RNA Structures with Pseudoknots

DIPLOMARBEIT

eingereicht von

Christian Haslinger

zur Erlangung des akademischen Grades

Magister rerum naturalium

an der Formal- und Naturwissenschaftlichen Fakultät

der Universität Wien

December 3, 1997

An dieser Stelle möchte ich allen danken, die mir bei der Entstehung dieser Arbeit geholfen haben.
PETER SCHUSTER danke ich für die Aufnahme in seine Arbeitsgruppe.
PETER STADLER war für mich ein kompetenter und sehr unkomplizierter Betreuer. Allen Kolleginnen und Kollegen vom Institut danke ich für die Hilfsbereitschaft und die angenehme Arbeitsatmosphäre. Meine Eltern ermöglichten mir dieses Studium und unterstützten mich wo sie nur konnten.

Wofür ich meiner Freundin BARBARA danke läßt sich kaum in Worte fassen. Diese würden auch eher Inhalt eines Liebesbriefes sein, der sicher dicker wäre als diese Diplomarbeit.

Abstract

Secondary structures of nucleic acids are a particularly interesting class of contact structures. Many important RNA molecules,however contain pseudoknots, which are excluded explicitly by the definition of secondary structures. We propose here a generalization of secondary structures that incorporates "non-nested" pseudoknots. We also introduce a measure for the complexity of more general contact structures in terms of the chromatic number of their intersection graph. We show that RNA structures without nested pseudoknots form a special class of planar graphs, the so called "bi-secondary structures". Upper bounds on their number are derived, showing that there are fewer different structures than sequences.

An energy function capable of dealing with bi-secondary structures was implemented into a generalized kinetic folding algorithm. Sterical hindrances involved in pseudoknot formation are taken into account with the help of two simplifications: stacked regions are viewed as stiff rods and unpaired bases are assumed to be very flexible. Three parameters are employed in the energy function to consider the sterical situation. The parameter adjustment demands a rather heuristic approach. A variety of experimentally determined bi-secondary structures were used as target structures in a series of folding procedures with different parameters. For short range pseudoknots a suitable parameter set was found.

Different parameter sets were used to study the mapping from sequences to bi-secondary structures. Statistics of bi-secondary structure elements as well as the frequency distribution of bi-secondary structures were computed. The frequency distribution can be described by a generalized form of Zipf's law, which means that there are view common structures and many rare ones. Neutral nets (neighboring sequences sharing the same structure) of bi-secondary structures were compared with those calculated for pure secondary structures and found to be less extended.

Zusammenfassung

Sekundärstrukturen von Nukleinsäuren bilden eine Klasse von Kontaktstrukturen die von großem Interesse ist. Viele wichtige RNA Moleküle beinhalten Pseudoknoten, welche aber durch die Definition von Sekundärstruktur explizit ausgeschlossen sind. Hier schlagen wir eine Verallgemeinerung von Sekundärstrukturen vor, die nicht verschachtelte Pseudoknoten miteinschließt. Die chromatische Zahl des Schnitt-Graphen von verallgemeinerten Kontaktstrukturen wird als Maß für deren Komplexität eingeführt. Wir zeigen, daß RNA Strukturen ohne verschachtelte Pseudoknoten eine spezielle Klasse von planaren Graphen bilden, die sogenannten Bisekundärstrukturen. Es wurden obere Grenzen für deren Anzahl berechnet, die zeigen, daß es weniger verschiedene Strukturen als Sequenzen gibt.

Eine Energiefunktion für Bisekundärstrukturen wurde in einen einfachen kinetischen Faltalgorithmus implementiert. Die sterische Hemmung die bei Pseudoknoten Bildung entsteht, wurde mit Hilfe zweier Vereinfachungen berücksichtigt: Helices werden als steife Stäbchen betrachtet während angenommen wird, daß ungepaarte Basen sehr flexibel sind. Drei Parameter werden in der Energiefunktion benutzt um die räumliche Situation zu berücksichtigen. Die Justierung der Parameter fordert eine mehr oder minder heuristische Vorgangsweise. Eine Reihe von experimentell ermittelten Bisekundärstrukturen wurden als Zielstrukturen in unterschiedlichen Faltexperimenten mit verschiedenen Parametern verwendet. Für Pseudoknoten mit kurzer Reichweite wurde ein geeignetes Parameter-Set gefunden.

Verschiedene Parameter-Sets wurden verwendet um die Abbildung von Sequenzen zu Bisekundärstrukturen zu untersuchen. Die Verteilung der Häufigkeit von Bisekundärstrukturen sowie eine Statistik der Strukturelemente wurde berechnet. Betrachtet man die Häufigkeit von Bisekundärstrukturen, findet man eine sehr ungleichmäßige Verteilung die einige wenige sehr häufige dafür viele extrem seltene Strukturen aufweist. Die Verteilung kann gut durch eine allgemeine Form des Zipf'schen Gesetzes beschrieben werden. Auch die Statistik der Strukturelemente zeigt im wesentlichen das selbe Bild wie für

Sekundärstrukturen. Neutrale Netze (benachbarte Sequenzen die die selbe Struktur bilden) von Bisekundärstrukturen die mit jenen von Sekundärstrukturen verglichen wurden, wiesen eine geringere Ausdehnung auf.

Contents

List of Figures

List of Tables

1 Introduction

Presumably one of the most important problems and greatest challenges in present day theoretical biochemistry is understanding the mechanism that transforms sequences of biopolymers into spatial molecular structures. One the one hand the sequence is the most easily accessible molecular biological information, on the other hand the spatial structure is the key to describe the molecule in its functional aspects.
A sequence is properly visualized as a string of symbols which together with the environment determines the molecular architecture of the biopolymer. In case of one particular class of biopolymers, the ribonucleic acid (RNA) molecules, decoding of information (stored in the sequence) can be properly decomposed into two steps.

- The formation of the secondary structure, i.e. the pattern of Watson-Crick (and GU) basepairs.
- The embedding of the contact structures in three-dimensional space (such as pseudoknots).

The classical definition of secondary structures excludes pseudoknots mostly for technical reasons [100] (the folding problem for RNA can be solved efficiently by dynamic programming in the absence of pseudoknots [100, 108]). The sequence structure relation of RNA was studied in detail in a series of papers at the level of secondary structures.[4, 18, 20, 19, 25, 26, 76, 85, 86] The most salient findings of these investigations are:

(i) There are many more sequences than (secondary) structures.

(ii) There are few frequent and many rare structures. Almost all sequences fold into frequent or common structures.

(iii) Sequences that fold into a common structure are distributed nearly uniformly in sequence space.

(iv) A sequence folding into a common structure has a large number of neutral neighbors (folding into the same structure) and a large number of neighboring sequences that fold into very different secondary structures.

(v) Neutral paths perlocate sequence space along which all sequences fold into the same secondary structure. In fact there are extended neutral networks of sequences folding into the same common structure.[26, 65]

(vi) Almost all common structures can be found close to any point in sequence space. This property is called shape space covering.

The impact of these features on evolutionary dynamic is discussed in [37, 75]. A population explores sequence space in a diffusion-like manner along the neutral network of a viable structure. Along the fringes of the population novel structures are produced by mutation at a constant rate [36]. Fast diffusion together with perpetual innovation makes these landscapes ideal for evolutionary adaptation. An increasing number of experimental findings, as well as results from comparative sequence analysis suggest, that pseudoknots are important structural and functional elements in many RNA molecules.

1.1 The Purpose of this Work

It is an important question, whether the findings (i) through (vi) remain true when pseudoknots are taken into account. Assertion (i), the existence of more sequences than structures, is a necessary prerequisite for all subsequent statements concerning the sequence-structure map of RNA. It is necessary therefore to estimate the number of RNA structures *with pseudoknots* in order to decide whether the results quoted above can in fact be true for "real" RNA molecules. Combinatorial aspects of RNA secondary structures have been studied in detail by Waterman and coworkers [56, 74, 82, 97, 99, 100, 98] and Hofacker *et al.* [32]. In section 3 we discuss generalizations of secondary structures and their graph-theoretical properties. The combinatorics of these objects is then studied by both analytical and numerical methods in section 6. In order to investigate if the findings (ii) through (v) apply we need an RNA-structure prediction method. That means a suitable energy function has to be designed and implemented into a folding algorithm. We describe in section 4 an energy function that deals with RNA-secondary structures comprising a wide variety of pseudoknots, namely the bi-secondary structure (introduced in section 3). The energy function is tested in section 5 using two interesting molecules which biological aspects are described in section 2.

2 Functional Aspects of RNA Pseudoknots

2.1 Why Does One Care about Pseudoknots ?

Recent work has indicated that pseudoknots are only marginally more stable than simple secondary structures (although thermodynamic data in this area are still scarce [61, 49]). This observation suggests a role for pseudoknots as conformational switches or control elements in several biological functions [73]. In molecules that lack an overall three-dimensional fold, pseudoknots fold locally and their positions along the sequence reflect their function [48]. For example, pseudoknots that are folded at the 5'-end of mRNAs tend to be involved in translational control whereas those at the 3'-end maintain signals for replication. In molecules with catalytic activities, pseudoknots are located at the core of the tertiary fold and involve nucleotides that are far apart in the sequence (RNAseP). The diversity of molecular biological functions performed by pseudoknots can be subdivided into tree different groups:

(i) **Translational control:** 5'-end pseudoknots appear to adopt two roles in the control of mRNA translation: either specific recognition of a pseudoknot by some protein is required for control, as described for the 5'-end of mRNAs in some prokaryotic systems [73, 58]; or, the presence of a folded pseudoknot is necessary with no requirements on the nucleotide sequence [5, 92, 8]. In several viruses, the expression of replicase is controlled either by *ribosomal frame shifting* [5, 92, 8, 13, 15] or by *in-frame read-through* of stop codons [104]. In both cases, pseudoknot formation is necessary [5, 92, 15]. The requirements appear, however, more strict for read-through than for frame shifting. Nevertheless, the correct position of the pseudoknot in the 3' direction with respect to the slip site in ribosomal frame shifting, and with respect to the AUG codon in read-through is an absolute requirement [5, 104]. The presence of three pseudoknots in 16S rRNA has been suggested on the basis of comparative sequence analyses [59]. In general these pseudoknots are assumed to show strong interactions with ribosomal proteins. One pseudoknot is known to be important for the binding of tRNA to the ribosomal A site [106, 52], and was shown to be essential for ribosomal function [60]. These observations are particularly interesting in view of the suggested conformational switch that involves the other two pseudoknots.

(ii) **Core pseudoknots:** are necessary to form the reaction center of ribozymes. Most of the enzymatic RNAs with core pseudoknots are involved in cleavage or self-cleavage reactions [51, 21, 6, 29]. One Example (RNAseP) is discussed in this section.

(iii) **3'-end pseudoknots:** replication control is the common function of tRNA-like motifs at the 3'-end of several groups of plant viral RNA genomes [48]. This structural similarity is paralleled in biological function as the tRNA-like motifs are recognized by many tRNA-specific enzymes such as aminoacyl-tRNA synthetases, nucleotidyl transferase, or RNAseP [48]. The tRNA-like structure has been shown to be necessary for the initiation of replication [48]. A telomeric function of the tRNA-like structure was also demonstrated [62] , in agreement with the genomic tag model associated with such 3'-terminal tRNA-like motifs [102]. Recently, the strech of three pseudoknots preceding the tRNA-like structure in tobacco mosaic virus was shown to act as the functional equivalent of a poly(A) tail, stabilizing a reporter mRNA and increasing gene expression up to 100-fold [23].

Throughout our work two rather complex molecules were picked out to demonstrate their pseudoknot folding behavior and the problems associated with structure prediction (section 5). One example chosen for short distance pseudoknots (almost exclusively H-type) is *tmRNA*, a molecule with interesting molecular biological features. Another example for long distance tertiary interactions is *RNAseP RNA*. Here we give a brief characterization of these species:

2.2 tmRNA

2.2.1 The Biological Relevance of tmRNA

Because of the action of nucleases, mRNAs may be truncated and lose their encoded stop codons at their 3' ends. Translation of those messages can still proceed but comes to a halt when the ribosome reaches the 3' end of the mRNA. In the absence of a stop codon, release factor cannot trigger the dissociation of nascent polypeptides from ribosomes, and all ribosomes engaged in translation of the same reading frame are stalled. Alanine-charged tmRNA may come to the rescue. It combines both transfer and messenger RNA properties. The tRNA-like domain enables the ribosome to catalyze the next peptidyl transfer of the nascent chain to the tmRNA-bound alanine. The defective mRNA can then be released, and the ribosome switches to the reading frame provided by the tmRNA. This process, message switching, might resemble frame shifting or ribosomal hopping, with the significant difference that it works in trans, as the new reading frame is provided by a second molecule. Charged tRNAs complementary to the codons of tmRNA are then used for translating the encoded tag. The tagged polypeptide can finally be released as the ribosome reaches the in-frame stop codon of the tmRNA. Tail specific proteases are now able to destroy the defective translation product.

2.2.2 Secondary Structure of E.Coli tmRNA

Experimental results: To determine the secondary structure several methods were used which complement each other. The structure of tmRNA was proposed on the basis of covariation of homologous sequences [10] and on its reactivity in solution toward enzymatic and chemical probes (RNAses, imidazol probing and lead-induced hydrolysis [17, 93]). The probing data allows discrimination between the single- and double-stranded regions within tmRNA. Because tmRNA is a large molecule (363nt), many secondary structure models are possible even if the probing data considerably restrict the number of realistic solutions. Although there is no evidence it may be that tmRNA sequences are posttranscriptionally modified. Structural elements H1, H3, H4, H5, H6 pK2.1, and both stems of PK1, PK3, and PK4 are supported by the data collected. Elements H2 and pK2.2 are questionable by probing, but supported robustly by covariations. Element R1 is drawn as a pseudoknot, but at least one alternative form could fit the data equally well. Many of these structural domains are connected by single-stranded links of variable length. However, in some cases, there is no connecting nucleotides. It seems plausible that tm-RNA might undergo conformational change during the transition from tRNA to mRNA. If so, then covariation of nucleotides among the various sequences might reflect either of the two conformations, because it only represents a functionally conserved base pair. Probing experiments as performed, only access the tmRNA as it exists in solution, and not a molecule that might be induced to a different form by interaction with a ribosome, for instance.

