Sequence-Structure Relations of Single RNA Molecules and Cofolded RNA Complexes

Dissertation
zur Erlangung des akademischen Grades
Doctor rerum naturalium

Vorgelegt der
Fakultät für Chemie
der Universität Wien

von
Ulrike Mückstein

Institut für
Theoretische Chemie

im November 2005

Danksagung

Danke an alle, die zum Enstehen dieser Arbeit beigetragen haben

von wissenschaftlicher Seite: Peter Schuster, Ivo Hofacker, Peter Stadler, Christoph Flamm, Stephan Bernhart, Hakim Tafer, Lukas Endler, Kurt Grünberger, Jörg Hackermüller, Michael Kospach, Stefan Müller, Andreas Svrcek-Seiler, Andrea Tanzer, Caroline Thurner, Stefan Washietl, Stefanie Widder, Christina Witwer, Michael Wolfinger.
meiner Familie: Lukas Endler, Raphael Mückstein,
Irene und Heinz Dorfwirth, Eva und Heinz Mückstein, Ingrid, Wolfgang und Katharina Mückstein, Margarte Luise Mückstein.
meinen Freunden: Irene Oberschlick, Alexandra Reisch, Ferdinand Eder, Günter Weber, Nils Leiminger und last but not least: fkk - schlicht wunderbar, dem furchtbaren karpfen klub.

Abstract

In this work we investigated the folding of RNA sequences into secondary structures from different perspectives. One way to describe the relation between single RNA molecules and their secondary structures is the mapping of sequence into structure. In this mapping the preimage is the set of all possible sequences of a given length and alphabet, the image is the set of secondary structures adopted by the sequences. When viewed in the context of biological evolution the sequence is the object under variation, whereas the structure is the target of selection. Thus RNA sequence to structure mapping provides a suitable mathematical model to extract robust statistical properties of the evolutionary dynamics based on RNA replication and mutation.

Within the last years RNA sequence structure maps were analyzed in great detail by the group of Peter Schuster. In the first part of my thesis the results of this analysis were reevaluated by exhaustive folding and enumeration of the sequence spaces $\mathcal{I}_{AUGC}^{(\ell=9)}$ and $\mathcal{I}_{AUGC}^{(\ell=10)}$, where $\{A, U, G, C\}$ is the alphabet and ℓ the sequence length. We were able to prove the results of previous studies by considering only the set of sequences that fold into stable secondary structures, i.e. structures with negative free energies: As expected there are more sequences than structures. The frequency distribution of secondary structures is highly biased. The majority of sequences fold into few common structures. Common structures form extended neutral networks. We examined the topology $\mathcal{I}_{AUGC}^{(9)}$ and $\mathcal{I}_{AUGC}^{(10)}$ by partitioning sequences into components defined by neighbourhood relation and structural criteria. Using stepwise less stringent criteria for the construction of components, we could show that one extensively connected network exists in each of the two sequence spaces. This fact is remarkable because an overwhelming percentage of sequences does not fold at all and we have to expect that the sequences forming stable structures are embedded in a sea of sequences having the open chain as image. The explanation of the apparent paradox is the high dimension of sequence space, nine for $\mathcal{I}_{AUGC}^{(9)}$ and ten for $\mathcal{I}_{AUGC}^{(10)}$: Distances are short in high-dimensional spaces and connected preimages of structures, which are infrequent compared to the open chain, can readily span distances of the diameter of sequence space. Furthermore we could demonstrate that shape space covering, which says that it is sufficient to screen a high dimensional sphere around an arbitrarily chosen sequence in order to find at least

one sequence for every common structure, holds in $\mathcal{I}^{(9)}_{AUGC}$ and $\mathcal{I}^{(10)}_{AUGC}$.

The role of structure neutral networks in evolutionary dynamics has been studied by Peter Schuster's group using computer simulations of RNA population in a flow reactor. We examined relay series of different evolutionary trajectories of the alphabet $\{A, U, G, C\}$ to extract common features of RNA structure optimization. We found that relay series may not only be monotonic sequences of structures with increasing fitness converging to a target structure but may contain structures that are visited more than once in the process of evolutionary optimization. Such structures differ from structures visited only once during the optimization process by having a restricted set of common neighbors showing the same fitness. Furthermore the accessibility relations between common neighbors are highly symmetric, favoring an easy conversion between these structures.

In the last years the importance of sequence specific interaction between two RNA molecules in the regulation of gene expression became increasingly apparent. We developed a method to study significant aspects of RNA-RNA interaction. By applying a modified version of McCaskill's partition function algorithm to RNA co-folding we can provide detailed information about the location of an RNA-RNA interaction, about the structural context of the binding site and also about the energetics of an RNA-RNA interaction. The application of the partition function to this problems allows not only a compensation of errors resulting from an inherent imprecision of secondary structure prediction algorithms but also a more exact description of the interaction than provided by sampling methods.

Zusammenfassung

Der Schwerpunkt dieser Arbeit ist die Faltung von RNA Sequenzen in Sekundärstrukturen. Eine Möglichkeit, die Relation zwischen RNA Sequenzen und ihren Sekundärstrukturen zu beschreiben, ist die Abbildung von Sequenzen auf Strukturen. In dieser Abbildung stellt die Menge aller möglichen Sequenzen einer bestimmten Länge und eines bestimmten Alphabets das Urbild dar, die Bildmenge umfasst alle Strukturen, in die diese Sequenzen falten können. Aus evolutionärer Sicht ist die Sequenz das Objekt unter Variation, während die Struktur der Selektion unterworfen ist. Die Abbildung von Sequenzen auf Strukturen liefert daher ein ideales mathematisches Modell, um robuste statistische Eigenschaften der evolutionären Dynamik, die auf RNA Replikation und Mutation basiert, vorherzusagen.

In den letzten Jahren wurde die Abbildung von Sequenzen auf Strukturen in der Gruppe von Peter Schuster eingehend analysiert. Im ersten Teil meiner Doktorarbeit reevaluieren wir die Ergebnisse dieser Analysen durch vollständige Faltung und Aufzählung der Sequenzräume $\mathcal{I}_{AUGC}^{(\ell=9)}$ und $\mathcal{I}_{AUGC}^{(\ell=10)}$, wobei $\{A, U, G, C\}$ das Alphabet und ℓ die Sequenzlänge ist. Wir beschränken uns in dieser Arbeit auf die Menge der Sequenzen, die eine stabile Sekundärstruktur bilden, d.h. eine Sekundärstruktur mit negativer freier Energie. Durch die Einengung der Datenmenge können wir zeigen, dass die Ergebnisse vorhergegangener Analysen den Tatsachen entsprechen: Wie erwartet gibt es mehr Sequenzen als Sekundärstrukturen. Die Mehrzahl der Sequenzen faltet in wenige häufige Strukturen. Häufige Strukturen bilden ausgedehnte neutrale Netzwerke. Wir untersuchen die Topologie der Sequenzräume $\mathcal{I}_{AUGC}^{(9)}$ und $\mathcal{I}_{AUGC}^{(10)}$ durch die Partitionierung der Sequenzen in Komponenten, die durch Nachbarschaftsbeziehungen und Strukturkriterien definiert sind. Durch die schrittweise Anwendung immer weniger stringenter Kriterien für die Definition einer Komponente, können wir zeigen, dass in beiden Sequenzräumen ein einziges ausgedehntes Netzwerk von stabilen Sequenzen existiert. Dieses Ergebnis ist bemerkenswert, da der überwiegende Anteil der Sequenzen keine stabile Sekundärstruktur bildet, was zu der Annahme berechtigt, dass Sequenzen mit stabiler Sekundärstruktur in einem Ozean von Sequenzen, die in die offene Kette falten, eingebettet sind. Die Erklärung für dieses scheinbare Paradoxon liegt in der hohen Dimension der betrachteten Sequenzräume. $\mathcal{I}_{AUGC}^{(9)}$ hat eine Dimension von 9, $\mathcal{I}_{AUGC}^{(10)}$ eine Dimension von 10. In solch hochdimensionalen Räumen sind Distanzen kurz und die verbundenen Ur-

bilder von Strukturen, die selten im Vergleich zur offenen Kette sind, können problemlos Abstände wie den Durchmesser des Sequenzraumes umfassen. Weiters konnten wir für die Sequenzräume $\mathcal{I}^{(9)}_{AUGC}$ und $\mathcal{I}^{(10)}_{AUGC}$ die Gültigkeit des Prinzips des "shape space covering" beweisen, welches besagt, dass es ausreicht eine hochdimensionale Kugel um jede beliebige Sequenz zu durchsuchen, um mindestens eine Sequenz für jede häufige Struktur zu finden.

Die Rolle neutraler Netzwerke in der evolutionären Dynamik wurde von Peter Schusters Arbeitgruppe durch Computersimulationen von RNA Populationen im Flußreaktor studiert. In dieser Arbeit untersuchen wir Relay Serien unterschiedlicher evolutinärer Trajektorien des Alphabets $\{A, U, G, C\}$, um gemeinsame Eigenschaften der RNA-Struktur Optimierung zu finden. Wir haben herausgefunden, dass Relay Serien nicht nur als monotone Abfolge von Strukturen mit zunehmender Fitness, die zu einer gemeinsamen Zielstruktur konvergiert, vorliegen, sondern auch Strukturen, die im Prozess der evolutionären Optimierung mehr als einmal auftauchen, enthalten können. Diese Strukturen unterscheiden sich von Strukturen, die im Laufe des Optimisierungsprozesses nur einmal gefunden werden, durch eine limitierte Anzahl häufiger Nachbarstrukturen mit gleicher Fitness. Ausserdem ist die Erreichbarkeitsrelation zwischen den häufigen Nachbarn hochgradig symmetrisch, was eine leichte Umwandlung dieser Strukturen ineinander unterstützt.

In den letzen Jahren wurde die Bedeutung der sequenzspezifischen Wechselwirkungen zweier RNA Moleküle in der Regulation der Genexpression erkannt. Wir haben eine Methode entwickelt, mit der man bedeutende Aspekte von RNA-RNA Wechselwirkungen untersuchen kann. Durch die Anwendung einer modifizierten Version von McCaskills "partition function" Algorithmus können wir detailierte Informationen bezüglich der Lage einer RNA-RNA Interaktion, bezüglich des strukturellen Kontexts der Bindestelle und bezüglich der Energetik der RNA-RNA Wechselwirkung geben. Die Anwendung der Zustandssumme auf unser Problem ermöglicht nicht nur eine Kompensation von Fehlern, die aufgrund der inhärenten Ungenauigkeit von Sekundärstrukturvorhersagealgorithmen auftreten, sondern erlaubt auch eine exaktere Vorhersage der Interaktion als statistische Methoden.

Contents

1 Introduction

Biological evolution formed and still forms the shape of the biosphere as we perceive it today. It deals with the effects of changes in the environment and the heritable information of organisms that result in the alternation of their appearance and leads to a diversity of biological species.
The current theory of biological evolution originates from two epochal contributions by Charles Darwin and Gregor Mendel. In 1865 Gregor Mendel published a paper describing how traits are passed through generations. He realized that many traits are passed to the next generation as pairs of discrete units, which he called genes. In 1985, Charles Darwin proposed natural selection as the basic mechanism for the origin of phenotypic variants and, ultimately, new species. However one great unsolved problem in Darwin's work was the question of how and why species originate. Darwin and his followers were confronted with a seeming paradox: Evolution was described as a continuous process, as a gradual change over time. But obviously distinct species show differences from each other, which suggests that some mechanism had created a discontinuity between them.
A solution was provided by Ernst Mayr, Theodosius Dobzhansky and others by achieving the "neo-Darwinian" synthesis. The neo-Darwinian synthesis brings together Charles Darwin's theory of the evolution of species by natural selection with Gregor Mendel's theory of genetics as the basis for biological inheritance [16,49]. Mayr proposed that species originate when a population of organisms become isolated from the main group by time or geography. During the period of isolation these organisms evolve traits that are different from the main group and the two groups can no longer interbreed [48]. In addition to offering an explanation for the mechanism of specification the neo-Darwinian synthesis incorporates genetics and population biology into studies of evolution, recognizing the importance of mutation and variation within a population. The inclusion of genetics provided the knowledge of how information is carried from one generation to the next. At the beginning of the 20^{th} century Morgan and his colleagues developed the notion

of mutation. They found that genes can change and that these mutations are permanent changes in genes, causing different phenotypes. In 1952 the Hershey-Chase experiments provided the proof that DNA (deoxyribonucleic acid) was the genetic material responsible for the hereditary transfer of information. In 1953 Watson and Crick determined the structure of DNA as two helical chains each coiled round the same axis [84]. Francis Crick also coined the term "Central dogma", which describes the theory that DNA encodes RNA (ribonucleic acid) which encodes proteins, but that information does not flow from proteins to DNA. According to the Central dogma the flow of genetic information can be described in the following way: The genetic information of a cell is contained within its DNA, where the sum of the DNA of an organism is called its genome. The genome is organized into chromosomes that come in pairs. Each pair contains functionally similar but not identical alleles at corresponding loci. An allele is a variant of a gene, the smallest unit of genetic information that is sufficient for the generation of a functional product. The DNA serves as the template for its own reproduction in a process termed replication. Genes are transcribed into RNAs, which are the templates for the production of proteins in a process called translation. The Central dogma postulates that proteins fulfill most structural, catalytic and regulatory functions within the cell. The inclusion of population biology in the neo-Darwinian synthesis results in the description of selection as a process that alters the frequency of genes in a population.

"More recently the classic Neo-Darwinian view has been replaced by a new concept which includes several other mechanisms in addition to natural selection. Current ideas on evolution are usually referred to as the Modern Synthesis" [55]. Futuyma [24] provided an abstract of the tenets of the evolutionary synthesis: In short he states that populations contain genetic variation that arises by random (i.e. not adaptively directed) mutation and recombination. Populations evolve by changes in gene frequency brought about by random genetic drift, gene flow, and especially natural selection.

Phenotypic changes are gradual and diversification comes about by speciation. The last sentence states that speciation is a gradual accumulation of small genetic changes. Today the theory of speciation as a gradual process is in contrast to the theory of "Punctuated Equilibrium", that states that long periods of stasis are followed by rapid speciation. However both models of evolution are able to explain important aspects of the evolutionary process and the debate focuses on questions about the relative contributions of gradual versus punctuated change, the average size of the punctuations, and the mechanism.
Another dogma of Neo-Darwinism has also been challenged in the last decades. The central dogma states that genetic information flows from the DNA via RNA to proteins and establishes proteins as the catalytically and regulatory active components of the cell. Today a vast amount of studies supply evidence that RNA is catalytically as well as regulatory active [2,3,47]. Furthermore the unidirectional flow of information from DNA to RNA to protein has been proven wrong [31,56].

1.1 General context

According to the modern synthesis heritable alternations of the genome occur by random mutation. Random molecular variation in proteins and DNA frequently results in changes that have no influence on the fitness of the individual organism, in other words are selectively neutral [39]. Neutral evolution facilitates adaptive evolution by increasing the number of phenotypes that can be reached with a single mutation from an original phenotype [35]. In this way neutral evolution supplies a mechanism for the maintenance of genetic diversity in natural populations, where diversity allows a parallelization of the search for better adapted variants. Additionally neutrality offers a mechanism of shielding the phenotype (i.e. the sum of traits of an organism exposed to selection) from alternations of the genotype (i.e. the genome) which allows even more variation within a population. The following pages give a short overview of the contributions to the neutral theory of evolution.

In the late 1960s Motoo Kimura introduced the neutral theory of molecular evolution [39]. According to Kimura the vast majority of single-nucleotide exchanges in the genomes of existing species are selectively neutral, i.e. have no influence on the fitness of an individual. As a result, fitness-neutral variants are not subjected to natural selection but spread in a population by genetic drift. Genetic drift is a stochastic process resulting from random sampling in the production of offsprings. Therefore neutral variants, by shielding the phenotype from genotypic diversity, provide a mechanism to maintain the diversity of a population. During the 1970s and 1980s Manfred Eigen, Peter Schuster and John McCaskill proposed a theory for optimization and maintenance of nucleotide sequences under error prone replication. Manfred Eigen introduced the quasi-species theory that describes the evolution of a finite population of asexual replicators at high mutation rates [12, 14]. The quasi-species theory states that the only prerequisite for evolution is an open system with replication far from thermodynamic equilibrium that lives on limited resources. In such a system every sequence is associated with a spectrum of mutants, called the molecular quasi species, which is completely defined by the sequence, assuming a sequence specific fitness distribution and a fixed mutation rate [13]. In this process selection itself assures the maintenance of variability, as selection of a sequence is tantamount to selection of a quasi-species.

In the 1980s and 1990s RNA was studied as a model system for molecular evolution. RNAs are biologically active nucleotide sequences that can fold into well defined structures, enabling them to specifically interact with other molecules and to catalyze biochemical reactions. The secondary structure of RNA molecules, which is the main determinate of their tertiary structure and therefore their function, is readily approximated by different folding algorithms [33, 46, 93]. Walter Fontana et. al. studied the evolutionary optimization of RNA molecules by selecting for specific features of secondary structures generated from different RNA sequences. They found that structure

neutral sequence variants play a key role in evolutionary optimization [19–22]. In the 1990s the group of Peter Schuster studied the importance of neutral variants for molecular evolution. They showed analytically that increasing neutrality leads to an increase of the phenotypic error threshold [23, 62]. The phenotypic error threshold puts a limit on the amount of information maintainable in Darwinian evolution where the important parameter is the average number of neutral base substitutions per replication. This leads to the concept of a neutral network. A neutral network is defined as the set of all genotypes that are neutral with respect to some phenotype. Together with the neighborhood relation "accessible by one mutation" the neutral set is turned into a graph. Walter Grüner et. al. [27, 28] studied the structure of neutral networks for the mapping of RNA sequences into secondary structures using binary alphabets by exhaustive folding and enumeration. Together with the statistical evaluation of RNA networks over the natural alphabet [78, 79] the following picture emerged: The RNA sequence space, constructed by connecting each sequence with all its one-error neighbors, is percolated by extensive structure neutral networks. The sequences of a population may diffuse over this network without loosing the currently optimal structure, until a non neutral mutant with increased fitness is found. The population will then switch to the neutral network of that structure. In this way neutral networks convey an ideal combination of search capacity and robustness to mutation.

To describe the evolutionary process it is important to define accessibility relations between phenotypes. Evolutionary modification of phenotypes, however, requires the modification of the underlying genotype. In the case of RNA, accessibility relations between sequences are readily provided using the Hamming distance (i.e. the number of point mutations between sequences of fixed length) as a natural metric. Accessibility relations between structures, on the other hand, cannot be quantified by a distance but have to be described using weaker notions of neighbourhood. Fontana et.al. [22] provided a measure for the accessibility between RNA structures by using the

probability that one step away from a random point in the neutral network of a structure S_α results in a sequence folding into a structure S_β. This accessibility distribution is not symmetric and therefore no distance. The accessibility distribution is converted into the binary attribute of nearness by defining the neighbourhood of a shape S_α as the set containing S_α and all shapes accessible from S_α above a certain likelihood [18]. The definition of structure neighbors obtained this way is consistent with the formalization of the neighbourhood concept in topology.
One consequence of the shape space topology described above is the fact that pairs of neutral networks approach each other closely at intersection points of the compatible sets in which they are embedded. Where the compatible set of a structure is the set of all sequences that can form all base pairs of this structure [78].The *intersection theorem* guarantees that for any two prescribed secondary structures there is always a non-empty set of compatible sequences [17].

A direct proof for the existence of extended neutral networks and the intersection theorem was provided by the experimental data from Schultes and Bartel [68]. Starting from two phylogenetically unrelated ribozymes with different catalytic activities, whose RNA-conformations had no base pair in common, Schultes and Bartel constructed a RNA sequence that is compatible with both secondary structures and showed both catalytic functionalities. Thus the authors were not only able to track neutral paths of constant structure and full ribozyme function from the mutants to the parents but also showed that two neutral networks approach each other very closely in the surrounding of the chimeric molecule.

1.2 Organization of this work

This work elaborates on different aspects of RNA sequence to secondary structure relations. In section 3.1 and section 4.1 the relation between single RNA sequences and their secondary structures is studied from an evolu-

tionary point of view. Section 3.1 addresses generic features of evolutionary dynamic of RNA molecules using a sequence to secondary structure mapping of RNA molecules. By exhaustive folding and enumeration of the sequence spaces $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ and $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ we can demonstrate that the results obtained by Schusters group hold for this sequence spaces, if a hamming distance of one ($d^h = 1$) and $d^h = 2$ are used as variation operators.

In Section 4.1 we examine common features of RNA structure optimization by simulating the evolution of a population of replicating and mutating RNA molecules in a flow reactor. The course of the evolutionary optimization process is represent by a series of phenotypes leading from the target structure to an initial shape, called the relay series. We find that relay series of a population of RNA sequences may not only be monotonic sequences of structures with increasing fitness converging to a predefined target structure but may contain structures recurring in evolutionary optimization process, even within one and the same relay series. We can show that this recurrence of specific structures in the evolutionary optimization process is governed by the neighborhood context of the recurring structures.

In section 3.2 we approach RNAs from a functional point of view. In the last year it became increasingly apparent that RNA-RNA interactions play an important role in the regulation of gene expression trough sequence specific interactions between RNA molecules. We introduce an variation of McCaskill's partition function algorithm that provides information about the location, the structural context and the free energy of the interaction between a small RNA and its target RNA.

2 Sequence-Structure Mapping of RNA

2.1 RNA secondary structure

Biopolymers as DNA, RNA and Proteins are linear strings composed of several distinct monomers. The monomers of nucleic acids are termed nucleotides, each nucleotide consists of a nitrogenous heterocyclic base (a purine or a pyrimidine), a pentose sugar and a phosphate group. Two classes of nucleic acids are distinguished according to the type of sugar present in the nucleotides: A nucleic acid that contains riboses as sugar residues is termed ribonucleic acid (RNA), a polymer composed of nucleotides including 2-deoxyribose is named deoxyribonucleic acid (DNA). RNA and DNA also differ partially in their base composition. Whereas RNA contains adenine (A) and guanine (G) as purine compounds and the pyrimidine bases cytosine (C) and uracil (U), DNA contains thymine (T) instead of uracil. Especially RNA but also DNA, often incorporates so-called rare, modified base. In DNA and RNA the atoms of the sugar residues are marked by a ' to distinguish them from the atoms of bases.

Nucleic acids polymers consist of nucleotides that are linked by phosphodiester bonds (polynucleotides). The process of polynucleotide synthesis is a polymerization where a pyrophosphate residue is split off. Nucleic acids are synthesized within the cell in a template dependent process, in which the polynucleotide chain grows from the 5' to the 3' terminus. The process of DNA synthesis, during which a double stranded DNA molecule is copied, is termed replication. However, in a process called reverse transcription DNA can also be synthesized using RNA as a template. The process of RNA synthesis is termed transcription, where the template is generally DNA.

The biological function of a biopolymer, such as an RNA molecule or a protein, is mostly determined by its three-dimensional structure. A single stranded nucleic acid sequence generally contains complementary region that

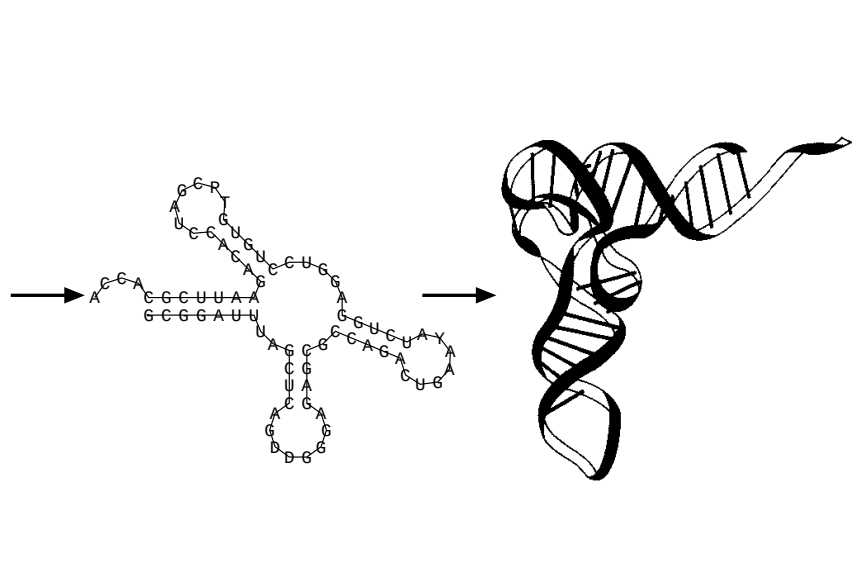


Figure 1: Sequence, secondary structure and schematic representation of the tertiary structure of $tRNA^{phe}$.

are able to form double helices. This results in a pattern of double helical regions interspersed with loops termed the secondary structure of an RNA or DNA. The 3-dimensional arrangement of this secondary structure elements is called the tertiary structure. As an example of RNA sequence to structure relation the first structure experimentally determined, the yeast $tRNA^{phe}$ [77], is shown in figure 1. The energetic contributions of tertiary interactions are weaker than the contributions of the interactions generating the secondary structure. RNA folding can therefore be viewed as a hierarchical process in which secondary structure formation precedes the formation of the tertiary structure.

