The Energy Landscape of RNA Folding

DIPLOMARBEIT
zur Erlangung des akademischen Grades
Magister rerum naturalium

Vorgelegt der
Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

von
Michael Wolfinger

am Institut für Theoretische Chemie und Molekulare Strukturbiologie

im März 2001

Dank an alle,

die zum Gelingen dieser Arbeit beigetragen haben:

Peter Stadler, Ivo Hofacker, Peter Schuster.

Ingrid Abfalter, Aderonke Babajide, Jan Cupal, Martin Fekete, Christoph Flamm, Dagmar Friede, Kurt Grünberger, Jörg Hackermüller, Christian Haslinger, Philipp Kobel, Michael Kospach, Ulli Mückstein, Stefan Müller, Bärbel Stadler, Roman Stocsits, Andreas Svrcek-Seiler, Caroline Thurner, Günther Weberndorfer, Andreas Wernitznig, Stefanie Widder, Christina Witwer, Daniela Dorigoni, Judith Jakubetz.

Frieda Wolfinger, Sepp-Jürgen Wolfinger.

Zusammenfassung

Durch die Entdeckung, daß RNA-Moleküle katalytische Aktivität besitzen können, wandelte sich die Sichtweise von der Bedeutung der RNA für die lebende Zelle drastisch. Wurde RNA einst als reiner Informationsvermittler zwischen DNA und den Proteinen gesehen, ist heute klar geworden, daß RNA eine aktive Rolle bei der Regulation vieler Vorgänge in lebenden Zellen spielt. Die Struktur von RNA wird durch Kontakte individueller Nukleotide, sogenannter Basenpaare, aufgebaut. Aufeinanderfolgende Basenpaare bilden helicale Bereiche, welche auf die Struktur im Gegensatz zu den ungepaarten Bereichen einen stabilisierenden Einfluß ausüben. Der Faltungsprozess von RNA gehorcht einer hierachischen Ordnung. Stabile Sekundärstrukturelemente falten schnell und bestimmen die anschließende Faltung in die dreidimesionale Struktur. Am Institut für Theoretische Chemie und Molekulare Strukturbiologie wurden über die letzten Jahre effiziente Algorithmen, die sich auf experimentell gemessene Energieparameter stützen, entwickelt und der Allgemeinheit als `Vienna RNA Package` zugänglich gemacht. Mit dem `Vienna RNA Package` lassen sich die Sekundärstruktur sowie die thermodynamischen Eigenschaften von beliebigen RNA-Molekülen berechnen.

Neben den thermodynamischen Eigenschaften von RNA spielt die Kinetik der Faltung eine wichtige Rolle. Kürzlich wurde ein Algorithmus vorgestellt (Flamm et. al. 2000), der das Studium der Faltungsdynamik von RNA-Molekülen ermöglicht. Da es sich hierbei um ein stochastisches Verfahren handelt, muß eine große Anzahl von Trajektorien berechnet werden, was extrem rechen- und zeitintensiv ist. In der hier vorgestellten Arbeit wird das stochastische Verfahren durch ein deterministisches ersetzt. Durch eine Verringerung des Zustandsraumes auf die lokalen Minima der Energiehyperfläche lässt sich das Problem als Markov Prozess in kontinuierlicher Zeit formulieren. Die Dynamik des Moleküls, insbesondere die Besetzungswahrscheinlichkeiten der einzelnen lokalen Minima können so bequem und vor allem binnen kürzester Zeit berechnet werden.

Neben einer eleganten Methode zur Visualisierung und Berechnung der Energielandschaften wird auch die Faltungskinetik eines RNA-Moleküls, das eine metastabile Strukturn ausbilden kann (bi-stabile RNA oder RNA switch) gezeigt.

Abstract

Within the last years it became clear that RNA molecules do not only store and transfer genetic information but they can also act as catalytic units. On the one hand, RNA was seen once as an 'information-exchanging-unit' between DNA and proteins. On the other hand nowadays it has become evident that RNA fulfills important regulation tasks in living cells, which changed the point of view dramatically. The structure of RNA is determined by contacts of individual nucleotides, so called base pairs. Successing base pairs form helical regions which have a stabilizing effect on the structure. The structure formation process of RNA is thought to be of hierarchical order. Stable secondary elements fold fast and determine the three-dimensional fold. Within the last years, efficient algorithms, which are based upon experimentally measured energy parameters, have been developed at our institute and made available as the `Vienna RNA Package`. With aid of the `Vienna RNA Package`, secondary structures and thermodynamic properties of RNA molecules can be calculated.

Evidently, the kinetics of RNA folding plays an important role besides thermodynamic properties. Recently, an algorithm has been suggested, which allows studying the folding behavior of RNA molecules. Due to the fact that this is a stochastic model, many trajectories have to be calculated, which is very intensive in terms of time ans computer resources. Hence, we replace the stochastic model with a deterministic one in this thesis. By reducing the state space to the local minima of the energy landscape, the problem can be formulated as a continuous time Markov chain. With this ansatz, the dynamic behavior of the molecule, especially population probabilities of distinct local minima of the energy landscape can be calculated within seconds.

We present an elegant method for the computation and visualization of the energy landscape and the folding dynamics of a RNA molecule that can fold into a metastable secondary structure (bi-stable RNA or RNA switch).

Contents

1 Introduction **6**

2 Thermodynamic Folding **10**
2.1 RNA Structure 10
2.2 RNA Secondary Structure 11
2.3 Conformation Space: The Thermodynamic View 18

3 Kinetic Folding **22**
3.1 State of the Art 22
3.2 The Move Set 23
3.3 Conformation Space: The Kinetic View 28

4 Energy Barriers **31**
4.1 Definitions 31
4.2 The Algorithm of `barriers` 35
4.3 Degenerate Saddles 41

5 Markov Chains **45**
5.1 Stochastic Processes 45
5.2 Continuous Time Markov Chains 47
5.3 The Model 49
5.4 The Barrier Tree Approximation 54
5.5 Reliability and Comparison 57

6 Computational Results **59**
6.1 A first example: sexi 59
6.2 The medium-sized molecule: bertl 63
6.3 The switching molecule 66

7 Conclusion and Outlook **74**

Appendix A **78**

Appendix B **79**

List of Figures **84**

References **85**

1 Introduction

When Charles Darwin put forward a first empirical theory of biological evolution in his famous book "*The Origin of Species*" in 1859, he suggested that the diversity and complexity of present day organisms can be explained on the basis of two key principles: inheritable *variation* and natural *selection*. His theory quickly became one of the most influential contributions to natural science and although the laws and mechanisms of variation had not been accepted in the nineteenth century, he set a cornerstone to a modern view of the basis of life.

More than a century later the *molecular* basis of life has become clear: Biopolymers like DNA, RNA and proteins are the essential ingredients in the 'cookbook of life'. This thesis focuses on RNA. RNA molecules do not only serve as carriers of information, but also as functionally active units. The three dimensional shape of tRNA molecules plays a crucial role in the process of protein synthesis. RNA is known to exhibit catalytic activity [8, 24, 23, 33]. While the activity of these so called "ribozymes" is usually restricted to cleavage and splicing of RNA itself, recent evidence suggests that RNA also plays a predominant role in ribosomal translation. These discoveries have given much support to the idea that an *RNA World* [21, 34, 35] stood at the origin of life, in which RNA served both as carrier of genetic information as well as catalytically active substance. RNA may not necessarily have been the first step in prebiotic evolution, but the idea that RNA preceded not only DNA, but also the invention of the translational system, seems widely accepted. Furthermore, RNA provides an ideal, currently the only, system to study genotype-phenotype relationships. Following [52], the phenotype for an RNA molecule can be defined as its spacial structure.

An interesting aspect concerning biomolecules is structure prediction. The structure prediction problem for both proteins and RNA can be solved with reasonable accuracy at the level of secondary structures. Due to the fact that the process of RNA folding is thought to be of hierarchical nature [6], secondary structures can be seen as a coarse grained approach to the three dimensional structures. In the protein case, the secondary structure is defined as the local conformation of the backbone and is formed by hydrogen bonds between backbone atoms. In the RNA case, the secondary structure is defined as a pattern of base pairs, which is determined by hydrogen bonds between the four bases Adenine (A), Guanine (G), Cytosine (C) and Uracil (U). It is important to realize that there is a major difference between protein and RNA secondary structures: While in the protein case, secondary structures are formed by the backbone, RNA secondary structures are determined by *side chains*. Powerful algorithms have been suggested within the last 25 years to make the computational treatment of RNA feasible [46, 59, 65]. A freely available implementation of these algorithms is the `Vienna RNA Package` [27, 28].

An important contribution to the understanding of the behavior of RNA molecules was given by Stefan Wuchty [60] who introduced a tool which allows the computation of *all* suboptimally folded RNA molecules within a desired energy range above the ground state. This opened the door for a more thorough investigation of the dynamical behavior of RNA chains. More generally, with knowledge about all suboptimal secondary structures, an insight into the energy landscape of RNA was given (for a thorough introduction to characterization and computation of general landscapes, see [53]). It became necessary to introduce a metric between different secondary structure, called *move set*. This move set influences the shape of the energy landscape dramatically: Depending on which combinations of opening and closing of base pairs are allowed, the energy landscape can be very rugged or more or less smooth (figure 1). Nevertheless, RNA landscapes are thought to be extremely rugged.

Recently, a stochastic model of the *kinetic* folding of RNA molecules was suggested by Christoph Flamm [18]. His algorithm uses the most elementary move set for the inter-conversion of RNA secondary structures, consisting of the insertion or the removal of single base pairs, as well as the exchange of one pairing partner in a base pair. As the structural changes made during

(a) Rugged landscape: Bryce Canyon, UT

(b) Smooth landscape: Capulin Volcano, NM

Figure 1: A rugged landscape (a) is rocky and separated into many local maxima and minima, whereas a smooth landscape (b) can be traversed without many uphill and downhill climbs

one simulation are small, a realistic concept of folding paths arise. With aid of his tool it was possible to show that the folding simulation of SV11, a RNA molecule which can form a metastable structure, is in excellent agreement with experimentally measured data. Molecules like SV11, which can form one or more metastable secondary structures are considered as so called *molecular switches* which can fulfill essential regulation tasks in living cells.

The energy landscape of a given RNA sequence is determined by (a) all legal secondary structures the molecule can fold into, (b) the energies of all these secondary structures and (c) the move set. The properties of the energy landscape affect the folding kinetics of the RNA sequence. Energy landscapes can be visualized with so called *barrier trees*, see section 4.1 for the exact definitions. Due to the fact that the algorithm presented in [18] is a stochastic process, a large number of folding trajectories has to be calculated to make a reasonable prediction of the kinetic folding behavior of the molecule, which is very intensive in terms of computer time and resources. This is evident, because *all* suboptimal secondary structures have to be considered within such a simulation.

A reasonable approach to overcome this problem is to reduce the size of the state space of the system, e.g. by just allowing local minima of the energy landscape as legal states. With transition rates between different local minima depending on the energy berrier separating them, it is possible to formulate a *Markov process* describing the dynamic behavior of the RNA molecule. With this ansatz it is possible to assign specific population probabilities to different local minima at the beginning of the simulation. Depending on these initial conditions and the energy ratio in terms of the barrier height between different states, several other local minima are being populated as time elapses ending in a predetermined equilibrium distribution. In other words, the *whole* dynamic behavior of a desired RNA sequence can be visualized with this new method.

All software tools described in this thesis are written in `ANSI C` and tested under the free operating system Linux.

2 Thermodynamic Folding

2.1 RNA Structure

RNA is transcribed (or synthesized) in cells as single strands of (ribose) nucleic acids. However, these sequences are not simply long strands of nucleotides. Rather, intra-strand base pairing will produce structure motives. The structure formation process of RNA can conceptually be partitioned into two consecutive stages. First, the specific sequence (the string of bases) or *primary structure*, is transformed into a pattern of complementary base pairings called the *secondary structure*. Second the secondary structure distorts, to form a three dimensional spatial structure or *tertiary structure*.

It is hard to solve the structure prediction problem for RNA structures since the number of degrees of freedom of the RNA chain is very high (indeed it is much higher than in the protein case). There are several facts that support the consideration of the secondary structure of RNA as a coarse grained approach to the three dimensional spatial structure:

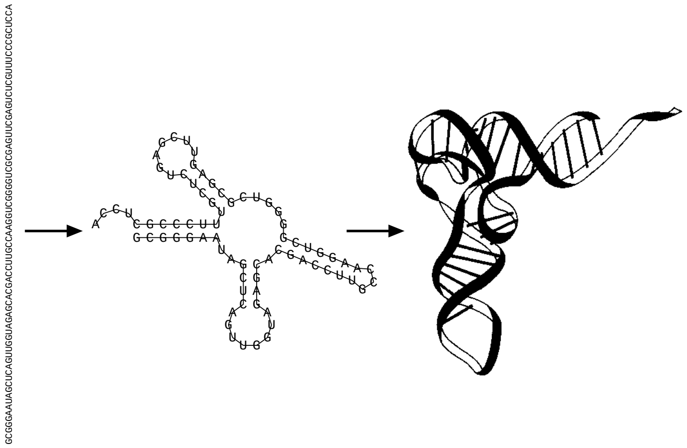

GCGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUUCCCGCUCCA

Figure 2: Folding of the phenylalanyl-transfer-RNA tRNA$^{\mathrm{phe}}$ into its spacial structure.

- The conventional base pairing and the base stacking cover the major part of the free energy of folding.

- The secondary structure provides a scaffold of distance constraints to guide the formation of the tertiary structure.
- In contrast to the protein case, the secondary structure of RNA is well defined and assigns all bases to secondary structure elements.
- RNA secondary structure is conserved in evolution and has been used successfully to interpret RNA function and reactivity.

The secondary structure of RNA is formed by aggregation of planar complexes, or *base pairs* of purine and pyrimidine bases. There are four naturally occurring bases: Adenine (A), Guanine (G), Cytosine (C) and Uracil (U). G and C, respectively A and U are complementary bases which can form strong hydrogen bonds, a weaker base pair is also possible between G and U, often referenced as “wobble” base pair.

The tertiary structure as shown in figure 3 is the three-dimensional configuration of the molecule. Tertiary interactions are hydrogen bonding or stacking interactions between structure elements.

2.2 RNA Secondary Structure

A secondary structure $\mathcal{S}$ is formally defined as the set of all base pairs (i, j) with $i < j$ such that for any two base pairs (i, j) and (k, l) with $i \leq k$ the two following conditions hold [59]:

1. $i = k$ if and only if $j = l$.
2. There are no knots or pseudo knots allowed. For any two base pairs (i, j) and (k, l) the condition $i < k < l < j$ or $k < i < j < l$ must be satisfied.

The first condition simply means that each nucleotide can take part in at most one base pair. Several examples of tertiary interactions breaking this condition are known, including base triplets, G-quartets and A-platforms.

The second condition guarantees that the secondary structure can be represented as a planar graph. The most abundant structural elements, which break this condition are pseudoknots. A pseudoknot is governed by *Watson-Crick* base pairing between a hairpin loop and a single-stranded stretch or between two single-stranded stretches. Consequently, a pseudoknot can be considered as either a secondary structural element or a tertiary interaction. While pseudoknots are important in some natural RNAs, they can be considered as part of the tertiary structure for our purposes. Not all secondary structures can be formed by a given biological sequence, since not all combinations of nucleotides form base pairs.

Let $\mathcal{A}$ be some finite alphabet of size κ, let Π be a symmetric Boolean $\kappa \times \kappa$-matrix and let $\Sigma = [\sigma_1 \ldots \sigma_n]$ be a string of length n over $\mathcal{A}$. A secondary structure is *compatible* with the sequence Σ if $\Pi_{\sigma_p,\sigma_q} = 1$ for all

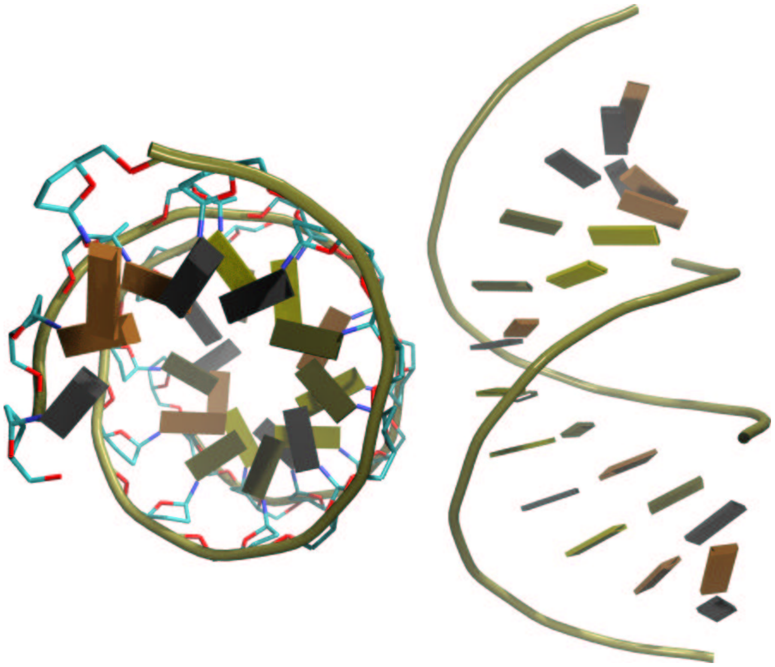


Figure 3: Illustration of the molecular structure of a typical A-RNA. Views are parallel (l.h.s.) and perpendicular (r.h.s.) to the helix axis.

base pairs (s_p, s_q). Following [30, 59] the number of secondary structures $\mathcal{S}$ compatible with a specific string can be enumerated as follows: Denote by $S_{p,q}$ the number of structures compatible with the substring $[\sigma_p \dots \sigma_q]$. Then

$$S_{l,n+1} = S_{l,n} + \sum_{k=l}^{n-m} S_{l,k-1} S_{k+1,n} \Pi_{\sigma_k,\sigma_{n+1}} \tag{1}$$

A secondary structure compatible with a given sequence with maximal number of base pairs can be determined by a dynamic programming algorithm [47]. The restriction to knot-free structures is necessary for efficient computation.

Usually, only Watson-Crick (**AU** and **GC**) and **GU** pairs are allowed. The secondary structure indicates the position of base paired helices. These are linked by single-stranded regions that can form hairpins, internal bulges within helices, multi-branched loops or link helices. The complexity and design variability of such structures is stunning and revals those present in proteins.

Secondary structures can be represented as strings composed of the symbols `(`, `)`, and `.` representing nucleotides that are paired with a partner towards the 3' end, towards the 5' end, and that are unpaired, respectively. Pairs of matching parentheses therefore indicate base pairs. A short hairpin structure, consisting of 4-loop and a helix of length 3 will therefore be written as `(((....)))`, see [27]. Figure 4 shows the secondary structure of tRNA$^{\mathrm{phe}}$ and its corresponding bracket-dot-representation.

There are several other ways to represent RNA secondary structures: In the particularly easy *Circular representation* (figure 5), the bases of the sequence are placed equidistant to one another on a circle and for each base pair a chord is drawn between the two bonded bases. Since the structures are knot-free by definition, no two chords will intersect.

Another useful approach for the comparison of RNA secondary structures is called *mountain representation* where `'('`, `')'`, and `'.'` is identified with “up”, “down”, and “horizontal”, respectively. See Figure 6 for mountain representation.

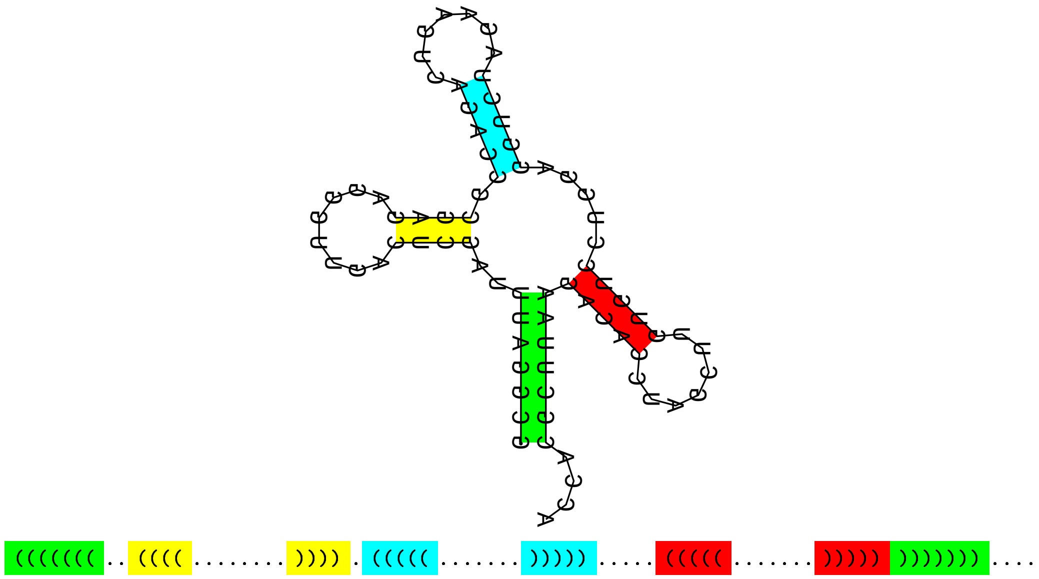


Figure 4: Secondary structure of tRNA$^{\mathrm{phe}}$ and the corresponding bracket-dot-notation. Same colors represent the same base-pairing regions.