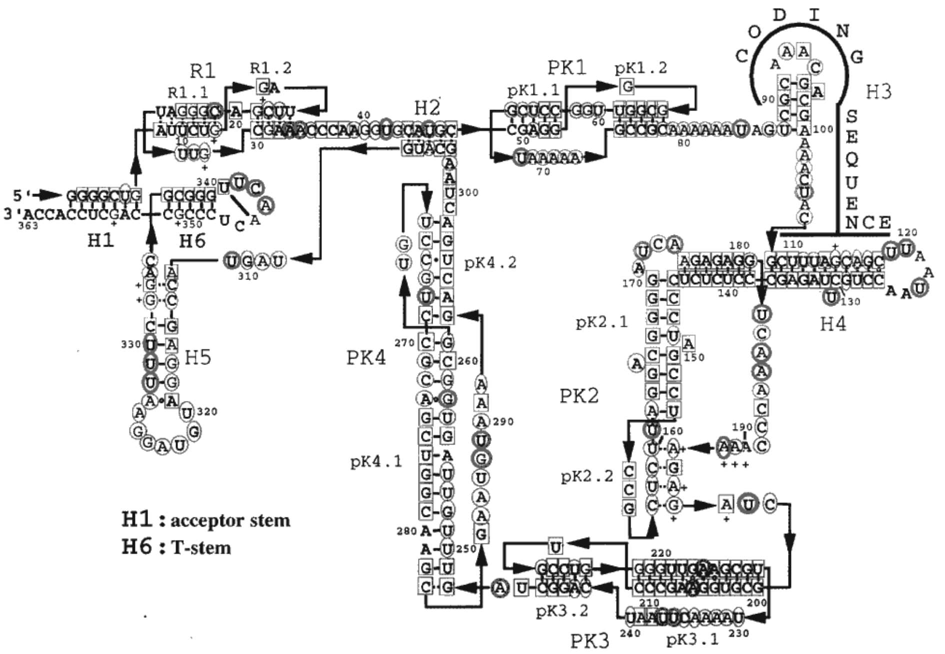


Figure 1: The secondary structure of tmRNA

2.3 RNAseP

2.3.1 The Biological Relevance of RNAseP

Ribonuclease P (RNase P) is a key enzyme in the biosynthesis of tRNA [2, 14, 69]. It is an RNA processing endonuclease that specifically cleaves precursors of tRNA, releasing 5' precursor sequences and thereby forming the mature 5' ends of the tRNAs. RNase P is involved in processing all species of tRNA and is present in all cells and organelles that carry out tRNA synthesis. It is a particularly interesting enzyme because of its composition: it is a ribonucleoprotein [41]. In Bacteria the RNase P holoenzyme is composed of a large RNA (usually 350-400 nucleotides) and a single molecule of a small protein (ca.120 amino acids in known instances). The bacterial RNase P RNA is clearly the catalyst in the reaction. In contrast, archaeal and eucaryal RNase P RNA subunits have not yet been found to exhibit catalytic activity after the removal proteins. With the possible exception of the ribosome [54], RNase P is the only known example of an RNA that in vivo truly acts as an enzyme, in the sense that it reacts with multiple substrates. Other known catalytic RNAs, for instance self-cleaving introns or satellite RNAs, naturally perform only a single intramolecular reaction [84].

2.3.2 Secondary Structure of E. coli RNAseP RNA

Experimental results: The determination of the secondary structure of the bacterial RNase P RNA was a challenge because of the substantial sequence and length variation in the molecule from diverse organisms (figure 2).

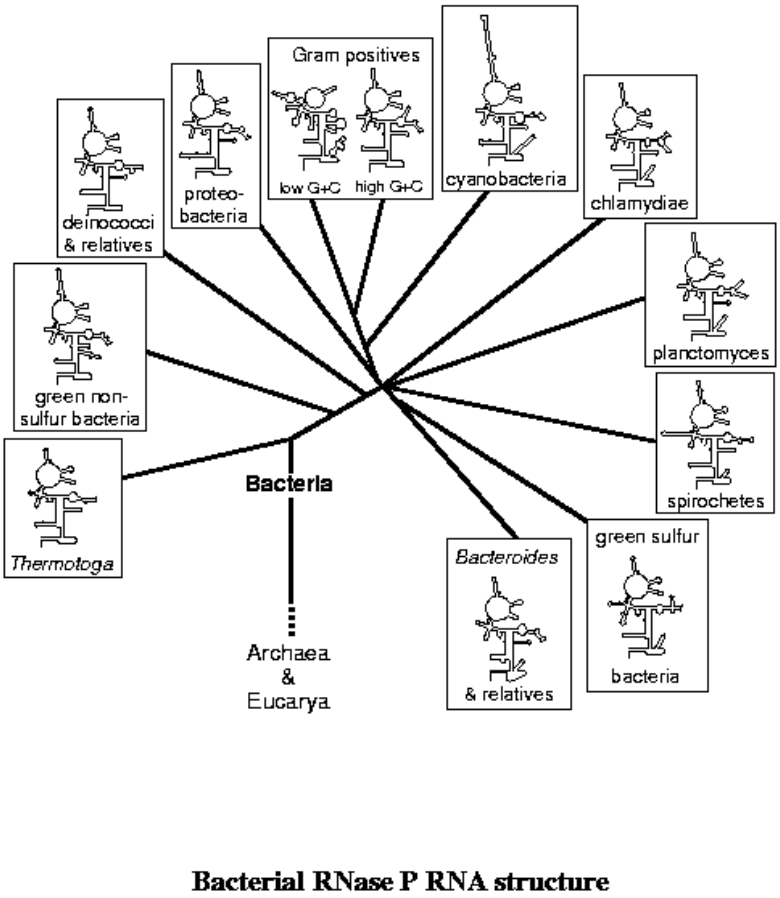


Figure 2: Phylogenetic tree of bacterial RNaseP RNA

The first RNase P RNA sequences determined, from E. coli [63]and B. subtilis [64], differ so much from one another that it was initially impossible to align the sequences over most of their lengths (figure 3). The alignment process, ordering homologous nucleotides, usually in columns in a sequence set, is an essential step in comparative analysis: covariation of bases is meaningful only if the covarying bases are homologs. The alignment of these sequences was resolved by the determination and comparison of additional sequences, from organisms of intermediate evolutionary relatedness to E. coli and B. subtilis. Phylogenetic trees, such as those based on ribosomal RNA (rRNA) sequence comparisons, are important resources for the selection of such source-organisms for comparative sequence studies.

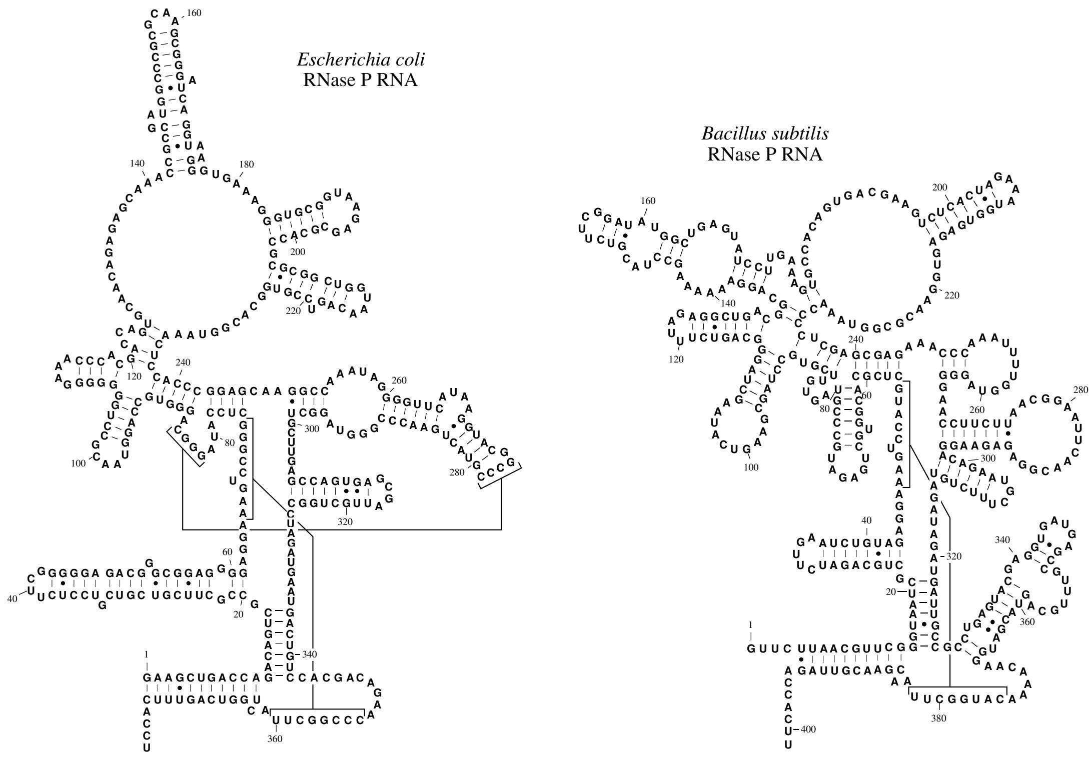


Figure 3: Comparison of RNAseP RNA secondary structure of Escherichia coli and Bacillus subtilis

3 Contact Structures and Diagrams

3.1 Secondary Structures

The three-dimensional structure of a linear biopolymer, such as RNA, DNA, or a protein can be approximated by its *contact structure*, i.e., by the list of all pairs of monomers that are spatial neighbors. A contact structure is represented by the *contact matrix* $\mathbf{C}$ with the entries $\mathbf{C}_{ij} = 1$ if the monomers i and j are spatial neighbors without being adjacent along the backbone, and $\mathbf{C}_{ij} = 0$ otherwise. Hence $\mathbf{C}_{ij} = 0$ if $|i - j| \leq 1$. We shall use the notation $[n] := \{1, \ldots, n\}$.

Definition. A *diagram* $([n], \Omega)$ consists of n vertices labeled 1 to n and a set Ω of *arcs* that connect non-consecutive vertices. A closely related class of diagrams which allow also arcs between consecutive vertices are the *linked diagrams* introduced by Touchard [90]. These are studied in some detail in the references [35, 40, 80, 81]. It is customary to arrange the vertices along the x-axis and to draw the vertices in such a way that they are confined in either the upper or the lower half-plane. The diagram of a contact structure with contact matrix $\mathbf{C}$ has the set of arcs

$$\Omega := \left\{ \{i, j\} \,\middle|\, \mathbf{C}_{ij} = 1 \right\}. \tag{1}$$

The contact matrix is thus the adjacency matrix of the corresponding diagram.

Definition. A *diagram graph* is a simple vertex labeled graph Γ with the following properties:

(i) The $n + 1$ vertices of Γ are labeled $0, 1, \ldots, n$.

(ii) Γ contains the Hamiltonian cycle $[0, 1, \ldots, n, 0]$.

(iii) The "root" vertex 0 has degree 2.

Lemma 1. There is an isomorphism ι between the diagrams with n vertices and the diagram graphs with $n+1$ vertices.

Proof. Let $\mathbf{B}$ be the matrix with the entries $\mathbf{B}_{i,i+1} = \mathbf{B}_{i+1,i} = 1$, $i = 0, \ldots, n-1$, and $\mathbf{B}_{0n} = \mathbf{B}_{n0} = 1$. Furthermore let $\mathbf{C}$ be a symmetric $n \times n$ matrix with entries 0 or 1 and $\mathbf{C}_{ij} = 0$ whenever $|i-j| \leq 1$. In other words $\mathbf{C}$ is the contact matrix of a diagram on n vertices. The adjacency matrix of a diagram graph with $n+1$ vertices is of the form

$$\mathbf{A} = \mathbf{B} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{C} \end{pmatrix} \tag{2}$$

since the "root" vertex 0 has only the neighbors 1 and n. Since $\mathbf{B}$ is already defined by n, this decomposition is unique and $\mathbf{A}$ can be constructed if $\mathbf{C}$ is known and *vice versa.*

Essentially the same construction can be used for contact structures of molecules with a circular backbone, i.e., for circular ssRNA or ssDNA. The only restriction is that $\{1, n\}$ cannot be an arc in the case of a circular molecule. It is convenient in this case to define the corresponding diagram graph without the artificial root 0. Each graph Γ with a Hamiltonian cycle is then the diagram graph of a contact structure with a circular backbone. The results in the following discussion hold for both linear and circular nucleic acids.

Definition. A diagram is called an 1-diagram if for any two arcs $\alpha, \beta \in \Omega$ holds $\alpha \cap \beta = \emptyset$ or $\alpha = \beta$.

Lemma 2. A diagram Δ is a 1-diagram if and only if the vertices of the diagram graph $\iota(\Delta)$ have vertex degree less or equal to 3.

Proof. By definition no vertex of the diagram is contained in more than a single arc. Adding the two edges along the Hamiltonian cycle $\mathcal{H}$, we find the $\deg(v) \leq 3$ for all vertices of the diagram graph. The diagram graphs of 1-diagrams are closely related to cubic Hamiltonian graphs. The latter are studied in detail in section 9.4 of reference [94].

Definition. Let $\alpha = \{i, j\}$ with $i < j$ be an arc of a diagram. We write

$\bar{\alpha} :=[i, j]$ for the associated interval.

Definition. Two arcs of a diagram are *consistent* if they can be drawn in the plane without intersection in the same half-plane. An obvious algebraic characterization is

Lemma 3. Two arcs $\alpha, \beta \in \Omega$ of a diagram are consistent if either one of the following four conditions is satisfied.

(i) $\bar{\alpha} \cap \bar{\beta} = \emptyset$.

(ii) $\bar{\alpha} \subseteq \bar{\beta}$.

(iii) $\bar{\beta} \subseteq \bar{\alpha}$.

(iv) $\bar{\alpha} \cap \bar{\beta} = \{k\}$, a single vertex.

Case (iv) is ruled out by definition in 1-diagrams. 1-diagrams can be used to model the base pairing interactions in nucleic acids. Indeed, the classical definition of a secondary structure [97] requires that each base pairs with at most one other nucleotide. The second defining condition is the absence of *pseudoknots* which can be expressed in terms of the contact matrix in the following from: If $\mathbf{C}_{ij} = \mathbf{C}_{kl} = 1$ and $i < k < j$ then $i < l < j$. That is, if the intervals of two arcs $\{i, j\}$ and $\{k, l\}$ have non-empty intersection then one is contained in the other [74]. Using the above terminology we have the following

Definition. A secondary structure is a 1-diagram in which any two arcs are consistent. A graph that can be embedded in the plane such that all its vertices lie on the exterior region is called *outerplanar*. This class of graphs was introduced and characterized in terms of subgraphs in ref. [9]. Another interesting characterization in terms of a spectral invariant is discussed in [12].

Lemma 4. A 1-diagram Δ is a secondary structure if and only if its diagram graph $\iota(\Delta)$ is outerplanar.

Proof. A secondary structure is a 1-diagram Δ in which all arcs can be drawn without intersection in the same half-plane. Equivalently, all arcs can be drawn inside the Hamiltonian cycle $\mathcal{H}$ in $\iota(\Delta)$. Each secondary structure can be encoded as a string s of length n in the following way: If the vertex i is unpaired, then $s_i =$ '`.`'. Each arc $\alpha = \{p, q\}$ with $p < q$ translates to $s_p =$ '`(`' and $s_q =$ '`)`'. Since the arcs are consistent their corresponding parentheses are either nested, `(( ))`, or next to each other, `()()`. As there are no arcs between neighboring vertices in a 1-diagram there is at least one dot contained within each parenthesis. A variant of this notation is the *mountain representation* of RNA secondary structures [33]. The "dot-parenthesis" notation is used as a convenient notation in input and output of the `Vienna RNA Package`, a piece of public domain software for folding and comparing RNA molecules [31].