The assembly of RNAs into biologically functional structures is specified no only by the sequence of the RNA. In addition, counter ions are needed to neutralize the negative charge of the phosphates and to enhance the specificity of the RNA interactions. Multivalent metal ions such as Mg^{2+} stabilize RNA structure more efficient than monovalent ions, because the number of ions that must be localized around the RNA is smaller [88]. Other parame-

ters that influence RNA stability are temperature and the pH value.

Except in some viruses RNA molecules are single stranded. In RNA double helical regions consist mainly of Watson-Crick (GC and AU) base pairs or the slightly less stable GU pairs. The stacking energy of these *allowed* base pairs is the major driving force for RNA structure formation. Other combinations of pairing nucleotides, which are called *non-canonical* pairs and occur especially in tertiary structure motifs, are not considered in secondary structure prediction [32].

2.1.1 Loop decomposition of secondary structures

As discussed in the preceding section the secondary structure of RNA molecules contributes the dominant part of the three dimensional folding energy. It can be successfully used in the interpretation of RNA function and reactivity and is frequently conserved in evolution. [40, 43]. In this work we will describe the structure of an RNA molecule by its secondary structure.
From a mathematical point of view a secondary structure is defined as a set S of base pairs (i, j) where $i < j$. Two base pairs (i, j) and (k, l) with $i \leq k$ are compatible with a secondary structure S if

(i) $i = k$ iff $j = l$, and

(ii) $i < k$ implies $i < k < l < j$

In the classical definition of an RNA secondary structure [83] each base interacts with at most one other nucleotide, which is stated in (i). Condition (ii) specifies that base pair are not allowed to cross, excluding pseudo-knots.

Any secondary structure S can be uniquely decomposed into stems, loops, and external elements. A vertex i is called interior to the base pair (k, l) if $k < i < l$. If there is no base pair (p, q) $k < p < q < l$ such that $p < i < q$

we will say that i is immediately interior to the base pair (k, l). A base pair (p, q) is said to be (immediately) interior if p and q are (immediately) interior to (k, l).

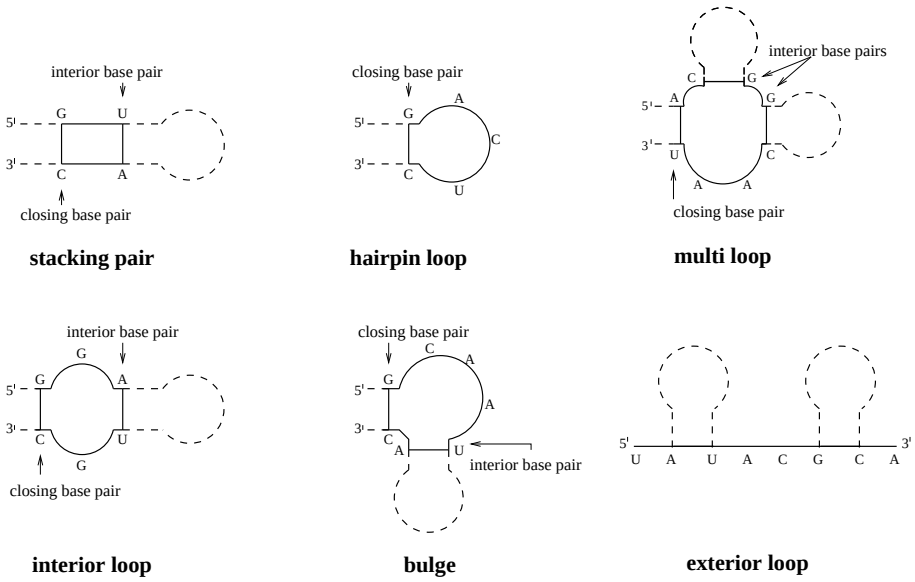


Figure 2: Basic loop types of RNA secondary structure. The different loop types in RNA are distinguished by their degree, where the degree of a loop is given by 1 plus the number of terminal base pairs of stems which are interior to the closing pair of the loop. A loop of degree 1 is called hairpin (loop), a loop of a degree larger than 2 is called multiloop. A loop of degree 2 is called bulge if the closing pair of the loop and the unique base pair immediately interior to it are adjacent; otherwise a loop of degree 2 is termed interior loop. Two stacked base pairs form an interior loop with size 0.

A stem consists of subsequent base pairs $(p-k, q+k), (p-k+1, q+k-1), ..., (p, q)$ such that neither $(p-k-1, q+k+1)$ nor $(p+1, q-1)$ form a base pair. The length of the stem is $(k+1)$, $(p-k, q+k)$ is the terminal base pair of the stem. Notice that isolated single base pairs are considered as stems of length one.

A loop consists of all unpaired vertices which are immediately interior to some base pair (p, q), which is termed the "closing" pair of the loop. The size of the loop is given by the number of vertices immediately interior to the closing pair (p, q). Figure 2 gives an overview of the basic loops types in RNA secondary structures.

A vertex is termed external if it is unpaired vertex and does not belong to a loop. An external element is a collection of adjacent external vertices. If it contains the vertex 1 or n it is a free end, otherwise it is called joint.

2.1.2 RNA secondary structure representation

RNA secondary structures can be depicted using a variety of different representation forms, see Fig.3. Conventionally RNA secondary structures are draw as a planar graph constructed from combinations the of different loop types shown in figure 2.
Secondary structure graphs are outer-planar, i.e., they can be drawn in such a way that the backbone forms a circle and all base pairs are represented by chords that must not cross each other.
An alternative representation method for RNA secondary structures is the *dot-bracket* or *string* notation: The secondary structure is encoded as a linear string with balanced parentheses, '(' and ')', representing base pairs and dots, '.', representing unpaired positions. The "dot-bracket" notation is used as a convenient notation in input and output of the `Vienna RNA Package`, a software for folding and comparing RNA molecules [33].
Secondary structures can also be represent as a *dot plot*. A base pair (i, j) is indicated by a square in row i and column j in the upper right side of the dot plot, the area of the square is proportional to the predicted base-pairing probability. A square in row j and column i in the lower left side of the dot plot indicates a base pair (i, j) which is part of the minimum-free-energy structure of the sequence. The *mountain*-representation (or *mountain plot*)

is especially useful to compare large structures [34]. The three symbols of the string representation '.', '(' and ')' are assigned to three directions "horizontal', 'up' and 'down' in the plot. The structural elements match certain secondary structure features. *Peaks* correspond to hairpins. The symmetric slopes represent the stems enclosing the unpaired bases in the hairpin loop, which appear as a plateau. *Plateaus* represent unpaired bases. When interrupting sloped regions they indicate bulges or interior loops. *Valleys* indicate the unpaired regions between the branches of a multi loop or, when their height is zero, they indicate external vertices.

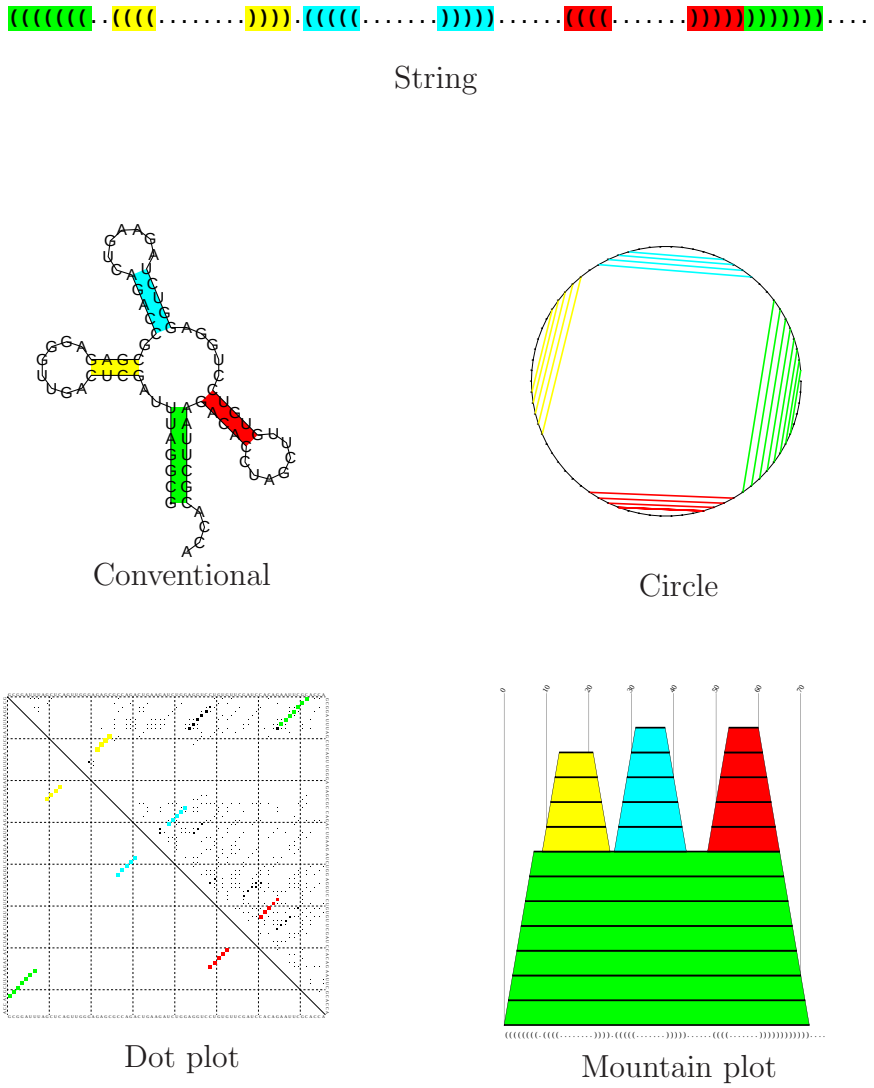


Figure 3: Different representations of RNA secondary structure. All drawings show the same structure and use the same colors to mark the different stems.

2.1.3 Calculation of the minimal free energy

The prediction of the secondary structure of a single RNA molecule is a classical problem in computational biology. The simplest approach to predicting the secondary structure of RNA molecules is to find the configuration with the greatest number of paired bases. A dynamic programming algorithms for RNA folding using this simple approach was introduced by Nussinov [59]. The basic idea is that each base pair in a secondary structure divides the structure in an interior and an exterior part that can be treated separately, see figure 4. The problem of finding, say, the optimal structure of a subsequence $[i, j]$ can thus be decomposed into the subproblems on the subsequence $[i+1, j]$ (provided i remains unpaired) and on pairs of intervals $[i+1, k-1]$ and $[k+1, j]$ (provided i forms a base pairs with some position $k \in [i, j]$).

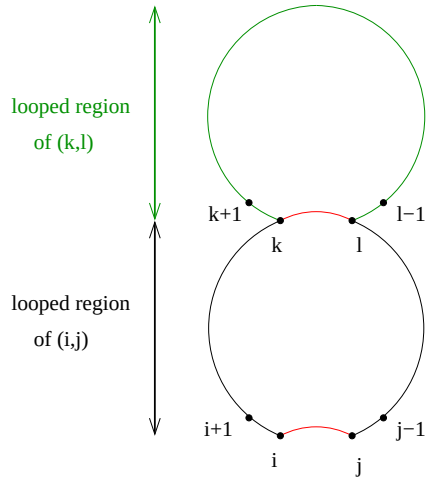


Figure 4: Looped regions in RNA secondary structures are defined by their corresponding closing base pair. Each base pair divides the structure into distinct parts: In the figure base pair (k, l) divides the structure into a part interior to (k, l), indicated in green, and an exterior part, shown in black.

In the more realistic "loop-based" energy model the same approach is used. The simplest loop considered in this model are two adjacent stacked base

pairs. These stacked pairs confer most of the stabilizing energy to a secondary structure, whereas most other loop types destabilize the structure by entropic contributions. In addition, one has to distinguish between the possible types of loops that are enclosed by a base pair because hairpin loops, interior loops, and multiloops all come with different energy contributions. The energy parameters for small loops are strongly dependent on the sequence of the loop, these parameters are tabulated exhaustively [44]. For larger loops sequence dependence is conferred only through the base pairs closing the loop and the unpaired bases directly adjacent to the closing pair. For these loops the energy is given by

$$F_L = F_{\text{mismatch}} + F_{\text{size}} + F_{\text{special}}$$

where F_{mismatch} is the contribution from unpaired bases inside the closing pair and the base pairs immediately interior to the closing pair. The last term is used e.g. to assign bonus energies to unusually stable tetra loops. Polymer theory predicts that for large loops the size should grow logarithmically. The energy contribution of multiloops, F_{L_m}, is approximated by a linear decomposition to keep the recursive scheme tractable.

$$F_{L_m} = a + b(x-1) + cu, \tag{1}$$

where u is the number of unpaired bases in the loop and x the degree of the loop, a, b and c are constants. The "loop-based" energy model assumes that stacking base pairs and loop entropies contribute additively to the free energy of an RNA secondary structure [45, 46, 80]. The free energy $F(S)$ of a given secondary structure S is thus the sum of the terms for its loops.

$$F(S) = \sum_{L \in S} F_L$$

For the computation of the minimum free energies, we use the following recursion: Let F_{ij} be the minimum free energy of the sequence interval $[i, j]$, and let C_{ij} be the minimum free energy under the condition that (i, j) form

a base pair, F_{ij}^M holds the minimum free energy given that $[i, j]$ lies within a multi loop and contains at least one helix, while F^{M1} includes one and only one helix. For easier comprehensibility, dangling end contributions are neglected in these formulas.

$$\begin{aligned}
F_{ij} &= \min \begin{cases} F_{i+1,j} \\ \min_{i<k\leq j} \{C_{ik} + F_{k+1,j}\} \end{cases} \\
C_{ij} &= \min \begin{cases} \mathcal{H}(i,j) \\ \min_{i<k<l<j} \{C_{kl} + \mathcal{I}(i,j;k,l)\} \\ \min_{i<k<l<j} \left\{F_{i+1,k-1}^M + F_{k,j-1}^{M1} + a\right\} \end{cases} \\
F_{ij}^M &= \min \begin{cases} F_{i+1,j}^M + c \\ \min_{i<k\leq j} \left\{C_{ik} + b + F_{k+1,j}^M\right\} \end{cases} \\
F_{ij}^{M1} &= \min \begin{cases} F_{i,j-1}^{M1}; C_{ij} + b \end{cases}
\end{aligned} \tag{2}$$

The free energy for a Hairpin loop closed by base pair (i, j) is given by $\mathcal{H}(i, j)$, the free energy of an interior loop closed by pairs (i, j) and (k, l) is given by $\mathcal{I}(i, j; k, l)$ and a, b and c are contributions for closing a multi loop, extending it by one stack and one base, respectively, see equation (2.1.3).
This recursion is $\mathcal{O}(n^4)$, but by restricting the length of allowed Interior loops, which is a physically reasonable constraint, the time complexity is reduced to $\mathcal{O}(n^3)$. Thermodynamic folding algorithms used for free energy minimization were originally conceived by Zucker, Stiegler and Sankoff [81, 93, 94] and are based on an application of dynamic programming to the RNA problem [83].

2.1.4 Equilibrium partition function

As discussed in subsection 2.1.3 the approximate decomposition of the free energy of a secondary structure into a sum of terms involving adjacent stacked pairs and constrained loops allows the prediction of a single optimal free energy secondary structure. It is, however, also possible to examine the full

ensemble of probable alternative equilibrium structures. Mc Caskill [50] introduced an algorithm for the calculation of the full equilibrium partition function by dynamic programming, which scales as $\mathcal{O}(n^3)$, where n is the length of the sequence. The knowledge of the equilibrium ensemble of structures allows the computation of the matrix of base pair binding probabilities between all nucleotides in the RNA sequence. This base pair binding probabilities represent the sum over all equilibrium weighted structures in which a selected base pair occurs. Consequently the base pair probability matrix summarizes the information about the global ensemble of structures in equilibrium.
In this section we will discuss the dynamic programming algorithm for the calculation of the partition function and the back-summing calculation for base pair probabilities. Algorithms that are designed to enumerate all structures (with a below-threshold energy) [89], that compute averages over all structures [50], or that sample from a (weighted [9] or unweighted [79]) ensemble of secondary structures, need to make sure that the decomposition of the structures into substructures is unique, so that each secondary structure is counted once and only once in the dynamic programming algorithm.

For the dynamic programming calculation of the partition function Z we will apply the same notation as in chapter 2.1.3. The partition function Z is given by

$$Z = \sum_{S} e^{-[F(S)\beta]} \tag{3}$$

where $\beta = 1/kT$ and F_L is the contribution of a given loop. To convert the recursions for the calculation of the minimal free energy in equation (2) into recursion permitting the calculation of the partition function the maxima are transformed into sums and the additive contributions in products. Equation (2) is phrased in a way that ensure that each secondary structure is counted once and only once in the dynamic programming algorithm.
The sum over all structures S involves an exponentially increasing number

of terms as the sequence length increases. To handle the increasing number of terms, Mc Caskill [50] introduces the restricted partition function, $Z^b[i,j]$. $Z^b[i,j]$ is the sum over all possible loops closed by the base pair (i,j), including the contributions of their sub-loops, it corresponds to C_{ij} in equation (2). The calculation of the full partition function for the segment $[i,j]$ is then given by:

$$Z[i,j] = 1.0 + \sum_{i \leq k \leq j} Z^1[i,k]Z[k+1,j]. \tag{4}$$

where the ancillary quantities $Z^1[i,j]$ are defined as

$$Z^1[i,j] = \sum_{i \leq l \leq j} Z^b[i,l], \tag{5}$$

with the initial conditions $Z^b[i,i] = 0$, $Z[i,i] = 1.0$ and $Z[i+1,i] = 1.0$. Starting with the shortest segments one proceeds iteratively to calculate $Z^b[i,j]$ and $Z[i,j]$ until one reaches $Z[1,N]$, which is the full partition function Z [50].

The number of base pairs in a loop closed by (i,j) increases exponentially with the size of the loop. A further decomposition of the free energy, F_L, of a loop closed by (i,j) is therefore necessary. To keep the recursive scheme tractable the energy contribution of multiple loops, F_{L_m}, is approximated by a linear decomposition:

$$F_{L_m} = a + b(x-1) + cu. \tag{6}$$

where u is the number of unpaired bases in the loop and x the degree of the loop, a, b and c are constants. The partition function over all possible loops closed by the base pair (i,j), $Z^b[i,j]$, can then be partitioned into the contributions of all possible hairpin loops, all kinds of interior loops and multiple loops within region $[i,j]$, see equation (7) and figure 5.

$$
\begin{aligned}
Z^b[i,j] &= e^{-[\mathcal{H}(i,j)\beta]} + \sum_{\substack{k,l \\ i<k<l<j}} e^{-[\mathcal{I}(i,j,k,l)\beta]} Z^b[k,l] \\
&+ \sum_{i<k<j} Z^m[i+1,k-1] Z^{m1}[k,j-1] e^{-[(a)\beta]} \qquad (7)
\end{aligned}
$$

A reduction to a cubic algorithm may be obtained if the calculation of the free energy of large interior loops, in which the number of unpaired bases $u > u_m$, only depends on u and not on the position of the pair $(h.l)$ within the loop.

$$
\begin{aligned}
\sum_{k,l} e^{-[\mathcal{I}(i,j,k,l)\beta]} &= \sum_{\substack{k,l \\ u \le u_m}} e^{-[\mathcal{I}(i,j,k,l)\beta]} \\
&+ \sum_{u>u_m} e^{-[\mathcal{I}(i,j,u)\beta]}. \qquad (8)
\end{aligned}
$$

The calculation of the multiloop contribution to the partition function can be reduced to order $\mathcal{O}(n^3)$ by introducing ancillary quantities. The first ancillary quantity is the restricted partition function $Z^{m1}[i,j]$, which yields the partition function over all segments $[i,j]$ that contain one and only one structured section.

$$
Z^{m1}[i,j] = Z^{m1}[i,j-1] + Z^b[i,j] e^{-b} \qquad (9)
$$

The second ancillary quantity $Z^m[i,j]$ comprises the partition function over all segments $[i,j]$ that contain one or more structured sections. The partition of the multiloop contributions into $Z^{m1}[i,j]$ and $Z^m[i,j]$ furthermore ensure that no multiloop contribution is counted twice.
In summary the partition function $Z = Z[1,N]$ can be calculated with a dynamic programming algorithm of cubic order by a recursive application of equation (4) to equation (9) [50].

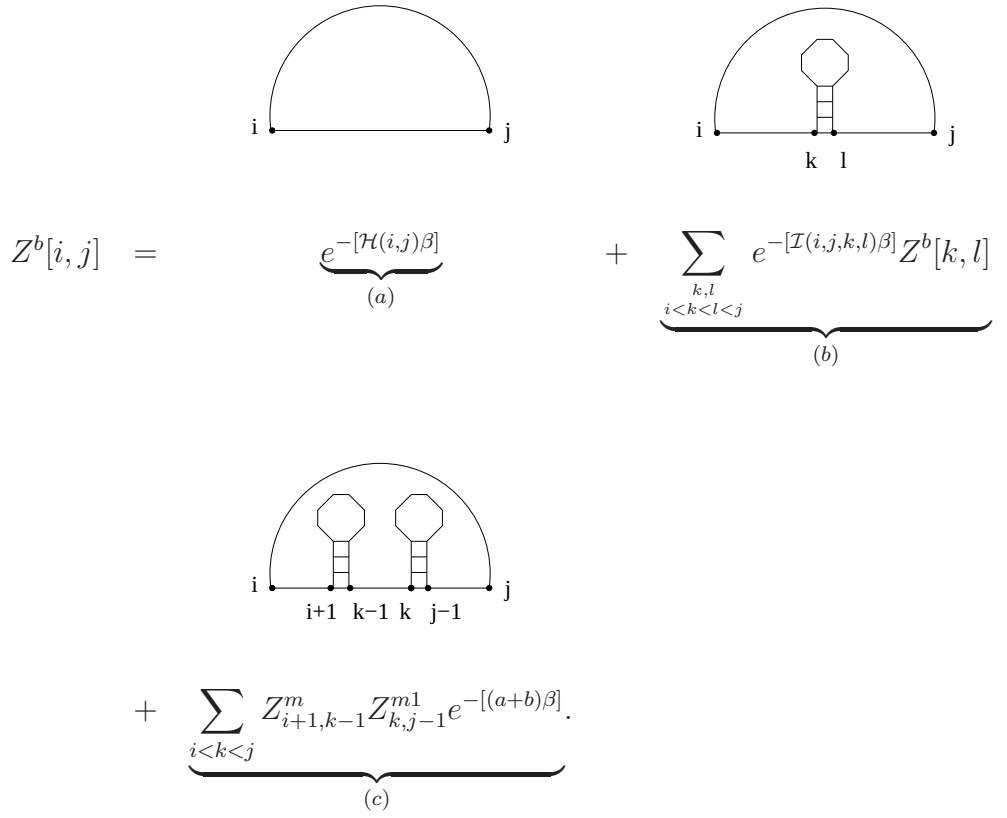


Figure 5: The partition of $Z^b[i,j]$, the partition function over all possible loops closed by base pair (i,j), into the contributions of the different loop types is illustrated in more detail: $\mathcal{H}(i,j)$, a and $\mathcal{I}(i,j,k,l)$, b, are functions that compute the loop energies of hairpin and interior loops given their enclosing base pairs. The contribution shown in c is for the calculation of multiloops. The computation of multiloop requires two additional types of restricted partition functions: $Z^m[i,j]$ is the partition function of all conformations on the interval $[i,j]$ that are part of a multiloop and contain at least one component, i.e., that contain at least one substructure that is enclosed by a base pair. Z^{m1} counts structures in multiloops that have *exactly* one component, see text for details.

The probability of a give structure S can be calculated once the partition function is know.

$$P(S) = \frac{1}{Z} e^{-[F(S)\beta]}. \tag{10}$$

In biological systems, however, not only the mfe-structure of a RNA but also sub-optimal structures play a functional role, e.g. [37]. In order to get an overview of biological relevant structural properties of a specific RNA we have to focus on the equilibrium probabilities of substructures that are common to a whole class of related structures. The formation a given base pair is such a substructure that displays the most significant features of the equilibrium ensemble.