- *Peaks* correspond to hairpins. The symmetric slopes represent the stems enclosing the unpaired bases in the hairpin loop, which appear as a plateau.
- *Plateaus* represent unpaired bases. When interrupting sloped regions they indicate bulges or interior loops, depending on whether they occur alone or paired with another plateau on the other side of the mountain at the same height respectively.
- *Valleys* indicate the unpaired regions between the branches of a multi-stem loop or, when their height is zero, they indicate unpaired regions separating the components of secondary structures.

The height of the mountain at sequence position k is simply the number of base pairs that enclose position k; *i.e.*, the number of all base pairs (i, j) for which $i < k$ and $j > k$.

Any secondary structures can be uniquely decomposed into loops as shown in figure 7 (note that a stacked base pair may be considered as a loop of size zero). A secondary structure graph is equivalent to an ordered rooted tree. An internal node (black) of the tree corresponds to a base pair (two nucleotides), a leaf node (white) corresponds to an unpaired nucleotide. Contiguous base pair stacks translate into "ropes" of internal nodes, and loops appear as bushes of leaves.

The energy of an RNA secondary structure is assumed to be the sum of the energy contributions of all loops. Energy parameters for the contribution of individual loops have been determined experimentally and depend on the loop type, size and partly its sequence.

The additive form of the energy model allows for an elegant solution of the minimum energy problem through dynamic programming, that is similar to sequence alignment. This similarity was first realized and exploited by Michael Waterman [58, 59].

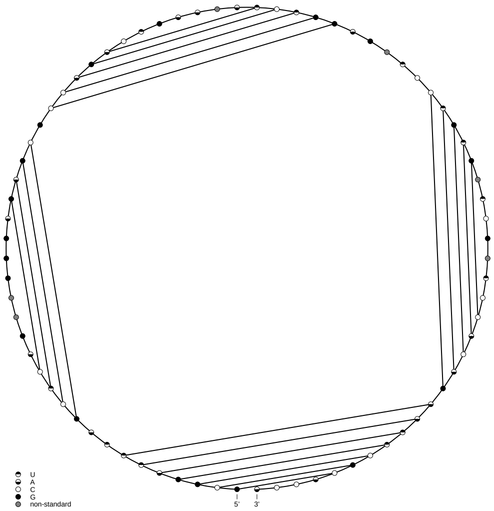


Figure 5: The secondary structure of tRNA$^{\mathrm{phe}}$ in *circular representation*

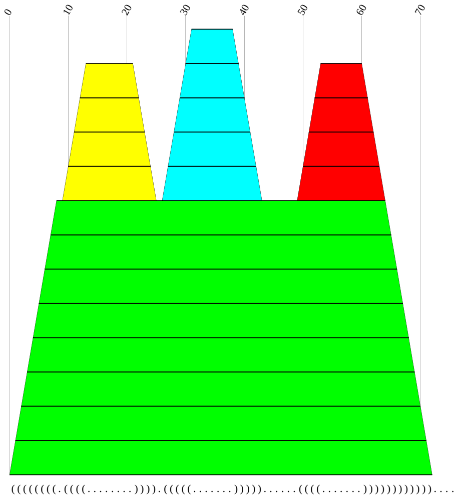


Figure 6: The secondary structure of tRNA$^{\mathrm{phe}}$ from Yeast (see Figure 2) in *mountain representation*. Colors are the same as in figure 4 and represent the same base-pairing regions

The first dynamic programming solution was proposed by Ruth Nussinov [46, 47] originally for the "maximum matching" problem of finding the structure with the maximum number of base pairs. Michael Zuker and Patrick Stiegler [64, 65] formulated the algorithm for the minimum energy problem using the now standard energy model. Since then several variations have been developed: Michael Zuker [63] devised a modified algorithm that can generate a subset of suboptimal structures within a prescribed increment of the minimum energy. The algorithm will find any structure $\mathcal{S}$ that is optimal in the sense that for every pair b in $\mathcal{S}$ there is no structure $\mathcal{S}_b$ that contains the pair b and has lower energy than $\mathcal{S}$. As shown by John McCaskill [42] the partition function over all secondary structures $Q = \sum_S \exp(-\Delta G(S)/kT)$ can be calculated by dynamic programming as well. In addition his algorithm can calculate the frequency with which each base pair occurs in the Boltzmann weighted ensemble of all possible structures, which can conveniently be represented in a so called "dot-plot". Figure 8 shows such a dot-plot of tRNA$^{\mathrm{phe}}$.

The equilibrium frequency p of a base pair (i, j) is represented by a square of area p in position i, j of a triangular matrix. The lower left triangular matrix shows the optimal fold , namely the ground state. In contrast the upper right triangular matrix displays the base pair frequencies within the structure ensemble at the thermodynamic equilibrium as obtained from the partition function. Note that in this example a large number of base pairs from suboptimal folds are visible. While the helix is very well defined, the loop region can can fold into various alternatives. This indicates, that the loop region of the ground state is flexible in a structural sense.

The memory and CPU requirements of these algorithms scale with sequence length n as $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$, respectively, making structure prediction feasible even for large RNAs of about 10000 nucleotides, such as the entire genomes of RNA viruses [29, 31]. A for academic use freely available implementation of these algorithms is the `Vienna RNA Package` [27, 28] (available from `http://www.tbi.univie.ac.at/ ˜ivo/RNA/ViennaRNA-1.4.tar.gz`). Energy parameters used there can be found in [41].

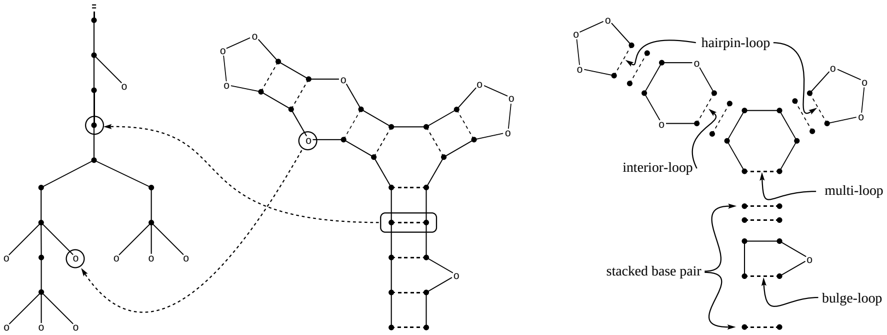


Figure 7: Various representations of RNA secondary structure: The tree representation of the secondary structure graph in the middle (l.h.s); Representation of an RNA secondary structure as a planar graph (middle); The loop decomposition of the secondary structure graph in the middle (r.h.s). The closing base pairs of the various loops (base pair, hairpin, bulge, interior, multiloop) are indicated by dotted lines (Note that a helix of length n decomposes in n-1 stacked base pairs).

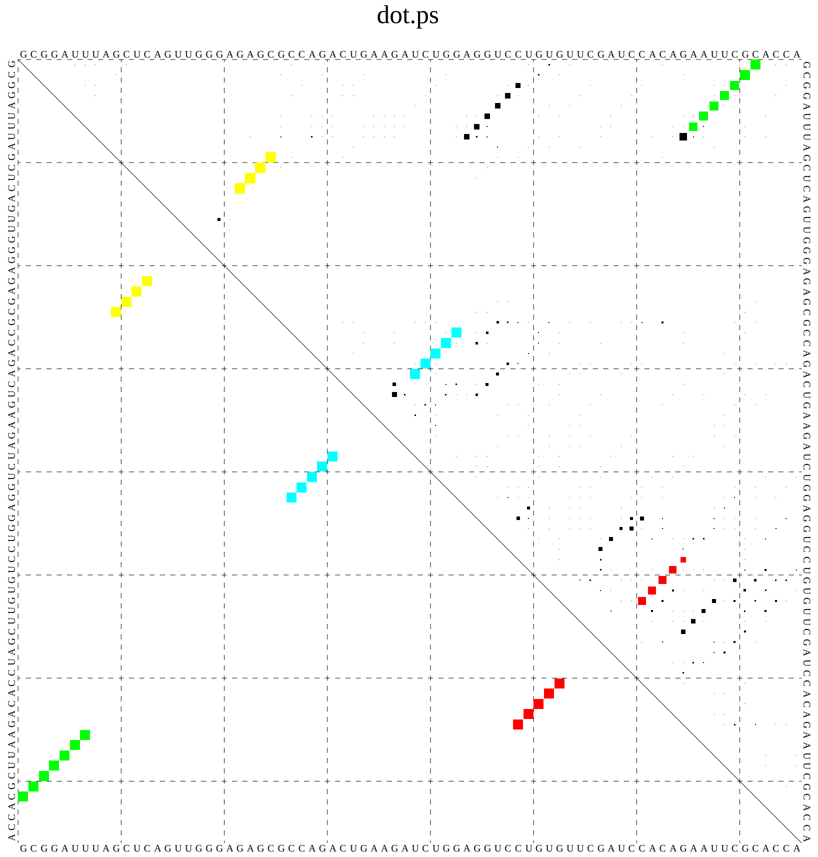


Figure 8: Dot-plot of tRNA$^{\mathrm{phe}}$; The equilibrium frequency p of a base pair (i, j) is represented by a square of area p in position i, j and j, i of the matrix. The lower left triangle shows only base pairs contained in the ground state, which occur with significant frequency. The upper right triangle displays the frequencies within the thermodynamic equilibrium. A large number of base pairs from suboptimal structures are visible. Again, same colors represent the same base-pairing regions as in figures 4 and 6

2.3 Conformation Space: The Thermodynamic View

The *conformation space* $\mathcal{C}$ of a given sequence is the total set of secondary structures $\mathcal{S}$ compatible with this sequence. As mentioned each secondary structure $\mathcal{S} \in \mathcal{C}$ itself is a list of base pairs (i, j) in a way, that any two base pairs from $\mathcal{S}$ do not cross each other, if $\mathcal{S}$ is represented as a graph in the plain. From the total recursion (equation 1) an asymptotic formula for the growth of the number of secondary structures with chain length n can be derived.

$$S_n \sim n^{-\frac{3}{2}} \cdot \alpha^n \tag{2}$$

Counting only those planar secondary structures that contain hairpin loops of size three or more (steric constraint), and that contain no isolated base pairs one finds $\alpha = 1.8488$ for the total number of secondary structures. The size of the conformation space increases exponentially with the chain length. The density of states $g(\varepsilon)$ is a convenient measure to get a survey of the conformation space $\mathcal{C}$ of a given sequence. It displays the energies of the individual structures $\mathcal{S}$, and their distribution with regard to the ground state. Furthermore $g(\varepsilon)$ is the basis for the equilibrium statistical mechanics of any system, because the average of any physical property $\mathcal{P}$, depending on the energy, is given by the Boltzmann-weighted sum,

$$\langle \mathcal{P} \rangle_{eq} = \frac{1}{Z} \cdot \sum_{\varepsilon} \mathcal{P}(\varepsilon) \cdot g(\varepsilon) \cdot e^{-\varepsilon/k_B T} \tag{3}$$

where k_B is the Boltzmann's constant, T is the absolute temperature and

$$Z \equiv \sum_{\varepsilon} g(\varepsilon) \cdot e^{-\varepsilon/k_B T} \tag{4}$$

is the partition function, giving a complete thermodynamic description of the system.

A variation of John McCaskill's algorithm can be used to compute the complete density of states [13] for a given sequence. In figure 9 the density of states is shown for yeast tRNA^{phe}. The conformation space of yeast tRNA^{phe}, a molecule of only 76 nucleotide length, has the astronomical size of $\sim 14.9 \cdot 10^{16}$ secondary structures (By comparison the human brain is built up of $\sim 1 \cdot 10^{10}$ neurons). The overall shape of the density of states for this example is Gaussian. This is not surprising since ε is composed of a large number of additive contributions. The overwhelming majority of the secondary structures however has positive energy. Hence only a small subset of all possible structures is physically important. These approximately 2 million structures have negative energy, the reference state being the open chain. The folding process of RNA molecules is believed to operate mostly on this small subset of $\mathcal{C}$. Unfortunately $g(\varepsilon)$ provides almost no information about the folding landscape, with respect to dynamics. If the kinetic progress in folding of a biopolymer is modeled, it is helpful to define a reaction coordinate. The reaction coordinate serves as measure, to gauge the "closeness to the native structure". A thermodynamic reaction coordinate defines closeness to the native state in terms of the *energy* of the conformation, whereas a kinetic reaction coordinate defines closeness to the native structure in terms of how quickly that conformation can transform to the native state. For instance the density of states defines "closeness" between two states of the energy landscape in terms of *energy*. In this sense all states which take energies similar to the ground state, seem to be close to the ground state.

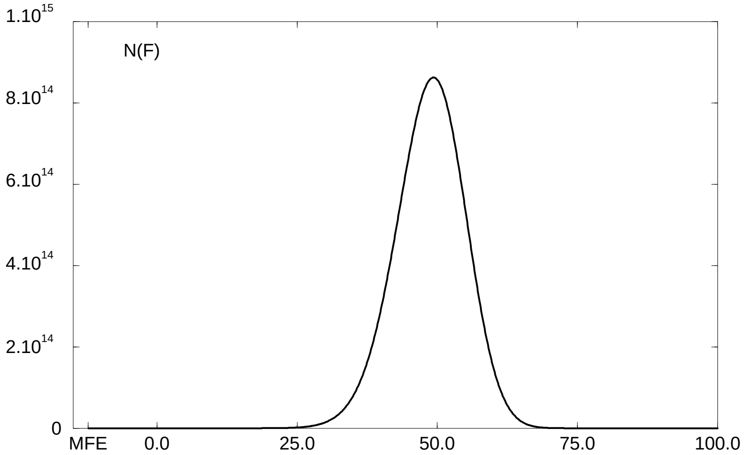


Figure 9: Density of states of the yeast tRNA$^{\text{phe}}$ with an energy resolution of 0.1 kcal/mol. Less than 2 million structures have negative energy, the reference state being the the open structure. For details see [12, 13].

No information is obtained whether the ground state and these "energetically close" states are structurally similar enough to allow a rapid interconversion. This information however is of utmost importance, since it elucidates the local features of the folding landscape, which have a feed back onto the folding dynamics. Figure 10 illustrates the problem. A thermodynamic reaction coordinate sees some deeply trapped conformation B as being "nearly native", because B has low energy, even though such conformations must overcome high-energy barriers to reach the native state. But a kinetic progress coordinate should describe, at least at some rudimentary level, the fraction of *time* that has elapsed, or that remains, for the folding, rather than the fraction of *energy* that remains. By using a thermodynamic reaction coordinate, B in figure 10 is closer to native N than A is. But by using a kinetic reaction coordinate, A is closer to N, since A has to climb a smaller energy barrier to reach N than B. For landscapes with kinetic traps, thermodynamic reaction coordinates do not characterize the kinetics well, because they completely neglect energy barriers.

Therefore a measurement called *move set*, which captures "structural vicinity" in a kinetic sense, needs to be developed before the relationship between the folding dynamics and the topology of the underlying energy landscape can be studied. The move set and its influence on the topology of the folding landscape will be discussed in further detail in section 3.2.

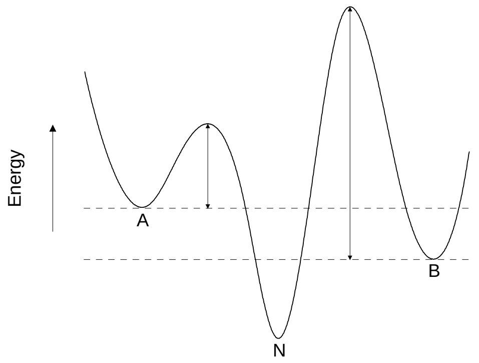


Figure 10: Thermodynamic *versus* kinetic reaction coordinate. State B is *energetically* closer to N (lower energy), but state A is kinetically closer to N (smaller barrier to cross). For didactic reasons a continuous reaction coordinate is used as abscissa. In the realm of RNA secondary structures energy and reaction coordinate are discrete.

3 Kinetic Folding

3.1 State of the Art

The present understanding of RNA folding is still largely based on classic studies of tRNA. In the 1970s the crystal structure of tRNA^{phe} [11, 50] became available. Temperature jump and NMR experiments were used to identify the conformations of intermediates on the path to the equilibrium fold of different tRNAs [2, 10, 14, 22, 38, 54]. More structural information and insight into RNA catalysis came from the first crystal structure of a hammerhead ribozyme [48] in 1994. A great impact on the understanding of RNA spatial structure came from high-resolution cristallography of one of the two structural domains of the catalytic core of a group I intron [7].

Recently, kinetic studies [61, 62] of a ribozyme derived from the Tetrahymenea group I intron, a considerably more complex molecule than tRNA or hammerhead ribozyme, introduced some previously unexplored features of RNA folding. As pointed out by Patrick Zarrinkar and James Williamson, the Tetrahymenea ribozyme folds by a hierarchical pathway with successively larger structures generally requiring longer time scales. Short range secondary structure appears to form rapidly to yield a state in which much of the secondary structure is present, but which is still very flexible and lacks stable tertiary contacts. The native structure is then formed from this “quasi fluid” state by the successive formation and stabilization of larger folding units, which generally correspond to identifiable structural subunits. These subunits seem to form in a hierarchical manner, where the presence of the fast forming elements is required for the formation of the slower folding subunits. The formation of specific long range contacts that allow the folding units to interact then occur late on the folding pathway. The sequential folding of domains in the ribozyme show striking parallels to the way how the α-subunit of the protein tryptophane synthetase achieves its fold.

Several groups developed kinetic folding algorithms for RNA secondary structure, mostly in an attempt to get better structure predictions than their thermodynamic counterparts. Only little effort has been put into the reconstruction of folding pathways or the consideration of pseudo-knots. The great majority of these algorithms are based on Monte-Carlo methods [43]. In general these algorithms start from some initial structure (e.g. the open chain) and progress, by incorporation of whole helices, through a series of nearly optimal structures to the most probable one at the end of the folding process.

The first attempts modeled the folding process as a strictly *sequential* process. Different criteria for choosing the next stem for incorporation, like choosing the stem with the maximal number of base pairs [32] or the stem with the largest equilibrium constant [40] have been tested. A disadvantage of the sequential methods is their inability to destroy already constructed stems, and hence simulations get easily stuck in local minima.

Next, the folding process was modeled as a *Markovian* random process [5, 44, 45, 55] to circumvent the problems of sequential methods. These algorithms differ mainly in the method how they reduce the state space to make the calculation of the transition probability matrix computationally feasible. Helix formation rates are approximated through models using parameters derived from experimental results, helix fusion rates are deduced from the formation rates by using a *Boltzmann* distribution hypothesis on the structure space.

Another fruitful approach was suggested by Christoph Flamm [16, 18] and his tool `kinfold`, which is capable of calculating trajectories describing the folding behavior of single secondary structures. See section 5.3 for a more detailed description of his work.

3.2 The Move Set

The conformation space $\mathcal{C}$, as has been illustrated in section 2.3, is a multi-dimensional space. Depending on the coarsegraining of the energy, conformation space can being highly degenerated. *A priori* it is not clear how to move in such a complex space, therefore a set of rules is needed to control the movement. Such a set of rules is called a *move set* (for an example see figure 11). It is basically a collection of operations, which, applied to an element of $\mathcal{C}$, transforms this element into another element of $\mathcal{C}$. Strictly spoken a *move set* is an order relation on $\mathcal{C}$, defining *adjacency* between the elements of $\mathcal{C}$. It fixes the possible conformational changes that can take place in a single step during the simulation of folding and thus defines the topology of the conformational space. The following properties are important for move sets:

1. Each move has an inverse counterpart. At thermodynamic equilibrium the quotient of forward and backward reaction rates must give the microscopic equilibrium constant (If there is no backward reaction, the law of microscopic reversibility is broken).

2. The outcome of an operation always leads to an element of the underlying state space (Any operation yielding an element outside the state space is illegal).

3. The move set has to be ergodic. In other words starting from an arbitrary point of the state space every other point must be reachable by a sequence of legal operations (If this property is not fulfilled, and only a subset of the state space is accessible to the system the expectation $\langle \mathcal{F} \rangle$ of any state function $\mathcal{F}(\mathcal{S})$ will be incorrect or at least biased).

4. Every move set defines a metric on the underlying state space.

Two more terms are of importance for the further discussion. A *trajectory* is defined as a sequence of consecutive states of the state space generated by a series of legal operations from some initial state. A *path* (or *folding path*) is defined as a cycle free trajectory, more concrete, each state occurs only once within the sequence of adjacent states. In other words any trajectory can be transformed into a path by eliminating the cycles.

