Kinetic Folding of RNA

DISSERTATION
zur Erlangung des akademischen Grades
Doktor rerum naturalium

Vorgelegt der
Formal- und Naturwissenschaftlichen Fakultät
der Universität Wien

von
Christoph Flamm

am Institut für Theoretische Chemie und Strahlenchemie

im August 1998

Abstract

The ability to fold into well-defined native conformation is a prerequisite for biologically functional biopolymers. Since RNA secondary structure can be computed relatively easily and accurately it provides an ideal model system for theoretical investigations. The topology of the energy landscape of a given sequence strongly influences its folding pathways and mechanisms. The energy landscape for RNA molecules is believed to be rugged, exhibiting many deep local optima, in which the folding process may become trapped. In this thesis a kinetic folding algorithm has been developed in order to study the dynamics of RNA folding on such an energy landscape.

The new algorithm uses the most elementary move set possible for the inter conversion of RNA secondary structure. It consists in the insertion or the removal of single base pairs, as well as the exchange of one pairing partner in a base pair. Since the changes made during one simulation step are small, no unrealistic assumptions about the transition rates have to be made. Furthermore a more realistic concept of a folding path arises, if the introduced structural changes are small.

Folding simulations of natural and artificial tRNA sequences exhibit cases where the sequence finds the native state efficiently and often via the same intermediate structures, as well as cases where a large fraction of runs get trapped in local minima from which they cannot escape on the time-scale of the simulation. By prohibiting base pairing for a few crucial nucleotides, the base modifications present in natural tRNAs strongly bias the folding kinetics as well as the equilibrium ensemble towards the native state.

An analysis of the folding behaviour of various tRNAs shows, that the folding process is hierarchically organized. Local secondary structure elements form early and progressively reorganize into larger sub-domains during the folding process. Often secondary structure elements near the 5'-end form faster, than comparable ones near the 3'-end. This might well be a result of evolutionary selection of sequences to support efficient folding during transcription.

Information about folding paths can be inferred directly from folding simulations. In particular, important kinetic traps can be easily identified. For small RNA molecules it is possible to observe the escape from such traps within the simulation time. Simulations of SV-11, an RNA molecule with a known metastable structure, are in excellent agreement with experimentally measured data.

Zusammenfassung

Die Faltung von Biopolymeren in einen wohl definierten Grundzustand, ist eine Voraussetzung dafür, daß diese Moleküle ihre biologischen Funktionen erfüllen können. Die relative Leichtigkeit, mit der sich die Sekundärstruktur von RNA berechnen läßt, macht diese Molekülklasse zu einem idealen Studienobjekt für theoretische Untersuchungen. Die Topologie der Energielandschaft, die dem Faltungsprozeß eines RNA-Molekäs zugrunde liegt, beeinflußt sowohl den Mechanismus als auch den Faltungsweg im Speziellen. Es wird vermutet, daß die Energielandschaft von RNA auf Grund vieler lokaler Optima von sehr rauher Gestalt ist. Der Faltungsprozeß kann deshalb in einem der vielen lokalen Minima zum Stillstand kommen.

In der vorliegenden Arbeit wird ein neuer kinetischer Faltungsalgorithmus vorgestellt, der das Studium der Faltungsdynamik von RNA-Molekülen gestattet. Der Algorithmus benützt einen Satz elementare Transformationen um RNA Sekundärstrukturen in einander umzuwandeln, die alle auf einzelnen Basenpaaren operieren. Solche Transformationen sind beispielsweise das Einsetzen eines Basenpaars in eine gegebene Sekundärstruktur oder das Entfernen eines Basenpaars aus derselben. Da die strukturellen Veränderungen bei solchen Transformation im allgemeinen klein sind, wird ein plausibler Faltungsweg erhalten, ohne das unrealistische Näherungen für die Übergangsraten zwischen zwei Strukturen gemacht werden müssen.

Faltungssimulationen von natürlichen und künstlichen tRNA-Molekülen zeigten, daß die faltende Kette dem Grundzustand, effizient und oftmals über eine Kaskade ähnlicher Zwischenstrukturen, zustrebt. In manchen Fälle bleibt der Faltungsprozeß allerdings in einem lokalen Minimum hängen. Einige wenige modifizierte Basen in den natürlichen tRNA Sequenzen reichen aus, um das Auffinden des Grundzustandes zu erleichtern.

Aus Faltungssimulationen läßt sich sehr einfach Information über Faltungswege erhalten, die dazu benützt werden kann, kinetische ‘Fallen’ aufzuspüren, sowie jene Wege zu studieren, auf denen RNA-Moleküle aus diesen ‘Fallen’ entkommen. Der Faltungsprozeß selbst scheint hierarchisch organisiert zu sein. Lokale Sekundärstrukturelemente bilden sich früh aus und ordnen sich später zu strukturell größeren Einheiten um. Sekundärstrukturelemente am 3’-Ende bilden sich schneller als solche am 5’-Ende, was effiziente Faltung während der Transkribtion gestattet. Die Faltungssimulationen von SV-11, einem RNA-Molekül das eine metastabile Struktur ausbildet, stehen in exzellenter Übereinstimmung mit dem Experiment.

Contents

1 Introduction

1.1 General Context

Polymers are macromolecules, consisting of a linear arrangement of building blocks. The building blocks, or *monomers*, are linked together by covalent bonds to form the *sequence*. A “homopolymer” is built up by one type of monomer. The physical properties of homopolymers are essentially determined by the nature of the monomer, the length of the sequence and the nature of the junction between the monomers (e.g. cis-trans-isomery).

Nearly all biopolymers are “heteropolymers”, hence the sequence is built up by a hand full of different monomers. For example the building blocks for proteins are 20 amino acids, and those for RNA are 4 nucleotides Adenin (**A**), Guanin (**G**), Cytosin (**C**) and Uracil (**U**). In addition to the length and the nature of the monomer junction, the physical properties of heteropolymers are strongly influenced by the succession of the monomers along the sequence.

Polymers have the ability to fold back on themselves, due to interactions between individual residues of the sequence. If the interaction between individual residues is weaker than the interaction between residues and solvent molecules, than both homo- and heteropolymers tend to form arbitrary compact conformations called “random coils”. For most homopolymers the interactions between residues are unspecific. In contrast, for biopolymers like proteins and RNA these interactions are specific, and can lead to the adaption of a “unique” compact conformation called “native state”. During the structure formation process both, RNA and proteins, try to minimize the solvent exposure of hydrophobic residues by burying these residues in the interior of the structure. But it is self-evident from the different chemical nature of RNA and proteins, that the ways how these macromolecules achieve their compact conformation is different. For proteins the driving force of the collapse into compact conformations is the formation of a hydrophobic core. For RNA the formation of compact conformation is promoted by the tendency to maximize the stacking interaction between base pairs.

Due to the close inter-relation between function and structure, it is essential for living cells, that a folding macromolecule not only adopts its correct and functional conformation, but does this also in a biological relevant sufficiently short time. To gain more insight into performance and control of biochemical reactions, it is indispensable to understand the physical principles and mechanisms that underlie the folding process of biological macromolecule.

Unfolded proteins contain numerous solvent-exposed hydrophobic regions and therefore have a great tendency to form both intramolecular and intermolecular aggregates. *Molecular chaperones* are proteins that function to prevent or reverse such improper associations. The molecular chaperones comprise several unrelated classes of proteins including heat shock proteins Hsp70, chaperonins of the Hsp60 family (GroEl in *E. coli*, Cpn60 in chloroplasts), chaperonins of the Hsp10 family (GroES in *E. coli*, Cpn10 in chloroplasts) and nucleoplasmins.

The mechanism by which molecular chaperones carry out their functions is not yet understood in detail. However, many of them are ATPases, which bind to unfolded polypeptides and apparently apply the free energy of ATP hydrolysis to effect their release in a favorable manner. For example, certain Hsp70 proteins bind to not yet fully synthesized polypeptide chains as they emerge from the ribosome. Ulrich Hartl [53] has demonstrated that GroEL and GroES act in concert in an ATP-driven process to enclose unfolded proteins in a protected environment that prevent their non-specific aggregation while they spontaneously fold to their native conformations. The energy provided by ATP hydrolysis is used to disrupt incorrect interactions allowing a "misfolded" protein to escape from kinetically trapped conformations [85]. This mechanism is supported by the observation that chaperonins do not increase the rate of protein folding [77, 103, 128, 129] but, rather increase the yield of correctly folded product [17, 68, 78, 120, 139]. Some of them where shown can to slow down folding.

Although *in vivo* protein folding can be guided by molecular chaperones many proteins fold to their native state in the absence of accessory proteins, albeit with low efficiency. Moreover the molecular chaperones are not components of the native state of the proteins whose folding they facilitate. Hence they mediate the proper folding of a polypeptid to a conformation governed solely by the polypeptide's amino acid sequence.

In vitro folding experiments of several tRNAs, self-splicing group I introns and 5S rRNA showed, that RNA-molecules as-well can get kinetically trapped in non-active alternative conformations [133]. Such "misfolded" RNAs can be renatured to their active conformation by non-specific RNA-binding proteins [57]. The RNA folding problems observed *in vitro* could be of relevance to the *in vivo* behaviour of RNA. These experimental results brought Richard Karpel [73] to suggest the hypothesis, that non-specific RNA-binding proteins act as a kind of "RNA chaperones" in the living cell, to facilitate proper RNA folding. Till today there exist no established examples supporting the existence and action of such "RNA chaperones" *in vivo*. For a more detailed review on the hypothesis of RNA chaperones see a review by Kevin Weeks [138].

From a theoretical point of view, the problem of how biopolymers achieve their native state splits up into two aspects. The first aspect is the structure prediction problem. The second aspect deals with the dynamics of the folding process itself.

Since the sequence of a biopolymer specifies its three-dimensional structure, it should be possible, at least in principle, to predict its native structure solely from the knowledge of its sequence. The fact that experimental methods like X-ray crystallography or NMR-spectroscopy yield time-averaged "snapshots" of the structure of a biopolymer may leave the false impression that biopolymers have fixed and rigid structures. In fact, as is becoming increasing clear, biopolymers like proteins or RNA are flexible and rapidly fluctuating molecules whose structural mobilities have functional significance. The native states of proteins and RNA consists of a large ensemble of closely related and rapidly inter-converting conformational sub-states of nearly equal stabilities.

Theoretical methods for the three-dimensional structure prediction of proteins and RNA are still kind of an art and require extensive computation. Besides the sequence these methods require additional information from spectroscopy, chemical probing or biochemical degradation. The enormous difficulty in making such calculations reliable, sufficiently accurate and computational tractable has, so far, limited their success.

However the structure prediction problem for both proteins and RNA can be solved with reasonable accuracy on the level of secondary structure. The secondary structure of proteins is defined as the local conformation of the backbone, and is formed by hydrogen bonds between backbone atoms. The secondary structure of RNA is defined as the pattern of base pairs, which is formed by hydrogen bonds between atoms of the four bases. Thus, in contrast to proteins, the secondary structure of RNA is formed by the "side chains". In the following structure means always secondary structure and we shall mention explicitly when the 3D structure is considered. For RNA powerful algorithms [100, 146] based on the method of dynamic programming [10] and experimentally measured energy parameters [40, 55, 69, 131] have been developed. Using these algorithms the sequence to structure map for RNA [38, 39, 118] and its consequences for evolutionary adaption [65] have been characterized in detail.

For the protein secondary structure prediction Peter Chou and Gerald Fasman [23] devised the most popular algorithm. The propensity for an individual amino acid to adopt a local conformation (α helical, β strand or coil) is evaluated from a database of known structures as a ratio of the occurrences in one local conformation to the number of examples not in that local conformation. This method has the advantage of being easy to use and relatively accurate (∼50–55%). It suffers from the slow increase in accuracy with the increasing data base. Other methods are "homology modeling [76, 96]", "threading [37, 79]" or use of "knowledge based potentials [119]".

These algorithms, however, use heavy input of known protein structures from databases, and yield in many cases only approximate structures. In many situations, such an approximate structure may be useful for answering the biologically relevant questions, or for designing mutagenesis experiments. The structural information obtained by the methods mentioned above is seldom detailed or reliable enough to investigate the protein folding landscape. In contrast to the RNA case, it is therefore not possible to study the sequence to structure map for proteins by explicit folding. However the topology of the sequence to structure map for proteins can be probed using inverse folding techniques. Such investigations [8] reveal surprising similarities between the generic properties of the sequence to structure map of RNA and proteins.

The second question concers the kinetics of folding, the approach to the essentially unique folded state, which has been extensively investigated within the protein field. Analytic studies based on "beads on a string" models [16, 41, 102, 116] and simulations of simplified lattice [30, 52] or off-lattice protein models [11, 48, 67] uncovered fundamental aspects of protein folding dynamics. Theory and experiment have converged to yield the basic principles and the particular mechanisms for initiation of folding. The ability to analyse structure at a level of individual residues in polypeptides and denatured states using NMR spectroscopy as well as, in unstable intermediates and transition states using protein engineering methods, has permitted detailed analyses of folding pathways. The results from experiments and simulations resulted in a synergistic agreement between experiment and theory.

While a lot of theoretical and experimental questions concerning protein folding kinetics have been extensively investigated, the information available on similar questions for RNA is rather sparse. Reasons for this difference in knowledge may be rooted in the fact, that for a long time RNA molecules have been viewed as a largely passive class of molecules within the interplay of metabolism.

About a decade ago the discovery of RNA molecules with catalytic activities [20, 72] (ribozymes) and the evidence for an active role of messenger and ribosomal RNAs in gene expression [28, 66], provided convincing proof that RNA molecules are much more functionally sophisticated than previously assumed. The great ability of RNA to catalyze chemical reactions lies within its propensity to fold into three dimensional structures by forming helices via *Watson-Crick* base pairing that delineates single-stranded regions or loops capable of creating binding sites for various substrates and metal ions.

In vitro selection and evolution methods [82, 130] have proved to be very successful for generating "new ribozymes" (at least in part). The relative ease with which RNA binds metal ions can explain the success of all natural and most artificial ribozymes. The catalytic repertoire of ribozymes include reactions like phosphorylation, ligation, polymerization transesterification or cleavage of bonds. For a detailed discussion on the structural and functional complexities of ribozymes see a recent review by Luc Jaeger [70].

1.2 The Folding Problem

Biopolymers achieve their native conformation by spontaneous folding. The native state seems to be the most stable one, since it is commonly adopted through folding from different starting conformations as has been shown experimentally for proteins [3, 4]. For the great majority of biopolymers the folding itself happens under physiological conditions on time scales of less then a minute. However an exhaustive search of the conformational space to find the native state (based on equal probabilities of conformations) would take a biopolymer of moderate length at least billions of billions of years. This puzzle of finding the 'needle' (native state) in the haystack (conformation space) and doing so quickly is called the "Levinthal paradox" [29, 147]. The solution of the puzzle is unequal probabilities of conformations leading to conformational landscapes supporting fast approach to the global or at least a local minimum. The landscape perspective readily explains the process of reaching a global minimum in free energy and doing so quickly by multiple folding routes on funnel-like energy landscapes [15, 31, 80, 127]. Instead of viewing folding as a process in which all chains perform essentially the same sequence of events to reach the native state, the landscape perspective envisions folding as representing the ensemble average of a process that is microscopically more heterogeneous. Each individual polymere chain may follow its own trajectory, but just like skiers down a mountain, they all may enentually reach the same point at the bottom, the native state. Understanding the folding mechanism is highly relevant for understanding how these molecules carry out their function.

On it's way from the denatured state to a compact conformation, the folding chain follows only instructions encoded on the sequence. During this self-assembly process frequently competing interactions between residues happen until the specific frame of interactions, resembling the native conformation, is formed.

The driving force of protein folding is the formation of a compact hydrophobic core reflecting the preference of hydrophobic groups to be buried inside the protein to minimize solvent exposer. Since the hydrophobic residues are dispersed throughout the primary sequence, it is clear that all hydrophobic residues cannot be satisfied simultaneously. In RNA, for instance, the strong hydrophobic stacking interaction between base pairs promotes the formation of compact structures while the high negatively charged phosphate backbone works against this tendency. Systems exhibiting such behavior are considered to be energetically "frustrated", in a sense that notall favorable interactions can be satisfied simultaneously. According to this conflict between local requirements and global tendencies the free energy landscape of biopolymers is rugged. Several minima exist separated by barriers of various heights. Distribution of minima and barriers over several orders of energies indicates (limited) self-similarity. Assuming however, that natural selection "designed" biopolymers, it is probable that frustration of the energy landscape has been minimized during evolution. For instance free energy bias toward the native conformation [16] could prevent the folding chain from exploring an astronomical number of possible conformations in order to find the native one in reasonable time. In the protein field such a type of landscape is discussed as the "folding funnel".

High thermodynamic stability of RNA double helixes can unfortunately trap the folding molecule in a deep local minimum. For example, the stacking free energy for the formation of a helix of 5 base pairs can easily be around 10 kcal/mol, whereas the thermal energy kT is only 0.6 kcal/mol at a temperature of 300 K. When stacking free energies are large compared to kT it is difficult to open helices once they are formed. Such misfolded structures are believed to play an important role in the kinetics of folding, especially for longer RNA sequences [94]. To a certain extent the RNA folding problem [32, 108] shows a lot of parallels to the much more intensively studied protein folding problem.

1.3 Organization of this work

In the following chapter, the basic concepts of the RNA secondary structure model are introduced. Various commonly used algorithms for RNA secondary structure prediction, based on thermodynamic methods are discussed and applied to explore the conformation space.

