

# Estimation of low-energy refolding paths Visualization of Lattice Protein Dynamics

Michael Wolfinger

Institute for Theoretical Chemistry  
University Vienna

May 23, 2006

# Outline

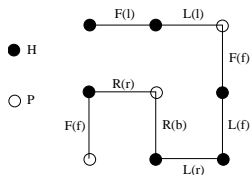
- 1 Lattice Proteins
- 2 Conformation space
- 3 Energy landscapes
- 4 Refolding paths
- 5 Dynamics
- 6 Visualization

# The HP-model

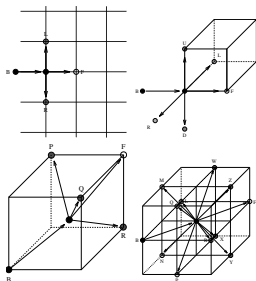
Suggested by Dill, Chan and Lau in the late 1980ies. In this *simplified model*, a conformation is a *self-avoiding walk (SAW)* on a given lattice in 2 or 3 dimensions. Each bond is a straight line, bond angles have a few discrete values. The 20 letter alphabet of amino acids (monomers) is reduced to 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



FRRLFLF



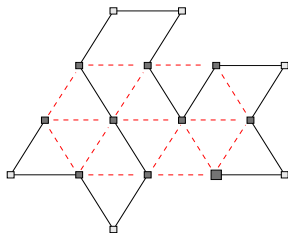
# Contact Potentials

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

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

# Lattice proteins

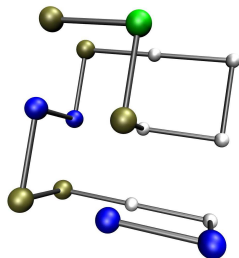
$\mathcal{S} = \text{HPRPHHHPPRHHHPHPH} \quad n = 16$



$$E = -15$$

	<i>H</i>	<i>P</i>
<i>H</i>	-1	0
<i>P</i>	0	0

$\mathcal{S} = \text{NNHHPPNNPHHHHHPXP} \quad n = 16$

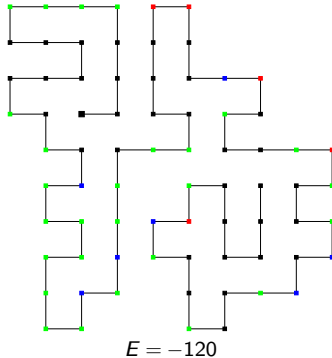


$$E = -16$$

	<i>H</i>	<i>P</i>	<i>N</i>	<i>X</i>
<i>H</i>	-4	0	0	0
<i>P</i>	0	1	-1	0
<i>N</i>	0	-1	1	0
<i>X</i>	0	0	0	0

# Lattice proteins - interaction scheme II

$\mathcal{S} = \text{HHHHNNNNHHHHHHHHNHPNNNNNNNPNPNHNNHHHHXXHHPXHNHHNXNHHNPHPNHHNHHNPXNHHHHHH}$   
 $n = 74$



	<i>H</i>	<i>P</i>	<i>N</i>	<i>X</i>
<i>H</i>	-4	0	0	0
<i>P</i>	0	0	-1	0
<i>N</i>	0	-1	0	0
<i>X</i>	0	0	0	0

# 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 lattice protein structures

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

problem is NP-hard

In the RNA case

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

dynamic programming algorithms available

dim	Lattice Type	$\mu$	$\gamma$
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
- a symmetric neighborhood relation  $\mathfrak{N} : X \times X$
- an energy function  $f : X \rightarrow \mathbf{R}$

# 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 lattice protein structures

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

problem is NP-hard

In the RNA case

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

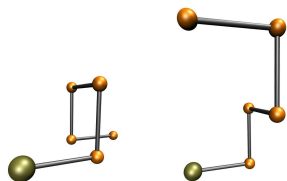
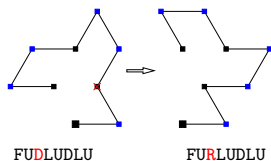
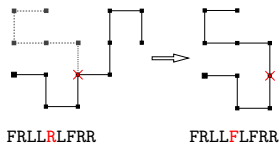
dynamic programming algorithms available

dim	Lattice Type	$\mu$	$\gamma$
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:

