Molecular Evolution of Short RNA Molecules - Neutral Nets in Sequence Spaces and Kinetic Properties of RNA

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften

eingereicht von

DI Michael Kospach

an der Fakultät für Naturwissenschaften und Mathematik

der Universität Wien

im Jänner 2003

Abstract

RNA can serve as an ideal model for evolution. In a simple way it combines genotype and phenotype in a single molecule. The minimum free energy (mfe) structure of a RNA molecule is directly derivable from its sequence. Neutrality in terms of constant fitness plays a major role in evolution to overcome local maxima in the fitness landscape and can again be realised in the RNA model. Many sequences fold into the same secondary structure. As fitness can be derived directly from structure a population drifting on such a set of equally fit sequences undergoes neutral evolution. Random drift on the other hand is only possible if there exists a net of neutral neighbours that are accessible by single mutations which are chosen from a set of mutation operators. Those components can be derived by breadth-first-traversal algorithms from the larger neutral net of all sequences that fold into the same secondary structure. Sequence to structure mappings can be derived by folding all sequences of a certain chain length and over a given alphabet into their mfe secondary structure. In this work this was done for the sequence spaces of the alphabets GC and AU for chain length up to 30, for AUG and UGC alphabets up to chain lengths of 20 and for the natural alphabet of AUGC for chain lengths up to 16. Sequence to structure maps were computed for different folding parameters, compared and partitioned into components. The generic features are comparable to those obtained in previous less extensive calculations using an older set of folding parameters. The total number of structures formed increase exponentially with the chain length. There is a small number of common and many rare structures roughly following a generalised Zipf's law. The fraction of sequences that do not form a stable mfe structure other than the open chain decreases with an increasing chain length. The fraction of common structures also decreases whereas the fraction of sequences folding into common structures increases with a growing chain length so that at large chain lengths nearly all sequences fold into a small number of stable secondary structures. As far as the relatively small number of cases studied allows for general conclusions, these features are not bound to certain alphabets or folding parameter sets. We find

the neutral networks of the higher ranked and rarer structures to be more often split into a large number of components. Neutral nets were ranked by their size starting at rank one for the largest network. Common structures tend to show a single giant component or up to four large components. If a structure decomposes into two to four almost equal sized components they often differ clearly in their base composition which can be explained by structural features. Such structural elements, that would allow the formation of additional base pairs whenever the sequences carry complementary bases at the corresponding positions, lead to systematic biases from an even distribution of sequences folding into the same structure in the space of compatible sequences. The second part of the work examined the kinetics of RNA folding using an already established algorithm. The components of the small tested neutral networks did not show a clear difference in their overall mean folding time according to their size but statistics of data obtained from kinetic folding of a large number of sequences of different length and secondary structure and from analysing their energy landscapes showed some clear results. The wide-spread prejudice that a sequence's mfe determines its folding kinetics could be disproved. The mean folding time does not correlate with the minimum free energy. It rather strongly depends on the height of the energy barrier that separates the most important local minimum which often is associated with the highest barrier, from the mfe structure on the folding path. An analysis of the distribution of folding times suggests a logarithmic normal distribution. Depending on the energy landscape and the occurrence of important local minima several of such log-normal distributions can overlap, one for each major barrier to all other important local minima visited on the folding path. Whereas the resulting overall distribution often differs from a log-normal distribution, but in simple cases it is well described by this distribution.

Zusammenfassung

Ribonukleinsäure (RNA) ist ein ideales Modell der Evolution. Sie kombiniert auf eine einfache Art und Weise Genotyp und Phänotyp in einem einzigen Molekül. Die Sekundärstruktur mit minimaler freier Energie (mfe) ist direkt aus ihrer Sequenz ableitbar. Neutralität in Bezug auf konstante Fitness spielt eine wichtige Rolle in der Evolution um lokale Maxima in der Fitnesslanschaft zu überwinden und kann ebenfalls im RNA-Modell realisiert werden. Viele Sequenzen falten in die selbe Sekundärstruktur. Da die Fitness direkt aus der Struktur abgeleitet werden kann, unterzieht sich eine Population, die auf dieser Menge von Sequenzen gleicher Fitness umherdriftet neutraler Evolution. Zufallsdrift ist wiederum nur möglich, wenn es ein Netz von neutralen Nachbarn gibt, die in Einzelmutationen, die aus einer Menge von Mutationsoperatoren gewählt werden können, erreichbar sind. Diese Komponenten kann man durch einen Breitensuchalgorithmus aus dem größeren neutralen Netz der Sequenzen, die in die selbe Struktur falten extrahieren. Sequenz-Struktur-Abbildungen kann man dadurch erreichen, daß man alle Sequenzen einer konstanten Kettenlänge und eines bestimmten Alphabets in ihre mfe-Sekundärstruktur faltet. In dieser Arbeit wurde das für die Sequenzräume der Alphabete GC und AU bis zu einer Kettenlänge von 30, für das AUG und das UGC Alphabet für Kettenlängen bis zu 20 und für das natürliche Alphabet von AUGC für bis zu 16 Nukleotid lange Ketten durchgeführt. Die Sequenz-Struktur-Abbildungen wurden mit unterschiedlichen Parametern berechnet, verglichen und ihre Komponenten bestimmt. Ihre allgemeinen Eigenschaften sind mit früheren weniger ausführlichen Berechnungen vergleichbar, bei denen ältere Parameter verwendet wurden. Die Gesamtzahl der gebildeten Strukturen steigt exponentiell mit steigender Kettenlänge an. Es gibt eine kleine Anzahl an häufigen und viele seltene Strukturen die grob einem verallgemeinerten Zipf-Gesetz folgen. Der Anteil der Sequenzen, die keine mfe-Struktur außer der offenen Kette bilden, sinkt mit der Kettenlänge genauso wie der Anteil an häufigen Strukturen, während der Anteil von Strukturen die in häufige Strukturen falten steigt, sodaß bei großen Kettenlängen nahezu alle Sequenzen in eine kleine Anzahl von stabilen

Sekundärstrukturen falten. Soweit die relativ geringe Anzahl an untersuchten Fällen allgemeine Schlußforderungen zulassen, sind diese Eigenschaften nicht an bestimmte Alphabete oder Faltungsparameter gebunden. Wir finden, daß sich neutrale Netzwerke der höheren Ränge öfters in eine Vielzahl von Komponenten aufspalten. Neutrale Netze wurden beginnend mit Rang eins für das größte Netzwerk der Größe nach gereiht. Häufige Strukturen zeigen eher eine einzige Riesenkomponente oder bis zu vier große Komponenten, die sich dann deutlich in ihrer Basenzusammensetzung unterscheiden, was man auf strukturelle Eigenschaften zurückführen kann. Solche strukturellen Elemente, die die Bildung eines weiteren Basenpaares ermöglichen würden wann immer sich komplementäre Basen an den entsprechenden Positionen befinden, führen zu systematischen Abweichungen von einer gleichmäßig Verteilung der Sequenzen, die in die gleiche Struktur falten, im Raum der kompatiblen Sequenzen.

Im zweiten Teil der Arbeit wurde die Faltungskinetik von RNA mit bereits bestehenden Algorithmen untersucht. Die Komponenten der untersuchten kleinen neutralen Netzwerke zeigten keinen deutlichen Unterschied in ihrer gesamten mittleren Faltungszeit in Abhängigkeit von ihrer Größe aber die Statistik von kinetischen Faltungsdaten einer großen Menge von Sequenzen verschiedener Länge und Sekundärstruktur und die Analyse ihrer Energielandschaften ergab eindeutige Ergebnisse. Das weit verbreitetes Vorurteil, daß die minimale freie Energie einer Sequenz ihre Faltungskinetik festlegt, konnte widerlegt werden. Die mittlere Faltungszeit ist von der minimalen freien Energie unabhängig. Vielmehr hängt sie hauptsächlich von der Höhe der Energiebarriere ab, die das wichtigste lokale Minimum, das oft zu der höchsten Barriere gehört, von der mfe-Struktur am Faltungsweg trennt. Die Analyse der Faltungszeiten legt eine logarithmische Normalverteilung nahe. Abhängig von der Energielandschaft und dem Auftreten wichtiger lokaler Minima können mehrere solche Verteilungen überlappen. Wohingegen sich die resultierende Gesamtverteilung oft von einer logarithmischen Normalverteilung unterscheidet, aber in weniger komplizierten Fällen wird sie gut von dieser Verteilung beschrieben.

Contents

List of Figures

List of Tables

1 Introduction

1.1 General Context

Evolution of RNA is especially important as it plays a major role at an very early stage of life. Cells contain batteries of protein based enzymes for manipulating DNA but few for processing RNA, however in contrast to proteins nucleic acids can direct their own synthesis and many coenzymes are ribonucleotides. There exists a broad range of catalytic RNAs that are also known as ribozymes [70]. In the same manner DNAs called deoxyribozymes show catalytic function, for example in site specific cleaving of single stranded DNA similar to restriction enzymes [6]. On the other hand RNA is an essential contributor in protein expression. Messenger RNA (mRNA) carries the genetic information stored in DNA to the ribosomes that also contain RNA where transfer RNA (tRNA) enables the translation into proteins. In other words RNA serves as information carrier and as catalyst. These observations led to the RNA-world hypothesis [20, 75]. Another hint supporting this hypothesis is that ribosomes the places of protein synthesis are made of 2/3 RNA and only 1/3 protein which makes it imaginable that protoribosomes were self replicating molecules entirely made of RNA that evolved the ability to influence the synthesis of proteins. The RNA-world hypothesis proposes that in the beginning before the occurrence of cells life was based on RNA.

Both key functions: catalysis and the storage of genetic information were fulfilled by RNA before the occurrence of proteins and DNA that mainly serve this function nowadays. The role of RNA is reduced to an intermediate at protein synthesis and some minor functions in modern eucaryotes but it is still the most important component of some viruses today. Nowadays RNA serves as the hereditary material of many viruses and was found to be the only substance viroids are made of. Viroids are single stranded circular RNAs that cause infectious diseases [14]. They are supposed to be molecular fossils of precellular self-replicating RNAs. An alternative hypothesis suggests that because of the similarities to self-splicing group I introns that occur in

mitochondrial and ribosomal RNA genes, viroids are 'escaped introns'.

Compared to proteins or double stranded DNA that stores genetic information in organisms, RNA which occurs mainly single stranded is chemically less stable but more active and can form a huge range of structures. Because of its instability RNA proved to be less useful to act as a storage molecule for genetic information than DNA. On the other hand the same instability led to an increased error rate at the process of reduplication. This disadvantage at the storage function can also be considered as an increased variation which is the motor of evolution. In living organisms of today the shortest generation time is about 30 minutes in some bacteria whereas RNA can have generation times of less than one second. This makes RNA a perfect target for evolutionary studies which were first experimentally performed by Sol Spiegelman [67] in his serial transfer tests and later in SELEX (systematic evolution of ligands by exponential enrichment) [40, 71, 84] experiments where evolutionary trends can be observed in a short time range which would take at least some thousand times longer in organisms living today. SELEX is an in vitro selection of functional nucleic acids and makes use of large populations of random RNA or DNA sequences as the raw material for the selection of rare functional molecules. For example RNAs can be evolved to bind a protein. This would involve the following steps:

- RNA transcripts with a sequence which confers binding specificity for a target protein are selected for by incubating a RNA library with the target protein which has been immobilised on nitrocellulose filters.
- The filters are washed to remove non-specifically bound RNA molecules.
- Specific binding RNA molecules are eluted from the target protein and are collected.
- The eluted RNA molecules are converted to single strand cDNA. Duplex DNA is produced from the single strand cDNA by PCR.
- In vitro transcription produces a library of RNA molecules enriched for sequences with binding affinity for the target protein.

- The process of selection and enrichment is continued until a collection of high affinity binding sequences has been produced.

Small interfering RNA (siRNA) is a tool that is used for systematically deciphering the function and interactions of genes [45]. It is a promising and already widely used technology to knock down the expression of any gene in vertebrate cells using double stranded RNA (dsRNA) which is called post-transcriptional gene silencing. This makes it important to explore properties of such small RNAs. A siRNA typically consists of two 21 nucleotides single stranded RNAs that are complementary to both strands of the silenced gene and form a 19 base pair duplex. This is a length that falls in the range of small RNA molecules whose properties are studied in this work. The anti-sense strand of the siRNA guides mRNA cleavage and degradation of the mRNA of the gene to be silenced. In promising studies virus infections of mammalian cells could be inhibited by directing siRNAs against viral mRNA.

Structural information and insight into RNA catalysis came from the first crystal structure of a hammerhead ribozyme [54]. This is the first known structure of a complex RNA besides those of tRNAs. Hammerhead ribozymes are embedded in the RNAs of certain plant viruses and are named due to the shape of their secondary structures.

1.2 RNA Secondary Structures

The possible conformations of RNA are delimited by a random coil or open chain on one hand and by the thermodynamically most stable or minimum free energy structure on the other hand. This minimum free energy (mfe) structure often but not always is also the native structure. Compared to the open chain structure a molecule folded into another structure is more resistant to degradation and often this defined structure is important for a biochemical function. In three dimensional structures there are several types of contacts that are not well understood or studied. An example are pseudo-knots [28, 53] where some data is already available. On the other hand there are algorithms to calculate the two dimensional minimum free energy structure of single stranded

RNA [31, 44, 50, 81, 89]. This secondary structure is a coarse grained model of the real tertiary structure. It shows the contacts which occur by specific base pairing. These base pairs are the result of hydrogen bonding interactions between complementary bases. Like DNA, RNA contains four different bases. Guanine (G) and cytosine (C) and adenine (A) are also found in DNA but instead of thymine RNA contains uracil (U). The possible base pairs in RNA are the Watson-Crick pairs namely A–U and G–C that are also found in DNA but also G–U pairs which do not occur in DNA.

This secondary structure model only gives information of pairing and non pairing regions but not about distances. Stacking is the force that drives the folding into secondary structures, but the formation of an energetically favourable double stranded region implies the formation of a energetically unfavourable loop at the same time. There is progress in including some frequent types of pseudo-knots into the folding algorithm [29] and the co-folding of two RNA strands. The secondary structure of RNA can be seen as an intermediate state in the folding process from random coil to tertiary structure. This can be shown by increasing temperature: tertiary interactions disappear fist, secondary structure elements dissociate later [3].

1.3 Sequence Space

A sequence space is a discrete point space that has as many points as there are different possible sequences. There is one and only one point for each sequence and all points are ordered by the Hamming distance. The term Hamming distance [27] named after Richard W. Hamming was first introduced in information theory and equals the number of positions that differ in two aligned sequences. This Hamming distance d_H satisfies the properties of a metric:

- $d_H(X, Y) = 0$ if and only if $X = Y$
- $d_H(X, Y) = d_H(Y, X)$
- $d_H(X, Z) \leq d_H(X, Y) + d_H(Y, Z)$ for any strings X, Y, Z

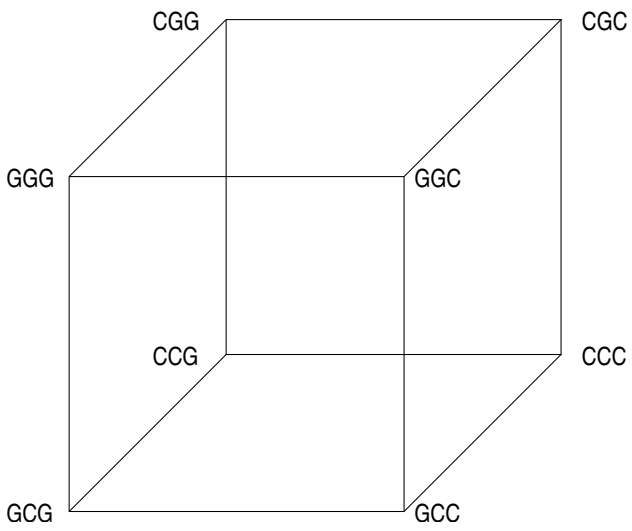


Figure 1: Sequence space over the alphabet GC and sequence length 3. A binary sequence space is a hypercube, each corner representing one of the 2^n possible binary sequences of length n [9]. The lines connect nearest neighbours, i.e. mutants that differ in only one position.

Figures 1, 3 and 4 give illustrations of sequence spaces over different alphabets. The term landscape was introduced by Sewall Wright [85, 86]. Similar to a natural landscape a fitness landscape consists of many local suboptimal maxima. An evolving system may reach the global maximum and escape the local maximum if there exits a ridge of points having the same fitness by randomly travelling across this neutral ridge until a point of higher fitness is found (see figure 5). Landscapes are often constructed in two steps. In our case this is obtained by a first mapping from sequence space, which is formed by all possible sequences of a certain chain length and alphabet, into (secondary) structures and a second mapping from the space of structures (i. e. phenotype space) into real numbers representing the fitness of those structures. There are more RNA sequences than secondary structures and many RNA sequences form the same secondary structure which is called neutrality.

In general there are four different approaches to study sequence-structure maps of RNA. Those methods and their limits are shortly described in the following:

- Random graph theory [56] which uses a mathematical model. In general the random graph approach works rather well. Exceptions are for example directly related to structure. See figure 2 for examples of such structural elements.

- Statistical evaluation of random walks in sequence space or by means of inverse folding [15, 63]. This allows the examination of longer sequences but statistics also means limited accuracy.

- Simulation of evolutionary dynamics [16, 17, 36, 74] making use of chemical kinetics of replication and mutation which is restricted to small parts of sequence space.

- Exhaustive folding and enumeration [23, 24, 64] using a folding algorithm that brings about exact results. As a disadvantage this method can only be applied to short chain lengths.

The strategy used in this work is complete enumeration. It is a possible way to find properties of neutral networks of RNA in respect to neutral evolution. Neutrality in evolution means that mutations occur that do not affect the fitness. Kimura [39] assumed that most mutants are selectively neutral and adaptive mutations are rare. His theory is supported by recent experimental data [52] from bacteria. The structure of a RNA molecule is often linked to a biological function therefore sequences folding into the same (secondary) structure are assumed to have the same fitness. In the same manner as in neutral evolution of whole organisms a neutral mutation in respect to RNA secondary

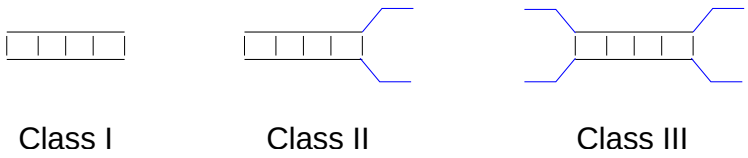


Figure 2: Structures can be classified according to their ability of forming additional base pairs. Class II and III structures can lead to deviations from predictions of random graph theory. Figure 33 gives more details.

structure leads to a different sequence that has still the same phenotype (i.e the same minimum free energy structure) as before the mutation occurred. This of course implies the new sequence to be compatible to the structure i.e. it consists of two bases that are capable of forming a pair at any two positions

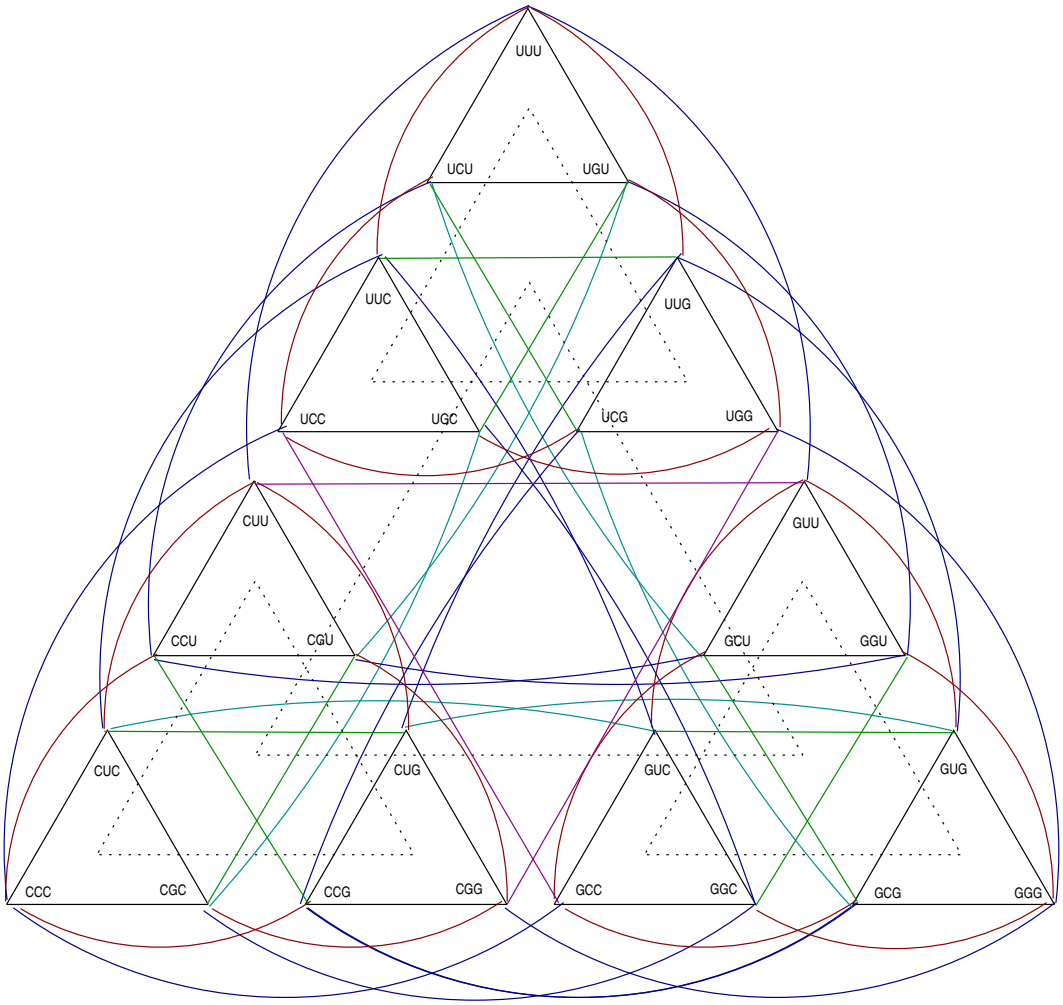


Figure 3: Sequence space over the alphabet UGC and sequence length 3. In contrast to the sequence space of a binary alphabet which can be constructed as a line for sequence length $n = 1$, a quadrat for $n = 2$ and a cube for $n = 3$, the sequence space of a three letter alphabet is iteratively constructed as a triangle (large dotted triangle above) for $n = 1$, three connected triangles (small dotted triangles above) for $n = 2$ and nine triangles for $n = 3$. The corners represent the sequences and solid lines and arcs connect nearest neighbours.

that are paired in the target structure. All sequences that fold into the same secondary structure spread out a neutral network.

Computer simulations of neutral evolution were extensively performed [16, 36, 72, 73]. It could be shown that neutral changes can set the stage for a mutation that leads to a better adapted structure [36]. Neutral evolution gives

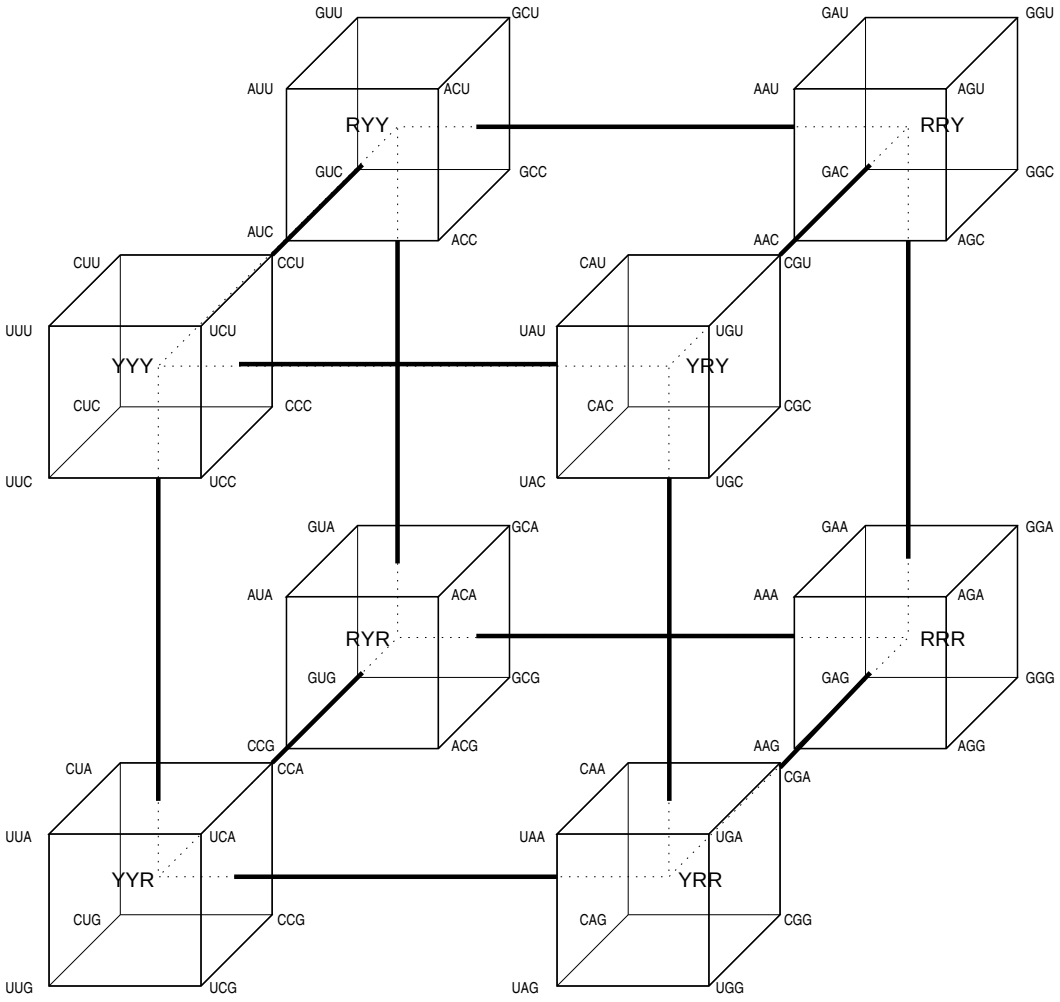


Figure 4: Sequence space over the alphabet AUGC and sequence length 3. In analogy to figure 1 a hypercube builds the sequence space of a two letter alphabet consisting of purines (R) and pyrimidines (Y) [9]. The purines consist of either adenine (A) or guanine (G), and the pyrimidines can be either uracil (U) or cytosine (C) leading to additional eight hypercubes as subspaces at each corner of the binary purine pyrimidine hypercube.

access to a virtually unlimited number of structures and can thus play an important role in adaptive evolution. A scenario of adaptive evolution, where a population evolves over a suboptimal neutral net until it encounters another net with a better secondary structure after which it relocalises to this new net, was observed in simulations on a fitness landscape that was based on RNA secondary structure [35]. Simulations of replicating and mutating RNA populations under selection show that sudden adaptive progress coincides mostly, but not always, with discontinuous shape transformations [16] and confirm the importance of neutral genetic drift periods between steps of sudden increases of fitness. Eric van Nimwegen et. al. analysed a model of a population evolving over neutral networks of RNA secondary structures and showed that the tendency to evolve toward highly connected parts of the network is solely determined by the network topology [73]. A population must explore large portions of neutral networks before it discovers a rare connection to fitter phenotypes [72].

All possible RNA sequences of a certain chain length form the corresponding sequence space of this chain length and the used alphabet. By exhaustive folding of all its sequences all minimum free energy structures of this sequence space can be achieved. The neutral network of all sequences folding into a certain structure decompose into components according to given rules of possible mutations (e.g. only point mutations are allowed). All components of a given structure form the sequence of components.

The sequence spaces of two letter alphabets (AU and GC) and the four letter alphabet (AUGC) where studied previously by Walter Grüner *et al.* [23, 24] and Ulrike Göbel [22], respectively (see figures sequencespaceGC and sequencespaceAUGC for illustrations). The four letter alphabet because it is the most important one as it occurs naturally, the two letter alphabets mainly because they are easier to study because of the smaller sequence space at a given chain length. But there is not only this practical aspect why to explore sequence spaces of reduced alphabets. In the early days of life the original genetic material probably contained less than four different subunits. Even a system involving only adenine and inosine has been proposed [7]. Cytidine is

the least stable of the four nucleosides occurring in RNA because it tends to spontaneously deaminate to uridine and there are examples of natural RNAs that have a very low cytidine content. For these reasons RNAs consisting of three different kinds of nucleotides, especially those made of AUG could play a role in vivo. It was shown that the folding of RNA into a catalytic active structure is possible in a three nucleotide system [58]. An RNA ligase ribozyme that lacked cytidine was achieved through in vitro evolution and still showed activity. This class I ligase has a chain length of about 140 nucleotides and catalyses the joining of the 3'-hydroxyl of a template-bound oligonucleotide substrate to the 5'-triphosphate of the ribozyme.

