# A Sampling Approach for the Exploration of Biopolymer Energy Landscapes

Richter, A.S.; Will, S.; Backofen, R.

University of Freiburg, Bioinformatics Group, Georges-Köhler-Allee 106, 79110 Freiburg, Germany

**Abstract**
The folding of biopolymers is crucially determined by the properties and the topology of the underlying energy landscape. A reduced representation of these energy landscapes is provided by barrier trees, which can be used to study the dynamical behavior of the folding. We presented a generic, problem-independent approach for the generation of barrier trees of discrete biopolymer models. In contrast to previous studies, the approach used does not rely on enumeration, which is limited to smaller molecules due to the amount of available memory. The algorithm has been applied to RNA and a lattice protein. The results show that the approach can be used to compute both all local minima and the exact barrier tree of an energy landscape. The presented method does not restrict the investigated conformation space to certain regions.

## 1. Introduction

The three-dimensional structure of RNA and proteins is vital for their biological function. An insight into the structure formation process can be gained by the study of the energy surface, on which the folding proceeds. These energy landscapes exhibit the same geometrical features as natural occurring landscapes, like mountains, valleys, plains, ridges and so on, but they are multidimensional. Typical characteristics of a landscape, like the number of local optima, the basin distribution as well as the transition states between the optima, can be conveniently visualized in the manner of a barrier tree. These barrier trees provide a reduced representation of energy landscapes, and they are a very useful description for the study of biopolymer folding pathways. They also give an appropriate impression of the overall topology of the energy landscape. Having the underlying energy landscape of a biopolymer at hand, the folding dynamics of the molecule can be investigated, and properties of the folding landscape like kinetic traps can be unveiled. See Figure 1 for a schematic energy landscape representation and the associated barrier tree.

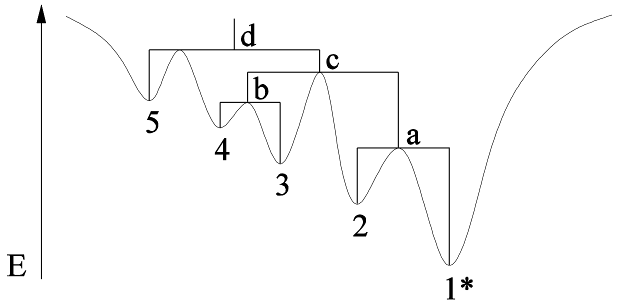


**Figure 1** Schematic representation of an energy landscape and its associated barrier tree. The local minima are marked with numbers, and their connecting saddle points are marked with lowercase letters. The global minimum of the energy landscape is marked with an asterisk (taken from ref. [12]).

A stochastic algorithm to simulate the folding kinetics of RNA sequences into secondary structures was presented by Flamm et al. [3]. The stochastic simulation considers all legal RNA structures, which makes it very time-consuming and computationally intensive. The barrier tree of an energy landscape can be computed by exhaustive enumeration of all conformations as implemented in the program `barriers` [3]. Since the search space grows exponentially with the length of the molecule even in simplified models [6,9], its use is limited to landscapes of modest size. Different kinetics for the study of RNA folding based on barrier trees have been presented in [11]. A generic algorithm to generate and explore the lower part of energy landscapes was presented by Wolfinger et al. [12]. The approach was applied to lattice protein energy landscapes. However, the method is limited by the available memory and enumerates only structures below a given energy threshold. The resulting barrier trees represent just a partial landscape.

In this work, we present a method to approximate the barrier tree of an energy landscape. The used sampling approach does not restrict the search space, enables arbitrary approximation accuracy and can be implemented memory efficient.

## 2. Methods

### Energy Landscapes

A biopolymer energy landscape can be described formally by the following three parts (cf. [8] for a more general definition):

- a set $X$ of conformations,
- an operator $N : X \to \mathrm{P}(X)$, which defines the neighborhood of a conformation $x$ in $X$ and assigns to each conformation a set of accessible neighbors $N(x)$, and
- an energy function $E : X \to \Re$.