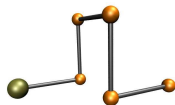
- A set  $X$  of configurations
- a symmetric neighborhood relation  $\mathfrak{N} : X \times X$
- an energy function  $f : X \rightarrow \mathbf{R}$

# The move set



FU**U**RR

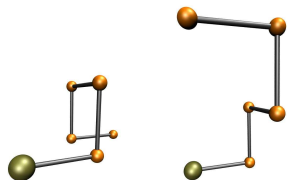
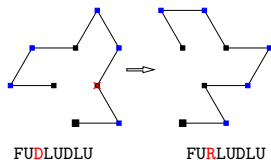
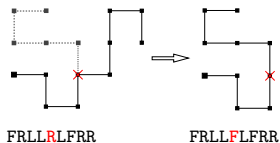
FU**D**RR



FUD**L**R

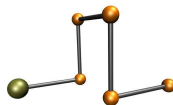
- For each move there must be an inverse move
- Resulting structure must be in  $X$
- Move set must be *ergodic*

# The move set



FUUR

FUDR



FUDL

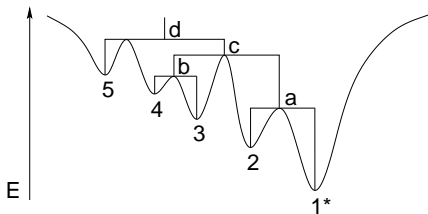
- For each move there must be an inverse move
- Resulting structure must be in  $X$
- Move set must be *ergodic*

# 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 that 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  $E(\mathfrak{G}, x) \leq \eta$  (with a predefined threshold  $\eta$ ).



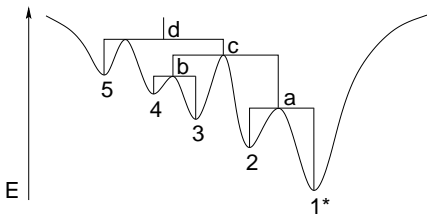
C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.  
Barrier trees of degenerate landscapes.  
*Z. Phys. Chem.*, 216:155–173, 2002.

# 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 that 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  $E(\mathfrak{G}, x) \leq \eta$  (with a predefined threshold  $\eta$ ).



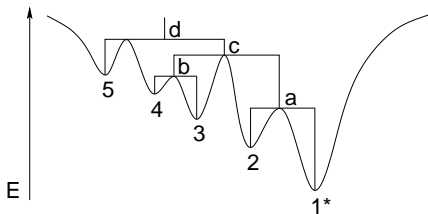
C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.  
Barrier trees of degenerate landscapes.  
*Z. Phys. Chem.*, 216:155–173, 2002.

# 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 that 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  $E(\mathfrak{G}, x) \leq \eta$  (with a predefined threshold  $\eta$ ).



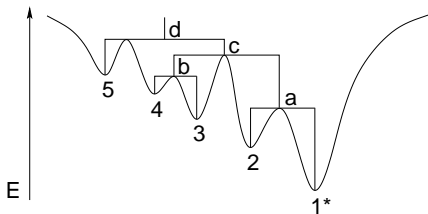
C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.  
Barrier trees of degenerate landscapes.  
*Z. Phys. Chem.*, 216:155–173, 2002.

# 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 that 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  $E(\mathfrak{G}, x) \leq \eta$  (with a predefined threshold  $\eta$ ).



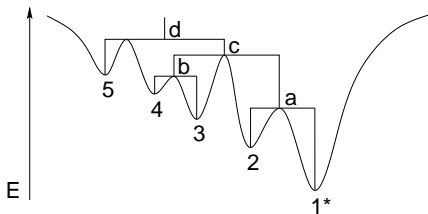
C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.  
Barrier trees of degenerate landscapes.  
*Z. Phys. Chem.*, 216:155–173, 2002.

# 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 that 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  $E(\mathfrak{G}, x) \leq \eta$  (with a predefined threshold  $\eta$ ).



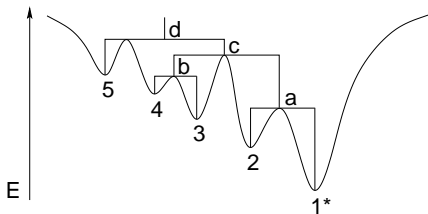
C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.  
Barrier trees of degenerate landscapes.  
*Z. Phys. Chem.*, 216:155–173, 2002.

