

Funnelns in Energy Landscapes

Konstantin Klemm
Christoph Flamm
Peter F. Stadler

SFI WORKING PAPER: 2007-09-033

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu



SANTA FE INSTITUTE

Funnels in Energy Landscapes

Konstantin Klemm¹, Christoph Flamm², and Peter F. Stadler^{1,2,3}

¹Bioinformatics Group, Department of Computer Science, University of Leipzig

²Institute for Theoretical Chemistry, University of Vienna, Austria

³Santa Fe Institute, Santa Fe, New Mexico, USA

{klemm,stadler,xtof}@bioinf.uni-leipzig.de

Received: date / Revised version: date

Abstract. Local minima and the saddle points separating them in the energy landscape are known to dominate the dynamics of biopolymer folding. Here we introduce a notion of a “folding funnel” that is concisely defined in terms of energy minima and saddle points, while at the same time conforming to a notion of a “folding funnel” as it is discussed in the protein folding literature.

PACS. 87.10.+e General theory and mathematical aspects – 87.15.Cc Folding and sequence analysis – 89.75.Fb Structures and organization in complex systems

1 Introduction

The dynamics of structure formation (“folding”) of biopolymers, both protein and nucleic acids, can be understood in terms of their *energy landscapes*. Formally, a landscape is determined by a set X of conformations or states, a neighborhood structure of X that encodes which conformations can be reached from which other ones, and an energy function $E : X \rightarrow \mathbb{R}$ which assigns the folding energy to each state. In the case of nucleic acids it has been demonstrated that dynamics features of the folding process can be derived at least in a good approximation by replacing the full landscape by the collection of local minima and their connecting saddle points [1].

The notion of a “*folding funnel*” has a long history in the protein folding literature [2–8]. It arose from the observation that the folding process of naturally evolved proteins very often follows simple empirical rules that seem to bypass the complexity of the vast network of elementary steps that is required in general to describe the folding process on rugged energy landscapes. Traditionally, the funnel is depicted as a relation of folding energy and “conformational entropy”, alluding to the effect that the energy decreases, on average, as structures are formed that are more and more similar to the native structure of a natural protein [9]. It may come as a surprise, therefore, that despite the great conceptual impact of the notion of a folding funnel in protein folding research, there does not seem to be a clear mathematical definition of “funnel”. Intuitively, one would expect that a funnel should be defined in terms of the basins and barriers of the fitness landscape (since, as mentioned above, these coarse-grained topological features determine the folding dynamics). Furthermore, it

should imply the “funneling” of folding trajectories towards the ground state of the molecule.

Methods to elucidate the basin structure of landscapes by means of trees that represent local minima and their connecting saddle points have been developed independently in different contexts, among them $\pm J$ spin models [10], potential energy surfaces (PES) for protein folding [11,12] and molecular clusters [13,14], and the kinetics of RNA secondary structure formation [15].

2 Folding Dynamics as a Markov Chain

We consider here only finite discrete conformation spaces X with a prescribed set of elementary moves of transitions that inter-convert conformations. In the following we write $M(x)$ for the set of conformations accessible from $x \in X$. For example, $X = \{-1, +1\}^N$ in spin-glass setting, where flipping single spins is the natural definition of a move. In the case of RNA or protein folding, the breaking and formation of individual contacts between nucleotides or amino acids, resp., is the most natural type of move set [15].

The dynamics on X is modeled as usual by the 1st order Markov chain with Metropolis transition probabilities

$$\begin{aligned} p(y|x) &= \frac{1}{|M(x)|} \min\{1, \exp(-\beta(E_y - E_x))\} & \text{for } y \in M(x) \\ p(x|x) &= 1 - \sum_{y \in M(x)} p(y|x) \end{aligned} \tag{1}$$

All other transition probabilities are zero.

We will be interested in the average time τ_x the system takes to reach a pre-defined target state $x_0 \in X$ when starting at state $x \in X$, given by the recurrence

$$\tau_x = 1 + \sum_{y \in M(x)} p(y|x)\tau_y + p(x|x)\tau_x \quad (2)$$

with $\tau_0 = 0$ (starting from the target state).