Chapter 3 presents a novel and efficient algorithm for the simulation of the folding dynamics of RNA secondary structure. Starting with a brief overview of the state of the art of RNA kinetic folding, the physical model, underlying the algorithm, is developed. Afterwards the crutial components of the algorthm and their computational implementation are discussed in detail.

In chapter 4 the results of various simulations are showes. A Discussion and an outlook in chapter 5 concludes the work.

2 Thermodynamic Folding

2.1 RNA Structure

The structure formation process of RNA can be partitioned conceptually into two consecutive stages. First, the string of bases, called sequence 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 object referred to as the *spatial structure*. The consideration of the secondary structure of RNA as a coarse grained approach to the three dimensional spatial structure is supported by several facts:

- RNA secondary structure formation covers the major part of the free energy of folding.
- As opposed to the protein case, the secondary structure of RNA is well defined and assigns all bases to secondary structure elements.
- The secondary structure provides a scaffold of distance constraints to guide the formation of the tertiary structure.
- RNA secondary structure is conserved in evolution and has been used successfully by biochemists to interpret RNA function and reactivity.

The secondary structure of RNA is formed by aggregation of planar complexes of purine and pyrimidin bases. The geometry of such a complex, or *base pair*, is determined by hydrogen bonds between the two bases. The original set of base pairs, namely the *Watson-Crick* base pairs G≡C and A=U, was soon complemented by a G–U "wobble" base pair, which is admissible within RNA double helices.

Depending on their biological function, naturally occuring RNAs either display long, double helical structures or they are globular, with short double helical domains connected by single stranded streches. RNA double helices display two major, structurally similar conformations, depending on the salt concentration of the solvent. At low ionic strength the A-RNA double helix with 11 base pairs per helix turn predominates. If the salt concentration is raised, A-RNA is transformed into A'-RNA with a 12-fold helix [6]. Both A- and A'-RNA structures exhibit features typical of Watson-Crick base pairs. A typical A-RNA helix is shown in figure 1.

Especially double-stranded structural elements like helices can be very stable. For instance, a RNA duplex of 10 base pairs has a half-time for dissociation of ∼30 min, and G/C-rich duplexes of 10 base pairs have dissociation half-times of up to ∼100 years at a temperature of 300 K [132]. By comparision the most stable protein α-helices dissociate on the sub-microsecond time scale [44].

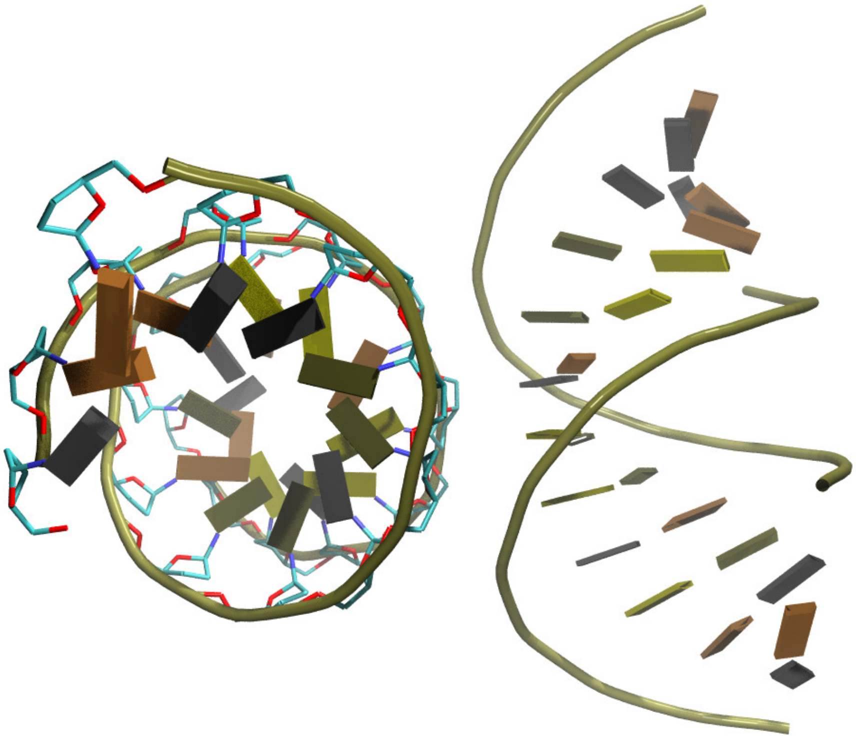

Figure 1: Illustration of the molecular structure of A-RNA. For A-RNA the number of nucleotides per helix turn is 11, the axis rise per residue is 2.73 to 2.81 Å and the base pair tilt 16° to 19°. Views are parallel (l.h.s.) and perpendicular (r.h.s.) to the helix axis.

2.2 Definition and Computation of 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 [137]:

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. Prominent examples of tertiary interactions breaking this condition are base triples [22, 126], G-quartets [1, 7, 74] and A-platforms [18].

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 strech or between two single-stranded streches. Consequently, a pseudoknot can be considered as either a secondary structural element or a tertiary interaction. While pseudoknots are important in some natural RNAs [104, 140], 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 $\Sigma = [\sigma_1 \dots \sigma_n]$ 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 base pairs (s_p, s_q). Following [63, 137] 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 $[\sigma_p \ldots \sigma_q]$. 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 [101]. 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 [62, 59].

Any secondary structures can be uniquely decomposed into loops as shown in figure 2 (note that a stacked base pair may be considered a loop of size zero). A secondary structure graph is equivalent to an orderd rooted tree. An internal node (black) of the tree corresponds to a base pair (two nucleotides), a leaf node (white) corresponds to one unpaired nucleotide. Contiguous base pair stacks translate into "ropes" of internal nodes, and loops appear as bushes of leaves. The tree representation will be of special importance if the implementation of the kinetic folding algorithm is discussed in section 3.5.

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 (see e.g. [40, 69, 134]) 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 [135, 137]. His observation was the starting point for the construction of reliable energy-directed folding algorithms [59, 145].

The first dynamic programming solution was proposed by Ruth Nussinov [100, 101] originally for the "maximum matching" problem of finding the structure with the maximum number of base pairs. Michael Zuker and Patrick Stiegler [145, 146] formulated the algorithm for the minimum energy problem using the now standard energy model. Since then several variations have been developed: Michael Zuker [144] 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 there is no other structure $\mathcal{S}'$ with lower energy containing all base pairs that are present in $\mathcal{S}$. As shown

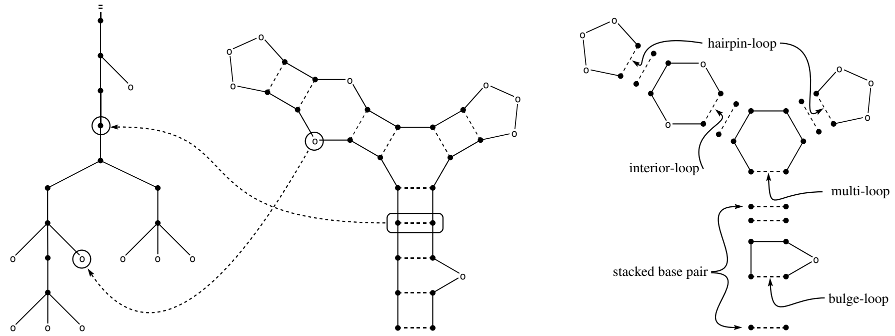

Figure 2: Various representations of RNA secondary structure: The tree representation of the secondary structure graph in the middle (l.h.s); Representation of an RNA secondary structure as a planar graph (middle); The loop decomposition of the secondary structure graph in the middle (r.h.s). The closing 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).

by John McCaskill [88] 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 3 shows such a dot-plot of the Sarkin-Ricin-loop, the longest conserved ribosomal RNA sequence, located in the principal RNA of the large ribosomal subunit [49, 99].

It is the site of attack of two protein toxins, ricin and α-sarcin, that kill cells by inactivating ribosomes. The two toxins recognise the Sarcin-Ricin-loop specificaly and damage it [35, 36]. Once damaged, ribosomes do not bined elongation factors properly [54], and that failure results in the cessation of protein sythesis. The conformation of the Sarcin-Ricin-loop has

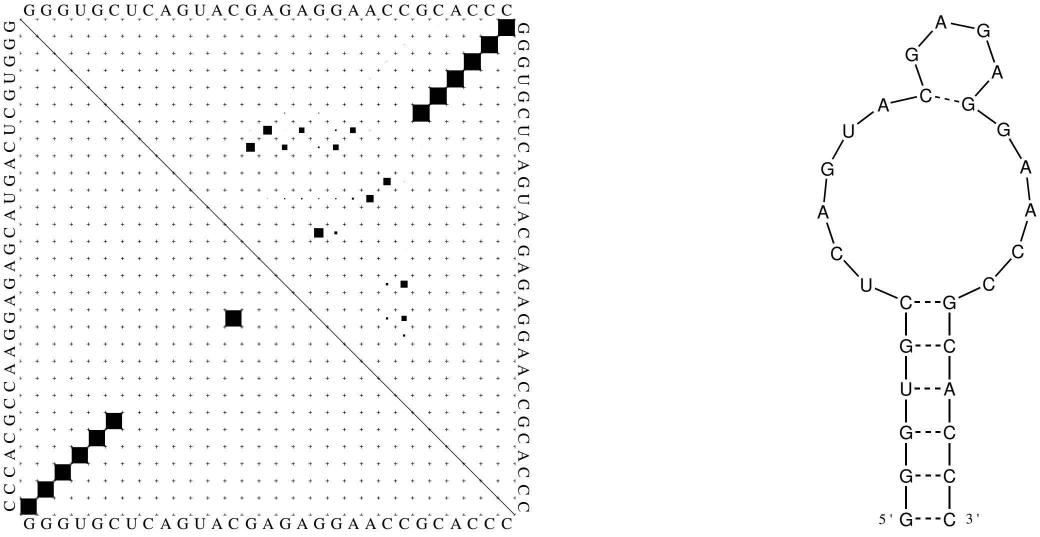

Figure 3: Dot-plot (l.h.s.) and minimum free energy structure (r.h.s.) of the Sarcin-Ricin-loop; The equilibrium frequency p of a base pair (i, j) is represented by a square of area p in position i, j and j, i of the matrix. The lower left triangle shows only base pairs contained in the ground state, which occure with segnificant frequency. The upper right triangle displayes the frequencies within the thermodynamic equilibrium. A large number of base pairs from suboptimal structures are visible.

been determined in solution by NMR-spectroscopy [123]. 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 of the Sarkin loop, 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 optained from the partition funktion. 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 [61, 64]. A for academic use freely available implementation of these algorithms is the `Vienna RNA Package` [59, 60].

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 $\alpha = 1.8488$. 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,

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

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 [27] for a given sequence. In figure 4 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 smal 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. No information is obtained whether the ground state and these "energetically close" states are structurally similar enough to allow a rapid inter-conversion. 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 5 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

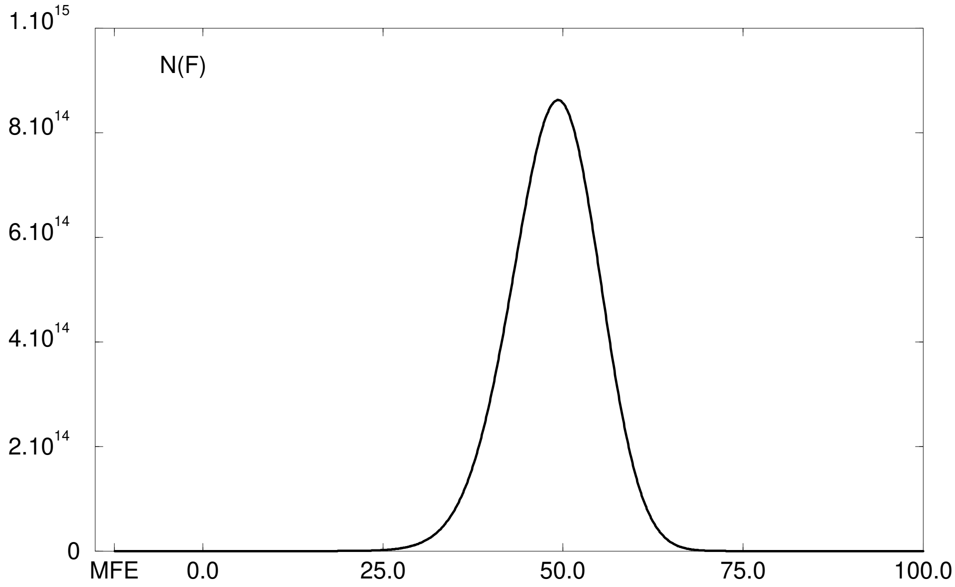

Figure 4: Density of states of the yeast tRNA$^{\text{phe}}$ with an energy resolution of 0.1 kcal/mol. Less than 2 million structures have negative energy, the reference state being the the open structure.

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 5 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 well the kinetics, 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 the sections 3.2 and 3.3.

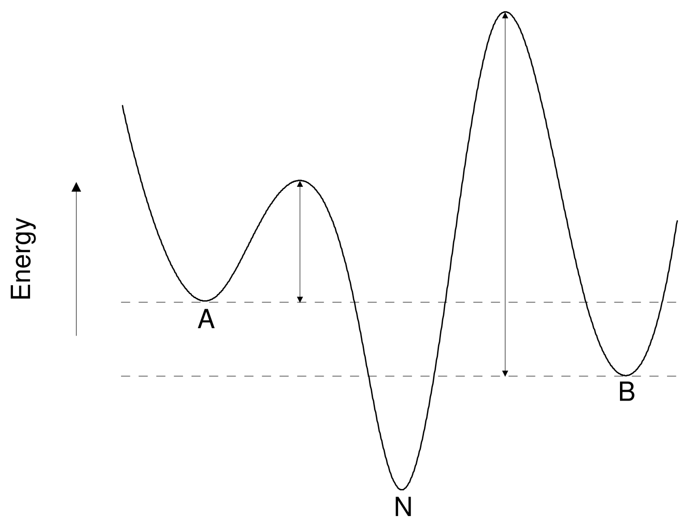

Figure 5: Thermodynamic *versus* kinetic reaction coordinate. State B is *energetically* closer to N (lower energy), but state A is kinetically closer to N (smaller barrier to cross). For didactic reasons a continous reaction coordinate is used as abscissa. In the realm of RNA secondary structures energy 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} [25, 111] 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 [9, 24, 33, 43, 84, 121]. More structural information and insight into RNA catalysis came from the first crystal structure of a hammerhead ribozyme [105] 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 [19].

Recently, kinetic studies [142, 143] 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 occure 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 [46, 47, 117, 124], or the consideration of pseudo-knots [2, 45]. The great majority of these algorithms are based on Monte-Carlo methods [89]. In general these algorithms start from some initial structure (e.g. the open chain) and progress, by incorporation of whole helixes, 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 [71] or the stem with the largest equilibrium constant [86] 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 [14, 90, 92, 122] 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 [5], helix fusion rates are deduced from the formation rates by using a *Boltzmann* distribution hypothesis on the structure space. With the appearance of experimental evidence for the fact that the folding of an RNA molecule takes place simultaneously during transcription [13, 93], various algorithms [91, 117] have been altered in order to consider this "history-based" aspect of RNA folding too.

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*. 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 two 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 [106]". 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.

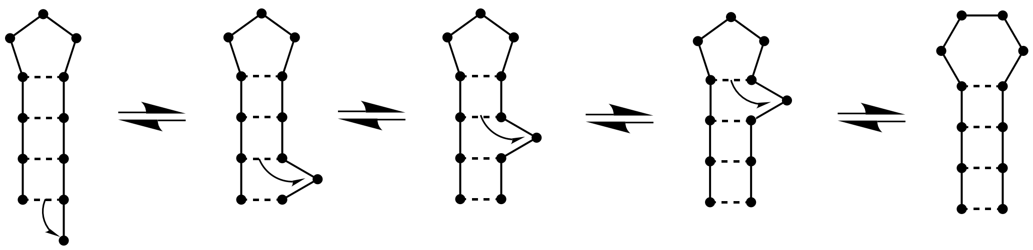

Figure 6: Defect diffusion: The bulge can easily migrate along the helix. For the left to right transformation the shift moves are indicated by arrows

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 [106] 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 6). In the framework of 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 7 illustrates this special

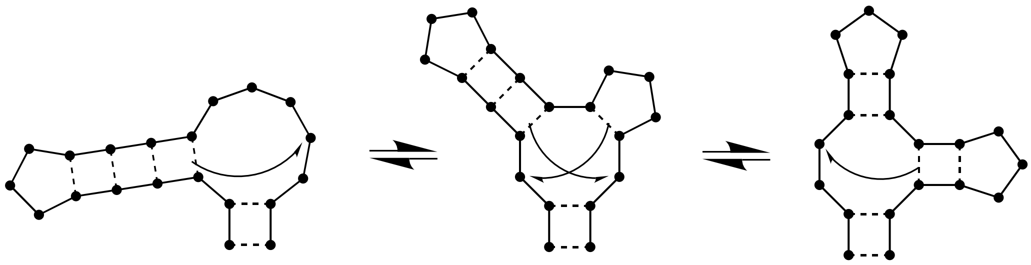

Figure 7: Inter-conversion of overlapping helices is facilitated by shift moves (indicated by arrows).

"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.

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. The degrees of freedom are the transformations provided by the move set. A concept similar to *sequence space*, which was originally invented in information theory [50], can be used for representing the ordering of conformations. There are several examples of applications of the concept of sequence space to problems in biophysics and biology [34, 87]

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 neighbours of each other according to the move set. The vertices of the conformation space corresponding to neighbouring 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 alway a regular one (generalized *hypercube*).