The most elementary move set, on the level of RNA secondary structures consists of insertion and deletion of a single base pair (i, j). This move set will be designated as MS1 in the further discussion. It is always possible to construct a path between any two $S_i, S_j \in \mathcal{C}$ by using operations from MS1. To find such a path, remove from S_i all base pairs that do not occur in S_j, and insert afterwards into this intermediate structure S_k all base pairs from S_j that do not occur in S_i. (Note, that $S_k = S_i \cap S_j$ can be the empty set, which resembles the open chain, being as well an element of $\mathcal{C}$).

It is easy to see, that the path, constructed by the rule given above is also the path of minimal length connecting the tRNA structures S_i, S_j. Deleting base pairs from a legal structure always returns a legal structure. This means that the intermediate structure S_k is a legal structure as well. S_j is also a legal structure by definition. Hence inserting the missing base pairs into S_k to transform this structure into S_j in an arbitrary succession, must run through a cascade of legal structures. Because of the restriction to legal intermediate structures, any other combination of moves to transform a structure into another one must result in a longer path. Since every element of $\mathcal{C}$ can be connected to every other element of $\mathcal{C}$ by a path, it follows that MS1 is an ergodic move set on $\mathcal{C}$.

A dominant mechanism for helix formation is the highly cooperative *zipper mechanism* [49]'. Starting from a suitable nucleus which can still dissociate easily into its components, addition of new base pairs stacked to the nucleus leads to favorable, negative free energy contributions. From then on, growth of the helix is spontaneous and leads to stepwise construction of the helix just as a zipper is closed. MS1 is capable to describe this helix formation process properly.

An other important mechanism in the dynamics of RNA is believed to be "defect-diffusion". Since helix nuclei will be formed statistically along the RNA chain, intermediate formation of helices with incomplete base pairing is expected. Such intermediate mismatched helices can be annealed by a fast chain slide mechanism. For instance the loop base of a bulge loop present in a helix will be subjected to a rapid base pair formation and dissociation process. According to experimental data [49] defect-diffusion is some orders of magnitude faster then zippering. As a consequence of this rapid equilibration a bulge loop may move quite rapidly along the helix sequence. If a bulge loop forms at one end of the helix and disappears at the opposing end, the bulge loop diffusion results in a shift of the nucleotide strands by the nucleotide residues of the loop against each other (see figure 12). In the framework of

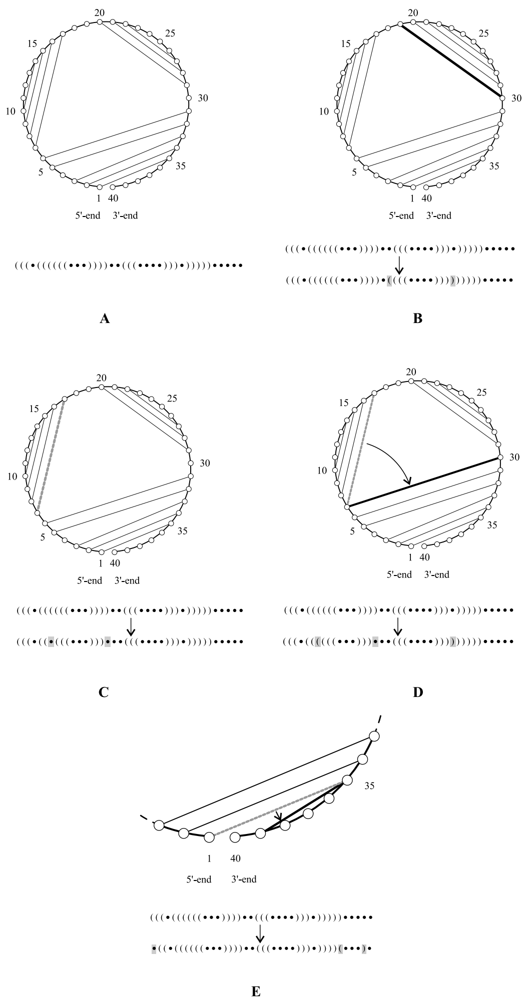


Figure 11: Elementary moves in the RNA folding algorithm. Secondary structures are shown in circle and parenthesis representation. The Structure A is changed by the formation B or the removal C of a base pair. A shift move of a base pair can occur either within the structure D or by flipping over the gap between the 3' and the 5' end E. The base pair after a move is shown in bold, the one being changed is shown by a gray dotted line.

MS1 the defect-diffusion is in most cases not a favorable process. It can only be achieved be a double move in contrast to zippering and therefore does not reflect the experimental results correctly.

To facilitate chain sliding MS1 must be extended by a further move called "shift". The shift converts an existing base pair (i, j) into a new base pair (i, k) or (l, j) in one step. The resulting move set will be referred to as MS2 in the following sections. Besides, defect diffusion, MS2 facilitates the metamorphosis of overlapping helices into each other. Especially if the two helices are located within a multi-loop the energetic profile of this process using the simple move set MS1 is unfavorable. Figure 13 illustrates this special "macro movement". Every ergodic move set that is extended by new moves naturally results in an ergodic move set again.

The algorithms cited in the section 3.1 generally operate on a list of all possible helices and consequently use move sets that destroy or form entire helices in a single move. The physical model of such a move set seems unrealistic because ad hoc assumptions about the rates of helix formation and disruption have to be made to cope with the introduction of large structural changes per time step. Furthermore the concept of "folding pathway" looses it's physical meaning, if structural changes are to large. For this reasons a more local move set like MS1 or MS2 is preferable if one aims at observing realistic folding trajectories.

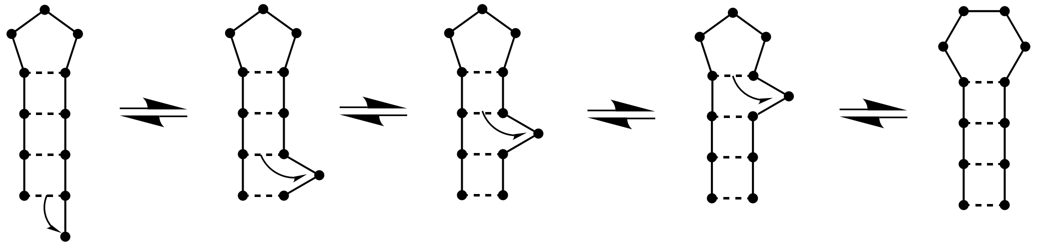


Figure 12: Defect diffusion: The bulge can easily migrate along the helix. For the left to right transformation the shift moves are indicated by arrows

3.3 Conformation Space: The Kinetic View

The energy landscape of a RNA molecule is a complex surface of the (free) energy versus the conformational degrees of freedom. In our case our allowed conformations are the secondary structures which are compatible with a particular sequence.

Like sequence space, the conformation space of secondary structures is a discrete space. Every secondary structure, a particular sequence can fold into, is represented by one *vertex* in the conformation space of the sequence. As has been illustrated in section 3.2 the move set induces a metric onto conformation space. If two conformations can be converted into each other, by applying a single move from the move set, the two conformations are neighbors of each other according to the move set. The vertices of the conformation space corresponding to neighboring conformations are connected by an *edge*. The object obtained in that manner is a complicated *graph*. In general, the graph representing conformation space is irregular, while the graph representing sequence space is always a regular one (generalized *hypercube*).

Figure 14 illustrates the conformation space for a short RNA molecule, which can form only 3 base pairs and 8 legal structures. The neighborhood of any vertex of the conformation space can easily be displayed in two dimensions. The entire conformation space, however, can be displayed only in

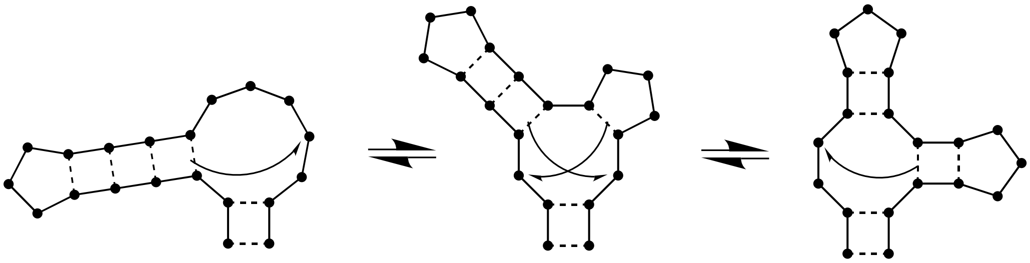


Figure 13: Inter-conversion of overlapping helices is facilitated by shift moves (indicated by arrows).

two dimensions and for very small sizes.

A *value landscape* is obtained by taking the graph of conformations as the support of a function that assigns a value to every conformation. In particular, a representation of the energy landscape of a RNA molecule is obtained by plotting the energy of a conformation according to the standard energy model over conformation space. Two factors characterize the shape of an energy landscape: (1) the density of states, and (2) a measure of structural similarity or kinetic "nearness" of one conformation to another. For the construction of the conformation space it is necessary to generate all possible secondary structures in a given energy range. The density of states gives only the number of conformations in a certain energy range, but not their explicit structures. Therefore suboptimal folding techniques are needed to provide this information.

Several approaches for the computation of suboptimal structures have been suggested. The development of these methods was motivated by several facts:

- Under physiological conditions RNA sequences may exist in alternative conformations whose energy difference is small.
- Aside from their possible biological significance, the density and accessibility of suboptimal conformations may determine how well-defined the ground state conformation actually is.
- The energy parameters on which the minimum free energy folding algorithms rely are inevitably inaccurate.

In contrast to many other suggested algorithms for the calculation of suboptimal RNA secondary structures, the program `RNAsubopt` [60], implemented by Stefan Wuchty generates **all** suboptimal folds of a sequence within a desired energy range. This is needed here for the construction of the conformation space.

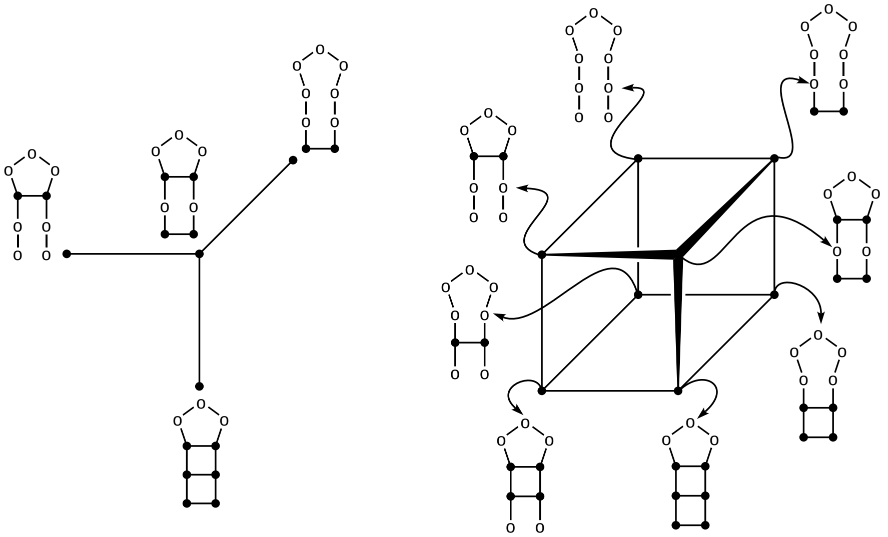


Figure 14: One move neighborhood of a vertex of the conformation space (l.h.s.) and its embedding in the graph representing the conformation space (r.h.s) for a small RNA molecule which can exhibit 3 base pairs.

4 Energy Barriers

Now, as we have both features at hand, namely all suboptimal structures within a given energy range and a metric (move set), a more detailed investigation of the energy landscape of RNA is possible. We will first consider some general features of the landscape and uncover topological details like *local optima* or *saddle points*. After that we will take a closer look at the main algorithm of the program `barriers`, which was originally written by Christoph Flamm [18] and modified for the requirements of this work.

4.1 Definitions

A structure is a *local minimum* if its energy is lower than the energy of **all** neighboring structures. A structure is called a *local maximum* if its energy is higher than the energies of **all** legal neighboring structures. Figure 15 illustrates which criteria the neighborhood of a point of the conformation space must fulfill to be a local optimum.

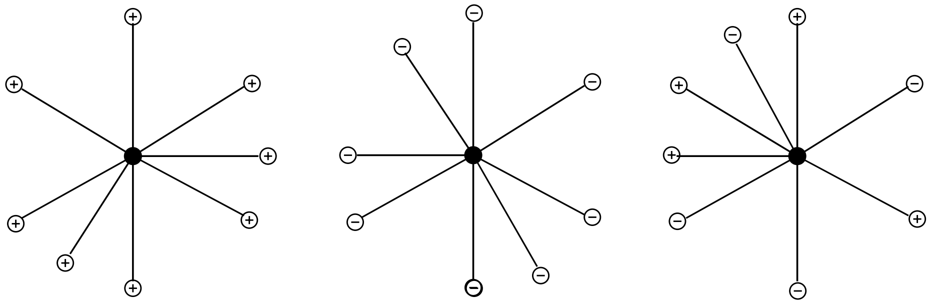


Figure 15: Illustration of the simple neighborhood of a local minimum (l.h.s), a local maximum (middle) and a saddle point (r.h.s). The signs within the circles denote neighbors with higher (+) or lower (–) energy compared to the structure in the center.

Saddle points are of special importance: A saddle point of the energy surface is a point that is neither a local minimum, nor a local maximum. However it is more convenient to use a more restrictive definition of a saddle point: A secondary structure S is a saddle point if there are at least two local minima that can be reached by downhill walks starting at S. Evidently, the saddle point with lowest energy that separates the basins of two local minima i and j is of particular importance. Those saddle points can be found by applying a flooding algorithm to the energy landscape (section 4.2). Figure 16 shows the tree representation of the energy landscape of a random RNA sequence with length $n = 42$. Leaves correspond to the valleys of the landscape, while saddle points are displayed by internal nodes. Saddle point energies can be read off easily. Figures 16 and 17 were calculated in such a way that any two local minima are joined by the saddle point with the lowest energy connecting the two local minima.

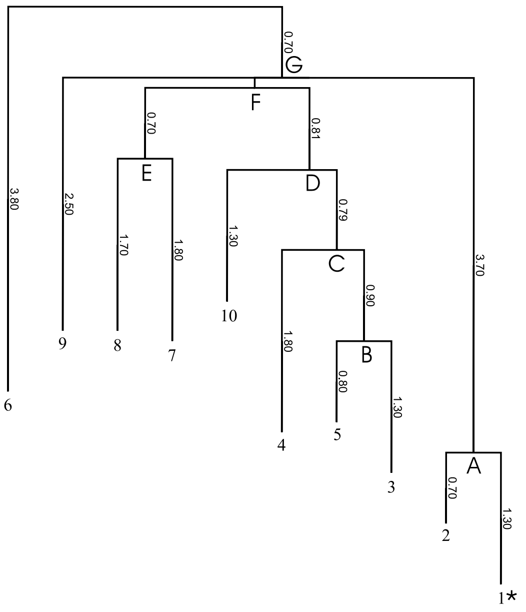


Figure 16: A typical barrier tree of the random RNA sequence `CCGCUCUACUGAGCGAAUCGACUAGAAAUCGCGAUACGAUCG` with length $n = 42$ as calculated with `barriers`. The leaves 1-10 denote the 10 lowest local minima of the energy landscape, the global minimum 1 on the right hand side of the plot is marked with an asterisk. Saddle points connecting different local minima are labeled with capital letters from A to G. The Energy barrier of 3 is $B(3) = E(B) - E(3)$, whereas the Energy Barrier to reach 10 from 3 is $E(3 \rightarrow 10) = B(3) + (E(C) - E(B)) + (E(D) - E(C)) = 1.30 + 0.90 + 0.79 = 2.99$ kcal/mol ($T = 310.15K$)

There is still a fundamental question concerning the energy landscape: What influences the ruggedness? In fact, the *definition* of neighborhood strongly influences this feature of the surface. In other words the choice of the move set critically forms the topology of the energy landscape. Figure 17 illustrates this strong metric dependency of the energy landscape. By changing the move set the connectivity of the local optima is changed dramatically. The barrier heights as well seem to lower in general if the “shift” move is used, which facilitates the annealing of defects. Since move set MS1 is subset of move set MS2, as has been explained in the section 3.2, all local optima of move set MS2 are also local optima under MS1, but not vice versa.

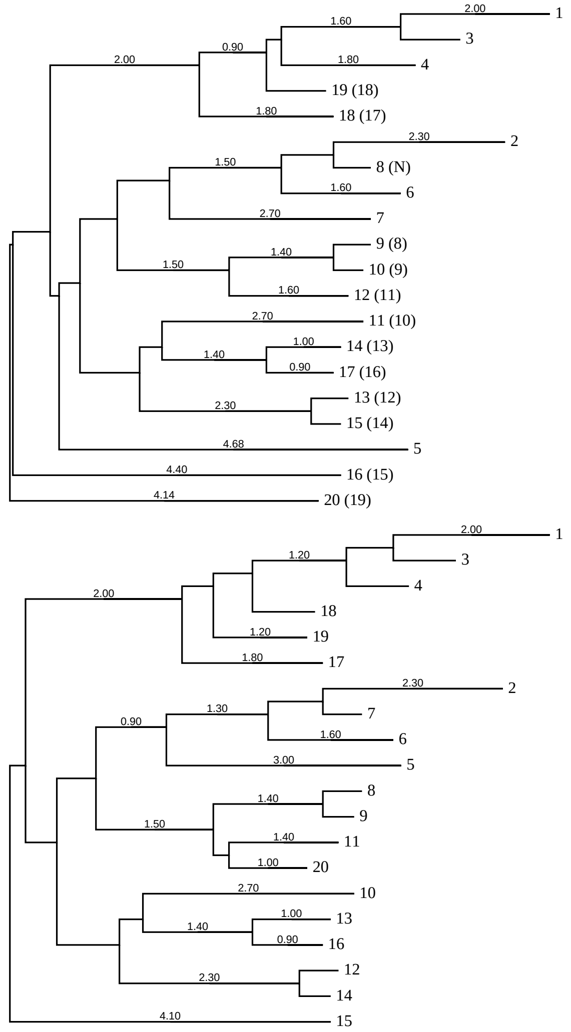


Figure 17: The tree representation of the 20 lowest local minima (leaves) and the saddle points (nodes) in the energy landscape of a typical RNA sequence. The lowest saddle points connecting two local minima are shown for move set MS1 (upper plot: insertion/deletion) and move set MS2 (lower plot: insertion/deletion/shift). The local minima are labeled in ascending order starting with the ground state. Equivalent minima are labeled identically in both trees. **Upper plot:** Cooresponding local minima from MS2 are given in brackets. Local minimum 8 occurs only within MS1. **Lower plot** Local minimum 20 just occurs here because is has not been seen yet in the upper plot, i.e. it has a higher energy up to which we algorithm couldn't get in the upper plot. Generally, The barrier heights and the connectivity is strongly influenced by the move set.

4.2 The Algorithm of `barriers`

The construction of the barrier tree starts from an energy-sorted list of all suboptimal structures in a certain energy interval which can be calculated with `RNAsubopt` (bundled with the `Vienna RNA package`). During the calculation two lists of valleys are needed, an active an an inactive one. The global minimum $x[1]$ belongs to the first active valley $V[1]$, while the list of inactive valleys is empty initially. Going through the energy-sorted list of secondary structures in increasing order there are three possibilities for each structure $x[k]$ at step k:

- $x[k]$ has one or more neighbors in exactly one of the active valleys $V[i]$. In this case $x[k]$ belongs to $V[i]$.
- $x[k]$ has no neighbors in either the active or the inactive valleys that have been found so far. Then $x[k]$ is a local minimum and determines a new active valley V_l.
- $x[k]$ has neighbors in more than one active valley, say$\{V_{i_1}, V_{i_2}, ..., V_{i_q}\}$. In that case $x[k]$ is a *saddle point* connecting those local minima. In the barrier tree $x[k]$ becomes an internal node and is added to the valley V_{i_1} with the lowest energy. All structures from the valleys $V_{i_2}, ..V_{i_q}$ are then copied to V_{i_1} while the status of the valleys $V_{i_2}, ..V_{i_q}$ is changed from active to inactive. Let us denote this instance with: Valleys $V_{i_2}, ..V_{i_q}$ are being *merged* with V_{i_1} (which will be called the 'father' from now on). Due to this one can say that from the point of view of a structure with an energy higher than the saddle point $x[k]$, $V_{i_1}, ..., V_{i_q}$ appear as a single valley that is subdivided only at lower energies. Consequently, after the highest saddle-point energy has been calculated, all valleys except for the global minimum V_{i_1} are in the inactive list.

It is important to realize the fact that as soon as a valley V_{i_k} has been merged with its father (and therefore copied to the inactive list), a 'deeper' local minimum V_{i_j} is is not accessible any longer as *the* original valley V_{i_k}.

Whenever a new structure that would belong to V_{i_k} is found, a recursion is started that looks for the father of V_{i_k}.