The conformation space is formed by the conformation set $X$ in combination with the neighborhood operator $N$. An energy function $E$ is called *non-degenerate,* if $E(x) = E(y) \leftrightarrow x = y$. A conformation $\hat{x}$ in $X$ is called *local minimum*, if for all $y$ in $N(\hat{x}) : E(\hat{x}) \leq E(y)$. A conformation $\hat{x}$ in $X$ is called a *global minimum*, if for all $y$ in $X : E(\hat{x}) \leq E(y)$. A list of conformations $x = x_1, \ldots, x_k = y$ with $x_i$ in $X$ for all $1 \leq i \leq k$ and $x_i$ in $N(x_{i+1})$ for all $1 \leq i < k$ is a *walk* between the conformations $x$ and $y$. The term *random walk* denotes an arbitrary, randomly chosen walk between two conformations. A walk is called an *adaptive walk*, if only neighbors with a lower energy are accepted. It is called a *gradient walk*, if in each step the neighbor with the minimal energy has to be chosen. The *saddle height* $E[\hat{x}, \hat{y}]$ between two local minima $\hat{x}$ and $\hat{y}$ is the minimum height that makes them accessible from each other, that is $E[\hat{x}, \hat{y}] = \min\{\max[E(s) \mid s \in w] \mid w : \text{walk from } \hat{x} \text{ to } \hat{y}\}$. A point *s* in *X* that satisfies this condition is called a *saddle point*. In non-degenerate landscapes, each saddle point is unique. The *barrier* of a local minimum is the height of the lowest saddle point that has to be overcome in order to reach a more favorable local minimum.

The local minima and the saddle points connecting the metastable states can be represented in a unique hierarchical structure called the *barrier tree* of the energy landscape. An example of a barrier tree with a schematic representation of the underlying energy landscape is given in Figure 1. In this example, the local minima marked with the numbers 2 and 3 are accessible to each other by the saddle point $c$. The saddle height $E[2,3]$ corresponds to $E(c)$. The energy barrier of the minimum 3 is $E(c) - E(3)$.

### RNA Secondary Structures

Coarse-grained discrete structure models reduce the level of detail to allow computational studies. An *RNA secondary structure S* is a list of Watson-Crick (A-U, G-C) or non-standard (G-U) base pairs $(i, j)$ with the conditions that (1) each base $i$ can pair with at most one other base $j$, and that (2) there are no two pairs $(i, j)$ and $(k,l)$ with $i < k < j < l$ [10]. A structure satisfying the second condition is called non-crossing and does not contain pseudo-knots. An RNA secondary structure can be visualized as a planar secondary structure graph or as a string in the bracket notation with two matching parentheses symbolizing a matching base pair and dots representing unpaired bases.

The abstract parts of energy landscapes are defined as follows for RNA: the conformation set $X$ of a given RNA sequence $s$ is the set of all secondary structures $S$, or conformations, that are compatible with $s$. The neighborhood of a conformation $S$ in $X$ is defined by single moves. A single move assigns a structure $S_x$ in $X$ a neighbor $S_y$ in $X$ by removal or insertion of a single base pair $(i, j)$ in compliance with the restriction that no pseudo-knots are allowed. The energy function is defined according to the nearest neighbor energy model, where the energy of an RNA structure is assumed to be equal to the summed up energy contributions of all secondary structural elements the structure can be decomposed into. Zuker and Stiegler formulated a dynamic programming algorithm for the calculation of minimum free energy structures using this energy model [14]. An algorithm that generates all suboptimal conformations below a certain energy threshold was presented by Wuchty et al. [13].

### Lattice Proteins

A simple and well-known coarse-grained protein model is the *HP-Model* proposed by Lau and Dill [5]. It reduces the 20-letter alphabet of the amino acids to a two-letter alphabet, consisting of H, which represents hydrophobic amino acids, and P, which represents polar or hydrophilic amino acids. Since it is commonly believed that the hydrophobic force is dominant in protein folding, the energy function only favors contacts between H-monomers. Only the backbone structure of the protein is modeled, that is one position for each amino acid. These positions are restricted to discrete positions on a geometrical structure that is known as lattice.

The conformation set consists of all self-avoiding walk structures that have the length of a given sequence $s$. The organization of the conformation space is described by pivot moves, which are rotations or reflections of the conformation behind a certain monomer. The energy function is given by the sum of the pair wise contact potentials of the structure. Optimal structures of simplified proteins on different lattices can be predicted using an approach based on constraint programming [2]. A structure is optimal, if there are as many contacts between H-monomers as possible. For the purpose of protein structure prediction, maximally compact hydrophobic cores are enumerated first. Subsequently, one tries to place the protein sequence on a compact core.

### Sampling of the Energy Landscape

In this study, the barrier tree of an energy landscape was constructed without exhaustive enumeration of all possible biopolymer structures. Instead, it was approximated by a sampling over the conformation space of the biopolymer.