Figure 8 illustrates the conformation space for a short RNA molecule, which can form only 3 base pairs and 8 legal structures. The neighbourhood of any vertex of the conformation space can easily be displayed in two dimensions. The entire conformation space, however, can be displayed only in 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: (1) Under physiological conditions RNA sequences may exist in alternative conformations whose energy difference is small. (2) Aside from their possible biological significance, the density and accessibility of suboptimal conformations may determine how well-defined the ground state conformation actually is. (3) The energy parameters on which the minimum free energy folding algorithms rely are inevitably inaccurate.

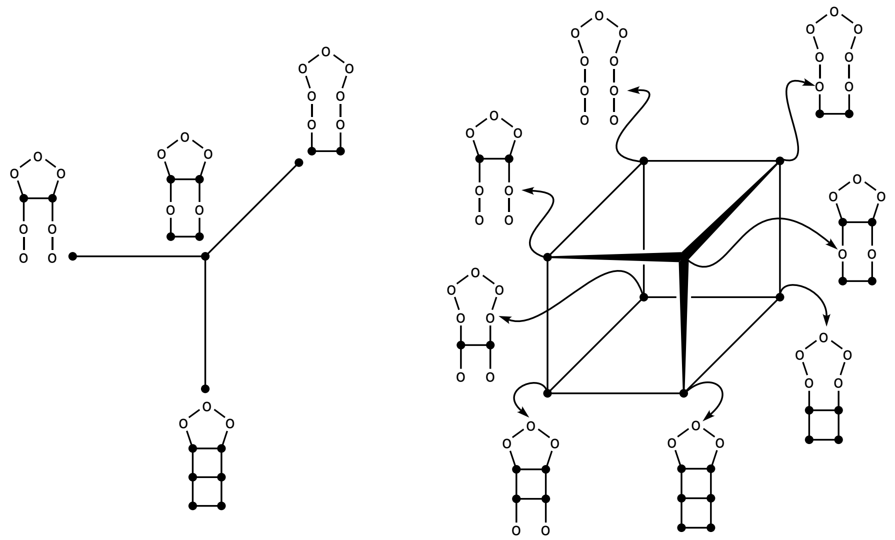

Figure 8: One move neighbourhood of a vertex of the conformation space (l.h.s.) and its embedding in the graph representing the conformation space (r.h.s) for a small RNA molecule which can exhibit 3 base pairs.

Akihiro Nakaya [97, 98] uses a combinatorial approach with a search tree pruning algorithm with dynamic load balancing across the processor elements in a parallel computer. Michael Zuker uses an extension [144] of his widely used dynamic programming procedure for the minimum energy problem [146], which generates for each admissible base pair in a given sequence the energetically best structure containing that base pair. While these approaches represent an improvement for the issues stressed above, they share a common problem: they do not compute **all** suboptimal structures within a given energy range, as needed for the construction of the conformation space.

The program `RNAsubopt` [141], implemented by Stefan Wuchty, does not share this problem, and generates all suboptimal folds of a sequence within a desired energy range. The Waterman-Byers scheme [136] forms the core of `RNAsubopt`. Michael Waterman and Thomas Byers developed their scheme in the context of suboptimal solutions to the shortest path problem in networks, and applied it later to obtain near-optimal sequence alignments. Table 1 lists the most stable structures of a typical RNA sequence of 30 nucleotide length as produced by `RNAsubopt`.

With both features at hand now, 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. Such a closer look at the energy landscape will uncover topological details like *local optima* or *saddle points*.

A structure is a *local minimum* if its energy is lower than the energy of **all** legal neighbouring structures. A structure is called a *local maximum* if its energy is higher than the energies of **all** legal neighbouring structures. Figure 9 illustrate which criteria the neighbourhood of a point of the conformation space must fulfill to be a local optimum.

All configurations that are not local minima or maxima of the energy surface are called *saddle points*. 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. Of course the saddle point with lowest energy that separates the basins of two local minima s and s' is of particular importance. Those saddle points can be found by applying a flooding algorithm to the energy landscape. Figure 10 shows the local minima and their connecting saddle points for a typical RNA sequence with length $n = 30$ as a tree representation. Figure 10 was constructed in such a way, that any two local minima are joined by the saddle point with the lowest energy, connecting the two minima. The ruggedness of the energy landscape is strongly influenced by the *definition* of neighbourhood. In other words the choice of the move set critically forms the topology of the energy landscape. Figure 10 illustrates this strong metric dependency of the energy landscape. With the change of the move set the connectivity of the local optima change 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.

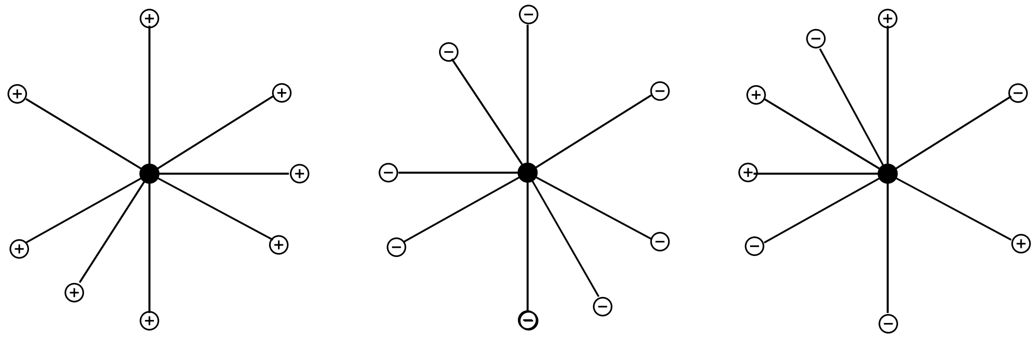

Figure 9: Illustration of the simple neighbourhood of a local minimum (l.h.s), a local maximum (middle) and a saddle point (r.h.s). The signs within the circles denote neighbours with higher (+) or lower (–) energy compared to the structure in the center.

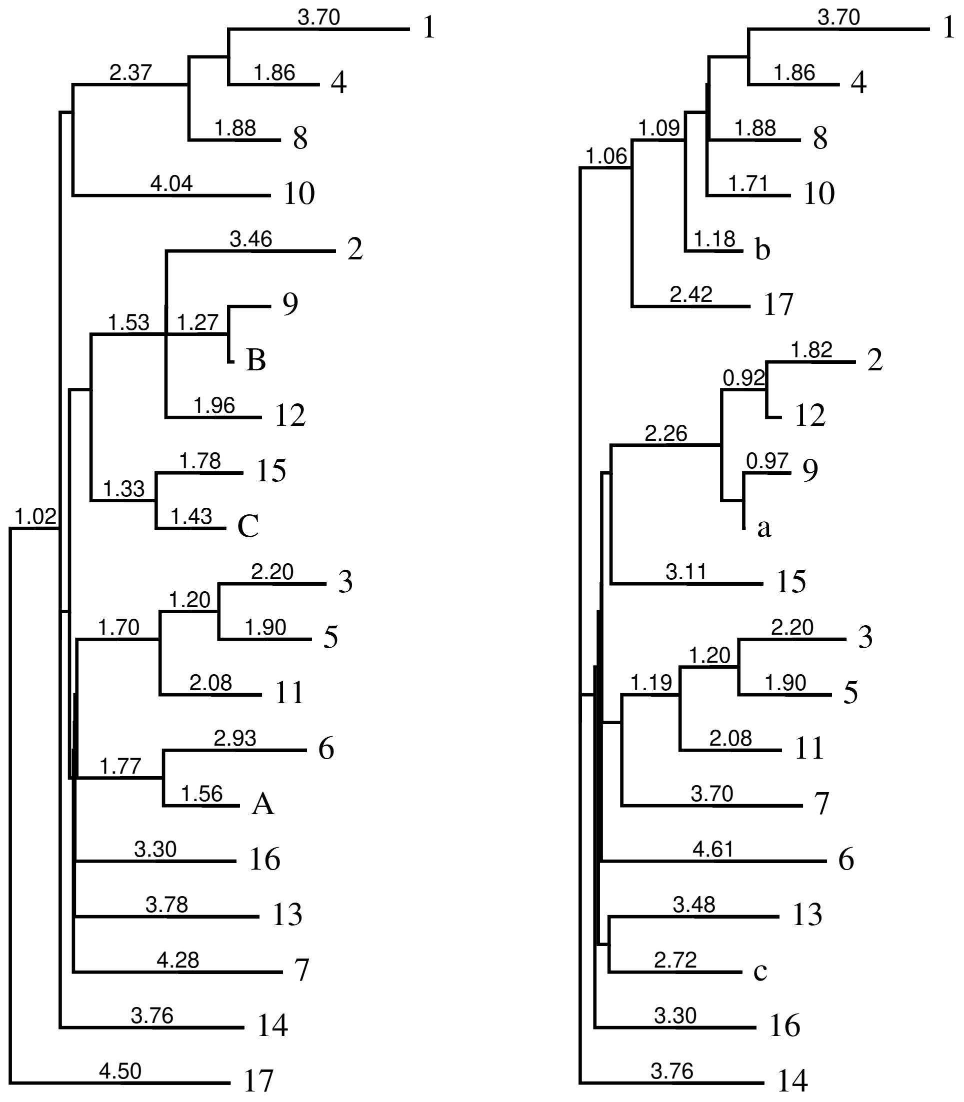

Figure 10: The tree representation of the 20 lowest local minima (leaves) and the saddle points (nodes) in the energy landscape of a typical RNA sequence. The lowest saddle points connecting two local minima are shown for move set MS1 (l.h.s: insertion/deletion) and move set MS2 (r.h.s: insertion/deletion/shift). The local minima are labeled in ascending order starting with the ground state. Equivalent minima are labeled identically in both trees. The length of the whiskers are scaled by the heights of the energy barriers. The barrier heights and the connectivity is strongly influenced by the move set.

It is interesting to compare the local minima with respect to move set MS1 or MS2 to the set of structures that are produced by Michael Zuker's suboptimal folding algorithms [144]. It generates for each admissible base pair in a given sequence the energetically best structure containing that base pair. Hence, for a sequence of length n at most $n(n-1)/2$ suboptimal structures are produced. Furthermore, each base pair present in the ground state regenerates by definition the ground state as the best structure containing it. It follows that no structures are generated with differ from the ground state by the absence of one or more base pairs. In addition, if the ground state consists of two substructures connected by a stretch of unpaired bases, no suboptimal alternatives will be produced that are suboptimal in both modules.

A secondary structure S is *Z-suboptimal* if there is no other secondary structure S' with lower energy containing all base pairs that are present in S. Obviously, the ground state is a local minimum with respect to any move set and it is also Z-suboptimal.

It is surprising to see, however, that a substantial fraction of the low energy structures are not Z-suboptimal. In fact, there are local minima with respect to the move sets MS1 and MS2 that are not Z-suboptimal, and conversely, some Z-suboptimal structures are not local minima, see Table 1.

The data compiled in figure 10 in combination with Table 1 can be used to extract possible folding pathways. In figure 11 the three most favorable pathways leading from the open (denatured) structure to the ground state are displayed. The first saddle point is determined by the nucleation of the first base pair. Adding base pairs to an established stack leads to lower energies. If the correct base pair is formed in the first step, the ground state is found without further obstacles. However, the energy barrier to the correct folding pathway is not the lowest in our example. Most saddle points encountered along the folding pathways contain an isolated base pair, i.e., they correspond to the nucleation of a novel stem. This is consistent with experimental findings on RNA folding. While the nucleation of a helix is a slow, closing additional base pairs is a fast cooperative process [106].

The examples given above illustrate that the folding landscape of an RNA molecule is a very complex hyper surface. As shown the "pathologic points" of the energy landscape are dramatically effected by the choice of the move set. Therefore, the move set, used for the simulation of the folding dynamics, should be a "natural one", if one hopes to capture a realistic folding dynamics. In this sense the move set must encode a set of structural changes a folding molecule can undergo with moderate activation energies. However the "static picture" of the folding landscape only gives a very coarse grained impression of how the folding dynamics itself can look like.

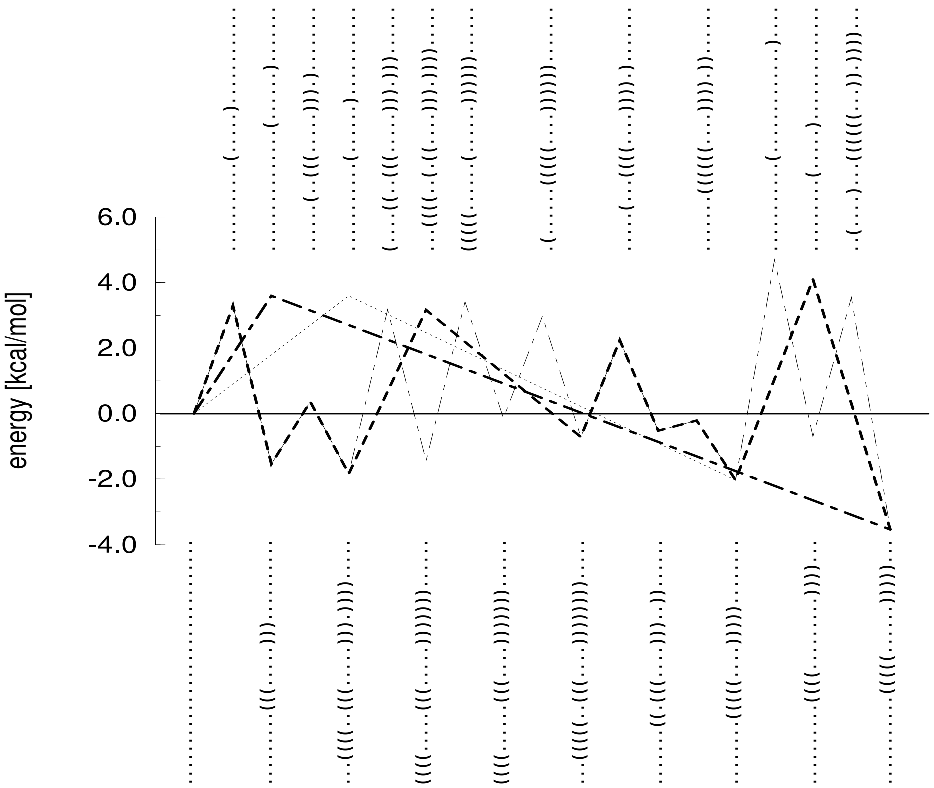

Figure 11: A variety of folding pathways starting with the open (denatured) structure lead to the ground state, among them a pathway with a single saddle point. Alternative foldings, however, have a lower energy barrier. Note that almost all saddle point contain an isolated base pair, i.e., they correspond to the nucleation of a novel stem.

F kcal/mol	Structure ACUAGUCGCGGGGAAUACCUUGGUUCCAAC	local minimum MS1	MS2	Z-sub-optimal
-3.54	...........(((((......)))))...	■	■	■
-2.14	.........(.(((((......)))))..)			■
-2.03	........(((((....))))))........	■	■	■
-1.84	...((((..(((.....))).)))).....	■	■	■
-1.70	.((......))(((((......)))))...	■	■	■
-1.54	.........(((.....)))..........	■	■	■
-1.45	...........((((........))))...			
-1.44	((((.....(((.....)))))))......	■	■	■
-1.43	...(((((.(((.....)))))))).....			■
-1.35	....(((..(((.....))).)))......			
-1.11	......(.(((((....)))))).......			■
-1.03	.....((..(((.....))).)).......			
-0.95	...(((((.((......)).))))).....	■	■	■
-0.94	.......(...(((((......)))))..)			■
-0.94	....((((.(((.....)))))))......			
-0.91	....((.....(((((......))))).))	■	■	■
-0.71	...(((((..(((....)))))))).....	■	■	
-0.70	..........((((.........))))...	■	■	■
-0.86	.........((((....)))).........			
-0.62	.....(((.(((.....)))))).......			
-0.56	............((((......))))....			
-0.52	.......((.(((....)))..))......	■	■	
-0.52	.......(((((.....)))..))......	■	■	■
-0.47	(((((....((......)))))))......	■	■	■
-0.46	....((((.((......)).))))......			
-0.44	..........(((....)))..........			
-0.44	..........((((((......)))))..)			■
-0.35	...........(((..........)))...			
-0.22	....((((..(((....)))))))......			
-0.21	.......((((((....))))).)......			
-0.21	.......((((((....)))).))......			
-0.18	...((((..((......))..)))).....			
-0.16	.......((.(((.....))).))......	■	■	■
-0.14	.(.....)...(((((......)))))...	■	■	■
-0.14	.....(((.((......)).))).......			
-0.13	...........(((((.(...))))))...			■
-0.10	(((......(((.....))).)))......			
-0.07	(((((.....((.....)))))))......	■		
-0.05	.........(.((((........))))..)			

Table 1: Energetically favorable Structures of a small RNA computed by `RNAsubopt`.

3.4 The Model

In the previous section the move set and its effect on the topology of the conformation space has been has discussed. The next important component in the construction of an algorithm capturing the folding dynamics of RNA is to pack the conception about the “real” folding process into a proper physical model.

Imagine a “gedanken experiment” where the folding of an RNA molecule is traced by a camera. Each time the conformation of the folding molecule changes, a snapshot is taken. The resulting series of snapshots allows to follow important properties of the folding process.

First, the resolution of the folding process is determined by the choice of the move set. A move set, which introduces only small changes when applied, yields a much longer and more detailed series of snapshots and hence captures the process in higher resolution.