The close resemblance of cubic Hamiltonian graphs [94] and diagram graphs of 1-diagrams suggests to investigate their relation in some more detail. A graph S is *homeomorphic from* a graph Γ if S can be produced from Γ by inserting vertices of degree 2 into some edges of Γ. S is also called a *subdivision* of Γ. Obviously each cubic Hamiltonian graph gives rise to a diagram graph on n vertices by subdividing the edges of a Hamiltonian cycle. On the other hand, not all diagram graphs are homeomorphic from a cubic Hamiltonian graph: Suppose $\{1, 3\}$ is an arc and 2 is an unpaired vertex. The corresponding diagram graph cannot be cubic since the triangle $1, 2, 3$ cannot be obtained from a cubic graph.

Definition. An arc α is an *undisturbed hairpin* if

(i) either there is no $\beta \in \Omega$ for which $\bar{\beta} \subset \bar{\alpha}$ is true, or $\bar{\beta} \subset \bar{\alpha}$ for all $\beta \in \Omega$.

(ii) α is consistent with all $\beta \in \Omega$.

Lemma 5. A diagram graph $\iota(\Delta)$ is homeomorphic from a cubic Hamiltonian graph Γ if and only if the set of arcs Ω is non-empty and Δ does not contain an undisturbed hairpin.

Proof. Consider a subset C of vertices such the induced subgraph in $\iota(\Delta)$ is a cycle C in $\iota(\Delta)$. It is easy to see that C can be contracted to a cycle that contains no vertices of degree 2 (in the original graph) if and only if it contains at least 3 vertices of degree larger than 2. These contractions can be performed independently in all such induced subgraphs, thus $\iota(\Delta)$ is homeomorphic from a cubic Hamiltonian graph if and only if each minimal cycle contains at least three vertices with degree 3. Observing that undisturbed hairpins are exactly the minimal cycles that contain two vertices of degree 3 completes the proof.

Corollary. A diagram graph of a secondary structure is not homeomorphic from a cubic Hamiltonian graph.

3.2 The Inconsistency Graph of a Diagram

Definition. Let $\Delta = ([n], \Omega)$ be a diagram. The *inconsistency graph* $\Theta(\Delta)$ of the diagram has vertex set Ω and $\{\alpha, \beta\}$ is an edge of $\Theta(\Delta)$ if and only if the arcs α and β are inconsistent in Δ. Essentially the same construction is used for the investigation of cubic Hamiltonian graphs in [94], where a result analogous to the following theorem is proved:

Theorem 1. Let Δ be a diagram. Then the following statements are equivalent.

(i) The diagram Δ can be drawn without intersecting arcs.

(ii) The diagram graph $\iota(\Delta)$ is planar.

(iii) The inconsistency graph $\Theta(\Delta)$ is bipartite.

Proof. (i $\iff$ ii) Δ can be drawn without intersection arcs if and only if $\iota(\Delta)$ is planar because the Hamiltonian cycle $\mathcal{H}$ of $\iota(\Delta)$ divides the plane into the interior and the exterior of $\mathcal{H}$ which correspond to the upper and lower half-plane of the diagram Δ, respectively.

(i $\Longleftrightarrow$ iii) If Θ is bipartite then there are two disjoint subsets Ω_U and Ω_L of Ω such that all arcs within the same subset are mutually consistent. Thus all arcs of Δ can be drawn without intersection. Conversely, if we can draw Δ without intersecting arcs then all arcs above and below the x-axis are mutually consistent, i.e., two arcs can be inconsistent only if they lie on different sides of the x-axis. Thus $\Theta(\Delta)$ is bipartite. Most of the literature on linked diagrams deals with *complete* diagrams, that is, each vertex $x \in [n]$ is incident with an arc [90, 40, 80]. It is straightforward to extend Touchard's definition of reducible diagrams to the incomplete diagrams considered here:

Definition. A diagram $([n], \Omega)$ is *reducible* if there is an interval $[p, q] \subset [n]$ such that

(i) For each $\alpha \in \Omega$ holds either $\alpha \cap [p, q] = \emptyset$ or $\alpha \subseteq [p, q]$.

(ii) There is an arc $\alpha \in \Omega$ such that $\alpha \cap [p, q] = \emptyset$.

(iii) There is an arc $\alpha \in \Omega$ such that $\alpha \subseteq [p, q]$.

If a diagram is not reducible, it is *irreducible*.

It will be convenient to say that an interval $[r, s]$ *supports* an arc α if $\bar{\alpha} \subseteq [r, s]$. Let $\Omega_{[r,s]}$ be the set of arcs supported by $[r, s]$. We shall say that $([r, s], \Omega_{[r,s]})$ is a sub-diagram of $([n], \Omega)$ if $[r, s]$ fulfills (i). The sub-diagram is non-trivial if $\Omega_{[r,s]}$ is neither empty nor equals Ω. A diagram is therefore reducible if and only if it contains a non-trivial sub-diagram. Let $([p, q], \Omega_{[p,q]})$ and $([r, s], \Omega_{[r,s]})$ be two sub-diagrams of Δ. Then $[p, q] \subset [r, s]$ implies $\Omega_{[p,q]} \subset \Omega_{[r,s]}$ and $([p, q] \cap [r, s], \Omega_{[p,q]} \cap \Omega_{[p,q]})$ is again a sub-diagram of Δ. The sub-diagrams of Δ therefore form a lattice with respect to inclusion.

Lemma 6. A diagram Δ is irreducible if and only if its inconsistency graph $\Theta(\Delta)$ is connected.

Proof. (i) Suppose $([n], \Omega)$ is reducible, and let $([p, q], \Omega_{[p,q]})$ be a non-trivial sub-diagram. Thus there are no arcs that are incident with vertices both in

$[p,q]$ and $[n] \setminus [p,q]$ and an arc α supported by $[p,q]$ is consistent with any arc β not supported by $[p,q]$. For all $\beta \in \Omega \setminus \Omega_{[p,q]}$ it is true therefore that β is not connected with any $\alpha \in \Omega_{[p,q]}$ in $\Theta(\Delta)$ and $\Theta(\Delta)$ decomposes into at least two non-empty components.
Now assume that $\Theta(\Delta)$ is not connected. The *support* of a component Θ' of $\Theta(\Delta)$ in $[n]$ is the union of all intervals $[r,s]$ where r and s are incident with arcs in Θ' and all vertices in $[r+1, s-1]$ are unpaired. Either the support of Θ' is connected, i.e., in which case it forms an interval fulfilling the conditions (i), (ii), and (iii), and Δ is a reducible diagram, or it contains a "hole" $[u,v]$ that contains a vertex x incident with an arc $\gamma \notin \Theta'$. Since γ is consistent with all arcs of Θ', it cannot be incident to any vertex outside $[u,v]$ (otherwise it would need to cross at least one arc of Θ'). Thus $[u,v]$ fulfills conditions (i), (ii), and (iii), and Δ is reducible.

The proof of lemma 6 implies an even stronger result: A sub-diagram corresponds to one or more components of the inconsistency graph. Reducible diagrams can therefore be viewed as being composed of substructures. These substructures do not conform the conventional decomposition into stems and loops, however, which form the basis of the standard energy model of nucleic acid secondary structures [22]. The notion of a stem trivially generalized to arbitrary 1-diagrams:

Definition. Two arcs $\alpha = \{i,j\}$ and β are *stacked* if $\beta = \{i-1, j+1\}$ or $\beta = \{i+1, j-1\}$. A *stem* is a subset Ψ of arcs α_0 through α_h such that α_p and α_{p+1} are stacked for $p = 0, \ldots, h-1$.

Lemma 7. Let Ψ be a stem in the 1-diagram Δ. Then the arcs of Ψ are either all isolated vertices or they are contained in the same component of the inconsistency graph $\Theta(\Delta)$; all arcs of a stem have the same adjacent vertices in $\Theta(\Delta)$.

Proof. It suffices to show that an arc β that is inconsistent with one arc $\alpha \in \Psi$ must be inconsistent with all arcs of the same stem. We observe that the arcs of

Ψ have the form $\alpha_p = \{i+p, j-p\}$ with $p = 0, \ldots, h$. Suppose β is inconsistent with α_p for some p. Then it involves a vertex k with $i + p < k < j - p$ and a vertex l with either $l < i + p$ or $l > j - p$. Since there is at most one arc attached to each vertex, β cannot involve the vertices between $i + p$ and $i + h$ or the vertices between $j - h$ and $j - p$ since they are already used by the arcs of the stem Ψ; hence we have $i + h < k < j - h$. A similar argument shows that l satisfies either $l < i$ or $l > j$, and β is therefore inconsistent with all arcs $\alpha_p \in \Psi$. Hence β is adjacent in $\Theta(\Delta)$ to all $\alpha_p \in \Psi$ and thus they belong to the same component of the inconsistency graph. If all arcs in Ω are consistent with (the arcs of) the stem Ψ they appear as isolated vertices in $\Theta(\Delta)$.

3.3 Bi-Secondary Structures

Definition. A *bi-secondary structure* is a 1-diagram that can be drawn without intersections of arcs. We may draw the arcs in the upper or lower half-plane, but they are not allowed to intersect the x-axis. Thus $\Omega = \Omega_U \dot{\cup} \Omega_L$ and the two diagrams $([n], \Omega_U)$ and $([n], \Omega_L)$ are secondary structures. Bi-secondary structures are therefore “superpositions” of two secondary structures.

Theorem 2. Let Δ be a 1-diagram. Then the following statements are equivalent:

(i) Δ is a bi-secondary structure.

(ii) $\iota(\Delta)$ is planar.

(iii) $\Theta(\Delta)$ is bipartite.

(iv) Among any three arcs of Δ at least two are consistent.

(v) $\Theta(\Delta)$ does not contain a triangle.

Proof. The equivalence of (i), (ii), (iii) is established in Theorem 1 for all diagrams. The equivalence of (iv) and (v) follows immediately from the definition of $\Theta(\Delta)$. The implication (iii$\Longrightarrow$v) is obvious. It remains to show that (iv) or (v) indeed implies (i), (ii), or (iii). We shall prove that $\neg$(ii) implies $\neg$(iv). Suppose $\iota(\Delta)$ is not planar. Then by Kuratowski's theorem implies that it contains a subgraph S that is homeomorphic from either the complete graph K_5 or the complete bipartite graph $K_{3,3}$ [42]. The vertex degree in $\iota(\Delta)$ is at most 3. Therefore it cannot contain a subdivision S of K_5 which would have has five vertices with vertex degree 4. Any subdivision S of $K_{3,3}$ contains exactly six vertices $1, 2, 3, 4, 5, 6$ of degree 3. Of the three edges incident with these vertices two must belong to the Hamiltonian cycle $\mathcal{H}$ because there is at most one arc of Δ attached to each vertex. Thus S contains a cycle $\mathcal{H}'$, containing the vertices 1 through 6, that corresponds to $\mathcal{H}$ in the sense that all edges of $\mathcal{H}$ that have been inherited by S belong to $\mathcal{H}'$. A remaining edge at 1 through 6 corresponds to an arc of Δ, hence it is directly connected to another vertex of degree 3. Thus $\mathcal{H}'$ is a Hamiltonian arc of S. Without loosing generality we may assume that the vertices 1 through 6 are ordered along $\mathcal{H}'$ and hence also along $\mathcal{H}$. Since S is a subdivision of $K_{3,3}$ the three remaining edges (i.e., arcs of Δ) must be $\{1, 4\}$, $\{2, 5\}$, and $\{3, 6\}$. Obviously, they are mutually inconsistent. Non-planarity of $\iota(\Delta)$ implies therefore the existence of three arcs that are mutually inconsistent.

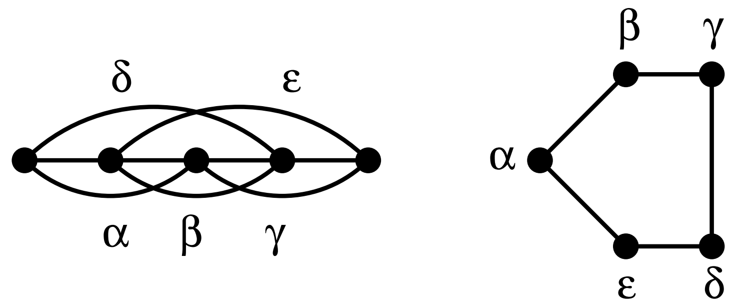


Figure 4: Theorem 2 is not valid for general diagrams. The inconsistency graph of the diagram Δ_5 is a pentagon and hence is neither bipartite nor does it contain a triangle

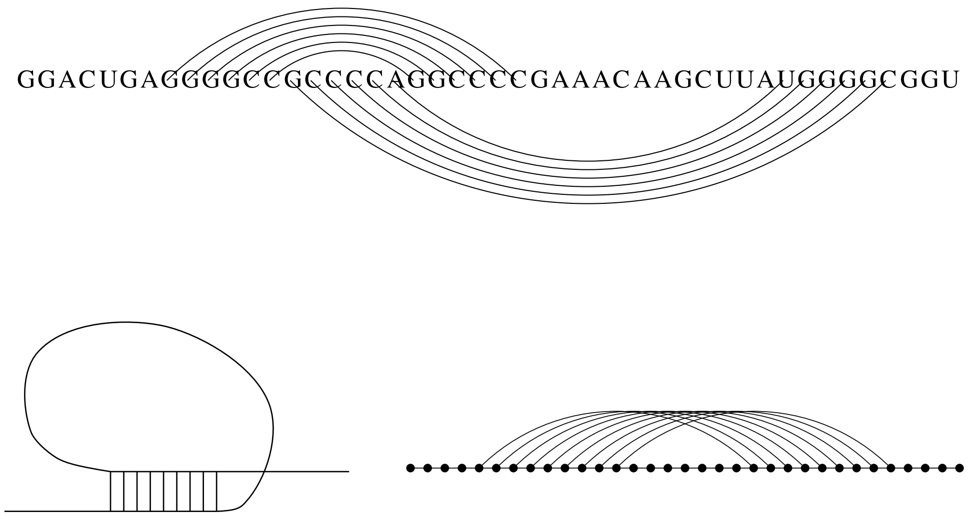


Figure 5: The contact structure of the proposed SRV-1 frameshift signal contains a pseudo-knot, see reference [89]. Pseudoknots such as this one belong to the class of bi-secondary structures.
Knots such as the one in the lower part of the figure do not belong to the class of bi-secondary structures.

The equivalence of (iii) and (v) does not hold for general diagrams. A counterexample is shown in figure 4. The virtue of bi-secondary structures is that they capture a wide variety of RNA pseudoknots, figure 5 (upper part), while at the same time they exclude true knots such as the structure in the lower part of figure 5.

Being the union of the two secondary structures $([n], \Omega_U)$ and $([n], \Omega_L)$ we can represent each bi-secondary structure as a string s using two types of parentheses: As in a secondary structure we write a dot '`.`' for all unpaired vertices. A pair $\{p, q\} \in \Omega_U$ becomes $s_p =$ '`(`' and $s_q =$ '`)`', while an arc $\{p, q\} \in \Omega_L$ becomes $s_p =$ '`[`' and $s_q =$ '`]`'.