# 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 that 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  $E(\mathfrak{G}, x) \leq \eta$  (with a predefined threshold  $\eta$ ).



C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger.

Barrier trees of degenerate landscapes.

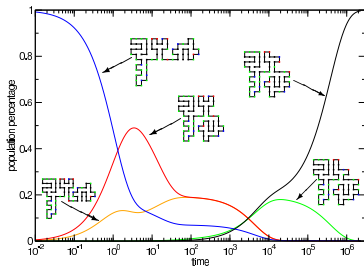
*Z. Phys. Chem.*, 216:155–173, 2002.

# Information from the barrier tree

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

This information can be used to approximate the dynamics of biopolymers, i.e. transition rates between different macrostates (basins in the barrier tree)

- $r_{\beta\alpha} = \Gamma_{\beta\alpha} \exp\left(-\frac{(E_{\beta\alpha}^* - G_{\alpha})}{kT}\right)$



# The lower part of the energy landscape

Two conformations  $x$  and  $y$  are mutually accessible at the level  $\eta$  (written as  $x \xleftrightarrow{\underline{\eta}} y$ ) if there is a walk from  $x$  to  $y$  such that all conformations  $z$  in the walk satisfy  $E(\mathcal{G}, z) \leq \eta$ . The **saddle height**  $\hat{f}(x, y)$  of  $x$  and  $y$  is defined by

$$\hat{f}(x, y) = \min\{\eta \mid x \xleftrightarrow{\underline{\eta}} y\}$$

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

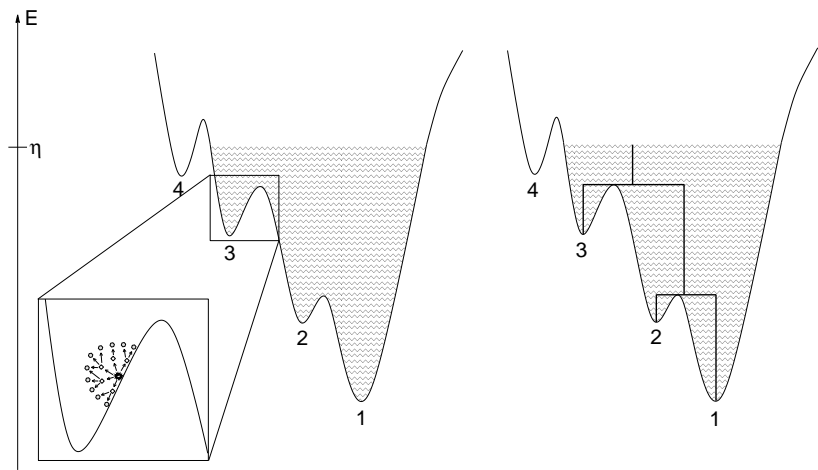
$$X^{\leq \eta} = \{y \mid \exists x \in X_{\min}^{\leq \eta} : \hat{f}(x, y) \leq \eta\}$$

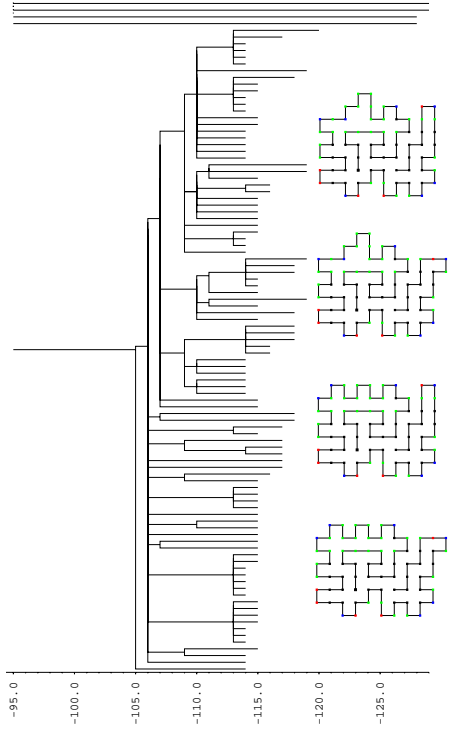
Given a restricted **set of low-energy conformations**,  $X_{\text{init}}$ , and a reasonable value for  $\eta$ , the lower part of the energy landscape can be calculated.



