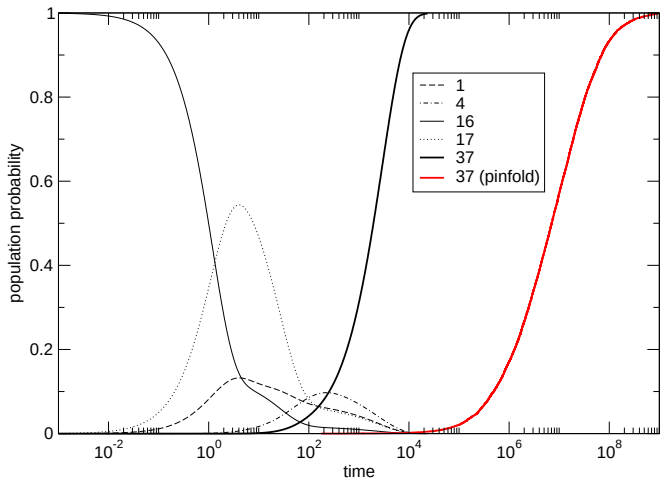


Dynamics of lattice proteins: HEX/TET lattice

NNHHPPNNPHHHHPXP $n = 16$







Barrier tree kinetics - problems and pitfalls

The method works fine for moderately sized systems.

Currently, we consider approx. 100 million structures within a single run of BARRIERS to calculate the topology of the landscape.

However, we are interested in larger systems:

- biologically relevant RNA switches
- large 3D lattice proteins

The next steps:

- use high-level diagonalization routines for sparse matrices
- calculate low-energy structures
- sample (thermodynamics properties of) individual basins
- sample low-enery refolding paths

Barrier tree kinetics - problems and pitfalls

The method works fine for moderately sized systems.

Currently, we consider approx. 100 million structures within a single run of BARRIERS to calculate the topology of the landscape.

However, we are interested in larger systems:

- biologically relevant RNA switches
- large 3D lattice proteins

The next steps:

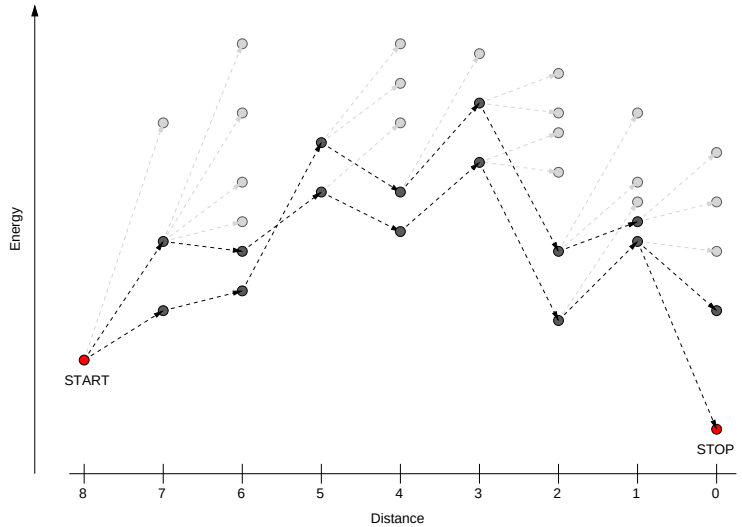
- use high-level diagonalization routines for sparse matrices
- calculate low-energy structures
- sample (thermodynamics properties of) individual basins
- sample low-enery refolding paths

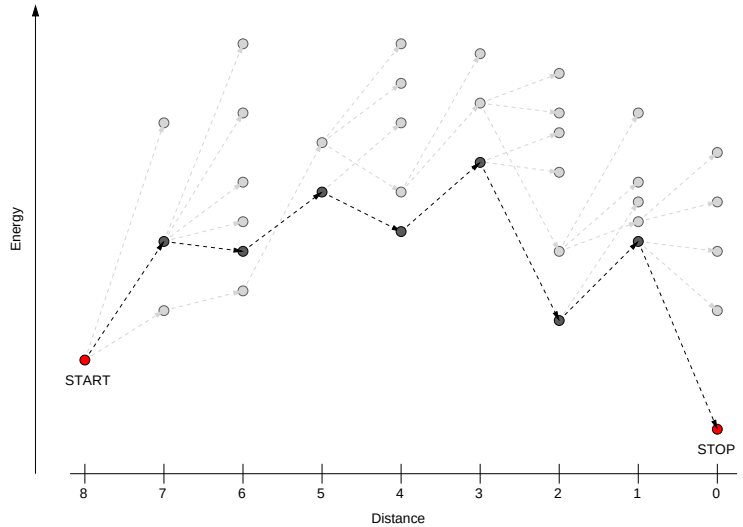
The PathFinder tool

A heuristic approach to efficiently estimate low-energy refolding paths

Overall procedure for direct paths:

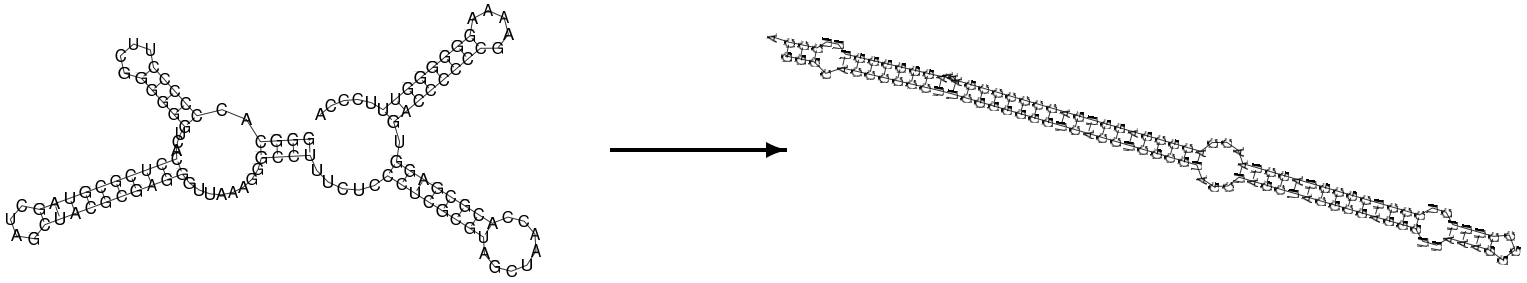
1. Calculate distance bewteen start and target structure
2. Generate all neighbors of the start structure whose distance to the target is less than the distance of the start structure
3. Sort those neighbor structures by their energies
4. Take the *n* energetically best structures, take them as new starting points and repeat the procedure until the stop structure is found
5. If a path has been found, try to find another one with lower energy barrier