The fact that $\Theta(\Delta)$ is bipartite allows us to define a *normal form* for this representation by means of the following rule:

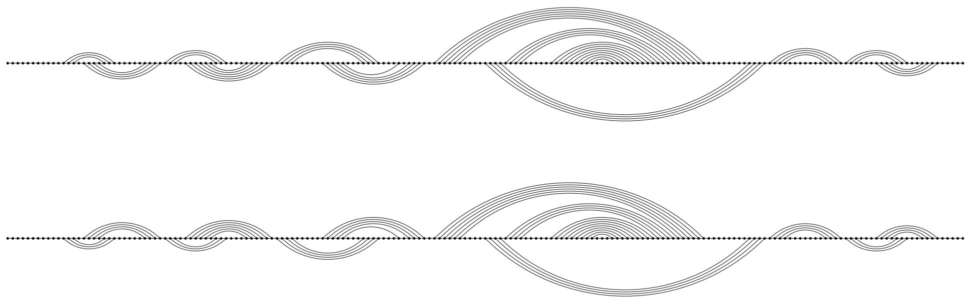


Figure 6: Two diagrams encoding the 3' non-coding region of tobacco mosaic virus RNA [1]. The upper diagram corresponds to the normal form, the lower diagram maximizes the number of upper arcs.

The leftmost arc of each connected component of $\Theta(\Delta)$ belongs to Ω_U. In particular, all isolated vertices of $\Theta(\Delta)$ are contained in Ω_U. The normal form of a secondary structure therefore contains only dots and (round) parentheses. Within each non-trivial connected component of $\Theta(\Delta)$ the distribution of arcs between Ω_U and Ω_L is unique since the component is bipartite. Lemma 7 implies that all arcs in a stack are written with the same type of brackets in normal form because they have a common neighboring vertex and hence they all belong to the same class of the partition.

Remark. If we compare the normal form of a molecule with any other possible representation it seems at the first sight as if two different molecules are given. For example in figure 7 the first picture (the normalform) shows stack 4 as pseudoknot whereas in the second picture stack 2 is the pseudoknot. In the first case the pseudoknot connects two interior loops in the second case a multiloop and a hairpin. But of course, if we use the bi-secondary structure as the basis of an energy function (section 4), the result has to be the same energy regardless of the distribution of arcs between the two half-plains Ω_U and Ω_L. So if we apply a folding algorithm which deals with bi-secondary structures, to a sequences with a known experimentally determined structure (section 5), we have to keep the representation in mind if we compare the results. However experimentally determined structures are not necessarily published in normalform. Therefore, if we compare two structures in `().[]` representation it is convenient to convert them into the normalform.

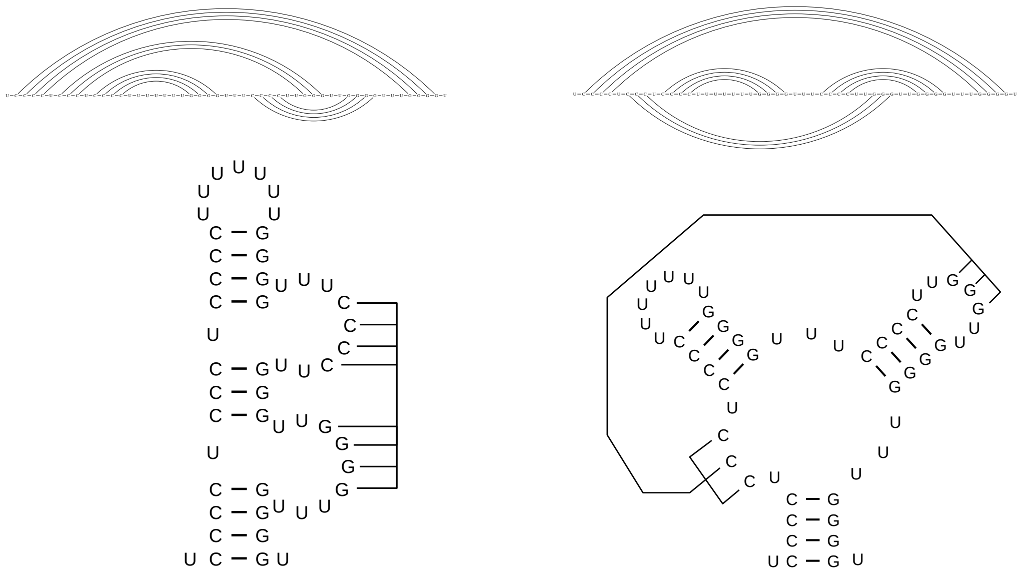


Figure 7: Different Representations of a Bi-Secondary Structure

3.4 Beyond Bi-Secondary Structures

A *color partition* of a graph Γ is partition $V = V_1 \cup V_2 \cup \ldots \cup V_c$ of its vertex set into c subsets V_i such that no two vertices in V_i are adjacent. The *chromatic number* $\chi(\Gamma)$ is the smallest number c of colors for which a color partition of Γ can be found. An arbitrary diagram Δ can be decomposed into substructures by means of the following obvious result:

Lemma 8. Let $\Delta = ([n], \Omega)$ be a diagram and let $\mathcal{V} : \Omega = \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_c$ be a partition of the set of arcs. Then the sub-diagram $([n], \Omega_i)$,

$i = 1, \ldots, c$, can be drawn without intersection if and only if $\mathcal{V}$ is a color partition of the inconsistency graph $\Theta(\Delta)$. Noticing that $\chi(\Gamma) = 1$ if Γ contains no edges and $\chi(\Gamma) = 2$ if Γ is bipartite with non-empty edge set the following characterization follows immediately:

Corollary. Let Δ be a 1-diagram. Then

(i) Δ is a secondary structure iff $\chi(\Theta(\Delta)) = 1$;

(ii) Δ is a bi-secondary structure iff $\chi(\Theta(\Delta)) \leq 2$.

The chromatic number $\chi(\Theta(\Delta))$ may therefore serve as a measure for the structural complexity of a contact structure. The following example shows that there are natural RNA structures that have a chromatic number $\chi(\Theta(\Delta)) > 2$. The *Escherichia coli* α-operon mRNA folds into a structure that is required for allosteric control of translational initiation [88]. Compensatory mutations have defined an unusual pseudo-knotted structure [87], the thermodynamics of which were subsequently investigated in detail [24]. The diagram of its contact structure cannot be drawn without intersections, see figure 8.

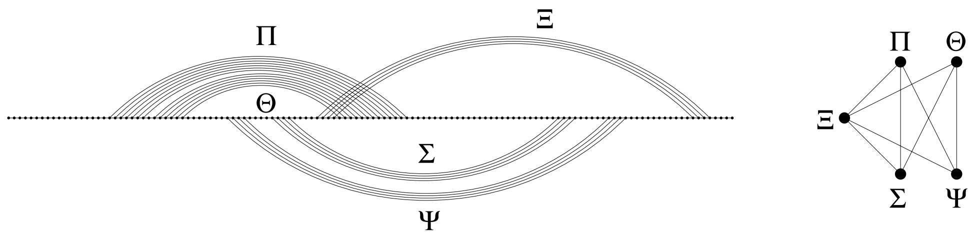


Figure 8: Diagram of the contact structure of E. coli α-mRNA. The structure contains 5 stems, labeled by uppercase Greek letters. As a consequence of lemma 7 we may choose the color partition if $\Theta(\Delta)$ such that all arcs in a stem have the same color. It therefore suffices to draw the inconsistency graph for stems (r.h.s. of the figure). It contains triangles, thus the diagram of this RNA structure is not a bi-secondary structure. It is easy to check that $\chi(\Theta(\Delta)) = 3$.

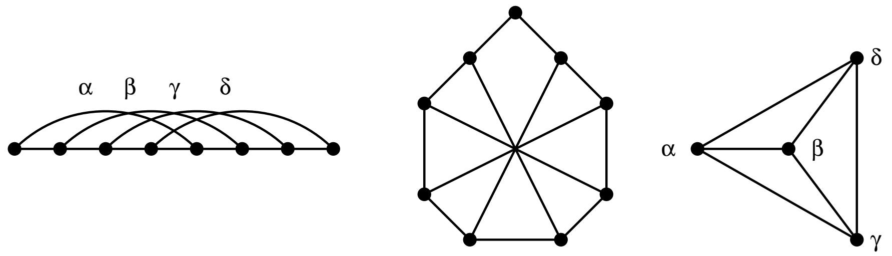


Figure 9: The graph V_8^* and its inconsistency graph.

Graphs with moderate chromatic numbers can be characterized by results similar to Kuratowski's theorem for planar graphs.

Proposition. [16] Let $k \leq 4$. A graph Γ with chromatic number $\chi(\Gamma) \geq k$ contains a subdivision of the complete graph K_k.

Remark. The generalization of this proposition to $k > 4$ is known as Hajós' conjecture. It is false for $k \geq 7$ and unsolved for $k = 5$ and $k = 6$ [34].

The graph invariant μ introduced by Colin de Verdière [12] leads to the same hierarchy of structures for small μ:

$\mu = 1$ $\iota(\Delta)$ is a circle, Δ has no arcs.

$\mu = 2$ $\iota(\Delta)$ is outerplanar, Δ is a secondary structure.

$\mu = 3$ $\iota(\Delta)$ is planar, Δ is a bi-secondary structure.

It is tempting therefore to conjecture that $\chi(\Theta(\Delta)) \leq 3$ might correspond to diagram graphs with Colin de Verdière invariant $\mu \leq 4$. The graphs with $\mu \leq 4$ have recently been identified as the *flat* or *linklessly embeddable* graphs [44]. A useful characterization of this class of graphs is proved in [67, 68]: "A graph is non-flat if and only if it has no minor in the so-called Petersen family". The graph V_8^*, figure 9, is a valid diagram graph. It is easy to check that V_8^* is flat and that its inconsistency graph is $\Theta(V_8^*) = K_4$. Hence there are flat diagram graphs for which $\chi(\Theta(\Delta)) \geq 4$. We do not know whether all diagram graphs with $\chi(\Theta(\Delta)) \leq 3$ are flat.

3.5 A Metric for 1-Diagrams

An interesting algebraic interpretation of secondary structures was proposed in [65]. Interpreting each arc $\{i, j\}$ as a transposition (i, j) on $[n]$ we may assign the permutation

$$\pi(\Delta) := \prod_{\alpha \in \Omega} (i_\alpha, j_\alpha) \tag{3}$$

to each diagram Δ.

Lemma 9. (i) If Δ a 1-diagram then $\pi(\Delta)$ is an involution.
(ii) An involution π is the permutation representation of a 1-diagram if and only if its cycle decomposition does not contain a canonical transposition, i.e., a transposition of the form $(i, i+1)$.
(iii) Different 1-diagrams give rise to different involutions.

Proof. (i) Since the arcs of an 1-diagram are disjoint we find only 1-cycles (the unpaired vertices) and 2-cycles (the arcs) in the cycle decomposition of $\pi(S)$. Thus $\pi(S)$ is an involution. The claims (ii) and (iii) are obvious. A natural set of generators for the symmetric group S_n is the set $\mathcal{T}$ of all transpositions. The corresponding length function is

$$\ell(\pi) = n - \text{cyc}(\pi), \qquad \pi \in S_n, \tag{4}$$

where $\text{cyc}(\pi)$ is the number of cycles into which π decomposes. We have $\ell(\tau) = 1$ if and only if $\tau \in \mathcal{T}$ is a transposition. The associated metric is the canonical metric on the Cayley graph $\Gamma(S_n, \mathcal{T})$, see [65] for a detailed discussion. Since the involutions form a subset of S_n we have

Theorem 3. The function

$$d(\Delta, \Delta') := \ell(\pi(\Delta)\pi(\Delta')^{-1}) = n - \text{cyc}(\pi(\Delta)\pi(\Delta')^{-1}), \tag{5}$$

where $\pi(\Delta)$ denotes the permutation representation of a diagram Δ, is a metric on the set of all 1-diagrams with n vertices. In particular, two 1-diagrams Δ and Δ' have distance $d(\Delta, \Delta') = 1$ if and only if they differ by a single arc. Metrics on "shape space" are necessary for a detailed quantitative study of sequence-structure maps. Applications to RNA secondary structures are reported for instance in [20, 76].

3.6 The Intersection Theorem

The virtue of equ.(3) is not limited to defining a metric on the set of structures. Suppose we are given an alphabet of monomers (for instance {**A**, **U**, **G**, **C**} in the case of RNA) and a rule that determines with pairs of monomers may form a base pair (**AU**, **UA**, **GC**, **CG**, **GU**, **UG** in the case of RNA).

Definition. A sequence s is *compatible* with a structure (1-diagram) Δ if for each arc $\{i, j\}$ the letters (monomers) s_i and s_j fulfill the pairing rule. The set of all sequences that are compatible with Δ is denoted by $\mathbf{C}[\Delta]$.

Theorem 4. (*Intersection Theorem*) Let Δ and Δ' be 1-diagrams. Then $\mathbf{C}[\Delta] \cap \mathbf{C}[\Delta']$ is nonempty.

The proof of this result in ref. [66] is valid for all 1-diagrams, not only for secondary structures. The intersection theorem sets the stage for shape space covering: it allow close-by sequences to fold into structures that are as different as desired — given a suitable folding potential. Further applications of equ.(3) can be found in [101]. Neutral networks in sequence space are modeled as random graphs in [66]. This ansatz generalizes from secondary structures to 1-diagrams without modifications. The only input parameter in this model, namely the fraction λ of neutral neighbors, must be determined computationally for a particular choice of the folding potential.

4 The Energy Model

As mentioned before the most essential part and core of every energy directed folding algorithm is the energy function. In other words, any result produced by a folding algorithm depends at first on the quality of the underlying energy function. And the reliability of the energy model depends on the quality of the empirical thermodynamic data. Unfortunately empirical energy parameters of sufficient accuracy for secondary structures with pseudoknots are not available. In this section we try to overcome this lack of data with the help of sterical considerations and three more or less intuitively introduced parameters.

For practical reasons all experimental data obtained for secondary structures are also used for bi-secondary structures. Therefore, we first describe the energy model for secondary structures and then proceed to bi-secondary structures and their additional features.

4.1 Thermodynamic Nearest Neighbor Parameters

The results of both quantum chemical calculations and thermodynamic measurements suggest that horizontal (base pairing) contributions to the total energy depend exclusively on the base pair composition, whereas vertical (base stacking) contributions depend on base pair composition *and* base sequence i.e. the upstream and downstream neighbors along the chain [71]. The *nearest neighbor model* introduces the assumption that the stability of a base pair, or any other structural element of an RNA, is dependent only on the identity of the adjacent bases and/or base pairs. The model is justified by the major contribution of short-range interactions (hydrogen bonding, base stacking) to the overall stabilizing energy of nucleic acid structures. In addition, it is natural to assign loop entropies to entire loops instead of individual bases. Treating stacks as special types of loops, one assumes therefore that the energy of an RNA secondary structure Φ is given by the sum of energy contributions ϵ of it's loops L.

$$E(\Phi) = \sum_{L \in \Phi} \epsilon(L) + \epsilon(L_{ext}), \tag{6}$$

where L_{ext} is the contribution of the "exterior" loop containing the free ends. Note that here stacked pairs are treated as minimal loops of degree 2 and size 0. In the following we shall discuss the individual contributions in some detail.