1.4 Sequence of Components

Neutrality in evolution is important to overcome local maxima of fitness. Fitness increases until a local maximum is reached. The 'valley' between this local maximum and the next higher state of fitness can be overcome by a random drift within a neutral network bridging this valley (figure 5).

In respect to RNA a sequence which has a minimum structural distance to a target secondary structure can be seen as a sequence of maximum fitness. Computer simulations of such an evolutionary process from sequences folding into random structures to sequences folding into a target structure performed in a flow reactor showed a discontinuous increase of fitness [16]. The alternation between adaptive walks that bring about a considerable increase of fitness and periods of unselected random drift within neutral networks where no noticeable change in fitness can be seen allow the escape from evolutionary traps in rugged fitness landscapes [62]. The widths of the valleys crossed rather than its depths influence the escape from local optima [73].

This means that the crossing of valleys of local minima is only possible if there exists a connected neutral network which is a network where every member can be reached by stepwise mutations chosen from the allowed set of mutational operators.

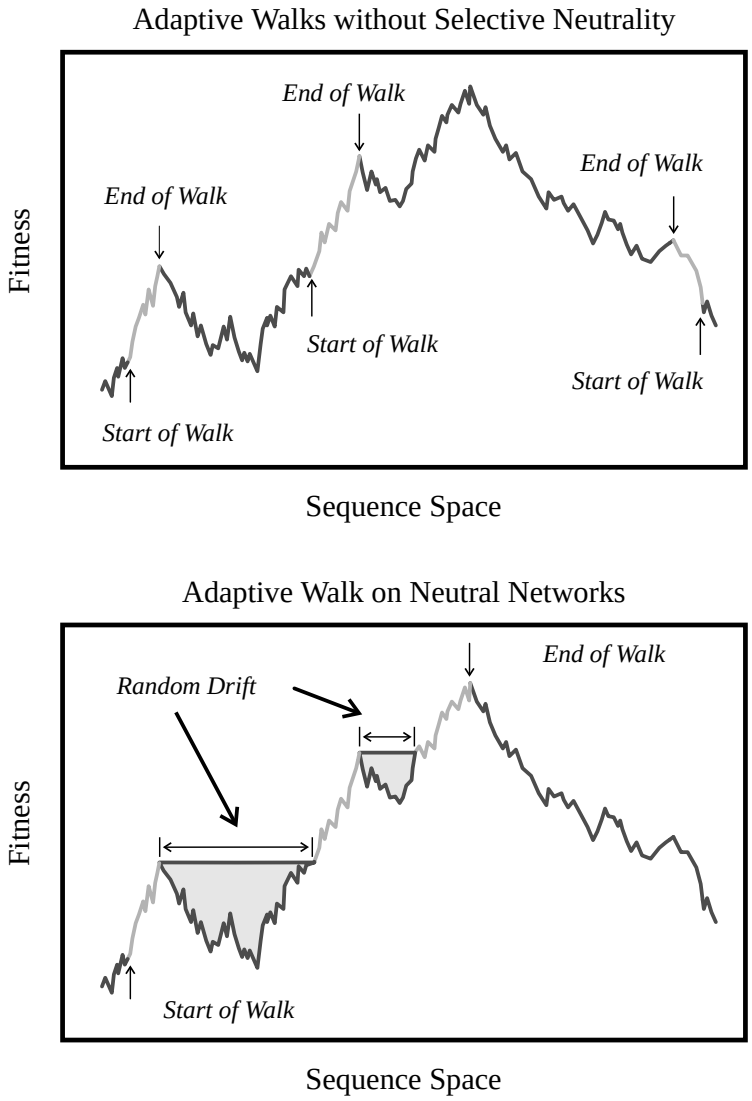


Figure 5: A schematic representation illustrating the importance of neutral nets to overcome local minima [61, 62]. Optimisation occurs through adaptive walks and random drift. Adaptive walks allow to choose the next step arbitrarily from all directions where fitness is non-decreasing. Populations can bridge over narrow valleys with widths of a few point mutations. In the absence of selective neutrality (upper picture) they are, however, unable to span larger Hamming distances and thus will approach only the next major fitness peak. Populations on rugged landscapes with extended neutral networks evolve along the networks by a combination of adaptive walks and random drift at constant fitness (lower picture). In this manner, populations bridge over large valleys and may eventually reach the global maximum of the fitness landscape.

A possible set of operators would consist of only point mutations leading to a Hamming distance $d_H = 1$ between the mutated and the original sequence. An extended set of mutational operators could also include base pair exchanges resulting in possible Hamming distances of $d_H = 1$ and $d_H = 2$. A point mutation at a pairing base that would break this base pair because the new base and remaining unchanged base cannot form one of the known base pairs, often brings about a compensatory mutation at the corresponding former pairing partner to reestablish the compatibility of the sequence to the original structure [30, 59]. This consecutive mutation is driven by selection pressure because the structure is often linked to an essential function or at least a selection advantage of the molecule.

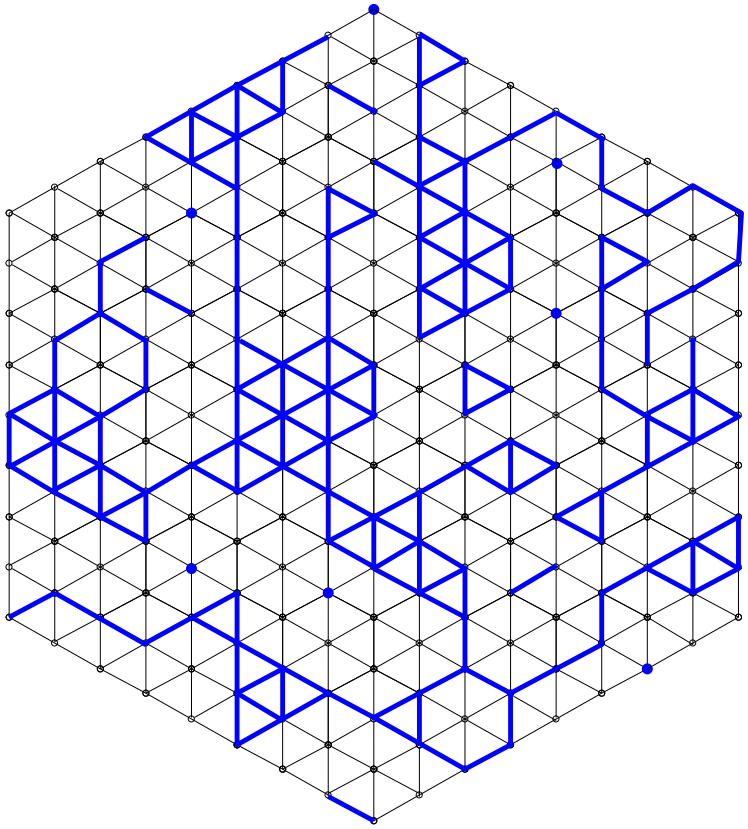


Figure 6: Neutral networks in sequence space. Neutral networks may be totally connected or they may consist of several components (blue networks).

Using such a set of mutation operators a neutral network can be parted into subsets of this network leading to components which are connected networks (see figure 6 for an illustration). A random drift is possible within such a component. For this reason the components were calculated for complete sequence spaces in this work. All components belonging to a neutral network are called the sequence of components.

The components were achieved by finding all sequences that can be reached by stepwise (neutral) mutation of one base or one base pair within the component.

1.5 Energy Landscapes and Kinetics of RNA Folding

The topology of an energy landscape greatly influences the folding kinetics and the possible folding paths. By calculating all suboptimal conformations of a certain energy range above the minimum free energy it is possible to reveal the underlying energy landscape and identify local minima and saddle points resulting in the energy barriers to the thermodynamic ground state [8, 12, 87]. RNA molecules are believed to exhibit rugged energy landscapes of many deep local minima. Those local minima can be mis-folded structures that consist of helices that are difficult to open once they are formed. It is believed that they play a major role in the kinetics of RNA folding [47].

The kinetics of energy folding can be simulated by random walk algorithms making use of a selected set of possible moves [12].

1.6 Organisation of this Work

First we will discuss some basic aspects of RNA and its folding into minimal free energy structures in chapter 2. The algorithms some of the used programs are based on and their background are discussed in chapter 3. In chapter 4 first the results obtained from numerous calculations of minimum free energy structures of whole sequence spaces and their neutral nets are presented. Then we will show the results of a great number of simulations of kinetic foldings of different RNA sequences and compare them to their analysed energy landscape. Finally the outcome of this work and further investigations that are to be done in future are discussed in chapter 5.

2 Basics: The Structure of RNA

In contrast to the primary structure which is simply the 5' to 3' list (or sequence) of covalently linked nucleotides, named by the attached base, the tertiary structure describes the positions of every atom in a RNA molecule in three dimensional space. An intermediate is the secondary structure that reveals the base pairings in a RNA sequence in two dimensions.

2.1 RNA Secondary Structure

A RNA secondary structure can be represented as an outer-planar graph i.e. a graph in which all vertices are arranged on a circle and all edges lie inside the circle and do not intersect. The nucleotides are represented by the vertices while the edges represent the backbone and the base pairing interactions. A formal definition of a secondary structure follows [80]:

> A secondary structure consists of a set of vertices $V = \{1, 2, ..., i, ..., N\}$ and a set of edges $E = \{i \cdot j, 1 \leq i < j \leq N\}$ fulfilling

(1) For $1 \leq i < n$, $i \cdot (i+1) \in E$.

(2) For each i there is at most one $h \neq i-1, i+1$ such that $i \cdot h \in E$.

(3) If $i \cdot j \in E$ and $h \cdot l \in E$ and $i < h < j$, then $i < l < j$.

The first condition simply means that RNA is a linear polymer, the second condition states that each base can be bonded to at most one other pairing partner, and the third forbids the formation of a three-dimensional or tertiary structure as it does not allow pseudo-knots and knots. While pseudo-knots are important structural elements in many RNA molecules [83], they are excluded from many studies mostly for a technical reason [80]. In the absence of pseudo-knots the folding problem for RNA can be solved efficiently by dynamic programming [80, 90]. Some efforts are taken to include pseudo-knots [1, 25] and in many cases they can be "added" to a predicted secondary structure

graph during a post-processing step. A current algorithm by Rivas [57] is able to deal with a large class of pseudo-knotted structures, but is extremely costly. In addition information about the energetics of pseudo-knots is still very limited [26].

A number of distinct structural motifs build the bases of secondary structures. Each secondary structure is composed of stacked base pairs, loops, and external elements which are neither part of a stack nor of a loop. Two stacked base pairs can be viewed as a special type of loop consisting of exactly four nucleotides. Five different types of loops can be distinguished. Hairpin loops and stacked base pairs have one and two base pairs, respectively. Bulges consist of two base pairs adjacent to each other and at least one unpaired base, interior loops have two base pairs which are not adjacent to each other and finally multi-loops which consist of three base pairs (see figure 7). The sum of energy contributions of all loops in a structure form the energy of a secondary structure. The energy contributions depend on the loop type, the composition of closing base pairs and interior base pairs as well as on the size of the loops.

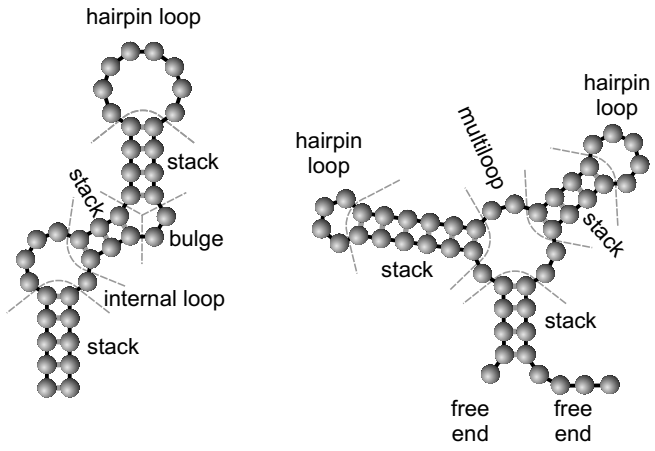


Figure 7: Secondary structures are composed of five distinct loop types namely stack, hairpin loop, interior loop, bulge and multi-loop. They build the basis of the additive energy model to calculate the energy of secondary structures. This calculation treats stacked pairs and bulges as special cases of interior loops. The contributed energies depend on the loop types, the composition of closing base pairs and interior base pairs as well as on the size of the loops.

2.2 Representations of Secondary Structures

Some commonly used graphical representations to visualise different properties of RNA secondary structures are discussed in the following.

- **The graph representation** shows the list of base pairs as a planar graph. It is the conventional representation in biochemistry. An example is given in figure 8.
- **The dot bracket notation** is more formalised. The nucleotides are symbolised by ".", "(" or ")". Where "." means an unpaired nucleotide, "(" a nucleotide that is paired with a base on its right side (opening of a base pair) and ")" a nucleotide that is paired with a base located on its left side (closing of a base pair). See figure 9 for an example of the dot bracket notation.
- **The circle representation:** [**51**] All bases are represented by a dot located on a circle. A chord connecting to dots denotes the pairing of the corresponding bases. If pseudo-knots are excluded none of the chords may cross another chord. Stacking regions are symbolised by groups of parallel chords as one can see in figure 10.
- **The tree representation:** permits a useful classification of structures according to their complexity [65,66,78,90]. Starting with a virtual root at the 5' end every node represents a unpaired nucleotide or a base pair. Every base pair leads to a lower level. Therefor a stack is represented by a vertical segment. Figure 11 shows an example tree.
- **The mountain representation** is directly derived from the dot bracket notation [34,41]. Every opening of a base pair (*or (*) increases the height of the mountain whereas a unpaired base (*or .*) results in keeping the current height constant and the closing of a base pair (*or)*) decreases the height. An example of the mountain representation can be found in figure 12. This representation is specially suitable to compare large RNA molecules [34, 41].

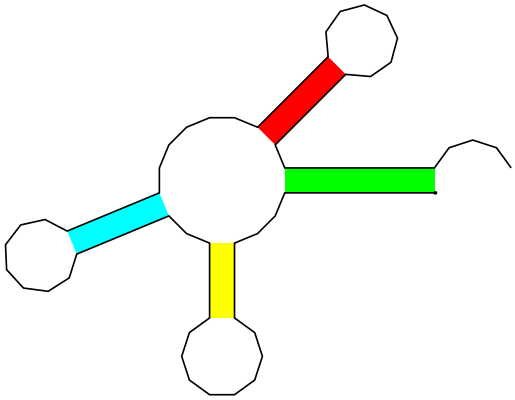


Figure 8: Secondary structure graph of the tRNA^{Phe} clover leaf structure. The stacks are marked with the same colour in every representation.

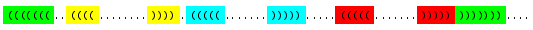


Figure 9: Dot bracket notation of the secondary structure of tRNA^{Phe}. The stacks are marked with the same color in every representation.

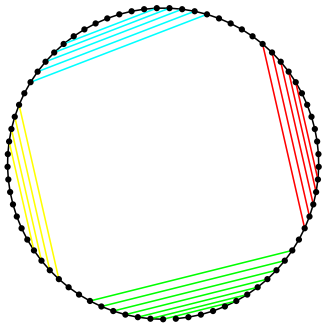


Figure 10: Circle representation of the secondary structure of tRNA^{Phe}. The stacks are marked with the same color in every representation.

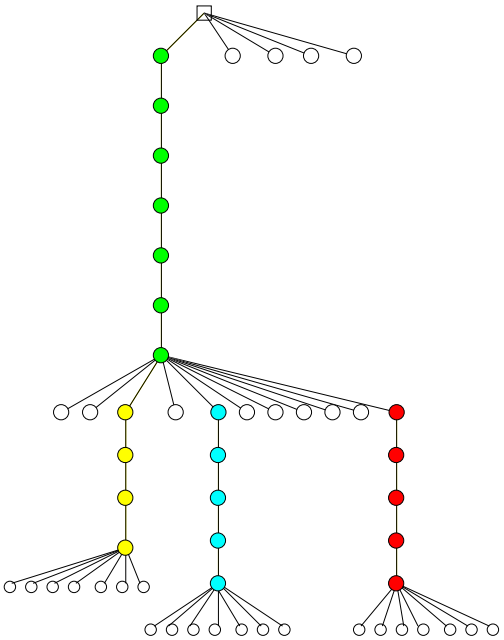


Figure 11: Tree representation of the secondary structure of tRNA^{Phe}. The stacks are marked with the same color in every representation.

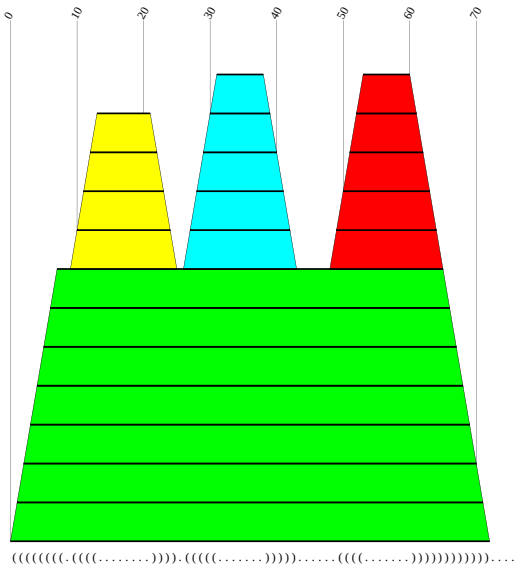


Figure 12: Mountain representation of the secondary structure of tRNA^{Phe}. The stacks are marked with the same color in every representation.

3 Algorithms

3.1 RNA secondary structure folding

The first efforts to find the secondary structure of a RNA strand by examining its sequence (primary structure) using dynamic programming were made by Ruth Nussinov [50, 51] in realising an algorithm to solve the maximum match problem. The output of this algorithm is the secondary structure that shows the maximum number of base pairs. The now standard energy model was formulated by Michael Zuker [90, 91] in his algorithm to find solutions to the minimum energy problem. The Vienna RNA package is an implementation of Zuker's algorithm plus many enhancements. MacCaskill's partition function algorithm and algorithms for inverse folding that suggest sequences folding into a given structure, calculating the specific heat of RNA sequences, calculating distances of RNA secondary structures and thermodynamic RNA secondary structure ensembles, calculating the energy of a RNA sequence on a given secondary structure, calculating suboptimal secondary structures of RNAs, and drawing RNA secondary structure graphs, make the Vienna RNA package a mighty tool for researchers working on the field of RNA. Each secondary structure can be uniquely decomposed into loops as discussed previously (see figure 7). A stacked base pair is considered a loop of size zero. The sum of the energy contributions of all loops is assumed to be the energy of a secondary structure. The minimum free energy (mfe) can be calculated recursively by dynamic programming [78, 80, 90, 91]. The structure leading to the mfe is retrieved later on by backtracking through the energy arrays. For individual loops the energy parameters have been determined in experiments [18, 37, 77] and were usually measured for T = 37 °C and $1M$ sodium chloride solutions. They depend on the loop type, loop size, and partly on its sequence. Only Watson-Crick pairs and GU and UG pairs are allowed since non-standard base pairs normally have context-dependent energy contributions. Those base pairs would not fit into the nearest neighbour model that states the energy contribution of a base pair in the interior of a helix to depend only on the

previous and the following pair.

Later John McCaskill [44] formulated an dynamic programming algorithm to calculate the partition function (see equation 1) and equilibrium probabilities over all possible base pairs.

$$Q = \sum_{i=1}^{n_S} \exp \frac{-\Delta G_0(S_i)}{RT} \tag{1}$$

The probability P_{kl} of each base pair k, j in the Boltzmann weighted ensemble of all structures can be calculated by this algorithm (see equation 2).

$$P_{kl} = P_{kl}^c + P_{kl}^i + P_{kl}^m \tag{2}$$

It is the sum of three independent terms:

- it closes a component with probability P_{kl}^c
- it is an interior base pair of an interior-loop, bulge, or stack with probability P_{kl}^i
- it is immediately interior to a multi-loop with probability P_{kl}^m.

The results are commonly visualised in a "dot plot" as shown in figure 13 where a square of the area P_{kl} represents the equilibrium frequency P for each base pair (k, l).

This square is plotted on position k, l in a two dimensional grid. The lower left triangular matrix shows the optimal fold whereas the upper right matrix shows the base pair frequencies within the structure ensemble at the thermodynamic equilibrium obtained by the partition function algorithm. This gives an impression of possible alternative foldings.

Memory becomes the bottleneck when folding larger molecules. A new implementation of the folding algorithm using parallel computers [10] can overcome this problem and allows the folding of a wider range of RNAs that where not accessible by the serial algorithm.

Both the standard energy model for calculating the minimum free energy structure and the partition function algorithm are implemented in the `Vienna`

`RNA package` [31, 32]. Other programs distributed with the package include `RNAsubopt` which is shortly described in chapter 3.4 and `RNAinverse` that derives its name from the inversion of the folding algorithm. It is a useful tool to find sequences folding into a specified structure starting from a given sequence or a random sequence. For each sequence found the Hamming distance [27] to the starting sequence is calculated which is the minimal number of point mutations required to convert two sequences into each other. The Hamming distance which is term that was originally introduced in information theory, is a important metric in the abstract sequence space.

If not stated differently, the `Vienna RNA Package` version 1.3 [31], which is freely available from `http://www.tbi.univie.ac.at/` was used to calculate mfe structures. This version uses the energy parameters taken from A. Walter

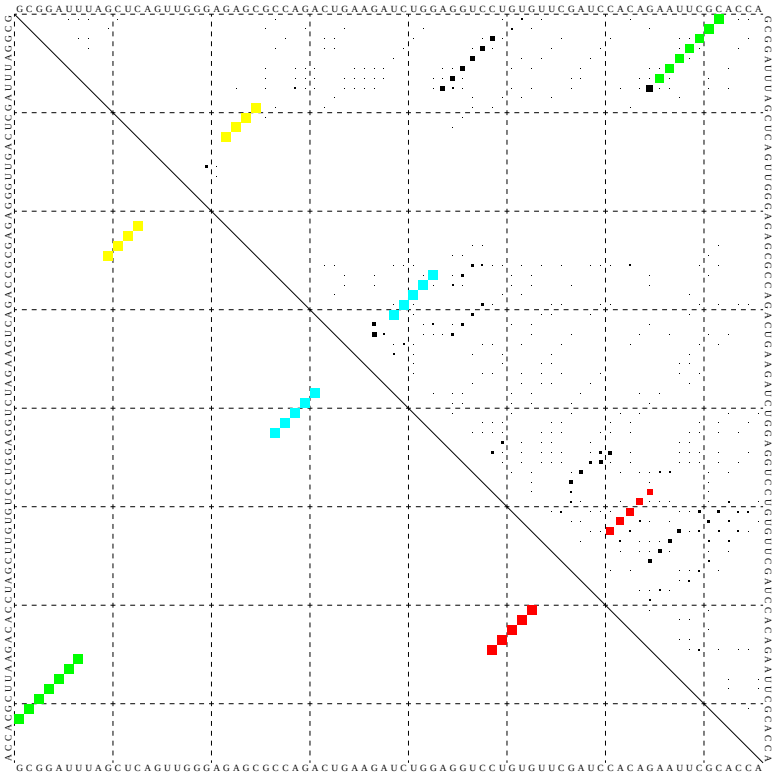


Figure 13: An example of a "dot plot" as generated by the partition function algorithm of the `Vienna RNA package`.

et al. [77]. To compare the results to the new parameters that were available only when this work was already ongoing, version 1.4 of the Vienna RNA Package was used. This new version implemented the most recent compilation of the energy parameters that can be found in Mathews et al. [43].

3.2 Components of a Neutral Net

The neutral net of a structure may be composed of several components. We are interested to obtain information about the number of the components and their sizes and composition.

First of all a suitable data structure is needed to store the sequences efficiently in terms of memory space and to perform fast searches. Different types and their advantages and disadvantages are discussed in the following.

3.2.1 Search Trees

Binary search trees

Binary search trees [5] use the following rules to store strings. For every node, all nodes down the left child have smaller values, while all nodes down the right child have greater values. To facilitate the imagination of a binary search tree an example is given in figure 14.

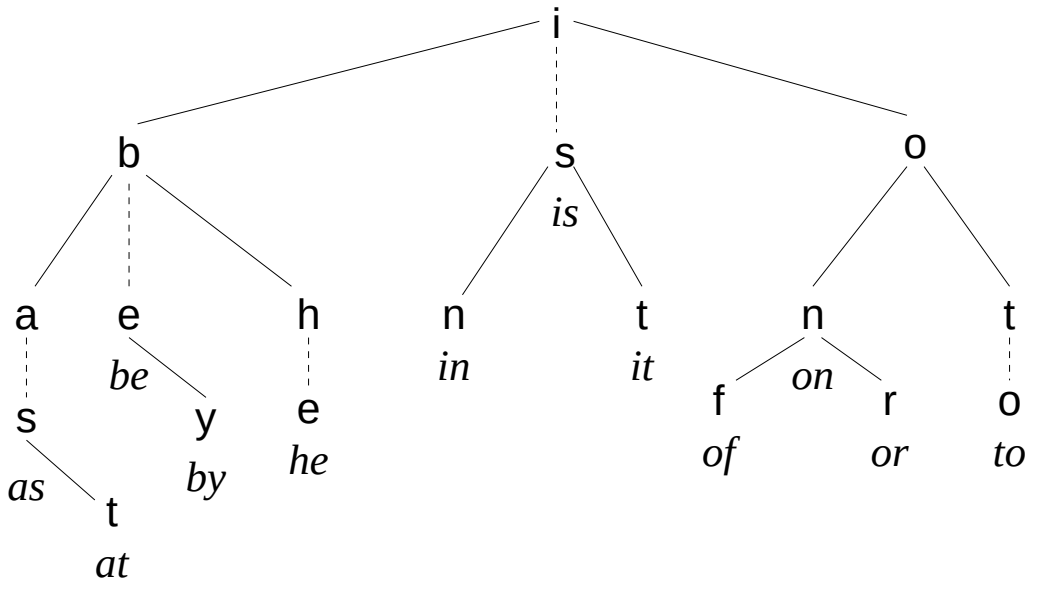


Figure 14: Example of a binary search tree. Every strings stored in the tree can be found below its last node in italic letters.

Digital search trees

Digital search trees or tries [5, 46] store data not at the nodes like binary trees but along the paths in the tree. In trees representing words of lowercase letters, each node has 26-way branching, with most branches are empty. To find a string, at every node we access one out of 26 array elements, test for null, and take a branch. An example is given in figure 15.

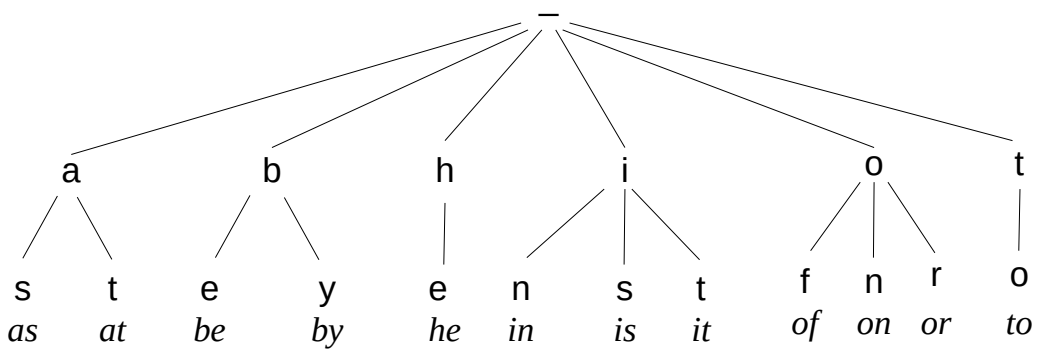


Figure 15: Example of a digital search tree. Only the relevant branches are shown here.

Ternary search trees

To find all neighbours ternary search trees [4, 5] were used. As digital search trees (or tries) they store strings character by character in contrast to binary search trees which store whole strings in each node. See figure 16 for an example of a ternary search tree.

They are space efficient, fast and capable to perform a near neighbour search on them which makes them a superior candidate for a sequence of components algorithm. They proceed character by character like tries and are space efficient like binary search trees but in contrast to binary search trees every node has 3 children (in binary search trees every node has 2 children). A search compares the current character in the search string with the character at the node. If the search character is less, the search goes to the left child; if the search character is greater, the search goes to the right child. when the

search character is equal, though, the search goes to the middle child, and proceeds to the next character in the search string. When searching for near neighbours of a given Hamming distance, all branches are searched until the given distance is reached or the node is null.