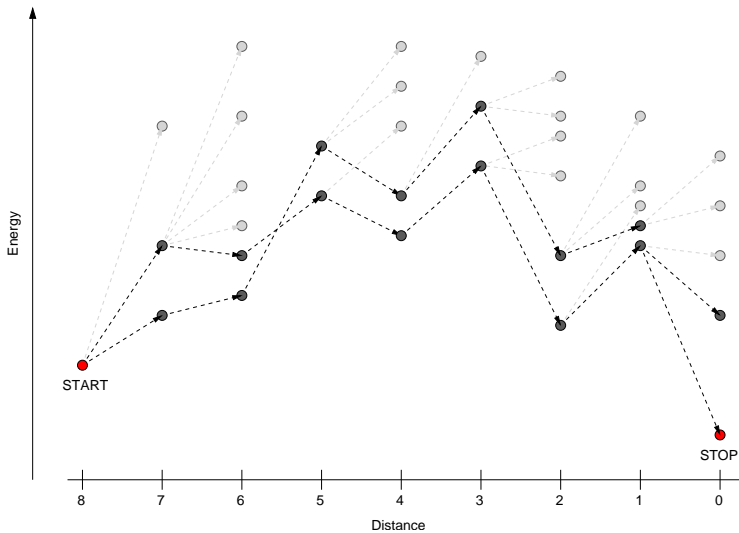
M. T. Wolfinger, S. Will, I. L. Hofacker, R. Backofen, and P. F. Stadler.  
Exploring the lower part of discrete polymer model energy landscapes.  
*Europhys. Lett.*, 2006.