In particular, the energy model contains the following contributions [91]:
Stacked pairs and G-U mismatches contribute the major part of the energy stabilizing a structure. Surprisingly, in aqueous solution parallel stacking of base pairs is more important than hydrogen bonding of the complementary bases. By now all 21 possible combinations of A-U G-C and G-U pairs have been measured in several oligonucleotide sequences with an accuracy of a few percent. The parameters involving G-U mismatches were measured more recently in Douglas Turner's group [30] and brought the first notable violation of the nearest-neighbor model: while all other combinations could be fitted reasonably well to the model, the energy of the $\substack{5'\text{G-U}\,3' \\ 3'\text{U-G}\,5'}$ stacked pair seems to vary form $+1.5$ kcal/mol to -1.0 kcal/mol depending on its context.
Unpaired terminal nucleotides and terminal mismatches: unpaired bases adjacent to a helix may also lower the energy of the structure through parallel stacking. In the case of free ends, the bases dangling on the $5'$ and $3'$ ends of the helix are evaluated separately, and unpaired nucleotides in multi-loops are treated in the same way. For interior and hairpin loops, the so called *terminal mismatch* energy depends on the last pair of the helix and both neighboring unpaired bases. While stacking of an unpaired base at the $3'$ end can be as stabilizing as some stacked pairs, $5'$ dangling ends usually contribute little stability. Terminal mismatch energies are often similar to the sum of the two corresponding dangling ends. Typically, terminal mismatch energies are not assigned to hairpins of size three. Few measurements are available for the stacking of unpaired nucleotides on G-U pairs, and for this reason they have to be estimated from the data for G-C and A-U pairs.

Loop energies are destabilizing and modeled as purely entropic. Few experimental data are available for loops, most of these for hairpins. The parameters for loop energies are therefore particularly unreliable. Data in the newer compilation by Jaeger et.al. [39] differ widely from the values given previously [22]. Energies depend only on the size and type (hairpin, interior or bulge) of the the loop. Hairpins must have a minimal size of 3, and values for large loops ($k > 30$) are extrapolated logarithmically:

$$\mathcal{H}(k) = \mathcal{H}(30) + \text{const.} \times \log(k/30) \tag{7}$$

Asymmetric interior loops are furthermore penalized [55], using an empirical formula depending on the difference $|u_1 - u_2|$ of unpaired bases on each side of the loop.

$$\Delta F_{\text{ninio}} = \min\left\{\Delta F_{\max}, |u_1 - u_2| \times \Delta F_{\text{ninio}}\left(\min\{4.0, u_1, u_2\}\right)\right\} \tag{8}$$

For bulge loops of size 1, a stacking energy for the stacking of the closing and the interior pair is usually added, while larger loops are assumed to prohibit stacking. Finally, a set of eight hairpin loops of size 4 are given a bonus energy of 2 kcal/mol. These tetraloops have been found to be especially frequent in rRNA structures determined from phylogenetic analysis. Melting experiments on several tetraloops [3] show a strong sequence dependence that is not yet well reflected in the energy parameters. No measured parameters are available for multi-loops, their contribution (apart from dangling ends within the loop) is approximated by logarithmic extrapolation. Energy parameters for the contributions described above have been derived mostly from melting experiments on small oligonucleotides. The first compilation of such parameters was done by Salser [72]. The parameters most widely in use today are based on work of D. Turner and coworkers . The current work uses the compilation of [22, 91, 30], who performed measurements at 37°C in 1 M NaCl. More recently the differences between symmetric and asymmetric loops have been reported to be only half the magnitude suggested by Papanicolau *et.al.* [55] and of higher sequence dependence [57]. Serra *et.al.* [78] found a dependence of hairpin loop energies on the closing base pair and presented a model to predict the stability of hairpin loops [77]. Walter and coworkers suggested a model system for the coaxial stacking of helices [95]. Wu and Walter studied the stability of tandem GA mismatches and found them to depend upon both sequence and adjacent base pairs [96, 45]. Ebel and coworkers measured the thermodynamic stability of RNA duplexes containing tandem G-A mismatches [70]. Morse and Draper presented thermodynamic parameters for RNA duplexes containing several mismatches flanked by C-G pairs. Mismatches are reported to have a wide range of effects on duplex stability; the nearest neighbor model is considered not to be valid for G-A mismatches [53]. These results are, however, not yet included into the parameter set used in this work.

4.2 Bi-secondary Structure Features

4.2.1 The Sterical Hindrance Involving Pseudoknot Formation

It is obvious that tertiary interactions like pseudoknots are subjected to sterical considerations. In contrast to the lack of thermodynamic data for pseudoknot forming, a lot of geometry information for RNA in general is available [71]. So the energy function deals mainly with sterical consideration beside the entropic contribution of loop formation. The basic idea rests on two simplifications: RNA stacks are viewed as stiff rods and unpaired regions are assumed to be very flexible. So if we want to close a loop, containing stiff rods and flexible chains the following parameter describes the sterical hindrance:

$$\nu = Ku - L_{i_{max}} + \sum_{i \neq i_{max}}^{n} L_i \tag{9}$$

- u ... unpaired bases. Based on neutral networks a model of evolutionary adaptation can be proposed.
- $L_{i_{\max}}$ ... number of bases in the biggest stack.

- $L_i \ldots$ number of bases in stack i.
- $K \ldots$ constant that determines how many stacked bases can be bridged by an unpaired base.
- $E_{ps} \ldots$ lowest possible energy contribution for pseudoknot generating loops.

The free energy values are estimated by extrapolation using the theory of Jacobson and Stockmayer [83]. If the free energy needed to join the ends of an unrestricted, zero volume polymer is known, the theory predicts the free energy needed to form a similar but larger loop. The required size of an RNA loop before it starts behaving as such a polymer still needs to be determined. We therefore introduced a threshold constant $\bar{\nu}$, which we used as a starting point for logarithmic extrapolation.

Three different cases can be distinguished:

(i) $\nu < 0$ loop formation is impossible

(ii) $\nu < \bar{\nu}$ loop energy is fixed at a constant value E_{ps}

(iii) $\nu > \bar{\nu}$ logarithmic extrapolation of destabilizing entropy loop contribution

$$\Delta G = E_{ps} + \text{const.} \times log(\nu/\bar{\nu}) \tag{10}$$

The pre-logarithmic multiplication factor was also used to extrapolate the energies of all other loops (equation 7). In order to get a notion which values ν can adopt, we give a view examples. For each example we calculate ν with to different values of K.

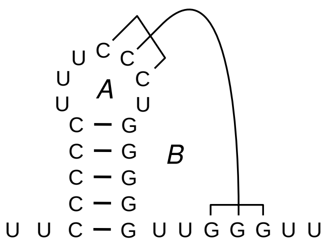


Figure 10:
$\mathbf{K} = \mathbf{2.5}$ $A : \nu = 7$ $B : \nu = 2.5$
$\mathbf{K} = \mathbf{3}$ $A : \nu = 9$ $B : \nu = 4$

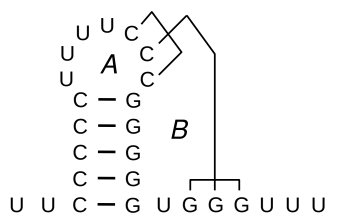


Figure 11:
$\mathbf{K} = \mathbf{2.5}$ $A : \nu = 7$ $B : \nu = -2.5$
$\mathbf{K} = \mathbf{3}$ $A : \nu = 9$ $B : \nu = -2$

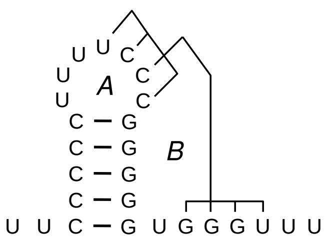


Figure 12:
$\mathbf{K} = \mathbf{2.5}$ $A : \nu = 3.5$ $B : \nu = -2.5$
$\mathbf{K} = \mathbf{3}$ $A : \nu = 8$ $B : \nu = -2$

The first three pictures show H-type pseudoknots, only in figure 10 pseudoknot formation is possible. The following four pictures display more complex situations. For instance in figure 15 and figure 16 a second pseudoknot can only be formed if $\mathbf{K} = \mathbf{3}$.

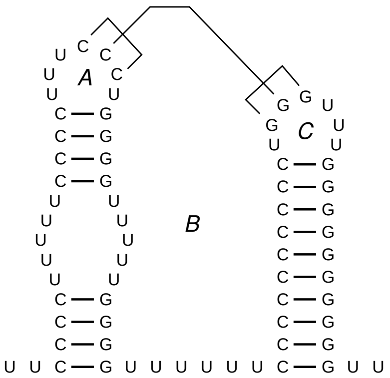


Figure 13:

K = 2.5 $A : \nu = 7$ $B : \nu = 30.5$

K = 3 $A : \nu = 9$ $B : \nu = 37$

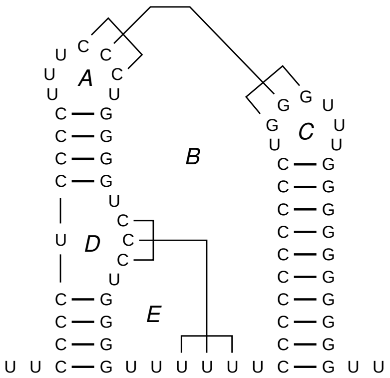


Figure 14:

K = 2.5 $A : \nu = 7$ $B : \nu = 4$

$C : \nu = 7$ $D : \nu = 4.5$ $E : \nu = 3.5$

K = 3 $A : \nu = 9$ $B : \nu = 6$:

$C : \nu = 9$ $D : \nu = 6$ $E : \nu = 5$

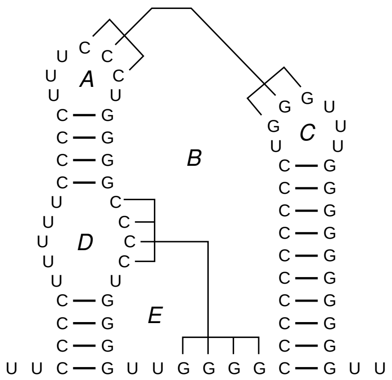


Figure 15:

K = 2.5 $A : \nu = 7$ $B : \nu = -1$

$C : \nu = 7$ $D : \nu = 11$ $E : \nu = 3.5$

K = 3 $A : \nu = 9$ $B : \nu = 0$

$C : \nu = 9$ $D : \nu = 14$ $E : \nu = 5$

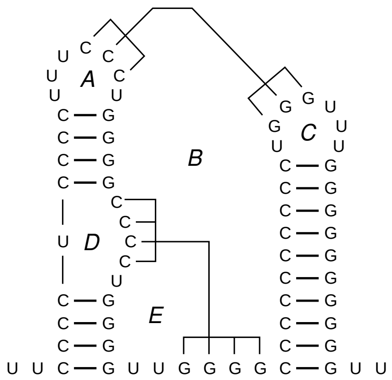


Figure 16:

K = 2.5 $A : \nu = 7$ $B : \nu = -1$

$C : \nu = 7$ $D : \nu = 1$ $E : \nu = 3.5$

K = 3 $A : \nu = 9$ $B : \nu = 0$

$C : \nu = 9$ $D : \nu = 2$ $E : \nu = 5$

5 Structure Prediction

There are several methods to deduce RNA-structures from a given sequence information and like almost all scientific prediction methods they make use of experimental results and simplifying assumptions. RNA prediction methods can be divided into two broad classes: Folding by *phylogenetical comparison* and *energy directed* (i.e.kinetic or thermodynamic) folding.

5.1 Phylogenetic Structure Analysis

Given a large enough number of sequences with identical secondary structure, that structure can be deduced by examining covariances of nucleotides in these sequences. This is the principle used for structure prediction through phylogenetic comparison of *homologous* (common ancestry) sequences [28]. Basically these methods just look for compensatory mutations such as an A change to C in position i of the aligned sequences simultaneously with a change from U to G in position j, indicating a base pair (i,j). So the sequence alignment is the most complicated theoretical part (if the sequences in the set are to dissimilar). The basic assumption is that structure is more conserved during evolution than sequence, since it is the structure that determines function. The only experimental information needed is a large enough number of sequences. Fortunately the sequence of nucleic acids is nowadays one of the best accessible molecular biological information. In fact the success of the method in the prediction of, for instance, the secondary structures of the 16S ribosomal RNAs [105], RNAseP or the clover-leaf structure of tRNAs provides an excellent justification for this method.

The advantages: Since no assumptions about pairing rules are necessary, non-canonical pairs and tertiary interactions can be detected as well.

The disadvantages: A sufficiently large set of sequences which exhibit the proper amount of variation has to be provided. So the sequences should be dissimilar enough to show many covariations while still yielding a good alignment. If there are strongly conserved regions (i.e. the function is sequence dependent) or parts of the structure are highly variable (because non-functional) our assumption holds not true. As a consequence, phylogenetically determined structures usually are incomplete, that means, they do not show all base pairs of the actual structures.

Nevertheless phylogenetic comparison can generate the most reliable structure models to date and are therefore frequently used for comparison of other folding algorithms.

5.2 Energy Directed Folding

There are to different approaches to energy directed folding: Algorithms that search for the structure of *minimal free energy* (or the equilibrium ensemble) and *kinetic folding algorithms*. It is not known if the biological relevant structure of a given RNA molecule is the structure of minimal free energy. The structure might be trapped in some local minimum during the folding process (this might be the case with long RNA-molecules). Kinetic algorithms therefore try to simulate the folding process. The folding of pseudoknots can be easily included and therefore we restrict the discussion to this method.

The Kinetic Algorithm

The first kinetic algorithm was proposed by Martinez [50] in 1984, mainly as an attempt to create a faster algorithm. As do many other algorithms it starts by compiling a list of possible helices. His idea was that the helix with the largest equilibrium constant (that is the lowest energy) would form first. All helices not compatible with this helix are then deleted from the list. The process is repeated until no helix is left whose incorporation would lower the energy of the structure. Such an algorithm will indeed execute in only $O(n^2)$ steps. The procedure implies that a helix that has once formed never opens again, so there is no refolding. Currently the only kinetic algorithms allowing sequence re-folding are nondeterministic. Furthermore, folding in vivo already starts during transcription, so that helices near the 5' end of the sequence should be formed first. Algorithms that take this fact into consideration are discussed in [1, 27].

The advantages: with a reliable energy function and an appropriate algorithm no additional experimental data is needed to fold any sequence. If the algorithm is fast enough a lot of interesting statistical properties concerning RNA folding can be revealed.

The disadvantages: the experimental data which is used to derive our thermodynamic parameters can never be completely satisfying (for instance pseudoknot data). Even with excellent parameters it is impossible to simulate the in vivo conditions during the folding process. So the simplifications which we are forced to do, have to be kept in mind when looking at the results.

In this work a kinetic folding algorithm generalized for bi-secondary structures was utilized, mainly because the implementation of pseudoknot folding is easily accomplished. In contrast to the original Martinez algorithm the generalized folding algorithm, of course, does not exclude pseudoknots as they occur in bi-secondary structures.

5.3 Parameter Adjustment

To evaluate the energy function parameters a rather heuristic approach was used. Experimental research provides a lot of RNA bi-secondary structures mainly derived from comparative sequence analysis and complementing probing experiments. Some of them were used to adjust the parameter set.