Second, certain conformational changes happen more frequently along the series of snapshots, some only rarely. Some changes are more likely then other ones. This fact leads to the conclusion that it is necessary to introduce a measure for the likelihood of a transitions between two conformational states.

Third, conformational changes, which are “far” from each other along the series of snapshots seem to happen independently in a sense, that the molecule apparently has no memory of what happened earlier on the trajectory.

The observations from the “gedanken experiment” above can be tied into a model of the following form. The chain moves from state to state in the conformation space governed by a transition probability law. The conformational changes are controlled by the chosen move set. The movement of the chain in the conformation space however seems to be arbitrary on a large time scale.

The model described above is called a *Markov chain.* A Markov chain is a random walk in an N-dimensional state space with a very short memory of only one step. In symbols, if X_n denotes the state of the system at time n, then

$$P\{X_{n+1} = j | X_n = i\} = p_{ij} \tag{5}$$

gives the probability to find the system at time $n+1$ in the state j. Since the transition probabilities do not depend on time the Markov chain is called *homogeneous*.

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. In this sense folding can be viewed as a very complex isomerisation reaction network.

In the stochastic formulation of first order chemical reaction kinetics the probability that a transition from a secondary structure i to a secondary structure j occurs within the infinitesimal time interval dt is given by $k_{ij}dt$, where k_{ij} is the rate constant in the deterministic description [42]. The probability $P(i,t)$ that a given RNA molecule will have the secondary structure i at time t is then given by the master equation

$$\frac{dP(i,t)}{dt} = \sum_j [P(j,t)k_{ji} - P(i,t)k_{ij}]. \tag{6}$$

Manfred Tacker *et al.* [124] have integrated numerically equation 6 on a very restricted subset of conformations, to assess the feasibility of a particular folding pathway. Since we will consider all secondary structures on a given sequence, our reaction network becomes combinatorial in nature. We resort to a numerical simulation of the situation described by the master equation 6. This simulation is based on a continuous time Monte Carlo method proposed by Daniel Gillespie [42]. More precisely, we will not be interested in the equilibrium solution of equation 6, but rather in computing the distribution of first passage times from some initial state to the thermodynamic ground state. In this framework the first passage time represents the folding time.

Let the probability for a transition from i to j in the interval dt be given by $k_{ij}dt$. We next compute the probability density, $p_{ij}(t)dt$, that this transition occurs between time t and $t+dt$. Consider first the probability $p_{ij}^{(0)}(t)$ that no transition to j has happened up to time t. This probability can be obtained by observing that

$$p_{ij}^{(0)}(t+dt) = p_{ij}^{(0)}(t)(1 - k_{ij}dt). \tag{7}$$

That is, the probability that no tranistion happens up to time $t + dt$ is the probability that no tranistion happened up to time t times the probability that none will occur within the following dt, which, by definition, is precisely $1 - k_{ij}dt$. This can be written as

$$[p_{ij}^{(0)}(t+dt) - p_{ij}^{(0)}(t)]/dt = dp_{ij}^{(0)}(t)/dt = -k_{ij}p_{ij}^{(0)}(t) \tag{8}$$

whose solution is $p_{ij}^{(0)}(t) = \exp(-k_{ij}t)$. The $p_{ij}(t)dt$ we are seeking is then simply given by the probability that no transition occurs up to time t times the probability that a transition occurs within the following dt, which yields the exponential density $p_{ij}(t)dt = k_{ij}\exp(-k_{ij}t)dt$.

To simulate the process described by equation 6 we need the probability $P(i \rightarrow j, t)dt$ that a transition occurs from conformation i to j between time t and $t + dt$. This is computed as the product between the conditional probability that a transition occurs to j (from i) given that some transition (from i) happens between t and $t + dt$ times the probability for the latter:

$$P(i \rightarrow j, t) = P(i \rightarrow j|t)P_i(t). \tag{9}$$

Since all reaction channels from i to its neighbouring conformations j are independent with exponential density $p_{ij}(t)dt = k_{ij}\exp(-k_{ij}t)dt$, the overall probability that some transition happens between time t and $t + dt$ is the product of these exponentials: $P_i(t)dt = a_i \exp(-a_i t)dt$, with $a_i = \sum_j k_{ij}$. The conditional probability $P(i \rightarrow j|t)$ is simply given by the relative weigths $P(i \rightarrow j|t) = k_{ij}/a_i$. This yields

$$P(i \rightarrow j, t)dt = k_{ij}\exp(-a_i t)dt. \tag{10}$$

(This can also be derived by observing that $\exp(-a_i t)$ is the probability that no channel fires up to time t, and $k_{ij}dt$ is the probabilty that the particular transition $i \rightarrow j$ fires at the following dt.) Thus, the Monte Carlo simulation of the folding process is simply performed by following the prescription 9. This means operationally to first advance our clock by a random number t distributed according to $P_i(t) = a_i \exp(-a_i t)$ (assuming we are in state i), and then to select a transition to one of i's neighbours j with probability k_{ij}/a_i. The current state is update to j, and the procedure repeats until we first hit the ground state.

What is still needed for completing our model of RNA folding, is a rule for calculating the rate constant k_{ij}, which characterizes the transition from conformation i to conformation j. A standard rule is the *Metropolis rule* [89], originally designed for studying equilibrium properties of matter. It was also applied successfully to kinetic problems like protein folding [125]. 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 $k_{ij}dt$ as given by the Metropolis rule is:

$$k_{ij} = \begin{cases} e^{-\frac{\Delta G}{kT}} & \text{if } G_j > G_i, \\ 1 & \text{if } G_j \leq G_i, \end{cases} \tag{11}$$

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 [75] 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:

$$k_{ij} := e^{-\frac{\Delta G}{2kT}} \tag{12}$$

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 [51]. For a detailed discussion of other possibilities to formulate the transition probabilities p_{ij}, see [21, 58]. 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 neighbouring states.

3.5 The Algorithm and its Implementation

As outlined in the previous section the folding dynamics is simulated by the Gillespie method. It is a variant of the standard Monte Carlo algorithm without rejections.

In standard Monte Carlo, the time spend in a certain state is proportional to the number of trials that have to be made until a new acceptable state is found. Imagine sitting on a pathologic point of the energy landscape for example at the bottom of a deep local minimum, then the rejection rate becomes rather high because all conformations in the neighbourhood posses obviously higher energy. A lot of trials have to be made, before a new conformation is accepted and hence for many steps no progress is made, slowing down the simulation reasonably.

As mentioned in the previous section, the Gillespie method provides an internal clock to measure time. Here, the time spent in a certain state is inversely proportional to the total flux Φ leading away from this state. If Φ is small, as for example at the bottom of a deep local minimum, the internal clock is advanced by a big time increment. For each step, the rate constants from the current state to all its neighbours are computed. Then, the time is advanced by an appropriate time increment adjusted to the sum of the rate constants. Finally, the current state is replaced by a state chosen from the set of neighbours. The consequence of this kind of procedure is progress at each step, because of the lack of waiting times due to rejection, resulting in a very efficient and fast simulation.

It should be mentioned, that the Monte Carlo method has a better performance than the Gillespie method if rejection rates are low. This roots in the fact, that the Gillespie method computes all the neighbours of the current state at each step, whereas the Monte Carlo method computes only one (acceptable) neighbour. However, the folding landscape of RNA molecules is supposed to be a rugged landscape with a lot of deep local minima. This suggests that the over-all behaviour of the Gillespie method should be superior to the Monte Carlo method.

The check for knot-freeness of the generated secondary structures is another factor, which can slow down the performance of the algorithm dramatically in a naive implementation. This problem can be circumvented by implementing the secondary structure as an ordered rooted tree. As explained previously (see section 2.1) each secondary structure can be uniquely decomposed into k-loops. Each k-loop is realized as a linked list in such a way that the list items are ordered corresponding to their 5' to 3' positions along the sequence. (Note that a base pair, a loop of $k = 0$, is a linked list with a single list item). The various k-loops are then linked together to yield an ordered rooted tree (see figure 12). Since a tree is *per definition* knot-free, the check for knots is obsolete if the move set operates on the tree itself.

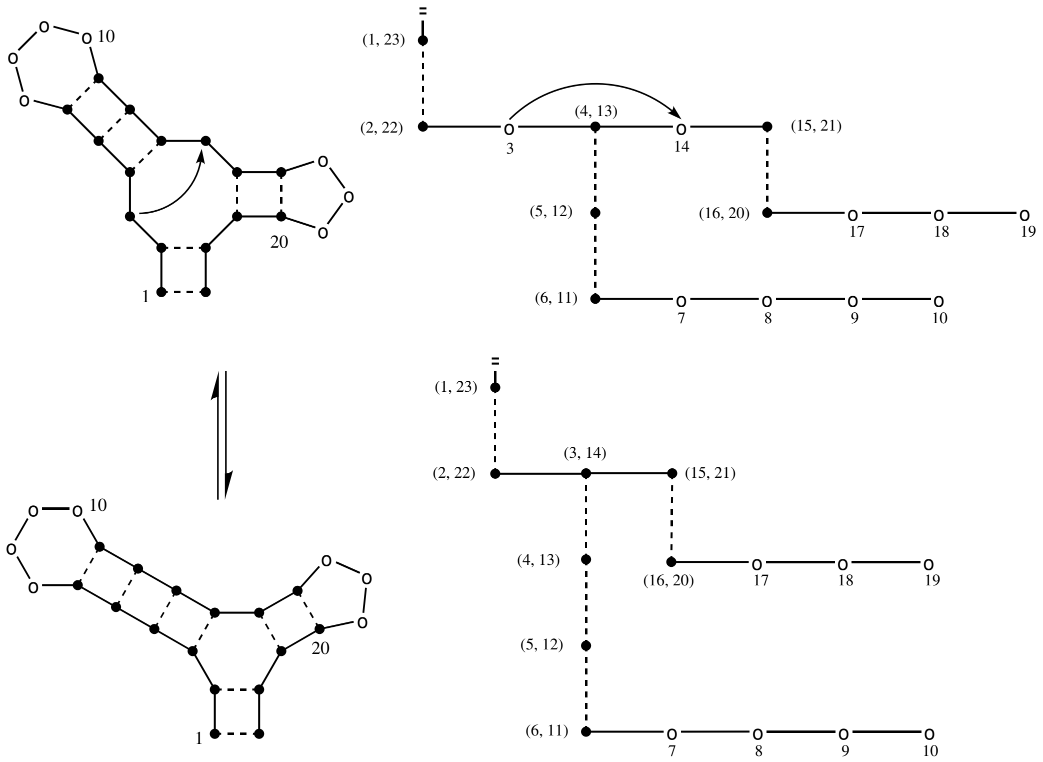

Figure 12: Illustration of the data structure used in `Kinfold` (filled cycles indicate base pairs, open cycles indicate unpaired bases; dashed lines within the tree representation signal links between different k-loops).

The tree can be traversed recursively by visiting the root, then visiting its subtrees in left to right order. During this traversal the various linked lists, the tree is composed of, can be examined for the insertion of possible "new" base pairs, or the deletion of "existing" ones, resulting in a simple reorganization of the tree.

Energy parameters for branched loops, based on experimental measurements are not available. The standard energy model extrapolates these multi loop parameters by the linear ansatz approach,

$$\Delta G = 4.6 \text{ kcal/mol} + 0.4 \text{ kcal/mol} \cdot u + 0.1 \text{ kcal/mol} \cdot m$$
$$(u \text{ ... number of branches, } m \text{ ... loop size})$$

allowing a faster prediction by a dynamic programming algorithm. This linear ansatz results in a high penalty for multi loops of moderate size. However folding simulations of tRNA with the developed program `Kinfold` showed, that a nucleation parameter for the formation of branched loops must be introduced to compensate for the linear ansatz of the standard energy model. To avoid such an "artificial" parameter, the linear size dependence of branched loops was changed to a logarithmic one, which seems more natural, since coaxial stacking of helices within the multi loop increases the stability of these structural motifs [134]. The basic algorithm of `Kinfold` works as follows:

Step 0. **(Initialization).**

(a) Set the time variable $t = 0$ and the "stopping time" t_{stop}.

(b) Specify the move set (MS1 or MS2) and the rule for calculating the channel weights R_i^α (e.g. Kawasaki, Metropolis).

(c) Specify the start structure and initialize the current structure S_{cur} with the start structure.

(d) Specify and store the stop structure S_{stop}.

Step 1. Generate the set of legal neighbour structures $\{S_n\}$ from S_{cur}.

Step 2. Calculate all the reaction channel weights $R_{curr}^{(\alpha)}$ and the total flux $\Phi_{cur} = \sum_\alpha R_{cur}^{(\alpha)}$. Afterwards normalize the $R_{cur}^{(\alpha)}$'s.

Step 3. Draw two random numbers $r_1,\ r_2 \in [0, 1]$ from a uniform number generator.

Step 4. Cumulatively adding the successive values $R_{cur}^{(1)},\ R_{cur}^{(2)},\ ...$ until their sum is observed to equal or exceed r_1. Choose the structure with the index of the last term added to the sum as the new S_{cur}.

Step 5. Calculate the time increment $t_{inc} = \frac{1}{\Phi_{cur}} \cdot \ln\left(\frac{1}{r_2}\right)$ and advance the clock $t = t + t_{inc}$.

Step 6. If $t > t_{stop}$ or if S_{cur} equals S_{stop}, terminate the calculation; otherwise, return to Step 1.

By following the above procedure from time 0 to time t, only one possible realization of the stochastic process is obtained. In order to get a statistically complete picture of the temporal evolution of the folding of a RNA molecule, several independent realizations or "runs" have to be carried out. Each run must start with the same initial conditions and should proceed to the same time t.

A small hairpin has been used to scale the time axis of the folding process simulations of `Kinfold`. This hairpin is formed by the `AAAAAACCCCCCUUUUUU` oligonucleotide and its folding kinetics has been measured experimentally [107]. Since the simulations have not yet been compared with measurements on longer RNA molecules, the times given in the figures of the sections below should only be taken as rough estimates.

`Kinfold` is written in `ANSI C`. Apart from the logarithmic multi-loop energies `Kinfold` uses the standard energy model as implemented in the latest version of the `Vienna RNA Package` [60].

4 Computational Results

4.1 Folding Kinetics of tRNA

As a first application of the algorithm we analysed the folding kinetics of the well known phenylalanine tRNA from yeast. Transfer RNA molecules from most organisms contain several modified bases, particularly methylations. These modified bases occur mostly in unpaired regions and often the modifications are such that base pairing is made impossible. Hence, one might speculate that the modified bases help to stabilize the correct fold.

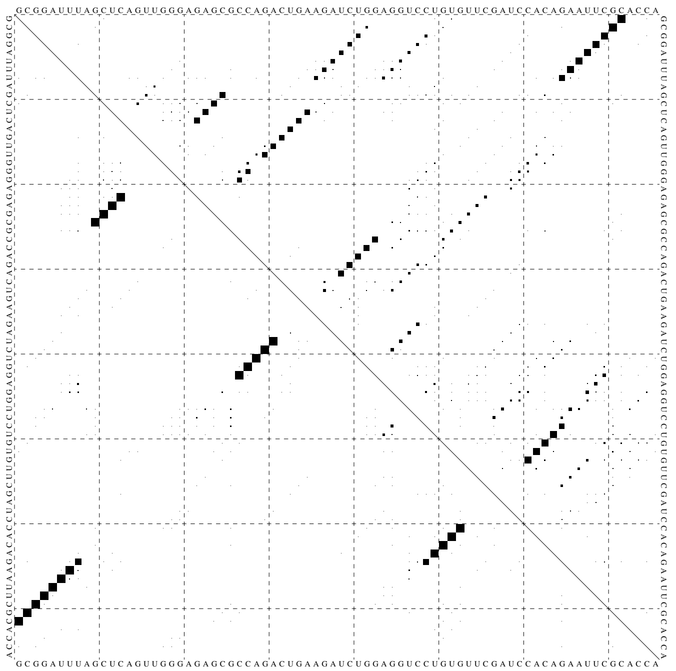

Figure 13: Base pair probabilities for an tRNA^{phe} with and without modified bases. The equilibrium frequency p of a pair $[i,j]$ is represented by a square of area p in position i,j and j,i of the matrix. Lower left: only base pairs contained in the ground state occur with significant frequency for the sequence with modified bases. Upper right: The unmodified sequence displays a large number of base pairs from suboptimal structures, although the ground state remains unchanged.

The phenylalanine tRNA from yeast used in the following, contains six modification in 76 nucleotides which prohibit base pairing. As can be seen in figure 13 the modifications have a strong effect on the equilibrium ensemble of structures. While for the sequence with modified bases only base pairs contained in the ground state occur with significant frequency, the unmodified sequence displays a large number of base pairs from suboptimal structures, although the ground state remains unchanged.

The frequency of the correct fold in the thermodynamic ensemble rises from 4.4% to 28% and suboptimal folding shows that the lowest six suboptimal structures are prohibited by the modifications and consequently the energy gap from the ground state to the next possible structures increases from 0.4 to 0.9 kcal/mol. The density of states and the density of local minima with respect to the move set MS1 for the modified sequence are shown in figure 14.

Local minima are of particular importance for the folding dynamics. All configurations within 15 kcal/mol of the ground state have been checked for local optima using the same move-set as in the folding simulation. The resulting distributions can be seen in the lower part of figure 14. For this plot all structures within 15kcal/mol of the ground state have been generated by

suboptimal folding and tested whether or not they are local minima. The tRNA sequence with modified bases used here displays only a few suboptimal structures within a few kT above the native state. The modified sequence exhibits very few local minima in the low energy region, there are only 10 local minima within 5 kcal/mol of the ground state compared to 173 for the unmodified sequence (not shown). Finding low energy local minima is an example for the analysis of the folding landscape made possible by the new suboptimal folding algorithm, without resorting to complete enumeration of structures [26].