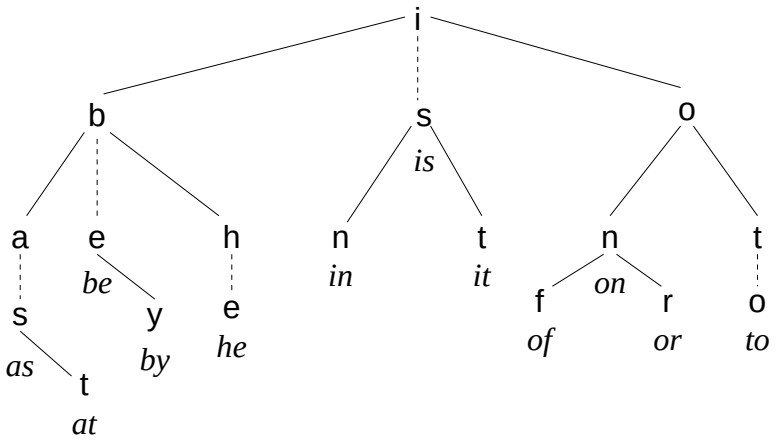


Figure 16: Example of a ternary search tree.

3.2.2 Algorithm to Obtain Components

When a convenient data structure is chosen the remaining part of the algorithm can be implemented. Earlier implementations used AVL trees and Btries which are tries over a binary alphabet and stored the sequences in a binary data or data base format respectively on disk [22–24]. For the reasons discussed previously the choice was made on ternary search trees. This new implementation utilises the near neighbour search of ternary trees and stores the sequences in a readable number format on disk. The *reduced* sequences are stored in files named of the structures they are folding into and before the search is performed in search trees. *Reduced* sequences are yielded by the function `reduce` that introduces new symbols for all possible base pairs. Each opening base or `(` in dot bracket notation is replaced by one of the new symbols according to the base pair they are affiliated with while bases at closing positions or `)` in dot bracket notation are not stored at all. This allows a reduction in disk space as well as in working memory space. On the other hand it enables a near neighbour search on the reduced sequences which yields not only the neighbouring sequences that differ by one base (Hamming distance $d_H = 1$) but also those that differ by on base pair ($d_H = 2$). Knowing the structure which is stored in the file name enables the algorithm to reconstruct the original sequences when they are need later.

Further space reduction is achieved by compressing the structure files with the `gzip` algorithm while they are saved to disk. Because of huge similarities of the sequences a reasonable compression ratio can be reached. The actions to reduce space are necessary because of the huge amount of sequences that has to be stored. The algorithm used to determine the sequence of components processes all immediate neighbours of a sequence before any remaining sequences are processed which is also called *breadth first traversal*. During a traversal the essential step involves finding all neighbours of the current node, which is implemented as a near neighbour search on a ternary search tree as already mentioned. The following iterative sequence of instructions describes this algorithm and shows how the neutral net of a structure is examined:

We use *algorithm 1* for nets that fit into memory as a whole (see also figure 17 for a graphical representation).

Algorithm 1:

1. Fold all sequences of a sequence space into the minimum energy structure. For each new structure found, create a file and store all the following sequences folding into this structure in this file.

   To save memory and search time, all pairing bases are substituted by one symbol representing this base pair.

   LIST is a list of all sequences folding into one structure.

2. Move a sequence TEMPSEQ from LIST into the current COMPONENTLIST.

3. Find all neighbours of the sequence TEMPSEQ in LIST and move them from LIST into the list of temporary sequences TEMPLIST.

4. Move a sequence TEMPSEQ from TEMPLIST into the current COMPONENTLIST.

5. Find all neighbours of the sequence TEMPSEQ in LIST and move them from LIST into the list of temporary sequences TEMPLIST.

6. Go back to 4 until TEMPLIST is empty.

7. The sequences stored in COMPONENTLIST now form the current component.

   Create a new COMPONENTLIST and make it the current one.

8. Go back to 2 until LIST is empty.

9. The complete sequence of components is obtained by enumeration of the sizes of all components.

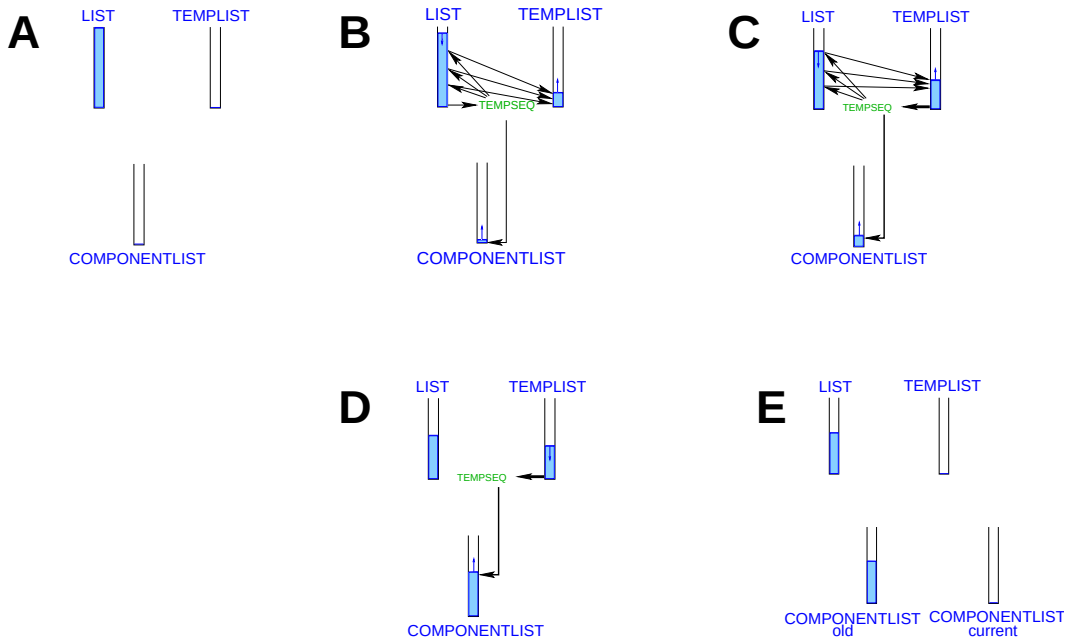


Figure 17: Algorithm 1 for calculating the sequence of components shown graphically: A: shows the initial state: the list (LIST) is filled with all sequences of an investigated neutral network, whereas the temporary list of sequences (TEMPSEQ) and of course the list of sequences belonging to the first component (COMPONENTLIST) are empty. B: a sequence (TEMPSEQ) is removed from LIST, all its neighbours are looked up in LIST and moved to TEMPLIST, and TEMPSEQ itself is pushed into COMPONENTLIST. LIST is shrinking whereas TEMPLIST and COMPONENTLIST are growing. C: a sequence (TEMPSEQ) is removed from TEMPLIST, all its neighbours are looked up in LIST and moved to TEMPLIST, and TEMPSEQ is pushed into COMPONENTLIST. This step is repeated and LIST keeps shrinking, whereas TEMPLIST and COMPONENTLIST keep growing until no neighbour of TEMPSEQ is found in LIST. D: if no neighbour of TEMPSEQ is found in LIST it is moved into COMPONENTLIST. This time TEMPSEQ is shrinking. E: if it happens that TEMPLIST is empty an new COMPONENTLIST is created and made the current one and the algorithm jumps to step B.

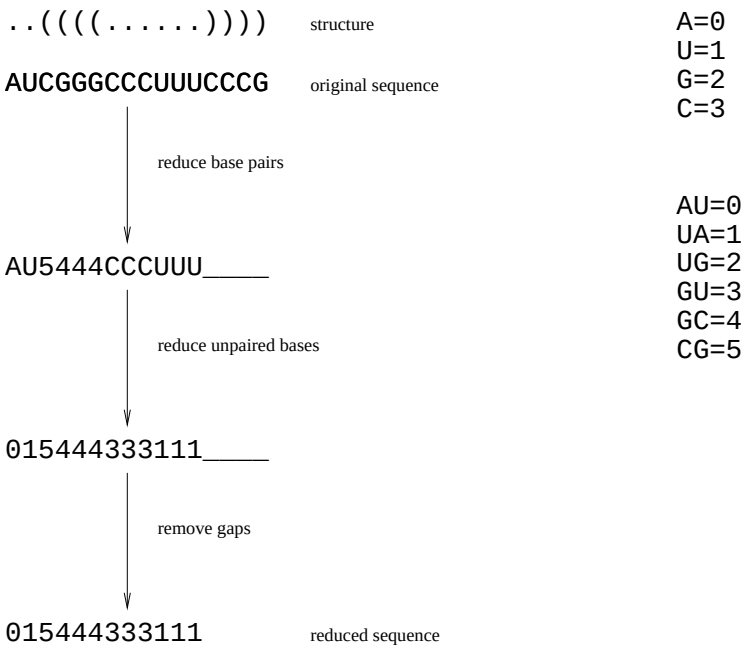


Figure 18: An example on how the `reduce` function processes sequences.

For large structure files not fitting into memory as a whole, a modified algorithm (*algorithm 2*) is used which processes the those files slice by slice. The size of the slices can be adapted according to the memory available on the computer in use. The modified algorithm works in two steps. First, one slice is read into memory and decomposed in the standard manner of algorithm 1. This is performed on all slices. As a matter of fact two components belonging to different slices may be connected in the whole network (figure 19). To find these connections a similar method to algorithm 1 is used in a second step. All slices are searched for connections of their components to components of different slices. Practically these means if two sequences of different components are found to be neighbours, the two components are connected and the search can be aborted. By joining the connected components this second stage yields the component structure of the complete network. This algorithm by-passes the problem of limited memory but the size of the slices has still to be adjusted to fit two slices into memory at the same time.

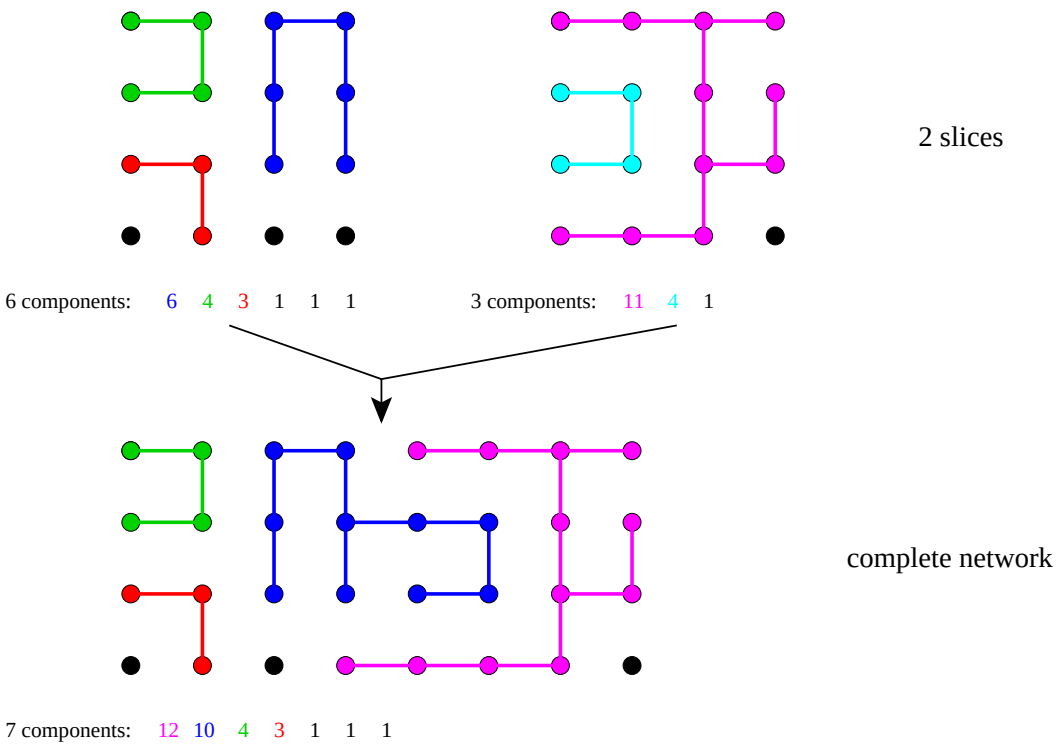


Figure 19: Graphical representation of algorithm 2 for an example neutral network. Each sequence is represented by a filled circle. Sequences belonging to one component are connected by a line of the same colour. First the components of the slices are determined. In a second step the components of the complete network are found by joining connected components belonging to different slices. Each sequence of components is shown below the corresponding network listing the size of each component in the respective colour.

Algorithm 2:

1. Push the maximum allowed number of sequences into LIST.
2. Calculate the sequence of components of LIST using algorithm 1.
3. Write each component of this slice into an own file.
4. Push all file names of the currently created files into FILENAMELIST.
5. Move a file name TEMPFILENAME from FILENAMELIST into the current COMPONENTLIST.
6. Find all neighbouring files of the file TEMPFILENAME in FILENAMELIST and move them from FILENAMELIST into the list of temporary file names TEMPFILENAMELIST.

   Finding all neighbouring files in FILENAMELIST means:

   (a) Load a file CURRENTFILENAME of FILENAMELIST into memory.

   (b) Test each sequence in TEMPFILENAME file for neighbouring sequences in CURRENTFILENAME. The two files are known to be neighbours after the first pair of neighbouring sequences is found. The rest of the sequences in both files need not be looked at.

   (c) Go back to 6a until all files in FILENAMELIST are tested.
7. Move a file name TEMPFILENAME from TEMPFILENAMELIST into the current COMPONENTLIST.
8. Find all neighbours of the file TEMPFILENAME in FILENAMELIST and move them from FILENAMELIST into the list of temporary file names TEMPFILENAMELIST.
9. Go back to 7 until TEMPFILENAMELIST is empty.

10. The sequences stored in the files of COMPONENTLIST now form the current component.

    Create a new COMPONENTLIST and make it the current one.

11. Go back to 5 until FILENAMELIST is empty.

12. The complete sequence of components is obtained by enumeration of all components.

The main difference between the two algorithms is that FILENAMELIST and TEMPFILENAMELIST in algorithm 2 contain file names instead of sequences which were stored in LIST and TEMPLIST in algorithm 1. Step 5 to 11 in algorithm 2 are essentially the same as step 2 to 8 in algorithm 1. For the lists in both algorithms a ternary search tree structure was used and a near neighbour search was performed on them to find neighbouring sequences. Without the possibility of the near neighbour search it would be necessary to create all mutants of the current sequence which are compatible with the structure under investigation in a first step. In a second step each of those candidate neighbours would have to be looked up in the set of sequences of the investigated network in order to find all neighbours. These two steps are obsolete when you are using ternary search trees which are competent in performing near neighbour searches.

The algorithms to obtain the sequence of components was implemented using the mighty Perl [76] programming language.

3.3 Kinetic Folding of RNA

RNA folding is modelled as a Markow process in conformation space using a given microscopic move set. Built up on these fundamentals Christoph Flamm implemented his `kinfold` [12] program to simulate the kinetic folding of RNAs into their minimum free energy secondary structures. This section describes the background and algorithm the program is based on.

To control the movement in the multidimensional conformation space $\mathcal{C}$ a set of rules is needed that defines the allowed moves. This set of rules is called move set. It is a number of operations that is needed to transform one element of $\mathcal{C}$ into another one. A move set defines the possible conformational changes that are allowed in a single step during the simulation of kinetic folding. Therefore it defines the conformational space.

A trajectory is a sequence of consecutive states of the state space that is generated by a series of legal operations from some initial state. Whereas a folding path is a cycle free trajectory, which means that each state occurs only once in the sequence of consecutive states.

Insertion and *deletion* of a single base pair form the most simple move set. Using these operations it is always possible to construct a path from any element S_j of the conformation space $\mathcal{C}$ to another element S_k of $\mathcal{C}$. The path of minimum length between S_j and S_k is found by removing all base pairs in S_j that do not occur in S_k followed by the insertion of all base pairs of S_k that do not occur in S_j. This move set can describe a frequent mechanism in helix formation called *zipper mechanism* [55]. New base pairs are stacked to a nucleus which is associated with favourable negative energy contributions. This leads to a spontaneous gradual growth of the helix that can be compared to the process of closing a zipper.

Another important mechanism is a called *defect diffusion*. Defect diffusion is much faster than zippering [55]. It can anneal mismatched helices that are the result of incomplete base pairing by a fast chain slide mechanism. As a result a bulge loop can rapidly move from one end of a helix to its opposite end leading to a shift of the two strands of the helix by one nucleotide. To reflect

this mechanism a further move called *shift* is introduced. The outcome of this move is a transformation from one base pair to a new one in a single step. The shift move also is conductive to the conversion of overlapping helices into each other. This is especially true for two helices within a multi-loop which would be energetically unfavourable using only insertions and deletions.

By adding the shift move to our basic move set we have determined the fundamentals of a realistic kinetic folding algorithm. See figure 20 for the complete move set.

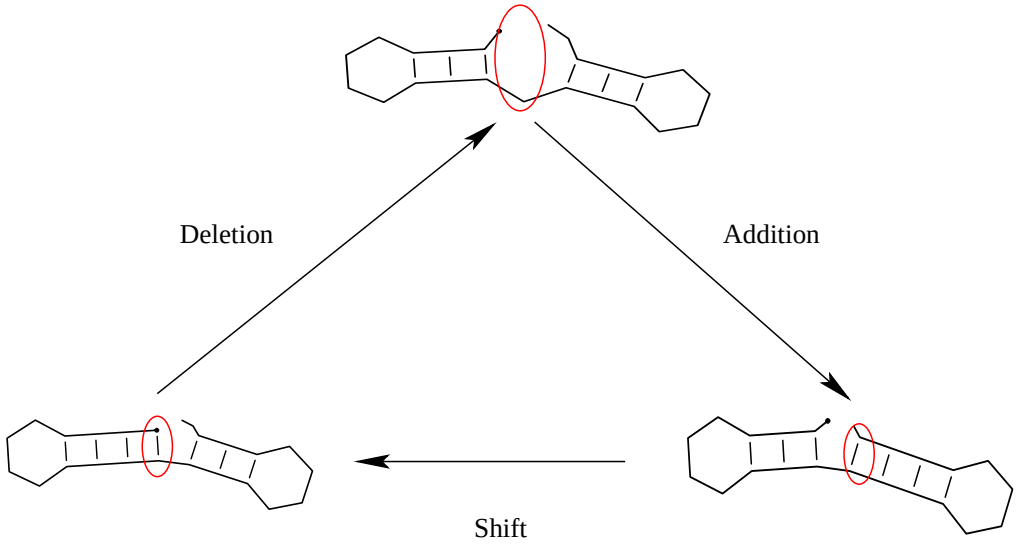


Figure 20: The moves allowed by the `kinfold` program. Starting at the upper structure in a clockwise order we first introduce a base pair by insertion followed by a shift move which reduces one stacking area while elongating another by one base pair in a single step. The next steps reveals the original structure by deleting one base pair. The changes made by each move are marked in red.

3.3.1 The Model and Algorithm of Kinetic Folding

The physical model of the folding process can be described in the form of a Markow chain which is a random walk in an N-dimensional state space with a very short memory of only one step. The following items underlie such a Markow chain:

Some conformational changes are more likely and therefore happen more frequently than others. A transition probability law controls the moves between states of the RNA chain in the conformation space.

The selected move sets determine the resolution of a folding trajectory. A higher resolution can be obtained by choosing a move set that produces only small conformational changes when applied. The higher resolution also leads to more detailed and longer trajectories.

Conformational changes lying "far" apart on the trajectory seem to happen independently. The apparently molecule does not memorise what happened earlier.

On a large time scale the movement of the chain within conformation space seems to be arbitrary.

According to conventional stochastic kinetics of chemical reactions [19] the probability $P(i,t)$ that a given RNA molecule will have the secondary structure i at time t is given by the master equation

$$\frac{dP(i,t)}{dt} = \sum_j [P(j,t)k_{ji} - P(i,t)k_{ij}] \tag{3}$$

where k_{ij} is the rate constant.

situation described by equation 3 is numerically simulated in `kinfold`. The algorithm used in `kinfold` is based on a continuous time Monte Carlo method proposed by Daniel Gillespie [21]. It calculates the distribution of first passage times which represent the folding times from an initial state to the thermodynamic ground state.

The Gillespie method is an efficient procedure at high rejection rates. Instead of rejections this method uses an internal clock. At each step the rate constants to all neighbours are computed and the time is advanced by a time

increment which is adjusted to the sum of the rate constants. At the bottom of a deep local minimum a higher rejection rate is found and the internal clock is advanced further than at a saddle point of the energy landscape.

A symmetric rule introduced by Kyozi Kawasaki [38] is used to calculate the rate constant k_{ij}, which characterises the transition between the two conformations i and j. This transition probability is formulated as:

$$k_{ij} = \exp \frac{-\Delta G}{2kT} \tag{4}$$

To simulate the stochastic process of folding not only the rate constants but also random numbers are involved at each step to choose the structure for the next step.

At the simulation each step involves the following phases:

- Generation of the set of legal neighbour structures according to the used move set
- Calculation of the rate constants from the current state to all its neighbours
- Drawing of random numbers and selection of a move
- Calculation of the time increment and advancing the clock

The simulation is a stochastic process. Therefore several simulations using the same initial conditions have to be carried out to show a realistic picture of the distribution of folding times.

3.4 Energy Barriers

The complex surface of the free energy versus the conformational degrees of freedom is called the energy landscape of a RNA molecule. The degrees of freedom are fixed by the allowed transformations of a move set whereas the allowed conformations are the secondary structures compatible to a particular sequence. An energy landscape can be represented by plotting the energy of a conformation according to the standard energy model over conformation space. To describe the shape of such an energy landscape we need to know all possible secondary structures within a certain energy range which can be generated by suboptimal folding techniques.

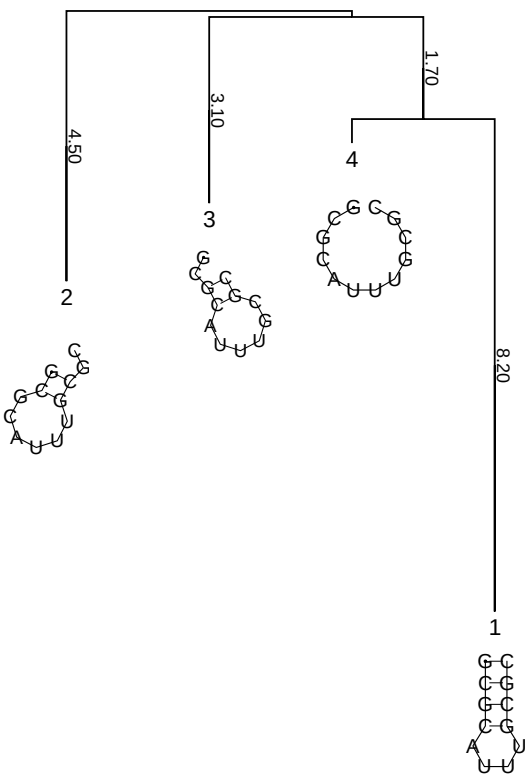
Stefan Wuchty's `RNAsubopt` [87] is a program that calculates all suboptimal secondary structures within a chosen energy range. It is based on the Waterman-Byers scheme [79] which was originally developed to find suboptimal solutions to the shortest path problem in networks.

Using all suboptimal structures within a given energy range and a move set it possible to explore topological details like local optima and saddle points.

A local minimum is a structure that has a lower energy than all legal neighbouring structures whereas a structure that has a higher energy than all legal neighbouring structures is called local maximum. Besides these local optima another characteristic describes a topology: saddle point.

A saddle point in a narrow sense is defined as a secondary structure that if used as a starting point allows to reach two local minima by downhill walks. Of special interest is the saddle point with the lowest energy that separates the basins of two local minima.

Using a flooding algorithm on the energy landscape reveals those saddle points. Together with the local minima they connect they can be presented in a tree representation. The resulting trees of local minima also show the barrier heights between two local minima and give an impression of the ruggedness of the energy landscape. A simple example of such a tree of local minima is given in figure 21. The key criterion that forms the topology of an energy landscape is the choice of the move set. A different move set changes the connectivity of




```
  GCGCAUUUGCGC
1 ((((....)))) ( -7.80)    0  17.70
2 ((......)).. ( -2.30)    1   4.50 .(......)...
3 ..((......)) ( -1.00)    1   3.10 ...(......).
4 ............ (  0.00)    1   0.40 ...(....)...
```

Figure 21: Example of a barrier tree produced by the program `barriers`. The structures of the local minima are drawn below the corresponding leafs of the tree. On the bottom the text output of `barriers` is shown. The first line lists the sequence. The following lines print the number of the local minimum, the structure, the energy, the local minimum it merges with, the barrier hight to the local minimum it merges with and the saddle structure. The number of the local minimum equals the rank when sorted by the energy. The minimum free energy structure is always number 1.

local optima and herewith the barrier heights, too.

A program that uses the flooding algorithm to calculate energy barriers and the resulting barrier trees is `barriers` [12, 13] implemented by Christoph Flamm. This program is available upon request from its author and was used

to create the barrier trees in this work. An example output of `barriers` is given in figure 21 while figures 22 to 24 visualise the flooding algorithm in a time series. To illustrate the connection between the branch lengths and the free energy according to figure 22 the barrier trees shown in this work were rotated to view the free energy at the ordinate.

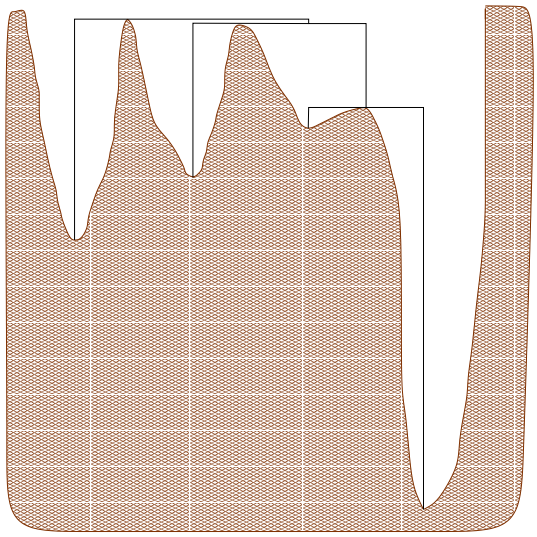


Figure 22: A visualisation of an energy landscape and its connection to the corresponding barrier tree using the same simple example found in figure 21. The barrier tree shown in black reveals four local minima and three saddle points between them. The local minima can be imagined as the deepest points of their basins. Those basins form the energy landscape shown as the brown contour of valleys and saddles.

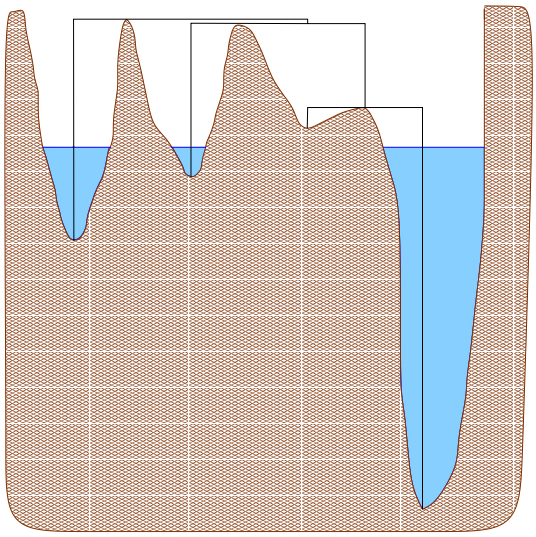


Figure 23: The landscape is flooded starting at the lowest energy level by continuously increasing the energy level.

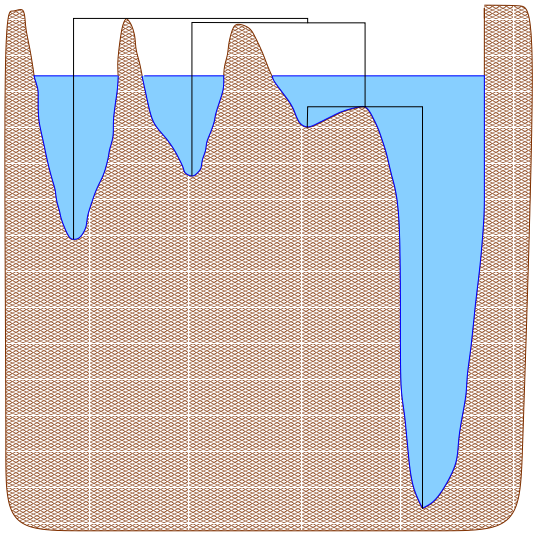


Figure 24: The two basins on the right side have merged while the landscape is continuously flooded until the given upper energy limit is reached where ideally all basins should have merged.

4 Computational Results

4.1 Sequence of Components

If not stated differently, the `Vienna RNA Package` version 1.3 [31], which is freely available from `http://www.tbi.univie.ac.at/` was used to calculate mfe structures. This version uses the energy parameters taken from A. Walter et al. [77].

To compare the results to the new parameters that were available only when this work was already ongoing, version 1.4 of the Vienna RNA Package was used. This version uses the parameters taken from Mathews et al. [43]. Only selected smaller sized sequence spaces were additionally folded using the newer version 1.4 of the Vienna RNA Package and its updated parameters because of the extremely time intensive folding and decomposition into components of whole sequence spaces of higher chain lengths.