PathFinder example - SV11

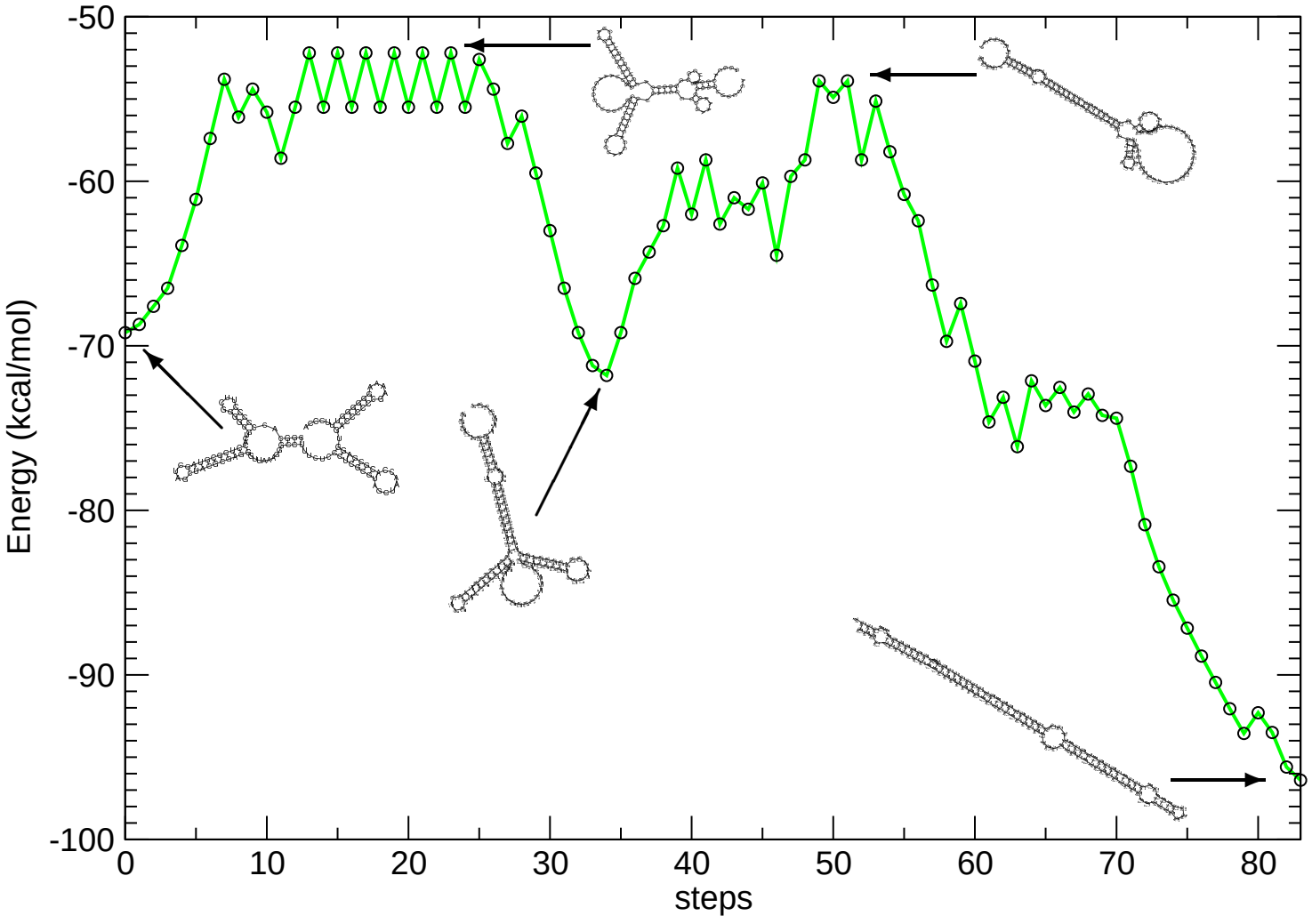
SV11 is a RNA switch of length 115



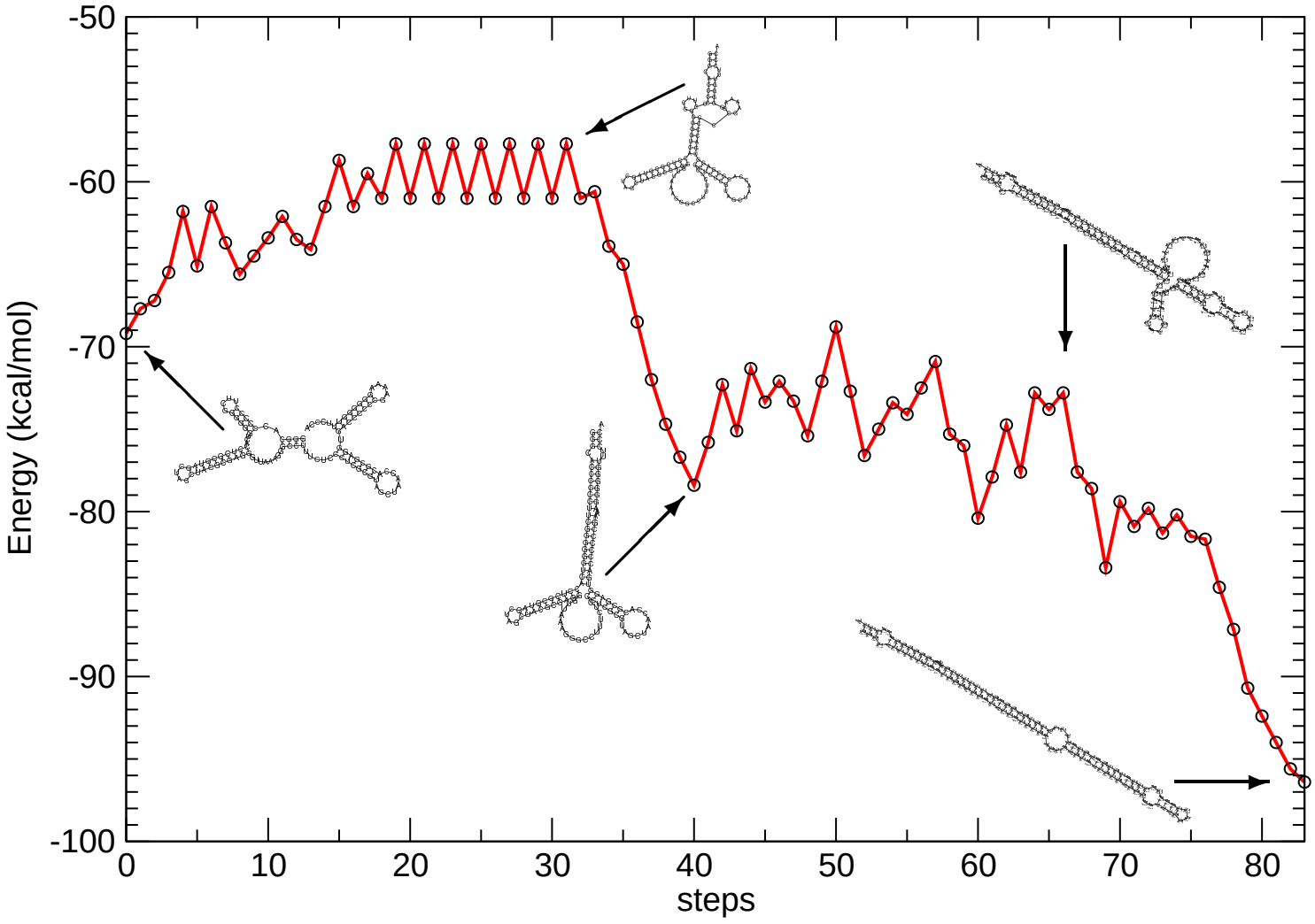
$E = -69.2$ kcal/mol
metastable
template for Qβ replicase