The flooding algorithm can can be visualized with the following 'gedanken experiment' (figure 19): Imagine a landscape with only two deep valleys A and B where A is energetically lower than B. Those two local minima are separated by the local optimum X, which is a saddle point. Water rises from bottom to top. In the first step (a), only the deeper valley A will be slightly filled with water. For our algorithm this means that all structures that are either below or exactly at the water surface belong to the local minimum A (all other structures are not accessible by now as we go through an energetically sorted list of secondary structures in ascending order). As the water still rises we encounter a different situation in step 2 (b). Not only A is filled with water, but also the deepest regions of B. From now on there are more possibilities for the secondary structures to belong to: Depending on which valley is the nearest (from the point of view of the conformation space), i.e. which local minimum contains structures that are legal neighbors of the actual one, a structure can either be added to A or to B.

Imagine the water rises further. The higher the water surface gets, the more structures are being seen. This means that with every increment (concerning the rise of the water) there are more possibilities for a structure which has not been seen so far to have neighbors in one or more of the valleys. Step 3 (c) shows this situation: The saddle point X has been found and there exists a structure which has neighbors in A *and* in B. In other words we can say the two lakes coincide. This is of special importance for the algorithm. As soon as X has been proved to be a saddle point, B is merged with its 'father' A and all structures from B can now be accessed as if they would be legal structures belonging to A. However, the algorithm does not stop here. As illustrated in step 4 (d), the water rises on and only valley A is still accessible. The end of the algorithm is reached as soon as either (a) all secondary structures have been processed or (b) a predefined amount of local minima has been found.

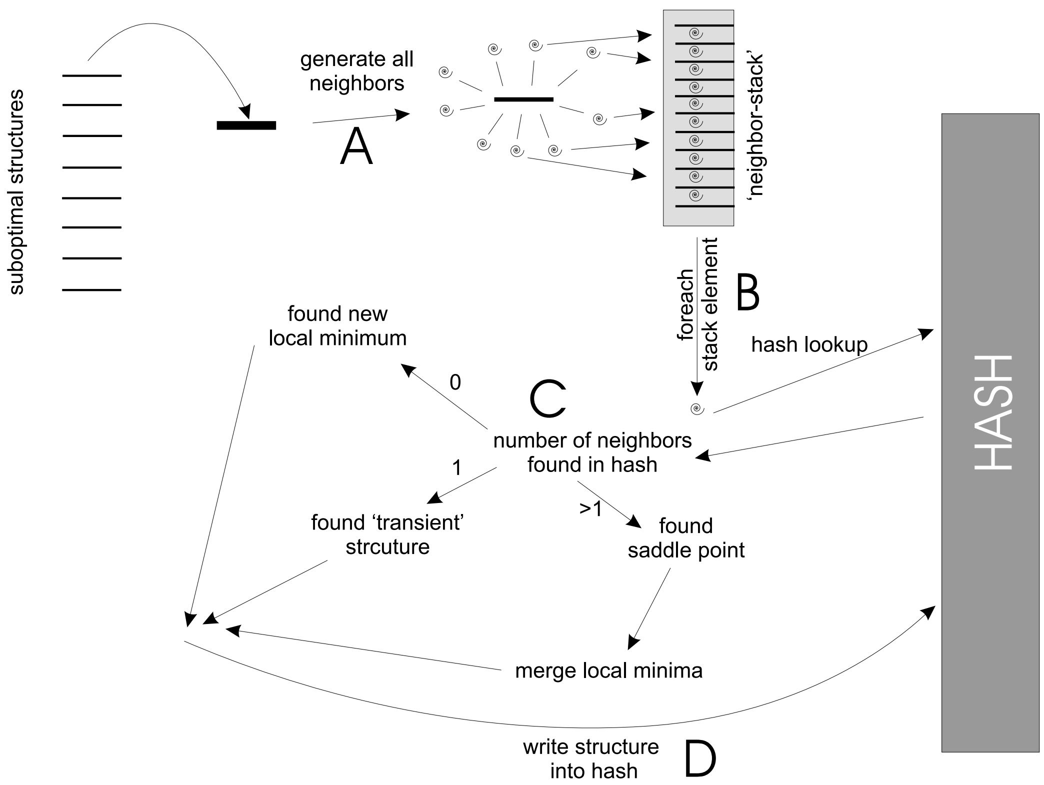
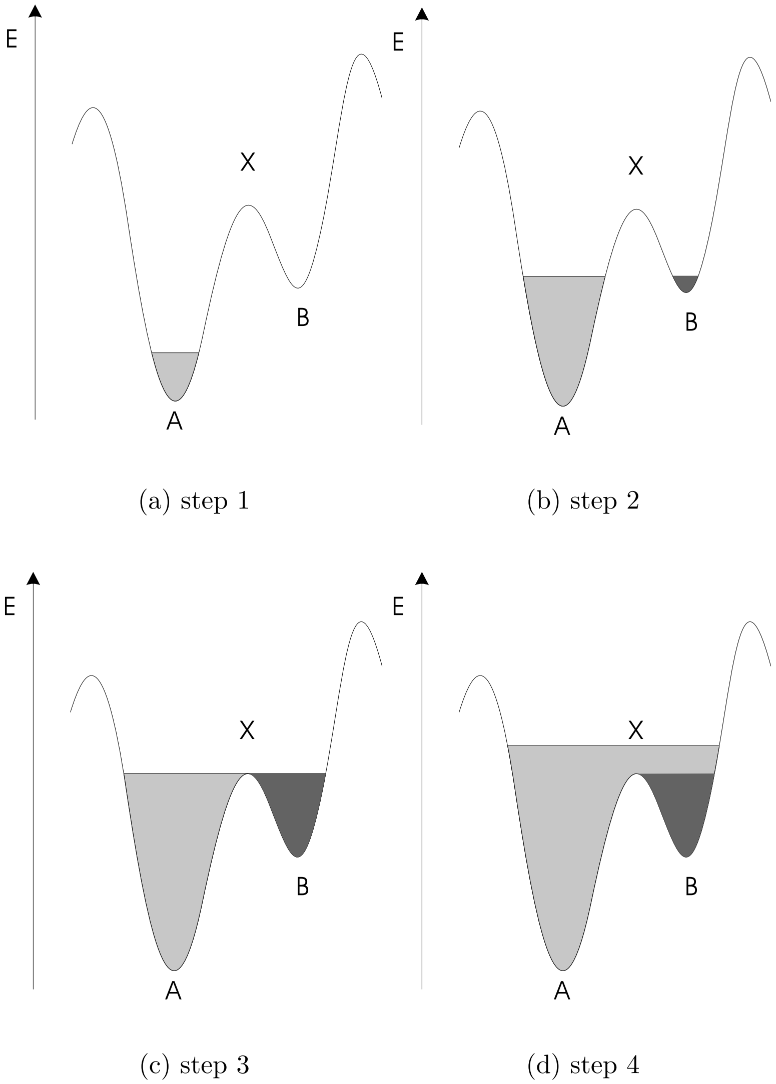


Figure 18: Schematic representation of the internal structure of `barriers`. The algorithm starts with an energy-sorted list of suboptimal secondary structures. They are processed in ascending order. First, all legal neighbors (in terms of the chosen move set) of the actual structure are generated and stored as a stack (**A**). Note that coils in the figure above correspond to neighboring structures. In the second step, each element of the 'neighbor stack' is processed (**B**). A routine searches a hash if the actual (neighbor) structure has yet been seen in a previous step of the computation. If this is true, the structure is remembered. When all elements of the stack have been processed, (**C**), there are three possibilities for the actual structure (whose neighbors were generated in **A**): If there was *no* adjacent structure resulting from the hash-lookup-procedure, the actual structure is a new local minimum and hence is added to the hash (**D**). If a neighboring basin has been found, the actual structure is assumed to be 'transient', which means it belongs to a certain basin of attraction. It is then added to the hash as well (**D**). The third possibility is that there have been found *two or more* local minima containing legal neighbors of the actual structure. If this is the case, then a saddle point connecting those local minima has been found. The energetically higher minimum is merged with the lower one (see text for details) and the saddle point structure is added to the hash again (**D**). At this point, the next structure is processed and the computation restarts with (**A**). This is repeated until no suboptimal structures are left.



(a) step 1 (b) step 2

(c) step 3 (d) step 4

Figure 19: 'Gedanken experiment' for the flodding algorithm where water rises in a landscape. For details see text.

The outcome of this procedure is the following information: There exist two local minima A and B which are connected by the saddle point X at a certain energy. All structures in A (as B is not accessible any longer) can be neighbors of other structures at higher energy. Evidently this is a very simplified 'gedanken experiment'. Real energy landscapes of RNA secondary structures do not only contain just two local minima, but several thousand. By applying this algorithm to RNA sequences typical barrier trees (see figure 16) can be obtained. The output produced by `barriers` is shown in figure 20. Evidently, the barrier tree can easily be reconstructed from this so called 'bar-file', as all barrier heights and connectivities among the local minima are listed there.

```
  ACUGAUCGUAGUCAC
1 ..((((....)))). ( -1.70)    0  13.60    0      142        0  -2.321       43  -2.010
2 ((((....)))... ( -1.50)    1   3.60    1 ...(........)..        11        8  -1.742       29  -1.667
3 ............... (  0.00)    2   1.80    1 ...(....)......         2        8  -0.032       60  -0.147
4 ((.....))...... (  0.10)    2   1.70    1 .(.....).......         1       10   0.100        6   0.099
5 .......((....)) (  1.70)    1   1.70    1 ........(....).         1       42   1.700        4   1.699
```

Figure 20: Output generated by `barriers` for the RNA sequence `ACUGAUCGUAGUCAC`. All information needed to reconstruct the barrier tree (figure 24 in section 6.1) is given. Generally, an output file (bar-file) like this consists of $m + 1$ lines where m is the number of local minima found in the system plus the first line containing the sequence. For details on how to read this bar-file, see the table below.

column	information
1	number of the local minimum
2	secondary structure of the local minimum in bracket-dot notation
3	energy of local minimum (in kcal/mol)
4	local minimum with which the actual local minimum was merged (the father)
5	energy barrier by which the actual local minimum and its father are separated (in kcal/mol)
6	multiplicity of the saddle point separating the actual local minimum and its father
7	secondary structure of the saddle in bracket-dot notation
8	basin size of the local minimum
9	basin size of the father at the time of merging
10	free energy of the actual local minimum
11	number of structures attracted by the actual gradient basin
12	free energy of the gradient basin

Note that the second line in the bar-file above looks slightly different from the subsequent ones: As the local minimum 1 represents the global minimum (and the mfe structure), there is no deeper valley it could merge with, so the father is 0 (which means there exists no father). Hence there is no saddle-point structure as well. The barrier height for the global minimum (13.00 kcal/mol in the example above) denotes that the calculation has been made up to an energy of $-1.70 + 13.00 = 11.3$ kcal/mol above the mfe structure. As the conformation space of the example sequence above does only consist of 142 secondary structures, and the *whole* conformation space has been considered for this calculation, no higher energy structures are available. When dealing with longer sequences, we are able to calculate suboptimal structures just up to a certain energy level and hence regulate up to which energy `barriers` should do its computation.

Additional information on the energy landscape can be gained during the construction of the barrier tree, i.e. we are interested not only in the local minima as calculated by the algorithm described above, but also in so called *gradient basins*. A gradient basin is the set of all initial points, from which a gradient walk (steepest descent) ends in the same local minimum. Evidently, this condition is just fulfilled, if gradient walks can be defined explicitly. Within our algorithm this is achieved within the loop over all neigboring secondary structures (step **B** in figure 18). A condition is evaluated, if a neighbor structure with more negative energy can be found. This is repeated for all neighboring structures. Finally, when all neighbors have been processed and the negighbor with lowest energy has been found, it is also stored in the hash. *Basin sizes* as well as *gradient basin sizes* can be calculated by adding up all all structures belonging to the same *basin* or *gradient basin* respectively. Additionally, we are able to calculate *partition functions* and *free energies* of basins and gradient basins. For details see section 5.4.

4.3 Degenerate Saddles

When talking about a single RNA molecule we always have to be aware of the fact that there is a very large number of secondary structures this molecule can fold into. This was shown in section 2.3. By applying $S_n \sim n^{-\frac{3}{2}} \cdot \alpha^n$ (equation 2) to an average sequence with $n = 120$ we get approximately 8.1×10^{28} structures. The number of possible secondary structures grows exponentially with the length of the sequence.

It is interesting to combine this fact with the concept of the barrier tree. A short consideration suggests that there should exist *degenerate* saddles, i.e. different secondary structures with the same energy, each of which has legal neighbor structures in the same local minima (see figure 21). We will also denote them *multiple* saddles here. This is also supported by the density of states concept, introduced in section 2.3: The overall shape of the density of states for many RNA molecules is gaussian and the overwhelming majority of structures is located within a small energy zone. This area should contain multiple saddles.

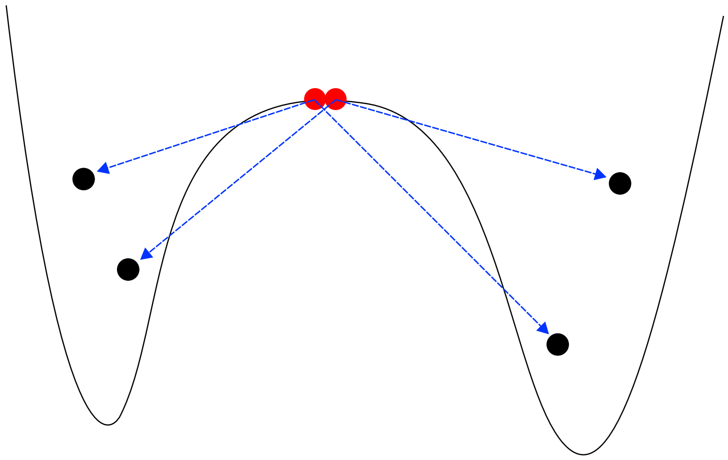


Figure 21: A schematic representation of degenerate saddle points with a multiplicity of 2 (red balls). Both of them have the same energy and have legal neighbor structures in each local minimum (black balls). Arrows point at the neighbors.

By strictly applying the above presentation of the algorithm of `barriers`, we encounter the following situation: The secondary structures are computed one-by-one in energy-ascending order. If the algorithm finds a saddle point, adjacent local minima are merged with their father and hence they are not any longer accessible as potential valleys that could be connected with other valleys via multiple saddles. We do not have any possibility to check if the saddle point just found is degenerate or not.

To circumvent this problem, we have modified the original version of `barriers` in a way that it also finds *multiple* saddle points if they exist in the conformation space $\mathcal{C}$. The major change in contrast to the original version is the issue that now at each step of the calculation an energy band (a series of structures within the range of some tenths kcal/mol) is being processed 'as-is' and if a saddle point connecting several local minima is found, they are not merged immediately but at the end of step, before the next energy band is being processed.

Conceptually, *degenerate* local optima should be of special importance when talking about transition rates between different local minima of the barrier tree (for details see section 5.3, we will only touch this here for the sake of completeness). They should influence a transition rate with a pre-exponential entropic factor Γ

$$k_{ij} = \Gamma e^{-\beta(E_S - E_i)} \qquad (5)$$

This is evident because the rate to reach one local minimum j from another local minimum i should be the bigger, the more possibilities (in our case the more saddle points) there are via which the transition has to take place. Unfortunately, figures 22 and 23 reveil that there are not too many multiple saddle points in the lower energy regions of the barrier trees (the regions where the 'interesting' folding kinetics take place). We therefore decided not to include degenerate saddles in our further investigations.

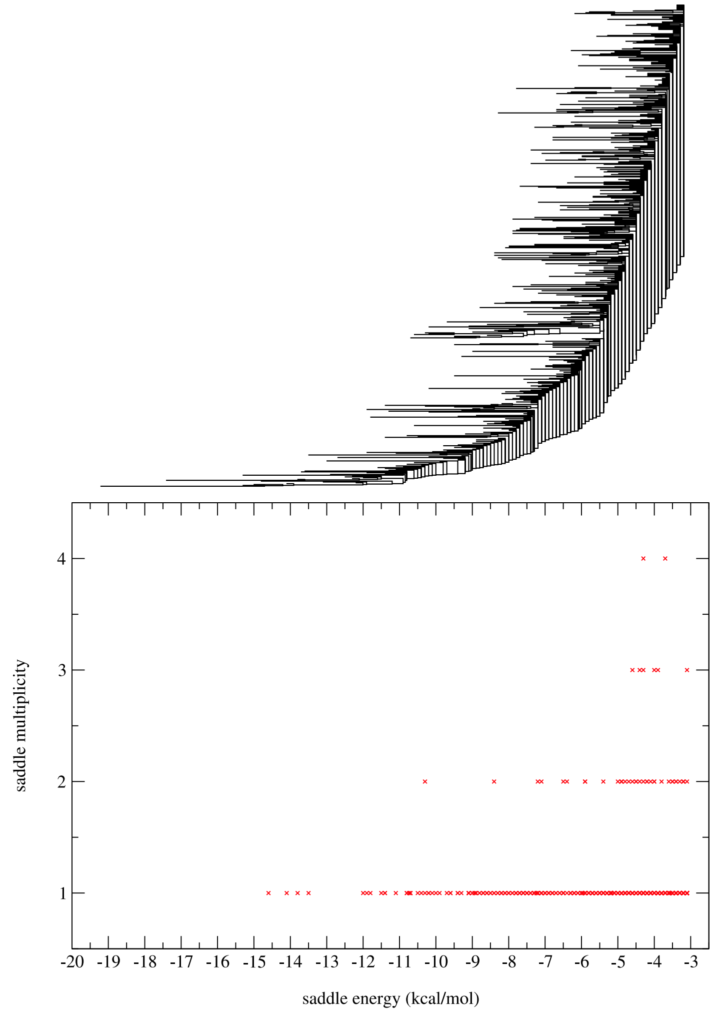


Figure 22: Saddle point energy versus saddle point multiplicity. Taken from a short artificial RNA chain `CCGGCGCGUCGCCGUAAGCGCGCUCGGGCAUAUAUAUUCAUAUGC` with a sequence length of $n = 45$. For this calculation, all suboptimal structures in the interval between -19.10 kcal/mol and -3.10 kcal/mol have been considered, structures were read in in energy intervals with $\Delta E = 0.01$ kcal/mol. Note that no multiple saddles occur within approximately 8.5 kcal/mol above the ground state.

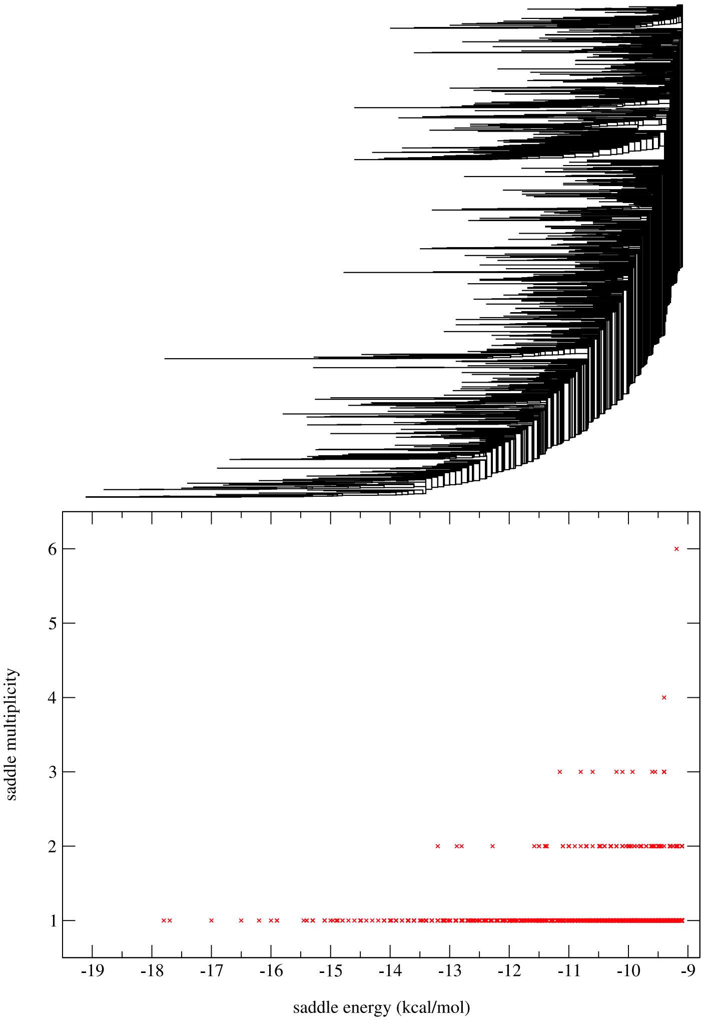


Figure 23: Saddle point energy versus saddle point multiplicity from a tRNA$^{\mathrm{phe}}$. All suboptimal structures in the interval between -19.10 kcal/mol (mfe) and -9.10 kcal/mol have been considered.The structures were read in energy intervals (energy bands) with $\Delta E = 0.01$ kcal/mol.

5 Markov Chains

This chapter deals with a particular class of stochastic models that form a cornerstone of this thesis. Stochastic models are widely used to describe phenomena that change randomly as time progresses. We focus on Markov chains, as simple and adequate models for many such phenomena.

