Estimation of low-energy refolding paths

Michael Wolfinger

Institute for Theoretical Chemistry
University Vienna

February 21, 2006

Outline

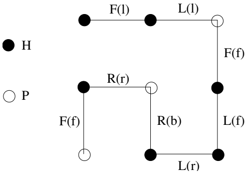
tbi

The HP-model

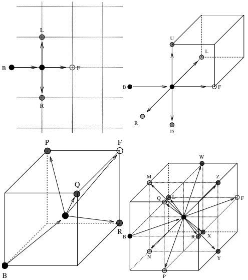
Suggested by Dill, Chan and Lau in the late 1980ies. In this *simplified model*, a conformation is a *self-avoiding walk (SAW)* on a given lattice in 2 or 3 dimensions. Each bond is a straight line, bond angles have a few discrete values. The 20 letter alphabet of amino acids (monomers) is reduced to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



FRRLLFLF



Contact Potentials

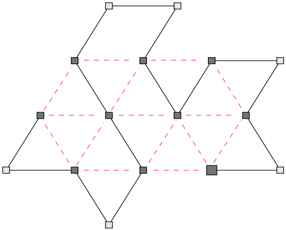
Generally, the energy function for a sequence with n residues $\mathfrak{S} = \mathfrak{s}_1\mathfrak{s}_2\ldots\mathfrak{s}_n$ with $\mathfrak{s}_i \in \mathscr{A} = \{a_1, a_2, \ldots, a_b\}$, the alphabet of b residues, and an overall configuration $x = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$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

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

tbi

Lattice proteins

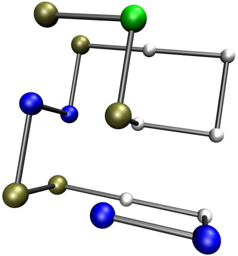
$\mathfrak{S} = \texttt{HPHPHHHPPHHHPHPH} \quad n = 16$



$E = -15$

	H	P
H	-1	0
P	0	0

$\mathfrak{S} = \texttt{NNHHPPNNPHHHHPXP} \quad n = 16$



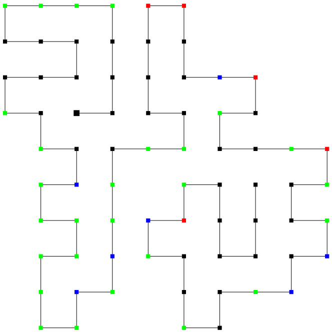
$E = -16$

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

Lattice proteins - interaction scheme II

$\mathfrak{S} = \texttt{HHHHNNNNHHHHHHNHNHPNNNNNNNNPNPNNHNNHHHHXXHHPXHNHHNXNHHNPHPNHHNHHNPXNHHHHHH}$

$n = 74$



$E = -120$

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

tbi

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

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

tbi

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

tbi

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	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
- 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

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

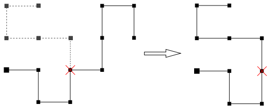
dynamic programming algorithms available

dim	Lattice Type	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

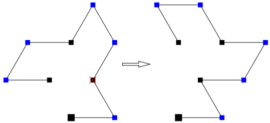
Formally, three things are needed to construct an energy landscape:

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

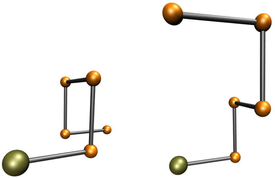
The move set



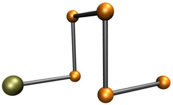
FRLLRLFRR FRLLFLFRR



FUDLUDLU FURLUDLU



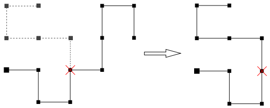
FUURR FUDRR



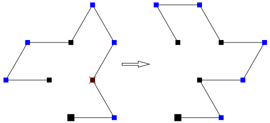
FUDLR

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

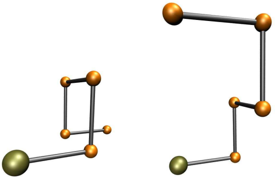
The move set



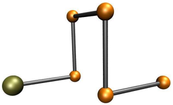
FRLLRLFRR FRLLFLFRR



FUDLUDLU FURLUDLU



FUURR FUDRR



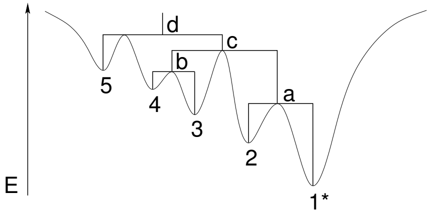
FUDLR

- 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

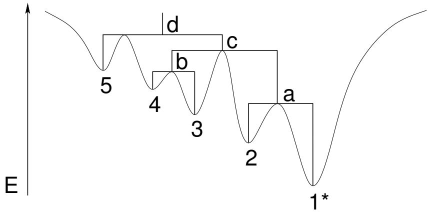
- a walk between two conformations x and y as a list of conformations $x = x_1 \ldots x_{m+1} = y$ such that $\forall 1 \leq i \leq m : \mathfrak{N}(x_i, x_{i+1})$
- the lower part of the energy landscape (written as $X^{\leq \eta}$) as *all* conformations x such that $E(\mathfrak{S}, x) \leq \eta$ (with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