The results from exhaustive folding of complete sequence spaces are discussed in the following. Different alphabets and chain lengths were examined. The feasible limits for the time consuming process of exhaustive folding were chain length 30 for two letter alphabets (AU and GC), chain length 20 for three letter alphabets (AUG and UGC) and chain length 16 for the complete four letter alphabet. Version 1.3 of RNAfold was used as this was the latest version available at the beginning of this work. Only for reasons of comparison selected smaller sized sequence spaces of the same alphabets were additionally folded using version 1.4 of RNAfold and its updated parameters.

Looking at figure 26 we find the fraction of sequences that fold into the open chain decreasing with the chain length so that at large chain lengths nearly all structures fold into a stable non-open-chain mfe structure.

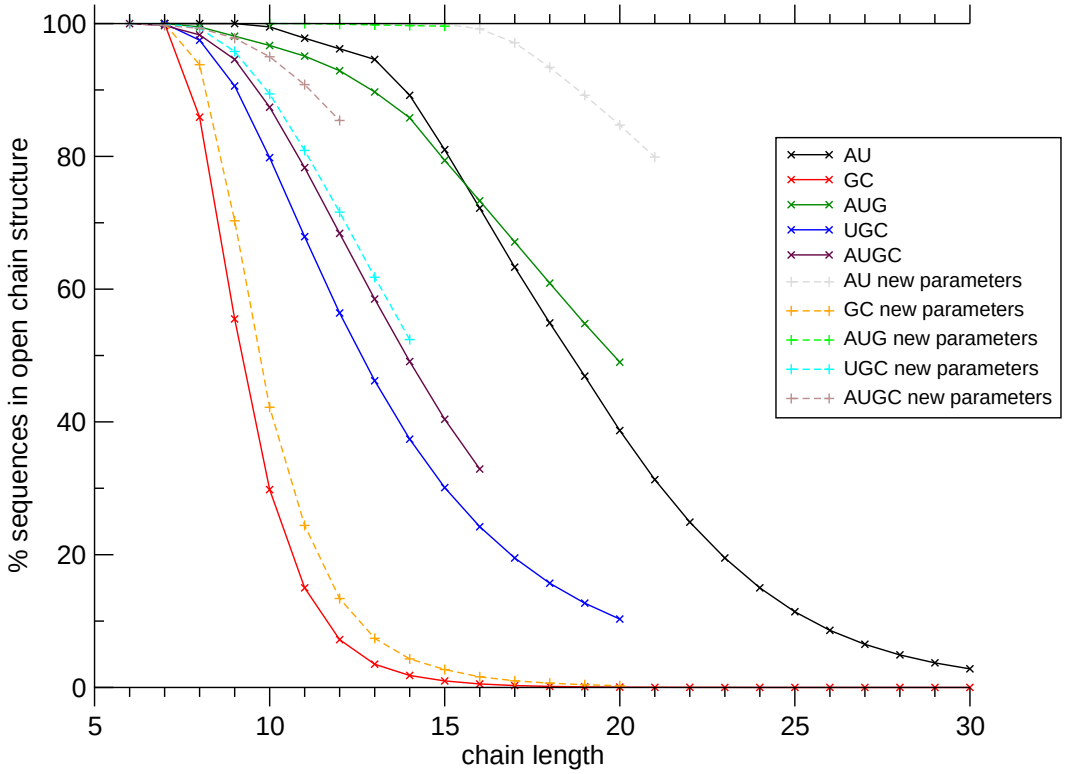


Figure 25: The percentage of sequences folding into the open chain structure for the complete sequence spaces of different alphabets and chain lengths computed using version 1.3 of RNAfold drawn in continuous lines. Only for reasons of comparison selected smaller sized sequence spaces of the same alphabets were additionally folded using version 1.4 of RNAfold and its updated parameters. They are marked as 'new parameters' in the legend and drawn in dashed lines.

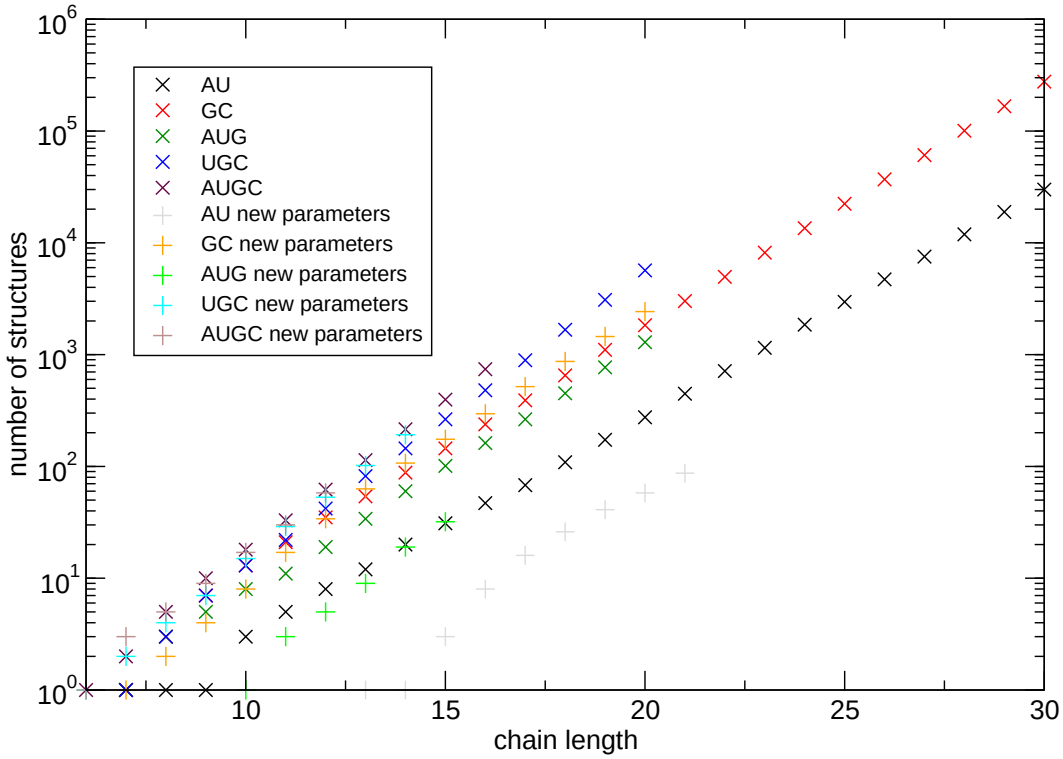


Figure 26: The number of structures as functions of the chain length. As in figure 25 The different alphabet sequence spaces were folded using version 1.3 of `RNAfold`. Again some examples were folded using version 1.4 of `RNAfold` and are label as 'new parameters'.

An upper bound to the number of folded structure was derived for a minimum number of three unpaired digits and a minimum stem length of two, i.e. counting only structures that do not contain isolated base pairs [33].

$S_n \sim 1.4848 \times n^{-3/2}(1.8488)^n$

The number of minimum free energy structures obtained from exhaustive folds can be extrapolated to estimate the expected number of structures for higher chain lengths. The estimates were obtained by a fit to the data shown in figure 26. The exact numbers can be found in tables 1-2 and tables 3-4 for parameters of version 1.3 and version 1.4 of `RNAfold` respectively. Version 1.3 foldings gave the following estimates where S_n is the number of structures for chain length n:

	AU			GC		
n	S	r_c	n_c	S	r_c	n_c
7	1	1	128	1	1	128
8	1	1	256	3	1	220
9	1	1	512	7	1	284
10	3	1	1019	13	4	635
11	5	1	2003	21	7	1293
12	8	1	3942	35	14	3115
13	12	1	7751	54	22	6773
14	20	1	14621	88	33	14276
15	31	1	26535	145	47	28653
16	47	1	47309	238	65	56370
17	68	4	89636	390	98	113111
18	109	19	198891	652	155	231406
19	173	44	436896	1104	231	463297
20	275	85	961700	1833	347	923722
21	447	114	1901674	3013	551	1861908
22	713	151	3725409	4963	855	3735008
23	1150	228	7538434	8169	1309	7469062
24	1852	332	15111655	13516	2079	15054472
25	2965	472	30223549	22351	3336	30493674
26	4713	668	60224307	36947	5134	61366034
27	7528	903	118957582	60894	7808	123090782
28	11900	1331	236869036	100634	12118	247327396
29	18898	2124	476408741	166804	18960	497464140
30	29950	3505	964766482	276569	29369	998818207

Table 1: The number of structures S including the open chain structure, the rank of the rarest common structure r_c and the number of sequences that fold into common structures n_c for different chain lengths n calculated by using the parameters of `Vienna RNA Package` version 1.3 .

	AUG			UGC			AUGC		
n	S	r_c	n_c	S	r_c	n_c	S	r_c	n_c
6							1	1	4096
7	1	1	2187	1	1	2187	2	1	16338
8	3	1	6526	3	1	6397	5	1	64442
9	5	1	19312	7	1	17833	10	1	247921
10	8	1	57116	13	1	47099	18	1	916038
11	11	1	168481	22	1	120233	33	1	3285249
12	19	1	493799	42	5	361683	62	4	12420860
13	34	1	1430221	82	22	1332692	114	25	58585513
14	60	1	4068148	145	39	4209228	215	46	245890150
15	101	2	11535339	264	67	13360447	396	74	1005111947
16	162	12	34885205	480	94	39877953	741	114	4075143785
17	264	40	114167811	891	140	119620338			
18	451	75	358073568	1669	208	355954349			
19	770	116	1096399800	3089	307	1050754221			
20	1291	156	3270115077	5668	558	3162248955			

Table 2: The number of structures S including the open chain structure, the rank of the rarest common structure r_c and the number of sequences that fold into common structures n_c for different chain lengths n calculated by using the parameters of `Vienna RNA Package` version 1.3 .

$$S_n^{\mathrm{AU}} \sim 0.949778 \times n^{-3/2}(1.67431)^n$$
$$S_n^{\mathrm{GC}} \sim 2.48359 \times n^{-3/2}(1.74616)^n$$
$$S_n^{\mathrm{AUG}} \sim 0.656429 \times n^{-3/2}(1.82938)^n$$
$$S_n^{\mathrm{UGC}} \sim 0.483069 \times n^{-3/2}(2.00026)^n$$
$$S_n^{\mathrm{AUGC}} \sim 0.434356 \times n^{-3/2}(2.06493)^n$$

Version 1.4 foldings gave less reliable results because of the smaller number of sequence spaces that were available:

$$S_n^{\mathrm{AU}} \sim 0.192778 \times n^{-3/2}(1.66434)^n$$
$$S_n^{\mathrm{GC}} \sim 1.32702 \times n^{-3/2}(1.82288)^n$$
$$S_n^{\mathrm{AUG}} \sim 0.042775 \times n^{-3/2}(2.03991)^n$$
$$S_n^{\mathrm{UGC}} \sim 0.249861 \times n^{-3/2}(2.13299)^n$$
$$S_n^{\mathrm{AUGC}} \sim 0.242539 \times n^{-3/2}(2.1529)^n$$

Due to the different valuation of the energy contributions, the two examined

	AU			GC		
n	S	r_c	n_c	S	r_c	n_c
7				1	1	128
8				2	1	240
9				4	1	360
10				8	3	697
11				17	5	1311
12				34	15	3491
13	1	1	8192	63	23	7027
14	1	1	16384	107	31	13376
15	3	1	32720	175	44	26253
16	8	1	65036	296	76	54875
17	16	1	127229	517	133	114578
18	26	1	244953	870	201	229413
19	41	1	467867	1453	311	461874
20	58	1	888362	2427	457	917432
21	87	1	1674658			

Table 3: The number of structures S including the open chain structure, the rank of the rarest common structure r_c and the number of sequences that fold into common structures n_c for different chain lengths n calculated by using the parameters of `Vienna RNA Package` version 1.4 .

parameter sets lead to a different number of structures formed. For example the unfavoured A–U pairs at the ends of a stack are often missing when using the new parameters leading in many cases to less but more realistic and stable structures for a given chain length. Moreover the first structure other than the open chain occurs at a longer chain length.

To compare these estimates to earlier calculations the results from Walter Grüner et al. [23] which were calculated using an older parameter set are given:

$S_n^{\mathrm{AU}} \sim (0.0097 \pm 0.0038) \times (1.489 \pm 0.0029)^n$

$S_n^{\mathrm{GC}} \sim (0.0853 \pm 0.0009) \times (1.6360 \pm 0.0007)^n$

	AUG			UGC			AUGC		
n	S	r_c	n_c	S	r_c	n_c	S	r_c	n_c
6				1	1	729	1	1	4096
7				2	1	2186	3	1	16340
8				4	1	6516	5	1	65067
9				7	1	18861	9	1	256497
10	1	1	59049	15	1	52762	17	1	995920
11	3	1	177105	29	1	143228	30	1	3806358
12	5	1	531020	53	6	440750	58	1	14325304
13	9	1	1591635	102	15	1350733			
14	19	1	4770376	192	35	4257662			
15	32	1	14287329						

Table 4: The number of structures S including the open chain structure, the rank of the rarest common structure r_c and the number of sequences that fold into common structures n_c for different chain lengths n calculated by using the parameters of `Vienna RNA Package` version 1.4 .

A structure is said to be common if its preimage is not smaller than the average size of a neutral network [60]. The average size of a neutral network is calculated by the number of sequences of a sequence space over the number of structures S of a sequence space. The number of sequences α^n is determined by the chain length n and the size of the alphabet α of a sequence space. Hence we can determine the rank of the rarest common structure r_c which means that the structure of rank r_c is still common, while the structure of rank $r_c + 1$ is already rare.

The number of structures S the rank of the rarest common structure r_c and the number of sequences that fold into common structures n_c are listed in tables 1-2 and 3-4 for the parameters of version 1.3 and 1.4 of `RNAfold`, respectively.

The fraction of common structures r_c/S decreases with an increasing chain

length. Looking at figure 27 we can see a short but abrupt decrease followed by a short increase ending in a constant decrease at longer chain lengths especially clear for the AU and the GC alphabet. This cannot be said for sure for all alphabets as some of them were only examined on shorter chain lengths but they all seem to follow this trend.

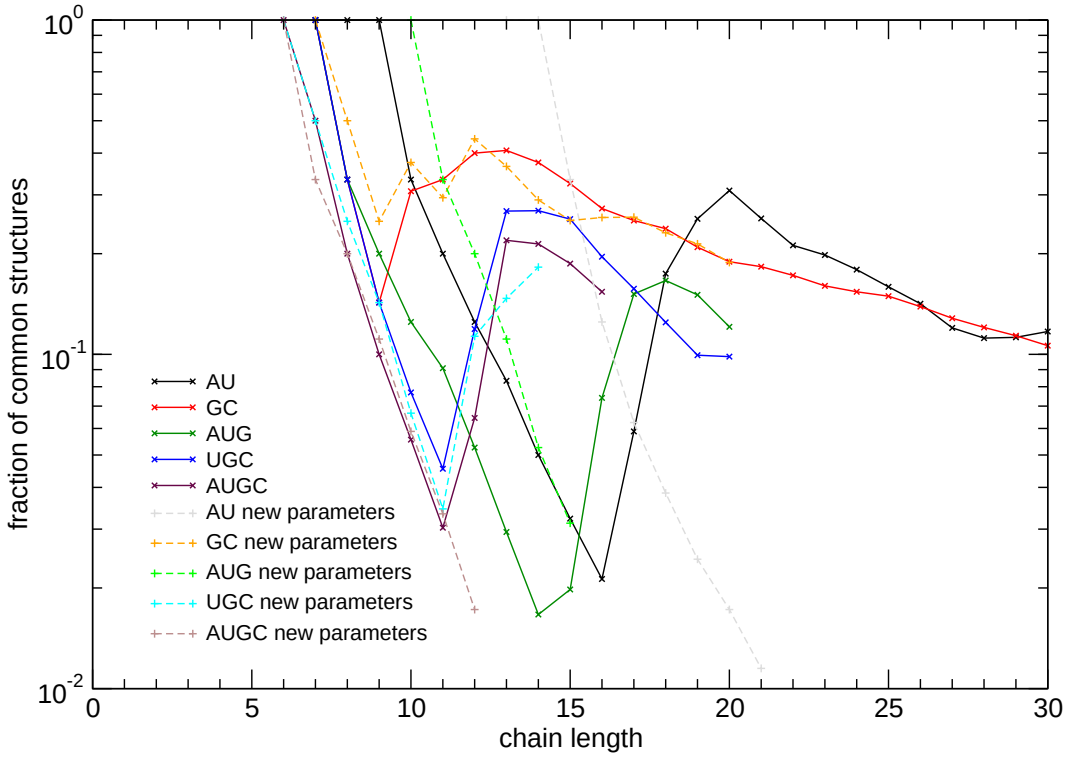


Figure 27: The fraction of common structures r_c/S versus the chain length n for the complete sequence spaces of different alphabets computed using version 1.3 of RNAfold drawn in continuous lines. Only for reasons of comparison selected smaller sized sequence spaces of the same alphabets were additionally folded using version 1.4 of RNAfold and its updated parameters. They are marked as 'new parameters' in the legend and drawn in dashed lines.

On the other hand the fraction of sequences that fold into common structures n_c/α^n increases with the chain length n. Figure 28 shows and strong decrease followed by a abrupt increase ending in a constant increase with increasing chain lengths clearly for the AU and the GC alphabets. Again data for large ns are missing for some other alphabets but as far as we can see we can assume a similar behaviour. At large chain length almost all sequences fold into common structures. Neither the used alphabet nor the folding parameter set seems to influence these general properties.

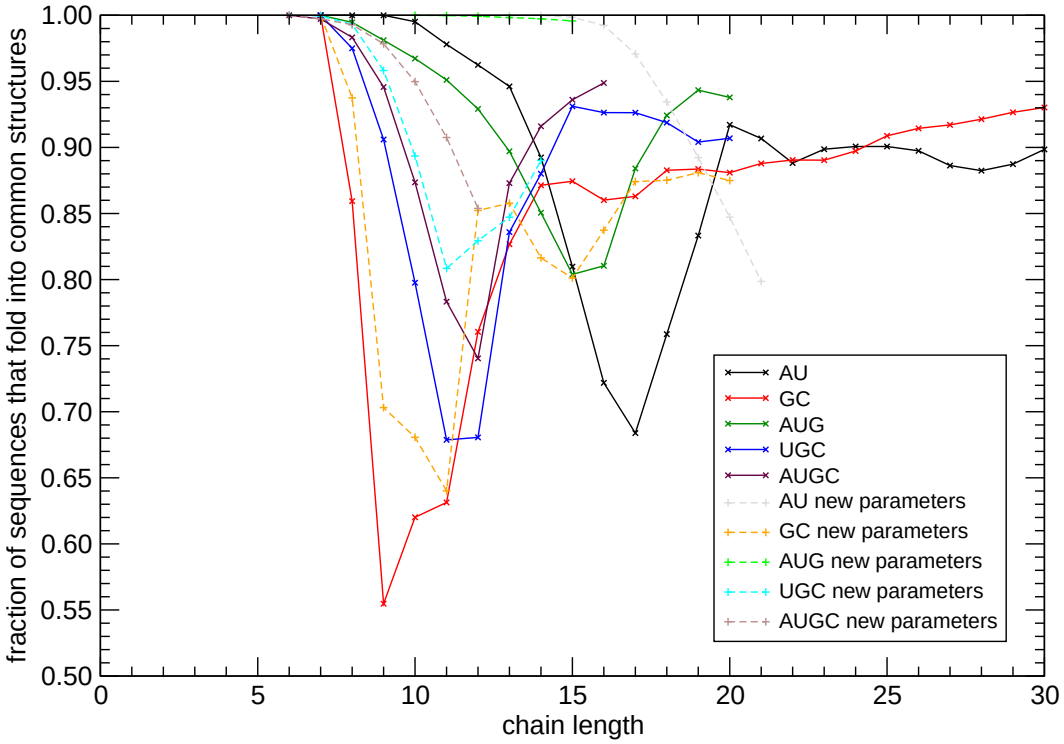


Figure 28: The fraction of sequences that fold into common structures n_c/α^n versus the chain length n for the complete sequence spaces of different alphabets computed using version 1.3 of RNAfold drawn in continuous lines. Only for reasons of comparison selected smaller sized sequence spaces of the same alphabets were additionally folded using version 1.4 of RNAfold and its updated parameters. They are marked as 'new parameters' in the legend and drawn in dashed lines.

RNA secondary structures behave like words in a natural language. There are a few common and may rare ones [23,68] which is known as Zipf's law [88]. In its simplest form

$F(r) = \frac{C}{r}$

it says that the frequency F of a word times its rank r is equal to a constant. A more generalised version of the law was proposed by Mandelbrot [42]:

$F(r) = \frac{C}{(r+b)^{\alpha}}$.

The structures obtained by exhaustive folding are ranked according to their frequencies. The ranking yields a distribution which follows a generalised Zipf's law

$F(r) = C(r+b)^{\alpha}$,

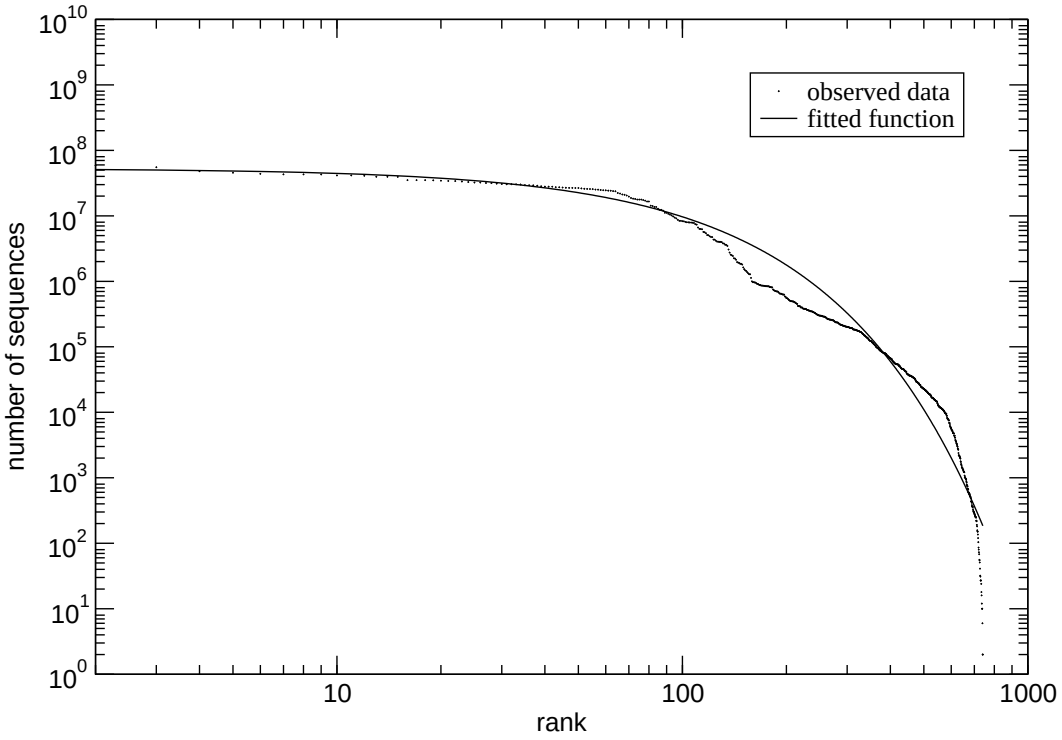


Figure 29: A fit to rank ordered network sizes of AUGC16. A generalised Zipf's law was formulated as $y = a * (1 + x/b)^{-c}$ as the fit function and the variables were fitted to $a = 5.26756 \times 10^7$, $b = 4.45264 \times 10^8$ and $c = 7.54683 \times 10^6$ using `xmgrace`'s non-linear fit function. The open chain (rank 1) is not shown in this figure.

where r and $F(r)$ are the rank and the frequency of the corresponding structure, respectively [69]. The constant C is a normalisation factor, b can be interpreted as the number of "very frequent" structures. The constant α determines the slope of the tail of the distribution. Distributions following this form of a generalised Zipf's law were found for all algorithms, parameter sets, and alphabets.

To give an overview on the number of components of whole sequence spaces, the *mean component size* $= \frac{size\ of\ network}{number\ of\ components}$ was calculated. This does not say much about the size of a component we are expecting, as there are often a few giant and many tiny components. The tiny components often consist of less than 0.1% of the neutral net and sometimes networks can be as small as only a single sequence. But we can see how split the networks are (see figure 30)

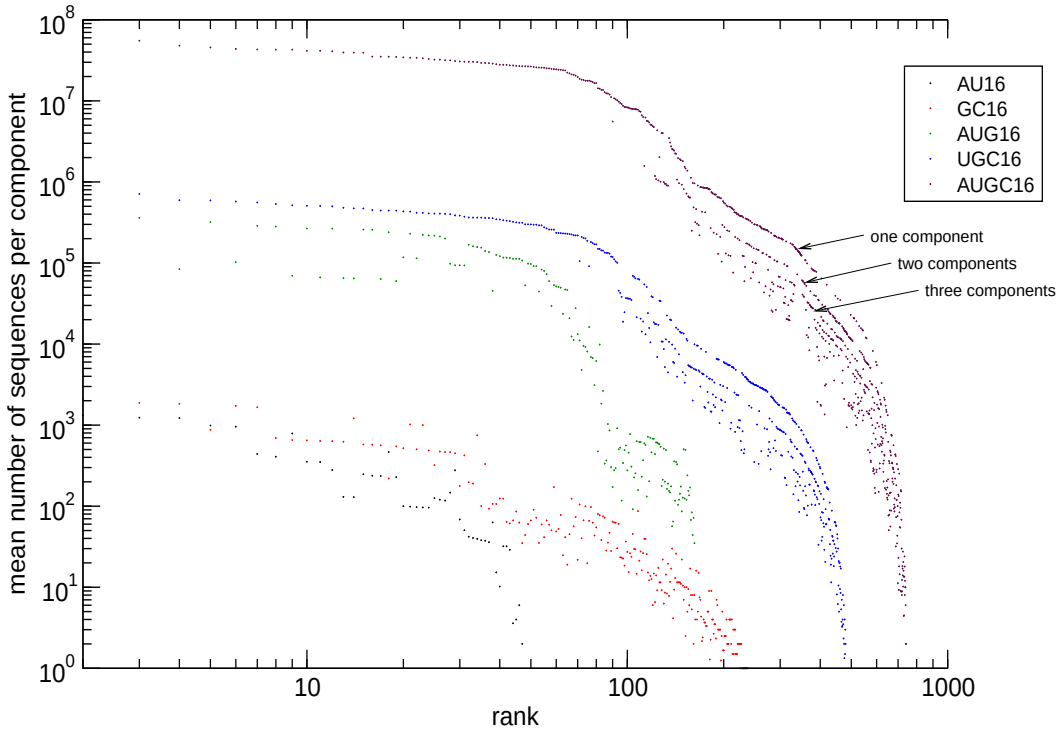


Figure 30: The mean component size of sequence spaces of chain length 16 folded using version 1.3 of `RNAfold`. We can see how split the networks are as networks of the same number of components form a graph as indicated by the arrows similar to figure 29.

