Energy Landscapes and Dynamics of Biopolymers

Michael T. Wolfinger[1], W. Andreas Svrcek-Seiler[1], Christoph Flamm[2], Ivo L. Hofacker[1], Peter F. Stadler[2]

[1]Institute for Theoretical Chemistry, University of Vienna, Austria

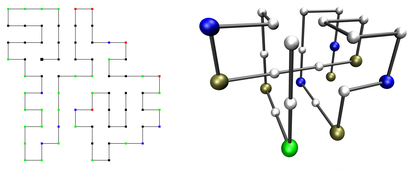
[2]Department of Computer Science, University of Leipzig, Germany

Tel: +43 1 4277 52747 Fax: +43 1 4277 52793 Email: {mtw,svrci,xtof,ivo,studla}@tbi.univie.ac.at Web: http://www.tbi.univie.ac.at/

The ability of biomolecules like DNA, RNA or proteins to fold into a well-defined native state is a prerequisite for biologically functional molecules. A reasonable level of coarse-graining is needed in order to treat biomolecules within a theoretical framework. Kinetics and structure formation processes of biopolymers are crucially determined by the topological details of the underlying (free) energy landscape. We present a generic, problem independent framework for exploration of the low-energy portion of the energy landscape of discrete systems and apply it to the energy landscape of lattice proteins.

Lattice Proteins

The HPNX model is used to study general properties of lattice heteropolymers. Within this simplified model, a conformation is regarded as a self-avoiding walk on a two- or three-dimensional lattice.



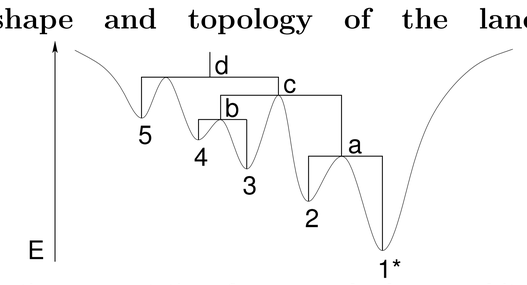
	H	P	N	X
H	−4	0	0	0
P	0	0	−1	0
N	0	−1	0	0
X	0	0	0	0

Left: 74-mer lattice protein on the 2D square lattice (SQ). *Middle:* 27-mer on the 3D simple cubic (SC) lattice. *Right:* Interaction scheme for the HPNX model used here.

The 20 letter alphabet of amino acids is reduced to a four letter alphabet: Hydrophobic (H), positive (P), negative (N) and neutral (X) residues. Energy is evaluated via a pair potential with attractive interactions when two beads are neighbors in the lattice but not along the chain. Lattice heteropolymers offer the advantage of modeling the general properties of proteins at relatively low computational cost. However, they represent a crude abstraction by implying fixed bond lengths and angles.

Energy Landscapes

The energy landscape of a biopolymer molecule is a complex surface of the free energy versus the conformational degrees of freedom. Energy landscapes are conveniently visualized by barrier trees that give an impression on the overall shape and topology of the landscape [2].

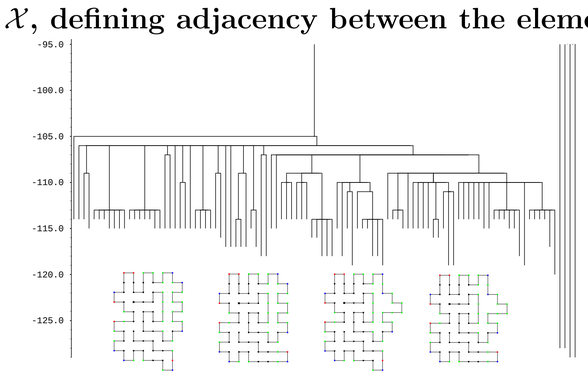


Schematics representation of an energy landscape and its associated barrier tree. Local minima are labeled with numbers (1-5), saddle points with lowercase letters (a-d). The global minimum is marked with an asterisk.