tbi

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

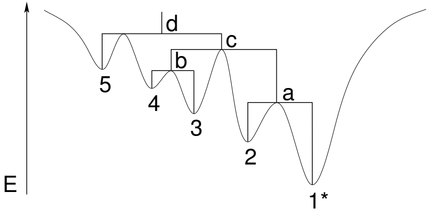
- a walk between two conformations x and y as a list of conformations $x = x_1 \ldots x_{m+1} = y$ such that $\forall 1 \leq i \leq m : \mathfrak{N}(x_i, x_{i+1})$
- the lower part of the energy landscape (written as $X^{\leq \eta}$) as *all* conformations x such that $E(\mathfrak{S}, x) \leq \eta$ (with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

tbi

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



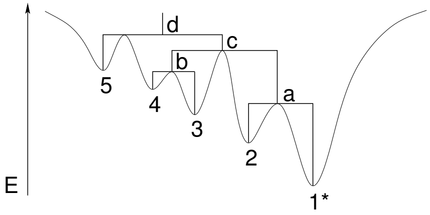
We further define

- 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).

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

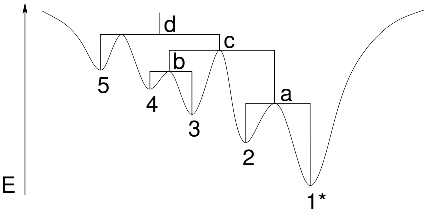
- a walk between two conformations x and y as a list of conformations $x = x_1 \dots x_{m+1} = y$ such that $\forall 1 \leq i \leq m : \mathfrak{N}(x_i, x_{i+1})$
- the lower part of the energy landscape (written as $X^{\leq \eta}$) as *all* conformations x such that $E(\mathfrak{S}, x) \leq \eta$ (with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

tbi

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).

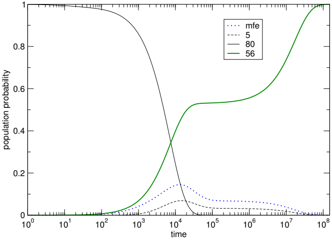
tbi

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)

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



The lower part of the energy landscape

Two conformations x and y are mutually accessible at the level η (written as $x \overset{\eta}{\longleftrightarrow} y$) 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

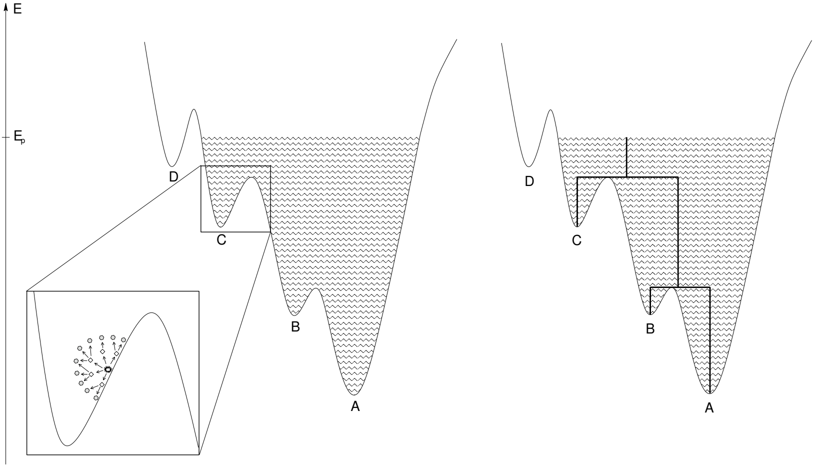
$$\hat{f}(x, y) = \min\{\eta \mid x \overset{\eta}{\longleftrightarrow} y\}$$