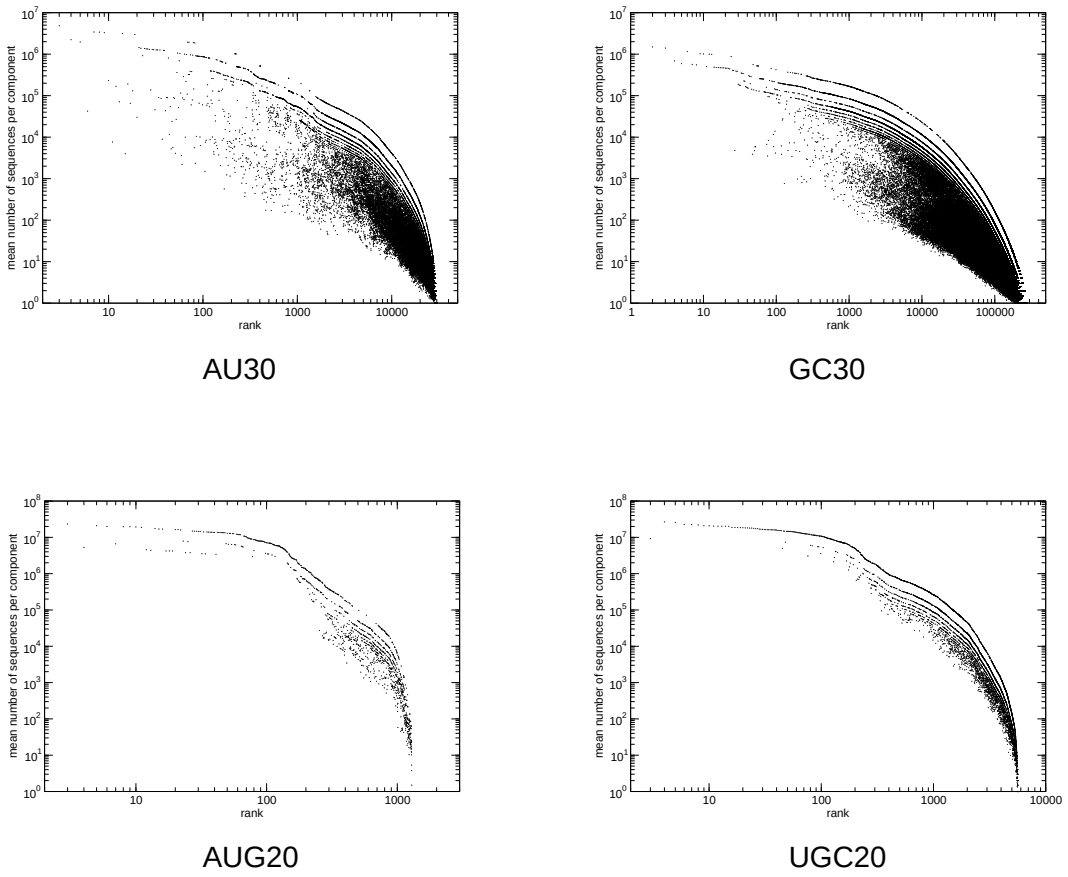


Figure 31: The mean component size of some large sequence spaces folded using version 1.3 of `RNAfold`. AU and AUG networks are more split into components than GC and AUG (rank scales differ for different chain lengths in this figure).

because networks of the same number of components form graphs similar to figure 29. Sequence spaces of chain length 16 (RNAfold version 1.3) are shown in figure 30, figure 31 gives and impression about the disunion of large sequence spaces calculated with RNAfold 1.3 and figure 32 shows some sequence spaces that were folded by RNAfold 1.4. While we find the first ranks to form a single giant or up to four components the number of components increases with decreasing network sizes. Of course there are less components on the last ranks as networks become smaller and the number of components cannot exceed their size. AU and AUG networks split into different components at smaller ranks

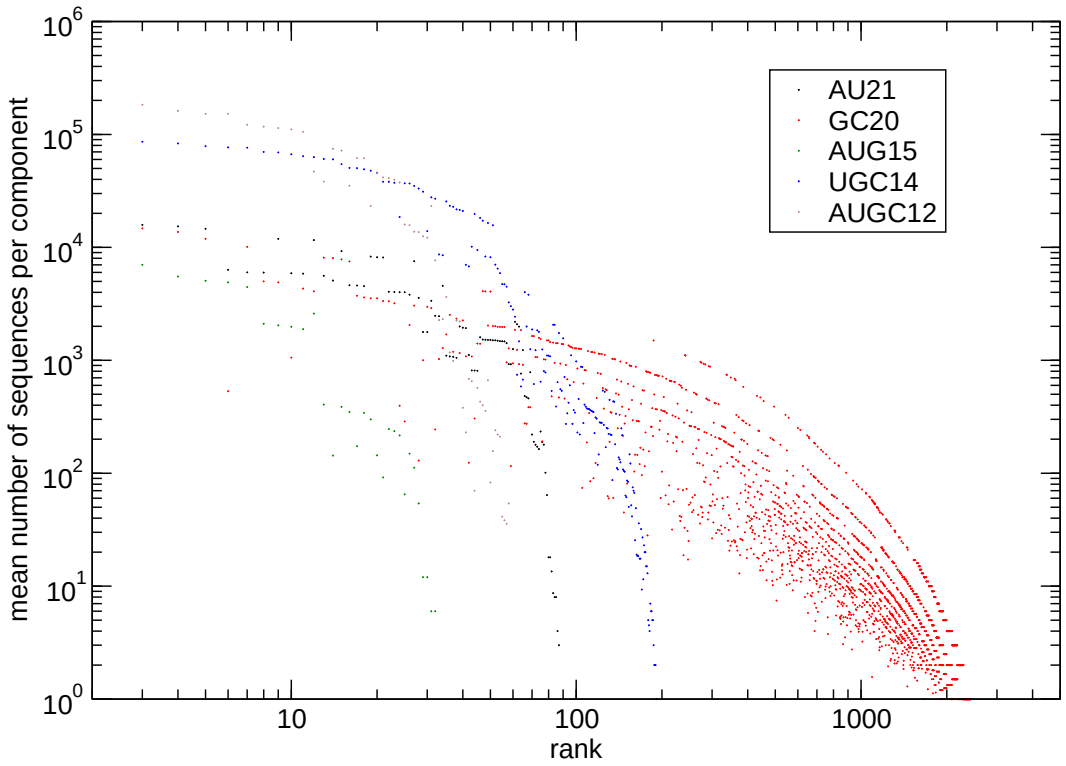


Figure 32: The mean component size of the largest sequence spaces examined using version 1.4 of `RNAfold`. Again we find AU and AUG networks to be more split than others.

and exhibit a larger number of components than UGC and AUGC networks with GC networks being in-between. This is comparable to the number of structures (see figure 26 and tables 1 to 4) that increases at the same chain length in the following order:

$$\#\mathrm{AU} < \#\mathrm{AUG} < \#\mathrm{GC} < \#\mathrm{UGC} < \#\mathrm{AUGC}$$

independently to the used folding parameters.

Looking at the number of components three classes of RNA secondary structures can be observed [24, 56, 61]. Corresponding to those classes secondary structures form one two or four components of almost equal size if they consist of none one or two structural elements respectively that allow the formation of additional base pairs. Examples for those elements are hairpin loops with five or more members, sufficiently large bulges, internal loops, multi-loop or stacking regions with two dangling ends. The three classes are shown in figure 33. For an example GC sequence space a G/C ration significantly different from one increases the probability to avoid the possible formation of an additional generally possible base pair. This means there is an excess of G or C in the base composition of the sequence. One structural element that allows the formation of an additional base pair changes the average sequence composition to larger C or a larger G content than the normal G:C = 1:1. This can be clearly seen in the example shown in figure 34.

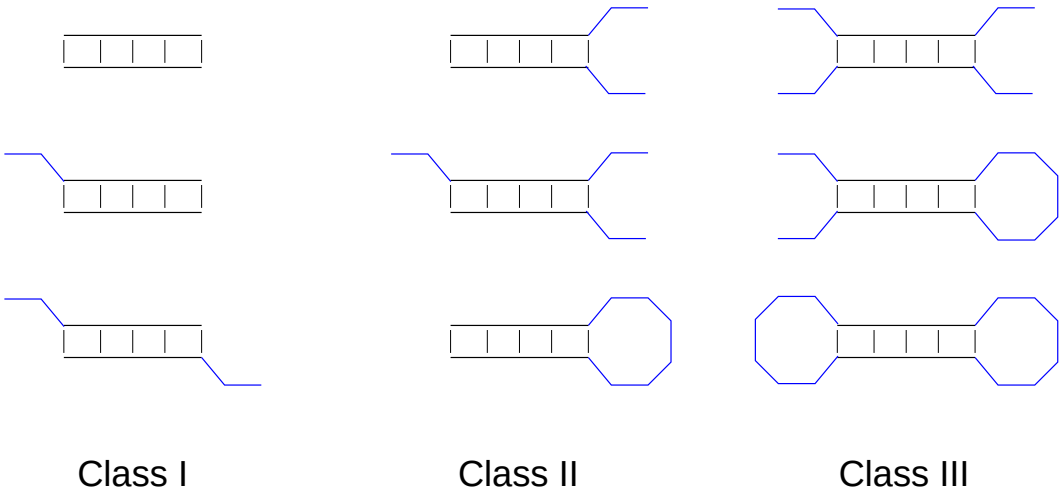


Figure 33: The 3 classes of RNA secondary structures according to the availability of structural elements that can form additional base pairs.

Class II structures consist of two independent structural elements that allow the formation of additional base pairs. In that case we see a distribution consisting of two sub-distributions. Each of the two structural elements leads to an excess of G or C in the first (G_1 or C_1) or the second structural element(G_2

or C_2). That means we see four combination: G_1G_2, G_1C_2, G_2C_1 and C_1C_2. The second and the third combination compensate the excesses to produce an average G:C ratio of 1:1 whereas the first and the last combination show a clear excess in G and C, respectively. This is mirrored in number of components and their base composition as one can see in the example of figure 35.

In the more sophisticate case of a larger alphabet those structural elements can only be investigated by looking at the base composition of the components, for example the roughly equal sized two components of rank 201 of the sequence space AUGC16. In figure 36 the appearance of each base A, U, G and C was counted at every position and their fraction plotted versus the position. The two base pairs of the hair pin are formed solely by G–C for the outer and

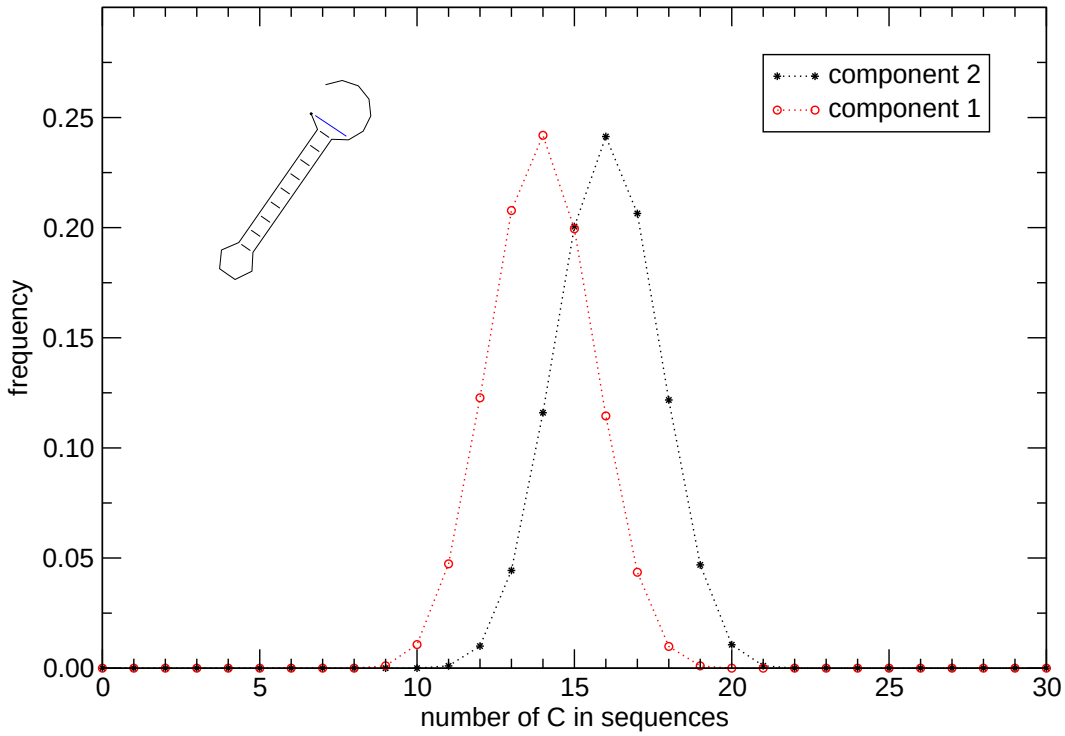


Figure 34: The C content in the two components of rank 17 of GC30 calculated using parameters of version 1.3. The structure graph shows the additional base pair in blue colour that could be formed assuming there are compatible bases in place. See table 5 for the sequence of components.

C–G for the inner pair of the larger component. To avoid the formation of an additional base pair, positions 3 and 13 consist of A–G and A–C in about equal parts. On the other hand the outer base pair of the smaller component 2 is composed of G–C and a smaller number of C–G, while the inner base pair is made of mainly C–G and some G–C pairs. The positions 3 and 13 that would allow to elongate the stack are solely set with G and A, respectively, that do not form base pairs. At both components the other unpaired positions consist of all four bases in about the same frequency of roughly 25% each.

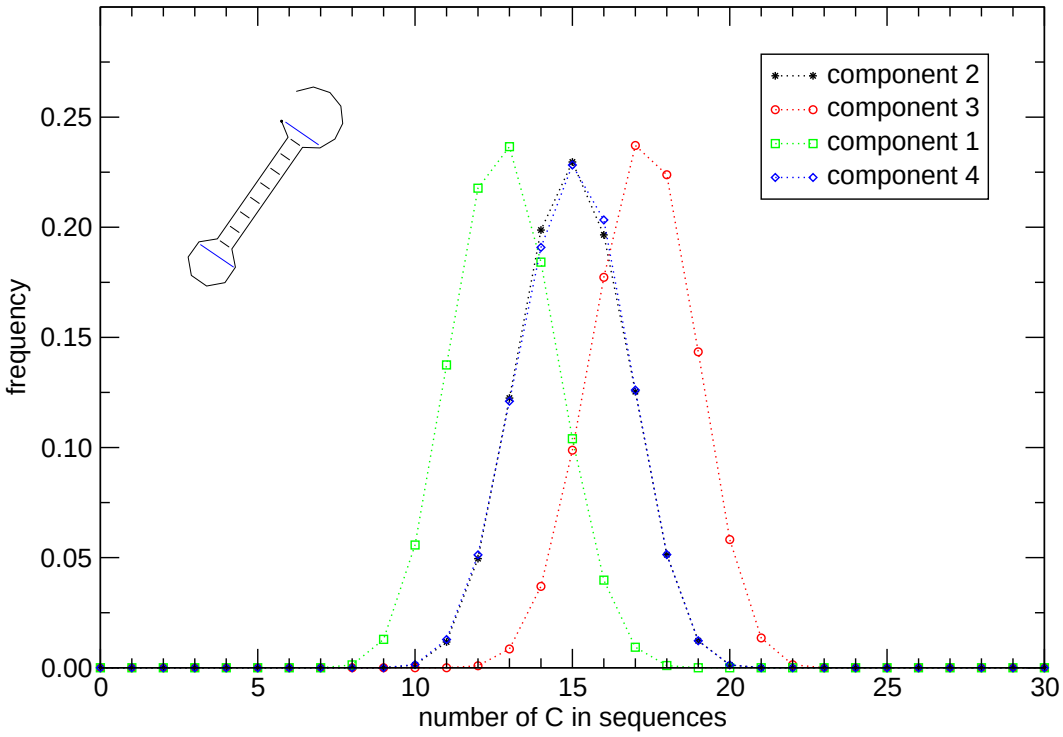


Figure 35: The C content in the four components of rank 31 of GC30 calculated using parameters of version 1.3. The structure graph shows the additional base pairs in blue colour that could be formed assuming there are compatible bases in place. See table 5 for the sequence of components.

Rank	Structure	No. of sequence	No. of components	Sequence of components
1	(((((((((....)))))))))........	1840277	1	1840277
2	........(((((((((....)))))))))	1495055	1	1495055
3	(((((((((...))))))))).........	1404647	1	1404647
4	(((((((((.....)))))))))..........	1377207	2	717214 659993
5	(((((((((.....)))))))))..........	1178339	2	647053 531286
6	(((((((((....)))))))))..........	1154181	1	1154181
7	(((((((((.....)))))))))..........	1136230	2	587540 548690
8	.(((((((((....)))))))))..........	1102317	2	555258 547059
9	((((((((((....))))))))))..........	1019189	1	1019189
10	((((((((((...))))))))))..........	1012695	1	1012695
11	........(((((((((.....)))))))))	1003459	2	521678 481781
12	......((((((((((....))))))))))	1001489	1	1001489
13	.......((((((((((...))))))))))	991964	1	991964
14	.......(((((((((.....)))))))))	970872	2	532637 438235
15	(((((((((.....)))))))))........	952702	2	482438 470264
16	........(((((((((....))))))))..	946925	2	536182 410743
17	.(((((((((....)))))))))........	941082	2	471354 469728
18	(((((((((.....))))))))).......	938924	2	474897 464027
19	.......(((((((((....))))))))).	934279	2	476172 458107
20	..(((((((((....)))))))))..........	921234	2	523675 397559
21	.(((((((((...)))))))))........	920508	2	462406 458102
22	.......(((((((((.....)))))))))	904162	2	458048 446114
23	......(((((((((.....)))))))))	882348	2	449422 432926
24	.........(((((((((...)))))))))	877159	1	877159
25	(((((((((.......))))))))).....	818050	2	441068 376982
26	........(((((((((...))))))))).	809568	2	421968 387600
27	(((((((.......))))))).........	794802	167	436832 357778 5 3 3 3 2 2 [. . .]
28	.........(((((((((....))))))))).	781780	2	394202 387578
29	.....(((((((((.......)))))))))	764465	2	415150 349315
30	.(((((((((.....)))))))))........	756850	4	203084 189651 187708 176407
31	.(((((((((.....))))))))).......	719041	4	192975 180078 179185 166803

Table 5: The first ranks of the sequence of components of GC30 (partly shortened). See figures 34 and 35 for the distribution of C in the different components of rank 17 and 31, respectively.

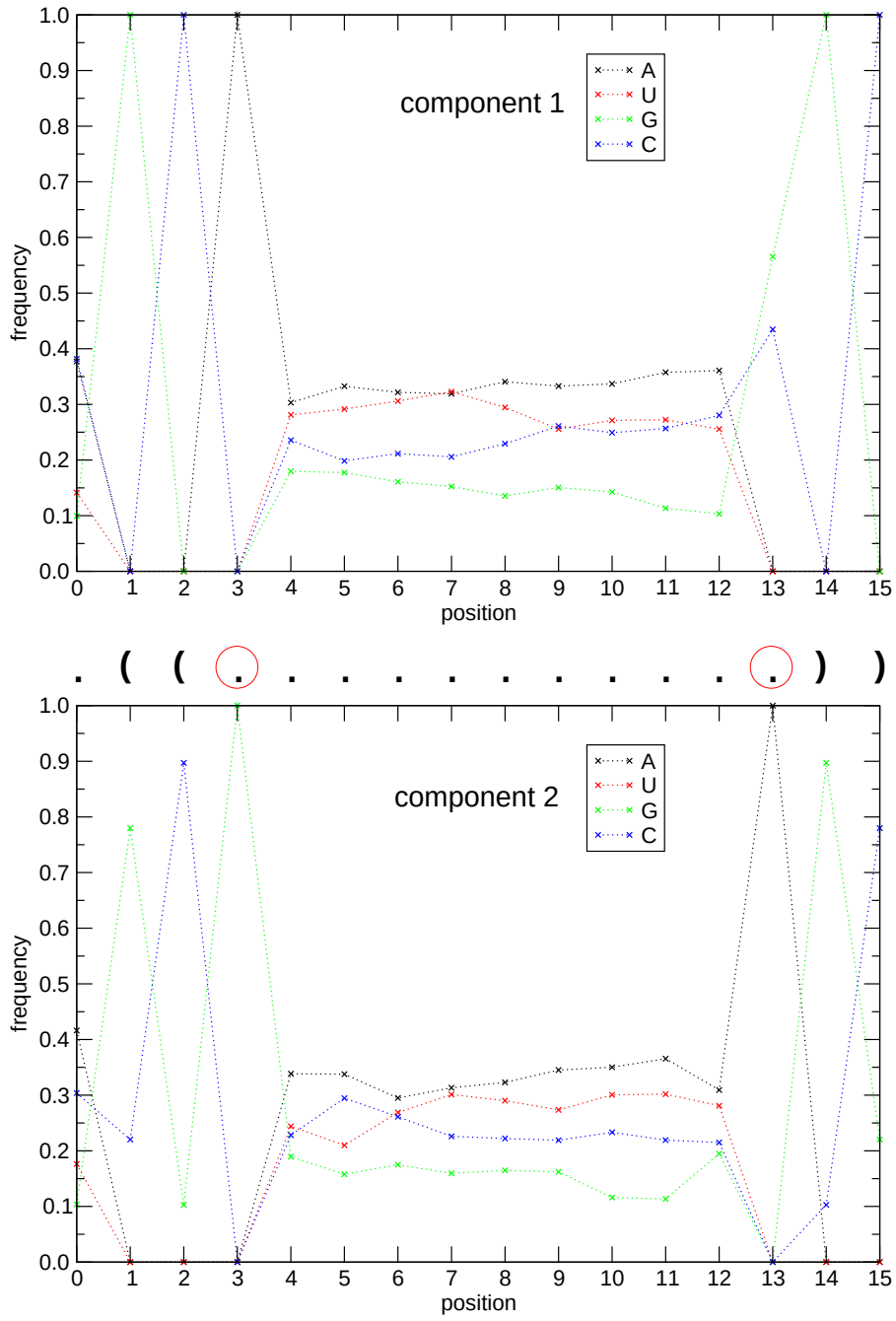


Figure 36: The histogram of the base composition of rank 201 of the sequence space AUGC16. The structure is shown between the histograms of the two components, that are roughly equally sized (289687 and 263865 sequences for components 1 and 2, respectively). Red circles mark the positions that could form an additional base pair if they are set with compatible bases.

4.2 Kinetics of RNA Folding

4.2.1 Influences to Folding Times

To test the influence of the minimal free energy of a RNA sequence on the kinetic folding behaviour about 1500 sequences folding into three different minimal free energy structures were tested. To achieve an adequate sample that incorporate the influence of size and structural differences the examined structures were selected from different classes and different complexity. A hairpin structure as a very simple example, the more sophisticated Y-shape structure containing already one multi-loop but being still equal in size to the hairpin, and a clover leaf structure as a more realistic bigger sized example that occurs in tRNAs. See figure 37 for details of those structures. The time dependent folding of the sequences was simulated using the program `kinfold` [12] using 1000 simulations per sequence. The complete flowchart is shown in figure 38.

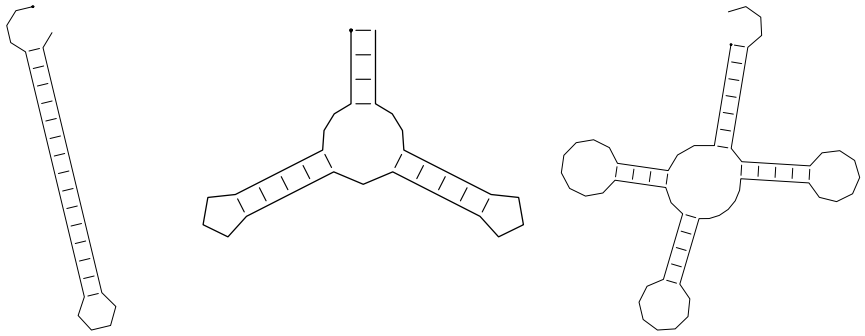


Figure 37: The three analysed structures. From left to right: hairpin structure: 39 bases (1000 sequences), Y-shape structure: 39 bases (788 sequences), $tRNA^{Phe}$ clover leaf structure: 76 bases (523 sequences).

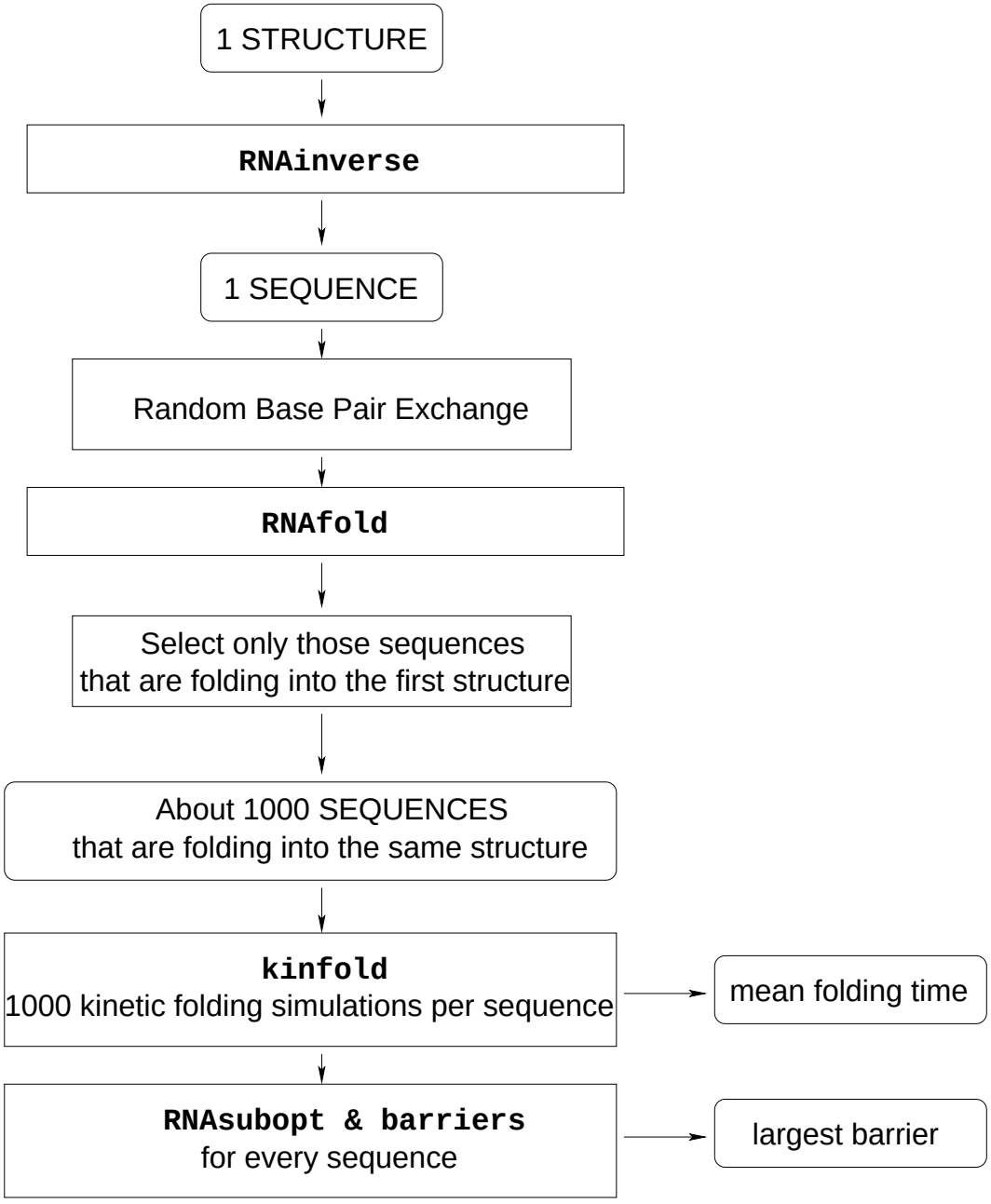


Figure 38: Flowchart of the steps involved to obtain statistically relevant data to analyse the kinetics of RNA folding.

This procedure resulted in enough data to allow a statistically relevant examination. In addition the barrier trees of local minima were calculated using `barriers` [12,13] and `RNAsubopt`. To calculate all suboptimal structures necessary to construct a complete barrier tree of a reasonable region it was necessary to choose a stepwise proceeding. This involved an increase of the energy range above the minimum free energy when using `RNAsubopt` at each step. This is done after the barrier tree calculated with barriers which uses RNAsubopt output as input was analysed in terms of enough local minima and missing saddles. The stepwise approach is necessary because of the time costly calculations of `RNAsubopt` at higher energy ranges.

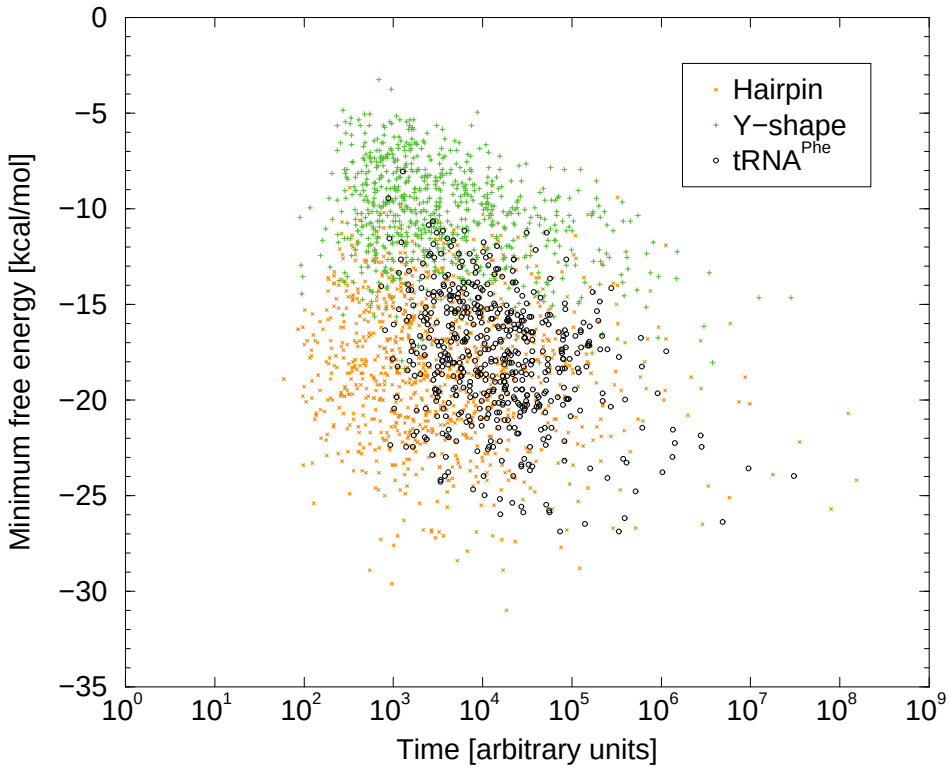


Figure 39: Minimum free energy versus mean folding time of sequences folding into three different structures showing no correlation. The structures are shown in figure 37.