# The Flooder approach

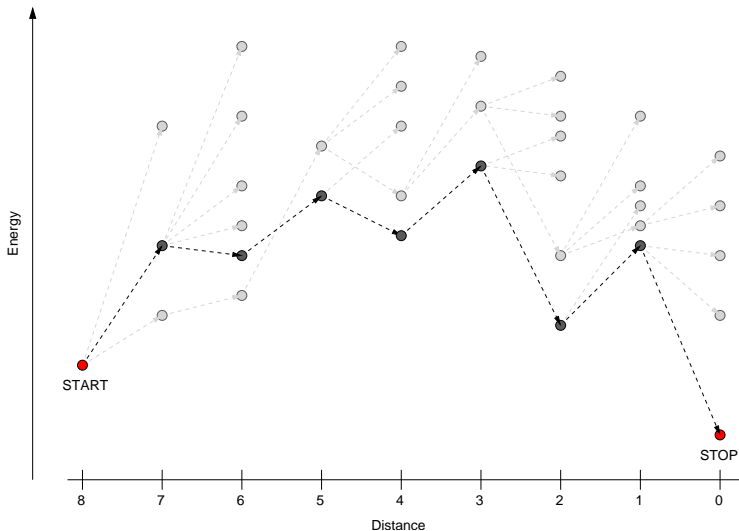




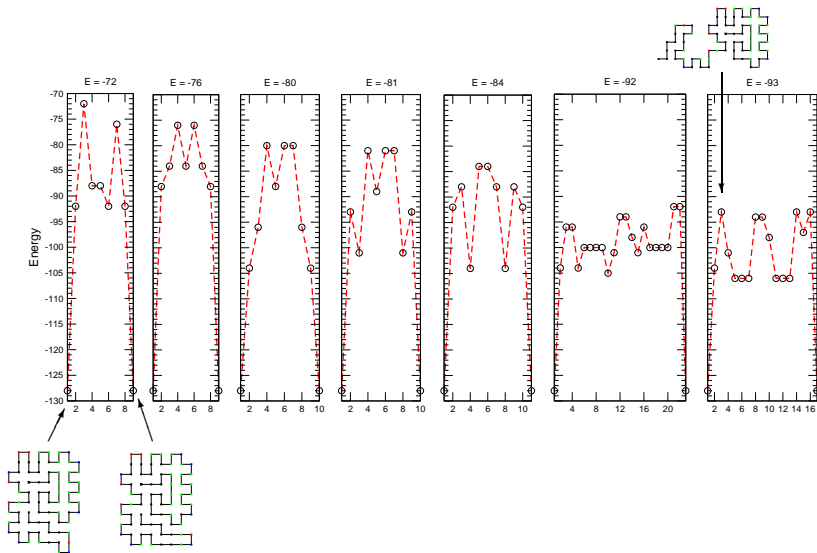
# PathFinder - illustration



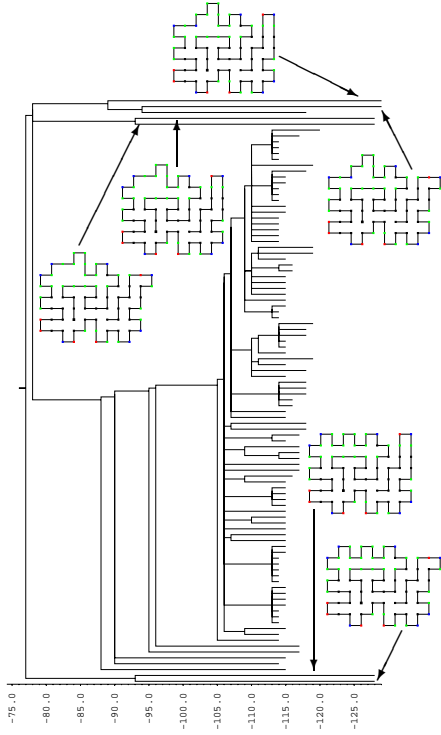
# PathFinder - illustration



# Refolding profiles



Connected!



# Dynamics of biopolymers

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?)*

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

# Barrier tree kinetics

For a reduced description we need

- **macro-states** that form a partition of full configuration space
- **transition rates** between macro-states, e.g.

$$r_{\beta\alpha} = \Gamma_{\beta\alpha} \exp\left(-\frac{(E_{\beta\alpha}^* - G_{\alpha})}{kT}\right) \text{ or}$$

$$r_{\beta\alpha} = \sum_{y \in \beta} \sum_{x \in \alpha} r_{yx} \text{Prob}[x|\alpha] \quad \text{for } \alpha \neq \beta \text{ with } r_{yx} = \begin{cases} e^{-\frac{\Delta E}{kT}} & \text{if } \Delta E > 0 \\ 0 & y \notin \mathcal{N}(x) \\ 1 & \end{cases}$$

All relevant quantities can be computed via the flooding algorithm.

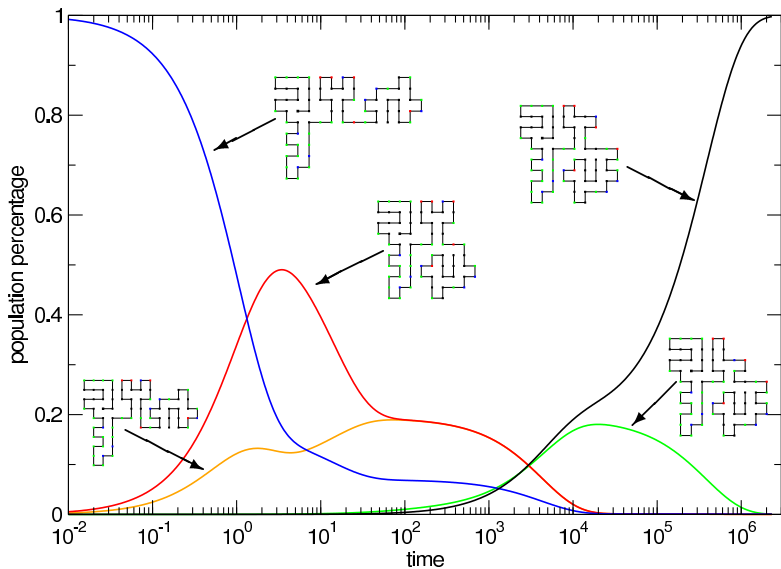


M. T. Wolfinger, W. A. Svrcek-Seiler, C. Flamm, I. L. Hofacker, and P. F. Stadler.

Efficient computation of RNA folding dynamics.

*J. Phys. A: Math. Gen.*, 37(17):4731–4741, 2004.