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

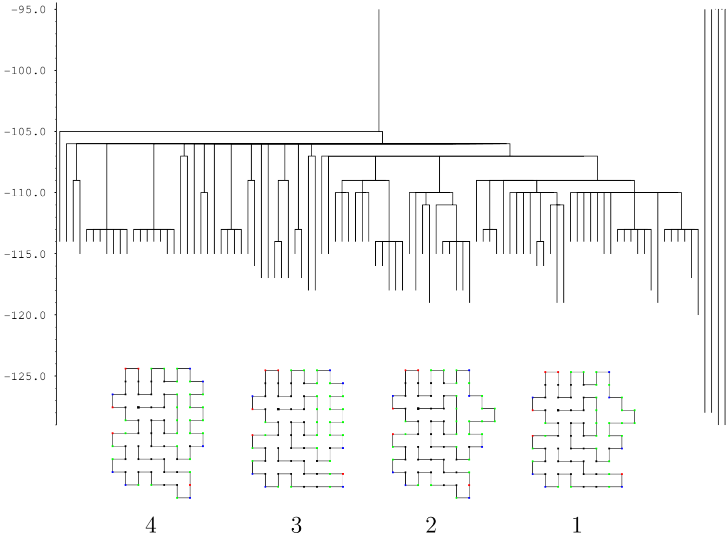
$$X^{\leq \eta} = \{y \mid \exists x \in X_{\text{min}}^{\leq \eta} : \hat{f}(x, y) \leq \eta\}$$

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.