5.3.1 The H-Type Pseudoknots

H-type pseudoknots are the most simple cases of tertiary interactions (figure 10,11,12). Therefore we started our attempt to adjust energy function parameters with a set of short RNAs with only one H-type pseudoknot. The selected RNAs are about 80nt long, all of them fragments of longer RNAs. They were taken from 7 closely related bacteriophages. The folding started with the open chain.

Table 1: Sample of closely related bacteriophages.

#	Phage	Length	Pk-position
1	LZ5	82nt	19 ..42
2	LZ3	81nt	19 ..42
3	T4	81nt	18 ..41
4	Tu1A	81nt	18 ..41
5	OX2	87nt	18 ..41
6	Tu1B	81nt	18 ..41
7	Baker	84nt	18 ..41

```
  |   1    |   10    |   20    |   30    |   40    |   50    |   60    |   70    |   80
1 ...........(((((....[[[)))))..((((((((.]]]........))))))))..((((((....))))))......
  UACUCAUUAAAGGUGUUGCUUGUGCACUACGUCAAGAUAGCAUUAUUAGAAUUUGGUUUCCGCAGAAGCCUGCGGUCACUU
2 ...........(((((....[[[)))))..((((((((.]]]........))))))))..((((((....)))))).....
  UACUCAUUAAAGGUGUUGCUUGUACACUACGUCAAAAUAGCAUUGUUAGAAUUUUGAUUUCCGCAGAAGCCUGCGGUCACU
3 ((((.......)))).......[[[[.....((((((((.]]]]......))))))))..((((((....))))))....
  UACUCAUUAAAGGUAUUGCUUGUGCACUACGUCAAGAUAGCAUUGUUAGAAUUUUGAUUUCCGCAGAAGCCUGCGGUCACU
4 ((((.......))))......[[[[....((((((((.]]]]......))))))))..((((((....)))))).....
  UACUCAUUAAAGGUAUUGCUUAUGCACUACGUCAAGAUAGCAUUGUUAGAAUUUUGAUUUCCGCAGAAGCCUGCGGUCACU
5 ((((.......)))).[[[[((.....))..((((((.]]]].........))))))..((((((....))))))..........
  UACUCAUUAAAGGUAUUGCUUGUACACUACGUCAAGAUAGCACUGUUAGAAUUUUUGAAACCAUAAAAUCCAAAAUUUUUAUCAACU
6 ((((.......)))).[[[[.........((((((((]]]].......))))))))..((((((....)))))).....
  UACUCAUUAAAGGUAUUGCUUGUACACUACGUCAAGAUAGCAUUAUUAGAAUUUUGAUUUCCGCAGAAGCCUGCGGUCACU
7 ((((.......)))).[[[[((.....))..((((((.]]]].........))))))..((((((....)))))).......
  UACUCAUUAAAGGUAUUGCUUGUACACUACGUCAAGAUAGCAUUGUUAGAAUUUUUGAAACCAUAGAAUCCAAAAUUUUUUAUC
```

In table 1 we display the experimentally predicted region for pseudoknotting, below the results of our folding experiments are shown. The H-type pseudoknot formation is relatively robust if the parameters are not varied to much. The results were obtained with the following set of parameters:

$$
\begin{aligned}
8 > K &> 4 \\
12 > \bar{\nu} &> 8 \\
E_{pk} &= 4200 \text{ cal/mole}
\end{aligned}
$$

The E_{pk} contribution for the lowest possible pseudoknot generating loop was chosen according to [1].

5.3.2 Beyond H-Type Pseudoknots

Two rather complex molecules were picked out to demonstrate their pseudoknot folding behavior and the problems dealing with them in more detail. One example chosen for short distance pseudoknots (almost exclusively H-type) is tmRNA, a molecule with interesting molecular biological features. Another example for long distance tertiary interactions is RNAseP RNA. Of course both RNAs fold into bi-secondary structures. In both cases the complete secondary structure without pseudoknots was used as start structure from which the folding process commenced. This is in accordance with the assumption, that at first the hole secondary structure is formed before tertiary interactions are established.

tmRNA

The five pseudoknots, of course show different behavior if we vary the parameter set. PK 3 is a good examples for the adjustment of parameter K. Because if K is smaller than 5, ν is smaller than 0, therefore PK 3 would be ruled out. In this case the assumption that stacks form stiff rods holds not true, because a single unpaired base is not likely to bridge five stacked basepairs. PK 1 is an example for a pseudoknot where K is not the most important constrain. In this case the logarithmic extrapolation together with $\bar{\nu}$ dominates the folding probability. Finally it turned out, that if we choose parameters that enable formation of PK 2, a view unintentional long range pseudoknots arise.

```
       <           R1           >               <            PK1           >
(((((((((((((..[[[[.)))))))...]]]].........((((((((((.[[[[[...)))))......]]]]]...
GGGGCUGAUUCUGGAUUCGACGGGAUUUGCGAAACCCAAGGUGCAUGCCGAGGGGCGGUUGGCCUCGUAAAAAGCCGCAAA
|   1    |   10    |   20    |   30    |   40    |   50    |   60    |   70    |
                                                        <                   PK2
...((((((....)))))).......((((((((((.........)))).))))))(((((((((((.((((...[[[[.)))
AAAUAGUCGCAAACGACUAAAACUACGCUUUAGCAGCUUAAUAACCUGCUUAGAGCCCUCUCUCCCUAGCCUCCGCUCUUAGG
  80    |   90    |   100   |   110   |  120    |  130    |   140  |  150    |  160
                                >   <                    PK3                     >
.)))))....))))))).........]]]]...((((((((((((.[[[[[)))))))))))).........]]]]]..
ACGGGGAUCAAGAGAGGUCAAACCCAAAAGAGAUCGCGUGGAAGCCCUGCCUGGGGUUUCCGCGUUAAAACUUAAUCAGGCUA
     |   170 |   180    |   190   |   200   |   210   |  220    |  230    |   240
<                      PK4                          >
((((((((((((((..[[[[[[)))))))))))))).........]]]]]]....))))).....
GUUUGUUAGUGGCGUGUCCGUCCGCCGCUGGCAAGCGAAUGUAAAGACGGACUAAGCAUGUAGUA
  |  250   |  260    |   270   |   280   |  290    |  300    |   310

((((((.........))))))..(((((.......))))))))))))....
CCGAGGAUGUAGGAACUUCGGACGCGGGUUCAACUCCCGCCAGCUCCACCA
      |   320  |  330    |   340   |   350   |  360
```

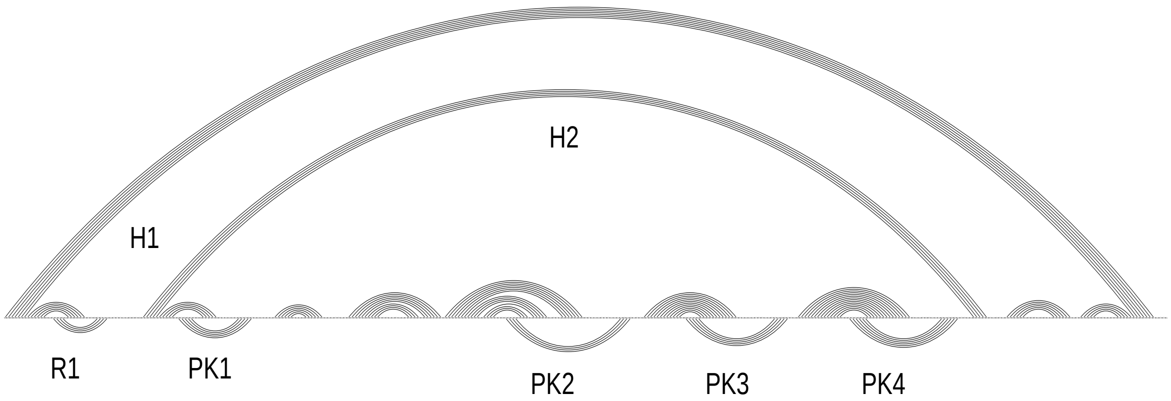


Figure 17: Sequence and structure of tmRNA in `().[]` representation and as graph

With the parameter set: $K = 5$, $\bar{\nu} = 10$, $E_{pk} = 420$ all pseudoknots except PK 2 emerge. The values lie within the intervals obtained from our H-type studies.

Table 2: PKs in tmRNA beginning with the most stable pseudoknot

PK4	PK 4.1	248-261	270-283
	PK 4.2	264-269	293-298
PK1	PK 1.1	49-53	63-67
	PK 1.2	55-60	73-78
PK2	PK 3.1	200-211	218-229
	PK 3.2	213-217	241-245
R1	R1 1.1	8-13	21-26
	R1 1.2	16-19	30-33

As mentioned above, the proposed structures H2 and pK2.2 are implied by covariation, but not supported by probing, and might then represent features of a second, functional conformation not present in the molecules studied in solution. Preliminary footprinting experiments of tmRNA with the ribosome are consistent with this suggestion, because the probing pattern of H2 and of loop2 of PK2 (as well as other parts of tmRNA) varies in the presence of ribosome. Despite the complexity and imponderables of a probing study, the combination of probing and covariation together strongly support some structural features, e.g., the pseudoknots PK3 and PK4. Some folded domains may only be stabilized by interaction with proteins (such as EF-Tu) or ribosomes. Probing data for R1 are consistent with breathing of the structure in solution, and these domains might be stabilized in vivo. So the predicted stability reflects experimental findings in a very convincing way.

RNAseP RNA

Comparative structure folding predicts two pseudoknots. None of them are H-type pseudoknots.

```
                                                                   <  PK1
((((((((((((((((((.((.((((((((((.(((((....))))))))))).)))))).))....[[[.[[[[[((((
GAAGCUGACCAGACAGUCGCCGCUUCGUCGUCCUCUUCGGGGGAGACGGGCGGAGGGGAGGAAAGUCCGGGCUCC
|   1    |   10    |   20    |   30    |   40    |   50    |   60    |    70
   < PK2
...[[[[.((((((((((.....)))))((((....))).)((...((..............(((((..((((((((
AUAGGGCAGGGUGCCAGGUAACGCCUGGGGGGGAAACCCACGACCAGUGCAACAGAGAGCAAACCGCCGAUGGCCCGC
 |   80   |   90    |   100   |   110   |  120    |  130    |   140   |  150
....))))).)))))))..)).......((((((.......)))))).((((((((....))).))))).........
GCAAGCGGGAUCAGGUAAGGGUGAAAGGGUGCGGUAAGAGCGCACCGCGCGGCUGGUAACAGUCCGUGGCACGGUAA
  |  160   |   170  |   180    |   190   |   200   |   210   |  220    |  230
                                           PK2  >
))))))))))))))...((((......((((((....(((((.]]]]))))))))))))).....))))......
ACUCCACCCGGAGCAAGGCCAAAUAGGGGUUCAUAAGGUACGGCCCGUACUGAACCCGGGUAGGCUGCUUGA
    |   240   |  250   |  260    |   270   |   280   |  290    |  300
                                                    PK1 >
((((((((....))))))))..........)))))))..........]]]]]]]].).))))))))))....
GCCAGUGAGCGAUUGCUGGCCUAGAUGAAUGACUGUCCACGACAGAACCCGGCUUAUCGGUCAGUUUCACCU
 |   310   |  320   |  330    |   340   |   350   |  360    |  370    |
```

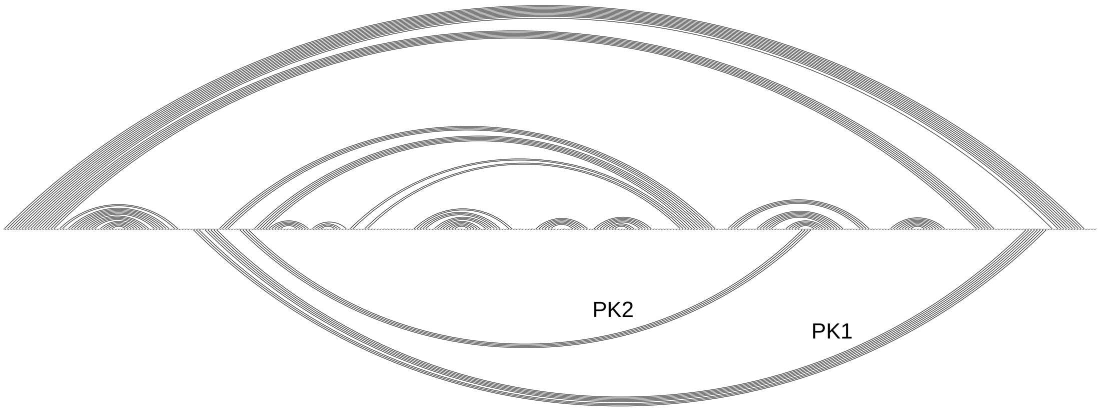


Figure 18: Sequence and structure of E.coli RNAseP in `().[]` representation and as graph

It turned out that it is almost impossible to get both of them predicted correctly. The short distance interaction of the PK 2 halfstem GGGC (82-85) with a fragment of PK 1 halfstem CCUG (71-68) prevents the formation of both PK 1 and PK 2. The problem is strengthened by the type of folding algorithm that is used. The kinetic folding algorithm does not treat PK 1 as a single stem because of the unpaired U in one halfstem. Instead two separate stems, which overlap and therefore exclude each other, are tested. So the only chance to get PK 1 and PK 2 is to modify the kinetic folding and to reduce the longrange costs extremely. To reduce the long range costs we can increase the threshold parameter $\bar{\nu}$ or decrease the constant factor for the logarithmic extrapolation. Both cases lead to parameters which are suitable for RNAseP RNAs taken from several different species (but give rise to additional pseudoknots):

- Escherichia coli
- Alcaligenes eutrophus
- Desulfovibrio desulfruicans
- Pseudomonas flurescens

Another sample of RNAseP RNAs form hairpin loop with large pseudoknot stacks. Here the simplification of sterical conditions is totally inappropriate.

- Agrobacterium tumefaciens
- Rhodospirillum rubrum
- Cyanophora paradoxa cyanelle
- Anacystis nidulans

In order to enable PK 2 we would have to assign the factor K a value of at least 7. But this leads to a high ν and to high energy contributions, unless we increment $\bar{\nu}$ dramatically, which in turn gives unrealistic numbers of pseudoknot.

Finally there is Lactobacillus acidophilus RNAseP RNA which lacks pseudoknots at all [7]. In this case experimental results are in accordance with our folding prediction.