The probability $P[kl]$ that base k is bound to base l in the equilibrium ensemble of structures is given by [50]:

$$P[k,l] = \sum_{S \in (k,l)} P(S). \tag{11}$$

For the computation of $P[k,l]$ the situation that base pair (k,l) is not enclosed by any other base pair has to be distinguished from the situation that (k,l) is enclosed by a base pair (i,j) with $i < k < l < j$, see figure 6.

If (k,l) is a non enclosed base pair then the partition function for segment $[1,k-1]$, is independent from the partition function for region $[l+1,N]$. If base pair (k,l) is enclosed by a base pair (i,j), the contributions of structures involving (k,l) in buldge, interior and multiple loops with closing pair (i,j), summed over all possible pairs (i,j), have to be considered. The contributions to $P[k,l]$ from structures with enclosed pairs (k,l) and exterior pairs (i,j) with $i < k < l < j$ can be computed recursively:

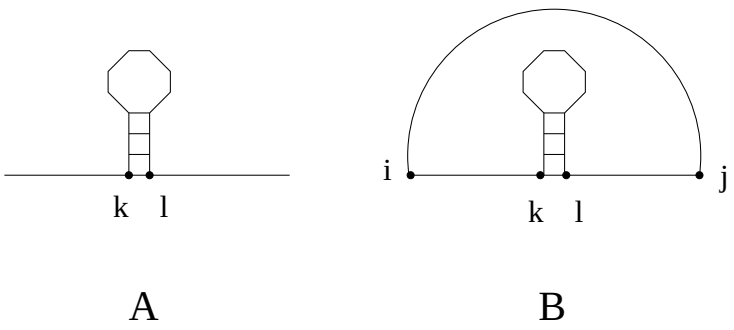


Figure 6: The figure shows the two situations that has to be distinguished for the calculation of $P[k,l]$. In A base pair (k,l) is exterior to all loops. In B base pair (k,l) is enclosed by a pairing (i,j) with $i<k<l<j$.

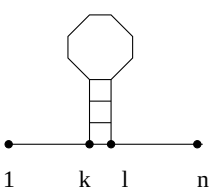
The calculation of $P[k,l]$ is in principle of order N^4, since summation over i, j, h and l are required. However, it is possible to reduce the calculations to order N^3 by using ancillary arrays and a careful distortion of the order in which the sums are performed. As for the calculation of the partition function, interior loops with $u > u_m$ are only dependent on u. For the calculation of multiloops the sum over i and j is reduced to a single sum by introducing ancillary arrays

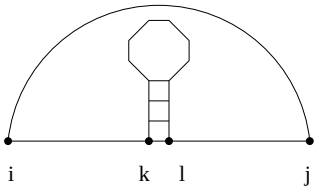
$$P^m[i,l] = \sum_{j>l} \frac{P[i,j]}{Z^b[i,j]} Z^m[l+1,j-1] \tag{12}$$

and

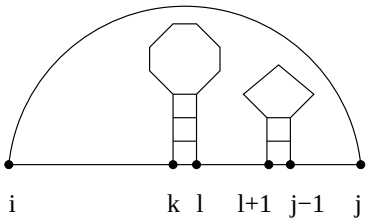
$$P^{m1}[i,l] = \frac{P[i,j]}{Z^b[i,j]} e^{((j-l-1)c)\beta} \tag{13}$$

These sum have to be calculated at the appropriate point in the recursion, when $P[i,j]$ has been defined. Since they are independent of k, they do not need recalculation if the values are stored. These modifications of the calcu-

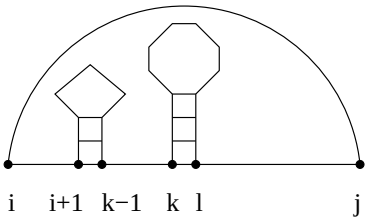


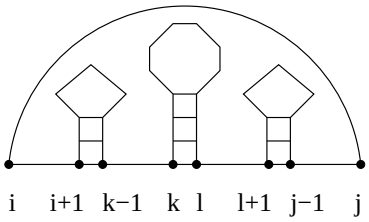


$$P[k,l] = \underbrace{\frac{Z[1,k-1]Z^b[k,l]Z[l+1,n]}{Z[1,n]}}_{(a)} + \underbrace{\sum_{\substack{u<u_m \\ i<k<l<j}} P[i,j]\frac{Z^b[k,l]}{Z^b[i,j]}e^{-[\mathcal{I}(i,j,k,l)\beta]}}_{(b)}$$



$$+ \quad \sum_{i<k} Z^b[k,l]e^{-[(a+b)\beta]} \quad * \quad \Big(\underbrace{e^{-[((k-i-1)c)\beta]}P^m[i,l]}_{(c)}$$





$$+ \quad Z^m[i+1,k-1]\big(\underbrace{P^{m1}[i,l]}_{(d)} \quad + \quad \underbrace{P^m[i,l]}_{(e)}\big)\Big).$$

Figure 7: In a base pair (k,l) is not enclosed. In the contribution shown in b the enclosed base pair (k,l) is part of an interior loop. Contributions c, d and e show the possible conformations in which (k,l) is enclosed by a multiloop with closing pair (i,j).

lation of $P[i,j]$ permit recursion 14, which is of order N^3 [50]. Recursion 14 is shown in more detail in figure 7.

$$\begin{aligned} P[k,l] &= \frac{Z[1,k-1]Z^b[k,l]Z[l+1,n]}{Z[1,n]} \\ &+ \sum_{\substack{u<u_m \\ i<k<l<j}} P[i,j]\frac{Z^b[k,l]}{Z^b[i,j]}e^{-[\mathcal{I}(i,j,k,l)\beta]} \\ &+ \sum_{i<k} Z^b[k,l]e^{-[(a+b)\beta]}\Big(e^{-[((k-i-1)c)\beta]}P^m[i,l] \\ &+ Z^m[i+1,k-1]\big(P^{m1}[i,l]+P^m[i,l]\big)\Big) \end{aligned} \tag{14}$$

2.2 RNA sequence-structure mapping and neutral networks

Essential aspects of evolutionary dynamics were studied by means of evolving populations of RNA molecules, e.g. [7, 71]. RNA combines both genotype and phenotype in a single molecule. This functional dichotomy makes RNA suitable as a model system for evolution. The genotype is the sequence, which is subjected to mutation. In what follows the secondary structure represents the phenotype, which is the target of selection [7]. Thus the folding of the RNA sequences into secondary structures establishes a set of rules describing how the genotype is converted into the phenotype, i.e. the sequence-structure map equals the genotype-phenotype map. In a second step the phenotype is evaluated to derive fitness values [70].
In the next section models for this mapping will be presented, which are suitable for analysis and exploration by computationally efficient algorithms and sufficiently realistic to incorporate important features of the evolutionary process [71].

2.2.1 Genotype-phenotype mappings

A RNA sequence can be described as a point in the space of all 4^n sequences with a fixed length n. The RNA sequence space, $\mathcal{I}$, has a natural metric [1] induced by one-error (point) mutants, known as the Hamming distance [30]. The Hamming distance between two end-to-end aligned sequences I_i and I_j, d_{ij}^h, is defined as the number of positions in which the sequences differ [71]. For the definition of selection criteria the distance between shapes, d_{ij}^S, can be described in terms of formal "edit" operations on secondary structure representations. This definition of shape-distance is not adequate for explaining patterns of phenotypic evolution, since there exist no physical operations that inter-converts structures heritably [18, 76]. The relation between sequences

[1] A metric on a set A is a map $D : A \times A \to \mathbb{R}$ with the following properties: for all $a, b, c \in A$ holds $D(a, b) = 0 \Longleftrightarrow a = b$, $D(a, b) = D(b, a)$ and $D(a, c) \leq D(a, b) + D(b, c)$.

(genotypes) and structures (phenotypes) may be formulated as a mapping from genotype space, $\mathcal{I}$, into phenotype space, $\mathcal{S}$, see figure 8 and equation (15).

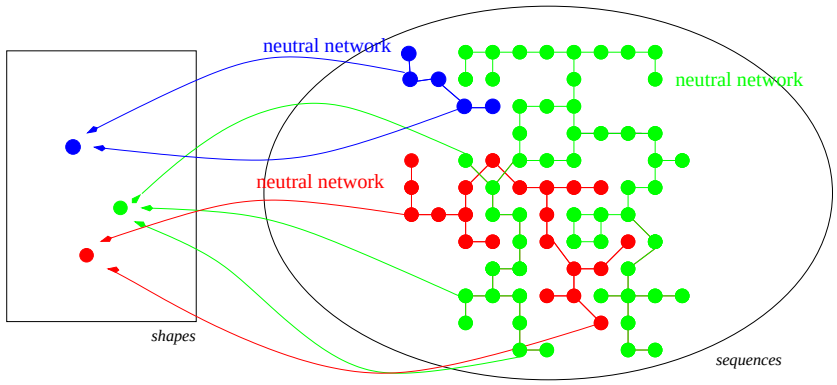


Figure 8: The figure shows a map from sequence (genotype) space to shape (phenotype) space. This mapping is many-to-one and hence non-invertible. Sequence and structure space are both high dimensional objects.

Equation (15) states that the sequence space $\mathcal{I}$, with the Hamming distance d_{ij}^h as distance measure is opposed by the shape space $\mathcal{S}$, with d_{ij}^S being the distance between phenotypes:

$$\Psi \quad : \quad \{\mathcal{I}; d_{ij}^h\} \quad \Rightarrow \quad \{\mathcal{S}; d_{ij}^S\} \quad \text{or} \quad S_k = \Psi(I). \tag{15}$$

The equation implies that a unique phenotype S_k is assigned to every genotype, while the inverse is not true. Generally a set of different sequences folds into the same secondary structure S_k. The many to one relationship between sequences and shapes is indicated by the fact that only the structure S_k has a fully specified subscript, whereas the set of sequences adopting S_k is denoted as 'I' [7, 71]. In evolution, the phenotype is evaluated according to its fitness value, where the fitness values are functions of evolutionary relevant properties of the phenotype [70]. The resulting mapping from shape space into the real numbers is called a landscape:

$$f \quad : \quad \{\mathcal{S}; d_{ij}^{S}\} \quad \Rightarrow \quad \mathbb{R} \quad \text{or} \quad f. = f(S_k). \tag{16}$$

As before many phenotypes may have the same fitness and again this mapping is non-invertible. Realistic fitness landscapes are complex objects with a very large number of local peaks and steep valleys [71].

A deeper understanding of the phenotype-genotype map is important for the examination of evolutionary mechanisms. In the equations introduced above fitness is a property of phenotypes, which are generated from genotypes using a set of predefined rules. A set of properly chosen rules makes it possible to extract relevant properties from real systems.
Currently a direct mathematical analysis of the prediction of phenotypes and their corresponding fitness values from the underlying genotypes is exceeding the computational scope. Therefore suitable mathematical models extracting robust statistical properties pertaining to the genotypephenotype mapping have been established [71, 72]. This model as described in the following sections, enables us to derive evolutionary relevant generalizations.

2.2.2 Generic properties of RNA folding

The simplest known example of evolutionary dynamics is RNA replication and mutation *in vitro* [75]. The genotype-phenotype map represents the relation between sequence and structure. RNA sequence-structure maps were analyzed by different techniques: construction of mathematical models based on random graph theory [63], exhaustive folding and enumeration of all sequences of a given chain length [27, 28], statistical evaluation by inverse folding and random walks in sequence space [19, 73] and by means of computer simulation of evolutionary dynamics [21, 22, 36, 82]. From these studies following generic results were derived:

More sequences than structures. An upper bound on the number of mfe structures of a fixed chain length n can be obtained recursively by counting

only those planar secondary structures that contain hairpin loops of size three or more and that include no isolated base pairs. The loop size constraint results from the fact that loops smaller than three nucleotides are unstable because of high steric strain energies. Single base pairs are often unstable since the dominating stabilizing contribution comes from base pair stacking. For large chain lengths ℓ the numbers of secondary structures, $N_S(\ell)$, are asymptotically approximated by the expression [7]:

$$N_S(\ell) \approx s(\ell) = 1.4848 l^{-3/2} (1.8488)^{\ell}. \tag{17}$$

The number of shapes computed from this expression is consistently smaller than the number of sequences, which is, for example, 4^{ℓ}, for the natural alphabet $\mathcal{A} = \{A, U, G, C\}$ [73]. Exhaustive folding and statistical evaluation of large sample of random RNA sequences of varying length, nucleotide alphabet and composition showed that the number of actually realized shapes is considerably smaller than the upper bound, $N_S(\ell)$, given in equation (17), [27]. The relation between RNA sequences and their structures is therefore highly degenerate.

Few common and many rare shapes. The frequencies of occurrence of individual structures in sequence space were obtained by folding large samples of random sequences of fixed chain length [27]. Analysis through exhaustive folding showed that the frequency of shapes is strongly biased. Ranking according to decreasing frequencies yields a distribution which obeys a generalized Zipf law, $f(r) = A(B+r)^{-\gamma}$, see figure 9. The rank of a shape is given by r (the most common structure has rank 1) and $f(r)$ denotes the fraction of sequences folding into the shape of rank r. The constants A, B and γ depend on sequence length and nucleotide alphabet. The Zipf-distribution assigns a high abundance to a tiny number of structures compared to those in the power tail [18]. This is best illustrated by an example: In the case of **GC**-sequences of length $\ell = 30$ more than 93% of all sequences fold into common shapes, which comprise only 10.4% of all shapes. A frequent shape may therefore be defined as one realized by more sequences than the average,

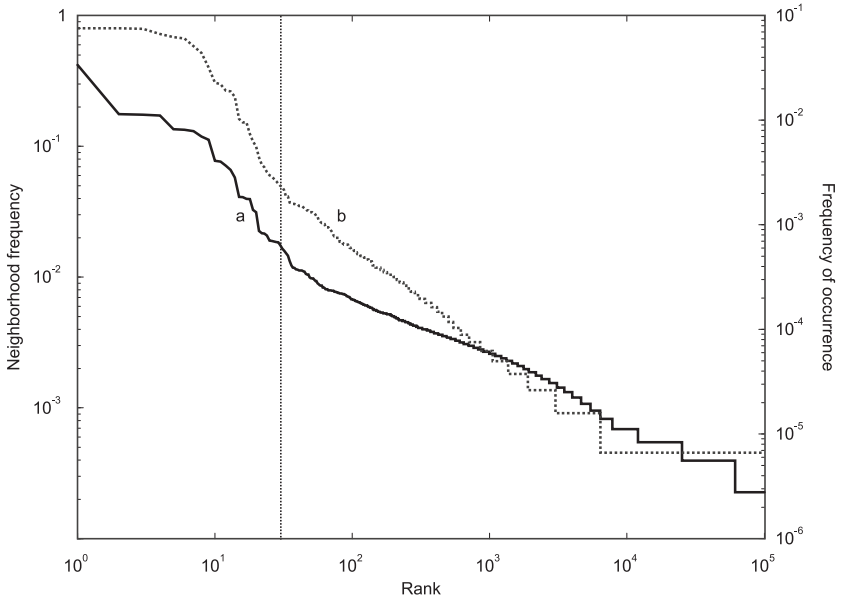


Figure 9: A log/log plot of the rank ordered structures in the boundary of tRNA^{phe}. Where the boundary B_α of a structure S_α consists of all sequences at hamming distance 1 from any sequence in the neutral set of structure S_α, see page 33 for a definition of neutral set. 28% of the neighbors of 2199 sequences folding into the clover-leaf structure formed the same shape than their reference sequence and thus belong to the neutral network. Curve **a** (right ordinate) shows the rank ordered frequency of occurrence $\vartheta(\beta, \alpha)$, which describes the total number of occurrences of structure S_β in the boundary of structure S_α, B_α. The neighborhood frequency $\nu(\beta, \alpha)$ is plotted in curve **b** (left ordinate). The neighborhood frequency reflects the likelihood of finding structure S_β in the one-mutation neighborhood of a randomly chosen sequence of S_α. The exact definition of $\vartheta(\beta, \alpha)$ and $\nu(\beta, \alpha)$ is given in section 2.2.3. The dotted vertical line separates the frequent structures in the boundary of tRNA^{phe} (right) from the hardly reachable shapes on the left. This is typical for a scaling according to Zipf's law, which implies that the log(frequency)/log(rank)-plot is a straight line [92].

$4^\ell/N_S(\ell)$. In the limit of long chains a vanishingly small fraction of shapes is frequent, but these are realized by almost all sequences [7]. Recapitulating, there are relatively few common structures and many rare ones.

Shape space covering. Sequences forming common structures are distributed (almost) randomly in sequence space. Schuster et al. [73] showed that it is sufficient to screen a (high-dimensional) sphere around an arbitrarily chosen reference sequence in order to find at least one sequence for every common structure. The radius of this shape space covering sphere, $r_{cov}(\ell)$ is much smaller than the radius of sequence space ($\ell/2$). For example, for a chain length $\ell = 100$ about 15 mutations are sufficient to find at least one sequence that folds into one of the common shapes [69]. The fact that all frequent structures are realized within a small neighborhood of any arbitrarily chosen sequence is referred to as "shape space covering".

Common structures form extended neutral networks. To discuss properties of a neighborhood in context of the RNA sequence - structure map, we have to start with some terminology. A *n-error neighbor* is a sequence that differers from a given reference sequence by n point mutations. A *neutral mutation* is a nucleotide substitution that preserves the mfe shape. The term *neutral neighbor* is used for an one-error neighbor that preserves the mfe-shape of its reference sequence. The *neutrality* of a sequence is defined as its fraction of neutral one-error neighbors [18].
Sequences folding into common shapes typically have a significant fraction of neutral one- or two-error neighbors. Such sequences form an extensive, mutationally connected network, that was termed *neutral network* [73]. The distribution of sequences belonging to a neutral network in sequence space was analyzed using random graph theory [62]. The set of all sequences, I_j, folding into a given structure, S_k, is termed the neutral set

$$\mathcal{G}_k = \Psi^{-1}(S_k) = \{I_j | \Psi(I_j) = S_k\}.$$

The neutral set $\mathcal{G}_k$ of a structure S_k can be converted in a graph by assigning sequences I_j to the vertices and drawing edges between all pairs of sequences with Hamming distance $d_{ij}^h = 1$. Modeling neutral networks described by $\mathcal{G}_k$ using random graph theory identified the average degree of neutrality of a given network $\overline{\lambda(\mathcal{G}_k)} = \overline{\lambda_j}$ as the parameter determining global network properties [63], where

$$\overline{\lambda_j} = \frac{\sum_{I_j \in G_k} \lambda_k(I_j)}{|\mathcal{G}_k|}.$$

$\lambda_k(I_j)$ describes the local connectivities of a network, in case of $\mathcal{G}_k$ it is the number of neutral nearest neighbors at individual nodes divided by the total number of nearest neighbors, $|\mathcal{G}_k|$ is the number of vertices in the neutral network.

"Neutral networks show a king of percolation phenomenon. They are connected and span the entire sequence space if $\overline{\lambda_j}$ exceeds a critical threshold value λ_{cr}. Below the threshold the networks are partitioned into many components with one dominating 'giant component' and many small islands" [7]:

$$\mathcal{G}_k = \begin{cases} \text{connected}: & \overline{\lambda_j} > \overline{\lambda_{cr}} = 1 - \kappa^{-\frac{1}{\kappa-1}}, \\ \text{partitioned}: & \overline{\lambda_j} < \overline{\lambda_{cr}} = 1 - \kappa^{-\frac{1}{\kappa-1}}, \end{cases}$$

where κ is the size of the alphabet ($\mathcal{A} = A, U, G, C : \kappa = 4$; $\mathcal{A} = G, C : \kappa = 2$;). Connected areas on neutral networks define regions in sequence space that are accessible to populations through genetic drift [36].

Neutral networks offer an ideal combination of search capacity and robustness to mutations: "the genotypes may diffuse over the network by single nucleotide exchanges without loosing the currently optimal structure, until a non neutral mutant is encountered with increased fitness. The population will then switch to the network of this structure" [22, 35, 36]. Neutral networks are of maximal use, if they comprise connected graphs and a maximal

number of new phenotypes is available in the 1-error neighborhood of the network or even a single sequence.
In RNA, a small number of mutations can result in a new secondary structure by destabilizing the parental fold and supporting one of the numerous alternative ones. As a result of this any single RNA sequence has a "neighborhood" of secondary structures that become favorable upon a few mutations [73]. On the other hand many RNA sequences can fold into the same secondary structure, or as Motoo Kimura stated, the vast majority of genetic change at the level of a population must be neutral, rather than adaptive [39]. The experimental data from Schultes and Bartel [68] provide a direct proof for the existence of extended neutral networks. Starting from two phylogenetically unrelated ribozymes with different catalytic activities, whose RNA-conformations had no base pair in common, Schultes and Bartel constructed a RNA sequence, that is compatible with both secondary structures and showed both catalytic functionalities, although with lower efficiency than the parent ribozymes. The chimeric molecule was afterwards optimized for both catalytic activities by mutation and selection: "Minor variants of this sequence are highly active for one or the other reaction, and can be accessed from prototype ribozymes through a series of neutral mutations" [68]. The authors were not only able to track neutral paths of constant structure and full ribozyme function from the mutants to the parents. Additionally this work shows, that two neutral networks approach each other very closely in the surrounding of the chimeric molecule. Thus this work can be viewed as an experimental example of the intersection theorem, that will be discussed in the next section.

2.2.3 Shadows and Intersections

For the evolutionary process it is important to understand which phenotypes are accessible from which genotypes. This notion of accessibility can be used to define a relation of nearness among phenotypes. The genotype-phenotype map, discussed in the previous sections, is ideally suited to implement these

concepts: The folding of RNA sequences induces a "statistical topology " on the set of mfe structures. The statistical topology organizing the set of RNA shapes explains why neutral networks in sequence space and neutral drift on these networks play a key role in evolutionary optimization" [22].

For situations conserving chain length where point mutations are the exclusive source of variation the Hamming distance, see page 28, is a natural metric for sequences. The definition of a natural metric for structures, however, is difficult. Evolutionary modification of a structure requires modifying its underlying sequence, therefore any definition of a metric based on a syntactic notion of (dis)similarity between structures is bound to be artificial. Fontana et al. [22] devised an accessibility relation between structures, that does not quantify a distance but expresses a weaker notion of neighborhood: They pointed out that "a structure S_β which is highly dissimilar from a structure S_α on syntactic ground might nonetheless be "near" to S_α on the count of being accessible from S_α by a small mutation in S_α's sequence. Alternatively, among two syntactically highly similar structures, one might nonetheless fail to be evolutionary "accessible" from the other."
Fontana et al. [22] defined the "boundary" B_α to consist of all sequences at hamming distance 1 from any sequence in $\mathcal{G}_\alpha$, where $\mathcal{G}_\alpha$ is the neutral set of structure S_α, see page 33 for an exact definition of $\mathcal{G}_\alpha$. Similarly, they called the set of sequences at distance d from $\mathcal{G}_\alpha$ its "d-boundary" and let boundary stand as a shorthand for 1-boundary. To obtain $\sum_\alpha \subset \sum$, the set of all 1-accessible structures of S_α, all sequences in B_α are folded into their mfe -structure. $\sum$ is the set of all mfe structures of fixed length over a given alphabet.

Even for moderate chain length, it is not possible to completely identify the set of structure neutral neighbors, $\mathcal{G}_\alpha$, for a given structure S_α [22]. To determine which structures are accessible from the neutral set $\mathcal{G}_\alpha$, one has to resort to sampling. This is done by fixing a secondary structure S_α of length

n and generate a sample of x sequences, that have S_α as their mfe structure by inverse folding [33]. For each sequence in this sample all the $3n$ neighbors of this sequence are folded, in this way the $3nx$ sequences in the boundary of S_α are generated. The structures of these sequences constitute a sample of $\sum_\alpha$ [22], we will call this sample a *local shadow* of structure S_α.

A structure S_β occurring in the shadow of a given structure S_α is accessible from S_α, that is $S_\beta \leftarrow S_\alpha$, if there exists a pair of sequences $I_a, I_b \in \mathcal{I}$ with $d^h_{I_a,I_b} = 1$ and $\Psi(I_a) = S_\alpha$ and $\Psi(I_b) = S_\beta$. In this notation the set of structures accessible from S_α is written as $\sum_\alpha = \{S_\beta | S_\beta \leftarrow S_\alpha\}$ [22].
We are not just interested in the structures, that occur in the 1-error neighborhood of sequences adopting structure S_α, but also in how often they occur. "Each structure $S_\beta \leftarrow S_\alpha$ has two multiplicities associated with it. One multiplicity, $N(S_\beta, S_\alpha)$, counts the total number of sequence-neighborhoods of S_α in which structure S_β occur-es at least once. We normalize it by the size N_α of $\mathcal{G}_\alpha$, and call it the *neighborhood frequency*:

$$\nu(S_\beta, S_\alpha) = \frac{N(S_\beta, S_\alpha)}{N_\alpha}. \tag{18}$$

The neighborhood frequency, $\nu(S_\beta, S_\alpha)$, reflects the likelihood of finding structure S_β in the one-mutation neighborhood of a randomly chosen sequence of S_α.
The other multiplicity refers to the total number of occurrences, $N_t(S_\beta, S_\alpha)$, of structure S_β in B_α. Each neighborhood of a sequence S_α is, therefore, weighted with the actual size of S_β in that neighborhood. We normalize it by $3nN_\alpha$, and call it the *occurrence frequency*:

$$\vartheta(S_\beta, S_\alpha) = \frac{N_t(S_\beta, S_\alpha)}{3nN_\alpha}. \tag{19}$$

$\nu(S_\beta, S_\alpha)$ and $\vartheta(S_\beta, S_\alpha)$ are estimated by sampling as mentioned above" [22].