# Barrier tree kinetics - example



# Kinetic Folding Algorithm

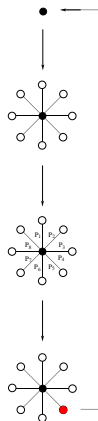
Simulate folding kinetics by a rejection-less Monte-Carlo type algorithm:

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  $1/\sum_i P_i$ .



C. Flamm, W. Fontana, I. Hofacker, and P. Schuster.

RNA folding kinetics at elementary step resolution.

*RNA*, 6:325–338, 2000.

- Visualize Pifold output
- Support data analysis
- Emphasize possible relationships
- Provide simulation comparison
- Provide data analysis and exploration
- Uncover regularities
- Expose the unseen
- Speed up cognition

Shneiderman's mantra:

" Overview first, zoom and filter, details on demand "

- Start with an overview
- Let the user filter out interesting data
- Show details only on demand for different data



S. Pötsch, G. Scheuermann, M. T. Wolfinger, C. Flamm, and P. F. Stadler.  
Visualization of lattice-based protein folding simulations.  
In *10th International Conference on Information Visualization (IV06)*, 2006.

- Visualize Pifold output
- Support data analysis
- Emphasize possible relationships
- Provide simulation comparison
- Provide data analysis and exploration
- Uncover regularities
- Expose the unseen
- Speed up cognition

Shneiderman's mantra:

**" Overview first, zoom and filter, details on demand "**

- Start with an overview
- Let the user filter out interesting data
- Show details only on demand for different data



S. Pötsch, G. Scheuermann, M. T. Wolfinger, C. Flamm, and P. F. Stadler.  
Visualization of lattice-based protein folding simulations.  
*In 10th International Conference on Information Visualization (IV06), 2006.*

# Overview first

Parold Visualization 1.1

File Actions Windows Help

Name: kh68\_21to5 Date: Wed Sep 14 11:07:26 2005

Sequence: N N H H P P N N P H H H P X P  
 Start structure: F L F F L F F F F R R L L L  
 Stop structure: F L F L F L L F R R L L L

Overview table

Simulation Index	Time	Steps	Found	Energy Max	Energy Min	
1	591	0.220	3	true	-11	-14
2	956	0.080	3	true	-11	-14
3	174	0.220	3	true	-11	-14
4	597	0.308	3	true	-11	-14
5	415	0.414	3	true	-11	-14
6	450	0.167	3	true	-11	-14
7	878	0.560	3	true	-11	-14
8	350	0.446	3	true	-11	-14
9	138	0.502	3	true	-11	-14
10	471	0.864	5	true	-3	-14
11	455	0.264	5	true	-11	-14
12	444	0.823	5	true	-11	-14
13	876	1.227	5	true	-11	-14
14	940	0.619	5	true	-11	-14
15	686	2.957	5	true	-3	-14
16	751	0.449	5	true	-11	-14
17	743	0.766	5	true	-11	-14
18	490	1.911	5	true	-11	-14
19	702	1.336	5	true	-3	-14
20	149	0.422	5	true	-3	-14
21	489	1.159	5	true	-3	-14
22	489	1.091	5	true	-3	-14
23	359	1.087	5	true	-3	-14
24	209	0.704	5	true	-3	-14
25	230	0.582	5	true	-11	-14
26	28	0.735	5	true	-11	-14
27	23	1.830	6	true	-11	-14
28	673	1.369	7	true	-10	-14
29	498	1.735	7	true	-3	-14
30	192	0.954	7	true	-11	-14

Energy model:  
 Model: HFNX  
 Damage: 2  
 Temp: 37.0  
 Logml: logarithmic  
 Par: VRNA-1.4

Move set:  
 Lattice: SQ  
 NoShift: off  
 NoLP: off

Simulation:  
 Num: 1000  
 Time: 10000.00  
 Seed: clock  
 Mc: Metropolis  
 Fpt: on