Things needed to construct an energy landscape:

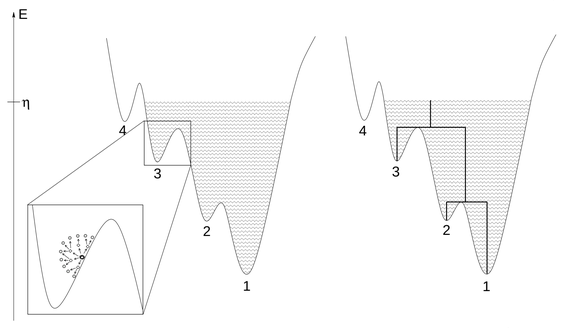
1. a set $\mathcal{X}$ of configurations
2. a notion $\mathcal{M}$ of neighborhood on $\mathcal{X}$ and
3. an energy function $f : \mathcal{X} \to \mathbf{R}$.

The conformation space $\mathcal{X}$ of a (biopolymer) sequence $\mathfrak{S}$ is the total set of configurations S compatible with this sequence. The move set $\mathcal{M}$ is an order relation on $\mathcal{X}$, defining adjacency between the elements of $\mathcal{X}$.



Energy landscape of a 74-mer lattice protein on the SQ lattice, calculated via the flooding algorithm with an energy threshold of -95. The lowest 4 local minima (corresponding structures listed below) on the right are not attached to the rest of the tree.

It crucially determines the topology of the underlying energy landscape. Here we use non-local, ergodic pivot moves that give rise to a fixed neighborhood relation $\mathcal{N} : \mathcal{X} \times \mathcal{X}$. A ***walk*** between two conformations x and y is a list of conformations $x = x_1 \dots x_{m+1} = y$ such that $\forall 1 \le i \le m : \mathcal{N}(x_i, x_{i+1})$. Given a threshold η, the lower part of the energy landscape (written as $\mathcal{X}^{\le\eta}$) consists of ***all*** conformations x such that $E(\mathfrak{S}, x) \le \eta$.



Schematic representation of the flooding algorithm (left plot). Starting from a certain conformation, all neighbor conformations are calculated repeatedly until all conformations in a certain region of the energy landscape are found.

Since exhaustive enumeration of all possible structures is only applicable to very short chains (the lattice protein folding problem was shown to be NP hard), we developed an algorithm for investigating the low energy part of the energy landscape selectively [5]. This approach starts at low energy conformations and enumerates all "accessible" conformations. To exemplify the idea, for generating the lower part completely one starts with ***all*** local minima x with $E(\mathfrak{S}, x) \le \eta$. Iteratively, one visits all conformations that are neighbors of already seen conformations and stay below the energy threshold η. Two conformations x and y are mutually accessible at the level η (written as $x \leftrightarrow^{\eta} y$) if there is a walk from x to y such that all conformations z in the walk satisfy $E(\mathfrak{S}, z) \le \eta$ [2]. The ***saddle height*** $\hat{f}(x, y)$ of x and y is defined by

$$\hat{f}(x, y) = \min\{\eta \mid x \leftrightarrow^{\eta} y\}.$$

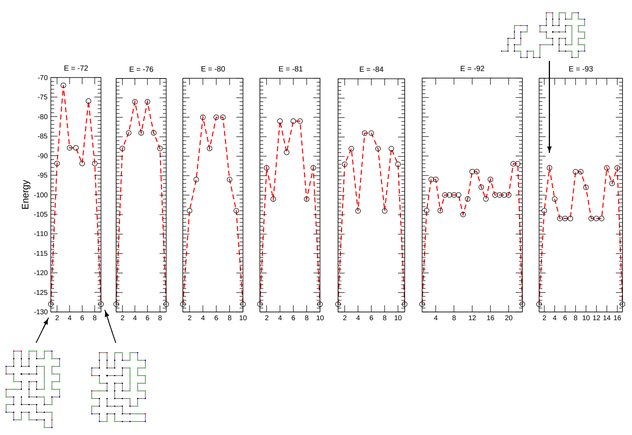
Given the set of all local minima $\mathcal{X}^{\le\eta}_{\min}$ below threshold η, the lower energy part $\mathcal{X}^{\le\eta}$ of the energy landscape is given by

$$\mathcal{X}^{\le\eta} = \{y \mid \exists x \in \mathcal{X}^{\le\eta}_{\min} : \hat{f}(x, y) \le \eta\}.$$

Since the complete set of local minima $\mathcal{X}^{\le\eta}_{\min}$ usually is not available, one can also start from a restricted set of low energy conformations $\mathcal{X}_{\text{init}}$ and hope to enumerate a large part of the low energy conformations.