$E = -96.4$ kcal/mol
stable
not a template

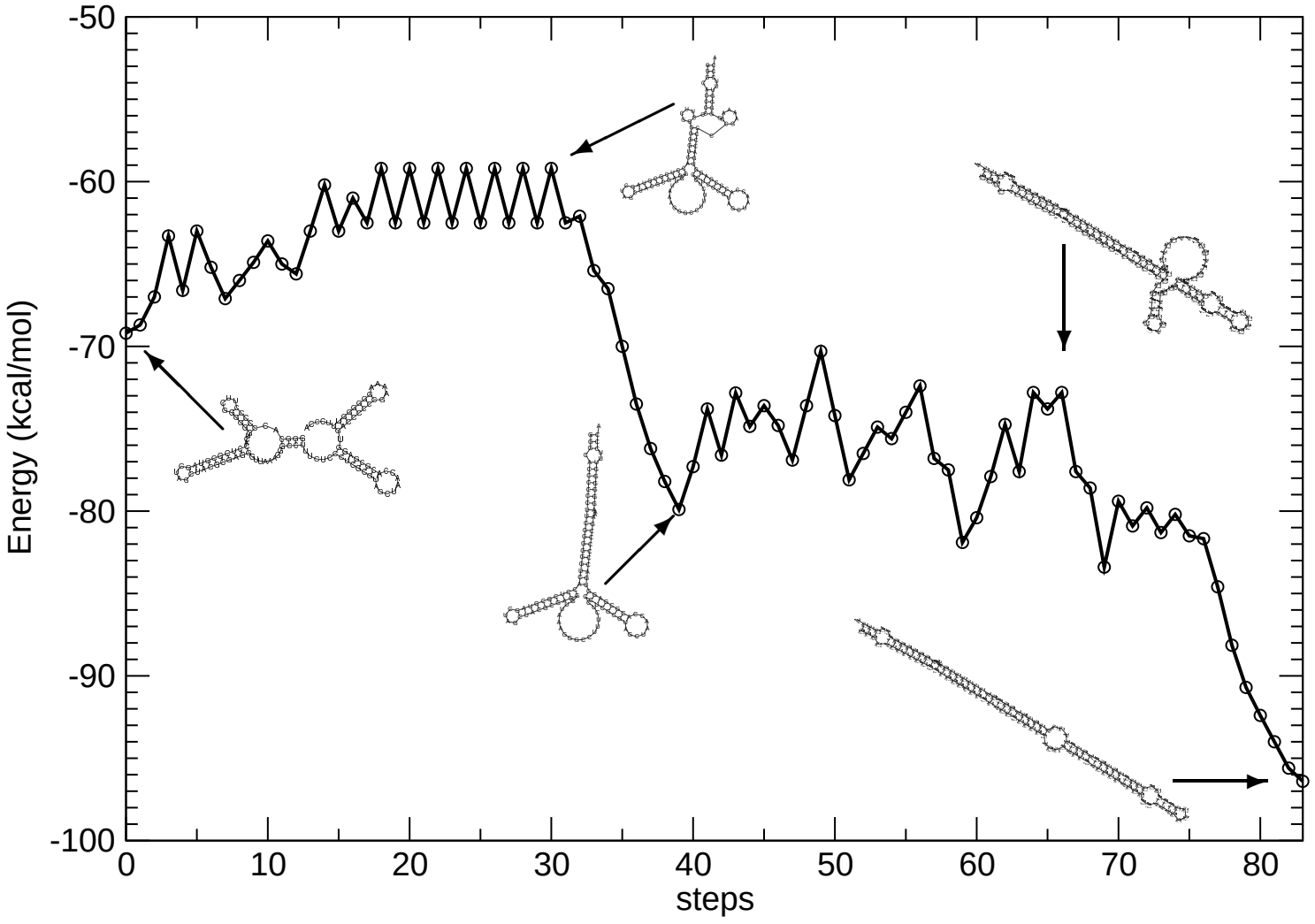
SV11 refolding path 1/3: $E_{\text{saddle}} = -52.2$ kcal/mol