The following sampling strategy was used:

- Choose a minimum $x$ out of all local minima that are already known.
- Perform a random walk of given length that starts from $x$ and terminates in conformation $x_n$. Save the highest energy value $E_{max}$ during the random walk.
- Perform an adaptive walk starting from $x_n$. The walk terminates in a local minimum $y$.
- If the minimum $y$ is not yet known, add $y$ to the barrier tree and add the mutual accessibility energy $E_{max}$ between $x$ and $y$. If the minimum $y$ is already known, and if $E_{max}$ is lower than the current mutual accessibility energy $E_{curr}$ between the minima $x$ and $y$ in the barrier tree, replace $E_{curr}$ by $E_{max}$.
- Iterate from step 1 while the termination condition is not fulfilled.

Figure 2 illustrates the whole sampling approach.

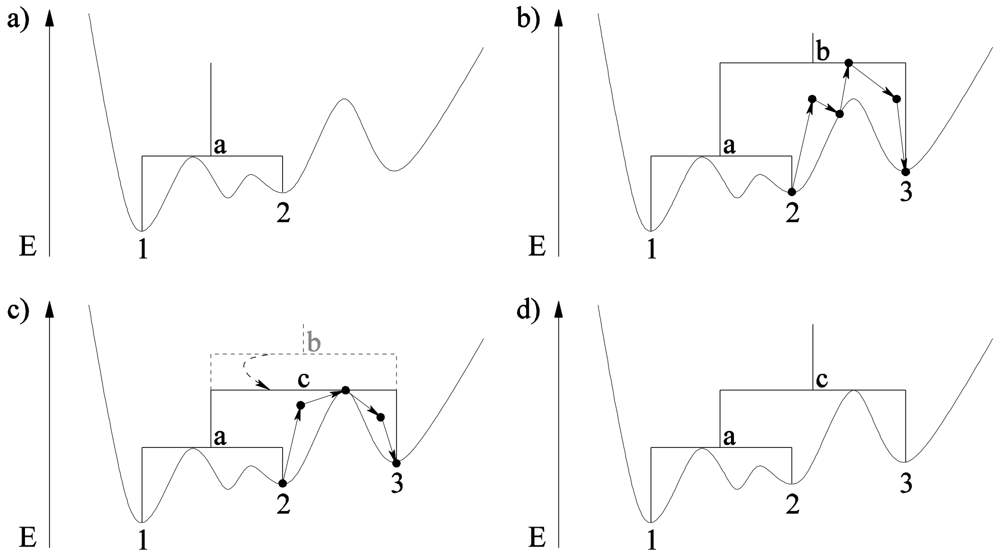


**Figure 2** Sampling of the energy landscape. a) A given barrier tree with two local minima 1 and 2, connected by the saddle point *a*. b) Choose minimum 2 randomly as random walk start conformation. The subsequent adaptive walk starts from the end conformation of the random walk and terminates in the local minimum 3. *E*(*b*) is the highest energy value during the random walk. Since minimum 3 is not yet known, add it to the barrier tree and connect it to minimum 2 by the saddle point *b* with the height *E*(*b*). c) Another sampled walk between minima 2 and 3 results in the saddle point *c* connecting 2 and 3. Since *E*(*c*) < *E*(*b*), update the estimated saddle height *E*(*b*) between 2 and 3 to the correct saddle height *E*(*c*). d) The resulting barrier tree of the energy landscape.

The resulting barrier tree is an approximation for the exact barrier tree of energy landscape. The approximation quality depends on the sampling and is increased with the number of sampling iterations. The whole sampling process can be controlled by the number of iterations, the length of the random walk, and the way of choosing the start conformation from the barrier tree for each iteration step. To favor low-energy conformations, the frequency of choosing an optimum as start conformation was proportional to the Boltzmann weight of the optimum. The sampling also requires a good set of start conformations. Optimal and near-optimal structures of the given RNA sequence were predicted by the approach of Wuchty et al. For proteins, constraint-based protein structure prediction was used.

The abstract barrier tree data structure utilized by the sampling algorithm has to meet the demand that the information gained by the sampling is stored memory efficient. Furthermore, the data structure should support several basic operations like, for instance, insertion of a new local minimum and update of an estimated saddle height with moderate time complexity. A binary tree meets these demands. When using a binary tree, the current barrier tree is always available for visual inspection and can be used to guide the sampling.