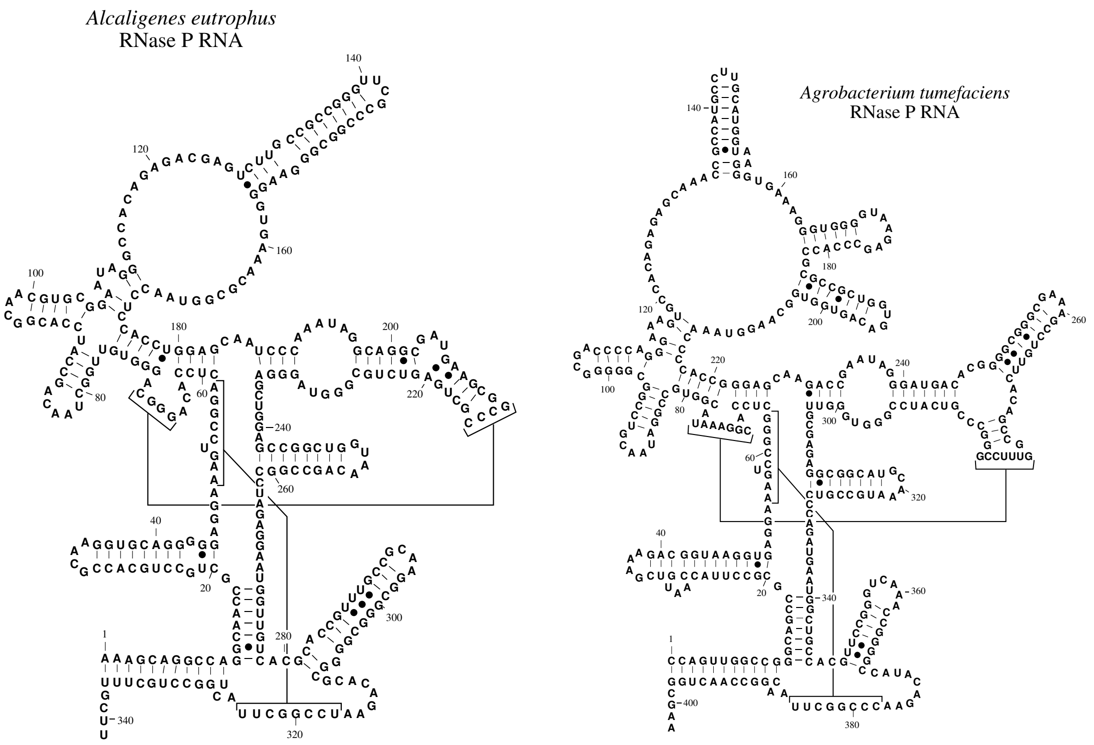


Figure 19: Secondary Structure of Alcaligenes eutrophus and Agrobacterium tumefaciens

The Results of our tmRNA and RNAseP Folding Experiments:

The results show that it is much easier to get parameters that lead to short distance (H-type) pseudoknots than to get parameters for long range interactions. That means if we choose parameters that enable the formation of long range interactions a lot more H-type pseudoknots than in the phylogenetical structure predicted will arise.

Particularly in RNAseP RNA additional non pseudoknot interactions are predicted which sometimes compete with pseudoknot stacks for bases. Since the parameters for secondary structures were not changed two possibilities are conceivable to avoid this additional stems. Either the algorithm only allows pseudoknot formation and uses the start structure as the most stable secondary structure, or the selected parameters prefer pseudoknot formation extremely. It always has to be kept in mind that both RNAseP RNA and tmRNA structure models were derived in large parts from comparative structure analysis. That means, that they may not show all base pairs of the real structure. However, changes in tertiary structure tend to be rarer than most compensatory changes in regular helices, so larger sets of sequences are required to detect them. Making comparative analysis of tertiary structure more difficult still, the specificities of bases that engage in tertiary pairs or triples tend to be less rigid than the canonical complementarities that establish secondary structure. Another crucial point concerning tertiary interactions are the RNA associated proteins. This proteins are thought to influence the tertiary structure much more than the pure secondary structure, particularly longrange interactions.

6 The Combinatory Map of RNA Bi-Secondary Structures

In this section we present estimations for the enumeration of secondary structures as well as bi-secondary structures. We also study the statistics of bi-secondary structures produced with different energy function parameter sets.

6.1 Enumeration of Bi-Secondary Structures

6.1.1 Enumeration 1-Diagrams

The number X_n of all diagrams on n vertices is $X_n = 2^{(n-1)(n-2)/2}$ since there are $(n-1)(n-2)/2$ possible arcs [79], which can be arbitrarily combined to form a diagram. In lemma 9 we have shown that all 1-diagrams correspond to involutions, therefore the number T_n of involutions on $[n]$ is an upper bound for the number D_n of 1-diagrams on $[n]$. The combinatorics of involutions is discussed for instance in the book [103]:

Proposition. The number T_n of involutions fulfills the recursion

$$T_n = T_{n-1} + (n-1)T_{n-2} \qquad n \geq 2 \qquad \text{and} \qquad T_0 = T_1 = 1\,,$$

and has the asymptotic form

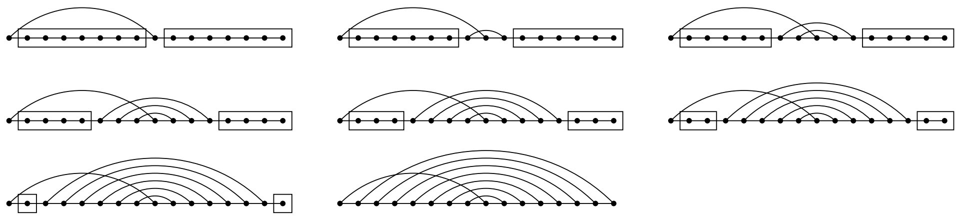
$$T_n \sim \frac{1}{\sqrt{2}} n^{n/2} \exp\left(-\frac{n}{2} + \sqrt{n} - \frac{1}{4}\right)\,.$$

The number of involutions T_n therefore grows faster than exponential in the sense that $\sqrt[n]{T_n} \to \infty$. 1-Diagrams can be counted by a very similar similar recursion as the following result shows:

Theorem 5. The number of 1-diagrams fulfills the recursion

$$D_{n+2} = D_{n+1} + (n+1)D_n - D_{n-1} + D_{n-2} \quad n \geq 2 \qquad D_0 = D_1 = D_2 = 1,\; D_3 = 2\,.$$

Proof. The first few values of D_n are obvious, $D_0 = 1$ is a convenient definition. The recursion is derived as follows: A 1-diagram on $n+2$ vertices can be formed either by adding a lone vertex to a 1-diagram on $n+1$ vertices or by adding an arc $\{1,k\}$ to a 1-diagram Δ on n vertices by inserting the vertex labeled k between the $k-1$st and the kth vertex of Δ. Note, however, that Δ must be a 1-diagram, but in addition it might have an arc $\{k-1,k\}$ in Δ, since these vertices are separated by the endpoint of the newly introduced arc in the new structure. Viewing this differently, we may either add the arc $\{1,k\}$ or the Ψ-like structure consisting of the arcs $\{1,k\}$ and $\{k-1,k+1\}$, which leaves us with a 1-diagram on $n-2$ vertices and the same problem. Repeating this argument we arrive at the following expansion:



Hence we have $D_{n+2} = D_{n+1} + nD_n + (n-1)D_{n-2} + (n-3)D_{n-4} + \ldots$ Observing that D_{n+1} can of course be written in the same form and substituting into the above equations yields

$$D_{n+2} = (n+1)D_n + nD_{n-1} + (n-1)D_{n-2} + (n-2)D_{n-3} + \ldots + 2D_1 + D_0 - D_{n-1}\,.$$

Subtracting the corresponding expansion for D_{n+1} yields

$$D_{n+2} - D_{n+1} = (n+1)D_n - D_{n-1} + D_{n-2}\,.$$

A simple rearrangement now completes the proof.

Corollary. $\lim_{n\to\infty} \sqrt[n]{D_n} = \infty$.

Proof. The series D_n is obviously monotonically increasing. Hence the series $a_{n+2} = (n+1)a_n$, $a_0 = a_1 = 1$ is a lower bound. It is well known that $a_n = (n-1)!!$ grows faster than exponentially.

Remark. A very similar formula is obtained for the case of a circular backbone. There are D_{n-2} diagrams with arc $\{1, n\}$ on n vertices. Thus the number of 1-diagrams with circular backbone is $D'_n = D_n - D_{n-2}$. An exponential upper bound can be found, however, on the numbers $D_n(c)$ of 1-diagrams whose inconsistency graph has chromatic $\chi(\Theta(\Delta)) \leq c$. We find

Theorem 6. $D_n(c) \leq (2c+1)^n$.

Proof. Consider a 1-diagram $\Delta = ([n], \Omega)$ with $\chi(\Theta(\Delta)) \leq c$. Then there is a color partition of Ω with c colors. As $([n], \Omega_i)$ is a secondary structure, it can be encoded in dot-parenthesis notation. Coloring the parenthesis with a different color for each class Ω_i of the color partition hence yields a unique representation of Δ. This representation can be interpreted as a string of length n over an alphabet consisting of '.' and c different pairs of brackets, i.e., with $2c+1$ letters. Theorem 6 is not a very good estimate as we shall see in section 3.3.

6.1.2 Secondary Structures

A secondary structure on $n+1$ digits may be obtained from a structure on n digits either by adding a free end at the right hand end or by inserting a base pair $1 \equiv (k+2)$. In the second case the substructure enclosed by this pair is an arbitrary structure on k digits, and the remaining part of length $n-k-1$ is also an arbitrary valid secondary structure. Therefore, we obtain the following recursion formula for the number S_n of secondary structures:

$$S_{n+1} = S_n + \sum_{k=m}^{n-1} S_k S_{n-k-1}, \quad n \geq m+1 \qquad S_0 = \ldots = S_{m+1} = 1 \qquad (11)$$

Table 3: The constants A_{ml} in equ.(12).

m	l		
	1	2	3
1	2.618	1.986	1.716
2	2.414	1.899	1.680
3	2.289	1.849	1.652
5	2.147	1.783	1.612

This expression has first been derived by Waterman [97]; m denotes the minimum number of unpaired digits in a hairpin loop. Similar recursions can be derived for the numbers $\Psi_n^{(m,l)}$ of secondary structures with minimum hairpin length m and minimum stack length l, see [32] for details. Asymptotically, these numbers behave as

$$\Psi_n^{(m,l)} \sim B_{m,l} n^{-3/2} A_{m,l}^n \,. \tag{12}$$

The most important numbers are collected in table 3. A more detailed table can be found in [32].

Detailed combinatorial studies on various aspects of secondary structure graphs are based on equ.(13), see for instance [56, 82, 97, 99, 100, 98, 32]. In the following we shall make use of the number

$$s(n,k) = \frac{1}{k}\binom{n-k}{k+1}\binom{n-k-1}{k-1} \tag{13}$$

of secondary structures of length n with k base pairs. This closed formula was recently derived in [74].

6.1.3 Bi-Secondary Structures

A first naive upper bound is $D_n(2) \leq S_n^2$, since on each side of the x-axis we have a secondary structure on n vertices. Theorem 5 implies $D_n(2) \leq 5^n$.

A slightly better bound can be derived using the enumeration of secondary structures:

Lemma 10. $D_n(2) \leq \max\limits_{\substack{0\leq k+l\leq n/2\\ l\leq k}} \frac{n}{2}\binom{n-k-1}{k-1}\binom{n-k}{k+1}\binom{n-2k}{2l}\binom{2l}{l}$.

Proof. We start with the $s(n,k)$ secondary structures with k arcs. In order to produce a bi-secondary structure we use $2l$ of the $n-2k$ unpaired positions for introducing l additional arcs. There are $\binom{n-2k}{2l}$ possible choices for these additional pairs, which may form any of the $C_l = \frac{1}{l+1}\binom{2l}{l}$ possible configurations of l matched parentheses. C_l is a Catalan number. Without loosing generality we may assume that $l \leq k$, i.e., the partial secondary structure with the larger number of pairs is drawn above the x-axis. Thus

$$D_n(2) \leq \sum_{k=0}^{n/2}\sum_{l=0}^{k} s(n,k)\binom{n-2k}{2l} C_l\,.$$

Replacing the sums by appropriate multiples of the maximum entry is trivial. Note that this bound is still a gross overestimate: (i) It contains all the redundancy of the `().[]`-representation. (ii) The number C_l also counts conformations of square brackets of the form `[]`, which do not correspond to a graph at all, and it counts conformations in which not all square brackets are inconsistent with an arc that is represented by a round bracket. These latter configurations are counted more than once.

Corollary. $\lim\limits_{n\to\infty} \sqrt[n]{D_n(2)} \leq 4.76136931.$

Proof. Let $A_n(k,l)$ denote argument of the maximum in lemma 10. It is straightforward to compute

$$\begin{aligned} A(x,y) &:= \lim_{n\to\infty}\frac{1}{n}\log A_n(nx,xy) \\ &= 2(1-x)\log(1-x) - 2x\log x - (1-2x)\log(1-2x) \\ &\quad -(1-2x-2y)\log(1-2x-2y) - 2y\log(y) \end{aligned} \tag{14}$$

Set $A := \max\{\, A(x,y) \,|\, 0 \le x+y \le 1/2 \wedge y \le x\,\}$. Then $\lim \sqrt[n]{D_n(2)} \le \exp(A)$. Solving the optimization problem that defines A is straight-forward. A short computation shows that $\hat{y} = 1/\sqrt{21}$ and $\hat{x} = (7-\sqrt{21})/14$ is the only local maximum with $x, y \le 1/2$. It violates the condition $y \le x$, however. The solution thus lies on the boundary of the triangle $(0,0)$, $(1/2,0)$ and $(1/4,1/4)$. Setting $y = 0$ one obtains the maximum $\hat{x} = 1/2 - 1/\sqrt{20}$. Along the edge $x+y = 1/2$ we find $\hat{y} = 1/\sqrt{12}$ violating the condition $y \le x$. With $x = y$ we arrive at the cubic equation $31x^3 - 31x^2 + 10x - 1 = 0$ which has a single real solution $\hat{x} \approx 0.1942$. We find $A(\hat{x},\hat{x}) \approx 1.5605329 = A$, because this value is much larger than the values of $A(x,y)$ at the three corners of the triangle. More sophisticated models of RNA take into account that (i) base pairs must enclose at least $m = 3$ other bases, and (ii) that isolated base pairs are energetically disfavored. In [32] the numbers $\Psi_n^{(m,l)}$ of secondary structures with stack size at least l base pairs and separation of the vertices incident with an arc at least m is derived. We define $\Psi_n^{(m,l;\kappa)}$ to be the number of 1-diagrams with $\chi(\Theta(\Delta)) \le \kappa$ and with the same restrictions, and set

$$A_{ml}^{(\kappa)} := \lim_{n\to\infty} \sqrt[n]{\Psi_n^{(m,l;\kappa)}} \tag{15}$$

Clearly we have $\Psi_n^{(m,l;2)} \le [\Psi_n^{(m,l)}]^\kappa$ since the 1-diagram Δ is a superposition of at most κ secondary structures. In particular we find the upper bound $A_{3,2}^{(2)} \le 3.418$ for the biophysical case.

We have not been able to derive an exact counting series for bi-secondary structures. Hence we resorted to a numerical survey. We pursued three different strategies for estimating the number of bi-secondary structures:

(1) Complete enumeration is feasible only for very small values of n because the number of structures grows faster than 2^n.

(2) As an alternative we produce random strings from the alphabet `().[]` and check each string if it is the normal form of a bi-secondary structure. The number of secondary structures is then estimated by $5^n \times$

Table 4: Best estimates for the constants $A^{(2)}_{ml}$. The counting data were fitted by the model $a\, n^{-b}\, c^n$.

m	l		
	1	2	3
1	4.42	2.49	2.00
2	4.03	2.43	1.94
3	3.81	2.35	1.89
5	3.44	2.22	1.74

$N_{\mathrm{nf}}/N_{\mathrm{sample}}$, where N_{sample} is the size of the random sample and N_{nf} is the number of detected normal forms in the sample.