To study the kinetic effect of the modifications, 1000 `Kinfold` simulations of the folding were performed for both modified and unmodified tRNA sequences. The resulting trajectories were then analysed for the existence of typical folding pathways. Data from a representative simulation are shown in figure 15. In this particular run the RNA folds somewhat slower than average, but nevertheless shows features common to all trajectories. A rapid collapse leads to a structure with almost as many base pairs as the native state but only small overlap with it. Folding then proceeds through a series of local minima that have more and more structural elements in common with the ground state. The waiting times in the local minima increase with decreasing energy. Many trajectories visit the same low energy intermedi-

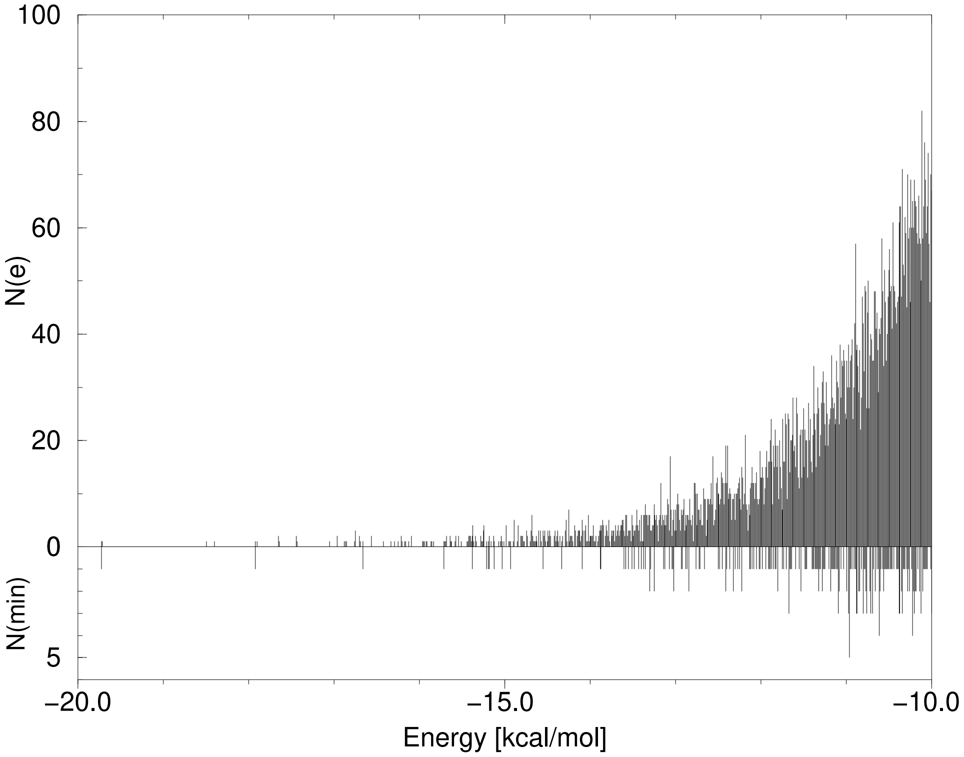

Figure 14: Density of states and the density of local minima for yeast $tRNA^{phe}$ in the region above the native state at higher resolution. For this plot all structures within 15kcal/mol the ground state were generated by suboptimal folding and tested for being local minima. The tRNA sequence with modified bases used here displays only a few suboptimal structures within a few kT above the native state.

ates, in particular, the stem closing the multi-loop forms latest in almost all simulations. Interestingly, the correct hairpins closest to the 5'-end are often formed first, which might support efficient folding during transcription.

As a measure of foldability we recorded the folding times, i.e. the times after which the ground state appears in the simulation for the first time. The resulting distribution can be seen in figure 16. Thick lines show the fraction of simulations that have found the ground state as a function of time. Thin lines show the distribution of folding times, scaled such that the maximum has height one. 1000 simulations were run for each sequence.

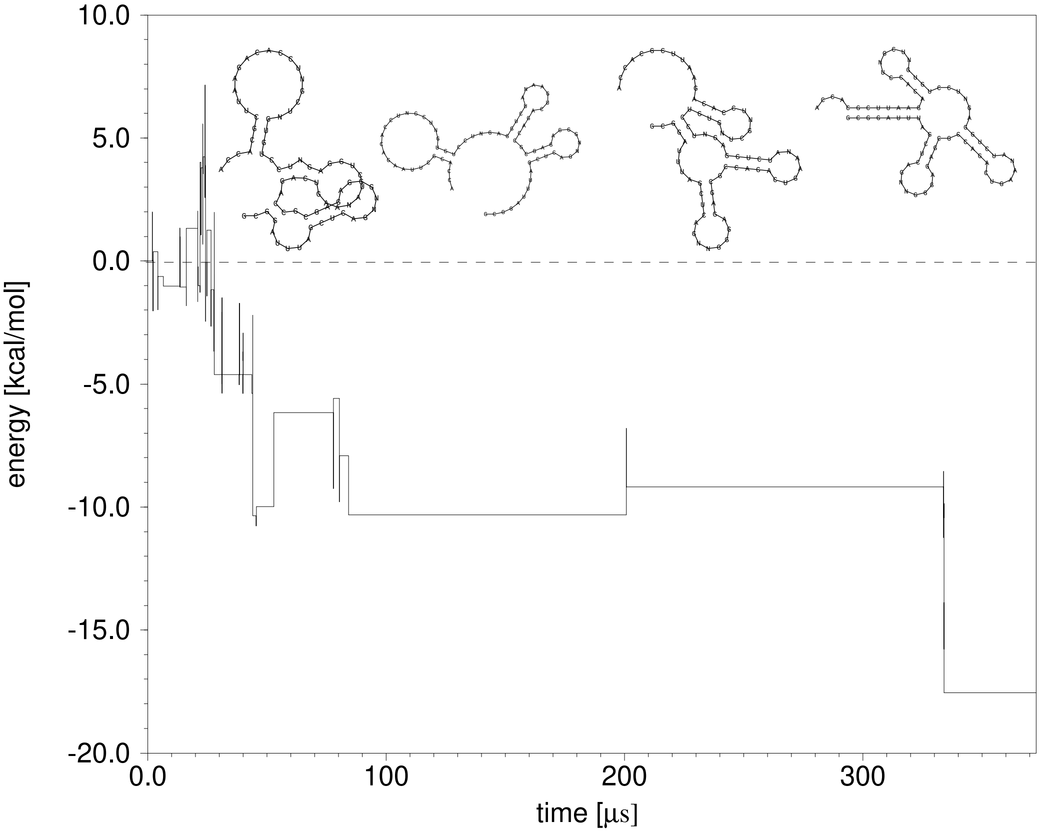

Figure 15: Energy as a function of time for a representative simulation of the modified tRNA^{phe}. A few intermediate structures are shown at the top, the last one being the native cloverleaf structure. The stem closing the multi-loop forms last in most simulations.

The modified sequence folds very efficiently and found the ground state in all of the 1000 simulations. This is consistent with recent analysis of experimental data by Devajaran Thirumalai [127], suggesting a directed pathway to the native state for tRNAs. The unmodified sequence folds much more slowly and only 46% of runs reach the ground state within the simulation time. The fraction of folded sequences is still rising at that point and longer simulation will be needed to decide whether the curve saturates at less than unity.

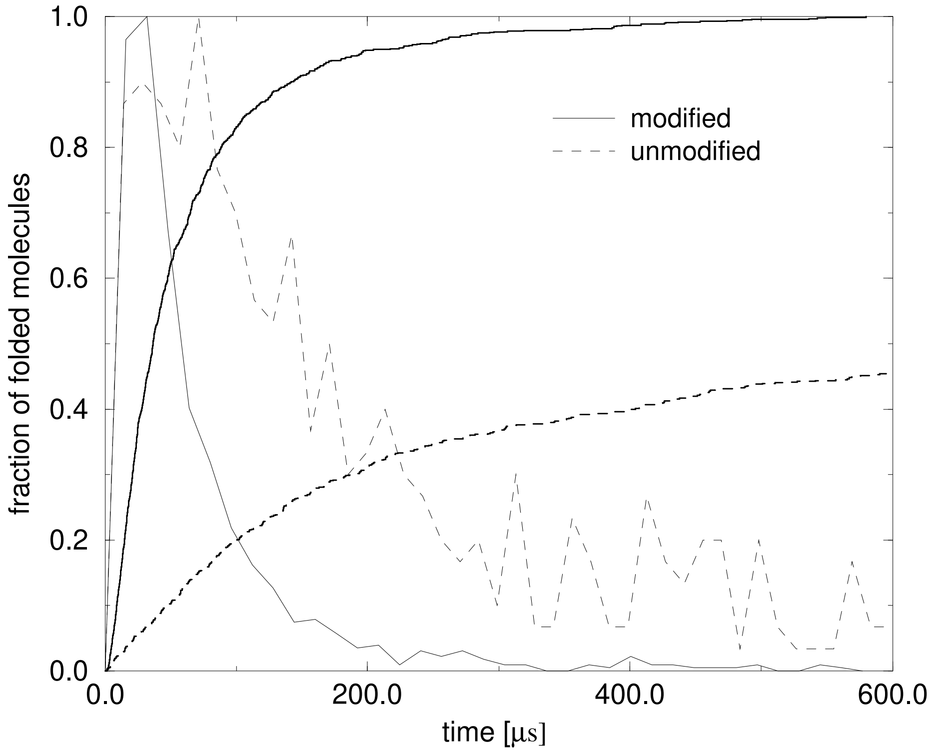

Figure 16: Folding kinetics of modified and unmodified Phenylalanine tRNA^{phe}. Thick lines show the fraction of simulations that have found the ground state as a function of time. Thin lines show the distribution of folding times, scaled such that the maximum has height one. While the modified sequence folds very efficiently, the unmodified sequences do not find the correct fold within the simulation time in over 50% of the cases.

In the tRNA data base six more sequences from *Escherichia coli.* exist, which fold into a “clover leaf” ground state as well. Table 2 shows the percentage of trajectories, which found the ground state. The simulations were performed under the same conditions established for $tRNA^{phe}$. Although all six sequences are modified, which has a stabilizing effect on their thermodynamic stability, their individual folding behavior is rather scattered. Only in two cases $\sim 98\%$ of the trajectories reach the ground state. Unfortunately, the corresponding unmodified sequences fold into an complete different ground state and can therefore not be compared to the runs with the modified sequences. Within the data there is no real trend visible showing that fast folding is a special property of tRNAs. In experiments *in vivo* RNA structures can be very sensitive to the procedures used, and even relatively simple molecules like tRNA can sometimes get trapped in other structures [133].

Sequence	Successful runs
RV1661	99%
RR1661	98%
RV1660	65%
RI1660	53%
RD1669	50%
RI1661	28%

Table 2: Percentage of trajectories which found the ground state for six other modified tRNAs possessing the “clover leaf” as ground state.

4.2 Foldability *versus* Thermodynamic Stability

In case of the phenylalanine tRNA the modified bases improved both thermodynamic stability, conferred by a large energy gap between native and mis-folded states, and foldability. The same relation has been claimed for lattice protein models by *Săli et. al.* [115]. To test this hypothesis two artificial sequences with the tRNA structure as ground state have been computed by means of the `RNAinverse` program from the `Vienna RNA Package`.

The thermodynamics of the first testsequence are average, the frequency of the ground state in the ensemble is about 7% and several alternative foldings can be seen in the base pair probability matrix, see inset of figure 17. The other testsequence had been designed to be especially stable. For this sequence the ground state dominates the ensemble with a frequency of 96% and no alternative foldings are recognisable in the dot plot.

1000 folding *Kinfold* simulations for each sequence have been made the results of which can be seen in figure 17. Surprisingly, it is the thermodynamically more stable sequence that folds poorly in this example. While $\sim$98% of the trajectories of the randomly chosen sequence reach the ground state, this fraction drops for the optimized sequence to $\sim$50%. Note, allthough the the randomly chosen sequence is a relatively good folder, the distribution of the folding times is much broader then for tRNA^{phe} (see figure 16).

Even an isolated example such as this one shows that it is easy to construct cases where the kinetics cannot be predicted from thermodynamic properties. More test cases will be needed in order to decide if and how strongly thermodynamic stability and foldability correlate on average.

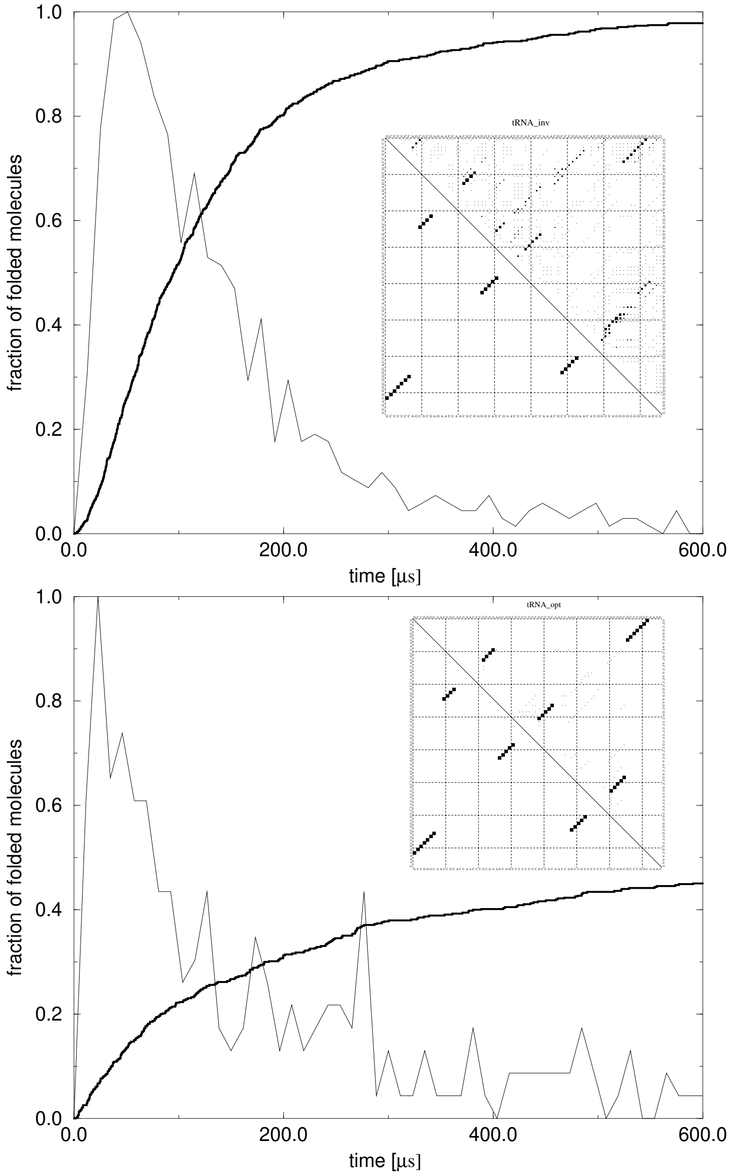

Figure 17: Thermodynamic stability and "foldability". A randomly chosen sequence with tRNA structure (Top). A sequence designed to be thermodynamically extra stable (Bottom). Although many alternative foldings are visible in the dot plot (see Insert) of the randomly chosen sequence it foldes efficiently to the ground state. In contrast the sequence designed to be thermodynamically extra stable (see Inset) folds only in less than 50% of the simulations.

4.3 Folding Paths

Another topic, being animatedly discussed with in the protein field, concerns the question whether or not the folding landscape of a biopolymer is structured in such a way that pronounced folding paths do exist. A folding pathway embodies the idea that the folding molecule goes through a sort of funnel on the folding landscape, like water flowing down a gutter, to the native structure. This process is more directed than a random search. According to this idea, a pathway of folding means that there exists a well-defined sequence of events which follow one another [81]. The gutter represents a particular series of conformational changes. It may have valleys representing intermediate states and hills exhibiting transition states on its way to the native state. While the pathway idea handily " solves" the random search problem of the biopolymer, the physical basis for such specific sequences of events is unclear. A pathway leads from a specific point (e.g. the denatured state) on the folding landscape to another one (e.g. the native state).

The problem with the concept of a pathway is, that the denatured state is not a single point on the landscape, but rather all the points on the landscape except for the ground state. A pathway is too limited an idea to explain the flow from everywhere else, the denatured ensemble of structures, to one point, the native conformation. To throw light on the question of folding pathways for RNA molecules, the frequency with which individual trajectories visit the same local minima along their way to the native state, has been investigated.

Figure 18 shows a compilation of the data for a couple of 1000 folding simulations, of tRNA^{phe}. In the energy range between -9.0 kcal/mol and 0.0 kcal/mol the same local minima are visited with a very low frequency. This finding suggests, that in this region of the folding landscape, a lot of equally efficient but distinct "folding paths" exist. The folding chain has no apparent preference for a special path, but rather proceeds down the folding landscape by many different routes. The dynamic behavior of the folding chain in this energy range of the folding landscape, is the observed and previously mentioned rapid collapse of the chain to a "compact" conformations possessing little overlap to the native structure.