In order to investigate the physical basis of the “funneling effect” we start with a simple 1-dimensional toy model with landscapes defined over the integers $\{0, \dots, n\}$, see Table 1 and Figure 1. The time τ to target crucially depends on the ordering of barriers. The time to target is shortest when barriers are decreasing towards the ground state as in panel (c) of Fig. 1. At the same time, the property of decreasing barriers towards the ground state matches the intuition of folding funnels. In the following section we therefore generalize this picture to arbitrary landscapes.

3 Geometric Funnels

A conformation $x \in X$ is a local minimum if $E_x \leq E_y$ for all $y \in M(x)$. Allowing equality is a mere mathematical convenience [16]. Let \mathbb{P}_{xy} be the set of all walks from x to y . We say that x and y are *mutually accessible at level η* , in symbols

$$x \stackrel{\eta}{\rightleftarrows} y, \quad (3)$$

if there is walk $\mathbf{p} \in \mathbb{P}_{xy}$ such that $E_z \leq \eta$ for all $z \in \mathbf{p}$, respectively. The *saddle height* $\hat{f}(x, y)$ between two configurations $x, y \in X$ is the minimum height at which they are accessible from each other, i.e.,

$$\hat{f}(x, y) = \min_{\mathbf{p} \in \mathbb{P}_{xy}} \max_{z \in \mathbf{p}} E_z = \min\{\eta | x \stackrel{\eta}{\rightleftarrows} y\} \quad (4)$$

The saddles between x and y are exactly the maximal points along the minimal paths in the equation above. We say, furthermore, that a saddle point s *directly connects* the local minima x and y , if (i) $E_s = \hat{f}(x, y)$ and (ii) s has neighbors $s', s'' \in M(s)$ such that there are gradient descent paths $\mathbf{p}_{s'x}$ and $\mathbf{p}_{s''y}$ starting from s' and s'' that end in x and y , respectively. Note that this includes the case that $s' = x$ and/or $s'' = y$.

For simplicity, we assume weak non-degeneracy for our energy landscape as follows. For every local minimum x there is a unique saddle point s_x of minimal height $h(x) =$

Table 1. Definitions of the one-dimensional landscapes in Figure 1 (a)-(d).

Landscape	$E(x) =$	
	x even	x odd
(a)	$x/(N-1)$	$x/(N-1)$
(b)	$x/(N-1)$	$x/(N-1) + 0.35$
(c)	$x/(N-1)$	$1.5x/(N-1)$
(d)	$x/(N-1)/2 + 0.75$	$x/(N-1)$