The results of our simulations are presented in the following. The often accepted hypothesis that the main factor in folding kinetics of an RNA sequence is its minimum free energy can be clearly defeated. Figure 39 shows the minimum free energy plotted versus the mean folding time. The result is unmistakable: no correlation at all can be found between them. The mean folding time definitely does not depend on the minimum free energy. In this plot we can only see the tendency of larger RNAs (tRNA^{Phe} shown in black colour which has a size of 76 bases) to be slower folders than shorter ones (hairpin and Y-shape shown in orange and green colour respectively with a size of 39 bases).

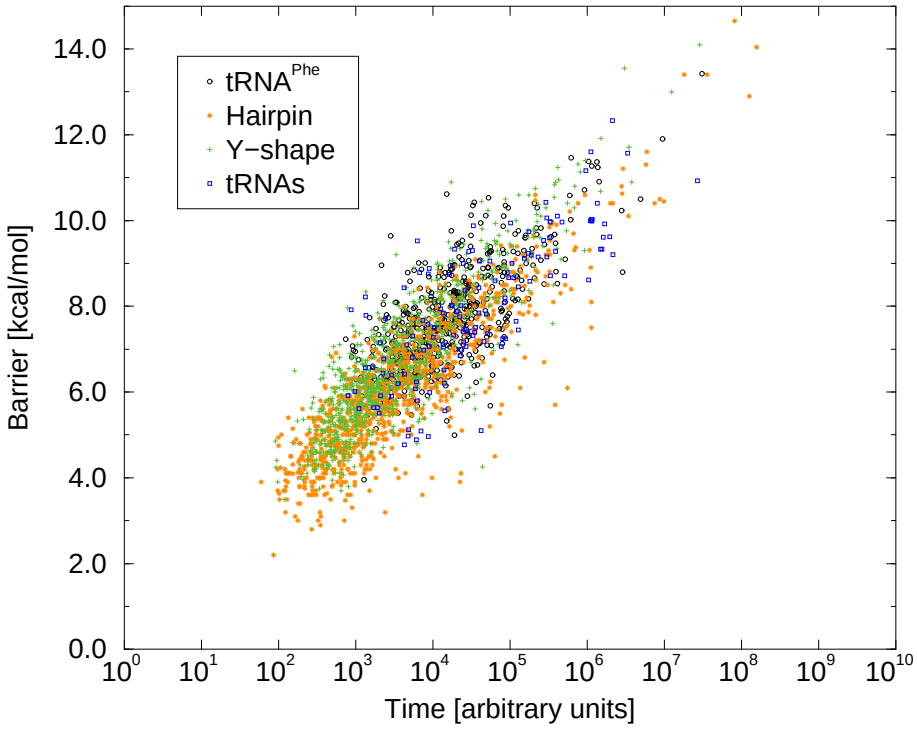


Figure 40: Highest barrier versus mean folding time of kinetic simulations starting at the open chain and ending at the mfe structure. Additional kinetic data of different tRNAs were provided by Dagmar Friede and are shown in blue colour.

First we examine the folding process from the open chain to the minimum free energy structure. Plotting the barrier height of the greatest barrier versus the mean folding times shows positive correlation (see figure 40). Starting at the open chain structure many local minima are visited and the barriers between them have to be overcome. The higher a barrier the longer it takes to escape from its basin. In general the time needed to overcome the greatest barrier is the main factor that determines the overall folding time. Smaller barriers are taken quickly and do not count a lot in terms of the overall mean folding time. These facts express themselves well in figure 40. Kinetic foldings of additional different tRNAs were performed by Dagmar Friede. Her data are shown in figure 40 in blue colour and fit well into the picture drawn by kinetic

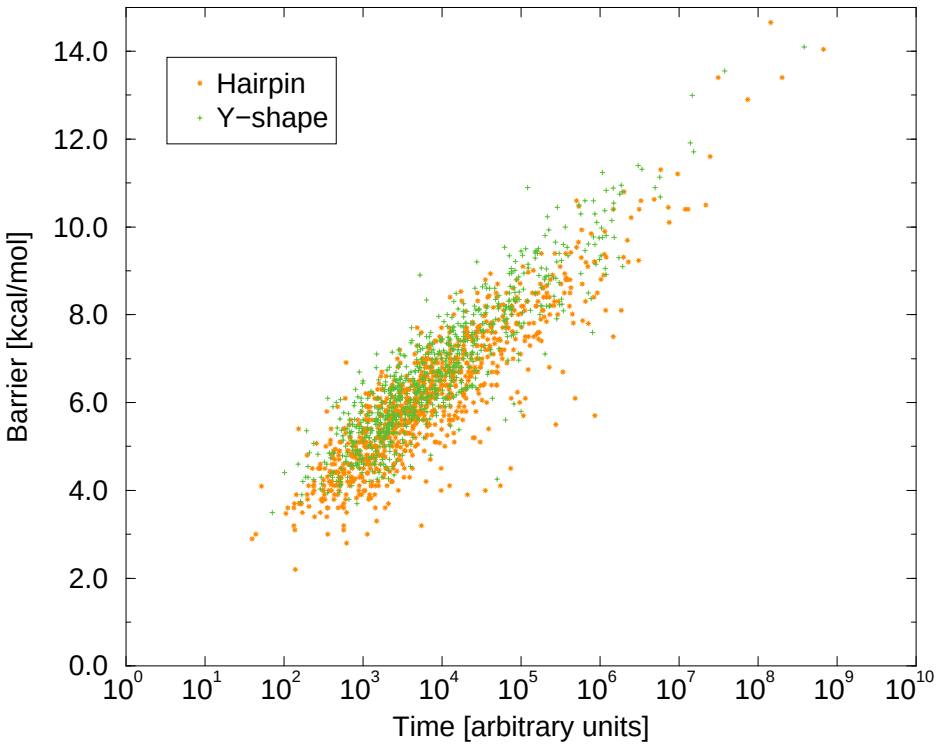


Figure 41: The highest barrier versus the mean folding time for a folding path starting at the local minimum associated with the highest barrier and ending at the mfe structure.

foldings of hairpin Y-shape and tRNA^{Phe} structures. Looking at this plot we also find that longer RNAs (tRNA^{Phe} and tRNAs) have higher maximum barriers and longer folding times than shorter RNAs (Hairpin and Y-shape).

In contrast to figure 40 figure 41 shows the refolding from a local minimum to the minimum free energy structure. We see a similar picture with a positive correlation between mean folding time and height of the greatest barrier. The folding process in these simulations did not start at the open chain structure but at the bottom of the basin of the highest barrier. This normally involves only a single barrier (the highest one) and excludes the influence of all smaller local minima and their barriers to the mean folding time. In fact their influence is only small as discussed previously that is why figures 40 and 41 are quite similar. The main statement of this experiment is that a higher (maximum) barrier results in a longer mean folding time.

The same simulations were performed on the hairpin and the Y-shape structure using a reduced alphabet that allowed only Gs and Cs in the sequences. For time efficiency reasons only some fast folding sequences were used. Figure 42 compares the barrier height in dependence to the mean folding time for the two structures in different alphabets. In spite of the small number of tested sequences using the GC alphabet we find that they fit well into the picture of sequences using the AUGC alphabet. The chosen alphabet does not seem to influence the general features of folding kinetics.

In the following we will discuss some example sequences and their kinetics. The most frequent trap and the mfe structure are shown and the barrier between them discussed as it strongly influences the mean folding time. The most important stop structures were obtained by selecting them by the following properties: the 30 greatest barriers to the mfe structure and the 30 largest basin sizes. Not all sequences showed 30 local minimum structures that merge with the mfe structure, and doubles were of course expunged. This procedure most times left less than 60 stop structures. After 1000 kinfold simulations the frequency each stop structure was reached could be counted.

Figure 43 gives an example of a hairpin structure that is an average folder in terms of kinetics. There is one major trap with a barrier of 12.90 kcal/mol

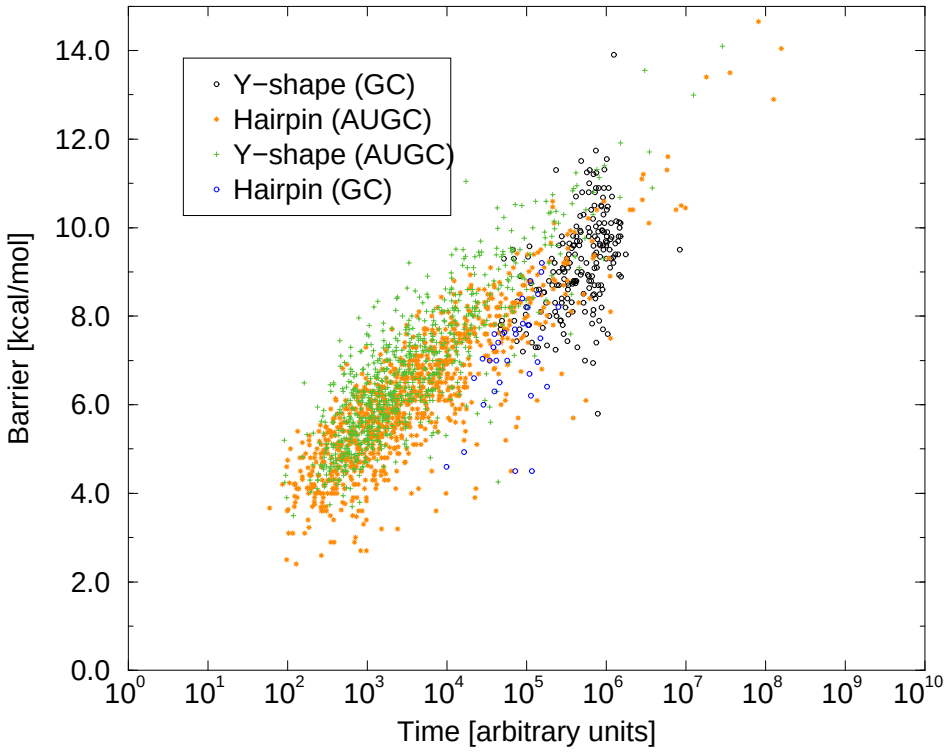


Figure 42: Barrier (of open chain to mfe structure) versus mean folding time. A hairpin and a Y-shape structure in two different alphabets (AUGC and GC) are compared.

to the minimum free energy structure which correlates well with the mean folding time of $1.25 \cdot 10^8$ arbitrary units.

Another average folding RNA is examined in figure 44. It folds into a clover leaf structure of tRNA^{Phe}. There is one major trap that is visited in 81.9 percent of the folding paths simulated. It has a barrier of 7.82 kcal/mol to the minimum free energy structure which correlates well with the mean folding time of $1.9 \cdot 10^4$ arbitrary units.

An example of a sequence that can be found below the regression curve of figure 40 is shown in figure 45. It folds into a hairpin structure. The most important trap is still not visited frequently and has a rather small barrier of 2.60 kcal/mol to the minimum free energy structure. It has neither the highest barrier nor the largest basin size. There are many local minima with similar

small barriers. The local minimum with the largest barrier of 4.10 does not have much influence on the mean folding time because it is only visited in 9.1%. This local minimum cannot be stated as a major trap. That is why the mean folding time of $2.3 \cdot 10^4$ arbitrary units is longer than expected from the height of the largest barrier.

A clover leaf structure example of a sequence that is located below the regression curve of figure 40 is shown in figure 46. The most important trap is still not visited frequently and has a rather small barrier of 3.41 kcal/mol to the minimum free energy structure. It has neither the highest barrier nor the largest basin size. There are many local minima with similar small barriers. The local minimum with the largest barrier of 4.99 does not have much influence on the mean folding time because it is only visited in 1.9%. This local minimum cannot be stated as a major trap. The long folding time of $1.9 \cdot 10^4$ arbitrary units is longer than expected from the height of the largest barrier and is in this case rather influenced by the large number of minor traps that are visited on the folding path.

Figure 47 shows a sequence that folds into a Y-shape structure and can be found above the regression curve in figure 40. The most important trap has not the highest barrier which is 9.20 kcal/mol but a rather small barrier of 4.14 kcal/mol to the minimum free energy structure. It also does not have the largest basin size. Still it is visited most often and therefore it's barrier has much more influence to the mean folding time than the largest barrier which cannot be stated as a major trap. That is why the mean folding time of $3.3 \cdot 10^4$ arbitrary units is shorter than expected from the height of the largest barrier.

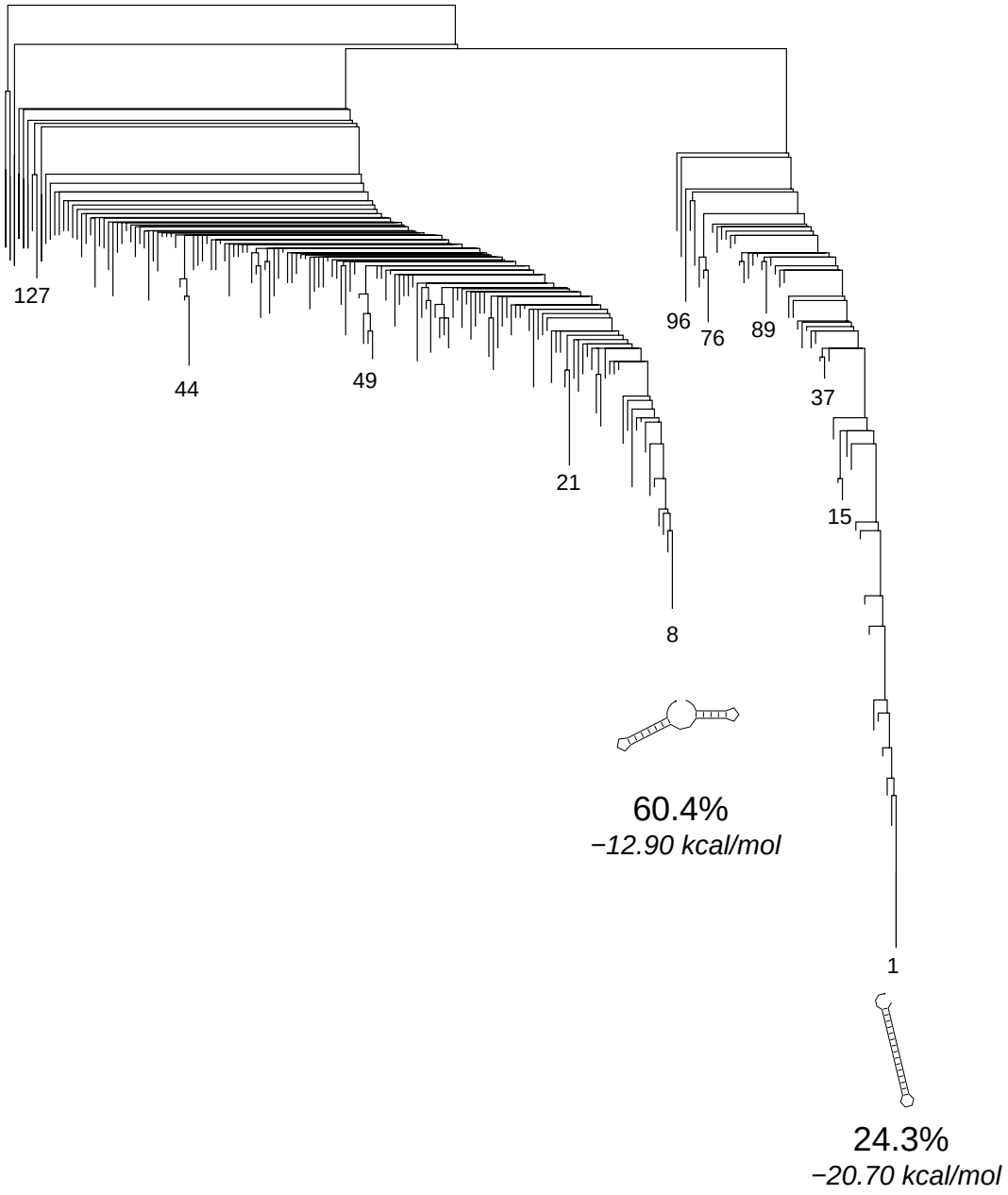


Figure 43: An example of a sequence (GGGAAUUCGGCAUAGCCGAAUCGUUGGUUGUGCCGAGUC) that folds into a hairpin structure. Selected structures are labeled by their number. This is the rank in the list of all suboptimal structures sorted by their free energy. The trap structure (number 8) and the minimum free energy structure (number 1) are drawn below their corresponding branch of the barrier tree. The frequency each of the structures is visited and in italic letters their free energy can be found next to the structures. This example shows an average folder with only one major tarp.

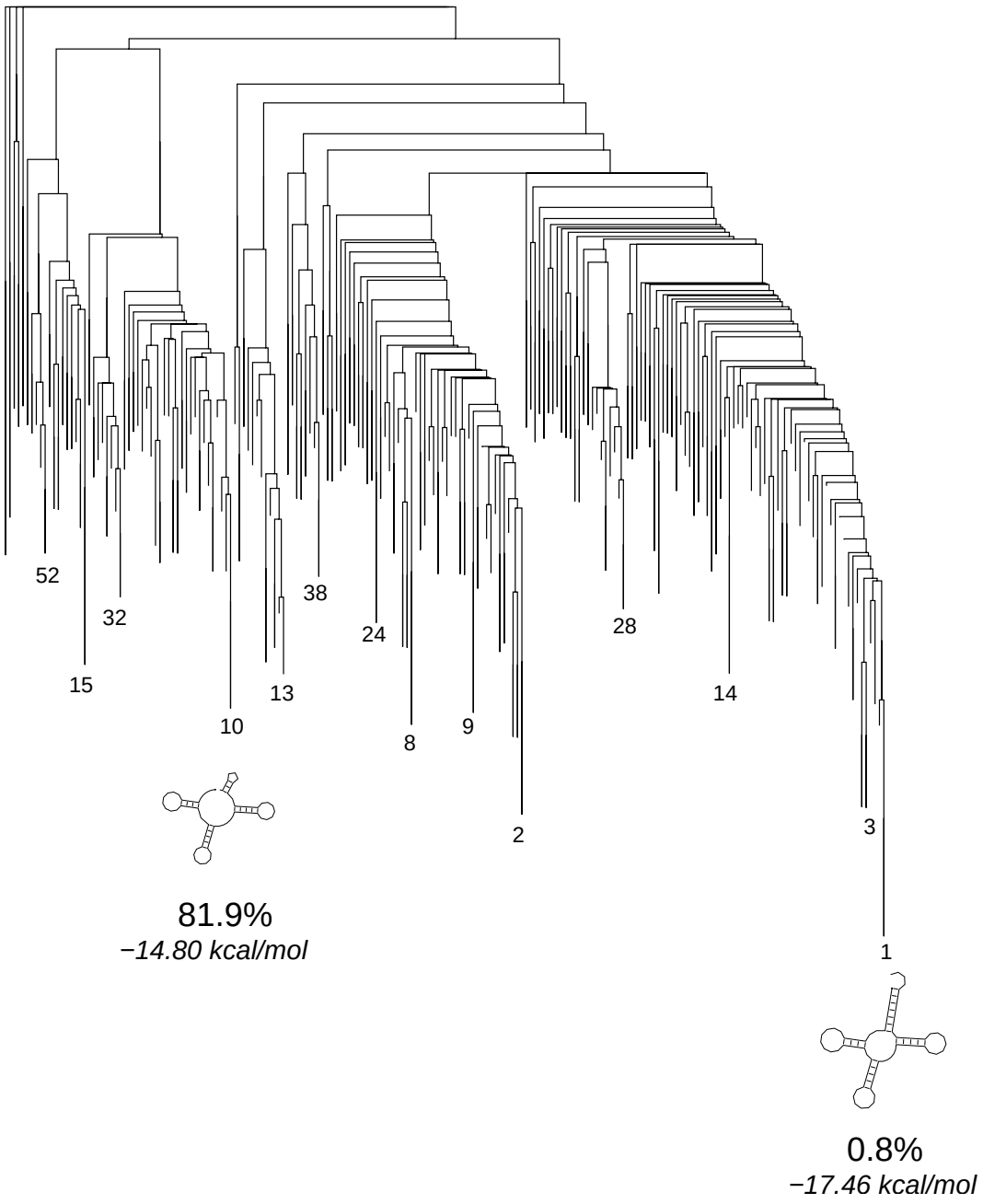


Figure 44: An example of a sequence (GAGACUUUAGAGCAGNNGGGAGCUCGCAUGACUGAANAUUAU GAGNUCGCGGGNUCGNUCUCCGUAGGUCUCACCA) that folds into a clover leaf structure of tRNA^{Phe}. This average folding RNA shows one major trap (number 10) that was visited 819 times in thousand simulations.

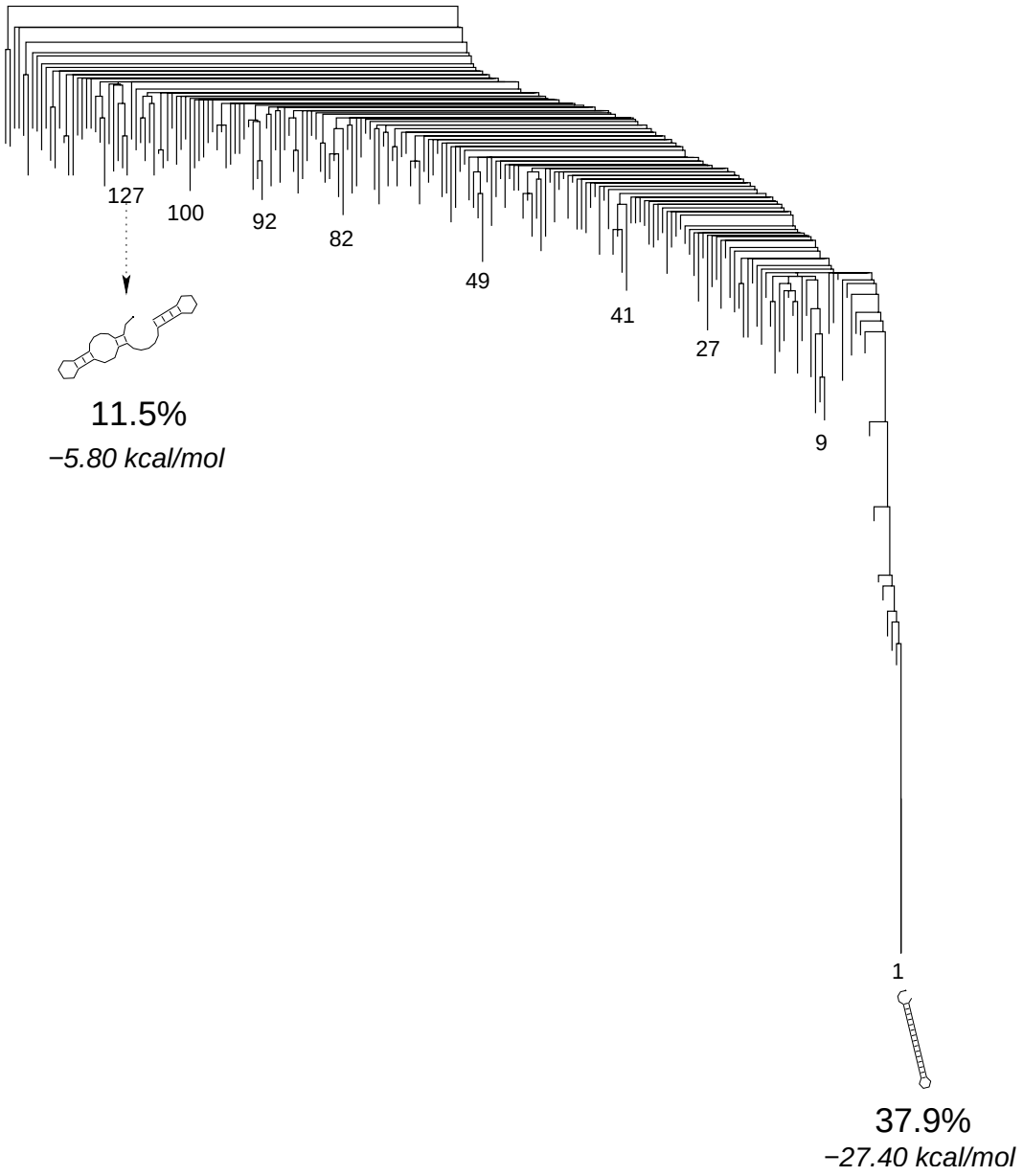


Figure 45: An example of a sequence (GGGAAGCUUGGGCUCCCGUAUCGACGGGAGCUCAAGCUC) that folds into a hairpin structure. The most important trap (number 127) is still not visited frequently and has a rather small barrier of 2.60 kcal/mol to the minimum free energy structure. It has not the highest barrier and there are many similar local minima. Many different ones are visited on the folding path, that is why the mean folding time is longer than expected from the highest barrier.

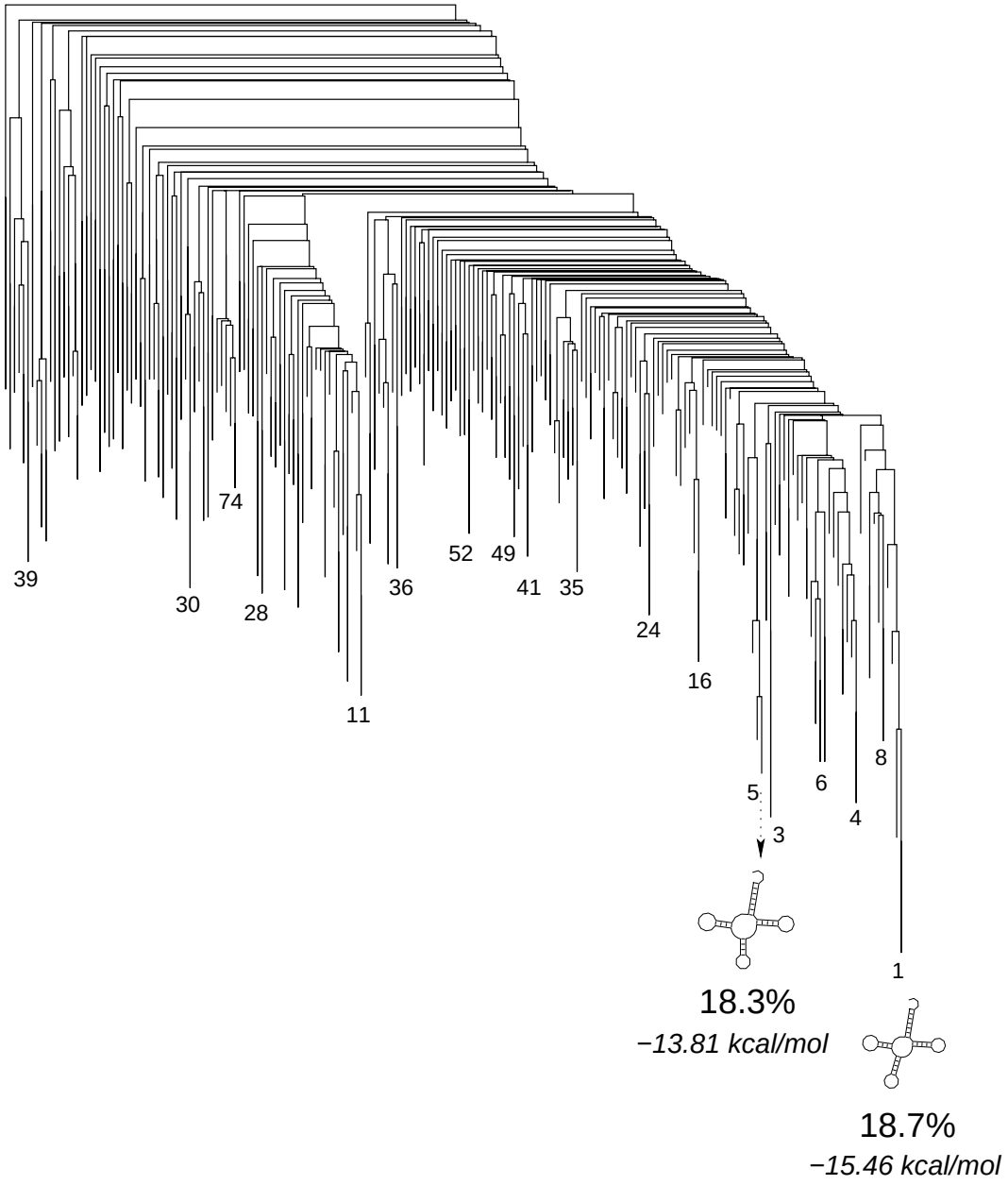


Figure 46: An example of a sequence (GCGCGGUUAGUGUAGNNGGGAACACGGGUAGCUGAANACUAUU AGNUCGAUUUNUCGNUCAAAUCAUCGCGCACCA) that folds into a clover leaf structure. The most frequent local minimum (number 5) still was only visited in 18.3 percent of the simulated foldings. The local minimum of the highest barrier is visited even rarer. They both cannot be stated as major traps. The long folding time in this case is rather influenced by the large number of minor traps that are visited on the folding path.