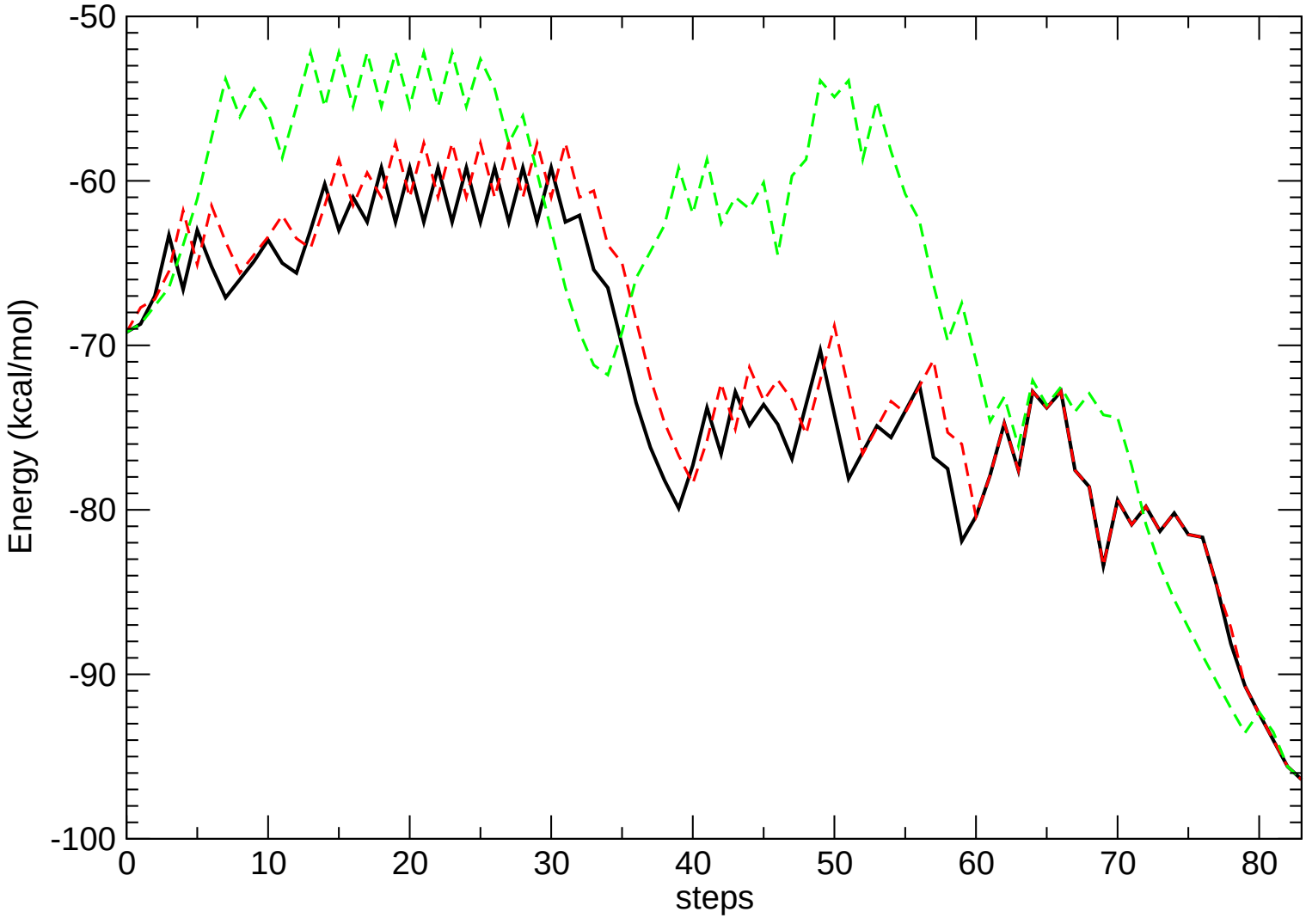
SV11 refolding path 2/3: $E_{\text{saddle}} = -57.7$ kcal/mol