Recapitulating the information provided above, $\nu(S_\beta, S_\alpha)$ and $\vartheta(S_\beta, S_\alpha)$ are measures of the probability that one step away from a random point in the

neutral network of structure S_α results in a sequence folding into structure S_β. The accessibility relation defined by $\nu(S_\beta, S_\alpha)$ and $\vartheta(S_\beta, S_\alpha)$ is not symmetric and therefore not a distance [18,22]. An example for this asymmetry is the loss and formation of a stack. A stack in shape S_α will be only marginally stable in most sequences realizing S_α. Therefore, the loss of the stack in the one-error neighborhood of sequences folding into S_α is very likely. In contrast, the reconstitution of this stack by a single point mutation in a structure S_β, that differs from S_α only by the absence of this stack, requires a special sequence context. Therefore, shape S_β may be significantly easier to access from S_α than the other way around [18].
This accessibility distribution is converted into a binary attribute of nearness by defining the *neighborhood* of structure S_α as the set containing S_α and all shapes accessible from S_α above a certain likelihood [21, 22]. "Because accessibility is asymmetric, shape S_β may be near (read: in the neighborhood) of S_α, but S_α may not be near S_β. This construction of shape-neighborhood is technically consistent with the formalization of the neighborhood concept in topology" [18, 76].

One consequence of the shape space topology described above is the fact that pairs of neutral networks approach each other closely at intersection points of the compatible sets in which they are embedded. A sequence is termed compatible with a secondary structure if all positions in paired regions are occupied by bases that are capable of forming a base pair [78]. The number of compatible sequences for a given structure S_k with n_u unpaired bases and n_p paired bases is $4^{n_u} * 6^{n_p}$. The number of compatible sequences is certainly larger than the number of sequences that actually form the given structure as their mfe structure [78]. The compatible set of a secondary structure S_α, $\mathcal{C}_\alpha$, is the set of all sequences that are compatible with structure S_α.
The *intersection theorem* guarantees that for any two prescribed secondary structures there is always a non-empty set of compatible sequences [17]. Extensive computational studies showed that the neutral set $\mathcal{G}_\alpha$ of a structure S_α

is approximately embedded in $\mathcal{C}_\alpha$ [27, 73]. In the case of common secondary structures S_γ, the neutral sets, $\mathcal{G}_\gamma$ are densely embedded in $\mathcal{C}_\gamma$. Therefore, the neutral networks of two structures S_k and S_l come very close together on the set $\mathcal{C}_k \bigcap \mathcal{C}_l$ of sequences that are compatible with both structures [63,64].

2.2.4 Computer simulations of RNA evolution

The role of neutral networks in evolutionary dynamics has been investigated by computer simulation of RNA populations in a flow reactor [22]. The flow reactor used to study the evolution of a population of replicating and mutating RNA species is implemented as a continuously stirred tank reactor [20]. In this setup, a continuous influx of material necessary for replication is compensated by a homogeneous outflux of reaction mixture. The flow rate is adjusted to obtain a population size fluctuating around an expectation value of N with a standard deviation of $\sqrt{N}$ [7]. Additionally to flow terms the chemical reaction mechanism includes replication and mutation steps. Manfred Eigen introduced the quasispecies theory that permits the application of of chemical reaction kinetics to molecular evolution [12]. According to Eigen's model, which was further developed in subsequent studies [13, 14], populations of nucleic acid sequences drift through sequence space gaining information by variation and selection. The distribution of genotypes in these populations is metastable but structured around a master sequence. Populations explore new environments through a mutation selection process, the information gained is laid down in their genotypes [72]. In this work the exploration of genotype space is described as a stochastic process with a predefined target, which forms an absorbing barrier and thus sets an end point.

The replication-mutation system introduced by Eigen can be described by ordinary differential equations. Here x_i denotes the relative frequencies of the n different genotypes I_i in the population. Defining the number of sequences adopting genotype I_i as N_i the population size is given by $N = \sum_{j=1}^{n} N_i$. The time dependence of the sequence distribution is then described by the

kinetic equations

$$\frac{dx_k}{dt} = x_k\Big(Q_{kk}a_k - \Phi(t)\Big) + \sum_{j=1,j\neq k}^{n} a_j Q_{kj} x_j, \quad k = 1,\ldots,n. \tag{20}$$

where a_k is the replication rate constant, $Q = \{Q_{ij}; i,j = 1,\ldots,n\}$ is the mutation matrix and the excess production rate $\Phi(t) = \overline{a}(t) = \sum_{j=1}^{n} a_j, x_j(t)$ is the average fitness of the reactor population [7]. The production of identical and mutant offsprings from the actual population is equal to the excess production leading to a constant population size $\sum_{j=1}^{n} dx_j/dt = 0$. In the mutation matrix Q diagonal entries stand for error free replication, while off diagonal entries denote mutation. In this work a uniform error rate is assumed. This implies that the probability of a mutation is independent of the nature and position of the base exchange and different base exchanges are independent from each other. Only point mutations are allowed, excluding indels, translocations or cross-overs [87]. Nevertheless the creation of any RNA molecule that is part of the sequence space is possible. Q_{kj}, the frequency at which a genotype I_k is synthesized as an error copy of I_j, is therefore defined by the replication accuracy per base q and the derived mutation rate $p = 1 - q$

$$Q_{kj} = q^{\ell}\Big(\frac{1-q}{q}\Big)^{d^h_{kj}} = (1-p)^{\ell}\Big(\frac{p}{1-p}\Big)^{d^h_{kj}} \tag{21}$$

where ℓ denotes the length of the sequence and d^h_{kj} is the Hamming distance between a template sequence and its offspring.

Since biochemical systems have an important stochastic component, the quantitative study of such systems requires stochastic simulation, which accounts for random biochemical events and their consequences. Therefore the reaction mechanism is implemented using the stochastic method for modeling chemical kinetics conceived by Gillespie [25, 26]. Gillespie's method is used to numerically simulate the time evolution of the system as a kind of random walk that is governed by the Master Equation, in this work equation (20).

Further details of the implementation of the flow reactor used in this work can be found in the work of Schuster & Fontana, see e.g. [7,20,72], a complete description of the implementation of the reactor used in this work is given in the thesis of Andreas Wernitznig [87].

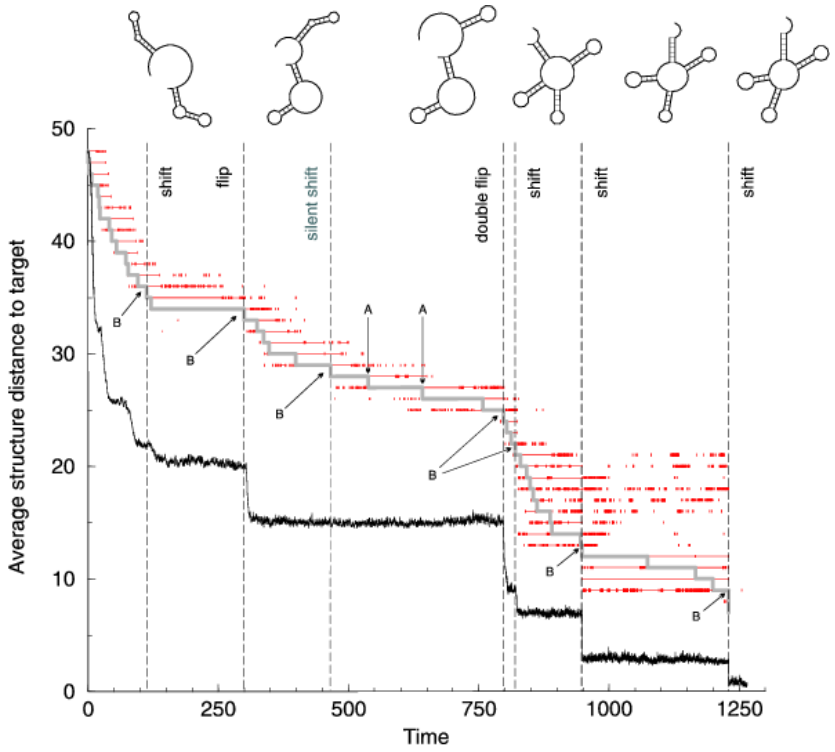


Figure 10: Results of an RNA structure optimization experiment with a predefined target structure in a flow reactor. A population size of $N = 1000$ and a mutation rate $p = 0.001$ were used. Fitness is computed as a function of the distance between current structure S_j and target structureS_τ, $d^s_{j\tau}$. The black curve shows the mean distance to the target structure of the entire population, $\overline{d^s_{j\tau}}$, plotted against time. The time scale represents the "real time" of the simulation experiment in arbitrary units. The grey step function indicates the relay series, where equal height indicates the same shape. The six most important shapes of the relay series are shown at the top of the figure. Continuous transitions between phenotypes are marked by **A**, discontinuous ones by **B**. Lower case letters denote individual transitions. Figure taken from [7].

"Previous computer simulations confirmed three basic features of molecular

evolution:(i) Population sizes of a few thousand molecules are sufficient for RNA optimization, (ii) stochastic effects dominate in the sense that the sequence of events recorded in one particular trajectory were never observed again in subsequent identical simulations, and (iii) sharp error thresholds as predicted by the quasispecies concept were observed in computer runs with different mutation rates" [7]. The trajectory of a typical simulation is shown in figure 10. The evolutionary trajectory describes the time course of the average distance of the population from a predefined target secondary structure. The course of the evolutionary optimization process is reconstructed through determination of a series of phenotypes leading from the target structure to an initial shape, called the relay series. Transitions between the shapes in a relay series can be continuous or discontinuous. Continuous transitions are the loss and formation of a base pair adjacent to a stack and the opening of a constrained stack, while the closing of a constrained stack and structure transformations that require the coordinated change of several part of the molecule at once are discontinuous transitions. Discontinuous transitions reflect those transformations that are difficult to achieve by the mechanisms underlying the genotype-phenotype map. These transitions permit key innovations that enable a population to reach the target in a particular experiment [21]. In this context discontinuous transitions can be viewed as gateways in evolutionary optimization. Discontinuous transitions are intimately connected with punctuation: "A population of replicating and mutating sequences drifts on the neutral network of the current best shape until it encounters a gateway to a networks that conveys some advantage or is fitness neutral. That encounter, however, is evidently not under the control of selection" [18]. An example of punctuation in evolving RNA population is demonstrated in figure 10: The average fitness shows periods of stasis punctuated by sudden improvements.

2.3 Classification on sequence-structure mapping

A variety of systems ranging from physics to biology are described by complex networks. Examples include metabolic networks, networks of chemicals linked by chemical reactions or the Internet, a network of routers and computers linked by physical links. Traditionally these systems have been modeled as random graphs, however in recent years it has been increasingly recognized that the topology and evolution of real networks are governed by robust organization principles [1]. As neutral network in RNA sequence space corresponds to an undirected graph we will apply the theory described in this section to neutral networks in sequence space.

A graph is a pair of sets $G = \{P, E\}$ where P is a set of n nodes and E is a set of k edges that connect two elements of P. The degree γ of a vertex is the number of edges in a graph, which touch the vertex, also called the local degree. A graph is said to be regular of degree γ if all local degrees are the same number γ. A random graph, on the other hand, is a collection of vertices and edges, connecting pairs of them at random. In some cases random graphs with appropriate distribution of vertex degrees can predict real world problems with surprising accuracy [58]. Ordinarily the connection topology of a network is assumed to be either completely regular or completely random. However many real networks lie somewhere in between the extremes of order and randomness. The description of the transition from locally ordered systems to a random network revealed the existence of so called "small-world" networks, that are highly clustered, like regular graphs, yet have small characteristic path length, like Erdös-Rényi random graphs [85].
The algorithm used to interpolate between regular and random networks is based on a two level process. It starts with a regular ring lattice with n vertices and k edges per vertex. Subsequently, each edge is rewired at random with probability p without altering the number of nodes or edges, self-edges and duplicate edges are forbidden. An example of this process is

given in figure 11. Watts and Strogatz [85] described graphs, that, at a small rewiring probability p, are sparse but not so sparse that the graph would become disconnected. These graphs can be highly clustered, like regular networks, but show small characteristic path lengths, like random graphs. Specifically, these graphs, which where termed *small world* networks, have to meet the following condition:

$$n \gg k \gg \ln(n) \gg 1, \tag{22}$$

where $k \gg \ln(n)$ guarantees that a random graph will be connected [5].

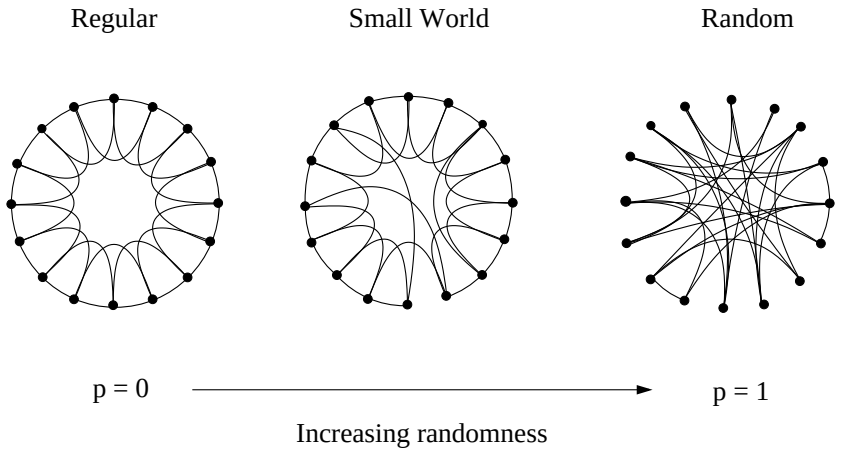


Figure 11: Random rewiring procedure for the transition between a regular ring lattice and a random network, figure taken from to Watts and Strogatz [85]. Details see text.

The structural properties of small-world graphs can be quantified by their characteristic path length $L(p)$ and clustering coefficient $C(p)$. L is defined as the number of edges in the shortest path between two vertices, averaged over all pairs of vertices. The clustering coefficient of a vertex, which measures the 'cliquishness' of the neighborhood of this vertex, determines what fraction of the vertices adjacent to this vertex are also adjacent to each other. The clustering coefficient C_i of a node i with k_i edges is given by

$$C_i = \frac{2E_i}{k_i(k_i-1)}, \tag{23}$$

where E_i is the number of edges that actually exist between the k_i neighbors of node i and $k_i(k_i-1)$ is the maximal number of edges that can exist between the k_i neighbor of node i. $C(p)$ is defined as the average of C_i over all i.

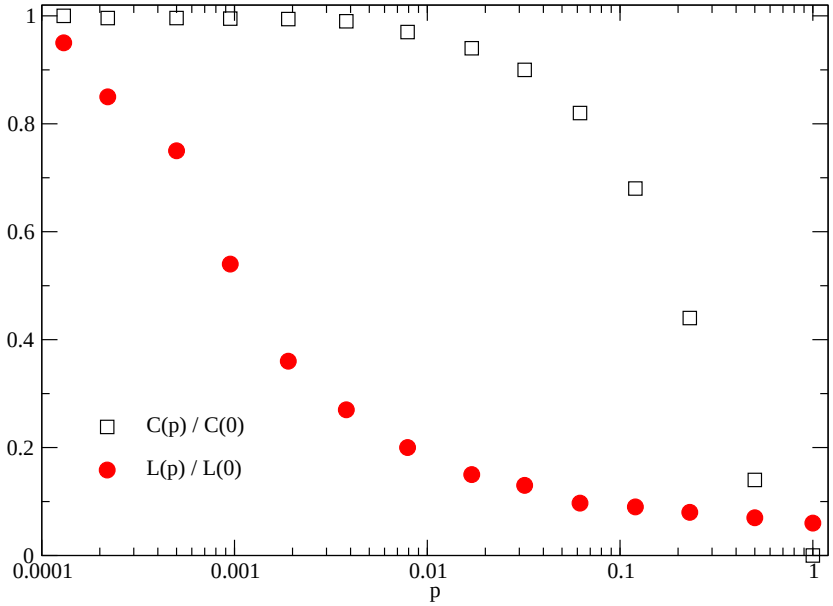


Figure 12: Characteristic path length $L(p)$ and clustering coefficient $C(p)$ over 20 random realizations of the rewiring process described in figure 11. Figure taken from to Watts and Strogatz [85].

Figure 12 shows the change of $L(p)$ and $C(p)$ for the family of randomly rewired graphs described in figure 11. Obviously, a rewiring probability of $p = 0$ leaves the graph regular, while $p = 1$ results in a random network topology. Watts and Strogatz [85] found that in the regime

$$L(p) \sim \frac{n}{2k} \gg 1, C(p) \sim \frac{3}{4} \text{ if } p \to 0, \tag{24}$$

while

$$L(p) \approx L_{\text{random}} \sim \frac{\ln(n)}{\ln(k)}, C(p) \approx C_{\text{random}} \sim \frac{k}{n} \ll 1 \text{ if } p \to 1. \tag{25}$$

Thus the regular lattice at a rewiring probability of $p = 0$ is highly clustered resulting in a linear growth of $L(p)$ with the number of vertices n. In contrast the random network at $p = 1$ is a poorly cluster graph where $L(p)$ grows logarithmically with n.

Figure 12 shows that there is an interval $0.001 < p < 0.01$ in the model, that exhibits small-world properties: The network obtained is sparse showing a characteristic path length $L(p)$ in the range of of random graphs L_{random}. However, the clustering coefficient $C(p) \gg C_{\text{random}}$, which results in a network that is more clustered than a random graph of equal size. Watts and Strogatz [85] stated that small world networks result from the immediate drop in $L(p)$ which is caused by the appearance of a few long range edges. The introduction of such 'short cuts' connects vertices that would otherwise be much farther apart than L_{random}. Therefore each shortcut has a highly nonlinear effect on L, it reduces not only the distance between the two vertices it connects, but also between their immediate neighborhoods. On the other hand the removal of an edge from a clustered neighborhood to make a shortcut, has, at most, a linear effect on C [85]. As a result of their unique topology models of dynamical systems with small-world coupling display enhanced signal-propagation speed, computational power and synchronizability.

3 Results on Sequence-Structure Mapping

3.1 Folding a single RNA molecule

RNA combines both the genotype (sequence) and the phenotype (structure) in a single molecule. Previous studies of the genotype-phenotype map, which is modeled by folding RNA sequences to secondary structures, yields four general results, see section 2.2: As there are far more sequences than secondary structures, many RNA sequences adopt the same secondary structure, which is termed neutrality. The neutral network of a structure S_j is the set of all sequences of a given length, which have S_j as their minimum free energy (mfe) structure. The neutral network of a structure S_j is connected and spans the entire sequence space if the average number of neutral neighbors at individual nodes of the network is greater than $1 - \kappa^{(1/(\kappa-1))}$, where κ is the size of the alphabet. A connected neutral network corresponds to an undirected connected graph. Common structures form extended neutral networks.
We study the topological structure of neutral networks using exhaustive folding and enumeration. Furthermore we examine the emergence of small-world properties in connected graphs corresponding to different neutral networks.

3.1.1 Exhaustive folding and enumeration of small RNA sequences

For a detailed study of the topological structure of phenotype space it is mandatory to know the sequence of all RNA molecules of a given length that can adopt a stable mfe structure. A mfe structure is defined as stable if it has a negative free energy. RNA sequences are folded into secondary structures using the Vienna RNA package. The parameter set for the calculation of the mfe structure does not include special stabilizing energies for certain tetra loops. Bases adjacent to helices in free ends and multiloops can participate in all possible dangling ends. Compared to the parameter set used in previous studies [41], these parameters reduce the number of sequences that can fold into a stable secondary structure. Furthermore structures with lonely base

pairs are penalized. For very short sequence lengths the exclusion of special stabilizing energies for certain tetra-loops is necessary to avoid structural artifacts - for example structures containing only a single isolated base pair.

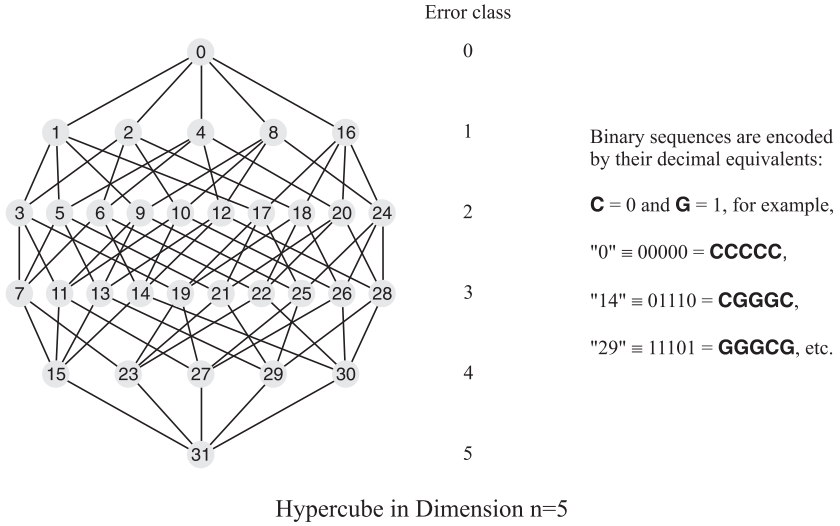


Figure 13: In high dimensional spaces distances are short. For example in a hybercube of dimension five, the shortest path percolating the space has a length of 5. Distance is measured in error classes, where one error corresponds to a hamming distance of one.

We investigate the sequence spaces $\mathcal{I}_{AUGC}^{(\ell=9)}$ and $\mathcal{I}_{AUGC}^{(\ell=10)}$. For our studies, all sequences of chain length $\ell = 9$ and chain length $\ell = 10$ and the natural alphabet $\mathcal{A} = \{A, U, G, C\}$ are exhaustively folded and enumerated. In such small sequence spaces, the majority of sequences forms no stable secondary structure, i.e. a secondary structure with a free energy below zero. In $\mathcal{I}_{AUGC}^{(\ell=9)}$ and $\mathcal{I}_{AUGC}^{(\ell=10)}$ only 1.3% and 3.8%, respectively, of all sequences form a mfe structure with a free energy below zero. Despite the low percentage of sequences forming stable structures, neutral networks may span the whole sequence space in dimension 9 and 10. This is demonstarted in figure 13, using the binary $\{G, C\}$ alphabet as an example.

The number of possible sequences of a chain of length ℓ and the natural alphabet that comprises $\kappa = 4$ letters is given by κ^ℓ. Although we only examine the properties of sequences that fold into stable secondary structures, the number of sequences under consideration is too large for an efficient usage of plain text file formate. We therefore need a system for storing, retrieving and managing large amounts of data. PostgreSQL [54], an Open Source database, was used to store all sequences of a given length that fold into a stable mfe structure.

The database setup was designed to allow fast and efficient analysis of the stored data: Each connected component of the neutral network(s) of a structure is stored in a separate table. A connected component is constructed by drawing edges between all neighbors in sequence space, which have either a Hamming distance of one ($d^h = 1$) or one compensatory mutation ($d^c = 1$). A compensatory mutation or base pair substitution is a mutation preserving a base pair. For easy retrieval of sequences adopting structure S_j, a list containing the names of the component tables and the mfe structure that they represent is maintained.

The setup of the database proceeds in two steps: In the first step all RNA sequences of a given length are generated and the mfe structure of each sequence is determined. The program RNAfold from the Vienna RNA package version 1.4 [33] was employ to determine the mfe of the secondary structure for a given sequence. The Vienna RNA package version 1.4 uses the energy parameters provided by Mathews et.al. [46]. For folding we applied the command-line-options $-d2$ and -4, $-d2$ causes bases adjacent to helices in multiloops and free ends to give a stabilizing energy contribution, regardless whether they are paired or unpaired, -4 prevents the usage of special stabilizing energies for certain tetra loops. Sequences that fold into the same mfe structure are collected in a table. To reduce the size of memory required to store the sequences each sequence is encoded as a single 32 bit unsigned integer.