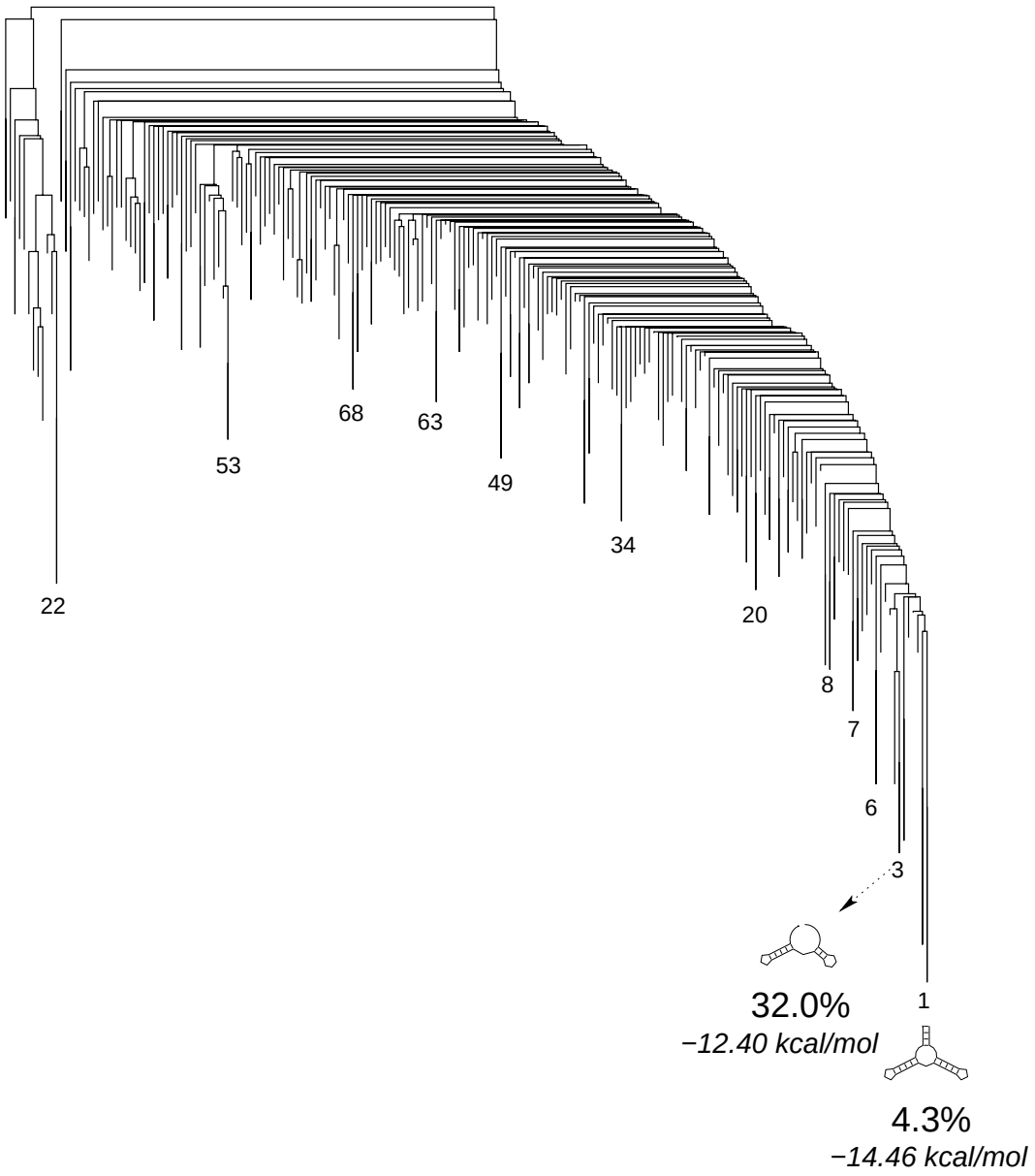


Figure 47: An example of a sequence (GCUCAAGGGGUUUUACCCCAGUUGUAAAACGAUAAGAGC) that folds into a Y-shape structure. The folding time is more influenced by the most frequently visited trap (number 3), than by the one with the highest barrier that is rarely found on the folding path. This results in a faster folding sequence than suggested by the size of the maximum barrier.

Not only the highest barriers but also the barriers of the most frequently visited trap structures influence the mean folding time. In figure 48 those barrier heights are plotted versus the mean folding times and again we find positive correlation between them. There are no more extremes at the upper left corner because local minima with high barriers that are seldom visited disappear. On the other hand at the lower right corner the values are more scattered because less frequently visited local minima of much higher barriers seem to have more influence on the total folding time. This is due to a large difference between the barrier of the most frequent local minimum and the highest barrier. To overcome a really high barrier takes proportionally more time than crawling over a much smaller barrier many times.

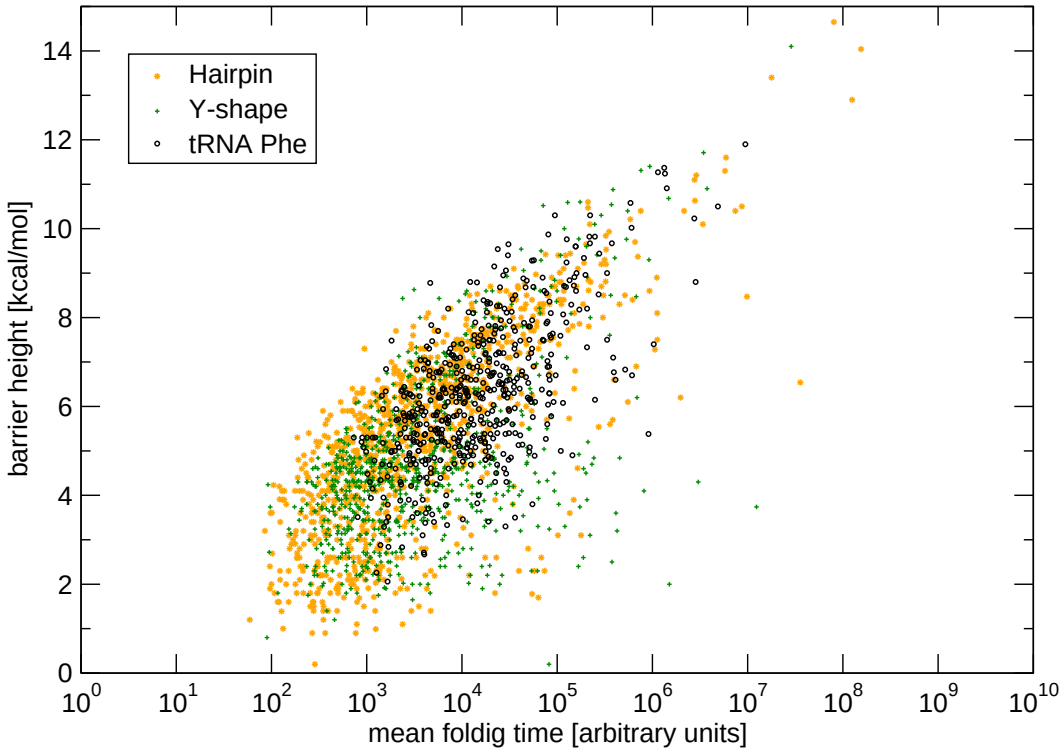


Figure 48: Barrier height of most frequent trap vs. mean folding time.

Some properties related to barrier trees are presented in the following. Parts of the data of different sized RNA sequences were kindly provided by Ivo Hofacker (unpublished). The barrier trees of the randomly generated RNA sequences were calculated using the `barriers` program and analysed. The maximum barrier height grows with the size of RNA (see figure 50). As larger RNA molecules can form more suboptimal structures there is also a larger number of possible local minima, saddle points and their corresponding energy barriers. Very often the local minimum located at the bottom of the largest basin also has the largest barrier, but this does not always have to be true. Nevertheless a larger basin has in general a higher barrier as shown in figure 49. The barrier height also grows with an increasing structure distance calculated as the base pair distance (figure 51).

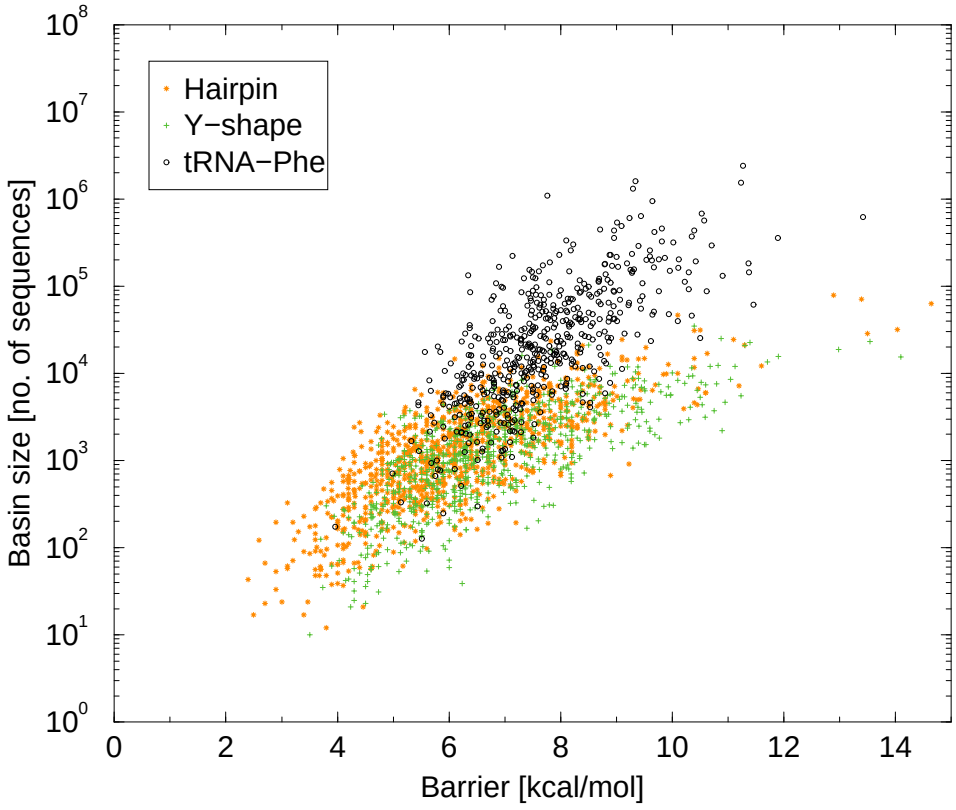


Figure 49: The barrier height in dependence on the basin size.

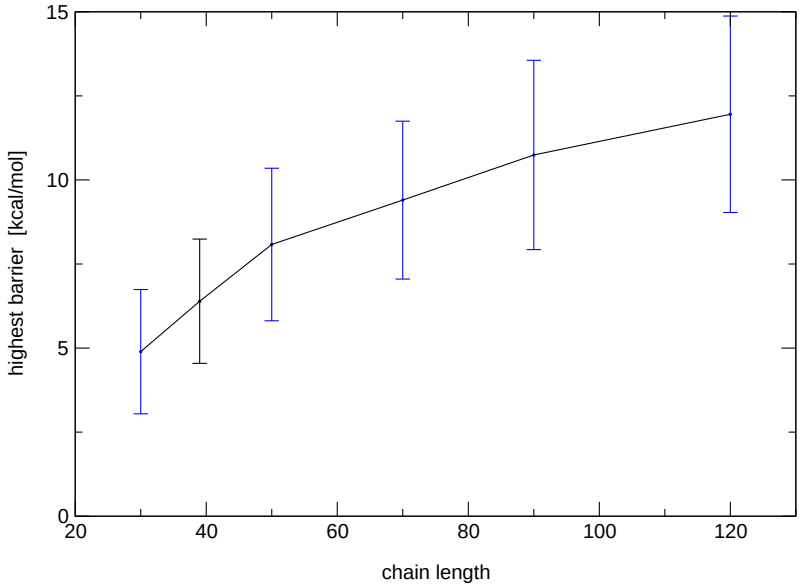


Figure 50: The maximum barrier height between the lowest 100 local minima in dependence on the size of RNA shown on randomly generated sequences (data provided by Ivo Hofacker is plotted in blue).

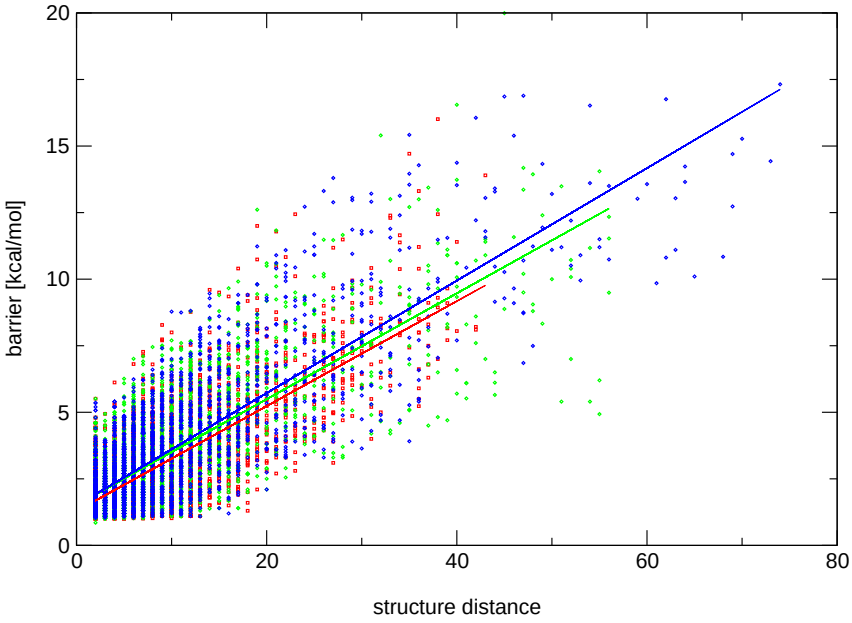


Figure 51: The structure distance calculated as base pair distance in dependence on the size of RNA shown on randomly generated sequences (data provided by Ivo Hofacker).

4.2.2 Kinetics of a Complete Sequence Space

The sequences of AUGC12 (new parameters version 1.4) folding into multi component structures were kinetically analysed. The largest components were expected to be the neutral nets that are mainly used in neutral evolution and therefor we would expect a shorter mean folding time for the larger component. The two component structures can be found in figure 52. Only at three out of 13 structures the larger component, which has rank one, folded more quickly than the smaller component (of rank two). Table 6 lists the components in detail.

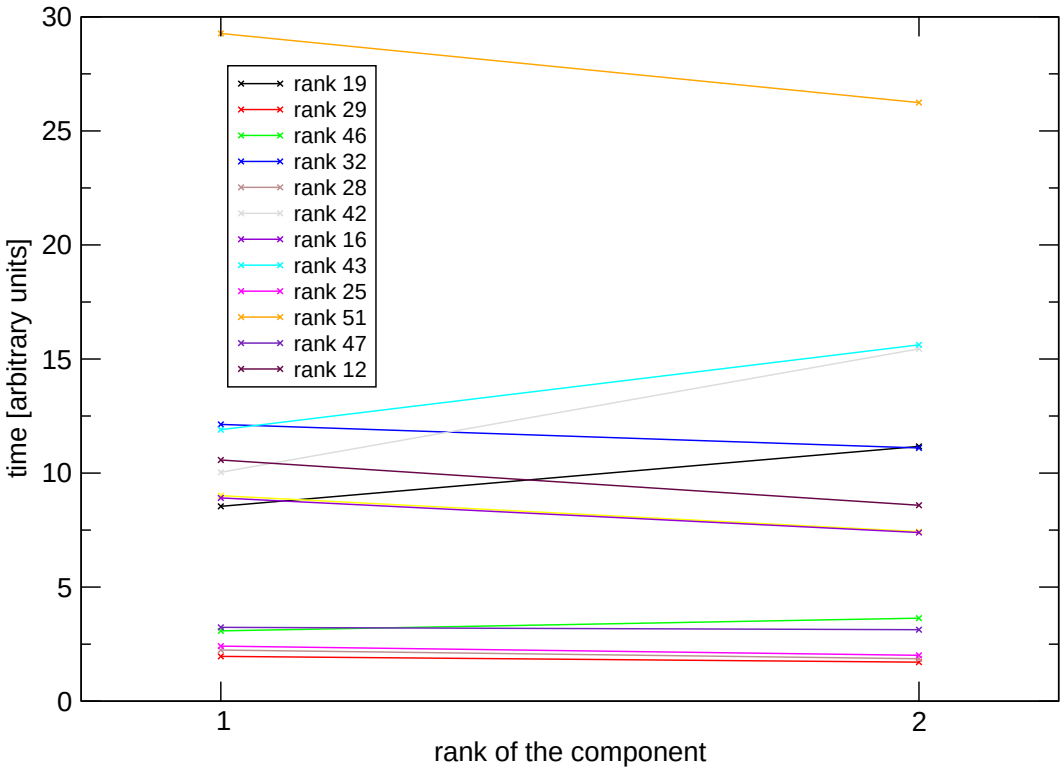


Figure 52: Kinetics of the two component structures of the sequence space AUGC12 calculated using the new parameters (version 1.4). The mean folding time of all sequences belonging to a component is plotted versus the rank (sorted by size) of that component. The legend gives the rank of the structure each component pair belongs to.

Rank	Structure	No. of sequence	No. of components	Sequence of components
1	.....	14325304	1	14325304
2	(((....)))..	218567	1	218567
3	.(((....))).	183335	1	183335
4	(((.....))).	161765	1	161765
5	((....))....	152393	1	152393
6	..(((....)))	152221	1	152221
7	...((....)).	121861	1	121861
8	((((....))))	117253	1	117253
9	.((....))...	113896	1	113896
10	.(((.....)))	110842	1	110842
11	..((....))..	105538	1	105538
12	((.....))...	93866	2	92066 1800
13	..((.....)).	76439	2	74879 1560
14	(((......)))	74626	1	74626
15	((......))..	71904	1	71904
16	.((.....))..	70375	2	68671 1704
17	.((......)).	61792	1	61792
18	((((...)))).	61613	1	61613
19	....((....))	46510	2	45629 881
20	.((((...))))	45288	1	45288
21	..(((...))).	41618	1	41618
22	(((...)))...	41092	1	41092
23	.(((...)))..	39740	1	39740
24	((.......)).	37472	1	37472
25	(....).....	31848	2	30010 1838
26	.....(....).	31498	2	28839 2659
27	....(....)..	27522	2	25365 2157
28	.(....).....	27312	2	24933 2379
29	..(....)....	25053	2	22782 2271
30	...(....)...	24366	2	22365 2001
31	...(((...)))	23260	1	23260
32	..((......))	15350	2	10700 4650
33	...((.....))	11365	5	7109 1584 904 902 866
34	......(....)	6940	3	3407 2632 901
35	((.(....))).	3638	1	3638

36	(((....)).).	3519	3	3455 32 32
37	((.(....).))	2963	3	2943 16 4
38	(.((....))).	2244	1	2244
39	((........))	2208	1	2208
40	.(.(....).).	1520	4	1452 48 13 7
41	(.(....).)..	1379	6	629 569 100 54 19 8
42	.((.......))	1368	2	905 463
43	.((.(....)))	1308	2	1284 24
44	(..(....)..)	1189	17	902 146 41 27 15 9 8 8 7 6 6 4 3 2 22 1
45	.(((....).))	1140	2	1111 29
46	..(.(....)).	860	2	799 61
47	(.(....))...	800	2	745 55
48	.(.(....))..	713	2	645 68
49	(.((....)).)	665	1	665
50	..(.(....).)	414	5	305 51 47 7 4
51	(..(...)..).	314	2	192 122
52	(.((...)).).	240	1	240
53	((((...)).))	220	1	220
54	((.((...))))	211	1	211
55	..((....).).	165	4	83 47 22 13
56	.((....).)..	153	4	73 43 24 13
57	(((....)).).	107	3	74 29 4
58	(.(.....).).	54	1	54

Table 6: The sequence of components of AUGC12 calculated using the new parameters (version 1.4).

The picture of more split neutral nets looks differently (see figure 53). Only in one structure (rank 57) the largest component showed the slowest mean folding time. In the rest (8 structures) the largest component was not the slowest and in 4 cases (ranks 50, 55, 56 and 41) the largest component had even the fastest mean folding time. Nevertheless no clear statement can be made on a difference in the fold-ability of components also because of the extreme variation in mean folding times of the sequences belonging to a single component.

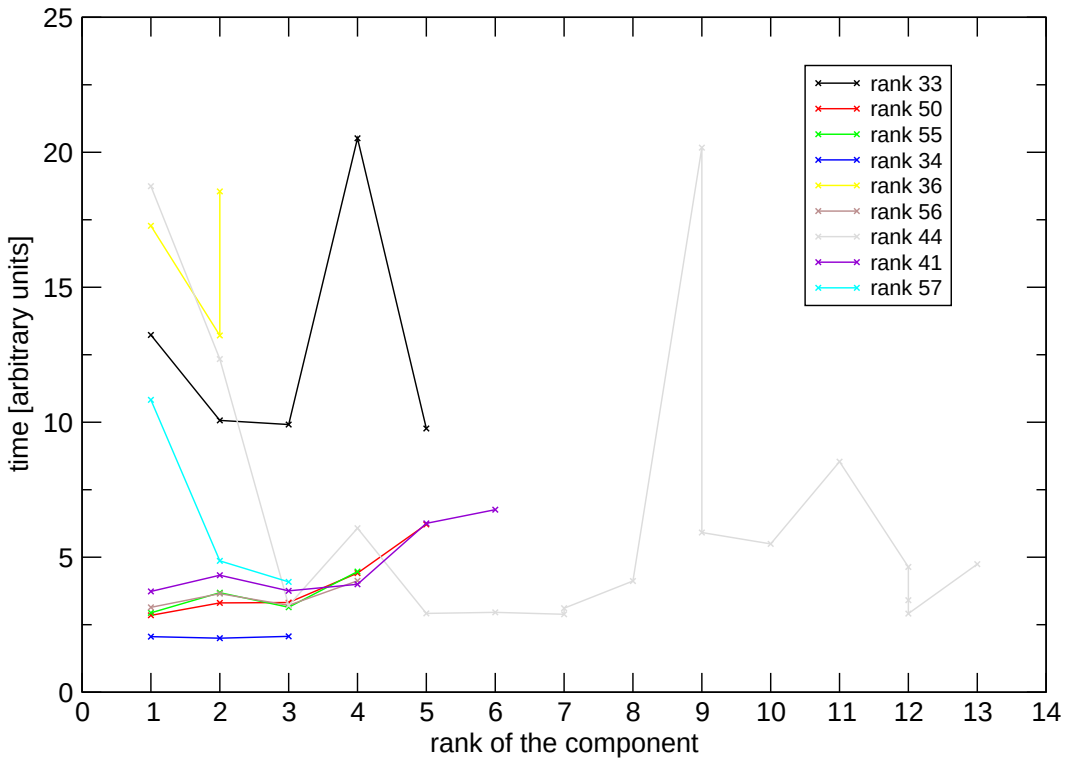


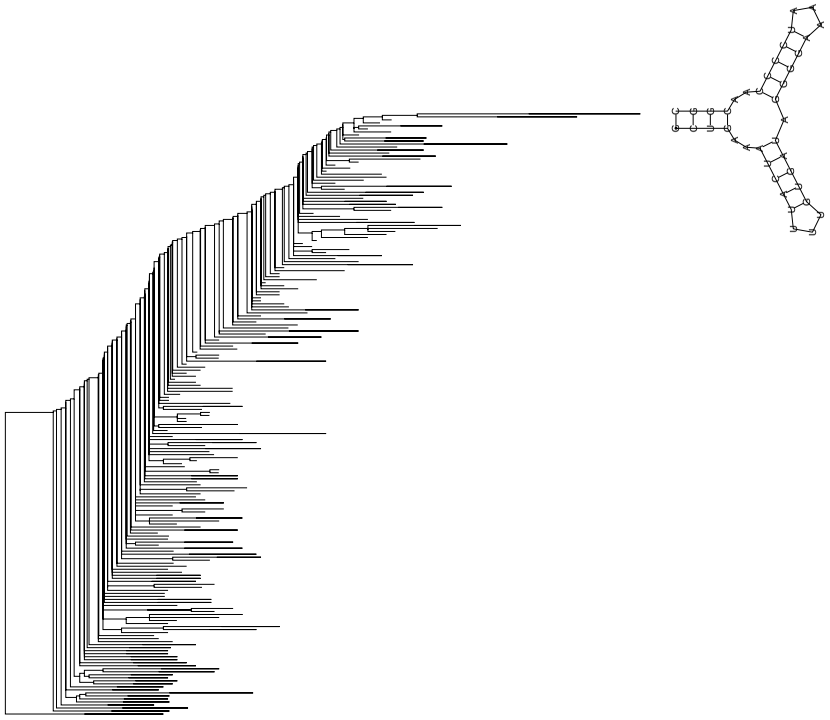
Figure 53: Kinetics of the multi (more than 2) component structures of the sequence space AUGC12 calculated using the new parameters (version 1.4). The mean folding time of all sequences belonging to a component is plotted versus the rank (sorted by size) of that component. The legend gives the rank of the structure each component pair belongs to.

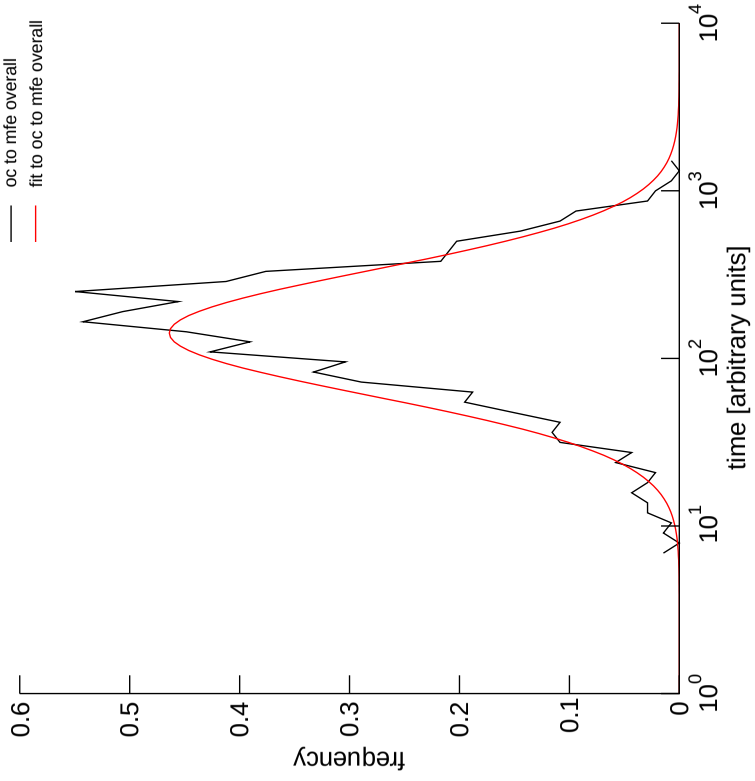
4.2.3 Distribution of Folding Times

An interesting question when investigating the kinetics of RNA secondary structure folding is which underlying distribution the folding times follow. When plotting the classified folding times versus their frequencies obviously nothing looking like a normal distribution can be seen. But if a logarithmic time scale is used the picture changes. It could be shown that the folding times from the open chain to the mfe structure follow a logarithmic normal distribution (in our example the logarithmic normal distribution is simply the result of logarithmising all time values). The following approach is taken to graphically superpose a fitted logarithmic normal distribution over the distribution of folding times derived from simulations. The folding times were classified first. When plotted on a logarithmic time scale the shape of a usual (non-logarithmic) normal distribution reappears. A logarithmic normal distribution can be fitted to the curve by transforming the time values into their logarithms. One can now fit a non-logarithmic normal distribution to the results. This is done by calculating the mean and the variance of the data to obtain estimators for location and variation parameters. Using those parameters, a normal distribution is calculated which can be scaled using a normalisation factor. This normalisation factor is the ratio of the area delimited by the normal distribution curve to the area delimited by the data curve. The latter is estimated by summing the area of the histogram bars of all the classes.

The folding times were created using the program `kinfold` [12]. The folding of every sequence was simulated 1000 times as described previously (see figure 38).

Figure 54: (next page) The sequence having the lowest maximum barrier and the lowest standard deviation of the tested sequences folding into the tested Y-shape structure. There is no major trap that obstructs a direct folding path from the open chain (oc) into the minimum free energy (mfe) structure. The mfe structure is shown below the corresponding leaf of the barrier tree on the right hand side. The distribution of folding times and its log-normal fit distribution is plotted as a histogram on the left hand side in black and red colour, respectively.