Within an energy of about -12.0 kcal/mol a whole bunch of local minima is visited on average by $\sim 25\%$ of the trajectories. Here the various folding trajectories, which arrived through different routes, must pass a kind of bottle-neck to proceed further towards the native conformation. The bottle-neck seems to be a kind of transition state ensemble with a relatively broad distribution of structural variety. The bottle-neck can be reached through various energetically not so favorable conformational rearrangements from the bunch of local minima. The activation energies are moderate, since all

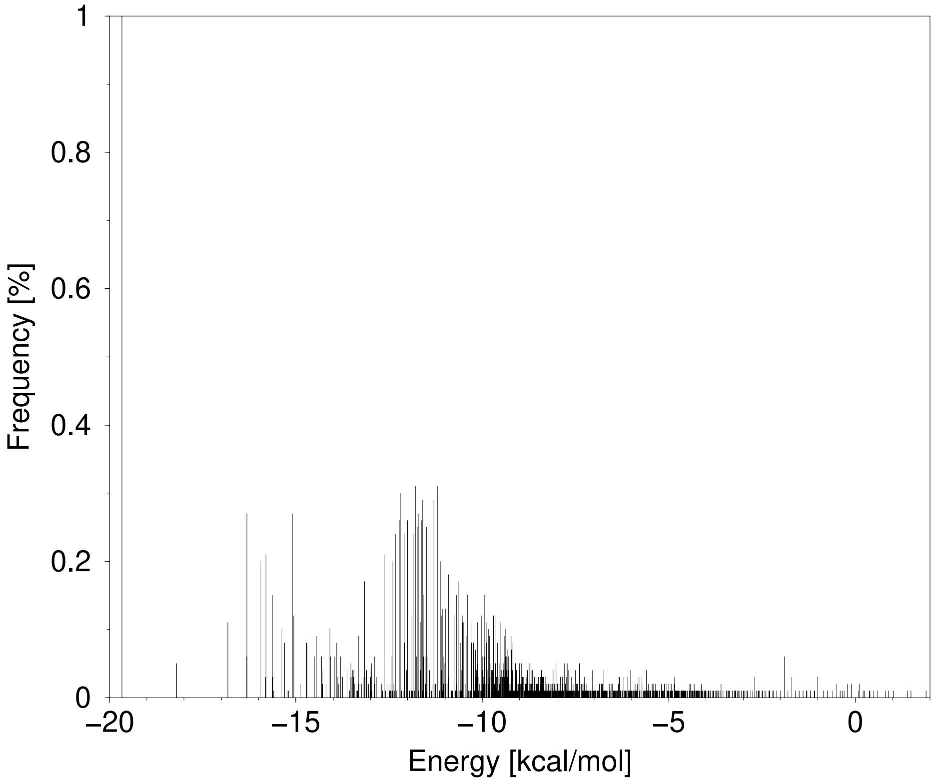

Figure 18: Distribution of local minima visited by the folding modified tRNA^{phe}. The folding paths are very heterogeneous, until the trajectories visit conformations with energie below $\sim$ –9kcal/mol. In that region about $\sim$30% of the trajectories visit the same local minima.

the trajectories reach the ground state and no fraction of trapped chains is observable within the allowed simulation time. The overall dynamic behavior of the folding chain can be summarized as a diffusive-like random search for the “escape pathways” to the native state.

To verify this folding behavior the distribution of the first passage times, the time the chain reaches for the first time the ground state, was reexamined. Figure 19 shows this distribution. Here the fraction of folding trajectories, which reach the ground state within the time interval $[t, t + \Delta t]$ multiplied by t is plotted *versus* the first passage time in a log-log-plot. The plot

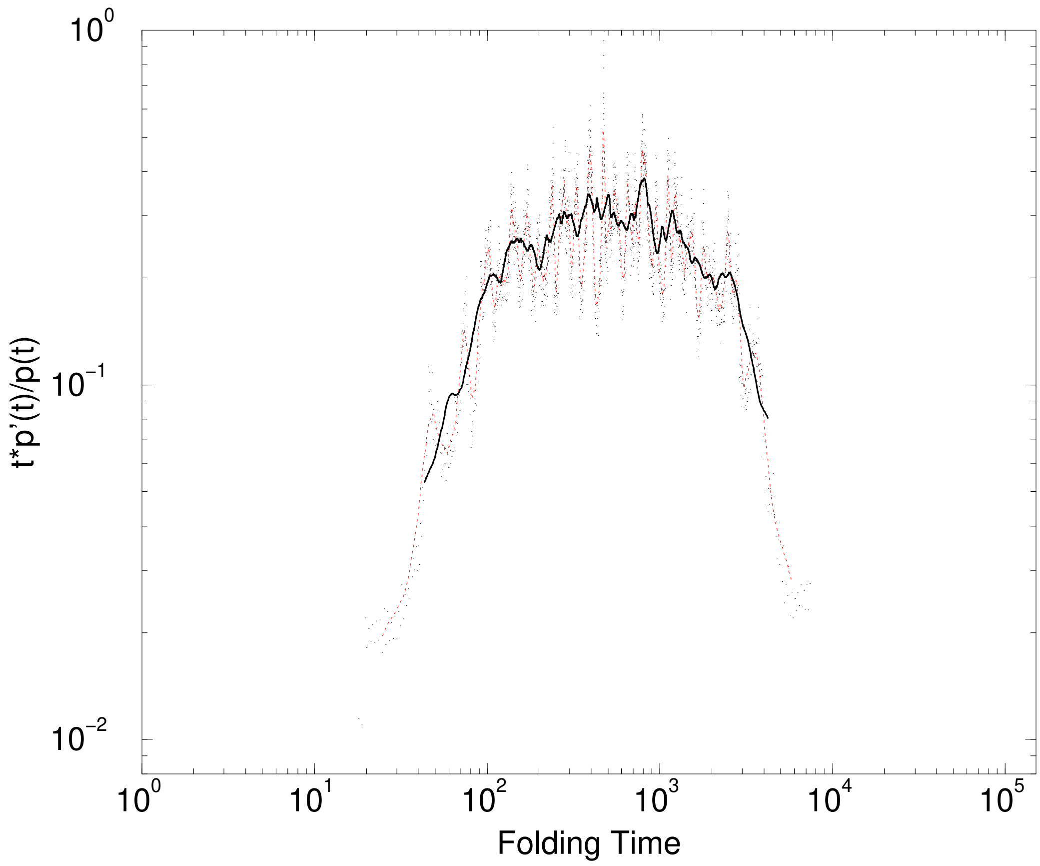

Figure 19: The kinetic signature of the modified tRNA^{phe} shows only a single peak. The time scale of folding is set by the closing of the multiloop.

shows one broad hump with nearly no fine-structure. This indicates, that the folding mechanisms of tRNA^{phe} form a quasi-continuous spectrum. The first passage times of this spectrum are nearly equal distributed within a broad time interval.

A complete different folding behavior is shown in Figure 20. Here an artificial RNA of 25 nucleotide length has been folded. The sequence was designed in such a way, that it can form either of two overlapping hairpins. One located against the 5' end of the sequence with an extra stable tetra loop. The other hairpin located near the 3' end of the sequence resembles

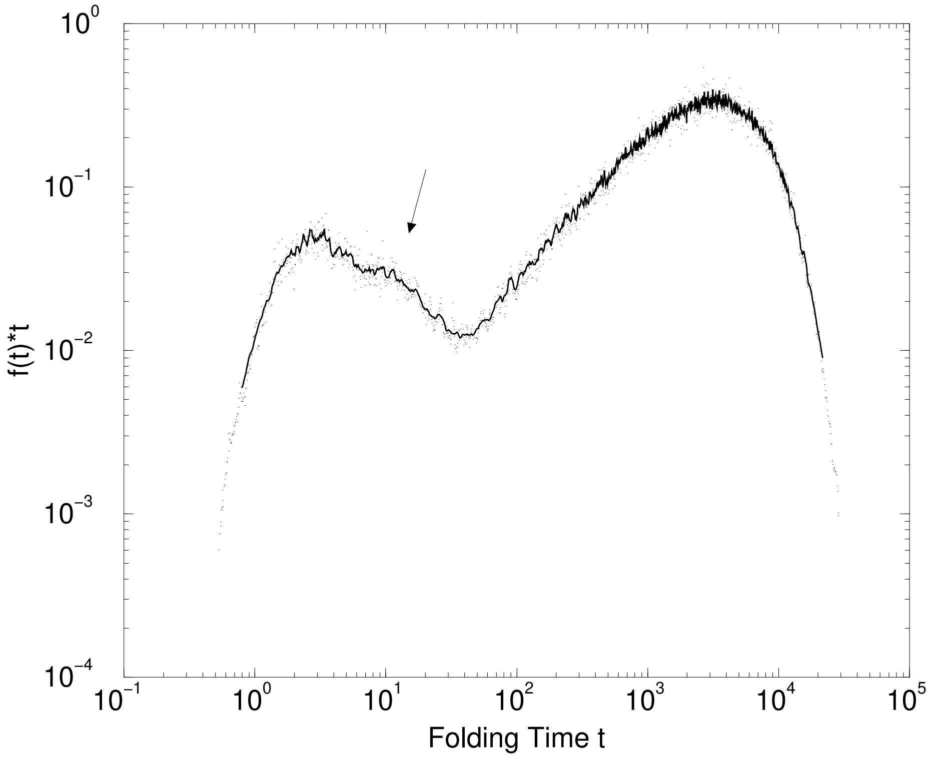

Figure 20: Distribution of folding times of a short RNA. The curve shows two distinct peaks corresponding to two different dominating folding pathsways. A less prominent folding pathway manifests itself as shoulder on the right hand side of the first peak (indicated by an arrow).

the ground state. Figure 20 shows to distinct humps. The left hump corresponds to the direct folding path from the open chain to the ground state (see left insert of figure 20). After insertion of the loop closing base pair, which is the rate limiting step, the ground state is reached without further obstacles by a smooth “zipper”. The energy profile for this folding path is shown in figure 22. In the right flank of this hump a shoulder is visible marked by an arrow. This shoulder depicts a slightly more complicated folding mechanism. Here the chain first misfolds into a set of local minima with slightly negative energies (see figure 21 bold lines). The escape from these local minima back to the unfolded state is fast, since the barriers on the way to the open

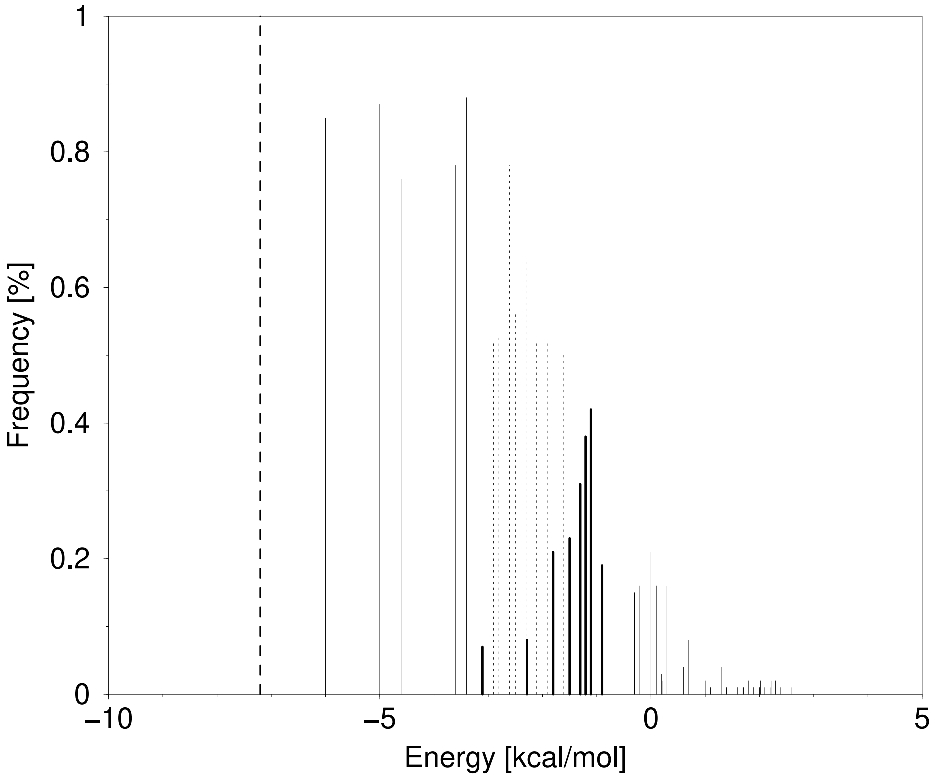

Figure 21: Distribution of local minima for different folding mechanisms (see text for details).

chain are relatively low. The chain then takes the direct folding path to the ground state. However the round about way to misfolded states results in an elongation of the folding time.

The hump to the right portraits pathways leading to local minima with energies near the ground state. Escape from these minima involves extensive “hill climbing” over high free energy barriers and therefore the folding times are resonably elongated. Figure 22 shows the escape pathway from the deepest local minimum. Of all the paths connecting this local minimum to the ground state, the path involving the smallest barrieres is shown. It is easy to see, that the escape pathway from this local minimum is energetically very expensive.

First, a misfolded hairpin at the 3’-end has to be molten. Second, the activation energy for the nucleation of the base pair, closing the hairpin loop of the ground state has to be overcome. Third, the overlapping extra stable hairpin at the 5’-end, has to be disrupted, involving again “hill climbing”, before the ground state can be formed (see figure 22). The energy profile of this last step is flattened by the shift move. The alternative path without base shifts, indicated by a dotted line in figure 22, showes two more free energy barriers.

Within the left flank of the second hump another folding mechanism is hidden. Here the chain forms parts of the two above mentioned hairpins. Before the helix of the ground state structure can be formed, the helix of the extra stable hairpin near the 5’ end must be melted. This is again an energetically expensive process.

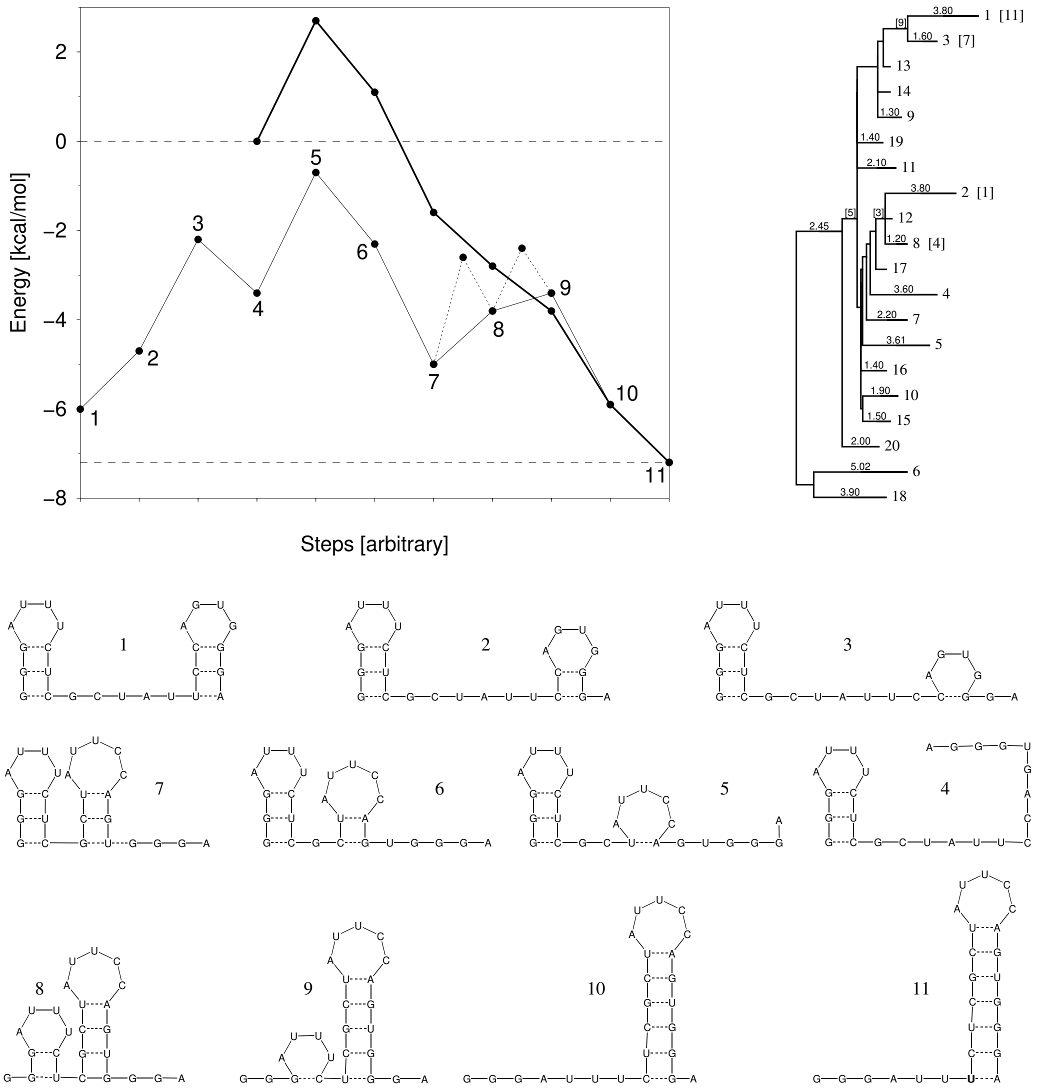

Figure 22: The upper l.h.s plot showes the energy profile of the two most prominent folding paths: The fast path a simple zipper (bold line) and the escape path from the folding trap (normal line); the dashed line indicate the energy barriers in the absence of shift moves. The upper r.h.s plot shows the energy barriers between the 20 lowest local minima. In the lower part the structures, which are realized along the escape path are shown.

4.4 Metastable Structures

Since biopolymers in living cells operate at temperatures far from 0 K, energy levels above the ground state are populated to a certain extend. Therefore the "real structure" of an RNA molecule fluctuates around the ground state structure. This ensemble of alternative foldings must be strictly distinguished from metastable conformations, because the later are **not** in thermodynamic equilibrium with the ground state conformation. Metastable conformations of RNA are conformations which are stable for a limited time span under certain environmental conditions.