In the second step we test if the sequences contained in a single table form

a connected component, or if they belong to scattered networks with many components. Each connected component of structure S_j is written into a separate table. The list of all component tables is automatically updated if tables are created or deleted. This database setup keeps table size as small as possible, while retaining the topological structure of the data set. Small tables improve the performance of the database.

For the forthcoming discussion we shall need the following definitions:
We consider the folding of an RNA sequence I_i into a secondary structure S_k.

$$S_k = \psi(I_i)$$

An RNA sequence $I_i = \{x_1 x_2 \dots x_\ell\}$ over an alphabet $\mathcal{A}$ with κ letters is *compatible* with a secondary structure S_k if base-pair $(i,j) \in S_k$ implies that letters $x_i x_j$ can form an allowed base-pair. This situation is expressed by $x_i x_j \in \mathcal{B}$, where the set of allowed base pairs $\mathcal{B} = \{AU, UA, GC, CG, GU, UC\}$. A specific sequence I_i may be compatible with more than one structure. The structure S_l with the minimal free energy, $F_{\min}(S_l, I_i)$, on sequence I_i is termed the mfe structure. Other structures S adopted by sequence I_i with a free energy $F(S, I_i) \geq F_{\min}(S_l, I_i)$ are termed suboptimal structures.

compatible set $\mathcal{C}_k$ The set all of sequences that are compatible with a structure S_k is then given by

$$C_k = \{I_i | (i,j) \in S_k \Longrightarrow x_i x_j \in \mathcal{B}\}$$

Note that a sequence I_i can be an element of $\mathcal{C}_k$ if S_k is a suboptimal structure of I_i.

intersection $\mathcal{X}$ The set of all sequences that are compatible with more than one secondary structure and can adopt these structures with a negative free energy, i.e. as stable structures, *intersection sequences*.

$$\mathcal{X} = \{I_i \in \mathcal{C}_k \bigcap \mathcal{C}_l | F(S_k, I_i) < 0 \wedge F(S_l, I_i) < 0\}$$

neutral set $\mathcal{G}_k$ The neutral set of a structure S_k, $\mathcal{G}_k$, is the set of sequences I_i that fold into structure $S_k = \psi(I_i)$ as their mfe structure, see section 2.2.2.

$$\mathcal{G}_k = \Psi^{-1}(S_k) = \{I_i | \Psi(I_i) = S_k\}.$$

Note that a sequence I_i can only be an element of the neutral set $\mathcal{G}_k$ if S_k is the mfe structure of I_i.

neutral network The neutral set $\mathcal{G}_k$ of a structure S_k can be converted in a graph by assigning sequences I_i to the vertices and drawing edges between all pairs of sequences that can be converted into each other using specified variation operators.

3.1.2 Generic features of the sequence-structure mapping for short sequences

In the sequence spaces $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ and $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ the open chain is the most frequent conformation, formed by 98.7% and 96.2% of all sequences, respectively. Out of the 4^9 possible sequences of the set of all sequences of length $\ell = 9$ and the alphabet $\mathcal{A} = \{A, U, G, C\}$ only 3280 have a stable secondary structure. These 3280 sequences fold into 4 different structures as shown in figure 14.

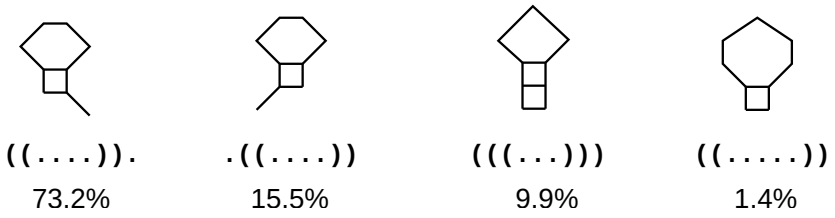


Figure 14: Structures realized in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ that adopt a stable mfe structure. Below the graph representation of each structure the dot-bracket notation is depicted. The frequencies written below the dot-bracket notation are the percentage of sequences folding into the indicated structure of all sequences with a stable structures in sequence space $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$.

As expected from the predictions of the RNA phenotype-genotype map the frequency of shapes is biased: 73.8% of all sequences in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ adopting a stable structure fold into secondary structure ((....))., these sequences are contained within one giant component. For the structure component decomposition of sequences folding in a specified structure S_k single base exchanges ($d^h = 1$) and base pair substitutions ($d^c = 1$) are used as "variation operators". To obtain the connected components of structure S_k we define the neutral set $\mathcal{G}_k$ of structure S_k as a graph: The graph $\mathcal{G}_k$ is a pair of sets $\mathcal{G}_k = \{V_k, E_k\}$ where $V_k = \{I|\psi(I) = S_k\}$ is the set of vertices and $E_k = \{I_i, I_j|I_i, I_j \in V_k \wedge (d^h_{i,j} = 1 \vee d^c_{i,j} = 1)\}$ the set of edges. We call the connected components of $\mathcal{G}_k$, $\mathcal{G}^l_k$, where $\mathcal{G}_k = \bigcup_l \mathcal{G}^l_k$ and $\mathcal{G}^l_k$ is connected.
Table 1 shows the results of the decomposition of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ in $\mathcal{G}^l_k$ components.

Table 1: Sequences of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ folding into the same secondary structure S_k are decomposed into mutationally connected structure components $\mathcal{G}^l_k$ using single base exchanges $d^h = 1$ and base pair substitutions $d^c = 1$ as variation operators.

	number of		
structure	sequences	$\mathcal{G}^l_k$ components	seq. per comp.
((....)).	2400	1	2400
.((....))	508	5	240, 220, 16, 16, 16
(((...)))	324	1	324
((.....))	48	1	48

One of the generic results of the analysis of the sequence-structure map is shape space covering, which states that it is sufficient to screen a (high-dimensional) sphere around an arbitrarily chosen sequence to find at least one sequence for every common structure. To test whether shape space covering holds in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$, we constructed hamming distance components $\mathcal{H}_i$. Each $\mathcal{H}_i$ is a pair of sets $\mathcal{H}_i = (V, E_h)$, where the vertices are all sequences with a stable secondary structures in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ $V = \{I_i \in \mathcal{I}^{(9)}_{\mathbf{AUGC}}|\psi(I_i) = S_i \wedge F_{\min}(I_i, S_i) < 0\}$ and the edges are $E_h = \{I_i, I_j \in V|d^h_{ij} = 1\}$. Therefore

an $\mathcal{H}_i$ component includes all sequences which can be accessed from each other by a path of successive $d^h = 1$ neighbors. A connected $\mathcal{H}_i$ component is termed a hamming distance graph. Table 2 shows the decomposition of all sequences of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ into $\mathcal{H}_i$ components. By combining the results of table 1 and table 2 we can easily demonstrate that shape space covering is realized in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$. The results of table 2 are related to table 1 by collating the $\mathcal{H}_i$ components of table 2 to the $\mathcal{G}^l_k$ components of table 1. Lets start with the first row of table 2: This row shows that the majority of all sequences that fold into a stable secondary structure are contained within one huge $\mathcal{H}_i$ component, $\mathcal{H}_1$. This huge $\mathcal{H}_1$ component comprises about 94% of all sequences with a stable secondary structure. Furthermore the sequences in this component fold into all 4 structures realized in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$. Therefore this huge $\mathcal{H}_1$ component contains parts of nearly all $\mathcal{G}^l_k$ components. It is important to keep in mind that the variation operators used to construct the $\mathcal{G}^l_k$ components are less stringent than the variation operator used for the construction of $\mathcal{H}_i$ components.

Next we merged the information contained in table 1 and table 2 to show that shape space covering also holds for the remaining $\mathcal{H}_i$ components. Consider, e.g. structure ((....))., it is contained in one $\mathcal{G}^l_k$ component in table 1, while in table 2 this structure is partitioned into two $\mathcal{H}_i$ components, $\mathcal{H}_1$ and $\mathcal{H}_8$, therefore $\mathcal{H}_1$ and $\mathcal{H}_8$ merge into one component, if base pair substitutions are permitted as additional variation operator. The same is true for $\mathcal{H}_i$ components $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_6$, they merge with the huge $\mathcal{H}_1$ component, if the variation operator includes base pair substitutions. The remaining sequences are partitioned into 3 small isolated $\mathcal{H}_i$ networks, $\mathcal{H}_4$, $\mathcal{H}_5$ and $\mathcal{H}_7$, each including 16 nodes. The minimal hamming distance between sequences in these small isolated networks and sequences that belong to the joint component described above, is $d^h = 2$. Therefore all sequences with a stable secondary structure are accessible from each other within $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$, if $d^h = 1$, $d^h = 2$ and $d^c = 1$ are allowed as variation operators. In other words, we don't need $d^h = 3$ and $d^c = 2$ to connect all stable sequences of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$.

Table 2: A $\mathcal{H}_i$ component is defined as the set of all sequences that differ either by $d^h = 1$ or are connected to each other by a path of sequences that are $d^h = 1$ neighbors. We do not consider the secondary structure of a given sequence for the construction of $\mathcal{H}_i$ components. $\mathcal{H}_i$ differ from $\mathcal{G}_k^l$ components in two respects: First all sequences in a given $\mathcal{G}_k^l$ component ave to fold into structure S_k as mfe structure. Second for construction of $\mathcal{G}_k^l$ components two variation operators $d^h = 1$ and $d^c = 1$ are used.

structure	structure components of Tab. 1							
	((....)).	.((....))					(((...)))	((.....))
No. of components	1	5.1	5.2	5.3	5.4	5.5	1	1
$d^h = 1$ component								
$\mathcal{H}_1$	2384	224	220	0	0	0	196	48
$\mathcal{H}_2$	0	16	0	0	0	0	0	0
$\mathcal{H}_3$	0	0	0	0	0	0	64	0
$\mathcal{H}_4$	0	0	0	16	0	0	0	0
$\mathcal{H}_5$	0	0	0	0	0	16	0	0
$\mathcal{H}_6$	0	0	0	0	0	0	64	0
$\mathcal{H}_7$	0	0	0	0	16	0	0	0
$\mathcal{H}_8$	16	0	0	0	0	0	0	0

To demonstrate that it is sufficient to screen a sphere around an arbitrarily chosen sequence to find at least one sequence for every common structure we decompose the huge $\mathcal{H}_1$ component of table 2 into the different secondary structures, see figure 15. Each colored circle symbolizes a $\mathcal{H}_i$ component of structure neutral sequences, $\mathcal{H}_i^{S_k}$. A $\mathcal{H}_i^{S_k}$ component is a graph $\mathcal{H}_i^{S_k} = (V_{hk}, E_{hk})$, where the vertex set $V_{hk} = \{I_i | \psi(I_i) = S_k\}$ and den edge set $E_{hk} = \{I_i, I_j \in V_{hk} | d_{ij}^h = 1\}$. The size of the circle is approximately proportional to the number of sequences in the corresponding $\mathcal{H}_i^{S_k}$ component. The black lines indicate that two $\mathcal{H}_i^{S_k}$ components are connected through $d^h = 1$ neighbors. It is immediately obvious that the connection between sequences belonging to one and the same structure is only possible by passing

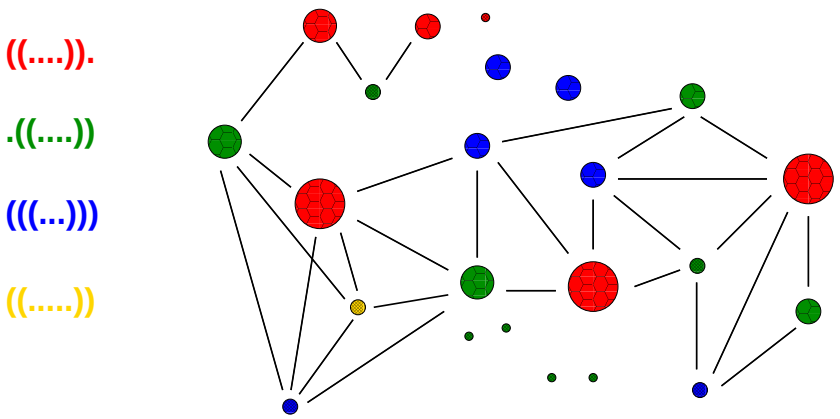


Figure 15: Decomposition of the largest hamming distance component of table 2, $\mathcal{H}_1$, into the different structures. Each colored circles indicates a component of sequences that share the same structure (indicated by the color) and are connected to each other by structure neutral paths of sequences with $d^h = 1$. A black line between two components implies that these components are connected through hamming distance $d^h = 1$ neighbors. The colored circles, that are not connected by black lines, depict the remaining $\mathcal{H}_i$ components

through a neutral network of another structure, if $d^h = 1$ is the only permitted variation operator. This fact proves that within the huge $\mathcal{H}_1$ component in the first row a table 2, shape space covering is fulfilled.
Through decomposition of the sequences in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ into components, which were constructed using stepwise less stringed criteria for the definition of a component, we have shown that the criterion of shape space covering holds for the set of all sequences of length $\ell = 9$ and the alphabet $\mathcal{A} = \{A, U, G, C\}$ that fold into a stable secondary structure.

We next measure the accessibility between the different secondary structures in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$. Accessibility relations between secondary structures are described in detail in Section 2.2.3. Figure 16 shows the accessibility between the four secondary structures realized in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$. The neighborhood frequency $\nu(S_\beta, S_\alpha)$, as well as the occurrence frequency, $\vartheta(S_\beta, S_\alpha)$ are shown, $\nu(S_\beta, S_\alpha)$ is defined in equation (18) and $\vartheta(S_\beta, S_\alpha)$ in equation (19). It is im-

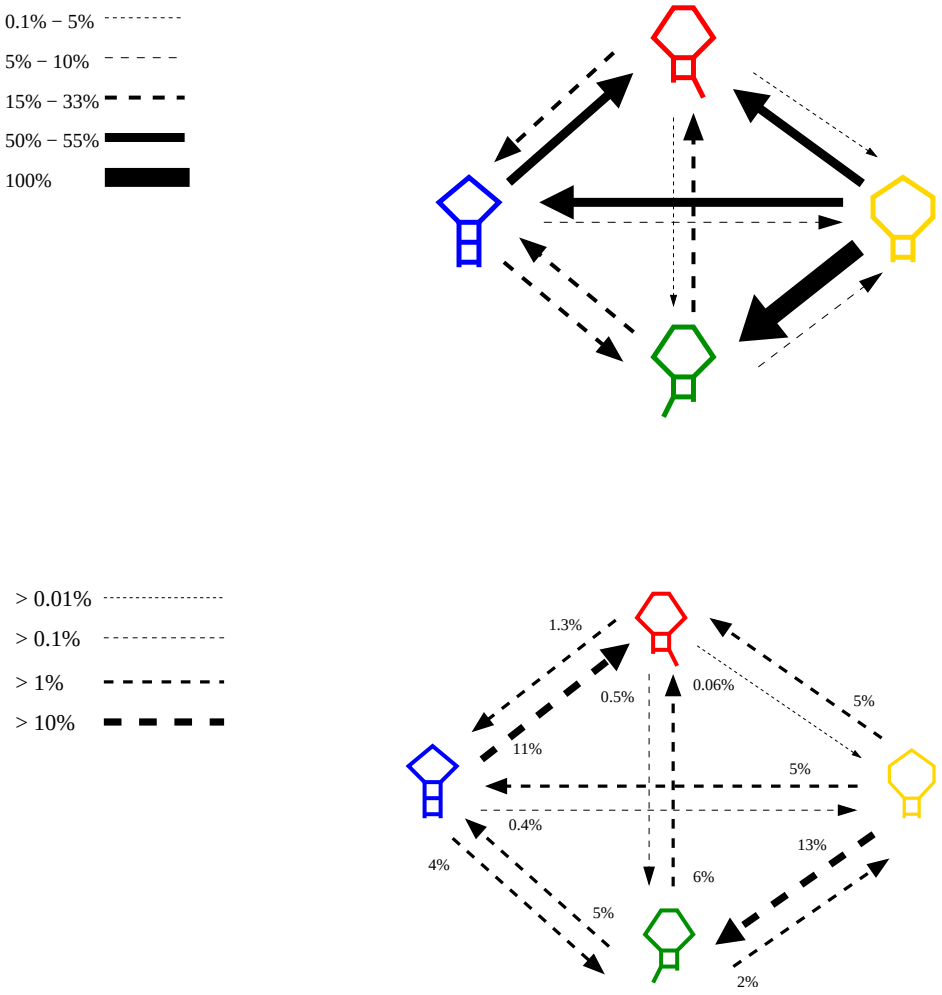


Figure 16: Accessibility relations for RNA secondary structures in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$. The accessibility between two secondary structures S_α and S_β is measured as described by [22], see Section 2.2.3. The neighborhood frequency $\nu(S_\beta, S_\alpha)$, shown in the upper part of the figure, is defined in equation (18) and reflects the likelihood of finding structure S_β in the $d^h = 1$ neighborhood of a randomly chosen sequence of S_α, the frequencies are shown as percentage of structures adopting the described feature. The definition of the occurrence frequency, $\vartheta(S_\beta, S_\alpha)$, show in the lower part of the figure, is given in equation (19) and refers to the total number of occurrences of a structure S_β in the boundary of S_α, B_α.

mediately apparent from figure 16 that the accessibility relations described by $\nu(S_\beta, S_\alpha)$ and $\vartheta(S_\beta, S_\alpha)$ are qualitatively identical. Another property of accessibility relations in RNA structure space is also immediately visible: In general, neighboring structures have very different frequencies of occurrence. This causes an asymmetric accessibility relation between their neutral networks:
Consider for example structures ((.....)), colored yellow in figure 16, and structure .((....)), colored green. All 48 sequences folding in structure ((.....)) have at least one $d^h = 1$ neighbor folding into structure .((....)), which corresponds to a neighborhood frequency $\nu(S_\beta, S_\alpha) = 1$. The situation is different for the set of sequences folding into structure .((....)) . In this set the 48 sequences that have $d^h = 1$ neighbors adopting structure ((.....)) comprise less than 10% of all sequences in this set. The probability that an offspring of a sequence folding into structure ((.....)) acquires structure .((....)) as a consequence of mutation, is therefore much higher than the probability that a sequence with .((....)) phenotype has an offspring folding in structure ((.....)) . Figure 16 also shows that the asymmetric accessibility relations between different neutral networks results in unequal neighborhood contexts. Structure ((.....)), which is the rarest structure realized in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$, is definitely nearer to all other structures, than any of them is to structure ((.....)) . The revers is true for the most common structure ((....))., all other structures in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ are nearer to ((....))., than it is to any of them. As a result of this asymmetric neighborhood relation the most common shape ((....)). is significantly easier to access from all other shapes in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ than vice versa.

One consequence of the shape space topology induced by the genotype phenotype map of RNA is the fact that the neutral networks of two structures S_k and S_l, $\mathcal{G}_k$ and $\mathcal{G}_l$, come very close together on the set $\mathcal{C}_k \bigcap \mathcal{C}_l$ of sequences that are compatible with both structures [63, 64]. We call the sequences that are compatible with more than one secondary structure and can adopt these structures with a negative free energy, i.e. as stable struc-

Table 3: Intersection sequences in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$, I. The first column of the table holds the number of intersection sequences that follow the pattern given in the second column. The subsequent columns indicate the structures that are compatible with the sequences within each row. The negative floating-point number is the free energy of a given sequence I_i, when folding into the indicated structure S_k, $F(S_k, I_i)$. The free energy of the mfe structure S_m, $F_{\min}(S_m, I_i)$ of sequence I_i is indicated by **bold print**. If two or more structures share the same minimal free energy **bold print** denotes which structure is used for structure component decomposition. A + states, that the sequence is compatible with a given structure, but that the free energy of this structure on the given sequence is positive, i.e. the structure is not stable. − indicates, that the sequence is not compatible with the given structure.

number of sequences	sequences	((....)).	.((....))	(((...)))	((.....))
72	$\begin{matrix}A\,A\,A\\ G\,G\,G\begin{matrix}U\,U\,U\\ G\,G\,G\end{matrix}C\,C\,C\\ C\,C\,C\end{matrix}$	+	+	**−0.9**	−0.6
64	$\begin{matrix}A\,A\,A\\ G\,G\,C\begin{matrix}U\,U\,U\\ G\,G\,G\end{matrix}G\,G\,C\\ C\,C\,C\end{matrix}$	−	−	**−1**	−0.1
64	$\begin{matrix}A\,A\,A\\ G\,C\,C\begin{matrix}U\,U\,U\\ G\,G\,G\end{matrix}G\,G\,C\\ C\,C\,C\end{matrix}$	−	−	**−1**	−0.7
16	$\begin{matrix}A\,A\\ G\,G\,G\,G\begin{matrix}U\,U\\ G\,G\end{matrix}A\,C\,C\\ C\,C\end{matrix}$	−	−0.1	−	**−0.1**
12	$\begin{matrix}A\,A\\ G\,G\,G\begin{matrix}U\,U\\ -\,G\end{matrix}C\,C\,C\,C\\ C\,C\end{matrix}$	**−1.4**	+	−0.9	−0.6
12	$\begin{matrix}A\,A\\ C\,C\,C\begin{matrix}U\,U\\ G\,G\end{matrix}G\,G\,G\,G\\ -\,C\end{matrix}$	**−1.9**	+	−0.9	−0.6

Table 4: Intersection sequences in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$, II. The first column of the table holds the number of intersection sequences, that follow the pattern given in the second column. The subsequent columns indicate the structures that are compatible with the sequences within each row. The negative floating-point number is the free energy of a given sequence I_i, when folding into the indicated structure S_k, $F(S_k, I_i)$. The free energy of the mfe structure S_m, $F_{\min}(S_m, I_i)$ of sequence I_i is indicated by **bold print**. If two or more structures share the same minimal free energy **bold print** denotes which structure is used for structure component decomposition. A + states, that the sequence is compatible with a given structure, but that the free energy of this structure on the given sequence is positive, i.e. the structure is not stable. − indicates, that the sequence is not compatible with the given structure.

number of sequences	sequences	((....)).	.((....))	(((...)))	((.....))
8	$\begin{matrix} A\,A \\ CCCC\overset{U}{\underset{G}{-}}GGG \\ C\,C \end{matrix}$	+	**−0.9**	−0.9	−0.6
8	$\begin{matrix} A \\ GGGG\overset{U}{\underset{G}{}}\overset{U}{\underset{G}{}}CCC \\ C \end{matrix}$	+	−0.6	**−0.9**	−0.6
4	$\begin{matrix} A \\ CCCC\overset{U}{\underset{G}{}}GGGG \\ C \end{matrix}$	**−1.9**	−0.9	−0.9	−0.6
4	$\begin{matrix} A \\ CCCC\overset{U}{\underset{G}{}}UGGG \\ C \end{matrix}$	+	**−0.9**	−0.9	−0.6
4	$\begin{matrix} A \\ GGGG\overset{U}{\underset{G}{}}ACCC \\ C \end{matrix}$	**−0.9**	−0.6	−0.9	−0.6
4	$\begin{matrix} A \\ GGGG\overset{U}{\underset{G}{}}CCCC \\ C \end{matrix}$	**−1.4**	−0.6	−0.9	−0.6

tures, *intersection sequences*. The set of all intersection sequences is then $\mathcal{X} = \{I_i \in \mathcal{C}_k \bigcap \mathcal{C}_l | F(S_k, I_i) < 0 \wedge F(S_l, I_i) < 0\}$. The 272 intersection sequences found in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ are listed in table3 and table4. We found 12 different combinations of compatible structures and different energy values for these structures when adopted by different $\mathcal{C}_k$ sequences. 75% of the intersection sequences are compatible with only two structures. $20,6\%$ of the intersection sequences are compatible with three different structures and can adopt these structures with a negative free energy. Only 4.4% of the intersection sequences are compatible with all four structures realized in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ and can fold into them with a negative free energy. The majority of intersection sequences adopts shape (((...))) as there mfe structure, but the remaining three shapes realized in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ are also found as the mfe structure of $\mathcal{C}_k$ sequences. The set of all intersection sequences of length $\ell = 9$ and the alphabet $\mathcal{A} = \{A, U, G, C\}$ comprises a connected graph if the variation operators are base substitutions $d^h = 1$ and base pair substitutions $d^c = 1$. The fact, that all intersection sequences of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ are contained within one connected component demonstrates that the neutral networks of two structures come very close together on the set of their intersection sequences.