5.1 Stochastic Processes

A *stochastic process* $\{X_t | t \in T\}$ is a family of random variables X_t defined over the same probability space and taking values in a set $\mathcal{S}$, usually referred to as the *state space* of a process. The parameter set T is often interpreted as time, and is sometimes called the *time range*. Each random variable X_t describes a snapshot random distribution on the state space $\mathcal{S}$ of the process at time t. The time range can be either discrete or continuous. This distinction separates two classes of stochastic processes, *discrete time stochastic processes* and *continuous time stochastic processes*. For simplicity, we assume that T is a subset of the nonnegative real numbers $\mathbb{R}^+$, in the discrete time case we identify T with the set of natural numbers $\mathbb{N}$ including 0.

A *Markov process* is a stochastic process that satisfies additional requirement. This *Markov property* requires that, for any given time instant (say t_n) the future behavior, for instance the value of $X_{t_{n+1}}$, is totally independent of its history, i.e. the values of $X_{t_{n-1}}$, $X_{t_{n-2}}$ and so on. It only depends on the state occupied at the current time instant t_n, given by the value of X_{t_n}.

In mathematical terms the Markov property requires that, for each sequence of time instances $t_{n+1} > t_n > t_{n-1} > t_{n-2} > ... > t_0$ (of arbitrary length n), we have that for each subset $\mathcal{A}$ of the state space $\mathcal{S}$,

$$\begin{aligned} & \text{Prob}\{X_{t_{n+1}} \in \mathcal{A} | X_{t_n} = P_n, X_{t_{n-1}} = P_{n-1}, ..., X_{t_0} = P_0\} \\ = \; & \text{Prob}\{X_{t_{n+1}} \in \mathcal{A} | X_{t_n} = P_n\} \qquad (6) \end{aligned}$$

Thus, the fact that the process was in state P_{n-1} at time t_{n-1}, in state P_{n-2}

at time t_{n-2}, and so on up to the fact that is was in state P_0 at time t_0 is completely irrelevant. The state X_{t_n} contains all relevant history information to determine the random distribution on $\mathcal{S}$ at time t_{n+1}. The above definition owes its name to A.A. Markov [39], who studied processes with this property at the beginning of the last century.

The above definition is tailored for continuous time Markov processes. In the discrete time case, the Markov property becomes somewhat simpler, since we do not have to bother about arbitrary sequences of time instances. Instead, we consider the (unique) sequence that contains all former time instances. Since we identified T with $\mathbb{N}$ we simply require for arbitrary $t \subseteq \mathbb{N}$,

$$\begin{aligned} & \mathrm{Prob}\{X_{t+1} \in \mathcal{A} | X_t = P_t, X_{t-1} = P_{t-1}, ..., X_0 = P_0\} \\ = \; & \mathrm{Prob}\{X_{t+1} \in \mathcal{A} | X_t = P_t\} \end{aligned} \tag{7}$$

It is worth to point out that the Markov property does not imply that the future behavior is independent of the current time instant t. If the value X_t does depend on t, the process is said to be *inhomogeneous*. However, throughout our discussion in the remainder of this thesis we shall assume that Markov processes are independent of the time instant of observation. In this case, a Markov process is said to be *homogeneous*; we gain the freedom to arbitrarily choose the origin of the time axis. In technical terms, homogeneity requires that we have (for $t' \geq t$ and $\mathcal{A} \subseteq \mathcal{S}$),

$$\mathrm{Prob}\{X_{t'} \in \mathcal{A} | X_t = P\} = \mathrm{Prob}\{X_{t'-t} \in \mathcal{A} | X_0 = P\} \tag{8}$$

The last simplification that we will impose concerns the state space $\mathcal{S}$ of a homogeneous (discrete or continuous time) Markov processes. Similar to the time range, the state space can be either discrete or continuous. We will only consider discrete state spaces here. This class of Markov processes is widely known as *Markov chains*.

By now, we have made three major restrictions for our requirements in contrast to the very general model of stochastic processes

- the Markov Property

- homogeneity and
- discrete state spaces.

Although the resulting class of homogeneous discrete and continuous time Markov Chains is one of the simplest classes of stochastic processes at all, it is adequate for our needs. Generally, Markov Chains are used to model a large variety of real world applications. The enormous amount of literature that exists on this subject testifies this, for example see [1]. Let us now consider a more exhaustive look at continuous time Markov Chains.

5.2 Continuous Time Markov Chains

A continuous time Markov chain X_t is a Markov process with discrete state space but continuous time range. We reformulate the Markov property (equation 6), with $t_n + \Delta t > t_n > t_{n-1} > t_{n-2} > ... > t_0$:

$$\begin{aligned} & \mathrm{Prob}\{X_{t_n+\Delta t} = P' | X_{t_n} = P, X_{t_{n-1}} = P_{t_{n-1}}, ..., X_{t_0} = P_{t_0}\} \\ = \; & \mathrm{Prob}\{X_{t_n+\Delta t} = P' | X_{t_n} = P\} \\ = \; & \mathrm{Prob}\{X_{\Delta t} = P' | X_0 = P\} \end{aligned} \tag{9}$$

If we substitute P with i and P' with j then the last expression can be rewritten as

$$\mathrm{Prob}\{X_{\Delta t} = j | X_0 = i\} = p_{ij} \tag{10}$$

This expression denotes the probability to reach state j from state i within the time step Δt. It is important to note that this probability (due to time homogeneity (equation 8) is independent of the actual time instant t_n (or t' or 0) of observation. Nevertheless there is a linear dependence on the length of the interval Δt. More precisely, for every pair of states i and j, there is a parameter k such that (for small Δt)

$$\mathrm{Prob}\{X_{\Delta t} = j | X_0 = i\} = k\Delta t + o(\Delta t) \tag{11}$$

where the sum over the probabilities to pass through intermediate states between i and j is given by $o(\Delta t)$. For us, the more important factor is k, which denotes a *transition rate* between i and j, a small nonnegative real number that scales how the transition probability increases with time. We shall assume here that i and j are different states. If, for any reason, i and j coincide, the probability to remain in state i during the time interval Δt (hence $\text{Prob}\{X_{\Delta t} = i | X_0 = i\}$) decreases with time, starting from 1 if $\Delta t = 0$. The corresponding transition rate is thus a negative real value which is implicitly determined by the increasing probability to leave state i.

In contrast to transition probabilities, transition rates do not depend on the length of time intervals. More generally, the probabilistic behavior of a continuous time Markov chain is completely described by the initial state (or distribution) and the transition rates between distinct states. We are now able to determine a continuous time Markov chain by means of a specific transition relation $i \overset{k}{\rightarrowtail} j$, defined over a certain state space $\mathcal{S}$ with an initial state i. They are called *Markovian chains*.

Definition 1 *A Markovian transition system is a tuple* $(\mathcal{S}, \rightarrowtail)$*, where*

- $\mathcal{S}$ *is a nonempty set of states, and*
- $\rightarrowtail$ *is a Markovian transition relation*

A Markovian chain is a triple $(\mathcal{S}, \rightarrowtail, i)$*, where*

- $(\mathcal{S}, \rightarrowtail)$ *is a Markovian transition system, and*
- $i \in \mathcal{S}$ *is the initial state*

Definition 2

A Markov chain (or Markovian chain) is said to be irreducible if it is possible to reach every state from every other state (not necessarily in one step).

A fundamental fact is that there exists a unique *stationary distribution* $\pi = (\pi_i : i \in \mathcal{S})$, i.e. a unique probability distribution satisfying the *balance equations*

$$\pi_j = \sum_i \pi_i p_{ij} \tag{12}$$

for all j. The point of stationarity is that, if the initial distribution X_0 of the chain is random with the stationary distribution π, then the position X_t at any subsequent non-random time t has the same distribution π, and the process $(X_t, t \geq 0)$ is then called the *stationary chain*. As a result of this stationarity, we can formulate an essential convergence condition: For any initial distribution

$$\text{Prob}\{X_t = j\} \rightarrow \pi_j \text{ as } t \rightarrow \infty \text{ or all } j \tag{13}$$

5.3 The Model

In the last sections two things were discussed which are crucial for the understanding of the remainder of this thesis. First, the move set and its influence on the topology of the energy landscape and second the algorithm of `barriers` which enables us to find valleys and local optima and hence allows for an efficient computational investigation of this landscape. Barrier trees were used to get an impression on local minima and saddle points.

Barrier trees have been considered recently for various models of disordered systems, including spin glasses and combinatorial optimization problems [3, 4, 15, 19, 37].

A very interesting approach for the understanding of kinetic folding of RNA was suggested by Christoph Flamm [16, 18]. He introduced the tool `kinfold` which is capable of calculating trajectories for the investigation of time evolution of RNA secondary structures. More generally, `kinfold` is capable of simulating the whole kinetic folding process of RNA molecules using the following ansatz:

Let I be a sequence which specifies a set of structures with which it is compatible,

$$\mathcal{S}(I) = \{S_0, S_1, ..., S_m\} \cup \{0\} \tag{14}$$

where S_0 is the minimum free energy (mfe) conformation, $S_1..S_m$ are energetically ordered suboptimal conformations and 0 is the denatured, open chain conformation. The set $\mathcal{S}(I)$ and the move set introduced in section 3.2 form the conformation space as mentioned earlier in section 2.3. A trajectory $\mathcal{T}(\mathcal{I})$ (as computed by `kinfold`) is a time-ordered series of secondary structures in $\mathcal{S}(\mathcal{I})$. Because the conformation space of secondary structures are always finite, every trajectory will reach S_0 after sufficiently long time. The *folding time* τ (associated with a trajectory) is defined as the first passage time, that is, the time elapsed until S_0 is encountered first. Due to the fact that τ may well be too long for a computer simulation, one can distinguish between trajectories that actually attain the ground state within the limits of a simulation from those that are trapped in a thermodynamically suboptimal conformation.

Translated into the language of chemical kinetics, the system is the RNA chain and a state of the system is a certain conformation of the RNA chain. Given the move set, RNA folding can then be modeled as a *Markov process* in conformation space as introduced in section 5.2. More precisely, if X_n denotes the state of the system at time n, the probability p_{ij} to find the system in state j after time Δt starting in state i is given by equation 10

$$\text{Prob}\{X_{\Delta t} = j | X_0 = i\} = p_{ij} = k_{ij}\Delta t + o(\Delta t)^2 \tag{15}$$

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

$$\frac{dP_t(i)}{dt} = \sum_{j \neq i} [P_t(j)k_{ji} - P_t(i)k_{ij}]. \tag{16}$$

Within this stochastic formulation, k_{ij} is the probability that a transition from a distinct state i to another distinct state j occurs within the infinitesimal time interval dt. For the soultion of the last equation (in matrical form), it is necessary to formulate a square intensity matrix (transition matrix) $\mathbf{U} = (u_{ij})_{i,j}$ which contains the transition rates between different states of the system

$$u_{ij} = \begin{cases} k_{ji} & \text{if } i \neq j, \\ -\sum_{l \neq i} k_{li} & \text{if } i = j \end{cases} \tag{17}$$

Equation 16 can be rewritten in matrical form:

$$\frac{d}{dt}P_t = \mathbf{U}P_t \tag{18}$$

We are interested in calculating the temporal distribution vector P_t, which can be calculated from the explicit solution of 18

$$P_t = e^{t\mathbf{U}}P_0 \tag{19}$$

where P_0 is the initial distribution vector.

In principle, equation 18 can be integrated numerically. Tacker et al. [56] used this technique to assess the feasibility of particular folding pathways of melting and refolding of $\mathrm{tRNA}^{\mathrm{phe}}$. Breton et al. [5] proposed a rigorous model of a sequential RNA folding process during transcription using this ansatz.

Our main problem is the calculation of population probabilities for local minima of the barrier tree. The structure probability distribution for the allowed local minima of the barrier tree can be calculated recursively from equation 19. Unfortunately, $\mathbf{U}$ is a matrix of dimension n where n is the number of local minima treated in the current simulation. As it is very difficult and inefficient to evaluate an expression like $e^{\mathbf{U}}$, similar to the right side of equation 19, the calculations are performed in the eigen space of the system as pointed out in the following.

Efficient diagonalization algorithms only exist for symmetric matrices. Due to the fact that $\mathbf{U}$ is not a symmetric matrix, it is necessary to symmetrize it. Generally, this can be achieved for self-adjoint matrices that satisfy detailed balance, such as Markov transition matrices (equations 23 and 24). A symmetric matrix $\mathbf{S}$ is obtained, which has the same eigenvectors and eigenvalues as $\mathbf{U}$.

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{+} \tag{20}$$

$$\mathbf{\Lambda} = \mathbf{V}^{+}\mathbf{S}\mathbf{V} \tag{21}$$

$\mathbf{V}$ is the matrix with the eigenvectors of $\mathbf{S}$ in column-order and $\mathbf{\Lambda}$ is a matrix with the eigenvalues of $\mathbf{S}$ in the diagonal. In matrical form, equation 21 can be written as

$$\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \cdot S \cdot \begin{pmatrix} v_1 & \dots & v_n \end{pmatrix}$$

Analogous we can write

$$e^{t\mathbf{S}} = \mathbf{V}e^{t\mathbf{\Lambda}}\mathbf{V}^{+} \tag{22}$$

S and **U** are associated with each other in the following way:

$$\begin{aligned} \mathbf{S} &= \pi^{1/2}\mathbf{U}\pi^{-1/2} && (23) \\ \mathbf{U} &= \pi^{-1/2}\mathbf{S}\pi^{1/2} && (24) \end{aligned}$$

where π is the equilibrium distribution as introduced before, i.e. it is the eigenvector associated with the biggest eigenvalue of **U**. The next step is to calculate $e^{t\mathbf{U}}$. Substitution of **S** in equation 24 with expression 20 gives

$$\mathbf{U} = \pi^{-1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{+}\pi^{1/2} \tag{25}$$

and analogously (note that $e^{t\mathbf{\Lambda}}$ is a diagonal matrix)

$$e^{t\mathbf{U}} = \pi^{-1/2}\mathbf{V}e^{t\mathbf{\Lambda}}\mathbf{V}^{+}\pi^{1/2} \tag{26}$$

We insert equation 26 in the right side of equation 19

$$\begin{aligned} e^{t\mathbf{U}}P_0 &= \pi^{-1/2}\mathbf{V}e^{t\lambda}\mathbf{V}^{+}\pi^{1/2}P_0 && (27) \\ P_t &= \pi^{-1/2}\mathbf{V}e^{t\lambda}\mathbf{V}^{+}\pi^{1/2}P_0 && (28) \end{aligned}$$

With substitution of

$$\pi^{-1/2}\mathbf{V} = \mathbf{M} \text{ and } \mathbf{V}^{+}\pi^{1/2} = \mathbf{N}$$

the original expression 19 reduces to

$$P_t = \mathbf{M}\exp(t\mathbf{\Lambda})\mathbf{N}P_0 \tag{29}$$

which can be calculated with moderate effort.

What still needs to be established is a rule for the rate-constant k_{ij} between two secondary structures i and j of the conformation space. We decided to chose the standard *Metropolis rule* [43], which was originally designed for studying equilibrium properties of matter and has also been applied successfully to kinetic problems like protein folding [57] on the one hand side, as well as the symmetric *Kawasaki rule* [36] on the other hand side. Both will be described in the following.

Let G_i be the free energy of the secondary structure i from which an allowed move to structure j with free energy G_j is made. Then, the transition probability k_{ij} as given by the Metropolis rule is:

$$k_{ij} = \begin{cases} e^{-\frac{\Delta G}{kT}} & \text{if } G_j > G_i, \\ 1 & \text{if } G_j \leq G_i, \end{cases} \tag{30}$$

where $\Delta G = G_j - G_i$.

The gradient of an energy landscape is an important determinant of the speed of moving uphill or downhill. The Metropolis rule only recognizes the uphill gradient. For uphill steps, by using the Boltzmann coefficient, sampling gets rarer as $\Delta G > 0$ increases. In contrast, all downhill steps ($\Delta G \leq 0$) are accepted with the same probability. This corresponds to the physical assumption that the spatial range of “favorable” contact interaction is literally zero, so residues along the chain would not “feel” any attraction to form a favorable contact. Since in Metropolis sampling the rates of forming a favorable contact does not increase with the contact’s favorability an intrinsic upper limit to downhill folding rates is set, which can be understood as a “diffusion limit” of the model.

A symmetric rule, which takes the gradient into account for both, uphill and downhill steps, is preferable, in order to avoid an intrinsic diffusion limit. Such a rule was first introduced by Kyozi Kawasaki [36] for studying time-dependent Ising models.

Due to Kyozi Kawasaki the symmetric rule evaluating the transition between the two states i and j connected by the reaction channel α is formulated as:

$$k_{ij} := e^{-\frac{\Delta G}{2kT}} \tag{31}$$

Note that the free energy difference ΔG between the two states i and j must be divided by $2kT$ to get the detailed balance right. The Kawasaki dynamics approaches the Boltzmann distribution at equilibrium because it satisfies microscopic reversibility [25]. For a detailed discussion of other possibilities to formulate the transition probabilities p_{ij}, see [9, 26]. As long as the law of detailed balance is satisfied by the rule, evaluating the transition probabilities, and the move set does not introduce too large conformational changes, the choice of a particular rule for the transition probabilities has only a small effect on the dynamics of the system, because then a state i quickly equilibrates with it's neighboring states.

5.4 The Barrier Tree Approximation

In the last section, the general model for the kinetic folding of RNA was introduced. In a previous section we learned that the conformation space grows exponentially with the chain length of the RNA molecule. Due to the fact that the algorithm of `kinfold` makes use of a stochastic model, very many trajectories have to be calculated to get a representative impression on the real folding behavior of the molecule. Furthermore one has to bear in mind that *all* secondary structures within a certain energy interval must be considered within such a simulation. These are the reasons which make realistic `kinfold`-simulations for longer, biologically more relevant sequences very intensive in terms of time and computer resources and this even leads so far that it is not possible to simulate the kinetic folding of RNA sequences with $n > 500$. As a matter of fact we have to replace this stochastic model with a deterministic one: It is necessary to reduce the conformation space.

Which states should be considered within the new, restricted conformation space? A short investigation of possible alternatives leads us back to the concept of barrier trees. As mentioned in section 4.1, a barrier tree represents the energy landscape of a RNA molecule, i.e. it shows local minima and saddle points. Why shouldn't we make use of exactly these structures and 'map' the original (very large) conformation space onto the barrier tree? We constitute that the energy landscape is represented 'as-is' by the barrier tree and its local minima and saddle points and there are no additional states (structures) that the system can attain. More formally, we can say that the state space $\mathcal{S}$ is partitioned into *macro states* (subsets of $\mathcal{S}$). In our case, such a macro state is characterized by a local minimum of the barrier tree.

With this ansatz, it is interesting to find out about the population probability of certain local minima on the barrier tree with respect to the fact that they are separated by more or less high energy barriers. In fact, we focus our investigations on the following questions:

- When starting the simulation at a specific local minimum of the tree (e.g. the denatured, open chain conformation), how long does it take for the system to reach an equilibrium state.
- To which extent are other local minima being populated on the way from the start structure to the minimum free energy structure.

To investigate these questions we make use of the concept of continuous time Markov chains introduced in section 5.2. As mentioned before, the conformation space is reduced in a way that we are only interested in local minima present in the barrier tree. In our special case, the *system* is the RNA chain and a *state* is the population probability of local minima of the energy landscape.

Let $\gamma(i)$ be the basin/gradient basin of local minimum i. Let further α and β be two different arbitrary macro states, π_α and π_β the equilibrium distribution for state α or β respectively. Then the partition function of $\gamma(i)$

is given by[1]

$$Z^*_{\gamma(i)} = \sum_{j \in \gamma(i)} \exp(-E_j/kT) \tag{32}$$

and the free energy is

$$G_{\gamma(i)} = -kT \ln Z^*_{\gamma(i)} \tag{33}$$

As we have to deal with a partition, it follows that

$$Z = \sum_i Z^*_{\gamma(i)} \tag{34}$$

From the last two equations we can derive

$$\sum_\alpha e^{-G_\alpha/kT} = Z \tag{35}$$

which illustrates that it is allowed to use macro states within our calculations. We are interested in formulating transition rates between different states of the system. To do this, it is necessary that the following requirements are fulfilled. First, we claim that

$$\pi_\alpha = \sum_{i \in \alpha} \pi_i \tag{36}$$

and second it is necessary that detailed balance (compare equation 12) must be fulfilled.

$$u_{\beta\alpha}\pi_\alpha = u_{\alpha\beta}\pi_\beta \text{ for all } \alpha, \beta \tag{37}$$

In fact, the detailed balance condition can also be seen as a condition for the *reversibility* of the system. With combination of

$$\pi_i = \exp(-E_i/kT)/Z \tag{38}$$

and equation 32 it follows that

$$\pi_\alpha = Z^*_\alpha/Z = (1/Z)\exp(-G_\alpha/kT) \tag{39}$$

---

[1] For the remainder of this thesis we will calculate with gradient basins.

which defines an equilibrium distribution for state α. This can be used (in combination with the detailed balance condition 37) to derive

$$\frac{u_{\beta\alpha}}{u_{\alpha\beta}} = \frac{e^{-G_\beta/kT}}{e^{-G_\alpha/kT}} \tag{40}$$

which can be extended to

$$\frac{u_{\beta\alpha}}{u_{\alpha\beta}} = \frac{e^{-(E_{S\alpha\beta}-G_\alpha)/kT}}{e^{-(E_{S\alpha\beta}-G_\beta)/kT}} \tag{41}$$

where $E_{S\alpha\beta}$ denotes the energy of the saddle connecting states α and β.
We are finally able to formulate an 'effective transition rate' $k_{\alpha\beta}$ from state α to state β. In our case the off-diagonal elements of matrix $\mathbf{U}$ (section 5.3) are given by

$$u_{\beta\alpha} = k_{\alpha\beta} = \Gamma_{\alpha\beta}\exp(-(E_{\alpha\beta} - G_\alpha)/kT) \tag{42}$$

G_α denotes the free energy of state α, k represents the Boltzmann-constant (which must not be confused with the transition rate $k_{\alpha\beta}$) and T the absolute temperature. $\Gamma_{\alpha\beta} = \Gamma_{\beta\alpha}$ is a prefactor that can be related to the entropy of the 'transition state'. In the simplest case it could be approximated by the multiplicity $\mu_{\alpha\beta}$ of the saddle point:

$$\Gamma_{\alpha\beta} = \Gamma_0\mu_{\alpha\beta} \tag{43}$$

The prefactor Γ_0 sets the time unit. At present we do not have a satisfactory model for Γ_0. Due to the fact that $\mu_{\alpha\beta}$ are small (section 4.3), we set it equal to 1, to simplify matters.