In some cases metastable conformations can be essential for cellular function of RNA as has been shown for the live cycle of viroids [56, 83, 109] or the processing of mRNA [113]. They can be identified experimentally by using temperature-gradient gel electrophoresis [110, 112, 114]. Another RNA which is known to fold into a metastable structure is SV-11.

SV-11 is an RNA species of 115 nt length. The molecule is able to fold into two alternative structure, a long hairpin resembling the ground state and an hairpin-hairpin-multi-loop motif resembling the metastable form (see

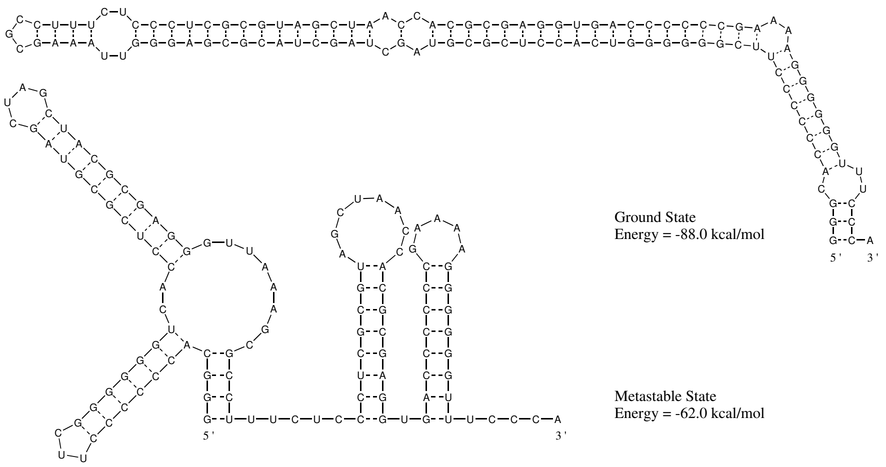

Figure 23: Ground state and metastable state of SV-11. While the metastable state is a template for the Qβ replicase, the ground state is not.

figure 23). While the less stable structure is an active template for the Qβ replication assay, the ground state of the molecule is unable to replicate. The transition from the metastable structure to the ground state is rather slow but has been observed experimentally [12]. This suggests that the lifetime of the metastable folding could be rather long.

SV-11 provides an excellent test case for `Kinfold` because its results can be compared with experimental results as well as with results achieved by kinetic folding simulations of other groups [46, 94]. The simulations with `Kinfold` show, starting from the open chain, only about 16% of the trajecto-

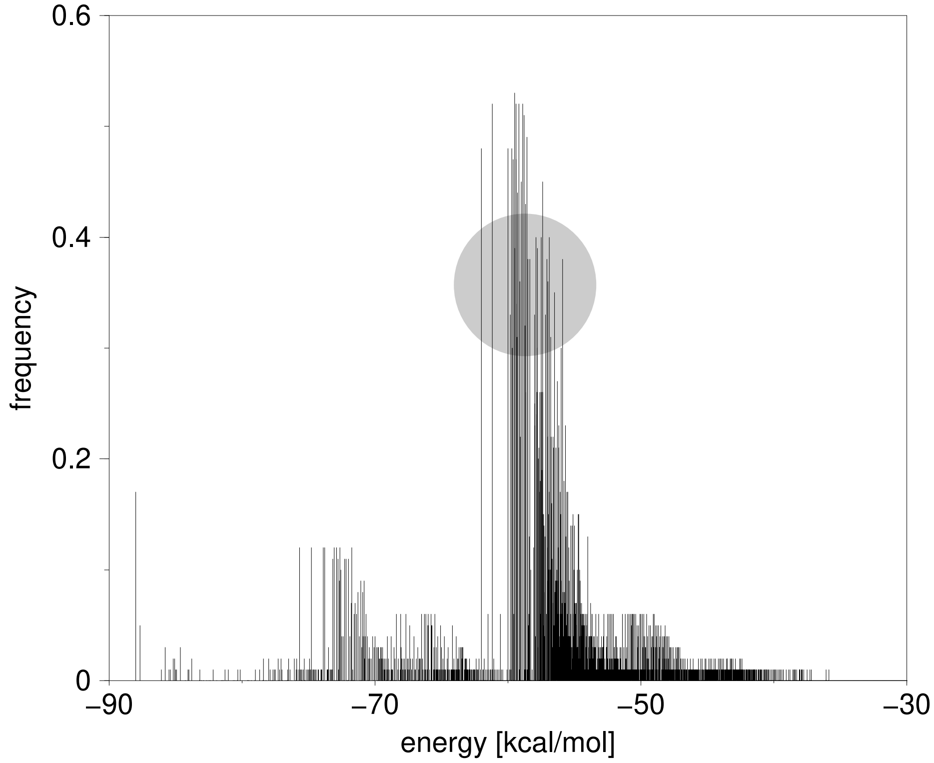

Figure 24: Fraction of local minima visited per trajectory. The metastable structure ensemble is marked by a filled cycle. Only about $sim16\%$ of the trajectories reach the ground state. The great majotity get trapped within the basin of the metastable structure (Dawn at Shiprock, New Mexico).

ries find the ground state within 500μs. The great majority of the trajectories get trapped in a cluster of states around -58 kcal/mol marked by a circle in figure 24. A closer look at these structures shows that they form a thermodynamically equilibrated ensemble around the proposed metastable structure of SV-11.

The simulation had been counter checked by starting from the metastable structure. In this case the trajectories remained in an energy interval around -58 kcal/mol, and reproduced exactly the distribution of the metastable structure ensemble. In the energy range between 0.0 kcal/mol and -45.0 kcal/mol different trajectories visit rarely the same local minima, witch means that in this region a lot of different and independent folding pathways seem to exist. Translated into the language of folding landscapes this behaviour corresponds to a very smooth funnel.

Steven Morgan and Paul Higgs [95] employed a *Monte Carlo* routine for addition and removal of whole helices to simulate the folding of SV-11. The molecule folded to the metastable structure, if folding occured during chain growth. Folding of SV-11 after chain growth was completed lead exclusively to the ground state structure.

Alexander Gultyaev *et al.* [46] used a genetic algorithm operating as well on a move set of whole helix insertion or deletion. His findings however critically depended upon intrinsic parameters of the genetic algorithm, like population size or chain growth rate. Starting from the full-length chain the genetic algorithm failed to predict the metastable structure.

In figure 25 the base pair probability matrix for the ground state and the metastable structure ensemble are shown. The frequency of occurrence of the ground state within the thermodynamic equilibrium is $\sim$ 30%, indicating that the native state is thermodynamically well defined. In contrast the metastable structure occurs only with a frequency of $\sim$ 0.07% and is therefore not visible in the dot-plot of the thermodynamic equilibrium.

SV-11 is a clear example where the structure prediction using thermodynamic prediction methods fails. Since the ensemble of metastable structures is invisible at thermodynamic equilibrium. In such cases only kinetic folding algorithms or, if enough homologous sequences are available, phylogenetic comparison can predict the structure with high accuracy.

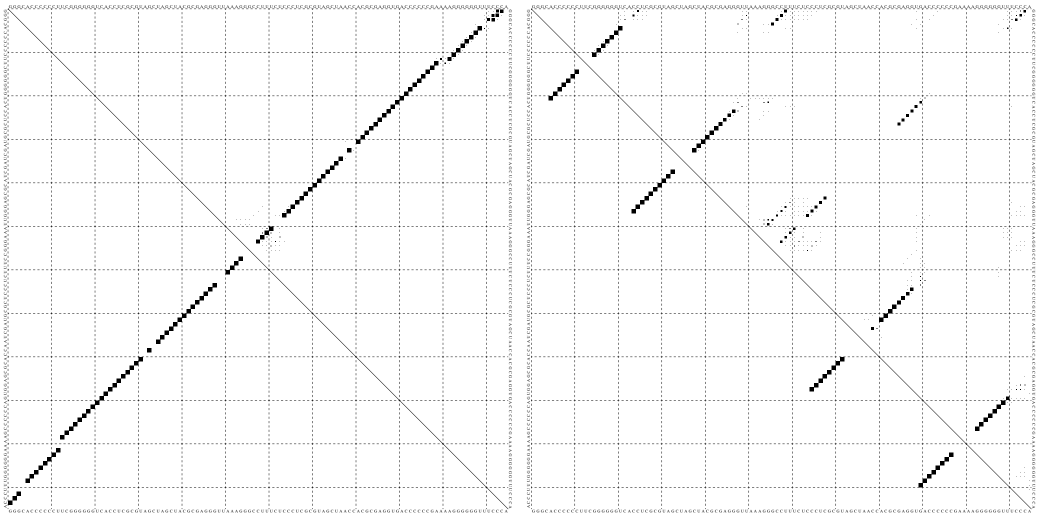

Figure 25: Dot-plot of the ground state (left) and the metastable structure ensemble (right) of SV-11. Note that the metastable structure ensemble posses such a low frequency within the thermodynamic equilibrium, that it is invisible in the dot-plot of the optimal fold.

5 Conclusion and Outlook

The ability of biopolymers to fold into a well-defined native state is a prerequisite for biologically functional biopolymers. RNA secondary structures provide an ideal model system to study the structure formation process. 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.

During this thesis an efficient program called `Kinfold` has been developed for the simulation of the folding dynamics of RNA secondary structure . Due to pairing rules and high stability of RNA secondary structure, the folding landscape of RNA molecules is the prototype of a rugged landscape. Therefore a rejection free Monte Carlo method is used by the algorithm, which is especially capable for the sampling on rugged landscapes.

A crucial component for the simulation of the RNA folding kinetics is the choice of the move set for inter-converting secondary structures. The move set lays down the topology of the folding landscape by defining which secondary structures are neighbours of each other. It also encodes the set of structural changes that RNAs can undergo at moderate activation energies. The move set is so to speak the basis of any kinetic algorithm for RNA folding.

The most elementary move set, on the level of secondary structure, consists of removal and insertion of a single base pair (while making sure that no knots or pseudo-knots are introduced into the structure). Beside this simple move set `Kinfold` makes use of an additional base pair “shift” move (in which one of the two positions of a given base pair is converted into a new one). These “shift” moves facilitate sliding of the two strands of a helix, bulge diffusion along a helix and the inter-conversion of partially overlapping helices, which are assumed to be important effects in the dynamics of RNA molecules. The move sets used in `Kinfold` are local move sets, allowing only small structural changes, which is of utmost importance if one hopes to observe “realistic” folding trajectories.

Folding simulations have been performed for various tRNAs. These simulations revealed cases where the folding molecule found efficiently the native state, and cases where a large fraction of folding molecules got trapped in local minima, from which they could not escape on a realistic time-scale. A closer examination of the folding trajectories of tRNA^{phe}, a "good" folding molecule, showed, that the folding process is hierarchically organized. This implies the early formation of small-scale secondary structure elements which progressively reorganize to larger sub-domains during the folding process. A slight tendency for a interdependence between sub-domain formation was observed as well. Secondary structure elements near the 5'-end of tRNA^{phe} formed in nearly all trajectories before others located farer to the 3'-end, which might support efficient folding during transcription. This might well be a result of evolutionary selection of sequences that fold better when produced in 5' to 3' direction.

Observations based on lattice protein models lead to the hypothesis, that high thermodynamic stability of proteins correlate with their good foldability. To test this hypothesis for RNA the effect of the modification of tRNAs onto their thermodynamic stability and their foldability has been investigated. The correlation found was not as distinct as in the protein case. In addition artificial RNAs with a completely different behaviour can be easily designed. However, in order to decide how strong thermodynamic stability and foldability correlate on average further investigations are required.

Information about folding paths can be inferred directly from `Kinfold` simulations. In particular, folding paths through metastable states can easily be identified in curves of the fraction $p(t)$ of molecules that have reached the the ground state at time t *versus* t. Such $p(t)$-curves are in principle experimentally measurable and provide an excellent check for the theoretical predictions of `Kinfold`. The analysis of $p(t)$-curves uncovered that some molecules have folding paths with very different time-scales. In general, this behaviour is determined by local minima with large basins of attraction on the folding landscape. For small examples it is possible to observe the escape from these basins within the simulation time, and analyse exactly the sophisticated pathway, that allows the molecule to escape from the trapped conformation.

The program `Kinfold` is organized in a modular fashion, making extensions of the move set or changes of the simulation method easy to handle. A conceivable extension would be the incooperation of 3D-contacts like pseudoknots, if experimental measured energy parameters for this kind of contacts become available in the future.

Several interesting evolutionary questions can be investigated using the kinetic folding algorithm. Examples are: How does the energy landscape of an RNA molecule evolve when selection pressure is put on shape conservation and foldability simultaneously? Are the heights of the energy barriers or the density of paths, connecting two points of the energy landscape targeted by the point mutations during such a process? The investigation of the sequence to structure map revealed that sequence space is percolated by extended neutral nets. Does the fine structure of this picture change if fast foldability is taken into account ? Does the structure of the neutral nets in the sequence space change if foldability is also taken into account ?

A fruitful area to which `Kinfold` can be applied for is the design of improved antisense RNA inhibitors on a theoretical basis. Partial or complete annealing between natural complementary RNAs is a crucial process in living cells including gene expression, splicing and antisense regulation. Despite a substantial number of successful applications of artificial antisense RNA in plant genetics or molecular medicine, as a tool to suppress aberrant gene expression or viral functions, little is known about the rules that govern the relationship between RNA structure and annealing kinetics. Computer supported structural design of antisense RNA can serve as a valuable tool to determine RNA–RNA association in *vitro* and biological effectiveness in living cells.

References

[1] F. Aboul-ela, A. I. Murchie D. G. Norman, and D. M. Lilley. Solution structure of a parallel-stranded tetraplex formed by d(TG4T) in the presence of sodium ions by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.*, 243:458–471, 1994.

[2] J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.*, 18:3035–3044, 1990.

[3] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[4] C. B. Anfinsen, E. Haber, M. Sela, and F. H. Jr. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 47:1309–1314, 1961.

[5] V. V. Anshelevich, A. V. Vologodskii, A. V. Lukashin, and M. D. Frank-Kamenetskii. Slow relaxational processes in the melting of linear biopolymers: A theory and its application to nucleic acids. *Biopolymers*, 23:39–58, 1984.

[6] S. Arnott, D. W. L. Hukins, S. D. Dover, W. Fuller, and A. R. Hodgson. Structures of synthetic polynucleotides in the A-RNA and A'-RNA conformation: X-ray diffraction analyses of the molecular conformations of polyadenylic acid·polyuridylic acid and polyinosinic acid·polycytidylic acid. *J. Mol. Biol.*, 81:107–122, 1973.

[7] G. Awang and D. Sen. Mode of dimerization of HIV-1 genomic RNA. *Biochemistry*, 32:11453–11457, 1993.

[8] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Fold. Des.*, 2:261–269, 1997.

[9] 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.

[10] R. Bellman. On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA*, 38:716–719, 1952.

[11] G. Berriz, A. Gutin, and E. Shakhnovich. Langevin model for protein folding: cooperativity and stability. *J. Chem. Phys.*, 1998. submitted.

[12] C. K. Biebricher and R. Luce. In vitro recombination and terminal elongation of RNA by Qβ replicase. *EMBO J.*, 11:5129–5135, 1992.

[13] J. Boyle, G. T. Robillard, and S.-H. Kim. Sequential folding of transfer RNA. A nuclear magnetic resonance study of successively longer tRNA fragments with a common 5' end. *J. Mol. Biol.*, 139:601–625, 1980.

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

[15] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Genet.*, 21:167–195, 1995.

[16] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*, 84:7524–7528, 1987.

[17] S. G. Burston, R. Sleigh, D. J. Halsall, C. J. Smith, J. J. Holbrook, and A. R. Clarke. The influence of chaperonins on protein folding. A mechanism for increasing the yield of the native form. *Ann. N.Y. Acad. Sci.*, 672:1–9, 1992.

[18] 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. RNA tertiary

structure mediation by adenosine platforms. *Science*, 273:1696–1699, 1996.

[19] 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.

[20] T. R. Cech and B. L. Bass. Biological catalysis by RNA. *Annu. Rev. Biochem.*, 55:599–630, 1986.

[21] 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.

[22] M. Chastain and I. Tinoco. Nucleoside triples from the group I intron. *Biochemistry*, 32:14220–14228, 1993.

[23] P. Y. Chou and G. D. Fasman. Prediction of protein conformation. *Biochemistry*, 13:211–245, 1974.

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

[25] 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.

[26] J. Cupal, Ch. Flamm, A. Renner, and P. F. Stadler. Density of states, metastable states, and saddle points. Exploring the energy landscape of an RNA molecule. In T. Gaasterland, P. Karp, K. Karplus, Ch. Ouzounis, Ch. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 88–91, Menlo Park, CA, 1997. AAAI Press.

[27] 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.

[28] A. E. Dahlberg. The functional role of ribosomal RNA in protein synthesis. *Cell*, 57:525–529, 1989.

[29] K. A. Dill. Theory of the folding and stability of globular proteins. *Biochemistry*, 24:1501–1509, 1985.

[30] K. A. Dill, S. Bromberg, K. Yue, K. M. Feibig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding - a perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

[31] K. A. Dill and H. S. Chan. From levinthal to pathways to funnels. *Nature Struct. Biol.*, 4:10–19, 1997.

[32] D. E. Draper. Parallel worlds. *Nature Struct. Biol.*, 3:397–400, 1996.