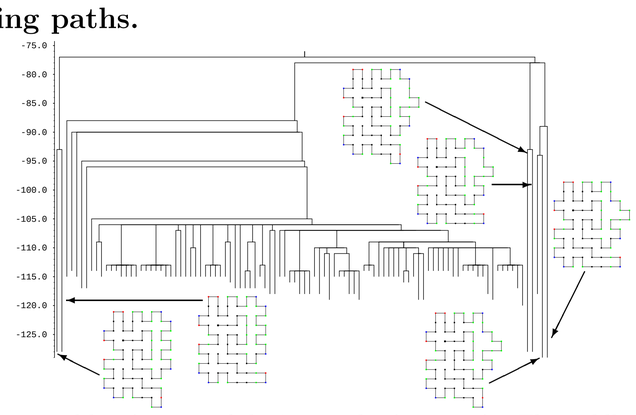
Refolding Paths

The figure at the bottom of the left column illustrates a common problem with the calculation of barrier trees based on the flooding approach: Saddle heights are not known a priori, resulting in non-connected trees. To overcome this, we developed a breadth-first-search heuristics for estimating minimal refolding paths between two arbitrary structures.



Energy profiles of the refolding process between two lattice protein structures from the barrier tree above.

Starting from a given conformation, we iteratively generate a predefined number of neighbor conformations with the constraint that adjacent structures have a lower (hamming) distance to the target. Optionally, we also allow a few indirect steps on the way to the target, i.e. those moves that result in a larger distance All visited structures are stored in a hash, enabling an iterative approximation of low-energy refolding paths.



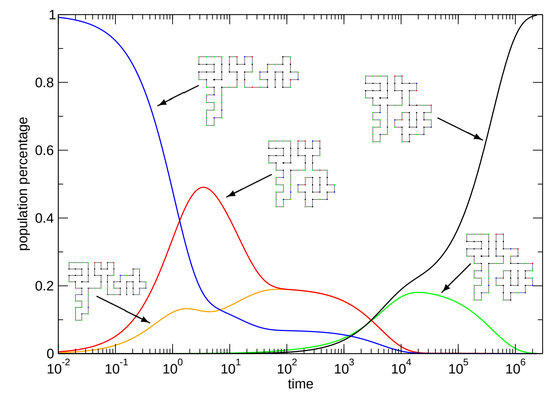
Connected barrier tree of the 74-mer lattice protein. The saddle between the two leftmost structures is at $E - 93$, the saddle connecting these states to the ground state is at $E = -77$.

Dynamics

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 [4]. The transition rate to reach state β from state α typically looks like

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

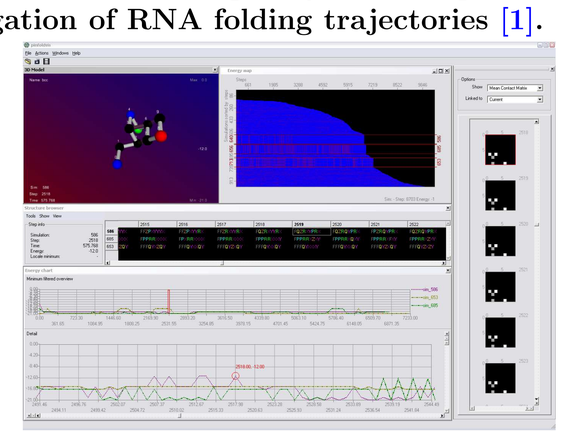
where Γ is a pre-exponential entropic factor, $E^*_{\beta\alpha}$ is the energy of the saddle point between states α and β and G_α is the free energy of basin α.



Reduced refolding dynamics between two selected states of the 74-mer. Several macrostates are populated temporarily, whereas all conformations find the target state after approx. 2 million time steps.

Visualization

Results from the macrostate dynamics are usually in good agreement with exact folding simulations obtained from Pinfold, a modified Monte Carlo type algorithm that has originally been implemented for investigation of RNA folding trajectories [1].



To facilitate the investigation of folding trajectories, we developed a graphical user interface for efficiently analyzing the results from Pinfold [3].

This novel framework allows not only for a rapid investigation of folding kinetics, but also provides a powerful method for further refinement of biopolymer folding landscapes.

References

[1] C. Flamm, W. Fontana, I. Hofacker, and P. Schuster. RNA folding kinetics at elementary step resolution. *RNA*, 6:325–338, 2000.

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

[3] S. Pötzsch, 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.

[4] 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.

[5] 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):726–732, 2006.