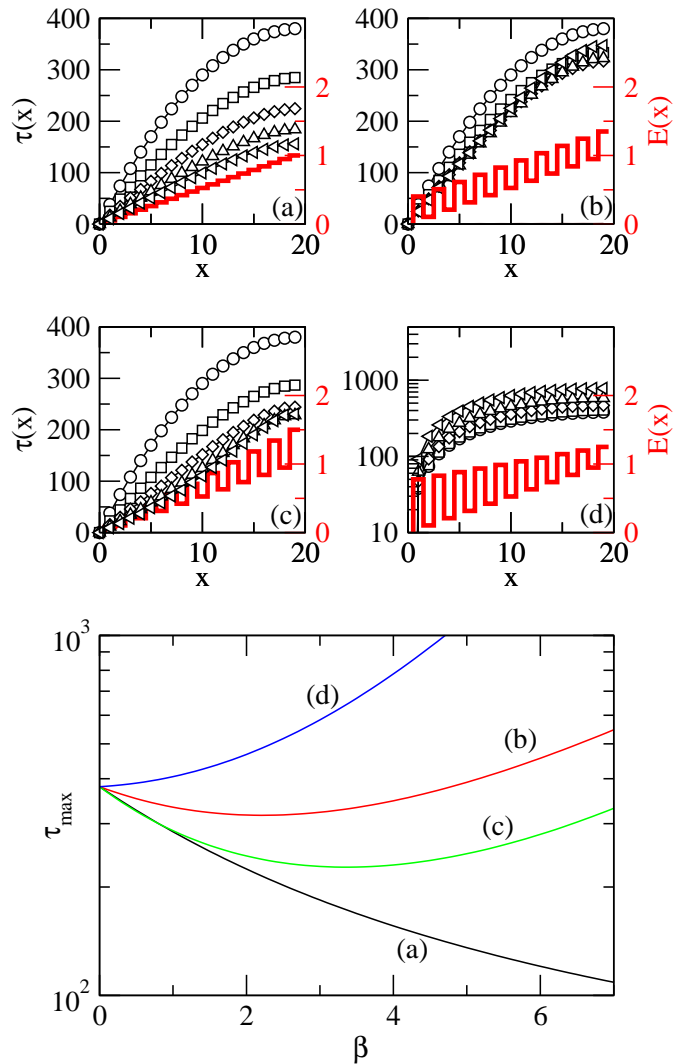


Fig. 1. (a-d) Dynamics on the one-dimensional energy landscapes $E(x)$ (thick curves) defined in Table 1. Thin curves show the average first passage time for state $x_0 = 0$ when starting from given state x , for temperatures $\beta = 0$ (circle), $\beta = 1$ (square), $\beta = 2$ (diamond), $\beta = 3$ (triangle up), $\beta = 4$ (triangle left). Bottom panel: temperature dependence of first passage times in the landscapes (a-d); τ_{\max} is the average time to reach x_0 for the first time starting at the “rightmost” state $x = 19$. Slight changes in the slopes or other details of the landscapes $E(x)$ do not change the qualitative behavior of τ_{\max} as long as the ordering of barrier heights is conserved.

$\min_z \hat{f}(x, z)$. Note that s_x is necessarily a direct saddle between x and some other local minimum z , which for simplicity we again assume to be uniquely determined. This condition is stronger than local non-degeneracy but weaker than global non-degeneracy in the sense of [16]. In particular, it implies that gradient descent paths are also uniquely defined for all initial conditions. In the degenerate case, we consider the set of all direct saddles and the set of the local minima directly connected to them.

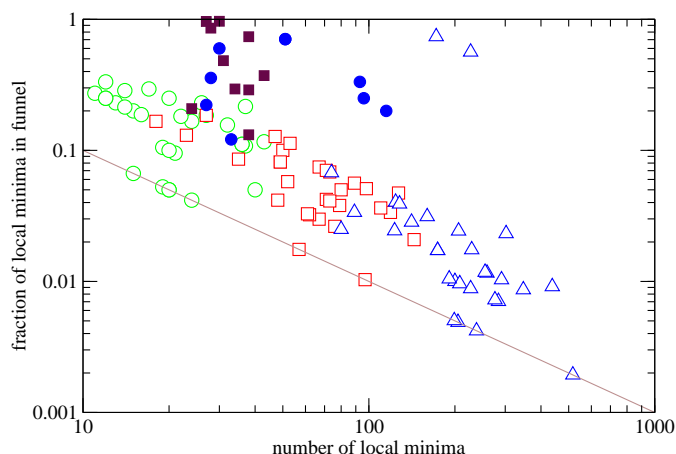


Fig. 2. Fraction of minima belonging to the funnel vs. the total number of minima found in the landscape. RNA hairpins (filled squares) and RNAs with two different near-ground state structures (filled circles). The straight line has slope -1. Landscapes with only the ground state in the funnel fall on this line. Open symbols are the results for the number partitioning problem with (NPP) with sizes $n = 8$ (circles), $n = 10$ (squares) and $n = 12$ (triangles). For each system size, 30 instances were generated by drawing random numbers a_1, a_2, \dots, a_n and c independently from the unit interval. The energy for a state $(x_1, \dots, x^n) \in \{-1, 1\}$ is $E(x) = |\sum_{i=1}^n x_i a_i + c|$. The c represents an extra “clamped” degree of freedom to break the symmetry under reversal of all spins. This ensures that almost all instances have a unique ground state.