Some of the results obtained from $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ were also calculated for the set of all sequences of length $\ell = 10$ and the alphabet $\mathcal{A} = \{A, U, G, C\}$ that fold into a stable secondary structure. In case of $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$, the number of sequences, that have a stable mfe structure is of course larger. 40345 sequences out of the 4^{10} possible sequences of length $\ell = 10$ over the natural alphabet fold into 9 different secondary structures, see figure 17.
While the set of stable structures of length $\ell = 9$ is dominated by one structure, which is formed by more than 70% of all sequences, more than 60% of the sequences of length $\ell = 10$ are nearly equally distributed between three structures, that occur with almost the same frequency. Together with structure ((.....))., which is the fourth common structure in $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$, and occurs with a frequency slightly below that of the three most common structures,

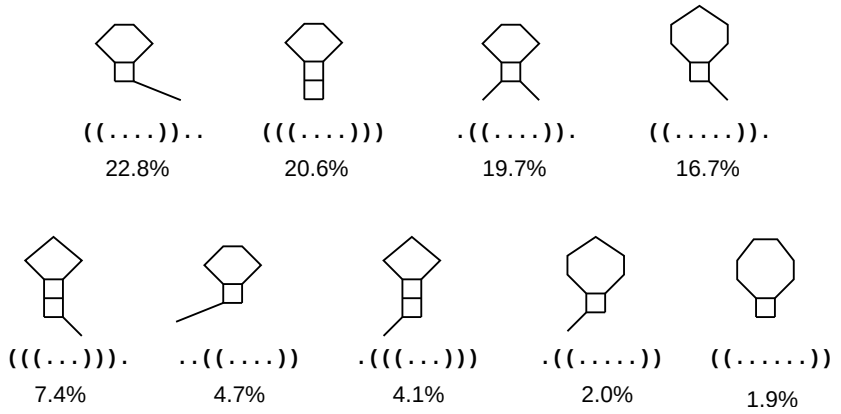


Figure 17: Structures in $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ that fold into a stable mfe structure in graph and dot-bracket notation. The frequencies given are calculated as described in figure 14.

the four highest ranking structures are realized by nearly 80% of the sequences adopting a stable secondary structure. The fact, that the highest ranking structures of $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ occur with almost the same frequency, is in agreement with the results obtained from large samples derived by folding random sequences of fixed chain length [7, 22, 73].
Table 5 shows the decomposition of $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ into $\mathcal{G}^l_k$ components. The three largest structure components are connected networks. Some of the structure components occupied by fewer sequences are scatted networks with many components.

In figure 18 the accessibility relations for RNA secondary structures in $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ are shown. For the majority of shapes accessibility relations are asymmetric. In a few cases, however, the accessibility relation between two structures is symmetric.

Next we examined the composition of the neighborhood of sequences comprising the different $\mathcal{G}^l_k$ components. We define sequence neighborhood in a graph theoretic sense. Sequence I_i is a neighbor of sequence I_j, if I_i can be

Table 5: Sequences of $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ folding into the same secondary structure are decomposed into mutationally connected components using single base exchanges and base pair substitutions as variation operators.

structure	number of sequences	number of components	sequences per component
((....))..	9214	1	9214
(((....)))	8312	1	8312
.((....)).	7964	1	7964
((.....)).	6752	2	6624, 128
(((...))).	2996	1	2996
..((....))	1883	5	892, 803, 64, 64, 60
.(((...)))	1656	1	1656
.((.....))	816	5	496, 128, 64, 64, 64
((.....))	752	1	752

converted into I_j by applying one of the variation operators once. We used $d^h = 1$ and $d^c = 1$ mutations as variation operators. The neighborhood of a sequence I_i, $\mathcal{N}(I_i)$, is then the set $\mathcal{N}(I_i) = \{I_j | d^h_{ij} = 1 \vee d^c_{ij} = 1\}$. Using this definition of sequence neighborhood we find two categories of sequences within a $\mathcal{G}^l_k$ component. On category is the set of sequences having at least one neighbor folding into a structure different from the mfe structure of this $\mathcal{G}^l_k$ component $\mathcal{O}(\mathcal{G}^l_k) = \{I_i \in \mathcal{G}^l_k | \exists I_y \in \mathcal{N}(I_i) : I_y \notin \mathcal{G}^l_k\}$. These $\mathcal{O}(\mathcal{G}^l_k)$ sequences, therefore, make a connection to another $\mathcal{G}^l_k$ component. Alternatively there is the set of sequences having only neighbors within their own $\mathcal{G}^l_k$ component, i.e. neighbors adopting the same mfe structure than these sequences. Were termed the set of these sequences the *interior* of a $\mathcal{G}^l_k$ component interior$(\mathcal{G}^l_k) = \{I_i \in \mathcal{G}^l_k | \nexists I_y \in \mathcal{N}(I_i) : I_y \notin \mathcal{G}^l_k\}$.
The set $\mathcal{O}(\mathcal{G}^l_k)$ of sequences having neighbors in other $\mathcal{G}^l_k$ components is split in two subsets. One subset of $\mathcal{O}(\mathcal{G}^l_k)$ are the intersection sequences, intersect$(\mathcal{G}^l_k) = \{I_i \in \mathcal{G}^l_k | \exists I_y \in \mathcal{N}(I_i) : I_y \notin \mathcal{G}^l_k \wedge I_i \in \mathcal{X}\}$. The other subset

Table 6: Sequence composition of the $\mathcal{G}_k^l$ components of $\mathcal{I}_{\mathbf{AUGC}}^{(10)}$, I. All sequences comprising the $\mathcal{G}_k^l$ components of $\mathcal{I}_{\mathbf{AUGC}}^{(10)}$ are sampled according to the secondary structures realized in their sequence neighborhood, $\mathcal{N}(I_i)$. If a structure $S_k = \psi(I_i)$ has more than one $\mathcal{G}_k^l$ component, each $\mathcal{G}_k^l$ component is shown separately. For the definition of the different neighborhood contexts see text. The columns labeled "stable neighbors" show mean and standard deviation of the number of neighbors with a stable secondary structure.

structure	neighborhood context	% of seq.	stable neighbors mean	σ.
((....))..	intersect	7.39	22.27	2.27
	$\mathcal{O}_k^{\mathrm{nc}}$	48.07	18.60	2.27
	interior	44.54	17.51	2.28
(((...))).	intersect	10.28	29.36	3.10
	$\mathcal{O}_k^{\mathrm{nc}}$	71.56	21.13	5.34
	interior	18.16	13.87	0.54
.((....)).	intersect	5.22	22.19	2.23
	$\mathcal{O}_k^{\mathrm{nc}}$	94.58	18.53	2.61
	interior	0.20	11.00	0.00
((.....)).	intersect	6.82	21.17	2.01
	$\mathcal{O}_k^{\mathrm{nc}}$	76.87	18.43	2.11
	interior	16.30	18.46	2.26
	$\mathcal{O}_k^{\mathrm{nc}}$	100.00	15.69	1.29
(((....)))	intersect	6.30	29.06	4.59
	$\mathcal{O}_k^{\mathrm{nc}}$	70.49	21.74	5.46
	interior	23.21	15.51	1.83
((......))	intersect	8.51	19.50	1.33
	$\mathcal{O}_k^{\mathrm{nc}}$	91.49	15.98	0.91

Table 7: Sequence composition of the $\mathcal{G}_k^l$ components of $\mathcal{I}_{\mathbf{AUGC}}^{(10)}$, II. All sequences comprising the $\mathcal{G}_k^l$ components of $\mathcal{I}_{\mathbf{AUGC}}^{(10)}$ are sampled according to the secondary structures realized in their sequence neighborhood, $\mathcal{N}(I_i)$. If a structure $S_k = \psi(I_i)$ has more than one $\mathcal{G}_k^l$ component, each $\mathcal{G}_k^l$ component is shown separately. For the definition of the different neighborhood contexts see text. The columns labeled "stable neighbors" show mean and standard deviation of the number of neighbors with a stable secondary structure.

structure	neighborhood context	% of seq.	stable neighbors mean	σ.
.((.....))	intersect	6.45	17.03	0.18
	$\mathcal{O}_k^{\mathrm{nc}}$	93.55	14.01	1.14
	intersect	3.12	20.00	0.00
	$\mathcal{O}_k^{\mathrm{nc}}$	96.88	13.31	0.88
	$\mathcal{O}_k^{\mathrm{nc}}$	100.00	10.00	0.00
	$\mathcal{O}_k^{\mathrm{nc}}$	31.25	10.00	0.00
	interior	68.75	9.00	0.00
	$\mathcal{O}_k^{\mathrm{nc}}$	75.00	10.08	0.28
	interior	25.00	9.00	0.00
..((....))	intersect	3.92	19.26	4.98
	$\mathcal{O}_k^{\mathrm{nc}}$	55.61	14.09	1.53
	interior	40.47	12.02	1.04
	intersect	2.37	27.05	2.01
	$\mathcal{O}_k^{\mathrm{nc}}$	59.90	15.35	2.55
	interior	37.73	13.23	1.18
	$\mathcal{O}_k^{\mathrm{nc}}$	28.12	11.28	0.96
	interior	71.88	9.00	0.00
	$\mathcal{O}_k^{\mathrm{nc}}$	6.25	11.00	0.00
	interior	93.75	9.00	0.00
	$\mathcal{O}_k^{\mathrm{nc}}$	53.33	11.00	1.44
	interior	46.67	9.00	0.00
.(((...)))	intersect	10.39	23.13	1.77
	$\mathcal{O}_k^{\mathrm{nc}}$	83.57	17.88	2.20
	interior	6.04	15.00	0.00

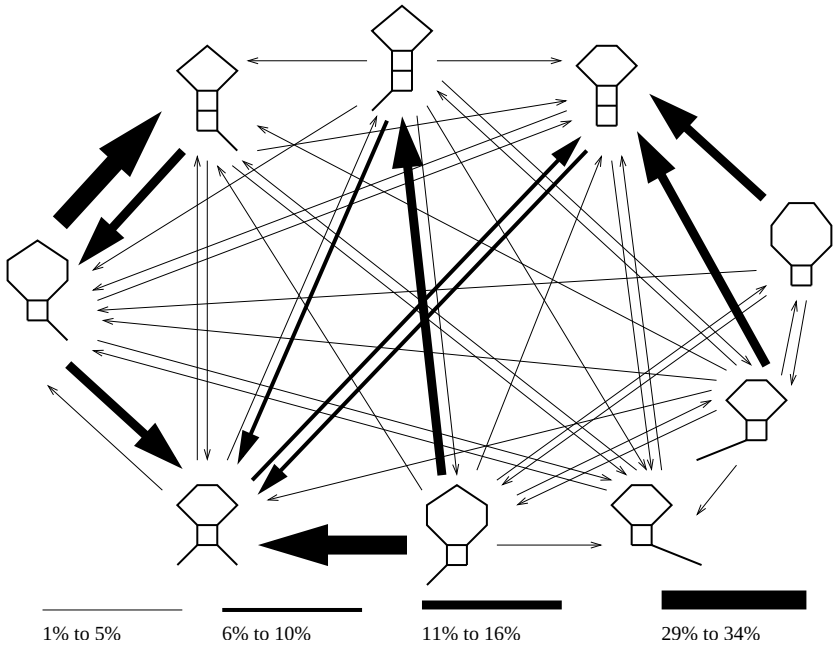


Figure 18: Accessibility relations for RNA secondary structures in $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$. The accessibility between two secondary structures S_α and S_β, is measured as described by [22], see Section 2.2.3. For $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ only the occurrence frequency, $\vartheta(S_\beta, S_\alpha)$ is shown. The definition of the occurrence frequency, $\vartheta(S_\beta, S_\alpha)$ is given in Equation (19) and refers to the total number of occurrences of a of structure S_β in the boundary of S_β, B_α. In the figure the frequencies are shown as percentages.

$\mathcal{O}^{\mathrm{nc}}_k(\mathcal{G}^l_k)$, comprises all sequences, that have neighbors folding into structures different from their mfe structure, but are not intersection sequences $\mathcal{O}^{\mathrm{nc}}_k(\mathcal{G}^l_k) = \{I_i \in \mathcal{G}^l_k l | I_i \in \mathcal{O}(\mathcal{G}^l_k) \backslash \mathrm{intersect}(\mathcal{G}^l_k)\}$.
In the majority of the $\mathcal{G}^l_k$ components shown in table 6 and table 7 more than 50% of all sequences have neighbors in other $\mathcal{G}^l_l$ components. In the $\mathcal{G}^l_k$ component of structure ((.....)). 100% of the sequences have neighbors in other $\mathcal{G}^l_k$ components. The three $\mathcal{G}^l_k$ components in which less than 50% of all sequences have neighbors in other $\mathcal{G}^l_k$ components, form small networks comprising maximal 64 sequences and might be an artifact of the folding algorithm. But even these three atypical $\mathcal{G}^l_k$ components include sequences

that have neighbors folding into other structures. The results of tables 6 and table 7 clearly show, that the set of all sequences of $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ that fold into a stable secondary structure represents a densely connected network, proving that shape space covering also holds for $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$. The set intersect$(\mathcal{G}^l_k)$ of all intersection sequences plays a decisive role in connecting the sequences of $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$. All intersection sequences have neighbors folding into stable structures different from their mfe structure. Furthermore, intersection sequences have on average more stable neighbors than sequences that are interior to one structure component.
By exhaustive folding and enumeration of all stable sequences in $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ and $\mathcal{I}^{(10)}_{\mathbf{AUGC}}$ we can show that the generic properties of the RNA sequence-structure map are realized in these two sequences spaces.

3.1.3 Small world features of neutral networks

To study the emergence of small world properties within the different connected components we implement a database interface for graph algorithms. PostgreSQL databases allow to create user-defined functions written in the C programming language. These functions are compiled into shared libraries and loaded by the database server on demand. Another feature of the PostgreSQL database is the Server Programming Interface (SPI), which gives users the ability to run SQL queries inside these user-defined C functions. We employ these features of PostgreSQL to implement an interface between the database and the ViGl graph library, which provides a variety of fundamental graph algorithms.

This interface between the database and the ViGl graph library was used to analyze the emergence of small-world features in the different structure neutral components of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$. As discussed in section 2.3, Watts and Strogatz [85] found graphs that can be highly clustered, like regular networks, but show small characteristic path length, like random graphs. These networks, which they termed small-world networks, have been shown to be pervasive

in a range of networks that arise from both natural sources and man-made technology [1, 85].

Small world properties of a connected undirected graph are quantified by two parameters. The mean path length $\langle L \rangle$ is defined as the number of edges in the shortest path between two vertices, averaged over all pairs of vertices. For the calculation of L a priority-first-search algorithm is employed. The clustering coefficient $\langle C_i \rangle$ of a vertex i, which measures the 'cliquishness' of the neighborhood of this vertex, determines what fraction of the vertices adjacent to i are also adjacent to each other. In extension, the clustering coefficient of a graph is defined as the average of C_i over all i. For the calculation of C we provide functions generating all $d^h = 1$ neighbors and all $d^c = 1$ neighbors containing one compensatory mutation. These functions works on unsigned integer encoded sequences using bit-shift operations. Using bit-shift operations for the implementation of the variation operators accelerated computation considerably. The combination of a database, which provides optimal storage management, with a graph library enables us to study the topological structure of large networks.

Table 8 shows the comparison of the $\mathcal{G}_k^l$ components of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ with random graphs. In order to compare a $\mathcal{G}_k^l$ component to a random graph the number of vertices and edges must be identical in the two graphs. Table 8 shows that largest $\mathcal{G}_k^l$ component, which contains all sequences, that have structure ((....)). as their mfe structure, exhibits small world features. The mean path length $\langle L \rangle$ of the $\mathcal{G}_k^l$ component of structure ((....)). is comparable L_{random} while the average clustering coefficient $\langle C_i \rangle$ is much larger as C_{random}. With decreasing component size the properties of the $\mathcal{G}_k^l$ components of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$ approach that of random graphs. The smallest structure component, comprising the set of all sequences that fold into structure ((.....)) as their mfe structure, exhibits values for both $\langle C_i \rangle$ and $\langle L \rangle$ that are essentially indistinguishable from that of comparable random graphs.

A difference in the topological structure of $\mathcal{G}_k$ components compared to the results from random graph theory has been described in previous studies [7].

Table 8: Characteristic path length and clustering coefficient for the structure components, $\mathcal{G}_k^l$, of $\mathcal{I}^{(9)}_{\mathbf{AUGC}}$. The mutational operators used to construct the neutral components are point mutations, $d^h = 1$, and compensatory mutations, $d^c = 1$. The rows labeled *random* contain average results for 100 realizations of random graphs with the same vertex and edge numbers as the $\mathcal{G}_k^l$ component directly above them. The column labeled "comp." contains the number of components of each structure, see table 1. σ_C and σ_L show the standard deviation for $\langle C_i \rangle$ and $\langle L \rangle$, respectively.

structure	comp.	nodes	$\langle C_i \rangle$	σ_C	$\langle L \rangle$	σ_L
((....)).	1	2400	0.128	-	4.640	-
random		2400	0.006	.0003	3.222	0.001
.((....))	5	240	0.188	-	3.669	-
random		240	0.037	0.003	2.678	0.003
		220	0.186	-	3.383	-
random		220	0.046	0.004	2.575	0.004
		16	0.400	-	1.500	-
random		16	0.406	0.053	1.545	0.026
(((...)))	1	324	0.162	-	3.591	-
random		324	0.033	0.002	2.689	0.003
((.....))	1	48	0.250	-	2.167	-
random		48	0.170	0.017	2.000	0.013

While random graph theory predicts a single *giant* component, in reality one, two, three or four large components (of almost equal frequency) were found frequently [28]. The quintessence of these studies is that at least large $\mathcal{G}_k$ components are not random graphs.

3.2 Cofolding RNA molecules

3.2.1 Algorithm for cofolding two RNA molecules

The basis of our algorithm is a modified version of the recursions for the equilibrium partition function introduced by McCaskill [50] as implemented in the Vienna RNA package [33]. We assume that the interaction between two RNA molecules is a stepwise process. In a first step, the target molecule has to adopt a structure in which a binding site is accessible. In a second step, the ligand molecule will hybridize with a region accessible to an interaction. Consequently we designed our algorithm as a two step process: The first step is the calculation of the probability that a region within the target is unpaired, or equivalently, the calculation of the free energy needed to expose a region. In a second step we compute the free energy of an interaction for every possible binding site [57]. Using the partition function enables us to consider the whole ensemble of possible structures. Our method is therefor not only more robust against the inherent imprecision of secondary structure prediction algorithms but also more exact than sampling methods.

3.2.2 Probability of an Unpaired Region

In the following let $F(S)$ denote the free energy of a secondary structure S, and write β for the inverse of the temperature times Boltzmann's constant. The equilibrium partition function is defined as $Z = \sum_S \exp(-\beta F(S))$. The partition function is the gateway to the thermodynamics of RNA folding. Quantities such as ensemble free energy, specific heat, and melting temperature can be readily computed from Z and its temperature dependence.
Since the frequency of a structure S in equilibrium is given by $P(S) = \exp(-\beta F(S))/Z$, partition functions also provide the starting point for computing the frequency of a given structural motif. In particular we are interested in the probability $P_u[i, j]$ that the sequence interval $[i, j]$ is unpaired. Denoting the set of secondary structures in which $[i, j]$ remains unpaired by

$\mathcal{S}^u_{[i,j]}$ we have

$$P_u[i,j] = \frac{1}{Z} \sum_{S \in \mathcal{S}^u_{[i,j]}} e^{-\beta F(S)} \tag{26}$$

Clearly, the set $\mathcal{S}^u_{[i,j]}$ will be exponentially large in general. The program `Sfold` [8,9] adds a stochastic backtracking procedure to McCaskill's partition function calculation [50] to generate a properly weighted sample of structures. One then simply counts the fraction of structures with the desired structural feature. This approach becomes infeasible, however, when $P_u[i,j]$ becomes smaller than the inverse of the sample size. Nevertheless, even very small probabilities $P_u[i,j]$ can be of importance in the context of interacting RNAs, as we shall see below.

We therefore present here an exact algorithm. In the special case of an interval of length 1, i.e., a single unpaired base, $P_u[i,i]$ can be computed by dynamic programming. Indeed, $P_u[i,i] = 1 - \sum_{j \neq i} P_{ij}$, where P_{ij} is the base pairing probability of pair (i,j), which is obtained directly from McCaskill's partition function algorithm [50]. It is natural, therefore, to look for a generalization of the dynamic programming approach to longer unpaired stretches[2].

We first observe that the unpaired interval $[i,j]$ is either part of the "exterior loop", (i.e., it is not enclosed by a basepair), or it is enclosed by a base pair (p,q) such that (p,q) is the closing pair of the loop that contains the unpaired interval $[i,j]$. We can therefore express $P_u[i,j]$ in terms of restricted partition functions for these two cases:

$$P_u[i,j] = \underbrace{\frac{Z[1,i-1] \times 1 \times Z[j+1,N]}{Z}}_{exterior} + \sum_{\substack{p,q \\ p<i\leq j<q}} \underbrace{P_{pq} \times \frac{Z_{pq}[i,j]}{Z^b[p,q]}}_{enclosed} \tag{27}$$

The first term accounts for the ratio between the partition functions of all sub-structures on the 5' and 3' side of the interval $[i,j]$ and the total par-

[2]Note that we cannot simply use $\prod_{k=i}^{j} P_u[k,k]$ since these probabilities are not even approximately independent.

tition function. In the second term, $Z_{pq}[i,j]$ is the partition function over all structures on the subsequence $[p,q]$ subject to the restriction that $[i,j]$ is unpaired and (p,q) forms a base pair, while $Z^b[p,q]$ counts all structures on $[p,q]$ that form the pair (p,q). Multiplying the ratio of these two partition functions by the probability P_{pq} that (p,q) is indeed paired yields the desired fraction of structures in which $[i,j]$ is left unpaired.

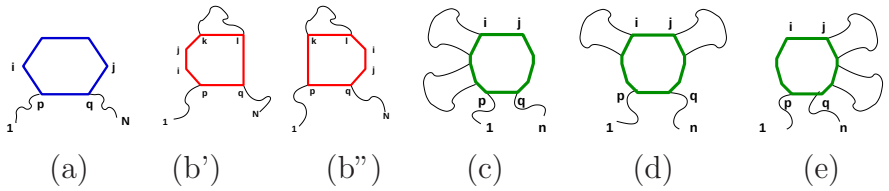


Figure 19: A base pair p,q can close various loop types. According to the loop type different contributions have to be considered. a) A hairpin loop is depicted in blue. b) In case of an interior loop, which is shown in red, two indepent contributions to $Q_{pq}[i,j]$ are possible: The unstructured region $[i,j]$ can be located on either side of the stacked pairs (p,q) and (k,l). c) If region $[i,j]$ is contained within a multiloop we have to account for three different conformations, indicated in the green structures, a more detailed description is given in the text.

The tricky part of the algorithm is the computation of the restricted partition functions $Z_{pq}[i,j]$. The recursion is built upon enumerating the possible types of loops that have (p,q) as their closing pair and contain $[i,j]$, see Fig. 19. From this decomposition one derives:

$$
\begin{aligned}
Z_{pq}[i,j] = & \underbrace{\exp(-\beta H(p,q))}_{(a)} \\
& + \sum_{\substack{p<i\le j<k \text{ or} \\ l<i\le j<q}} \underbrace{Z^b[k,l]\exp(-\beta I(p,q;k,l))}_{(b)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^{m2}[p+1,i-1]\exp(-\beta c(q-i))}_{(c)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^{m}[p+1,i-1]Z^{m}[j+1,q-1]\exp(-\beta c(j-i+1))}_{(d)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^{m2}[j+1,q-1]\exp(-\beta c(j-p))}_{(e)}
\end{aligned}
\tag{28}
$$

where $H(p,q)$ and $I(p,q;k,l)$ are functions that compute the loop energies of hairpin and interior loops given their enclosing base pairs; c is an energy parameter for multiloops describing the penalty for increasing the loop size by one. The computation of the multiloop contributions (c-e) requires two additional types of restricted partitions functions: $Z^m[p,q]$ is the partition function of all conformations on the interval $[p,q]$ that are part of a multiloop and contain at least one component, i.e., that contain at least one substructure that is enclosed by a base pair. These quantities are computed and tabulated already in the course of McCaskill's algorithm. There, the computation of Z^m requires an auxiliary array Z^{m1} which counts structures in multiloops that have *exactly* one component, the closing pair of which starts at the first position of the interval. For the one-sided multiloop cases (c) and (e) in Fig. 19 we additionally need the partition functions of multiloop configurations that have *at least* two components. These are readily obtained using

$$
Z^{m2}[p,q] = \sum_{p<u<q} Z^m[p,u]Z^{m1}[u+1,q]\,. \tag{29}
$$

It is not hard to verify that this recursion corresponds to a unique decomposition of the "M2" configurations into a 3' part that contains exactly one

component and a 5' part with at least one component.
It is clear from the above recursions that, in comparison to McCaskill's partition function algorithm, we need to store only one additional matrix, Z^{m2}. The CPU requirements increase to $\mathcal{O}(n^4)$ (assuming the usual restriction of the length of interior loops). In practice, however, the probabilities for very long unpaired intervals are negligible, so that $P_u[i,j]$ is of interest only for limited interval length $|j-i+1| \leq w$. Taking this constraint into account shows that the CPU requirements are actually only $\mathcal{O}(n^3 \cdot w)$.