5.5 Reliability and Comparison

One of the most important facts that we always must be aware of is the fact that the model presented here builds onto the assumption that the dynamic behavior of a folding RNA molecule should be simulated using a *barrier tree*. In contrast to our model, the conformation space of RNA molecules *in vivo* is not limited to some 'macro-states' (represented by the local minima of the landscape), i.e. the dynamics develops by making use of very many secondary structures. This leads us to the question: What *is* the actual dynamic behavior of a RNA molecule and how can we prove the correctness of our simulations?

To give a reasonably correct answer to that question, we need to include the whole conformation space of the observed molecule in our calculations. As we learned in section 2.3, the number of suboptimal secondary structures of a given RNA sequence grows exponentially with the chain length n. Even for small molecules it becomes very soon very big, even as big as it cannot be treated any more within a computer simulation. The limiting factor concerning computer resources is RAM, as the transition matrix (section 5.3) has to be stored as a whole during diagonalization. Nevertheless it is possible to calculate the dynamic behavior for some reasonably small conformation spaces with up to a few thousand secondary structures on modern machines with 1GB RAM. We call this the *full process* - in contrast to the *tree process* within our model. Due to the fact that the full process includes the entire conformation space of a given RNA molecule, it represents the 'real' dynamic behavior of the sequence and hence is an ideal reference for our simulations. Again, we modified `barriers` to gather information on the neighborhood relations among all secondary structures. Within the *full process* we formulated transition rates between the different secondary structures using the Metropolis and the Kawasaki rule introduced in section 5.3.

6 Computational Results

In the last sections the theoretical background of this thesis and the underlying model was introduced. With knowledge of the fundamental properties of RNA chains, the move set, the landscape described by barrier trees, the stochastic model of Markov chains and the formalism given in the last section we are now able to investigate the dynamic behavior of RNA molecules of moderate size, i.e. this allows us to calculate the time-evolution of population probabilities of local minima on the barrier tree. The tool which does the effective calculation is called `markov` and was written in `ANSI C`.
This section is divided into three subsections:

- The first one can be seen as an introduction to `markov`. We will demonstrate its capabilities with a small RNA sequence of length $n = 15$.
- In the second part we take a closer look at a slightly longer RNA chain whose entire conformation space consists of 876 structures and hence the full process can still be treated within `markov`.
- Finally we show the capabilities of the algorithm when investigating a longer RNA sequence, whose full process cannot be calculated any more (due to the fact that there are too many suboptimal secondary structures).

6.1 A first example: sexi

As a first application of the algorithm we will analyse the small artificially designed RNA chain with sequence `ACUGAUCGUAGUCA` and length $n = 15$ (the same sequence, which we will denote 'sexi' from now on, was used before in section 4.2 to explain the output of `barriers`). There is a simple reason why sexi is an ideal model sequence: As it is very short, its conformation space $\mathcal{C}$ (see section 2.3) does only consist of 142 structures. Figure 24 shows a barrier tree of sexi, which gives an impression on the simple shape of the associated energy landscape. From the bar-file we know that local minimum 3

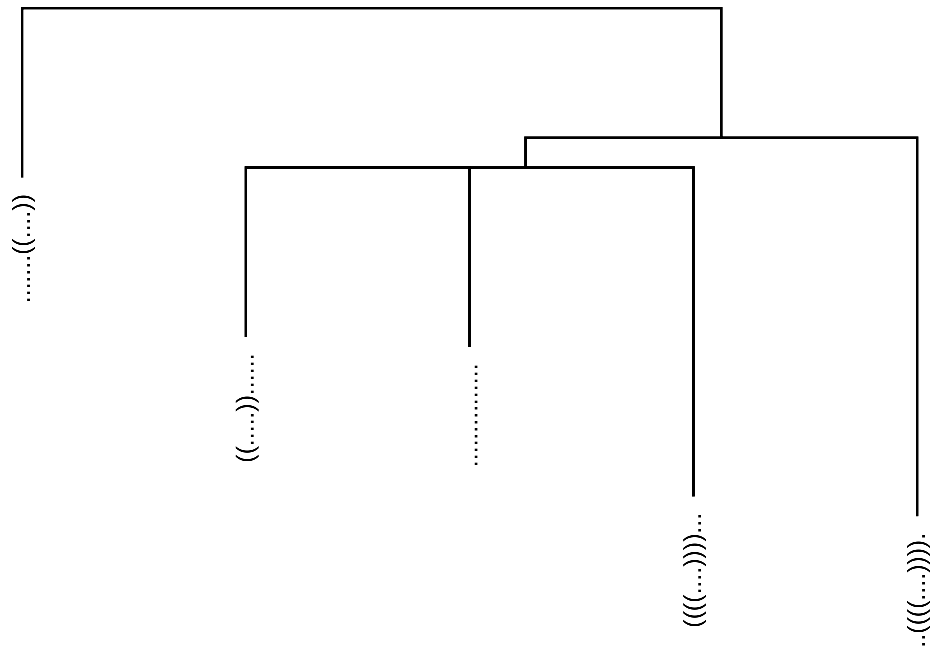


Figure 24: Barrier tree of the RNA sequence `ACUGAUCGUAGUCAC` with $n = 15$, illustrating that the conformation space is very small and there exist just 5 local minima (illustrated by bracket-dot-notation). Local minimum 3 denotes the open chain conformation. For details see text.

denotes the denatured open chain conformation. We further know that local minimum 3 was merged with local minimum 2, which itself was merged with the global minimum 1. Local minimum 4 is also connected with 2, whereas 5 is directly linked with the subtree containing the global minimum at higher energy. We start our simulation assigning local minimum 3 a population probability of 1, which means 100 percent of the population is situated in this local minimum initially. The upper part of figure 25 shows how the population probabilities of the lowest four local minima evolve with time, ending with the reach of an equilibrium distribution after approximately 1046 time steps. (Note that we use arbitrary time-units here). Figure 25 shows that at the beginning of the simulation, the population of 3 descends rapidly, allowing a population of the other local minima. Local minimum

4 reaches a population maximum at about 7 time-units, whereas 2 reaches its maximum after about 43 time-steps. At this time almost 100 percent of the total population is shared among local minima 1 and 2, just a little percentage is still populating 3 and 4.

The lower part of figure 25 reveils that the results gained within the simulation using the concept of the barrier tree are in general accordance with the 'real' dynamic behavior of the folding RNA molecule. Although we have to deal with a different time range (still in arbitrary units), the qualitative results are very similar. With the full process (and the standard Metropolis transition rates discussed in the last section), our model sequence reaches an equilibrium population distribution after approximately 2960 time units.

local minimum	population probability	secondary structure
1	0.6035	`..((((....)))).`
2	0.3459	`((((....)))...`
3	0.0294	`...............`
4	0.0197	`((.....))......`
5	0.0015	`.......((....))`

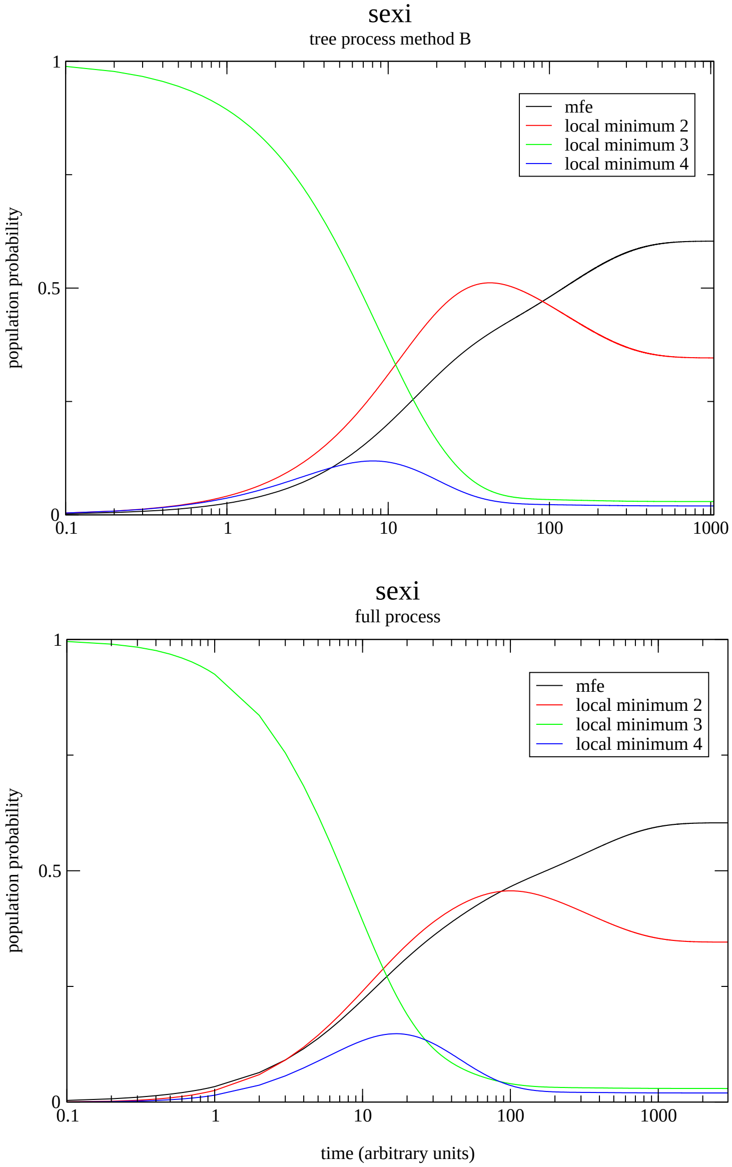


Figure 25: Dynamic behavior of the four lowest local minima of the model sequence sexi. The upper image shows the results of the simulated process from the barrier tree, whereas the lower image shows the results for the full process including all secondary structures assuming transition rates of the Metropolis-type. 'Method B' in the upper image denotes that transition rates between all local minima have been considered, not just between connectivities listed in the bar-file.

6.2 The medium-sized molecule: bertl

After illustrating the capabilities of `markov` with a small example in the last section we will now step onwards to a sequence with a bigger conformation space containing 876 secondary structures. The artificial sequence to be discussed here (denoted 'bertl') is `CGCGCUACUCCUAGAGCU` with $n = 18$. Although it is just slightly longer (3 bases) than sexi, the energy landscape now contains 11 local minima (figure 26). Appendix A lists the bar-file and a dot-plot.

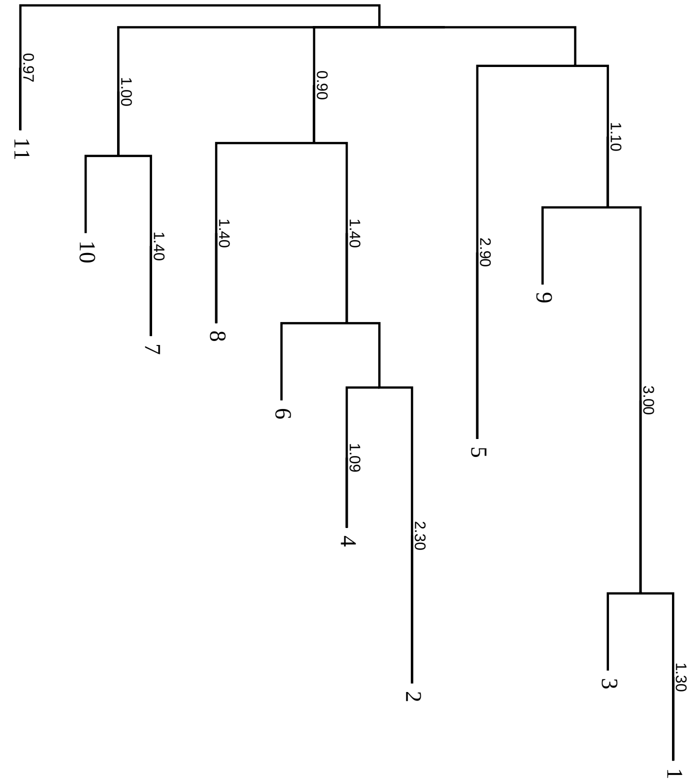


Figure 26: Barrier tree of the artificial RNA sequence bertl. Local minima 1 and 2 are separated by an energy barrier of 5.10 kcal/mol, local minima 7 on the left hand side represents the open chain conformation. Note that the subtrees containing 2 and 7 are connected with 1 at the same energy of 2.4 kcal/mol.

The ground state which folds into the the structure ...(((........))). has an energy of -3.30 kcal/mol. Figure 27 shows the dynamics of this second example molecule: As in the previous section, the upper picture shows the simulated process from the barrier tree, whereas the lower one shows the full process including all 876 suboptimal secondary structures calculated with the Metropolis rule.

We started the simulation with a population probability of 1 at the open chain conformation (local minimum 7). In both plots, 7 disappears so rapidly that its population probability can be neglected after approximately 100 steps. By observing the barrier tree we see that the lowest three local minima are situated in a rather narrow energy interval (in contrast to the other valleys). This elucidates that in the equilibrium case, there should be three minima which are observably populated: 1, 2 and 3. This fact is proved by the curves in figure 27.

Although generally the results of both runs (tree and full process) are very similar (especially the curve for local minimum 1), the curves for local minima 2 (red) and especially 5 (yellow) are slightly different. In the tree process, 2 has its population maximum at approximately 47 time steps (42,52 percent) and is (at this point in time) definitely more populated than 1 (36,60 percent). In contrast to that, 2 has its maximum after 270 time steps (37,85 percent) in the full process and hence is less populated than 1 at the same time (41,10 percent). Local minimum 5 is only slightly populated in the tree process, whereas it shows a population probability of even 16,25 percent after 37 time steps in the full process.

local minimum	population probability	secondary structure
1	0.5486	...(((........))).
2	0.2441	.((.(((....))).)).
3	0.1548	...(((.((...))))).
4	0.0294	.((............)).
5	0.0015	.......(((...)))..
7	0.034	....

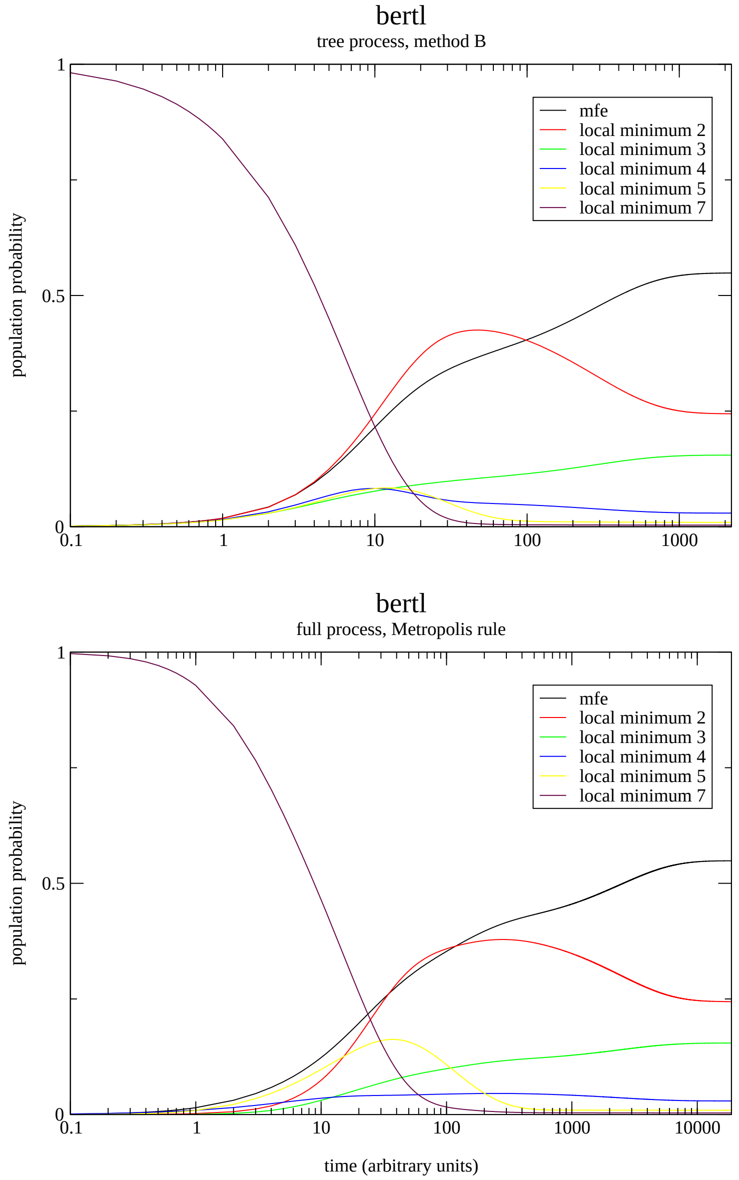


Figure 27: Dynamic behavior of the lowest local minima of the model sequence bertl. Similar to figure 25 in section 6.1, the upper image shows the results of the simulated process from the barrier tree, whereas the lower image shows the results for the full process including all secondary structures assuming transition rates of the Metropolis-type. Again, 'Method B' in the upper image denotes that transition rates between all local minima have been considered.

6.3 The switching molecule

After an exhaustive discussion of `markov` with small molecules in the last two sections, we will now leap to a RNA molecule with (1) a large conformation space of more than 60000 secondary structures and (2) a very interesting behavior. To be more precise, we will focus our investigations on the RNA sequence `GUGUUUGAGAGGAUAUGGCGUUUUUUUGGAUGC` which was used in [17] as an example of a bi-stable RNA sequence. Bi-stable RNA molecules (also denoted RNA switches) can fold into two or more thermodynamically stable secondary structures which are separated by a high energy barrier, which means that besides the subtree containing the global minimum, there are other dominating subtrees in the barrier tree (figure 28). Recently, artificial RNA switches have been designed. An impressive example is described in [51], where a sequence that can satisfy the base-pairing requirements of both the hepatitis delta virus self-cleaving ribozyme and an artificially selected self-ligating ribozyme, which have no base pairs in common, has been designed. Software tools for constructing RNA switches were introduced, see [17] and [20] for further details. Current research in our group focuses on RNA sequences that can fold into more than two stable secondary structures. We simulated two different scenarios with this RNA molecule.

- First, the same scenario which was used in the previous section: The open chain conformation is assigned a population probability of 1 at the beginning.
- Second, the refolding dynamics of this RNA switch when starting the simulation with the metastable structure.

Both runs show interesting dynamic behavior of the molecule.