With these preliminaries, we are now in the position to define *the funnel* of a landscape recursively as the following set F of states:

- (1) The ground state is contained in the funnel F .
- (2) The local minimum x belongs to the funnel F if a minimum saddle s_x connects x directly to a local minimum in the funnel F .
- (3) A state z belongs to the funnel if it is connected by a gradient descent path to a local minimum in F .

Using the above definition, we can recursively partition the landscape into “local funnels”: Simply remove F from X and recompute the funnel of the residual landscape.

The current implementation of the `barriers` program [15], which given an energy-sorted list of states computes the local minima and the saddle points separating them, also contains a routine that recursively decomposes a minimum height path between a given pair of local minima into a sequence of local minima and saddle points that directly connect two subsequent minima. This function can of course be used in particular to determine all pairs of local minima that are directly connected by a saddle point. It is not hard to see that the `barriers` program can be extended to compute the complete funnel-partitioning of the energy landscapes. We will report on this topic elsewhere in more detail.

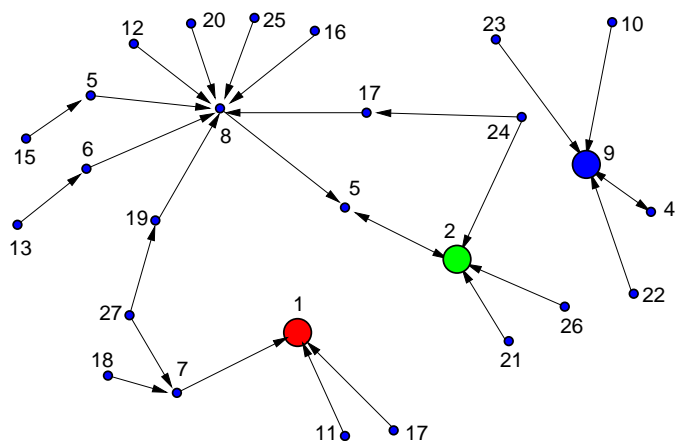


Fig. 3. Funnel partitioning for the folding landscape of the RNA sequence `xbix` (`CUGCGGGCUUUGGCUCUAGCC`). The landscape falls into three funnels. In [1] it was shown that a large part of the folding trajectories reach the metastable state 2 whose energy lies 0.8 kcal above the energy of the ground state 1.

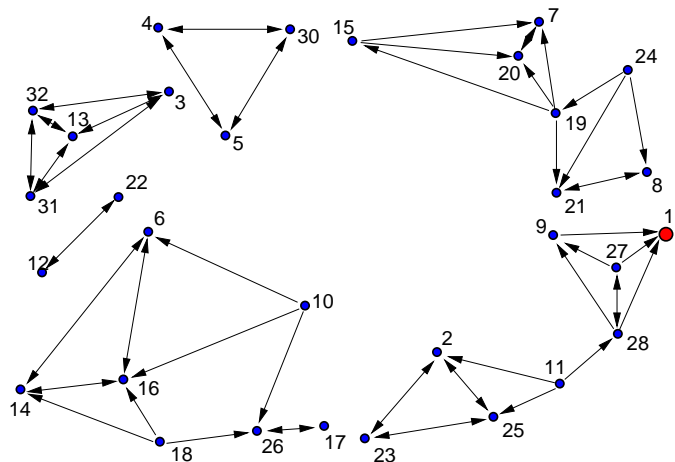


Fig. 4. Funnel partitioning for one instance of the number partitioning problem of size $n = 8$. State 1 is the unique ground state.

4 Examples

As one example of biopolymers we consider small artificial RNA sequences which have been designed either to fold into a single stable hairpin structure or to have two near-ground state structures that have very few base pairs in common. In the first case we expect landscapes dominated by funnels because the `RNAinverse` algorithm [17] tends to produce robustly folding sequences. In the second case we used the design procedure outlined in [18] to produce sequences that have decoy structures with moderate to large basins of attraction. The sequences we use here have a length of 30 nucleotides or less, shorter than most structured RNAs of biological importance.