## 3. Results

### Implementation of Energy Landscape Models

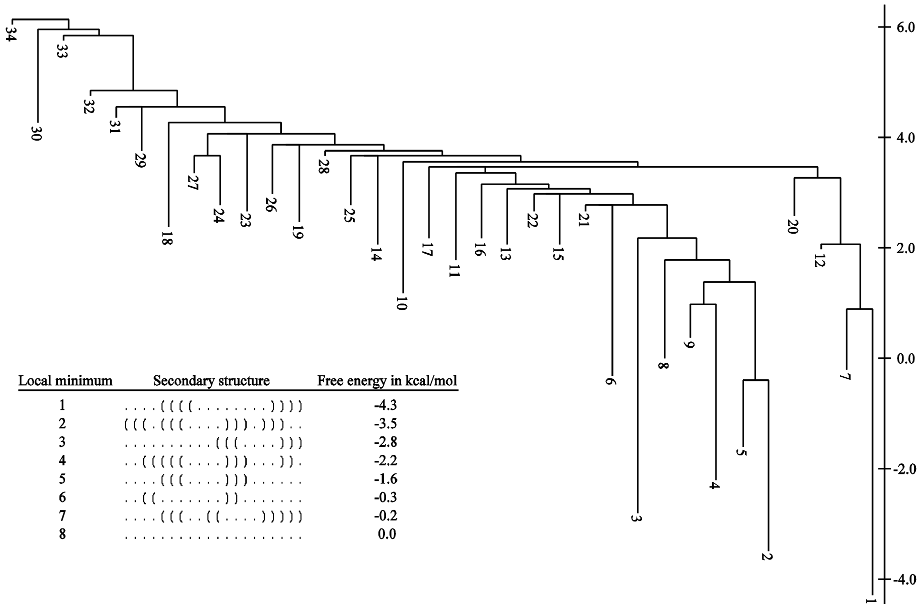
The presented sampling approach is generic; this means that it is not dependent on the underlying energy landscape model. Thus, a framework for the study of arbitrary landscapes is needed. The landscape models have to define at least a set of conformations and a neighborhood of the conformations in order to form the conformation space and an energy function over the conformations. The *Energy Landscape Library* (ELL) developed by our group meets these basic requirements [7]. The ELL currently implements energy landscape models for RNA secondary structures and for structures of simple lattice proteins. Due to the fact that the

ELL is highly modular, all available landscape models can be extended in a simple way and new models can be implemented in a straightforward manner.

## Experiments

An RNA example was used to assess if the sampling approach is capable of finding the exact barrier tree of the energy landscape. For RNA secondary structures, efficient algorithms exist to enumerate suboptimal structures. Thus, exact barrier trees for small molecules can be computed [3]. This provides the possibility to verify the barrier trees obtained by the sampling approach. A lattice protein example demonstrates the capabilities of the presented approach. In contrast to RNA secondary structures, no efficient algorithm for the enumeration of suboptimal structures below a certain energy level exist for lattice proteins. Hence, no exact barrier trees were available for comparison with the sampled ones.

As a first example, an artificially designed RNA molecule of length 20 with the sequence `CUGCGGCUUUGGCUCUAGCC` denoted `xbix` [11] was chosen. The conformation space of this molecule consists of 3886 secondary structures. Figure 3 shows the exact barrier tree of `xbix`. The tree was computed from a list of all conformations generated by `RNAsubopt`, which is part of the `Vienna RNA Package` [4]. The barrier tree has 34 local minima. The mfe structure `....((((........))))` has an energy of -4.3 kcal/mol and is represented by minimum 1. To assess our approach, a sampling of the energy landscape was performed, starting from the mfe conformation. The sampling was terminated as soon as the exact barrier tree has been found. On average over 10 runs, all 34 local minima were found after 1246 sampling iterations. The exact barrier tree was found after 25960 iterations on average.



| Local minimum | Secondary structure | Free energy in kcal/mol |
|---|---|---|
| 1 | `....((((........))))` | -4.3 |
| 2 | `(((.(((....))).)))..` | -3.5 |
| 3 | `..........(((....)))` | -2.8 |
| 4 | `..(((((....)))...)).` | -2.2 |
| 5 | `....(((....)))......` | -1.6 |
| 6 | `..((........))......` | -0.3 |
| 7 | `....(((..((....)))))` | -0.2 |
| 8 | `....................` | 0.0 |

**Figure 3** Barrier tree of the artificially designed RNA sequence `xbix` and the eight lowest local minima of the energy landscape.