3.2.3 Interaction Probabilities

The values of $P_u[i,j]$ can be of interest in their own right: Hackermüller, Meisner, and collaborators [29, 51] showed that the binding of the HuR protein to its mRNA target depends quantitatively on the probability that the HuR binding site has an unpaired conformation. While not much is known about the energetics of RNA-protein interactions, the case of RNA-RNA interactions can be modeled in more detail: The energetics of RNA-RNA interactions is viewed as a step-wise process, $\Delta G = \Delta G_u + \Delta G_h$, in which the free energy of binding consists of the contribution ΔG_u that is necessary to expose the binding site in the appropriate conformation, and contribution ΔG_h that describes the energy gain due to hybridization at the binding site. This additivity assumes that the energy of the original loop is unchanged by the binding of the oligo. For an unpaired binding motif in the interval $[i,j]$, we have of course $\Delta G_u = (-1/\beta)(\ln Z_u[i,j] - \ln Z) = (-1/\beta) \ln P_u[i,j]$. Since the energy gain from the hybridization can be substantial, it becomes necessary to deal also with very small values of $P_u[i,j]$. The sampling approach thus becomes infeasible.
The computation of the hybridization part is performed similar to `RNAduplex` or `RNAhybrid`: We assume that the binding region may contain mismatches and bulge loops. Thus the partition function over all interactions between a region $[i^*, j^*]$ in the small RNA and a segment $[i,j]$ in the target RNA is obtained recursively by summing over all possible interior loops closed by

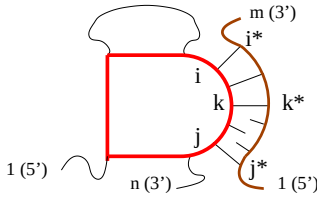


Figure 20: Calculation of the probability of an interaction between a short RNA and its target.

base pairs (k, k^*) and (j, j^*), see Figure 20.

$$Z^I[i,j,i^*,j^*] = \sum_{\substack{i<k<j \\ i^*>k^*>j^*}} Z^I[i,k,i^*,k^*]e^{-\beta I(k,k^*;j,j^*)}. \tag{30}$$

Since we are mostly interested in the binding of miRNAs and siRNAs to a target mRNA, we will neglect internal structures in the short RNA, and include unfolding of the mRNA target site. Thus only Z^I and $P_u[i,j]$ are needed to compute $Z^*[i,j]$, the partition function over all structures where the short RNA binds to region $[i,j]$, and for the computation of the corresponding binding probability, $P^*[i,j]$.

$$Z^*[i,j] = P_u[i,j] \sum_{i^*>j^*} Z^I[i,j,i^*,j^*]; \quad P^*[i,j] = Z^*[i,j] / \sum_{k<l} Z^*[k,l] \tag{31}$$

From $P^*[i,j]$ we can readily compute the probability P_k^* that a position k lies somewhere within the binding site. Note that these are conditional probabilities given that the two molecules bind at all. Furthermore $Z^*[i,j]$ can be used to calculate $\Delta G[ij] = (-1/\beta) \ln Z^*[i,j]$ the free energy of binding, where the binding site is in region $[i,j]$. For visual inspection $\Delta G[ij]$ can be reduced to the optimal free energy of binding at a given position i, $\Delta G_i = min_{k \le i \le l}\{\Delta G[kl]\}$. The memory requirement for these steps is $\mathcal{O}(n \cdot w^3)$, the required CPU time scales as $\mathcal{O}(n \cdot w^5)$, which at least for long target RNAs is dominated by the first step, i.e., the computation of the $P_u[i,j]$.

3.2.4 Interactions between small RNAs and their targets

In order to demonstrate that our algorithm produces biologically reasonable results, we compared predicted binding probabilities with data from RNA interference experiments. Small interfering RNAs (siRNAs) are short (21-23nt) RNA duplices with symmetric 2-3 nt overhangs [11, 52, 53]. They are used to silence gene expression in a sequence-specific manner in a process known as RNA interference (RNAi). Recently, there has been mounting evidence that the biological activity of siRNAs is influenced by local structural characteristics of the target mRNA [4, 42, 53, 60, 67, 90]: a target sequence must be accessible for hybridization in order to achieve efficient translational repression. An obstacle for effective application of siRNAs is the fact that the extent of gene inactivation by different siRNAs varies considerably. Several groups have proposed basically empirical rules for designing functional siRNAs, see e.g. [15, 66], but the efficiency of siRNAs generated using these rules is highly variable. Recent contributions [61, 67] suggest two significant parameters: The stability difference between 5' and 3' end of the siRNA, that determines which strand is included into the RISC complex [38, 74] and the local secondary structure of the target site [4, 42, 53, 60, 67, 90].

Schubert et al. [67] systematically analyzed the contribution of mRNA structure to siRNA activity. They designed a series of constructs, all containing the same target site for the same siRNA. These binding sites, however, were sequestered in local secondary structure elements of different stability and extension. They observed a significant obstruction of gene silencing for the same siRNA caused by structural features of the substrate RNA. A clear correlation was found between the number of exposed nucleotides and the efficiency of gene silencing: When all nucleotides were incorporated in a stable hairpin, silencing was reduced drastically, while exposure of 16 nucleotides resulted in efficient inhibition of expression virtually indistinguishable from the wild type.

We applied our methods to study the target sites provided by Schubert et al. [67]. Our predictions, shown in Fig. 21, are in perfect agreement with the

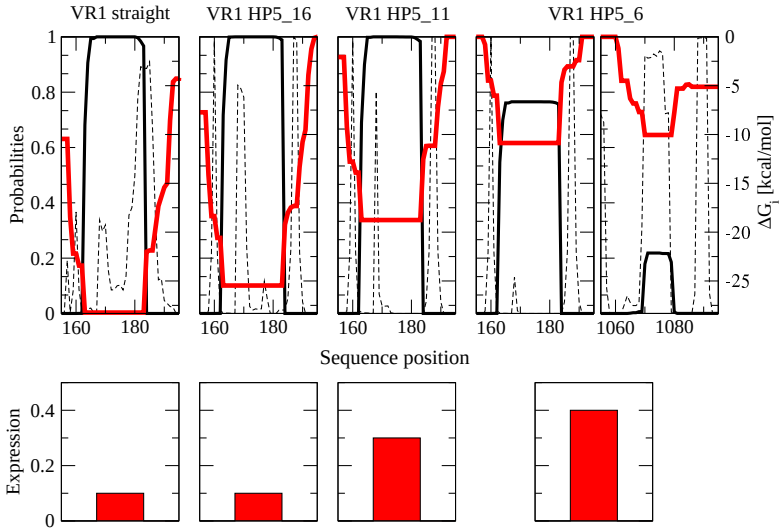


Figure 21: Probability of being unpaired $P_u[i,i]$ (dashed line), probability of binding to siRNA at position i, P_i^*, (thick black line) and ΔG_i, the optimal free energy of binding in a region including position i (thick red line) near the known target site of VsiRNA1. The scale for the probabilities is indicated on the left side, the scale for the minimal free energy of binding on the right side. At the bottom the protein expression levels in experimental data [67] are indicated. The isolated 21mer target sequence, displaying the same activity as the wild type mRNA, and 3 mutants are shown. A decreasing optimal free energy of binding is correlated with increasing expression. In the case of the HP5_6 mutant an alternative binding site becomes occupied as the optimal free energy of binding due to this alternative interaction nearly equals ΔG_i at the proposed target site.

experimental results. The target site of the "VR1straight" construct has a high probability of being unstructured, consequently ΔG_i, the optimal free energy of binding, is highly favorable and the siRNA will bind almost exclusively to the intended target site. The stepwise reduction of the target accessibility is directly correlated to a weaker optimal free energy of binding and decreasing silencing efficiency. In case of construct VR HP5_6 the optimal free energy of binding at an alternative binding site at positions 1066 to 1078 nearly equals that at the proposed target site. Since siRNAs can also function as miRNAs [10, 91], the siRNA might act in a miRNA like fashion

binding to this alternative target site and contribute to the remaining translational repression of this construct. The incomplete complementarity of the siRNA to the alternative target site should be no obstacle to functionality, since it was shown that miRNAs can be active even if the longest continuous helix with the target site is as short as 4 - 5 basepairs [6].
Our new accessibility prediction tool can thus be used to identify potential binding sites as well as explain differences in si/miRNA efficiency caused by secondary structure effects.

4 Results on RNA Evolution

4.1 Evolutionary trajectories of A, U, G, C

We analyzed the relay series of 1200 different evolutionary trajectories to extract common features of RNA structure optimization. The alphabet $\{\mathcal{A}, \mathcal{U}, \mathcal{G}, \mathcal{C}\}$ was used, the length of the sequences was 76 nucleotides. The flow reactor used for our studies was implemented by Andreas Wernitznig [87], it is based on the stochastic method of Gillespie [25,26]. A population size of 3000 RNA sequences was chosen for the run of the reactor. The fitness which is tantamount to the replication rate of a given sequence adopting structure S_i was calculated using following equation:

$$f = \frac{1}{0.01 + (\frac{d^S_{i\tau}}{\ell})^g} \tag{32}$$

where $d^S_{i\tau}$ is the hamming distance of the dot-bracket notations between a given structure S_i and the target structure S_τ, $\ell = 76$ as the length of the sequence and g ($g = 1$) as the weight.

4.1.1 Emergence of families of recurrent structures

These relay series contain about $14,000$ different structures. As expected from the stochastic nature of the simulations 80% of these structures occur only once. A remarkable feature of the remaining structures is that they can recur in a single relay series several times. In particular, 30% of the relay series contain at least one family of recurring structures. This shows that the relay series may not only be monotonic sequences of structures with increasing fitness converging to the target structure but may contain structures that are visited more than once in the process of evolutionary optimization, see Figure 22. Relay series containing families of recurring structures are on average longer than relay series without them. A relay series including at least one family of recurring structures passes on average 33 structures before reaching the target, whereas relay series without families of recurring

structures comprise 22 structures on average.
The structures forming the family of recurring structures (for example upper part of Figure 22, structures $\Omega_3, \Omega_4, \Omega_5, \Omega_6,$) have typically identical hamming distances to the target structure and consequently the same fitness. Therefore, selection does not discriminate between these structures. Another property of structures in these families is their similarity: Each structure differs from its immediate predecessor in the relay series only by closing or opening of a single base pair. Such transitions are termed continuous and occur frequently on point mutations. In the majority of the relay series containing families of recurring structures, the target structure is difficult to access from structures in the family. The transformation to the target structure involves simultaneous formation and cleavage of several base pairs. The probability of such a discontinuous transition is rather low. The target is therefore approached via a small number of "intermediate" structures, which permit a gradual convergence to the target structure.
In the lower part of figure 22 the accessibility between structures forming a family of recurring structures in a relay series is shown. For the determination of accessibility relations between two secondary structures we generated the local shadow of both structures. The local shadow of a structure S_i was obtained by enumerating the set of all secondary structures formed by all possible one-error mutants of 2000 sequences adopting structure S_i as their mfe structure. The accessibility between two structures $\varrho(S_i, S_j)$ is given as the fraction of sequences folding into structure S_j of the $2000*76*3$ sequences generated to obtain the shadow of structure S_i. A structure S_j was defined as accessible from a structure S_i, if S_j occurred in the local shadow of S_i with $\varrho(S_i, S_j) \geq 0.0001$. Note that the accessibility relation between two structures is usually not symmetric. In case of families of recurring structures, however, highly symmetric accessibility relations support the formation of a family. Furthermore, the accessibility of the target structure or of structures that have to be passed on the way to the target, is low in the shadows of members of a family of recurring structures. The low accessibility of

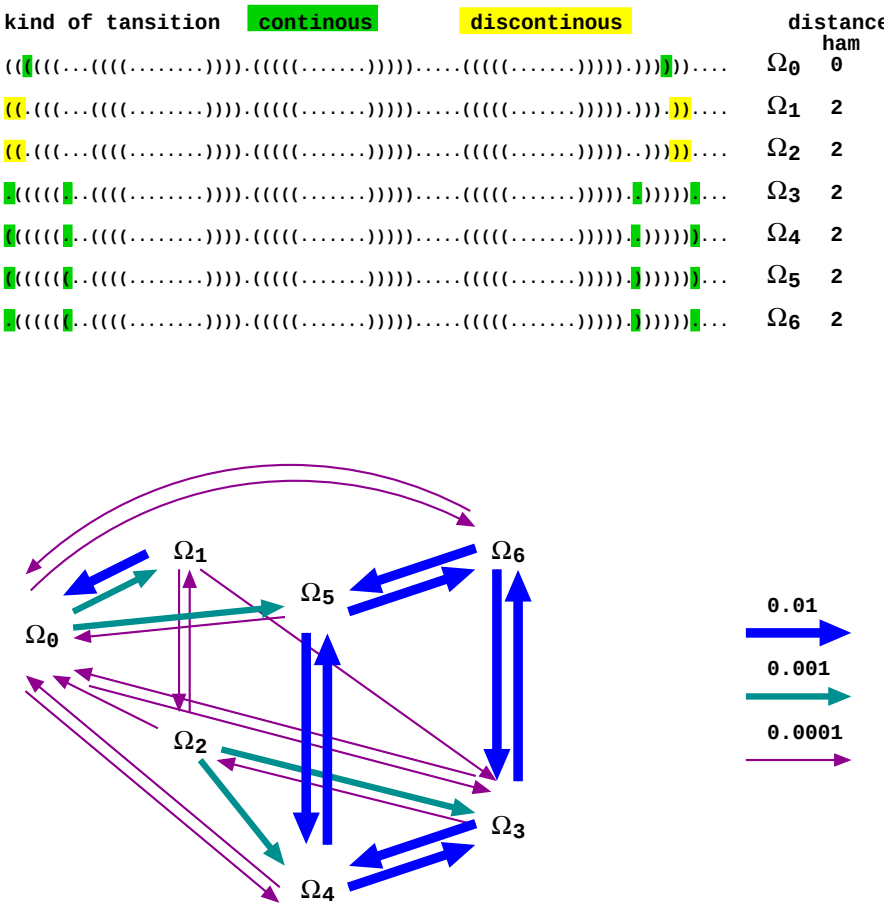


Figure 22: Example of a family of recurring structures in a relay series. In the upper part of the figure, the members of the family of recurring structures ($\Omega_3, \Omega_4, \Omega_5, \Omega_6$) are shown. Structures Ω_2 and Ω_1 allow the transition to the target structure Ω_0. In the lower part of the figure the accessibility between the members of the family of recurring structures, the target structure and the structures occurring between the target and the family of recurring structures are indicated by arrows. The numbers given above an arrow pointing from structure S_i to structure S_j indicate the frequency of a structure S_j in the local shadow of structure S_i, $\varrho(S_i, S_j)$.

structures that have to be visited on the way to the target also supports the formation of a family.

When studying the accessibility relation between the members of a family of recurring structures we noticed that, although the accessibility between two members of a family are highly symmetric, not every member of a family is accessible to all other members of the family, see figure 22. For example structure Ω_4 is accessible to structures Ω_5 and Ω_3 and vice versa, however structure Ω_6 cannot be reached from sequences adopting structure Ω_4. When comparing the non accessible structures we noticed that two family members S_i and S_j that are not accessible to each other differ from each other by a structure distance d^S_{ij} of two, whereas $d^S_{ij} = 1$ for family members that can access each other. But what does a structure distance $d^S_{ij} = 2$ tell us about the relation between the sequences folding into structures S_i and S_j? To answer this question we generated local shadows that contained additional information about the sequences that were analyzed to generate the shadow. Remember that a local shadow is constructed in the following way. 1) Find a sequence $I_{S_{\text{in}}}$ that adopts the input structure S_{in} of the shadow as its mfe structure. 2) Generate all hamming distance $d^H_{I_{S_{\text{in}}},I_x} = 1$ neighbors of this sequence, where we term a given neighbor as I_x. Notice that we are still on the level of sequences! 3) Fold all $d^H_{I_{S_{\text{in}}},I_x} = 1$ neighbors I_x to obtain their mfe structure S_{I_x} and make your statistics. For the current example we calculated the following quantities: For each sequence $I_{S_{\text{in}}}$ that adopts the input structure S_{in} we computed

- the compatibility between $I_{S_{\text{in}}}$ and each of its $d^H_{I_{S_{\text{in}}},I_x} = 1$ neighbors I_x. The compatibility tells us whether a neighbor I_x can fold into the secondary structure S_{in} of $I_{S_{\text{in}}}$. For a given neighbor structure S_{I_x} the compatibility is 1 if all sequences I_x can also fold into the input structure of the shadow S_{in}.

- the energy difference between the mfe of structure S_{in} on a given sequence $I_{S_{\text{in}}}$ and the mfe structure S_{I_x} on a neighbor sequence I_x. This

value is negative if the input structure S_{in} is more stable than the structure of its neighbors. It is positive if the neighbor structure S_{I_x} is more stable.

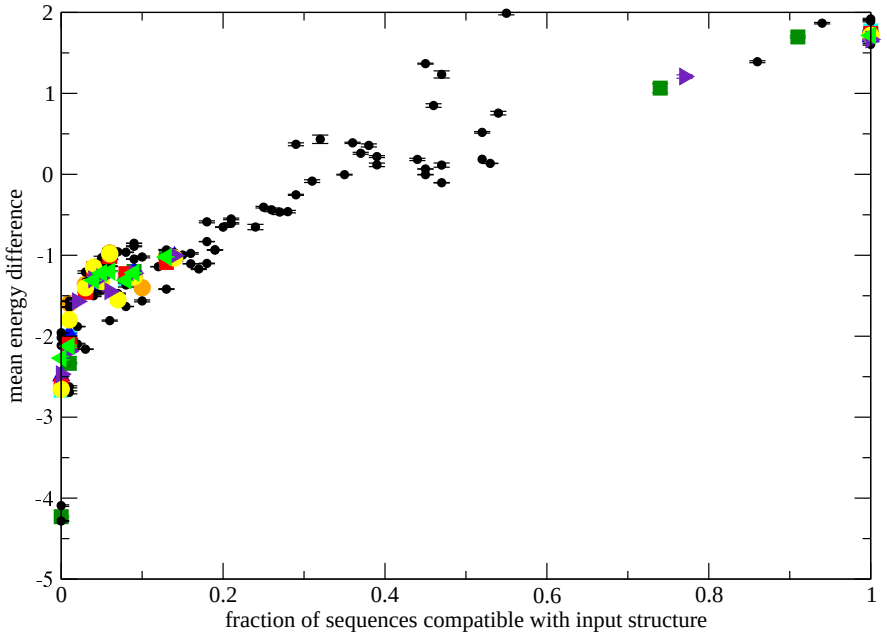


Figure 23: Plot of relative stability versus compatibility for shadows of sequences in a family of recurring structures compared to shadows of structures that occur only once within one of the evolutionary trajectories studied. The results for shadows of family members of recurring structures are depicted as large colored symbols. The results for shadows of structures that occur only once within one of the evolutionary trajectories are shows as small black dots. Only common structures that appear with a $\varrho(S_i, S_j) \geq 0.01$ times in any shadow are considered. Note that all shadows depicted in this figure have structures with a compatibility of 1.

Figure 23 shows a plot of the fraction of neighbor sequences I_x that are compatible with the input structure S_{in} versus the relative stability of the structure S_{I_x} compared to S_{in}. The relative stability of a structure S_{I_x} is given as the mean energy difference between the mfe of the input structure S_{in} and the mfes of each structure S_{I_x} occurring in the shadow. Figure 23 shows that their is a clear correlation between the fraction of compatible

sequences and the relative stability with reference to the input structure. If all sequences, that fold into a structure S_{I_x} are compatible with the input structure S_{in}, then structure S_{I_x} is on the average more stable than the input structure. On the other hand if nearly no sequence adopting structure S_{I_x} is compatible to the input structure, the stability of the input structure is notedly higher. Furthermore the distribution of compatibility values is different between members of families of recurring structures, shown as large colored symbols in figure 23, and structures that occur only once within one of the evolutionary trajectories, depicted as small black dots. Structures that are found within a family have either structural "neighbors" that are highly compatible with and notedly more stable than the input structure. Or they have neighbors that show nearly no compatibility and are articulately less stable. Structures that occur only once within one of the evolutionary trajectories, on the other hand, also have neighbors that show the same stability as the input structure combined with an intermediate compatibility value.
In the context of families of recurring structures the direct correlation between high compatibility to and higher stability of the input structure finds its expression in a super- to substructure relationship between accessible structures. As an example consider structures Ω_3 and Ω_4 in figure 22. Ω_3 is a substructure of Ω_4, i.e. all base pairs in structure Ω_3 are also realized in Ω_4. Every sequence in the shadow of structure Ω_3 that folds into Ω_4 as its mfe structure is compatible with Ω_3, that is the compatibility is one. In accordance with the results of figure 23, structure Ω_4 is more stable than structure Ω_3, an effect of the additional base pair. If we now consider the revers relation, the properties of structure Ω_3 as a neighbor of structure Ω_4, we find that these properties are reversed: Nearly no sequence that folds into Ω_3 is able to build the additional base pair necessary for formation of structure Ω_4, so the compatibility is near to zero and structure Ω_3 is clearly less stable than structure Ω_4. As the shadows of recurring structures contain only common structures that are in conformity with one of these two extremes, the structural variation within the set of easily accessible structure

is low. The variation between the common structures is limited to shrinking or extending one of the stacks or, rarely, losing a whole stack. The majority of the transitions are continuous. Structures that occur only once, on the other hand, allows a broader spectrum of phenotypic variation. The set of common structures also contains variants that can only be realized by a shift of some base pairs, these transitions are discontinuous. This fact is also reflected in the cardinality of the set of common structures which is markedly lower for recurring structures, $\#_{\text{rec}} \simeq 26$ than for structures accruing only once, $\#_{\text{once}} \simeq 33$.

Therefore we conclude that families of recurrent structures are a phenomenon that is based on the sequence structure map. Some Structures allow more variation in their immediate neighborhood than others. Structures that have a restricted set of common neighbors tend to recur in evolution, if the transitions within this set are mainly continuous, as is the case in the families we studied. The optimization process is then confined to members of the family, at least for a while. As a result of this the adaptive process visits more different structures than in a trajectory without a family of recurrent structures. However this imposes no obstacle to reaching the target structure.

5 Conclusion and Outlook

The central topic of this work is the folding of RNA sequences into secondary structure. This process is not only important for the functionality of biologically active RNA molecules, but also plays a role in the evolution of functional RNAs. Since the secondary structure contributes a major portion to the free energy of structure formation, it is an important intermediate in the folding process. Additionally secondary structures are often conserved in phylogenetic evolution. This facts indicate that the secondary structure of an RNA is a target of selection.

In the first part of this work evolutionary dynamics is described by replication and mutation of RNA molecules. Replication as well as mutation operate on the level of the RNA sequences. For a comprehensive theory of evolution, however, the secondary structure submitted to selection, has to be an integral part of the model. In this section the relation between sequence and secondary structure is formulated as a mapping from sequence space into structure space. Sequence structure maps were analyzed in great detail by the group of Peter Schuster [7, 18]. We could proof the results of this studies by exhaustive folding and enumeration of the sequence spaces of short sequences. For our studies the sequences spaces $\mathcal{I}_{AUGC}^{(\ell=9)}$ and $\mathcal{I}_{AUGC}^{(\ell=10)}$ were used, where $\{A, U, G, C\}$ is the alphabet and ℓ the sequence length. By exhaustive folding and enumeration of all sequences that form a stable secondary structure, i.e. a structure with a negative free energy, we showed that there are more sequences than structures. In $\mathcal{I}_{AUGC}^{(9)}$ 3280 sequences with a stable structure fold into 4 different secondary structures. In $\mathcal{I}_{AUGC}^{(10)}$ 40345 sequences adopt 11 different shapes. The frequency distribution of sequences adopting a given structure is highly biased. In $\mathcal{I}_{AUGC}^{(9)}$ 73.8% of all sequences adopt structure ((....))., in $\mathcal{I}_{AUGC}^{(10)}$ nearly 80% of the sequences are distributed with nearly equal frequencies between the four most common structures. This results show that there are few common and many rare shapes.