Let us first discuss the case when starting with the denatured, open chain conformation (upper plot in figure 29). From the bar-file (not listed) we know that local minimum 81 in figure 28 corresponds to the open chain conformation. This is represented by the green trajectory, which falls off very quickly, enabling a slight population of the deepest minimum in the left subtree, 8, as

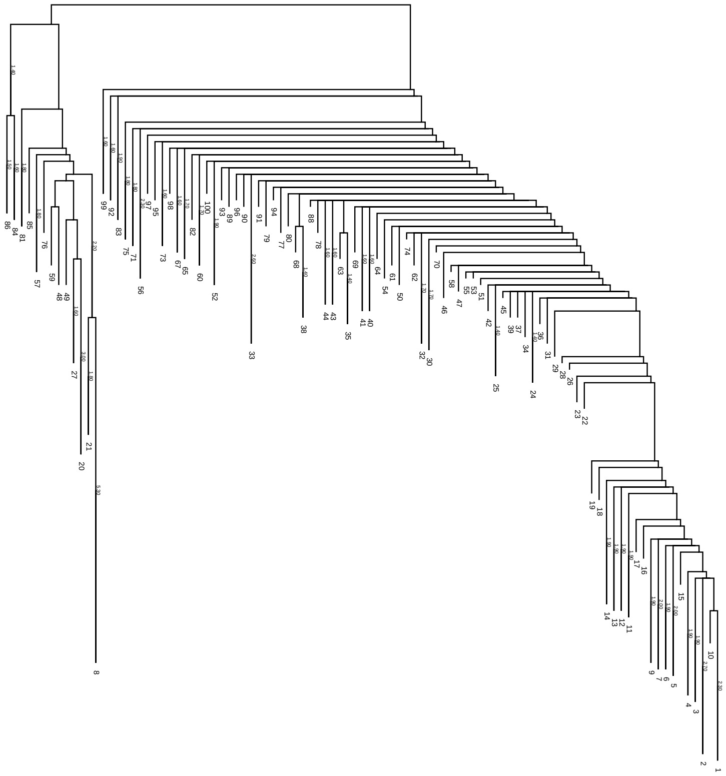


Figure 28: Barrier tree of the bi-stable RNA sequence `GUGUUUGAGAGGAUAUGGCGUUUU UUUGGAUGC`. The deep local minimum 8 (-6.7 kcal/mol) on the left hand side is separated from the subtree on the right hand side via an energy barrier of 10.10 kcal/mol. The energy of the mfe structure is -8.20 kcal/mol, directly followed by local minimum 2 with -8.10 kcal/mol. The denatured open chain conformation is represented by local minimum 81 in the very left part of the tree.

well as a population of local minimum 20. (Note that adjacent local minima are being populated in the region between approximately 2 and 50 time-steps as well, but due to the fact that neither of these local minima shows an effective population probability of more than 10 percent, they are not included in figure 29.) Nevertheless the more interesting region is between 10 and ap-

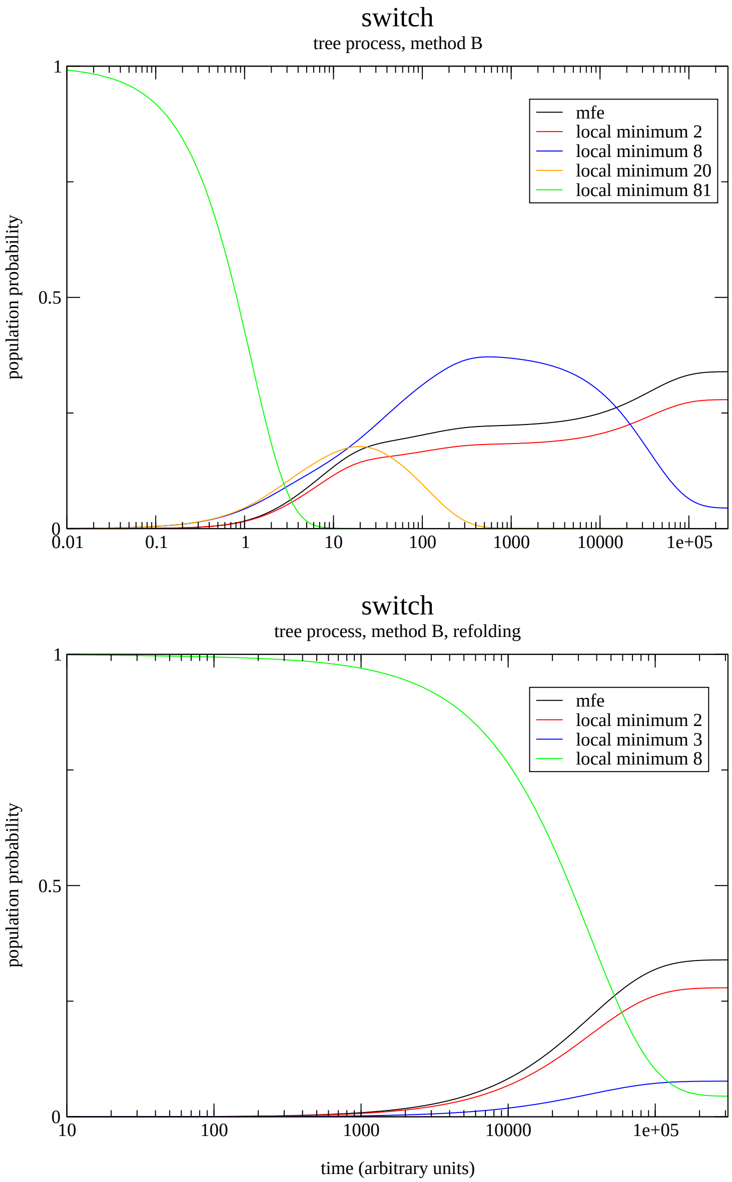


Figure 29: Dynamic behavior of the switching molecule. The upper plot shows the results of the simulation when starting with the open chain conformation (local minimum 81). Within this simulation, the left subtree from figure 28, i.e. local minimum 8 is being populated noticeably. The lower plot shows the results of a refolding-simulation from local minimum 8 to the mfe structure. Only the first three local minima in the right subtree are being populated clearly. Note that in both plots only noticeably populated local minima from the barrier tree are shown.

proximately 1000 time-steps. After its population maximum at 19 time steps, the trajectory of 20 falls off again, enabling a population of local minimum 8 containing the metastable structure `((((((....)))))).((((((....))))))` Besides 8, local minima 1 and 2 are being populated as well in this region. Note that the population probabilities of 1 and 2 have to be added to get an impression on the population ratio between the left and the right subtree in figure 28. Local minimum 8 has reached its population maximum after 530 time-steps. Finally, after 272100 time-steps the equilibrium has been reached. Local minimum 8 has fallen off and almost 62 percent of the total population is shared among local minima 1 and 2 in the right subtree and only 4.45 percent remain in local minimum 8.

A completely different scenario is given in the lower plot of figure 29. To be more precise, the *refolding* dynamics of our switching molecule is shown here. Refolding means that we start the simulation with a population probability of 1 in the lowest minimum of the left subtree in figure 28, local minimum 8. What stands out at first is the fact that during the first 1000 time steps, the whole population seems to remain in the left subtree. This is evident and can be explained by the high energy barrier (10.10 kcal/mol) that has to be overcome to change into the subtree containing the mfe structure. In the subsequent time region, say between 1000 and 100000 time steps, the population ratio changes dramatically: A large percentage crosses the energy barrier and hence the lowest local minima of the right subtree are populated. (Note that we have the same situation as in the upper plot here where only strictly populated local minima are shown.) Finally, after approximately 311000 time steps the equilibrium has been reached and more than 90 percent of the total population is situated in the right subtree. See the table below for the exact population values of selected local minima.

loc. min.	eq. pop. prob.	secondary structure
1	0.3394	`(((((..(((((((.....)))))))..)))))`
2	0.2789	`(..((..(((((((.....)))))))..))..)`
3	0.0770	`(((((..(((..((.....))..)))..)))))`
4	0.0633	`(..((..(((..((.....))..)))..))..)`
5	0.0411	`((((..((((((((.....))))))))..))))`
6	0.0349	`((((((((((((((.....))))))))))))))`
7	0.0338	`(..(..((((((((.....))))))))..)..)`
8	0.0445	`((((((....)))))).((((((....))))))`

Having the information about the folding kinetics within our barrier tree model, we are now able to compare these results with computational results of the same molecule generated with `kinfold` (section 5.3). This is shown in figure 31. The upper plot shows the results of 4000 `kinfold` - simulations of the observed molecule, i.e the first passage times of the folding starting with the open chain conformation and ending in the mfe structure. At first glance, two different folding mechanisms are visible: The first, *fast* mechanism describes the direct fold of a small percentage from the open chain conformation to the ground state (time ≤ 500). After the large plateau the majority of the folded molecules take times greater than 100000 time-units to reach the ground state (*slow* mechanism via the metastable structure). Note that the shape of this trajectory is similar to the trajectory for the mfe structure in the upper plot of figure 32, where we made the mfe an 'absorbing' state, meaning that in the transition matrix the rate to other states are extremely small.

The lower plot shows the *refolding* of our sequence from the metastable structure to the ground state. Evidently, this takes a long time as the high energy barrier separating local minima 8 and 1 on the barrier tree must be overcome. The lower plot of figure 32 shows the results of `markov` refolding-simulations from the metastable to the stable structure, yielding qualitative similar results.

Figure 30 shows the energy profile of the refolding from the metastable structure 8 to the mfe structure 1 (generated with `barriers`). Appendix B lists all secondary structures mentioned here. It is interesting that the refolding takes place via the open chain conformation (13). At the beginning, one of the two stems in (00) is opened, ending in an energetically favorable structure (07). Note that the open chain conformation must be visited because another nucleation center is needed. This is gained in (14). The subsequent, energetically unfavorable region can be explained with the fact that the nucleation region started in (14) is not yet optimal. After (23) has been visited, the way is opened for the formation of the mfe structure (36).

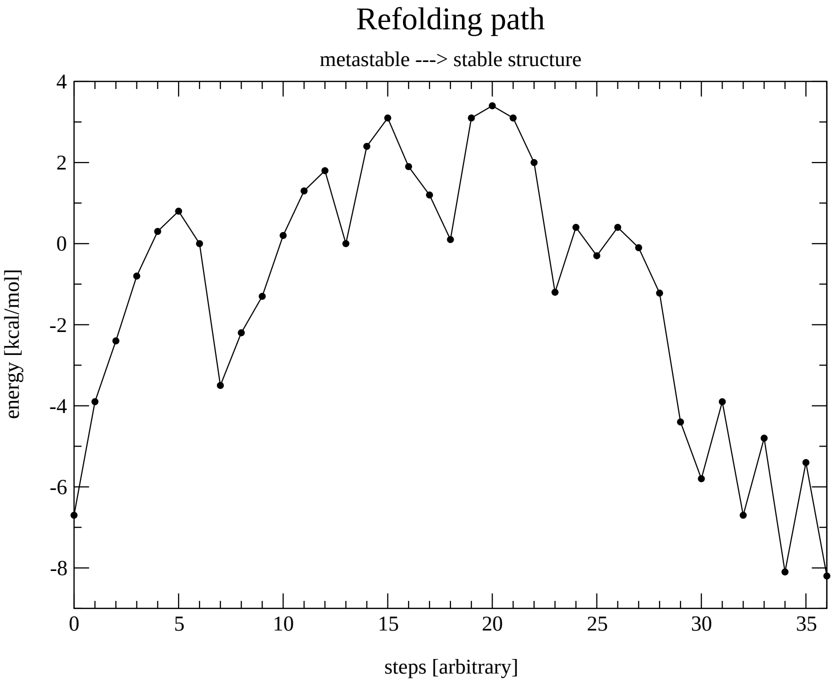


Figure 30: Energy profile showing the complete refolding path from the metastable structure to the mfe structure. Interesting steps are 7 (only one stem remains), 13 (open chain conformation) and 20 (unfavorable lonely pairs in structure). Appendix B lists all secondary structures displayed here.

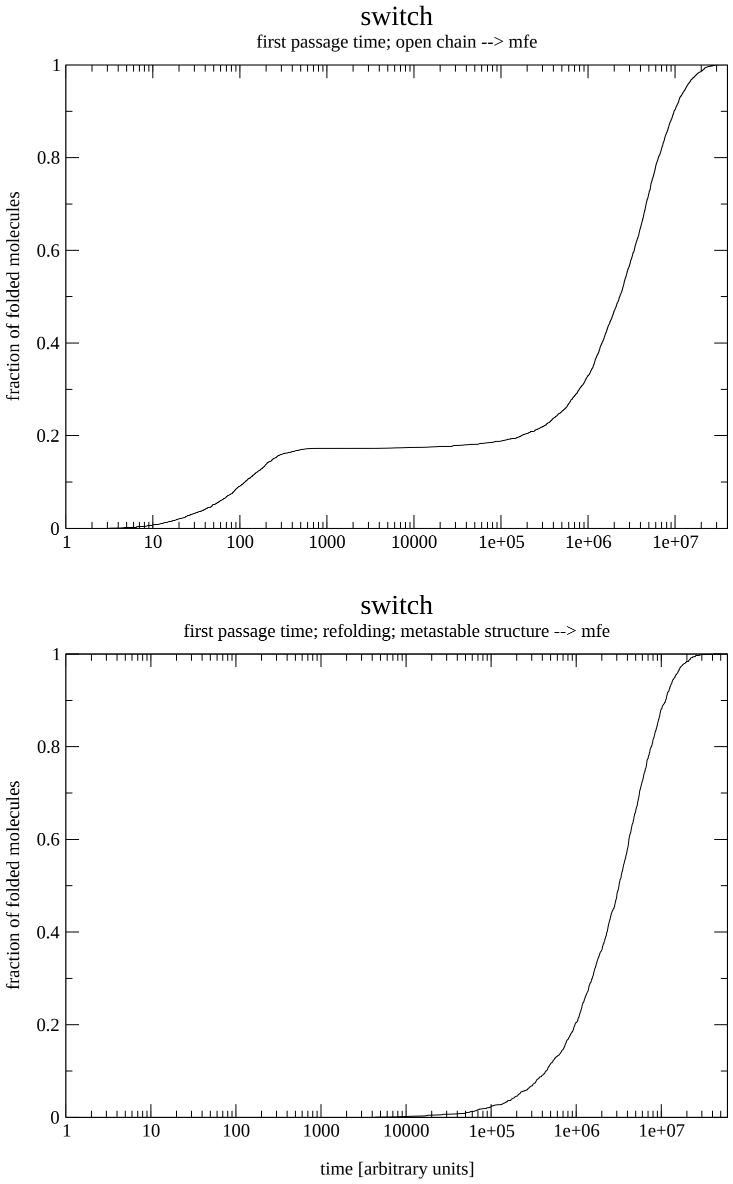


Figure 31: **upper plot:** 4000 `kinfold` - simulations of the switching molecule. Three dominating regions are visible: A small percentage folds directly (or within less than 500 time-units) to the ground state. Approximately 18-19 percent folds within 500 and 100000 time units (second area). Finally, the majority has folding times greater than 100000 time - units. **Lower plot:** 2000 `kinfold` - simulations showing the slow refolding dynamics of the RNA switch. For details see text.

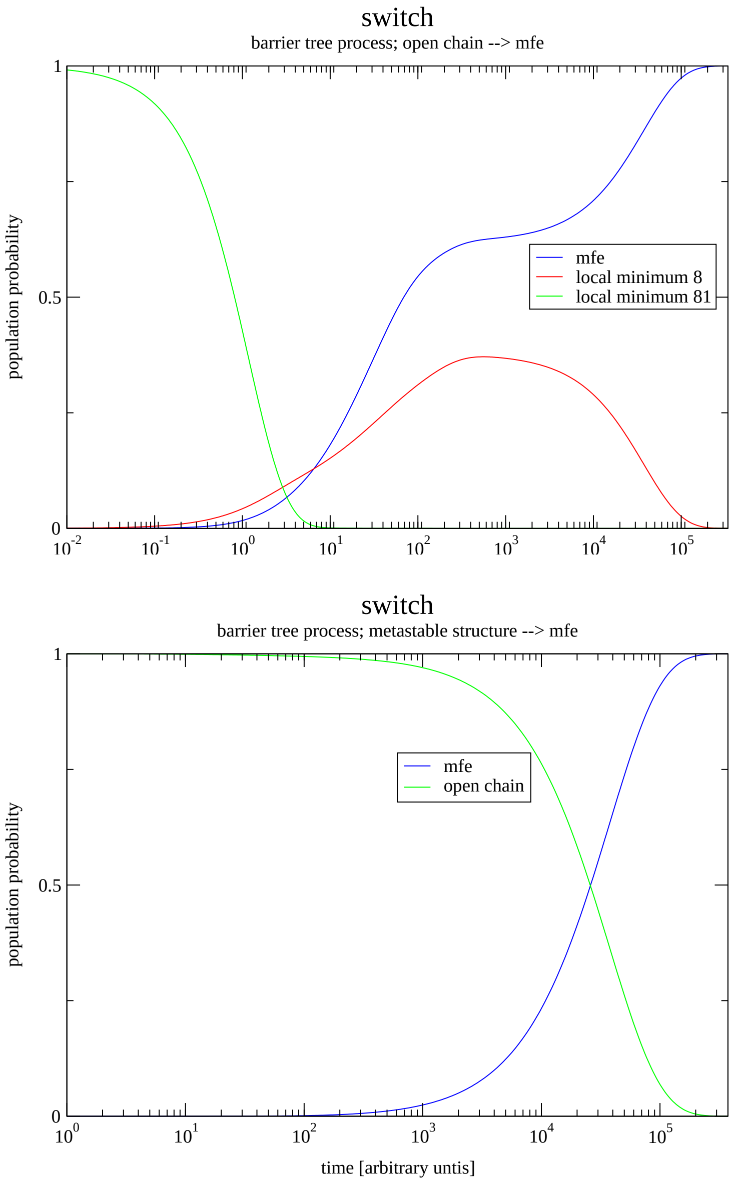


Figure 32: `markov`-simulations of the switching molecule with the assumption that the ground state is absorbing. Results gained here are qualitatively similar to the results gained from `kinfold`-simulations. Although the shape of all curves is similar fo figure 31, the time-depencency is different and can be explained with the pre-exponential factor within the transition rates.

7 Conclusion and Outlook

RNA is known to exhibit important tasks in living cells. It does not only serve as information-transmitting unit but also shows catalytic activity. RNA secondary structures provide a convenient form of coarse graining, hence their study yields information useful in the prediction of the full three dimensional structure as well as in the interpretation of the biochemical function of the molecules. The secondary structure model is sufficiently simple to allow efficient algorithms to compute (almost) any thermodynamic quantity of interest, yet it is still close enough to reality to address problems of particular interest.

The inter-conversion between different secondary structures is determined by a metric, called *move set*. The most elementary move set (at the level of secondary structures) consists of removal and insertion of a single base pair (with the assumption that no knots or pseudo-knots are inserted into the structure). Besides this simple move set, a slightly more sophisticated move set which enables additional base pair 'shift moves' is supported. These shift moves faciliate sliding of the two strands of a helix, bulge diffusion along the helix and the inter-conversion of partially overlapping helices, which are assumed to be important effects in the dynamics of RNA molecules.

Evidently, the structure of the energy landscape of a RNA molecule, i.e. the definition of ruggedness is associated inmost with the choice of the move set. In this thesis we introduced the tool `barriers` which allows an efficient computation of the energy landscape and yields a graphical representation of the landscape, so called *barrier trees*. A barrier tree gives an impression on the shape and ruggedness of the associated landscape and hence shows the distribution and energy ratios of local minima. With this tool at hand, it is possible to discuss *folding kinetics* of RNA at the level of barrier trees. To be more precise, it opens the door for a thorough investigation of the dynamic behavior of RNA molecules.

With this ansatz it was possible to formulate a *Markov process* in continuous time describing population probabilities of different local minima on the barrier tree, i.e. we were interested in the computation of trajectories describing the change of population rates while time elapses. Evidently, after a sufficiently long period of time, the dynamics end in a stable equilibrium distribution. The tool which does the effective calculations is called `markov` and was written in `ANSI C`.

To confirm the results of our simulations, we made use of two concepts: For smaller RNA chains we were able to compare the results gained from the tree process with a *full* process including *all secondary structures* on the energy landscape. For larger molecules we had to revert on the tool `kinfold` written by Christoph Flamm which calculates trajectories for single folding pathways using a stochastic ansatz.

As one would expect, the results gained from the full process are qualitatively very similar to the ones gained by the tree process. We were not able to detect significant discrepancies within our simulations. The only parameter which differs is time, but this is evident, since the rate constants between local minima in the barrier tree and those between secondary structures in the full process *are* different. We could also show that our results from the *barrier tree kinetics* are in general accordance with results calculated by `kinfold`. Again, the time constants are different.

The present work presents a first step towards a qualitative modeling of RNA folding kinetics. The main problems that still remain are:

- *Integration of multiple saddles into the model:* As mentioned above, the concept of multiple saddles on the energy landscape has to be integrated into the model of barrier tree kinetics. A transition between two states should be the more probable, the higher the saddle multiplicity connecting the two states is. In fact, it is necessary to include the *entropy* of the transition state into the model and hence draw conclusions about the border regions of basins / gradient basins. The remaining problem is that we hoped to find more multiple saddles than we actually did.

- *Calibration of the time axis:* The time-dependency of the trajectories calculated by `markov` is crucially connected with the choice of the pre-exponential factor in the transition rate. Although we tried to set up a model for Γ_0 (including the edges of basins / gradient basins) we were not able to define general rules for Γ_0. To be more precise, it is necessary to calibrate the time axis from the tree process in a way that it can be perfectly aligned to the trajectory calculated with `kinfold`, i.e. both curves should end at the same time.

- *We need a measure to qualify the difference between trajectories calculated with different models*: Although the trajectories from `markov` as well as those from `kinfold` show a qualitative similar run, there remain discrepancies. This is evident, as trajectories calculated with `kinfold` show the first passage times for reaching a pre-defined stop structure. On the other hand, trajectories from `markov` always end in a stable equilibrium distribution. Hence, it is necessary to refine the model in a way that the qualitative shape of both curves becomes more similar.