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

[34] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasi-species. *Adv. Chem. Phys.*, 75:149–263, 1989.

[35] Y. Endo, M. Mitsui, M. Motizuki, and K. Tsurugi. The mechanism of action of ricin and related toxic lectins on eucariotic ribosomes. the site and the characteristics of the modification in 28S ribosomal RNA caused by the toxins. *J. Biol. Chem.*, 262:5908–5912, 1997.

[36] Y. Endo and I. G. Wool. The site of action of α-sarkin on eukaryotic ribosomes. *J. Biol. Chem.*, 257:9054–9060, 1982.

[37] H. Flöckner, M. Braxenthaler, P. Lackner, M. Jaritz, M. Ortner, and M. J. Sippl. Progress in fold recognition. *Proteins*, 23:376–386, 1995.

[38] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[39] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Taranzona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47:2083–2099, 1993.

[40] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for prediction of RNA duplex stability. *Proc.Natl.Acad.Sci.USA*, 83:9373–9377, 1986.

[41] T. Garel and H. Orland. Mean-field model for protein folding. *Europhys. Lett.*, 6:307–309, 1988.

[42] D. T. Gillespie. A general method for nummerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

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

[44] B. Gruenewald, C. U. Nicola, A. Lusitg, G. Schwarz, and H. Klump. Kinetics of the helix-coil transition of a polypeptide with non-ionic side groups, derived from ultrasonic relaxation measurements. *Biophys. Chem.*, 9:137–147, 1979.

[45] A. P. Gultyaev. The computer simulation of RNA folding involving pseudoknot formation. *Nucl. Acids Res.*, 19:2489–2494, 1991.

[46] A. P. Gultyaev, F. H. D. van Batenburg, and C. W. A. Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, 250:37–51, 1995.

[47] A. P. Gultyaev, F. H. D. van Batenburg, and C. W. A. Pleij. Dynamic competition between alternative structures in viroid RNAs simulated by an RNA folding algorithm. *J. Mol. Biol.*, 276:43–55, 1998.

[48] Z. Guo and D. Thirumalai. Nucleation mechanism for protein folding and theoretical prediction for hydrogen-exchange labeling experiments. *Biopolymers*, 35:137–139, 1995.

[49] R. R. Gutell, M. Schnare, and M. Gray. A compilation of large subunit 23S– and 28S–like ribosomal RNA structures. *Nucl. Acids Res.*, 20(suppl.):2095–2109, 1992.

[50] R. W. Hamming. *Coding and Information Theory.* Englewood Cliffs, 2nd ed. prentice-hall edition, 1989. pp. 44-47.

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

[52] M.-H. Hao and H. Scheraga. Statistical thermodynamics of protein folding: comparision of mean-field theory with Monte-Carlo simulations. *J. Chem. Phys.*, 102:1334–1339, 1995.

[53] F.-U. Hartl, R. Hlodan, and T. Langer. Molecular chaperones in protein folding: the art of avoiding sticky situations. *Trends Biochem. Sci.*, 19:20–25, 1994.

[54] T. P. Hausner, J. Atmadja, and K. H. Nierhaus. Evidence that the G2661 region of 23S rRNA is located at the ribosomal binding site of both elongation factors. *Biochimie*, 69:911–923, 1987.

[55] L. He, R. Kierzek, J. SantaLucia, A. E. Walter, and D. H. Turner. Nearest-neighbour parameters for G-U mismatches. *Biochemistry*, 30:11124, 1991.

[56] R. Hecker, Z. Wang, G. Riesner, and D. Steger. Analysis of RNA structures by temperature-gradient gel electrophoresis: viroid replication and processing. *Gene*, 72:59–74, 1988.

[57] D. Herschlag. RNA chaperones and the RNA folding problem. *J. Biol. Chem.*, 270:20871–20874, 1995.

[58] 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.

[59] 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.

[60] I. L. Hofacker, W. Fontana, P. F. Stadler, and P. Schuster. `Vienna RNA Package`. `http://www.tbi.univie.ac.at/~ivo/RNA/`, 1994-98. (Free Software).

[61] 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.

[62] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucl. Acid. Res.*, 12:67–74, 1984.

[63] 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.

[64] 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.

[65] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, 93:397–401, 1996.

[66] J. Ilan. *Translational Regulation of Gene Expression*. Plenum Press New York, 1987.

[67] A. Irback and H. Schwarze. Sequence dependence of self-interacting random chains. *J. Phys. A*, 28:2121–2132, 1995.

[68] G. S. Jackson, R. A. Staniforth, D. J. Halsall, T. Atkinson, J. J. Holbrook, A. R. Clarke, and S. G. Burston. Binding and hydrolysis of nucleotides in the chaperonin catalytic cycle: Iplications for the mechanism of assisted protein folding. *Biochemistry*, 32:2554–2563, 1993.

[69] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, 86:7706–7710, 1989.

[70] L. Jaeger. The world of ribozymes. *Curr. Opin. Struct. Biol.*, 7:324–335, 1997.

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

[72] G. F. Joyce. In vitro evolution of nucleic acids. *Curr. Opin. Struct. Biol.*, 4:331–336, 1994.

[73] R. L. Karpel, D. G. Swistel, and J. R. Fresco. Mechanistic studies of ribonucleic acid renaturation by a helix-destabilizing protein. *Biochemistry*, 21:2102–2108, 1982.

[74] M. Katahira, K. Moriyama, M. Kanagawa, J. Saeki, M. H. Kim, M. Nagaoka, M. Ide, S. Uesugi, and T. Kono T. RNA quadruplex containing g and a. *Nucleic Acids Symp. Ser. 199534*, 34:197–198, 1995.

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

[76] P. Koehl and M. Delarue. A self-consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat. Struct. Biol.*, 2:163–170, 1995.

[77] A. A. Laminet, T. Ziegelhoffer, C. Georgopoulos, and A. Plückthun. The *e. coli* heat shock proteins GroEL and GroES modulate the folding of β-lactamase precursor. *EMBO J.*, 9:2315–2319, 1990.

[78] S. J. Landry and L. M. Gierasch. Polypeptide interactions with molecular chaperones and their relationship to in *vivo* protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, 23:645–669, 1994.

[79] C. M. R. Lemer, M. J. Rooman, and S. J. Wodak. Protein structure prediction by threading methods: evaluation of current techniques. *Proteins*, 23:337–355, 1995.

[80] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA*, 89:8721–8725, 1992.

[81] C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44–45, 1968.

[82] J. R. Lorsch and J. W. Szostak. Chance and necessity in the selection of nucleic acid catalysts. *Acc. Chem. Res.*, 29:103–110, 1996.

[83] P. Loss, M. Schmitz, G. Steger, and D. Riesner. Formation of a thermodynamically metastable structure containing hairpin II is critical for infectivity of potato spindle tuber viroid RNA. *EMBO J.*, 10:719–727, 1991.

[84] 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.

[85] J. Martin, T. Langer, R. Boteva, A. Schramel, A. L. Horwich, and F. U. Hartl. Chaperonin-mediated protein folding at the surface of GroEL through a 'molten globule'-linke intermediate. *Nature*, 352:36–42, 1991.

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

[87] J. Maynard-Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

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

[89] 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.

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

[91] A. Mironov and A. Kister. RNA secondary structure formation during transcription. *J. Biomol. Struct. Dyn.*, 4:1–9, 1986.

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

[93] J. A. Monforte, J. D. Kahn, and J. E. Hearst. RNA folding during transcription by escherichia coli RNA polymerase analyzed by RNA self-cleavage. *Biochemistry*, 29:7882–7890, 1990.

[94] S. R. Morgan and P. G. Higgs. Evidence for kinetic effects in the folding of large RNA molecules. *J. Chem. Phys.*, 105:7152–7157, 1996.

[95] S. R. Morgan and P. G. Higgs. Evidence for kinetic effects in the folding of large RNA molecules. *J. Chem. Phys.*, 105:7152–7157, 1996.

[96] S. Mosiman, R. Melishko, and M. G. N. James. A critical assessment of comparative molecular modelling of tertiary structure of proteins. *Proteins*, 23:301–317, 1995.

[97] A. Nakaya, K. Yamamoto, and A. Yonezawa. RNA secondary structure prediction using highly parallel computers. *Comput. Applic. Biosci.*, 11:685–692, 1995.

[98] A. Nakaya, A. Yonezawa, and K. Yamamoto. Classification of RNA secondary structures using the techniques of cluster analysis. *J. Theor. Biol.*, 183:105–117, 1996.

[99] H. F. Noller, J. Kop, V. Wheaton, J. Brosius, R. R. Gutell, A. M. Kopylov, F. Dohme, W. Herr, D. A. Stahl, R. Gupta, and C. R. Woese. Secundary structure model for 23S ribosomal RNA. *Nucl. Acids Res.*, 9:6167–6189, 1981.

[100] 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.

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

[102] V. Pande, A. Yu Grosberg, and T. Tanaka. Freezing transition of random heteropolymers consisting of arbitrary sets of monomers. *Phys. Rev. E*, 51:3381–3393, 1995.

[103] D. Peralta, D. J. Hartman, N. J. Hoogenraad, and P. B. Hø. Generation of a stable folding intermediate which can be rescued by the chaperonins GroEL and GroES. *FEBS*, 339:45–49, 1994.

[104] C. W. Pleij. Pseudoknots: a new motif in the RNA game. *Trends Biochem. Sci.*, 15:143–147, 1990.

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

[106] 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.

[107] D. Pörschke. Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix. *Biopys. Chem.*, 1:381–386, 1974.

[108] A. M. Pyle and J. B. Green. RNA folding. *Curr. Opin. Struct. Biol.*, 5:303–310, 1995.

[109] D. Riesner, T. Baumstark, F. Qu, T. Klahn, P. Loss, V. Rosenbaum, M. Schmitz, and G. Steger. *Physical basis and biological examples of metastable RNA structures*, pages 401–435. Structural Tool for the Analysis of Protein – Nucleic Acid Complexes. Birkhäuser, Basel, 1992.

[110] D. Riesner, K. Henco, and G. Steger. *Advances in Electrophoresis*, volume 4, pages 169–250. VCH Verlagsgesellschaft, Weinheim, 1991.

[111] 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.

[112] D. Riesner, G. Steger, R. Zimmat, R. A. Owens, M. Wagenhöfer, W. Hillen, S. Vollbach, and K. Henco. Temperature-gradient gel electrophoresis of nucleic acids: Analysis of conformational transitions, sequence variations, and protein-nucleic acid interactions. *Electrophoresis*, 10:377–389, 1989.

[113] V. Rosenbaum, T. U. Klahn, E. Lundberg, E. Holmgren, A. von Gabain, and D. Riesner. Co-existing structures of an mrna stability determinant. the 5' region of the escherichia coli and serratia marcescens ompa mrna. *J. Mol. Biol.*, 229:656–670, 1993.

[114] V. Rosenbaum and D. Riesner. Temperature-gradient gel electrophoresis. thermodynamic analysis of nucleic acids and proteins in purified form and in cellular extracts. *Biophys. Chem.*, 26:235–246, 1987.

[115] A. Šali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding. A lattice model study on the requirements for folding of native states. *J. Mol. Biol.*, 253:1614–1636, 1994.

[116] M. Sasai and P. G. Wolynes. Unified theory of collapse, folding and glass transition in associative-memory hamiltonian models of proteins. *Phys. Rev. A*, 46:7979–7997, 1992.

[117] M. Schmitz and G. Steger. Discription of RNA folding by simulated annealing. *J. Mol. Biol.*, 225:254–266, 1996.

[118] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. London B*, 255:279–284, 1994.

[119] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledgebased prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[120] R. A. Staniforth, S. G. Burston, T. Atkinson, and A. R. Clarke. Affinity of chaperonin-60 for a protein substrate and its modulation by nucleotides and chaperonin-10. *Biochem. J.*, 300:651–658, 1994.

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

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

[123] A. A. Szewczak and P. B. Moore. The sarcin/ricin loop, a modular RNA. *J. Mol. Biol.*, 247:81–98, 1995.

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

[125] 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.

[126] M. A. Tanner, E. M. Anderson, R. R. Gutell, and T. R. Cech. Mutagenesis and comparative sequence analysis of a base triple joining the two domains of group I ribozymes. *RNA*, 3:1037–1051, 1997.

[127] D. Thirumalai and S. A. Woodson. Kinetics of folding of proteins and RNA. *Acc. Chem. Res.*, 29:433–439, 1996.

[128] M. J. Todd, G. H. Lorimer, and D. Thirumalai. Chaperonin-facilitated protein folding: Optimization of rate and yield by iterative annealing mechanism. *Proc. Natl. Acad. Sci. USA*, 93:4030–4035, 1995.

[129] M. J. Todd, P. V. Viitanen, and G. H. Lorimer. Dynamics of the chaperonin ATPase cycle: Implications for facilitated protein folding. *Science*, 265:659–666, 1994.

[130] J. Tsang and G. F. Joyce. *In vitro* evolution of randomized ribozymes. *Methods Enzymol.*, 267:410–426, 1996.

[131] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Ann. Rev. Biophys. Biophys. Chem.*, 17:167–192, 1988.

[132] D. H. Turner, N. Sugimoto, and S. M. Freier. *Thermodynamics and Kinetics of Base-pairing of DNA and RNA Self-assembly and Helix Coil Transition*, chapter Nucleic Acids, pages 201–227. Springer-Verlag, Berlin, 1990.

[133] O. C. Uhlenbeck. Keeping RNA happy. *RNA*, 1:4–6, 1995.

[134] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of RNA folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.

[135] 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.

[136] M. S. Waterman and T. Byers. A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Math. Biosci.*, 77:179–188, 1985.

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

[138] K. M. Weeks. Protein-facilitated RNA folding. *Curr. Opin. Struct. Biol.*, 7:336–342, 1997.

[139] J. S. Weissman, Y. Kashi, W. A. Fenton, and A. L. Horwich. GroEL-mediated protein folding proceeds by multiple rounds of binding and release of nonnative forms. *Cell*, 78:693–702, 1994.

[140] E. Westhof and L. Jaeger. RNA pseudoknots. *Curr. Opin. Struct. Biol.*, 2:327–333, 1992.

[141] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structure. *Biopolymers*, 1998. *submitted.*

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

[143] 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.

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

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

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

[147] R. Zwanzig, A. Szabo, and B. Bagchi. Levintal's paradox. *Proc. Natl. Acad. Sci. USA*, 89:20–22, 1992.

Curriculum vitae

Mag. Christoph Flamm
1967-01-19

Schulbildung:	Volksschule in Wien Bundesrealgymnasium 9 in Wien Matura im Oktober 1986
Studium:	Chemie (1989-96) Universität Wien
Diplomarbeit:	Inst. für Org. Chemie (1994-95) bei Prof. Dr. Edda Gössinger (Naturstoffsynthese)
Titel:	Synthetische Untersuchungen zum Antibiotikum Nodusmicin
Dissertation:	Inst. für Theor. Chemie (1996-98) bei Prof. Dr. Peter Schuster)
Titel:	Kinetic Folding of RNA

List of Publication

[1] Christoph Flamm, Susi Rauscher, Christian Mandl and Peter F. Stadler. New conserved secondary structure motifs at the 3'end of rna genome of tick-borne flaviviruses. In *Symposium on Modern Approaches to Flavivirus Vaccines, Vienna, Austria*, page B18, 1996.

[2] Jan Cupal, Christoph Flamm, Alexander Renner, and Peter F. Stadler. Density of states, metastable states, and saddle points. Exploring the energy landscape of an RNA molecule. In T. Gaasterland, P. Karp, K. Karplus, Ch. Ouzounis, Ch. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 88–91, Menlo Park, CA, 1997. AAAI Press.

[3] Ivo L. Hofacker, Martin Fekete, Christoph Flamm, Martijn A. Huynen, Susanne Rauscher, Paul E. Stolorz, and Peter F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.*, 26:3825–3836, 1998.

[4] Susanne Rauscher, Christph Flamm, Christian Mandl, Franz X. Heinz, and Peter F. Stadler. Secondary structure of the 3'-non-coding region of flavivirus genomes: Comparative analysis of base pairing probabilities. *RNA*, 3:779–791, 1997.

An dieser Stelle möchte ich mich herzlich bei all jenen bedanken, die zum Entstehen der vorliegenden Arbeit beigetragen haben.

Prof. Peter Schuster, der mir nicht nur die Möglichkeit zur Dissertation in seiner Arbeitsgruppe bot und mich so in das wissenschaftliche Arbeiten einführte, sondern auch dafür sorgte, daß es nie an den Resourcen, den Ideen und der Motivation mangelte.

Dr. Walter Fontana, der sich immer Zeit für extensive Diskussionen nahm, bei denen Kreide und Tafel niemals fehlen durften.

Dr. Peter Stadler für unzählige Anregungen und Ratschläge.

Dr. Ivo Hofacker rettete mich oftmals aus den unendlichen Weiten des "Computer Universums".

Judith Jakubetz, die mir über so manche bürokratische Hürde hinweghalf.

Ronke Babajide, Jan Cupal, Martin Fekete, Thomas Griesmacher, Christian Haslinger, Stephan Kopp, Bärbel Krakhofer, Stefan Müller, Susanne Rauscher, Alexander Renner, Roman Stocsits, Norbert Tschulenk, Günther Weberndorfer, Andreas Wernitznig, Stefan Wuchty, die alle für ein angenehmes Arbeitsklima sorgten.

Eine innige Umarmung gilt meiner Freundin Mano.

Zum Schluß meinen Eltern, die mir durch ihre Unterstützung ein Studium ermöglichten.