The 27-mer HPNX sequence `HHXHPHHHNPHHPHHHHNHPHNHHHNP` in the three-dimensional cubic lattice was chosen as lattice protein example. The start conformations for the sampling were provided by constraint-based protein structure prediction [2]. The starting set consisted of the unique ground state with *E* = -80 and four suboptimal conformations on the first energy level with *E* = -79. The energy landscape sampling was stopped after 7 million iterations. The barrier tree resulting from a sampling is given in Figure 4. The tree shows the 150 lowest local minima. Altogether, 6444934 different local minima were found. Since symmetrical structures were not excluded, the tree shows 5 ground states due to rotations and reflections. These states have an energy of -80 and are labeled with 1 to 5 within the tree. Besides this, 20 suboptimal conformations with an energy of -79 were found. Previous studies of the landscape of this sequence gave a barrier tree with a single minimum with *E* = -80 and 4 minima with *E* = -79 [12]. The optimal and near-optimal conformations were connected via saddle heights in an energy range of -40 to -50. In the barrier tree found by the sampling approach, the optimal conformations are mutually accessible by energies between -30 to -40. This shows that the resulting barrier tree is just an approximation of the exact one.

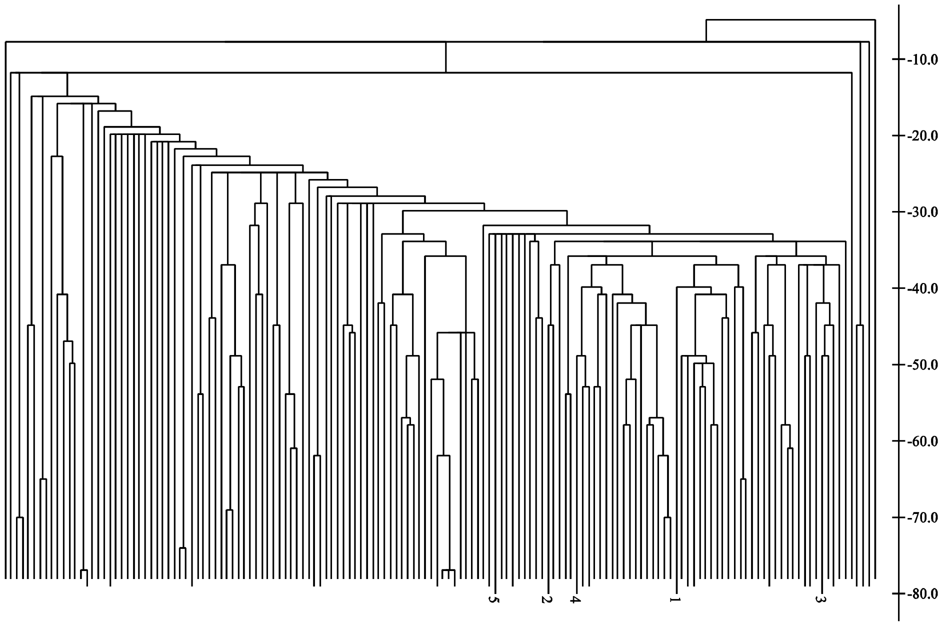


**Figure 4** Barrier tree of the HPNX-kind lattice protein after 7 million sampling iterations.

### Discussion of Results

In the RNA example, the approach for the sampling of energy landscapes yielded the exact barrier tree. With our method, all local minima of the landscape were found after a small number of sampling steps. Consequently, the sampling approach can be used to determine the local minima of biopolymer folding landscapes, at least for short sequences. To compare the barrier tree resulting from the sampling with the exact barrier tree, the root mean square deviation (RMSD) over their saddle heights was used. The RMSD soon reached the value zero, which means that the barrier tree obtained by the sampling agreed with the exact barrier tree. Accordingly, the sampling approach is capable of resulting in the correct barrier tree of an energy landscape. The runtime of the sampling implementation was a few seconds, which is by all means acceptable.

Applied to lattice proteins, the sampling approach found significantly more local minima than the previous “flooding” approach, which is based on enumeration of low-energy conformations [12]. Symmetrical structures of optimal and suboptimal conformations were found, and thus a larger region of the energy landscape was covered. Beyond this, all local minima are connected. The “flooding” approach usually gives non-connected subtrees. Therefore, direct or minimal refolding paths between non-connected minima have to be sampled. An drawback of the presented sampling method is that the estimated saddle heights are just an upper bound of the exact saddle heights.