Although there remain several problems which have not been solved so far, we could think of an biologically relevant implementation of our kinetic barrier-tree-model (figure 33). Imagine a Flow-Reactor where a replicating species (sequence) can coexist in two conformations: a metastable (M) and a stable (S) one. The metastable structure is being formed with a rate constant k_M, the stable structure can be attained via the metastable structure with a rate constant k_{MS} as well as directly with a rate constant k_S. The sequence can replicate via the metastable structure with a rate constant k_r. Within

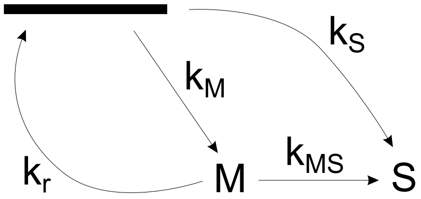


Figure 33: A schematic representation of a possible implementation of `markov`. For details, see the text

this simple model it would be interesting to let this system evolve for a certain period of time and to see how the replication behavior is changed. Will there be stable mutants? Evidently, k_M, k_S and k_{MS} can be calculates easily from the `markov`-simulations. The replication-rate k_r is connected with the rate-constant k_{MS} to reach the stable state from the metastable state. It is easy to see that, if a sequence remains in the metastable state for a long period of time, there is a high probability that many replication steps are made and 'fitter' species are generated.

Appendix A

Output of `barriers` for the RNA sequence bertl discussed in section 6.2. Data on basin sizes and partition functions are omitted.

```
     CGCGCUACUCCUAGAGCU
   1 ...(((........))). ( -3.30)    0  16.60    0
   2 .((.(((....))).)). ( -2.70)    1   5.10    1 ....(........)....
   3 ...(((.((...))))). ( -2.60)    1   0.60    1 ...(((.(.....)))).
   4 .((............)). ( -1.49)    2   1.09    1 .((.(........).)).
   5 .......(((...))).. ( -0.80)    1   2.90    1 ...(...(((...)))).
   6 .((....((...)).)). ( -0.50)    2   0.60    1 .((....(.....).)).
   7 .................. (  0.00)    1   2.40    1 .....(........)...
   8 ....(((....)))).... (  0.10)    2   1.40    1 .(..(((....)))..).
   9 ...((...((...)))). (  0.40)    1   0.60    1 ...(((...(...)))).
  10 .......((...)).... (  0.80)    7   0.60    1 .......(.....)....
  11 .(((...)((...)))). (  1.60)    1   0.97    1 .((......(...).)).
```

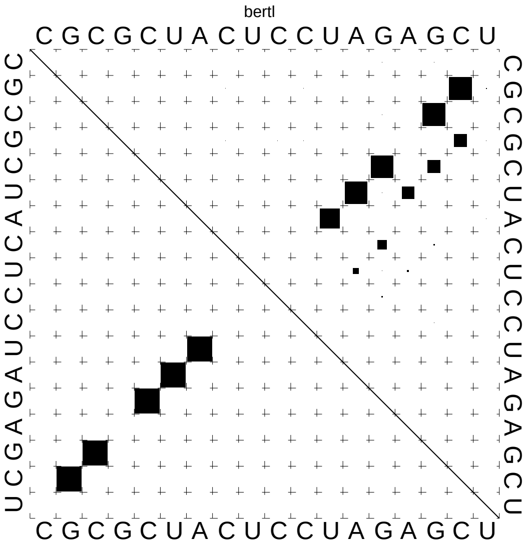


Figure 34: Dot plot of bertl. For details how to read it see figure 8

Appendix B

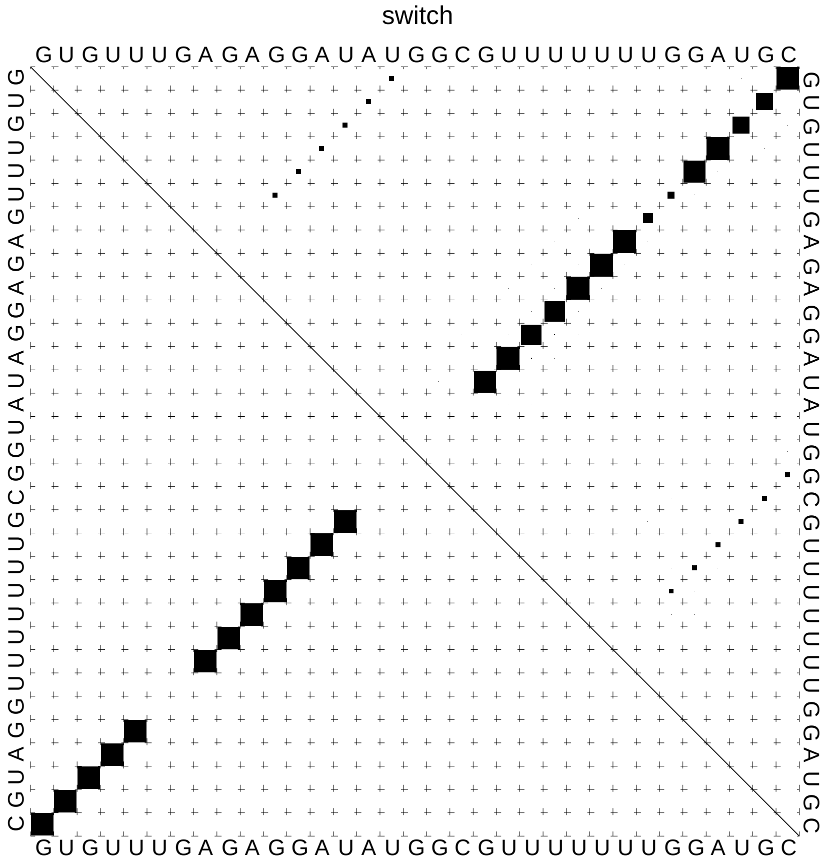


Figure 35: Dot plot of the swithing molecule (sequence `GUGUUUGAGAGGAUAUGGCGUUUUUUUGGAUGC`) treated in section 6.3. The lower left triangle shows the bape pairing probabilities within the thermodynamical equilibrium, whereas the upper right triangle displays few alternative suboptimal base pairing probabilities. The mfe structure in bracket-dot notation is `(((((..(((((((.....)))))))..)))))`

Additionally, we list the refolding path from the open chain conformation to the mfe structure here (energy in kcal/mol in brackets):

```
................................. (000.00)
...........(......).............. (002.40)
.........(.(......).)............ (003.10)
........((.(......).))........... (001.90)
.......(((.(......).))).......... (001.20)
......((((.(......).))))......... (000.10)
...(..((((.(......).))))....).... (003.10)
....(.((((.(......).))))....).... (003.40)
....((((((.(......).))))...)).... (003.10)
...(((((((.(......).))))...)))... (002.00)
(..(((((((.(......).))))...)))..) (-01.20)
(..(((((((.(......).)))...))))..) (000.40)
(..(((((((.(......).))...)))))..) (-00.30)
(..(((((((.(......).)...))))))..) (000.40)
(..(((((((.(......)....)))))))..) (-00.10)
(..(((((((.............)))))))..) (-01.22)
(..(((((((..(.......)..)))))))..) (-04.40)
(..(((((((..((.....))..)))))))..) (-05.80)
(..(((((((.(((.....))).)))))))..) (-03.90)
(..(((((((((((.....)))))))))))..) (-06.70)
(..((.((((((((.....)))))))).))..) (-04.80)
(..((..(((((((.....)))))))..))..) (-08.10)
((.((..(((((((.....)))))))..)).)) (-05.40)
(((((..(((((((.....)))))))..))))) (-08.20)
```

The *refolding path* between local minimum 8 (metastable structure) and local minimum 1 makes use of the following secondary structures:

```
00 ((((((....)))))).((((((....)))))) (-06.70)
01 ((((((....))))))..(((((....))))). (-03.90)
02 ((((((....))))))...((((....)))).. (-02.40)
03 ((((((....))))))....(((....)))... (-00.80)
04 ((((((....)))))).....((....)).... (000.30)
05 ((((((....)))))).....(......).... (000.80)
06 ((((((....))))))..(.........).... (000.00)
07 ((((((....))))))................. (-03.50)
08 .(((((....))))).................. (-02.20)
09 ..((((....))))................... (-01.30)
10 ...(((....))).................... (000.20)
11 ....((....))..................... (001.30)
12 .....(....)...................... (001.80)
13 ................................. (000.00)
14 ...........(......).............. (002.40)
15 .........(.(......).)............ (003.10)
16 ........((.(......).))........... (001.90)
17 .......(((.(......).))).......... (001.20)
18 ......((((.(......).))))......... (000.10)
19 ...(..((((.(......).))))....).... (003.10)
20 ....(.((((.(......).))))....).... (003.40)
21 ....((((((.(......).))))...)).... (003.10)
22 ...(((((((.(......).))))...)))... (002.00)
23 (..(((((((.(......).))))...)))..) (-01.20)
24 (..(((((((.(......).)))...))))..) (000.40)
25 (..(((((((.(......).))...)))))..) (-00.30)
26 (..(((((((.(......).)...))))))..) (000.40)
27 (..(((((((.(......)....)))))))..) (-00.10)
28 (..(((((((.............)))))))..) (-01.22)
```

```
29 (..(((((((..(.......)..)))))))..) (-04.40)
30 (..(((((((..((.....))..)))))))..) (-05.80)
31 (..(((((((.(((.....))).)))))))..) (-03.90)
32 (..(((((((((((.....)))))))))))..) (-06.70)
33 (..((.((((((((.....)))))))).))..) (-04.80)
34 (..((..(((((((.....)))))))..))..) (-08.10)
35 ((.((..(((((((.....)))))))..)).)) (-05.40)
36 (((((..(((((((.....)))))))..))))) (-08.20)
```

List of Figures

1 Rugged and smooth landscape 8
2 tRNA^{phe} structure 10
3 Molecular structure of a typical A-RNA 12
4 Secondary structure of tRNA^{phe} 14
5 Circular representation 15
6 Mountain Representation 16
7 Various representations of RNA secondary structure 17
8 Dot-plot of tRNA^{phe} 18
9 Density of states of the yeast tRNA^{phe} 20
10 Thermodynamic versus kinetic reaction coordinate 21
11 Elementary moves in the RNA folding algorithm 26
12 Defect diffusion 27
13 Shift moves 28
14 Conformation space of a short RNA molecule 30
15 Neighborhood of local extrema 31
16 A typical barrier tree 32
17 Influence of the moveset onto the barrier tree 34
18 Internal structure of `barriers` 37
19 The flodding algorithm 38
20 Output produced by `barriers` 39
21 Multiple Saddles 41
22 Multiple Saddles in a short RNA sequence 43
23 Multiple Saddles in tRNA^{phe} 44
24 Barrier tree and mountain plot of sexi 60
25 Dynamic behavior of sexi 62
26 Barrier tree of bertl 63
27 Dynamic behavior of bertl 65
28 Barrier tree of the switching molecule 67
29 Dynamic behavior of the switching molecule 68
30 Energy path for refolding metastable - stable structure 71

31 `kinfold`-Simulations of the switching molecule 72
32 `markov`-Simulations of the switching molecule 73
33 Schematic representation of an implementation of `markov` . . . 76
34 Dot plot of bertl 78
35 Dot plot of the switching molecule 79

References

[1] D. Aldous and J. A. Fill. Reversible Markov chains and random walks on graphs. available at `http://www.stat.Berkeley.EDU/users/aldous/book.html`.

[2] A. R. Banerjee, J. A. Jaeger, and D. H. Turner. Thermal unfolding of a group i ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry*, 32:153–163, 1993.

[3] O. Bastert, Dan Rockmore, P.F. Stadler, and G. Tinhofer. Landscapes on spaces of trees. submitted to Applied Mathematics and Computations.

[4] O.M. Becker and M. Karplus. The topolgy of multidimensional potential energy surfaces: Theory and applications to peptide structure and kinetics. *J. Chem. Phys.*, 106:1495–1517, 1997.

[5] N. Breton, C. Jacob, and P. Daegelen. Prediction of sequentially optimal RNA secundary structures. *J. Biomol. Struct. Dyn.*, 14:727–740, 1997.

[6] P. Brion and E. Westhof. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, 26:113–137, 1997.

[7] J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B . L. Golden, A. A. Szewczak, C. D. Kundrot, T. R. Cech, and J. A. Doudna. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science*, 273:1678–1685, 1996.

[8] T. R. Cech. RNA as an enzyme. *Scientific American.*, 11:76–84, 1986.

[9] Hue Sun Chan and Ken A. Dill. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins: Structure, Function, and Genetics*, 30:2–33, 1998.

[10] P. E. Cole, S. K. Yang, and D. M. Crothers. Conformational changes of transfer ribonucleic acid. equilibrium phase diagrams. *Biochemistry*, 11:4358–4368, 1972.

[11] D. M. Crothers, P. E. Cole, C. W. Hilbers, and R. G. Shulman. The molecular mechanism of thermal unfolding of escherichia coli formylmethionine transfer RNA. *J. Mol. Biol.*, 87:63–88, 1974.

[12] J. Cupal. The density of states of RNA secondary structures. Master's thesis, University Vienna, 1997.

[13] J. Cupal, I. L. Hofacker, and P. F. Stadler. Dynamic programming algorithm for the density of states of RNA secondary structures. In R. Hofstädt, T. Lengauer, M. Löffler, and D. Schomburg, editors, *Computer Science and Biology 96 (Prooceedings of the German Conference on Bioinformatics)*, pages 184–186, Leipzig, Germany, 1996. Universität Leipzig.

[14] D. E. Draper. Strategies for RNA folding. *Trends Biochem. Sci.*, 21:145–149, 1996.

[15] F.F. Ferreira, J.F. Fontanari, and P.F. Stadler. Landscape statistics of the low autocorrelated binary string problem. *J. Phys. A: Math. Gen*, 33:8635–8647, 2000.

[16] Ch. Flamm. *Kinetic Folding of RNA*. PhD thesis, University Vienna, 1998.

[17] Ch. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of Multi-Stable RNA Molecules. *RNA*, 7:254–265, 2001.

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

[19] P. Garstecki, T.X. Hoang, and M. Cieplak. Energy landscapes, supergraphs and "folding funnels" in spin systems. *Phys. Rev. E*, 60:3219–3226, 1999.

[20] R. Giegerich, D. Haase, and M. Rehmsmeier. Prediction and visualization of structural switches in RNA. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 4, pages 126–137, Singapur, 1999. World Scientific Press.

[21] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[22] T. C. Gluick and D. E. Draper. Thermodynamics of a pseudoknotted mRNA fragment. *J. Mol. Biol.*, 241:246–262, 1994.

[23] C. Guerrier-Takada and S. Altman. Catalytic activity of an RNA molecule prepared by transcription in vitro. *Science*, 223:285–286, 1984.

[24] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribunuclease p is the catalytic subunit of the enzyme. *Cell.*, 35:849–857, 1983.

[25] J. P. Hansen and I. R. MacDonald. *Theory of simple liquids.* Academic Press Inc., London, 2nd ed. edition, 1986.

[26] P. G. Higgs and S. R. Morgan. Thermodynamics of RNA folding. when is an RNA molecule in equilibrium. In F. Morán, A. Moreno, J.J. Merelo, and Chacón, editors, *Advances in Artificial Life*, pages 852–861, Berlin, 1995. ECAL 95, Springer Verlag.

[27] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[28] I. L. Hofacker, W. Fontana, P. F. Stadler, and P. Schuster. `Vienna RNA Package`. `http://www.tbi.univie.ac.at/~ivo/RNA/`, 1994-2001. (Free Software).

[29] I. L. Hofacker, M. A. Huynen, P. F. Stadler, and P. E. Stolorz. Knowledge discovery in RNA sequence families of HIV using scalable computers. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR*, pages 20–25, Menlo Park, CA, 1996. AAAI Press.

[30] J. A. Howell, T. F. Smith, and M. S. Waterman. Computation of generating functions for biological molecules. *SIAM J. Appl. Math.*, 39:119–133, 1980.

[31] M. A. Huynen, A. S. Perelson, W. A. Vieira, and P. F. Stadler. Base pairing probabilities in a complete HIV-1 RNA. *J. Comp. Biol.*, 3:253–274, 1996. SFI preprint 95-07-057, LAUR-95-1613.

[32] B. R. Jordan. Computer generation of pairing schemes for RNA molecules. *J. Theor. Biol.*, 34:363–378, 1972.

[33] G. F. Joyce. Amplification, mutation and selection of catalytic RNA. *Gene*, 82:85–87, 1989.

[34] G. F. Joyce. RNA evolution and the origins of life. *Nature*, 338:217–224, 1989.

[35] G. F. Joyce. The rise and fall of the RNA world. *The New Biologist*, 3:399–407, 1991.

[36] K. Kawasaki. Diffusion constants near the critical point for time-dependent Ising models. *Phys. Rev.*, 145:224–230, 1966.

[37] T. Klotz and S. Kobe. "valley structures" in the phase space of a finite 3d ising spin glass with $\pm i$ interactions. *J. Phys. A: Math. Gen*, 27:95–100, 1994.

[38] M. Lu and D. E. Draper. Bases defining an ammonium and magnesium ion-dependent tertiary structure within the large subunit ribosomal RNA. *J. Mol. Biol.*, 244:572–585, 1994.

[39] A.A. Markov. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. 1907. Reprinted in Appendix B of R. Howard: *Dynamic ProbabilisticSystem.* volume 1:Markov Chains, 1971.

[40] H. M. Martinez. An RNA folding rule. *Nucl. Acids Res.*, 12:323–324, 1984.

[41] D.H. Mathew, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288:911–940, 1999.

[42] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[43] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[44] A. Mironov and A. Kister. A kinetic approach to the prediction of RNA secondary structures. *J. Biomol. Struct. Dyn.*, 2:953–962, 1985.

[45] A. Mironov and V. F. Lebedev. A kinetic model of RNA folding. *BioSystems*, 30:49–56, 1993.

[46] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA*, 77:6309–6313, 1980.

[47] R. Nussinov, G. Piecznik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35:68–82, 1978.

[48] H. W. Pley, K. M. Flaherty, and D. B. McKay. Three-dimensional structure of a hammerhead ribozyme. *Nature*, 372:68–74, 1994.

[49] D. Pörschke. Model calculations on the kinetics of oligonucleotide double helix coil transitions. Evidence for a fast chain sliding reaction. *Biophys. Chem.*, 2:83–96, 1974.

[50] D. Riesner, G. Maass, R. Thiebe, P. Philippsen, and H. G. Zachau. The conformational transitions in yeast tRNA^{Phe} as studied with tRNA^{Phe} fragments. *Eur. J. Biochem.*, 36:76–88, 1973.

[51] E. A. Schultes and D. P. Bartel. One sequence, two ribozymes: A mechanism for the emergence of new ribozyme folds. *Science*, 289:448–452, 2000.

[52] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 17:213, 1971.

[53] P. F. Stadler. Towards a theory of landscapes. In R. Lopz-Pea, R. Capovilla, R. Garca-Pelayo, H. Waelbroeck, and F. Zertuche, editors, *Complex Systems and Binary Networks (Proceeding of the Guanajuato Lectures 1995)*, pages 77–163. Springer-Verlag, 1996.

[54] A. Stein and D. M. Crothers. Conformational changes of transfer RNA. the role of magnesium(II). *Biochemistry*, 15:160–167, 1976.

[55] A. A. Suvernev and P. A. Frantsuzov. Statistical description of nucleic acid secondary structure folding. *J. Biomol. Struct. Dyn.*, 13:135–144, 1995.

[56] M. Tacker, W. Fontana, P. F. Stadler, and P. Schuster. Statistics of RNA melting kinetics. *Eur. Biophys. J.*, 23:29–38, 1994.

[57] H. Taketomi, Y. Ueda, and N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. 1. the effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.*, 7:445–459, 1975.

[58] M. S. Waterman. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.*, 1:167–212, 1978.

[59] M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Math. Biosc.*, 42:257–266, 1978.

[60] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structure. *Biopolymers*, 49:145–165, 1998.

[61] P. P. Zarrinkar and J. R. Williamson. Kinetic intermediates in RNA folding. *Science*, 265:918–924, 1994.

[62] P. P. Zarrinkar and J. R. Williamson. The kinetic folding pathway of the tetrahymena ribozyme reveals possible similarities between RNA and protein folding. *Nature Struct. Biol.*, 3:432–438, 1996.

[63] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[64] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.

[65] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, 9:133–148, 1981.

Curriculum vitae

Michael Wolfinger
* 7. Juli 1976 in Linz, Oberösterreich

Ausbildung

09/1982 – 07/1986	Volksschule Goethestraße, Linz
09/1986 – 06/1994	AHS, Kollegium Aloisianum, Linz
06/1994	Matura
10/1994 – 06/1995	Präsenzdienst beim FlHB3, Hörsching, Oberösterreich
10/1995 – 03/2001	Chemiestudium an der Universität Wien
01/2000 – 03/2001	Diplomarbeit am Institut fuer Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien, bei Prof. Peter Stadler in der Gruppe von Prof. Peter Schuster