The Flooder approach

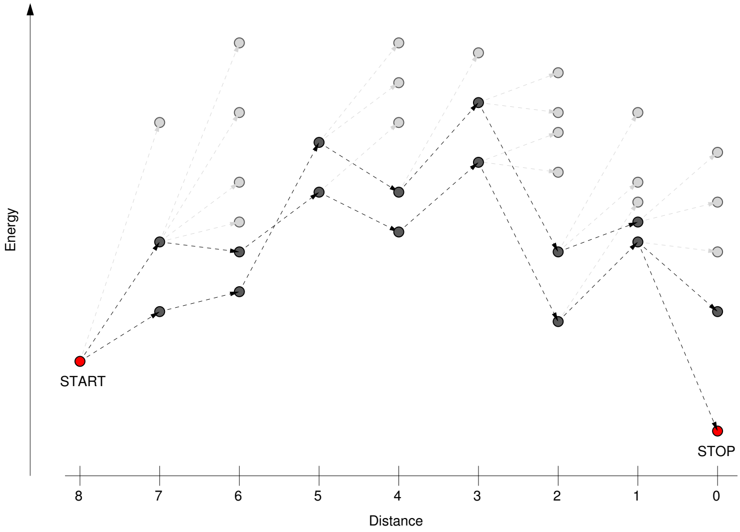


!Connected

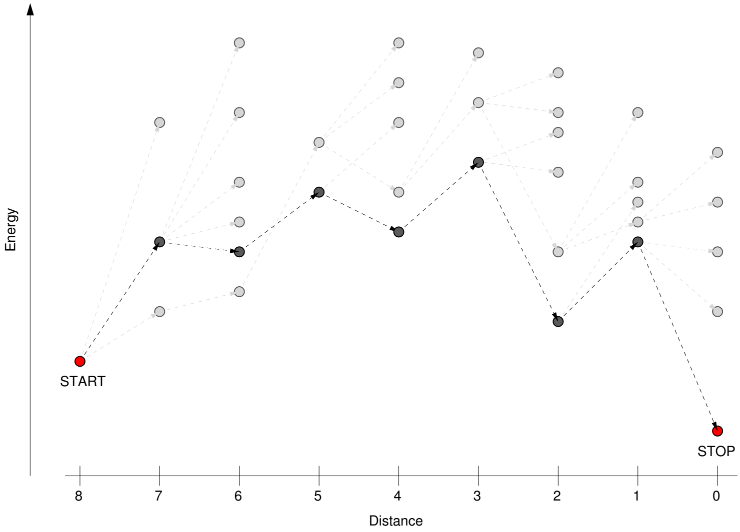


tbi

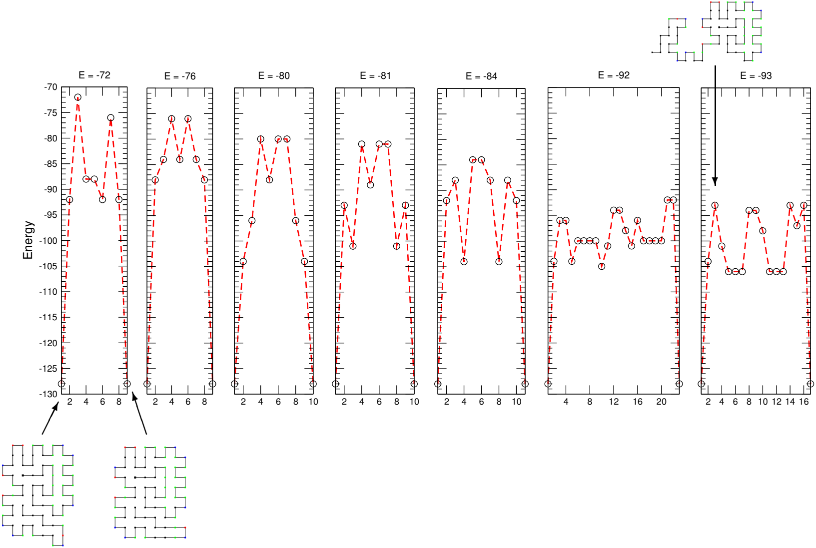
LatticePath - illustration



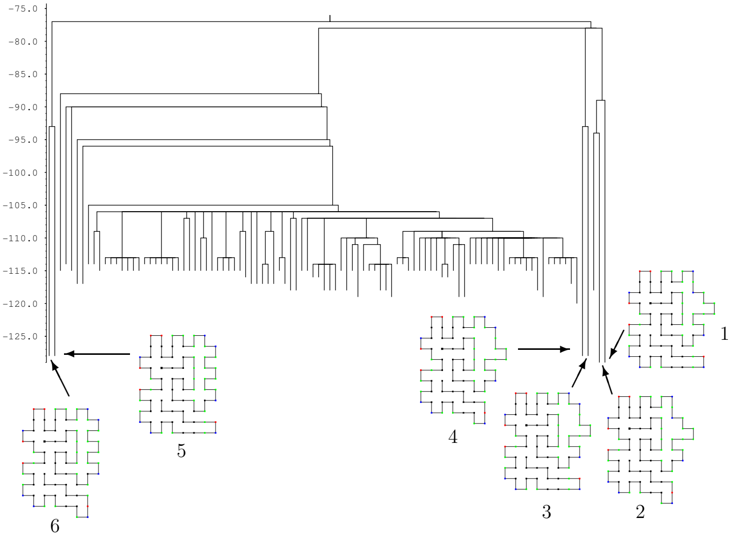
LatticePath - illustration



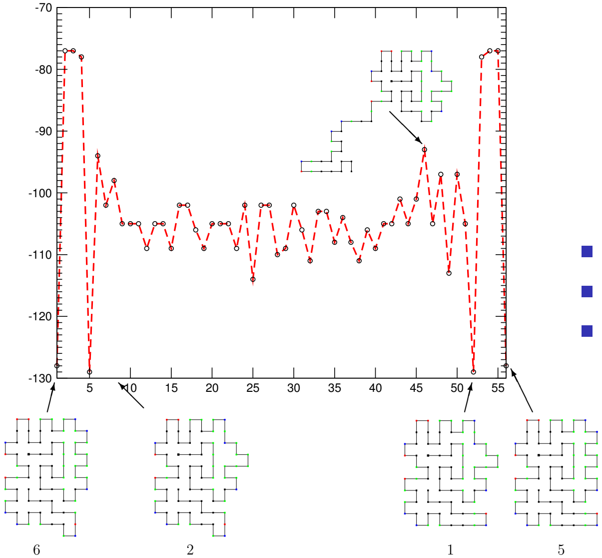
Refolding profiles



Connected!



A longer refolding profile



- 56 steps
- $E_{\text{max}} = -77$
- $E_{\text{min}} = -129$

Conclusion

- Discrete models allow a detailed study of the energy surface.
- Barrier trees approximate the landscape topology and folding kinetics.
- A heuristic approach allows to sample low-energy refolding paths between different structures
- This newly generated framework provides a powerful method for further refinement of biopolymer folding landscapes.

tbi

Conclusion

- Discrete models allow a detailed study of the energy surface.
- Barrier trees approximate the landscape topology and folding kinetics.
- A heuristic approach allows to sample low-energy refolding paths between different structures
- This newly generated framework provides a powerful method for further refinement of biopolymer folding landscapes.

Conclusion

- Discrete models allow a detailed study of the energy surface.
- Barrier trees approximate the landscape topology and folding kinetics.
- A heuristic approach allows to sample low-energy refolding paths between different structures
- This newly generated framework provides a powerful method for further refinement of biopolymer folding landscapes.

Conclusion

- Discrete models allow a detailed study of the energy surface.
- Barrier trees approximate the landscape topology and folding kinetics.
- A heuristic approach allows to sample low-energy refolding paths between different structures
- This newly generated framework provides a powerful method for further refinement of biopolymer folding landscapes.

Thanks

Sebastian Will

Rolf Backofen

Peter Stadler

Ivo Hofacker

Christoph Flamm

The electric, withouth whom this would not be possible