The barrier tree of the lattice protein showed that there are several optimal and suboptimal conformations with exactly the same energy. This high degree of degeneracy is a common feature of lattice protein energy landscapes. It can be seen as an artifact of the underlying model, which uses a limited alphabet size and fixed bond lengths and angles. At this point, it seems reasonable to ask whether it is correct to model proteins with reduced alphabets as two or four-letter alphabets. Several experimental studies have shown that functional and rapidly folding proteins do not necessarily require the full sequence complexity of naturally occurring proteins (see for example [1]). Since the approach presented here is generic and problem-independent, it can be readily applied to more realistic models with an alphabet that is larger than the four-letter HPNX alphabet or with more complex lattices like the face-centered cubic lattice (FCC). However, because the degrees of freedom increase in the FCC lattice, the size of the conformation space increases as well. Thus, more sampling iterations are required to obtain a good barrier tree approximation.

## 4. Conclusion

Barrier trees provide a coarse-grained representation of energy landscapes by organizing local minima and saddle heights in a hierarchical structure. They are a very useful tool for the study of biopolymer folding pathways. We developed a random sampling approach, which allows computing the exact or approximated barrier tree of the energy landscape. Using this method, the investigated conformation space of the landscape is not restricted to certain regions. Thus, in comparison to the current approaches for lattice proteins, more local minima were found, and the resulting barrier trees covered a larger region of the energy landscape. The sampling method has the advantage that, in contrast to methods based on enumeration, the number of sampled conformations for the barrier tree construction is not basically restricted by the available amount of memory. However, the estimated saddle heights within the resulting barrier trees are just an approximation of the saddle heights and can therefore be higher than within the exact barrier tree of the energy landscape. The sampling of direct paths between local minima of the barrier tree is a possible way to find better approximations of the saddle heights. The comparison of different approaches for the exploration of energy landscapes, namely selective enumeration and sampling of conformations, indicates that a strategy which combines the two methods could achieve very promising results. The sampling approach is capable of finding a huge number of minima and could therefore be used to roughly characterize the energy landscape. Afterwards, the “flooding” approach could be used to calculate the exact barrier tree of certain landscape regions by selective enumeration starting from minima that were found by the sampling.

Altogether, the sampling approach appears to be a promising technique for the computation of barrier trees as reduced representation of discrete biopolymer model energy landscapes. The barrier trees can be used as basis for the estimation of basin sizes. Moreover, they are a good starting point for the calculation of folding kinetics.

## References

[1] Akanuma, S.; Kigawa, T.; Yokoyama, S. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc. Natl. Acad. Sci. (USA), 99(21):13549–13553, 2002.

[2] Backofen, R.; Will, S. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. Journal of Constraints, 11(1):5–30, 2006.

[3] Flamm, C.; Fontana, W.; Hofacker, I.L.; Schuster, P. RNA folding at elementary step resolution. RNA, 6(3):325–338, 2000.

[4] Hofacker, I.L.; Fontana, W.; Stadler, P.F.; Bonhoeffer, S.; Tacker, M.; Schuster, P. Fast folding and comparison of RNA secondary structures. Monatshefte Chemie, 125(2):167–188, 1994.

[5] Lau, K.F.; Dill, K.A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules, 22(10):3986–3997, 1989.

[6] Madras, N.; Slade, G. The self-avoiding walk. Probability and its applications. Birkhäuser Boston, 1996.

[7] Mann, M.; Will, S.; Backofen, R. The Energy Landscape Library – a platform for generic algorithms. In Proc. of 1st BIRD'07, volume 217, pages 83–86. OCG, 2007.

[8] Stadler, P.F. Fitness landscapes. In Michael Lässig and Angelo Valleriani, editors, Biological Evolution and Statistical Physics, pages 187–207, Berlin, Germany, 2002. Springer-Verlag.

[9] Waterman, M.S. Introduction to computational biology: maps, sequences and genomes. Chapman & Hall, London, UK, 1995.

[10] Waterman, M.S.; Smith T.F. RNA secondary structure: a complete mathematical analysis. Math. Biosci., 42(3–4):257–266, 1978.

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

[12] Wolfinger, M.T; Will, S.; Hofacker, I.L.; Backofen, R.; Stadler, P.F. Exploring the lower part of discrete polymer model energy landscapes. Europhys. Lett., 74(4):726–732, 2006.

[13] Wuchty, S.; Fontana, W.; Hofacker, I.L.; Schuster, P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers, 49(2):145–65, 1999.

[14] Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research, 9(1):133–148, 1981.