By constructing connected components of structure neutral sequences, i.e. sequences adopting the same secondary structure, we could show that common structures form extended neutral networks. In contrast to previous studies, we used point mutations and compensatory mutations as variation operators for the construction of structure neutral connected components. Structure neutral networks show a kind of percolation phenomenon. They are connected and span the entire sequence space if the global connectivity of the network exceeds a certain threshold. A phenomenon intimately connected to this fact is shape space covering: Sequences forming common structures are distributed almost randomly in sequence space. Shape space covering in $\mathcal{I}^{(9)}_{AUGC}$ was proven by constructing networks of sequences connected by paths of hamming distance one neighbors. The secondary structure of a sequence was not considered for the construction of these hamming distance graphs. We found that 94% of all stable sequences of $\mathcal{I}^{(9)}_{AUGC}$ are contained within one huge hamming distance graph. The sequences in this graph fold into all four secondary structures realized in $\mathcal{I}^{(9)}_{AUGC}$. By partitioning this huge hamming distance graph into connected structure neutral networks we could show that the minimal distance between hamming graphs of rare structures and hamming graphs of the most common structure is hamming distance one. Although the huge hamming distance graph contains nearly all sequences, there are isolated hamming graphs of rare structures. The minimal distance between these isolated graphs and the huge hamming distance graph is hamming distance two. We conclude that shape space covering is realized in $\mathcal{I}^{(9)}_{AUGC}$, as the minimal hamming distance between any structure neutral hamming graph and the hamming graph containing the common structure is two.

The relationship between sequences and secondary structures of RNA molecules can also be modeled using random graph theory [63]. However, detailed studies showed differences in the topology of random subgraphs compared to structure neutral network [28,63]. We could confirm these results by studying small world features of structure neutral networks. In $\mathcal{I}^{(9)}_{AUGC}$ the majority of

the connected neutral components exhibit small world features, demonstrating that networks of structure neutral components are generally not random graphs.
Because of limitation in processing capability the examination of sequence spaces larger than $\ell = 10$ was not possible. A solution to this problem is parallelization. The algorithm used for the initial setup of the database can be parallelized by dividing the data into chunks and assigning different parallel processes one chunk of data to work on. Parallel versions for the Bellman-Ford-Moore algorithm for the single-source shortest path problem are available. Parallelization enhances the performance drastically and will enable us to study larger sequence spaces.

Structure neutral networks play an important role in evolutionary dynamics. Understanding the dynamics of evolution requires a model of populations searching trough sequence space by replication, mutation and selection. This model was provided by Peter Schusters group through simulation of the evolutionary process in a flow reactor [7, 20, 87]. In this work we studied 1200 relay series of different evolutionary trajectories of the alphabet $\{A, U, G, C\}$ to extract common features of RNA structure optimization towards a predefined target structure. 30% of these relay series contain structures that occurred repeatedly within one and the same relay series. This shows that the relay series may not only be monotonic sequences of structures with increasing fitness converging to the target structure, but may contain structures that are visited more than once in the process of evolutionary optimization. Structures that recur in evolution have a restricted set of common neighbors. Their common neighbors are highly similar to each other, the main difference being the opening or closing of a terminal base pair of a stack. The transition between such structures is continuous, supporting the recurrence of these structures within a single relay series. Evolutionary trajectories including recurring structures visit more different structures than trajectories in which every structure occurs only once.

Various kinds of RNAs play important roles in the regulation of gene expression, genomic organization and post-transcriptional modifications. Recently it has been shown that many of this processes require sequence specific interactions between two RNA molecules [60, 61, 65]. We applied a modified version of McCaskill's partition function algorithm to study the interaction between two RNA molecules. The interaction between two RNA molecules was modeled as a stepwise process: In the first step the target molecule has to fold into a conformation in which the interaction site is accessible. In a second step the ligand interacts with a region accessible for hybridization. Consequently we calculated the free energy needed to expose a region as a first step. Then we proceeded by computing the free energy of an interaction for every possible binding site. Our method provides detailed information about the location of an RNA-RNA interaction, about the structural context of the binding site as well as about the energetics of the interaction [57]. To test the quality of our method we compared our predictions with data from RNA interference experiments. There is mounting evidence that the biological activity of small regulatory RNAs is influenced by the structural context of the target site [4, 42, 60]. A target site within a mRNA must be accessible for hybridization in order to achieve efficient translational repression. We applied our methods to study the target sites provided by Schubert et al. [67]. Schubert et al. observed a significant obstruction of gene silencing caused by structural features of the target RNA. They could show a clear correlation between the number of accessible nucleotides within a target site and the efficiency of gene silencing. Our predictions are in perfect agreement with the experimental results: The stepwise reduction of the target accessibility is directly correlated to a weaker free energy of binding and decreasing silencing efficiency.

In this work we assumed that the influence of the secondary structure of a small ligand molecule can be neglected. To study interactions between two

larger molecules, however, we have to consider the secondary structure of both molecules. This would enable us to compute more complex interactions, as e.g. kissing hairpins. Furthermore we implicitly assumed that the energy change caused by binding of the ligand is a constant. A more realistic model should give consideration to the fact that the binding of the oligo to a loop will of course alter the energy contribution of the loop itself. A limitation for the implementation of this more realistic model is our lack of knowledge concerning the energetics of RNA-RNA interactions within loops. Additional measurement along the lines of the investigation of kissing-interactions [86] are required to improve the energy parameters for interacting RNAs.

Further improvements of our method are possible by including the kinetics and the concentration dependence of the interaction. The hybridization between two RNA molecules is influenced by the kinetics of the interaction. In cases where the target RNA contains two or more possible binding sites for the small regulatory RNA, the selection of the target site is governed by the interaction kinetics. Another factor that influences the interaction at a possible binding site is the concentration of the small RNA and the mRNA.

References

[1] R. Albert and A.L. Barabasi. Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[2] I. Alvarez-Garcia and E.A. Miska. MicroRNA Functions in Animal Development and Human Disease. *Development*, 21:4653–4662, 2005.

[3] A. Aravin and T. Tuschl. Identification and Characterization of Small RNAs Involved in RNA Silencing. *FEBS Lett.*, 579,(26):5830–5840, 2005.

[4] E. A. Bohula, A. J. Salisbury, M. Sohail, M. P. Playford, J. Riedemann, E. M. Southern, and V. M. Macaulay. The Efficacy of Small Interfering RNAs Targeted to the Type 1 Insulin-like Growth Factor Receptor (IGF1R) is Influenced by Secondary Structure in the IGF1R Transcript. *J. Biol. Chem.*, 278(18):15991–15997, 2003.

[5] B. Bollobas. *Random Graphs*. Academic Press, New York, 1985.

[6] J. Brennecke, A. Stark, R.B. Russell, and S.M. Cohen. Principles of MicroRNA-Target Recognition. *PLoS Biol.*, 3(3):e85, 2005.

[7] J.P. Crutchfield and P. Schuster, editors. *Evolutionary Dynamics: Exploring the Interplay of Selection, Accident, Neutrality, and Function*. Oxford University Press. Inc, 2003.

[8] Y. Ding, C. Y. Chan, and C. E. Lawrence. Sfold Web Server for Statistical Folding and Rational Design of Nucleic Acids. *Nucleic Acids Research*, 32(Web Server issue):W135–141, 2004.

[9] Y. Ding and C. E. Lawrence. A Statistical Sampling Algorithm for RNA Secondary Structure Prediction. *Nucleic Acids Res.*, 31:7280–7301, 2003.

[10] J.G. Doench, C.P. Petersen, and P.A. Sharp. siRNAs can Function as miRNAs. *Genes Dev.*, 17(4):438–442, 2003.

[11] D.M. Dykxhoorn, C.D. Novina, and P.A. Sharp. Killing the Messenger: Short RNAs that Silence Gene Expression. *Nat. Rev. Mol. Cell Biol.*, 4(6):457–467, 2003.

[12] M. Eigen. Self Organization of Matter and the Evolution of Biological Macro Molecules. *Naturwissenschaften*, 58(10):465–523, 1971.

[13] M. Eigen, J. McCaskill, and P. Schuster. The Molecular Quasispecies. *Adv. Chem. Phys.*, 75:149–263, 1989.

[14] M. Eigen and P. Schuster. The Hypercycle. *Naturwiss.*, 64:541–565, 1977.

[15] S.M. Elbashir, Harborth. J., K. Weber, and T. Tuschl. Analysis of Gene Function in Somatic Mammalian Cells using Small Interfering RNAs. *Methods*, 26(2):199–213, 2002.

[16] R.A. Fisher. *The Genetical Theory of Natural Selection. A Complete Variorum Edition.* Number ISBN 0-1985-0440-3. Oxford University Press., 2000/1930.

[17] C. Flamm, I.L. Hofacker, S. Maurer-Stroh, P.F. Stadler, and M. Zehl. Design of Multistable RNA Molecules. *RNA*, 7(2):254–65, 2001.

[18] W. Fontana. Modelling 'evo-devo' with RNA. *Bioessays.*, 14(12):1164–77, 2002. Review.

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

[20] W. Fontana, W. Schnabl, and P. Schuster. Physical Aspects of Evolutionary Optimization and Adaptation. *Phys. Rev. A*, 40(3301-3321), 1989.

[21] W. Fontana and P. Schuster. Continuity in Evolution: On the Nature of Transitions. *Science*, 280(5368):1451–5, 1998.

[22] W. Fontana and P. Schuster. Shaping Space: The Possible and the Attainable in RNA Genotype-Phenotype Mapping. *J. Theor. Biol.*, 194(4):491–515, 1998.

[23] C. V. Forst, C. Reidys, and J. Weber. *Advances in Artificial Life*, volume 929, chapter Evolutionary Dynamics and Optimization, pages 128–147. Springer, Berlin, Heidelberg, New York, 1995.

[24] D.J. Futuyma. *Evolutionary Biology*. Sinauer Associates, Massachusetts, 1979.

[25] D.T. Gillespie. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[26] D.T. Gillespie. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[27] 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. *Monatsh. Chem.*, 127:355–374, 1996.

[28] 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.Structure of Neutral Networks and Shape Space Covering. *Monatsh. Chem.*, 127:375–389, 1996.

[29] Jörg Hackermüller, Nicole-Claudia Meisner, Manfred Auer, Markus Jaritz, and Peter F. Stadler. The Effect of RNA Secondary Structures on RNA-Ligand Binding and the Modifier RNA Mechanism: A Quantitative Model. *Gene*, 345:3–12, 2005.

[30] R.W. Hamming. *Coding and Information Theory (2nd ed.)*. Prentice-Hall, Inc., 1986.

[31] A. Herbert. The Four Rs of RNA-directed Evolution. *Nat. Genet.*, 36(1):19–25, 2004.

[32] I.L. Hofacker and P. F. Stadler. RNA secondary structures. unpublished, 2005.

[33] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.

[34] P. Hogeweg and B. Hesper. Energy Directed Folding of RNA Sequences. *Nucl. Acids 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(1):397–401, 1996.

[37] R. Ishitani, O. Nureki, N. Nameki, N. Okada, S. Nishimura, and S. Yokoyama. Alternative Tertiary Structure of tRNA for Recognition by a Posttranscriptional Modification Enzyme. *Cell*, 113(3):383–94, 2003.

[38] A. Khvorova, A. Reynolds, and S. D. Jayasena. Functional siRNAs and miRNAs Exhibit Strand Bias. *Cell*, 115(2):209–16, 2003.

[39] M. Kimura. Evolutionary Rate at the Molecular Level. *Nature*, 217(624-626), 1968.

[40] 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.

[41] M. Kospach. *Molecular Evolution of Short RNA Molecules - Neutral Nets in Sequence Spaces and Kinetic Properties of RNA*. PhD thesis, University of Vienna, 2003.

[42] R. Kretschmer-Kazemi Far and G. Sczakiel. The Activity of siRNA in Mammalian Cells is Related to Structural Target Accessibility: a Comparison with Antisense Oligonucleotides. *Nucleic Acids Res.*, 31(15):4417–4424, 2003.

[43] S.-Y. Le and M. Zuker. Common Structures of the 5 Noncoding Rna in Enteroviruses and Rhinoviruses: Thermodynamical Stability and Statistical Significance. *J. Mol. Biol.*, 216:729–741, 1990.

[44] B. Liao and T. Wang. An Enumeration of RNA Secondary Structure. *Math. Appl.*, (15):109–112, 2002.

[45] D. H. Mathews. Using an RNA Secondary Structure Partition Function to Determine Confidence in Base Pairs Predicted by Free Energy Minimization. *RNA*, 10(8):1178–1190, 2004.

[46] D.H. Mathews, J. Sabina, and D.H. Zuker, M.and Turner. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol*, (288):911–940, 1999.

[47] J.S. Mattick. Challenging the Dogma: the Hidden Layer of Non-protein-coding RNAs in Complex Organisms. *Bioessays*, 10:930–939, 2003.

[48] E. Mayr. *Systematics and the Origin of Species*. Columbia University Press, New York, 1942).

[49] E. Mayr. *What Evolution is*. Number ISBN 0-465-04425-5. Basics Books, 2001.

[50] J.S. McCaskill. The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structures. *Biopolymers*, 29:1105–1119, 1990.

[51] Nicole-Claudia Meisner, Jörg Hackermüller, Volker Uhl, Andras Aszódi, Markus Jaritz, and Manfred Auer. mRNA Openers and Closers: A Methodology to Modulate AU-rich Element Controlled mRNA Stability by a Molecular Switch in mRNA conformation. *Chembiochem.*, 5:1432–1447, 2004.

[52] G. Meister and T. Tuschl. Mechanisms of Gene Silencing by Double-Stranded RNA. *Nature*, 431(7006):343–349, 2004.

[53] V. Mittal. Improving the Efficiency of RNA Interference in Mammals. *Nat. Rev. Genet.*, 5(5):355–365, 2004.

[54] B. Momjian. *PostgreSQL: Introduction and Concepts*. Addison Wesley Professional, 2001.

[55] L. Moran. What is Evolution? online.

[56] T. Mourier. Reverse Transcription in Genome Evolution. *Cytogenet. Genome. Res.*, 110(56-62), 2005.

[57] Ulrike Mückstein, Hakim Tafer, Jör Hackermüller, Stephan Bernhard Bernhard, Peter F. Stadler, and Ivo L Hofacker. Thermodynamics of RNA-RNA Binding. In Andrew Torda, Stefan Kurtz, and Matthias Rarey, editors, *German Conference on Bioinformatics 2005*, volume P-71, pages 3–13, Bonn, 2005. Gesellschaft f. Informatik.

[58] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random Graphs with Arbitrary Degree Distributions and their Applications. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.*, 64(026118):1–17, 2001.

[59] Ruth Nussinov, George Piecznik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for Loop Matching. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.

[60] M. Overhoff, M. Alken, R. K. Far, M. Lemaitre, B. Lebleu, G. Sczakiel, and I. Robbins. Local RNA Target Structure Influences siRNA Efficacy: A Systematic Global Analysis. *J. Mol. Biol.*, 348(4):871–881, 2005.

[61] J. S. Parker, S. M. Roe, and D. Barford. Structural Insights into mRNA Recognition from a PIWI Domain-siRNA Guide Complex. *Nature*, 434:663–666, 2005.

[62] C. Reidys, C. V. Forst, and P. Schuster. Replication and Mutation on Neutral Networks. *Bull. Math. Biol.*, (63):57–94, 2001.

[63] C. Reidys, P.F. Stadler, and P. Schuster. Generic Properties of Combinatory Maps: Neutral Networks of RNA Secondary Structures. *Bull. Math. Biol.*, 59(2):339–97, 1997.

[64] C.M. Reidys. Random Induced Subgraphs of Generalized n-Cubes. *Adv. Appl. Math.*, 19:360–77, 1997.

[65] B.J. Reinhart, F.J. Slack, M. Basson, A.E. Pasquinelli, J.C. Bettinger, A.E. Rougvie, H.R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–6, 2000.

[66] A. Reynolds, D. Leake, Q. Boese, Scaringe S., W.S. Marshall, and A. Khvorova. Rational siRNA Design for RNA Interference. *Nat. Biotechnol.*, 22(3):326–30, 2004.

[67] S. Schubert, A. Grunweller, V.A. Erdmann, and J. Kurreck. Local RNA Target Structure Influences siRNA Efficacy: Systematic Analysis of Intentionally Designed Binding Regions. *J. Mol. Biol.*, 348(4):883–93, 2005.

[68] E.A. Schultes and D.P. Bartel. One Sequence, two Ribozymes: Implications for the Emergence of new Ribozyme Folds. *Science*, 289(5478)):448–52, 2000.

[69] P. Schuster. How to Search for RNA Structures. Theoretical Concepts in Evolutionary Biotechnology. *J. Biotechnol.*, 41(2-3):239–57, 1995.

[70] P. Schuster. Evolution at molecular resolution. In Leif Matsson, editor, *Nonlinear Cooperative Phenomena in Biological Systems*, pages 86–112. World Scientific, Singapore, 1998.

[71] P. Schuster. Evolution in Silico and in Vitro: The RNA Model. *Biol. Chem.*, 382(9):1301–14, 2001. Review.

[72] P. Schuster and W. Fontana. Chance and Necessity in Evolution: Lessons from RNA. *Physica D:Nonlinear Phenomena*, 133:427–452, 1999.

[73] 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.Lond.B*, 255:279–284, 1994.

[74] D.S. Schwarz, G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P.D. Zamore. Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell.*, 115(2):99–208, 2003.

[75] S. Spiegelman. An Approach to the Experimental Analysis of Precellular Evolution. *Quart. Rev. Biophys.*, 4(2):213–53, 1971.

[76] B.M. Stadler, P.F. Stadler, G.P. Wagner, and W. Fontana. The Topology of the Possible: Formal Spaces Underlying Patterns of Evolutionary Change. *J. Theor. Biol.*, 213(2):241–74., 2001.

[77] J.L. Sussman and S. Kim. Three-dimensional Structure of a Transfer RNA in two Crystal Forms. *Science*, 192:853–858, 1976.

[78] M. Tacker, W. Fontana, P.F. Stadler, and P. Schuster. Statistics of RNA Melting Kinetics. *Eur. Biophys. J.*, 23(1):29–38, 1994.

[79] Manfred Tacker, Peter F. Stadler, Erich G. Bornberg-Bauer, Ivo L. Hofacker, and Peter Schuster. Algorithm Independent Properties of RNA Structure Prediction. *Eur. Biophy. J.*, 25:115–130, 1996.

[80] I. Tinoco Jr, P.N. Borer, B. Dengler, M.D. Levin, O.C. Uhlenbeck, D.M. Crothers, and J. Bralla. Improved Estimation of Secondary Structure in Ribonucleic Acids. *Nat. New. Biol.*, 246(150):40–1, 1973.

[81] D.H. Turner, N. Sugimoto, and S.M. Freier. RNA Structure Prediction. *Annu. Rev. Biophys. Biophys. Chem.*, 17:167–192, 1988.

[82] E. van Nimwegen, J.P. Crutchfield, and M. Mitchell. Finite Populations Induce Metastability in Evolutionary Search. *Phys. Letters*, A229:144–50, 1997.

[83] M.S. Waterman and T.F. Smith. RNA Secondary Structure: A Complete Mathematical Analysis. *Math. Biosc.*, 42:257–266, 1978.

[84] J.D. Watson and F.H.C. Crick. Molecular Structure of Nucleic Acids. *nature*, 171:737–738, 1953.

[85] D.J. Watts and S.H. Strogatz. Collective Dynamics of "Small-World" Networks. *Nature*, 393(6684):440–42, 1998.

[86] A. Weixlbaumer, A Werner, C. Flamm, E. Westhof, and R. Schroeder. Determination of thermodynamic parameters for HIV DIS type loop-loop kissing complexes. *Nucleic Acids Res.*, 32:5126–5133, 2004.

[87] A. Wernitznig. *RNA Optimization in Flow Reactors: A Study in Silico.* PhD thesis, University of Vienna, 2001.

[88] S.A. Woodson. Metal Ions and RNA Folding: a Highly Charged Topic with a Dynamic Future. *Curr. Opin. Chem. Biol.*, 2(104-109), 2005.

[89] S Wuchty, W Fontana, I L Hofacker, and P Schuster. Complete Suboptimal Folding of RNA and the Stability of Secondary Structures. *Biopolymers*, 49:145–165, 1999.

[90] K. Yoshinari, M. Miyagishi, and K. Taira. Effects on RNAi of the Tight Structure, Sequence and Position of the Targeted Region. *Nucleic Acids Res.*, 32(2):691–9, 2004.

[91] Y. Zeng, R. Yi, and B.R. Cullen. Micrornas and Small Interfering RNAs can Inhibit mRNA Expression by Similar Mechanisms. *Proc. Natl. Acad. Sci. USA.*, 100(17):9779–9784, 2003.

[92] G.K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Reading, MA, 1949.

[93] M. Zuker and D. Sankoff. RNA Secondary Structures and their Prediction. *Bull. Math. Biol.*, 46:591–621, 1984.

[94] M. Zuker and P. Stiegler. Optimal Computer Folding of Large RNA Sequences using thermodynamic and auxiliary information. *Nucl. Acids Res.*, 9:133–148, 1981.

List of Figures

List of Tables

A Appendix

A.1 List of Symbols

Table 9: List of Symbols

symbol	meaning
$\mathcal{A}$	an alphabet e.g. $\mathcal{A} = \{A, U, G, C\}$
κ	the size of a given alphabet
$\mathcal{I}$	sequence (genotype) space
I_k	a particular sequence
$\mathcal{S}$	secondary structure (phenotype) space
S_l	a particular secondary structure
$\mathcal{G}_l$	the set of all sequences folding into structure S_l, that is the neutral set of a structure S_l
B_l	the boundary of structure S_l
$\mathcal{C}_l$	set of structures compatible with S_l
$\sum$	set of all mfe structures of fixed length over a given alphabet
d^h_{kl}	Hamming distance between two sequences
d^S_{ij}	distance between two structures
$F(S)$	free energy of a given secondary structure S
Z	equilibrium partition function

Curriculum vitae

Persönliche Daten

Mag. Ulrike Mückstein
e-mail: ulim@tbi.univie.ac.at
Geb. am 02. 11. 1969 in Wien
Staatsbürgerschaft: österreichisch

Studium

seit 08/2005	Assistentin in Ausbildung bei Prof. Peter Schuster am Institut für Theoretische Biochemie der Universität Wien
seit 11/2001	Doktoratsstudium bei Prof. Peter Schuster am Institut für Theoretische Biochemie der Universität Wien
10/2001	2. Diplomprüfung Abschluß des Studiums der Biologie (Genetik) mit dem Titel Mag. rer. nat.
02/2000–10/2001	Diplomarbeit bei Prof. Peter F. Stadler am Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien: *A Variation on Algorithms for Pairwise Global Alignments*
06/1996	1. Diplomprüfung
03/1990	Beginn des Studium Irregulare der Molekulargenetik
09/1988	Beginn des Studiums der Biologie, Universität Wien

Praxis

03/2002–01/2004	Tutorin für die Übungen zur Strukturbiologie I und II, Prof. Peter Schuster, Universität Wien

Berufserfahrung

02/1999–01/2000	Wissenschaftliche Mitarbeiterin am Institut für Biochemie und Molekulare Zellbiologie, Universität Wien

12/1997–03/1998 Angestellte am Novartis Institute for Bio-Medical Research

Sprachen Deutsch (Muttersprache), Englisch (fließend in Wort und Schrift)

EDV

Betriebssysteme	UNIX (Linux), Windows
Sprachen	C, C++, Perl
Datenbanken	PostgreSQL
Anwendungen	LaTeX, Emacs, Word, Excel

Interessen Taoismus, Lesen, Wandern, Klettern, Tai-Chi

Publikationen

1. Mückstein U., Tafer H., Hackermüller J., Bernhard S., Stadler P.F. and Hofacker I.L. Thermodynamics of RNA-RNA Binding. In Andrew Torda, Stefan Kurtz, and Matthias Rarey, editors, *German Conference on Bioinformatics 2005*, volume P-71, pages 3-13,Bonn,2005, Gesellschaft f. Informatik

2. Mückstein U., Hofacker I.L., and Stadler P.F. Stochastic Pairwise Alignments. Bioinformatics 18: S153-S160, 2002

Konferenz Beiträge

1. Mückstein U. and Grünberger K. Shadows and Intersections of RNA Secondary Structure. Poster presentation at MATH/CHEM/COMP 2004, Dubrovnik, June 21-26, 2004

2. Mückstein U., Topological structure of RNA networks. Poster presentation at MATH/CHEM/COMP 2003, Dubrovnik, June 23-28, 2003

3. Mückstein U., Hofacker I.L., Stadler P.F., Stochastic Pairwise Alignments. Poster presentation at MATH/CHEM/COMP 2002, Dubrovnik, June 24-29, 2002

4. Fekete M., Flamm C., Hofacker I.L., Mückstein U., Rauscher S., Stadler P.F., Stocsits R., Thurner C. and Witwer C. Automatic Detection of Conserved Secondary Structure Elements in Viral Genomes. Poster presentation at the 52. Mosbacher Kolloquium, Mosbach 2001