SV11 refolding path 3/3: $E_{\text{saddle}} = -59.2\,\text{kcal/mol}$



SV11 refolding paths



libPF - a generic path sampling library

- In practice, this path sampling heuristics is implemented as a C library
- All structures along a path are stored in a hash and therefore available for the next iterations
- Heuristics routines are strictly separated from model-dependent routines, i.e. the library is completely generic
- Currently, RNA secondary structures and lattice proteins are implemented
- It is easy to extend the functionality to other discrete systems

Conclusion

- Discrete models allow a detailed study of the energy surface
- Barrier trees represent the landscape topology
- The lower part of the energy landscape is accessible by a flooding approach
- Macrostate folding kinetics reduces simulation time drastically
- A path sampling approach yields low-energy refolding paths and is a valuable tool for further kinetics studies

Conclusion

- Discrete models allow a detailed study of the energy surface
- Barrier trees represent the landscape topology
- The lower part of the energy landscape is accessible by a flooding approach
- Macrostate folding kinetics reduces simulation time drastically
- A path sampling approach yields low-energy refolding paths and is a valuable tool for further kinetics studies

Conclusion

- Discrete models allow a detailed study of the energy surface
- Barrier trees represent the landscape topology
- The lower part of the energy landscape is accessible by a flooding approach
- Macrostate folding kinetics reduces simulation time drastically
- A path sampling approach yields low-energy refolding paths and is a valuable tool for further kinetics studies

Conclusion

- Discrete models allow a detailed study of the energy surface
- Barrier trees represent the landscape topology
- The lower part of the energy landscape is accessible by a flooding approach
- Macrostate folding kinetics reduces simulation time drastically
- A path sampling approach yields low-energy refolding paths and is a valuable tool for further kinetics studies

Conclusion

- Discrete models allow a detailed study of the energy surface
- Barrier trees represent the landscape topology
- The lower part of the energy landscape is accessible by a flooding approach
- Macrostate folding kinetics reduces simulation time drastically
- A path sampling approach yields low-energy refolding paths and is a valuable tool for further kinetics studies

Conclusion

- Discrete models allow a detailed study of the energy surface
- Barrier trees represent the landscape topology
- The lower part of the energy landscape is accessible by a flooding approach
- Macrostate folding kinetics reduces simulation time drastically
- A path sampling approach yields low-energy refolding paths and is a valuable tool for further kinetics studies

References

Christoph Flamm, Ivo Hofacker, Peter Stadler
Rolf Backofen, Sebastian Will, Martin Mann


M. T. Wolfinger, W. A. Svrcek-Seiler, C. Flamm, I. L. Hofacker, and P. F. Stadler.
Efficient computation of RNA folding dynamics.
J. Phys. A: Math. Gen., 37(17):4731–4741, 2004.


M. T. Wolfinger, S. Will, I. L. Hofacker, R. Backofen, and P. F. Stadler.
Exploring the lower part of discrete polymer model energy landscapes.
Europhys. Lett., 74(4):725–732, 2006.


Ch. Flamm, W. Fontana, I.L. Hofacker, and P. Schuster.
RNA folding at elementary step resolution.
RNA, 6:325–338, 2000


C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.
Barrier trees of degenerate landscapes.
Z. Phys. Chem., 216:155–173, 2002.