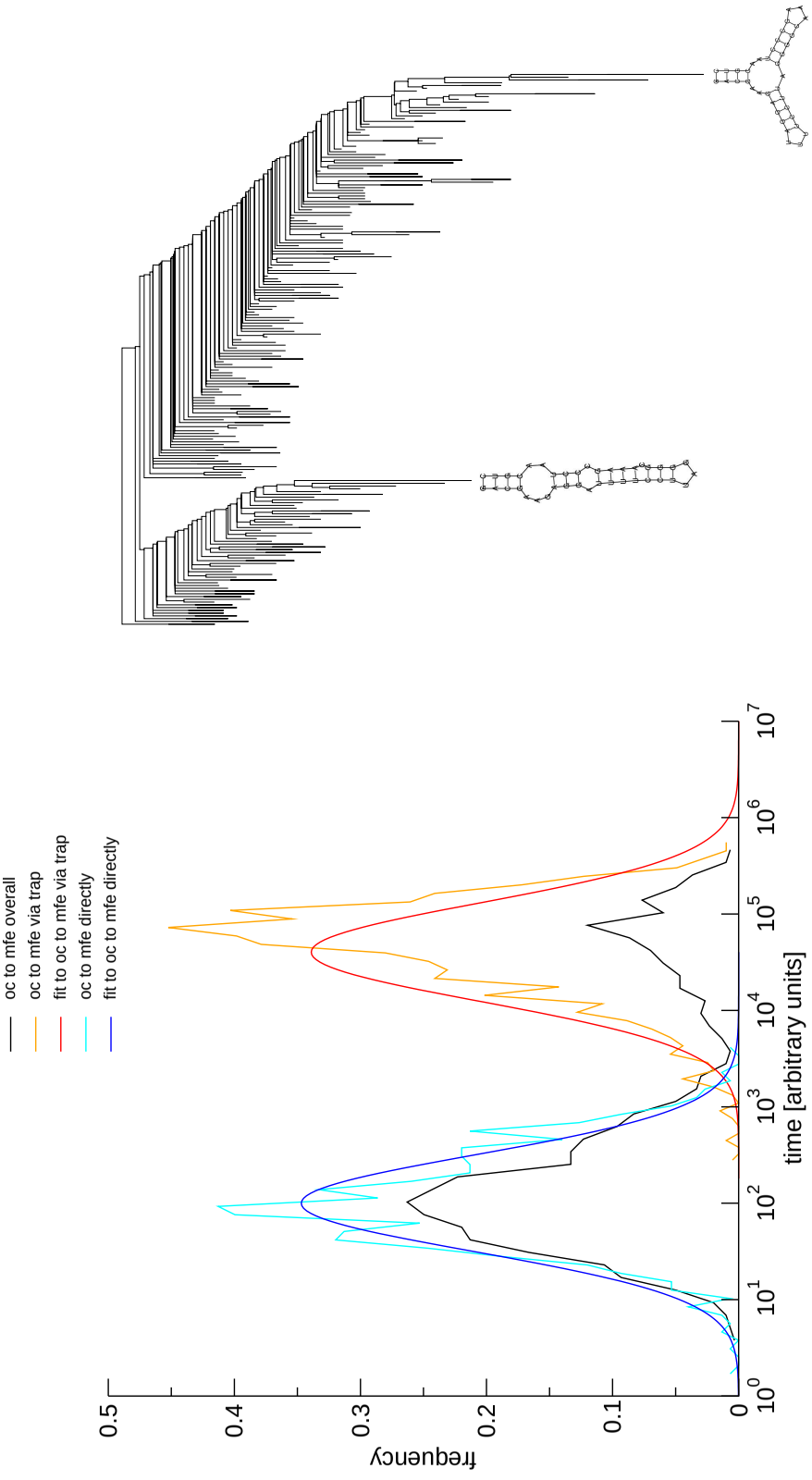

oc to mfe overall
fit to oc to mfe overall
frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
10^0
10^1
10^2
10^3
10^4
time [arbitrary units]

The logarithm of the folding times was calculated, the resulting values were classified and the frequency of each class was plotted against the upper class limit. A normal distribution was fitted to this distribution of logarithmic folding times. In the simple case this works perfectly for sequences whose barrier tree shows no major local minimum that could act as a trap. If other major local minima than the mfe structure exist on the barrier tree of a sequence, one can detect each of these local minima as separate peaks in the frequency distribution plot of logarithmic folding times. Each of these peaks follows its own normal distribution.

The simple case shown in figure 54 examined the sequence having the lowest maximum barrier and the lowest standard deviation of the tested sequences folding into the tested Y-shape structure. The barrier tree shows no major local minimum that could act as a trap. The maximum barrier is 3.60 kcal/mol which can be easily overcome.

An example where two separate peaks can clearly be distinguished can be seen in figure 55. This sequence folds into the tested Y-shape structure as the mfe structure but there is another local minimum structure that acts as a trap. This trap can be seen in the barrier tree and has a barrier of 7.50 kcal/mol to the mfe structure. The trap structure is located under its corresponding branch of the barrier tree. The representation of the mfe structure is located on the far right.

Figure 55: (next page) This sequence `GACGAAGAGGAUUUUCCUUAGGGGCAAAGCCCUAACGUC` folds into the tested Y-shape structure as the mfe structure but there is another local minimum structure that acts as a trap. The barrier tree on the right hand side clearly reveals two different folding funnels. The structures on the bottom of these funnels are shown below the corresponding leafs, the trap structure on the left and the mfe structure on the right hand side. Two separate distributions of folding times can clearly be distinguished in the plot on the left hand side (shown in black colour). The distributions obtained from folding simulations without visiting the trap structure and simulations where the structure is always visited are plotted in cyan and orange, respectively. Fitted log-normal distributions are plotted in blue and red, respectively.


oc to mfe overall
oc to mfe via trap
fit to oc to mfe via trap
oc to mfe directly
fit to oc to mfe directly
frequency
0
0.1
0.2
0.3
0.4
0.5
time [arbitrary units]
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7

The black line connects the upper right limits of the histogram bars representing a histogram of folding times from the open chain to the mfe structure. The cyan line represents the folding time distribution of another simulation of foldings from open chain to mfe directly without visiting the trap structure. The blue line shows a logarithmic normal distribution fitted to the cyan histogram. A simulation where the trap structure is always visited is represented by the orange histogram and its red fitted log-normal distribution. It is actually a combination of two separate simulations. A first one from open chain to the trap structure and a second one from the trap to the mfe structure. The mean folding time of the first simulation is then added to each time value of the second one to obtain the histogram shown in orange. The first peak represents the folding simulations of sequences that directly fold into the mfe structure without reaching a structure of the trap's folding funnel. The second peak represents the folding simulations of sequences that first fold into the trap structure and afterwards have to overcome the barrier to reach the mfe structure.

Another approach was tried to bring more evidence that we are looking at an logarithmic normal distribution by using a statistical test. A chi squared test should show if the distribution of the data from kinetic folding simulations and a chosen distribution (a normal logarithmic distribution in our case) differ significantly from each other or not. The test was carried out using data presented in figures 54 and 55. A perl module was written to calculate the critical values for the chi squared test. It is called `Statistics::Distributions` and is freely available at
`http://www.tbi.univie.ac.at/~michael/distributions.html` , or
`http://www.cpan.org` or upon request from the author. The test could not prove that there is no difference between the distribution of our example data and a logarithmic normal distribution. Obviously the influence of small local minima for the simple examples investigated were still too strong which resulted in a too large deviation form a logarithmic normal distribution to give a statistical insignificant difference between the two distributions.

Sequences that follow a more complicated folding kinetic often visit a large

number of different minor local minima. The barrier associated with these local minima are often only of an average height but they still show their own distribution of folding times. Looking at the overall distribution of the folding times over the complete folding path we are not able to distinguish single peaks of smaller local minima from each other. The distributions are overlapping and result in a smeared overall distribution that makes it impossible to clearly relate a single local minimum to a peak or even find a clear peak at the diagram of the folding time distribution. Still we can proceed on the assumption that kinetics of RNA secondary structure folding is based on the logarithmic normal distribution of the folding times.

5 Conclusion and Outlook

The goal of this work was to investigate neutral nets and their components by folding of complete sequence spaces and exhaustive enumeration. Using newer parameters and a broad range of sequence spaces differing in nucleotide alphabet and chain length, results similar to those of earlier simulations for a smaller range of sequence spaces and older parameter sets were obtained.

The limits of this work, set by available computer time and memory resources for exhaustive folding, are chain length up to 30 for two letter alphabets (AU and GC), chain length up to 20 for three letter alphabets (AUG and UGC) and chain length up to 16 for the complete natural four letter alphabet.

A small number of common and many rare structures are formed. The distribution of ranks roughly follows a generalised Zipf's law. The number of structures found increases for different alphabets at the same chain length in the following order: #AU < #AUG < #GC < #UGC < #AUGC independently to the folding parameters used. The fraction of sequences that do not form stable mfe structures other than the open chain decreases with an increasing chain length. Common structures are formed by a small fraction of all structures. With increasing chain length this fraction decreases further whereas the fraction of sequences folding into common structures increases with chain length. Extrapolation to large chain lengths suggests that almost all sequences fold into a relatively small number of stable secondary structures. As far as it can be seen from the sequence spaces examined these features are not bound to certain alphabets or parameter sets. We find the neutral networks of the higher ranked (ranked by the size starting at rank one for the largest network) and rarer structures to be more often split into a large number of components especially if very small components are ignored. Common structures tend to show a single giant component or a number of up to four components and sometimes some very small additional ones. If a structure decomposes into two to four almost equal sized components they often differ clearly in their base composition which can be explained by structural elements that can form additional base pairs.

Neutral evolution is possible on neutral nets of RNA sequences that fold into the same secondary structure especially on components of such neutral nets where random drift is possible. The set of compatible sequences includes such sets of neutral nets and their components as subsets. Single molecules can form two or more structures if they are elements of intersections of such sets of compatible sequences. Those molecules can fold into their minimum free energy structure and into suboptimal structures their sequence is compatible with. A population can on the other hand drift by mutation and selection on the component of its neutral net and become (additionally) an element of a set of sequences that is compatible to another structure. This structure may have a higher fitness which would make it likely that the population will switch to the corresponding neutral net (of sequences that have this mfe structure) if it can be achieved by an existing mutational operator. Therefore intersections of sets of compatible sequences are the places where steps of increasing fitness occur after periods of random drift on neutral nets. Building up on the neutral nets and their components obtained in this work those intersections can be identified and further investigated.

Another interesting topic to look at are generic properties of neutral networks. For example, one could test whether the nets resemble random, small world, or scale free features [2, 82]. Work on this subject [49] is already in progress.

In contrast to an often stated hypothesis efficiency of RNA folding, meaning mean folding times and fraction of successful trajectories, is not influenced by the minimum free energy of the examined RNA sequence. This could be clearly shown on 2311 different RNA sequences. The mean folding time depends largely on the maximum energy barrier on the folding path. A larger barrier takes more time to be overcome and therefore results in a longer folding time. The largest barrier itself depends on the size of a RNA. A longer RNA sequence results in a higher maximum barrier which has already been shown earlier [48]. This is compatible to the observation that longer RNAs have longer folding times. Another property, the barrier height, correlates with the basin size of the corresponding local minimum. A larger basin brings about a larger barrier.

This does not mean that the largest basin always has to be the one with the largest barrier but very often this is true. The choice of a particular alphabet used has no influence on the relation between the size of the largest barrier and the mean folding time. A reduced (two letter) alphabet shows the same behaviour compared to the complete four letter alphabet (AUGC).

The height of the largest barrier was found to be a reasonable measure for estimating mean folding times. This measure, however, ignores the fact that the local minimum associated with the highest barrier often is not visited at all. Some other trap structures may play a more important role on the folding path. Therefore the barrier height of the most frequently visited local minimum sometimes seems to be a better measure to describe the mean folding time. This of course requires knowledge of this local minimum which, with the current algorithm, involves costly kinetic simulations using selected candidate stop structures.

Further investigations should be made to find a more precise way to derive the distribution of folding times from the corresponding barrier trees. The maximum barrier is a starting point for estimates as it correlates well with the mean folding time. A way to estimate the variance is still to be developed. Maybe a heuristic can be found to improve the guesses of the two parameters. The distribution of folding times was fit to a log-normal distribution. Statistical tests however were not sufficiently certain to allow for a reliable identification of this fit because the influence of overlapping distributions lead to deviations from the expected overall distribution. Even in cases were only one major local minimum is reflected by a distribution with a single peak, the distributions may be overlay by effects resulting from a large number of small local minima. As a result the shape of the log-normal distribution is obscured. Analyses of folding dynamics on entire barrier trees by proper statistical treatments was recently shown to improve results [11].

6 Appendix

To give an impression of which mfe structures occur and their sizes as well as the number and sizes of their components the data obtained by complete enumeration of the sequence spaces of the chain length 12 of different alphabets are given in the following. The complete data of all calculated sequence spaces are available at `http://www.tbi.univie.ac.at/~michael/SOCdata.html`.

Rank	Structure	No. of sequence	No. of components	Sequence of components
1	`............`	3942	1	3942
2	`.(((....))).`	43	2	24 19
3	`(((....)))..`	39	2	25 14
4	`..(((....)))`	36	2	26 10
5	`((((....))))`	16	2	8 8
6	`.((....))...`	8	2	4 4
7	`...((....)).`	6	2	4 2
8	`..((....))..`	6	2	4 2

Table 7: The sequence of components of AU12.

Rank	Structure	No. of sequence	No. of components	Sequence of components
1	(((....)))..	332	1	332
2	(((...)))...	308	1	308
3	.....	294	1	294
4	..(((....)))	252	1	252
5	((((...)))).	252	1	252
6	.(((....))).	250	1	250
7	.((((...))))	230	1	230
8	(((.....))).	204	2	102 102
9	...(((...)))	192	1	192
10	..(((...))).	178	2	92 86
11	.(((.....)))	172	2	92 80
12	((((....))))	172	1	172
13	.(((...)))..	157	2	82 75
14	((....))....	122	1	122
15	..((....))..	100	2	70 30
16	(((......)))	98	2	56 42
17	((.....))...	91	4	44 26 14 7
18	..((.....)).	81	4	38 17 14 12
19	.((.....))..	76	4	24 24 16 12
20	.((....))...	66	2	46 20
21	...((....)).	64	2	48 16
22	((...)).....	62	3	31 27 4
23	.((......)).	61	2	31 30
24	((......))..	51	3	37 8 6
25	....((...)).	36	2	31 5
26	..((...))...	34	2	18 16
27	.((...))....	31	2	24 7
28	((.......)).	30	2	19 11
29	...((.....))	29	4	19 5 4 1
30	...((...))..	28	2	20 8
31	....((....))	18	2	15 3
32	..((......))	17	2	13 4
33	.((.......))	6	1	6
34	((.((...))))	1	1	1
35	((((...)).))	1	1	1

Table 8: The sequence of components of GC12.

Rank	Structure	No. of sequence	No. of components	Sequence of components
1	.....	493799	1	493799
2	((....))....	4562	4	4276 133 99 54
3	...((....)).	4259	3	3849 283 127
4	.((....))...	3691	4	3337 225 120 9
5	..((....))..	3599	4	3253 226 111 9
6	.(((....))).	3376	2	2460 916
7	(((....)))..	3090	2	2497 593
8	(((.....))).	2206	1	2206
9	.....(....).	1949	1	1949
10	..(((....)))	1877	3	1575 210 92
11	.(....).....	1715	1	1715
12	....((....))	1581	1	1581
13	..(....)....	1556	1	1556
14	....(....)..	1548	1	1548
15	...(....)...	1528	1	1528
16	((((....))))	638	2	505 133
17	((((...)))).	429	1	429
18	(..(....).).	20	2	10 10
19	(.(....)..).	18	2	10 8

Table 9: The sequence of components of AUG12.

Rank	Structure	No. of sequence	No. of components	Sequence of components
1	.....	299786	1	299786
2	(((....)))..	17063	1	17063
3	.(((....))).	15200	1	15200
4	(((.....))).	14999	1	14999
5	(((...)))...	14635	1	14635
6	..(((...))).	11253	1	11253
7	..(((....)))	10544	1	10544
8	.(((...)))..	9957	1	9957
9	.(((.....)))	9750	1	9750
10	((((...)))).	9663	1	9663
11	((.....))...	9616	1	9616

Rank	Structure	No. of sequence	No. of components	Sequence of components
12	((....))....	9281	1	9281
13	...(((...)))	8917	1	8917
14	..((.....)).	7848	1	7848
15	.((((...))))	7380	1	7380
16	...((....)).	7016	1	7016
17	.((.....))..	6743	1	6743
18	((((....))))	6242	1	6242
19	.((....))...	6076	1	6076
20	..((....))..	5750	1	5750
21	(((.....)))	5503	1	5503
22	.((......)).	5477	2	5423 54
23	((......))..	5145	1	5145
24	((...)).....	5027	1	5027
25	....((...)).	3684	1	3684
26	...((.....))	3493	1	3493
27	..((...))...	2945	2	2612 333
28	.((...))....	2866	3	2624 192 50
29	...((...))..	2819	2	2319 500
30	....((....))	2570	1	2570
31	((.......)).	2297	2	1990 307
32	..((......))	1521	1	1521
33	.((.......))	197	2	111 86
34	(((....).)).	44	7	25 8 4 2 2 2 1
35	((((...)).))	36	2	22 14
36	(.((...)).).	24	1	24
37	((.((...))))	20	1	20
38	((.(....).))	19	8	8 3 2 2 1 1 1 1
39	.(.((...))).	18	1	18
40	.(((....).))	9	2	8 1
41	(.((....))).	6	1	6
42	((.(....))).	2	1	2

Table 10: The sequence of components of UGC12.

Rank	Structure	No. of sequence	No. of components	Sequence of components
1	.....	11477941	1	11477941
2	(((....)))..	339220	1	339220
3	(((.....))).	303966	1	303966
4	.(((....))).	299733	1	299733
5	(((...)))...	246489	1	246489
6	((.....))...	239405	1	239405
7	((....))....	226298	1	226298
8	..(((....)))	217133	1	217133
9	..((.....)).	211976	1	211976
10	..(((...))).	209046	1	209046
11	.(((.....)))	202749	1	202749
12	...((....)).	199122	1	199122
13	.(((...)))..	195295	1	195295
14	.((.....))..	194965	1	194965
15	.((....))...	187071	1	187071
16	..((....))..	179793	1	179793
17	((((...)))).	174009	1	174009
18	...(((...)))	155100	1	155100
19	.((......)).	152020	1	152020
20	((......))..	148586	1	148586
21	.((((...))))	130752	1	130752
22	((((....))))	123341	1	123341
23	(((......)))	114691	1	114691
24	....((....))	88118	1	88118
25	((.......)).	84118	1	84118
26	...((.....))	72335	1	72335
27	((...)).....	70709	1	70709
28	.((...)))....	66657	1	66657
29	....((...)).	66605	1	66605
30	..((...))...	64633	1	64633
31	...((...))..	63278	1	63278
32	.....(....).	38659	4	34888 2457 858 456
33	..((......))	37196	1	37196
34	....(....)..	34091	4	30526 2062 1043 460
35	...(....)...	31069	4	27533 2018 1079 439
36	(....).....	30463	2	26719 3744

Rank	Structure	No. of sequence	No. of components	Sequence of components
37	..(....)....	30218	4	26564 2100 1140 414
38	.(....).....	29908	4	26150 2228 1096 434
39	.((.......))	19443	1	19443
40	((........))	8147	2	4574 3573
41	((.(....).))	2631	12	2326 158 61 36 28 7 5 4 3 1 1 1
42	(((....).)).	2085	5	2068 9 3 3 2
43	((.(....))).	1653	8	1393 116 115 14 8 3 2 2
44	((((...)).))	758	2	732 26
45	(.((....))).	755	1	755
46	.(((....).))	754	8	549 71 67 35 21 9 1 1
47	.((.(....)))	674	6	564 52 35 17 5 1
48	((.((...))))	555	1	555
49	.(.(....).).	412	2	396 16
50	..(.(....)).	384	2	340 44
51	.(.((...))).	366	2	238 128
52	.(.(....))..	356	2	314 42
53	(.(....))...	337	2	288 49
54	.(((...)).).	244	1	244
55	(((.....).))	182	2	127 55
56	(..(....).).	159	8	54 51 42 4 2 2 2 2
57	(.(....)..).	140	9	44 43 25 12 6 4 2 2 2
58	((.(.....)))	96	2	56 40
59	..((....).).	94	2	82 12
60	(.((...)).).	90	1	90
61	.((....).)..	75	2	64 11
62	(.(....).)..	68	1	68

Table 11: The sequence of components of AUGC12.

References

[1] 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.

[2] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[3] 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.

[4] J. L. Bentley and R. Sedgewick. Fast algorithms for sorting and searching strings. In *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms New Orleans*, pages 360–369, 1997.

[5] J. L. Bentley and R. Sedgewick. Ternary search trees. *Dr. Dobbs Journal*, 23(4):20–22, 24–25, 1998.

[6] N. Carmi, S. R. Balkhi, and R. R. Breaker. Cleaving DNA with DNA. *Proc. Natl. Acad. Sci. USA*, 95:2233–2237, 1998.

[7] F. H. C. Crick. Origin of the genetic code. *J. Mol. Biol.*, 38:367–379, 1968.

[8] J. Cupal, C. 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, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 88–91, Menlo Park, CA, 1997. AAAI Press.

[9] M. Eigen. The origin of genetic information: viruses as models. *Gene*, 135:37–47, 1993.

[10] M. Fekete, I. L. Hofacker, and P. F. Stadler. Prediction of RNA base pairing probabilities using massively parallel computers. *J. Comp. Biol.*, 7:171–182, 2000.

[11] C. Flamm. Personal communication.

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

[13] C. Flamm, I. L. Hofacker, and P. F. Stadler. RNA *in silico*: The computational biology of RNA secondary structures. *Adv. Complex Syst.*, 2:65–90, 1999.

[14] R. Flores. A naked plant-specific RNA ten-fold smaller than the smallest known viral RNA: the viroid. *C. R. Acad. Sci. III*, 324:943–952, 2001.

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

[16] W. Fontana and P. Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280:1451–1455, 1998.

[17] W. Fontana and P. Schuster. Shaping space: The possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.*, 194:491–515, 1998.

[18] 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.

[19] C. Gardiner. *Handbook of Stochastic Methods.* Springer-Verlag, Berlin, 2nd edition, 1990.

[20] R. F. Gesteland, T. R. Cech, and J. F. Atkins. *The RNA World.* Cold Spring Harbor Laboratory Press, second edition, 1999.

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

[22] U. Göbel. *Neutral Networks of Minimum Free Energy RNA Secondary Structures.* PhD dissertation, Universität Wien, 2000.

[23] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monath. Chem.*, 127:355–374, 1996.

[24] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Monath. Chem.*, 127:375–389, 1996.

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

[26] A. P. Gultyaev, F. H. D. van Batenburg, and C. W. A. Pleij. An approximation of loop free energy values of RNA H-pseudoknots. *RNA*, 5:609–617, 1999.

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

[28] C. Haslinger. RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bull. Math. Biol.*, 61:437–467, 1999.

[29] C. Haslinger. *Prediction Algorithms for Restricted RNA Pseudoknots.* PhD dissertation, Universität Wien, 2001.

[30] P. G. Higgs. RNA secondary structure: Physical and computational aspects. *Quart. Rev. Biophys.*, 33:199–253, 2000.

[31] 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.

[32] I. L. Hofacker, W. Fontana, P. F. Stadler, and P. Schuster. `Vienna RNA Package`. `http://www.tbi.univie.ac.at/~ivo/RNA/`, 1994-2000. (Free Software).

[33] I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics on RNA secondary structures. *Disc. Appl. Math.*, 89:177–207, 1998.

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

[35] M. A. Huynen. Exploring phenotype space through neutral evolution. *Journal of Molecular Evolution*, 43:165–169, 1996.

[36] 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.

[37] 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.

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

[39] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK, 1983.

[40] S. Klug and M. Famulok. All you wanted to know about SELEX. *Mol. Biol. Reports*, 20:97–107, 1994.

[41] D. A. M. Konings and P. Hogeweg. Pattern analysis of RNA secondary structure. similarity and consensus of minimal-energy folding. *J. Mol. Biol.*, 207:597–614, 1989.

[42] B. Mandelbrot. On the theory of word frequencies and on related markovian models of discourse. In R. Jakobson, editor, *Structure of Language and its Mathematical Aspects*. American Mathematical Society, Providence, 1961. pp. 190-219.

[43] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

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

[45] M. T. McManus and P. A. Sharp. Gene silencing in mammals by small interfering RNAs. *Nature Reviews Genetics*, 3:737–747, 2002.

[46] T. H. Merrett, H. Shang, and X. Zhao. Database structures, based on tries, for text, spatial, and general data. In *International Symposium on Cooperative Database Systems for Advanced Applications*, pages 316–324, Kyoto, Japan, 1996.

[47] 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.

[48] S. R. Morgan and P. G. Higgs. Barrier heights between groundstates in a model of RNA secondary structure. *J. Phys. A*, 31:3153–3170, 1998.

[49] U. Mückstein. Personal communication.

[50] 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.

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

[52] D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, and M. Blot. Genomic evolution during a 10 000-generation experiment with bacteria. *Proc. Natl. Acad. Sci. USA*, 96:3807–3812, 1999.

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

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

[55] 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.

[56] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997. SFI preprint 95-07-058.

[57] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[58] J. Rogers and G. F. Joyce. A ribozyme that lacks cytidine. *Nature*, 402:323–325, 1999.

[59] N. J. Savill, D. C. Hoyle, and P. G. Higgs. RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum likelihood methods. *Genetics*, 157:399–411, 2001.

[60] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnol.*, 41:239–257, 1995.

[61] P. Schuster. Genotypes with phenotypes: Adventures in an RNA toy world. *Biophys. Chem.*, 66:75–110, 1997.

[62] P. Schuster. Evolution in an RNA world. In L. Stocken and M. Ord, editors, *Foundations of Modern Biochemistry*, volume 4, pages 159–198. Jai Press, 1998.

[63] 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.

[64] P. Schuster and P. F. Stadler. Discrete models of biopolymers. In A. Konopka, editor, *Handbook of Computational Chemistry and Biology*. Marcel Dekker, New York, 2001. in press.

[65] B. A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *CABIOS*, 4:387–393, 1988.

[66] B. A. Shapiro and K. Zhang. Comparing multiple RNA secondarz structures using tree comparisons. *CABIOS*, 6:309–318, 1990.

[67] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 4:213–253, 1971.

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

[69] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA structure prediction. *Eur. Biophy. J.*, 25:115–130, 1996.

[70] N. K. Tanner. Ribozymes: the characteristics and properties of catalytic RNAs. *FEMS Microbiol. Rev.*, 23:257–275, 1999.

[71] C. Tuerk and L. Gold. Systematic evolution of high-affinity RNA ligands of bacteriophage T4 DNA polymerase in vitro. *Science*, 249:505–510, 1990.

[72] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, 96:9716–9720, 1999.

[73] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths? *Bull. Math. Biol.*, 62:799–848, 2000.

[74] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell. Finite populations induce metastability in evolutionary search. *Phys. Lett. A*, 229:144–150, 1997.

[75] D. Voet and J. G. Voet. *Biochemistry.* John Wiley & Sons, New York, 1995.

[76] L. Wall, T. Christiansen, and J. Orwant. *Programming Perl.* O'Reilly & Associates, Sebastobol, 3rd edition, 2000.

[77] 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.

[78] 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.

[79] 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.

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

[81] M. S. Waterman and T. F. Smith. Rapid dynamic programming algorithms for RNA secondary structure. *Adv. Appl. Math.*, 7:455–464, 1986.

[82] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-worl' networks. *Nature*, 393:440–442, 1998.

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

[84] D. S. Wilson and J. W. Szostak. In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.*, 68:611–647, 1999.

[85] S. Wright. The role of mutation, inbreeding, crossbreeding and selection in evolution. In D. F. Jones, editor, *Int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

[86] S. Wright. "Surfaces" of selective value. *Proc. Natl. Acad. Sci. USA*, 58:165–172, 1967.

[87] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structure. *Biopolymers*, 49:145–165, 1999.

[88] G. K. Zipf. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology.* Addison-Wesley, 1949.

[89] M. Zuker, J. A. Jaeger, and D. H. Turner. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucl. Acids Res.*, 19:2707–2714, 1991.

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

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

Curriculum vitae

DI Michael Kospach
* 16. 5. 1972

Schulbildung:	Volksschule in Wien (1978-82) Bundesrealgymnasium 19 in Wien (1982-90) Matura im Mai 1990
Studium:	Lebensmittel- und Biotechnologie (1990-97) an der Universität für Bodenkultur in Wien ERASMUS Programm in Biochemie (1994-95) an der University of Bath, Großbritannien
Diplomarbeit:	"In-vitro Evolution von Hefen" am Institut für Angewandte Mikrobiologie der Universität für Bodenkultur in Wien bei Prof. Dr. Florian Rüker
Dissertation:	"Molecular Evolution of Short RNA Molecules - Neutral Nets in Sequence Spaces and Kinetic Properties of RNA" (1999-2002) am Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien bei Prof. Dr. Peter Schuster