Figure 2 shows the fraction of local minima contained in the funnels of several landscapes. The RNA folding landscapes have folding funnels comprising a large part

of the landscape. The landscapes of RNA sequences forming hairpins have the largest funnels. For comparison, we plot the relative sizes of funnels for number partitioning problems (NPP) of different sizes, as defined in the caption of Fig. 2. These artificial landscapes have significantly smaller funnels than the RNA folding landscapes. Thus the latter have folding funnels much larger than expected for random rugged landscapes. Through these large funnels the folding polymer may be “guided” towards the native state.

Figure 3 shows an example of an RNA sequence with a strong kinetic trap studied in detail in [1]. In this landscape, a suboptimal structure has a local funnel that covers most of the landscape, while the ground state is separated by comparably high barriers from almost all other local optima. For comparison, the funnel partitioning of an instance of the NPP is shown in Figure 4.

5 Concluding Remarks

In summary, we have introduced here a rigorous definition of a folding funnel that is tractable computationally for arbitrary energy landscapes. In the case of RNA, where the lower fraction of the landscape can be generated without the need for exhaustively enumerating all configurations [19], funnels can be computed explicitly even for sequences that are of immediate biological interest. Our first computational results show that the energy landscapes of RNAs typically differ from the rugged landscapes of spinglass-style combinatorial optimization problems by exhibiting significantly larger funnels for the ground state. It remains to be investigated in future work whether this is also true e.g. for lattice protein models. A second important topic of ongoing research is the question which and to what extent evolutionary processes select molecules with funnel-like landscapes.

Acknowledgments. This work was supported in part by the EMBIO project in FP-6 (<http://www-embio.ch.cam.ac.uk/>).

References

1. M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, P.F. Stadler, *J. Phys. A: Math. Gen.* **37**, 4731 (2004)
2. J.D. Bryngelson, P.G. Wolynes, *Proc. Natl. Acad. Sci. USA* **84**, 7524 (1987)
3. P.E. Leopold, M. Montal, J.N. Onuchic, *Proc. Natl. Acad. Sci. USA* **89**, 8721 (1992)
4. J.N. Onuchic, Z. Luthey-Schulten, P.G. Wolynes, *Annu. Rev. Phys. Chem.* **48**, 545 (1997)
5. O.V. Galzitskaya, A.V. Finkelstein, *Proc. Natl. Acad. Sci. USA* **96**, 11299 (1999)
6. J.N. Onuchic, H. Nymeyer, A.E. Garcia, J. Chahine, N.D. Socci, *Adv. Protein Chem* **53**, 87 (2000)
7. B. Fain, M. Levitt, *Proc. Natl. Acad. Sci. USA* **100**, 10700 (2003)
8. K. i. Okazaki, N. Koga, S. Takada, J.N. Onuchic, P.G. Wolynes, *Proc. Natl. Acad. Sci. USA* **103**, 11844 (2006)
9. J.N. Onuchic, P.G. Wolynes, *Curr. Opin. Struct. Biol.* **14**, 70 (2004)
10. T. Klotz, S. Kobe, *J. Phys. A: Math. Gen* **27**, L95 (1994)
11. O.M. Becker, M. Karplus, *J. Chem. Phys.* **106**, 1495 (1997)
12. P. Garstecki, T.X. Hoang, M. Cieplak, *Phys. Rev. E* **60**, 3219 (1999)
13. D.J. Wales, M.A. Miller, T.R. Walsh, *Nature* **394**, 758 (1998)
14. J.P. Doye, M.A. Miller, D.J. Welsh, *J. Chem. Phys.* **111**, 8417 (1999)
15. C. Flamm, W. Fontana, I. Hofacker, P. Schuster, *RNA* **6**, 325 (2000)
16. C. Flamm, I.L. Hofacker, P.F. Stadler, M.T. Wolfinger, *Z. Phys. Chem.* **216**, 155 (2002)
17. I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster, *Monatsh. Chem.* **125**, 167 (1994)
18. C. Flamm, I.L. Hofacker, S. Maurer-Stroh, P.F. Stadler, M. Zehl, *RNA* **7**, 254 (2000)
19. S. Wuchty, W. Fontana, I.L. Hofacker, P. Schuster, *Biopolymers* **49**, 145 (1999)