Output:  
 Log: kh68\_21to5  
 Slant: on  
 Lmin: off  
 Cut: 20.00

Population chart

Energy map

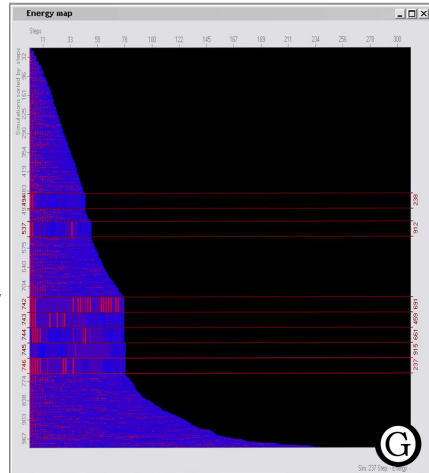
Energy histogram

Structure histogram

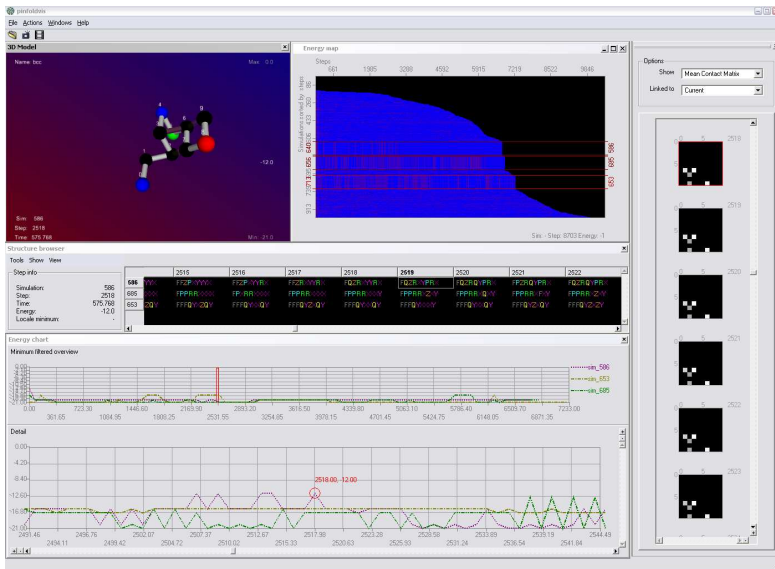
Labels A through G are placed around the interface elements.

# Zoom and filter

- Energy map
- Focus & context technique
- Huge data sets, limited screen size
- Details and overview in one window



# Details on demand



# Conclusion

- **Barrier trees** approximate the landscape topology and folding kinetics.
- A **heuristic approach** allows to sample low-energy refolding paths between different structures
- A **macrostate approach** of folding kinetics reduces simulation time drastically.
- A tool for **visualization of folding trajectories** enables a thorough investigation of folding kinetics simulations.
- This **newly generated framework** provides a powerful method for further refinement of biopolymer folding landscapes.

# Conclusion

- **Barrier trees** approximate the landscape topology and folding kinetics.
- A **heuristic approach** allows to sample low-energy refolding paths between different structures
- A **macrostate approach** of folding kinetics reduces simulation time drastically.
- A tool for **visualization of folding trajectories** enables a thorough investigation of folding kinetics simulations.
- This **newly generated framework** provides a powerful method for further refinement of biopolymer folding landscapes.

# Conclusion

- **Barrier trees** approximate the landscape topology and folding kinetics.
- A **heuristic approach** allows to sample low-energy refolding paths between different structures
- A **macrostate approach** of folding kinetics reduces simulation time drastically.
- A tool for **visualization of folding trajectories** enables a thorough investigation of folding kinetics simulations.
- This **newly generated framework** provides a powerful method for further refinement of biopolymer folding landscapes.

# Conclusion

- **Barrier trees** approximate the landscape topology and folding kinetics.
- A **heuristic approach** allows to sample low-energy refolding paths between different structures
- A **macrostate approach** of folding kinetics reduces simulation time drastically.
- A tool for **visualization of folding trajectories** enables a thorough investigation of folding kinetics simulations.
- This **newly generated framework** provides a powerful method for further refinement of biopolymer folding landscapes.

# Conclusion

- **Barrier trees** approximate the landscape topology and folding kinetics.
- A **heuristic approach** allows to sample low-energy refolding paths between different structures
- A **macrostate approach** of folding kinetics reduces simulation time drastically.
- A tool for **visualization of folding trajectories** enables a thorough investigation of folding kinetics simulations.
- This **newly generated framework** provides a powerful method for further refinement of biopolymer folding landscapes.

# Thanks

Ivo Hofacker

Christoph Flamm

Sebastian Will

Sebastian Pötzsch

Gerik Scheuermann

Rolf Backofen

Peter Stadler

Peter Schuster