(3) Using the recursion for secondary structures with given minimal stack length l and given minimal hairpin size m that is described in detail in [86], we randomly generate a sample of pairs of secondary structures. Interpreting these as the upper and lower part of bi-secondary structure we check their superpositions for being normal forms of bi-secondary structures. The number of bi-secondary structures is then approximately $\Psi_n^{(m,l)} \times N_{\mathrm{nf}}/N_{\mathrm{sample}}$, where the numbers $\Psi_n^{(m,l)}$ of secondary structures with hairpins of length at least m and minimal stack length m can be obtained recursively, see [32].

Our best estimates are compiled in table 4. In the biologically interesting case, $m = 3$ and $l = 2$, we find $A^{(2)}_{3,2} \approx 2.35$.

6.2 Statistics of Bi-secondary Structures

Most people who use folding algorithms are not interested in folding millions of sequences, they just want to know the secondary structure of specific sequences. But statistics can serve as a reference to compare structures obtained with different parameters used in our energy function. To generate such statistical reference we folded samples of 10^6 random sequences from the natural GCAU and the restricted GC alphabets. Throughout our statistical investigation we applied three different parameter sets,without varying the pre-logarithmic factor:

(i) Realistic parameters: $K = 4$, $\nu = 9$, $E_{pk} = 420$

(ii) Nonrealistic parameters: $K = 3$, $\nu = 10$, $E_{pk} = 200$

(iii) No pseudoknots possible: $K = 1000$, $\nu = 0$, $E_{pk} = 1000$

In the first case the parameters lie within the intervals obtained for H-type pseudoknots (section 5.3.1), the nonrealistic parameters are expected to make pseudoknot formation much to frequent and finally we use the plain secondary structure as a reverence. All folded sequences are constrained to have a minimum hairpin size of 3 and a minimum stack size of 2, i.e., there are no isolated basepairs.

6.2.1 Distribution of Pseudoknots

The natural GCAU sequences with realistic parameters shows that by far the most sequences are without pseudoknots and the probability for more than 2 pseudoknots is for all sequence lengths low. If we use parameter set (ii), of course more pseudoknots appear, particularly for longer sequences. Is remarkable that parameters (i) give positive curvature whereas parameters (ii) give negative. For the GC sequences with realistic parameters, the plot resembles

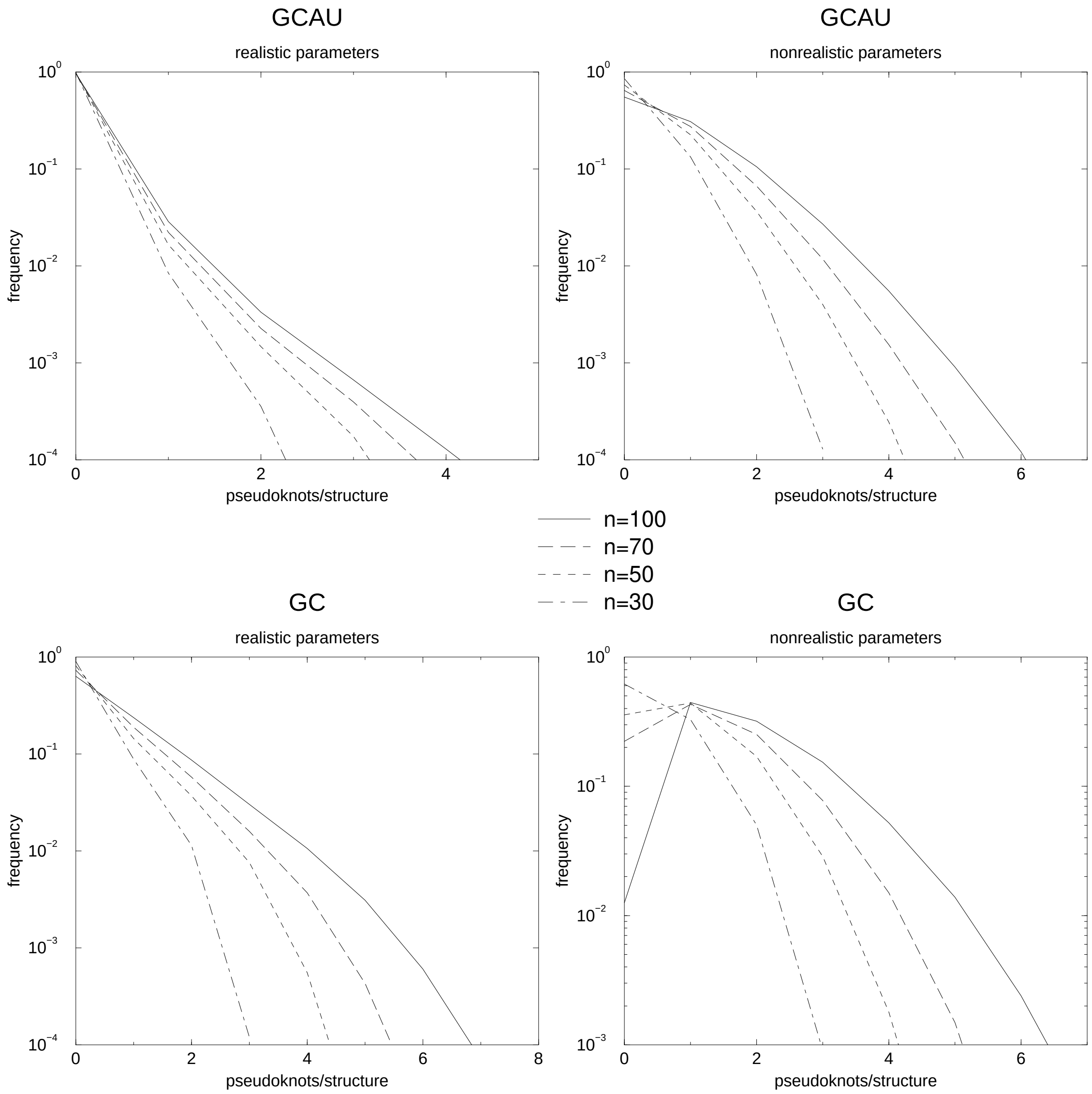


Figure 20: Number of pseudoknots for parameter set (i) (left side) and (ii) (right side), for sequence length 100,70,50 and 30.

the GCAU plot with parameter set (ii). The GC plot for nonrealistic parameters shows for longer sequences a maximum probability for one pseudoknot per structure.

6.2.2 Number of Stacks and Loops

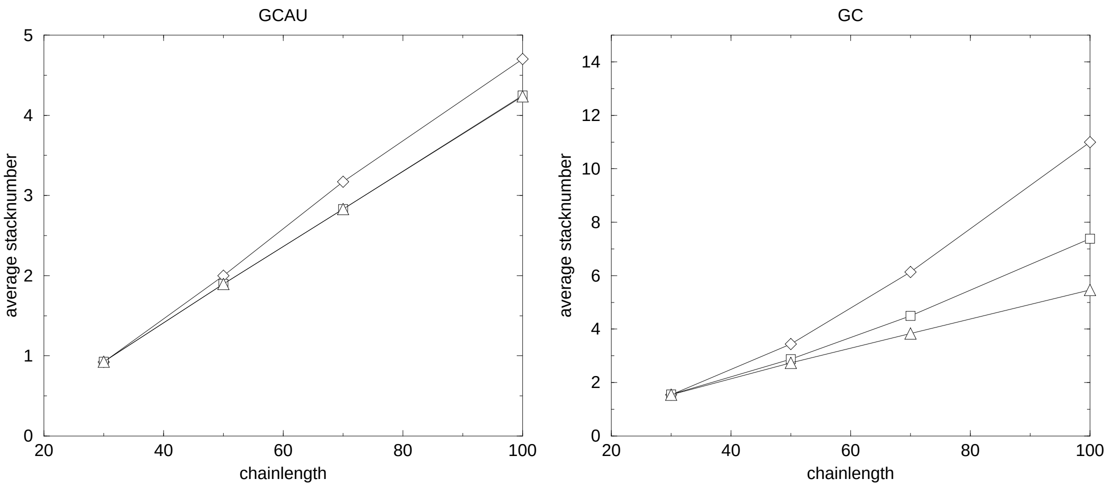


Figure 21: Mean values of the number of stacks and loops for parameter sets (i) $\square$, (ii) $\diamond$ and (iii) $\triangle$

The number of loops must equal the number of stacks because every loop must be closed by a stack. The mean number of loops and stacks for structures without pseudoknots scores linearly with the length. Dependence on the alphabet is weak. Structures folded with parameter set (ii) show at least for GCAU sequences the same behavior. In all other cases the dependency is slightly nonlinear, particularly for GC and parameter set (ii).

6.2.3 Number of Base Pairs

The mean number of base pairs increases linearly with sequence length n in all cases. Structures on the GC alphabet show much more base pairs. Parameter set (i) and (iii) produce similar plots, parameter set (ii) produces less base pairs.

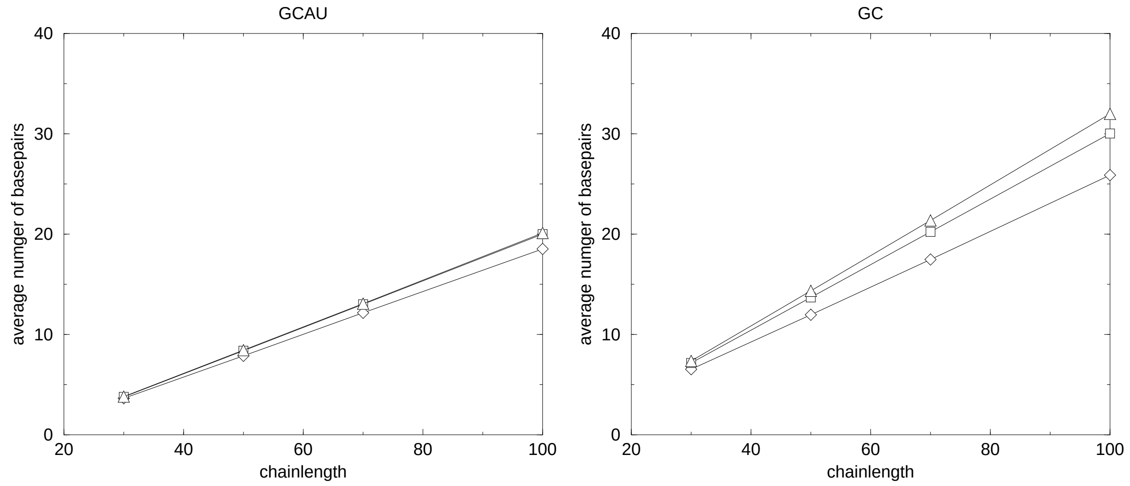


Figure 22: Mean values of the number of base pairs for parameter sets (i) □, (ii) ◇ and (iii) △

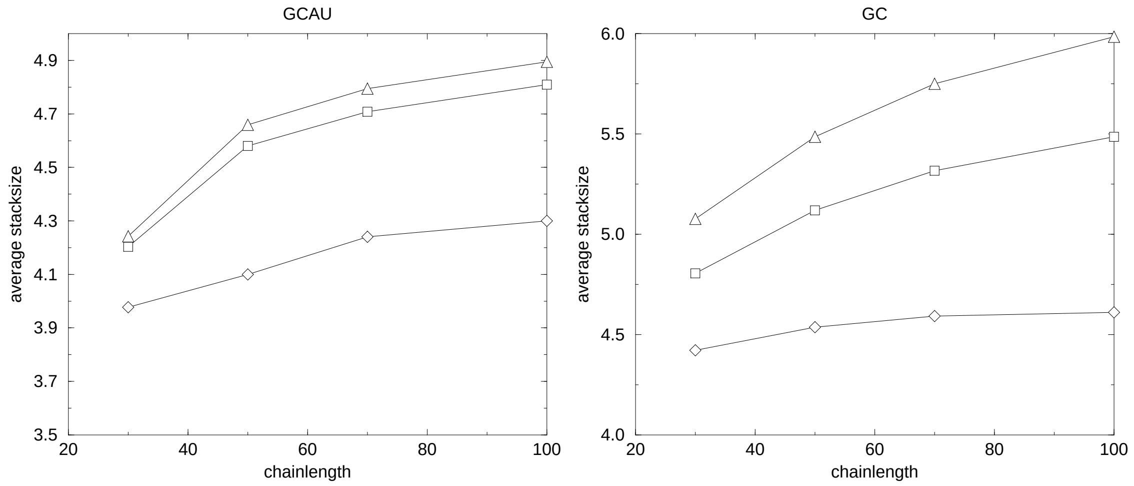


Figure 23: Mean values of stack sizes for parameter sets (i) □, (ii) ◇ and (iii) △

6.2.4 Stack Size

In all cases the mean stack size converges to a constant value although calculations for longer chain lengths would be needed. For both alphabets parameter set (ii) produces the smallest mean stack sizes but with a faster convergence, whereas parameter set (i) resembles more the plot without pseudo knots (but with smaller mean stack sizes).

6.2.5 Loop Size

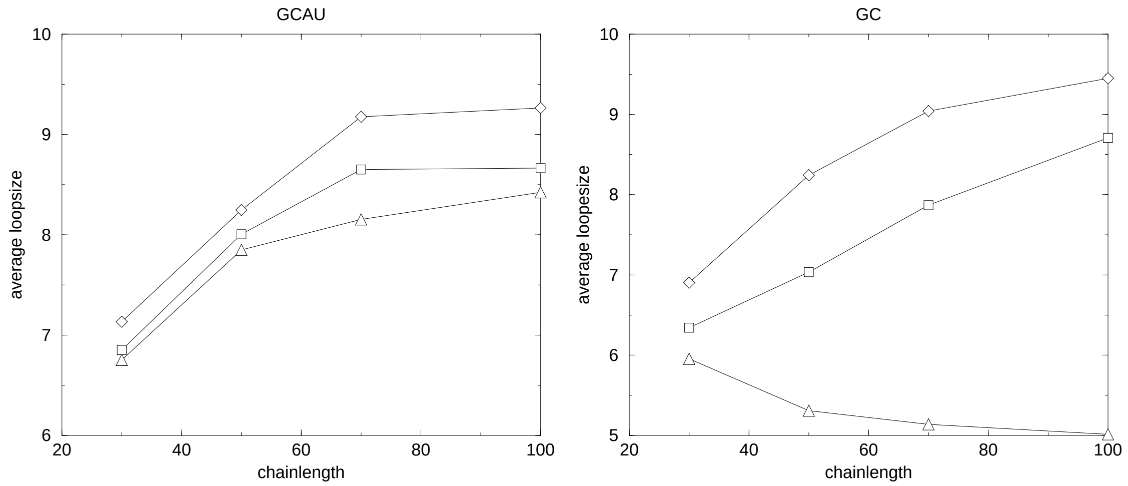


Figure 24: Mean values of loop sizes for parameter sets (i) $\square$, (ii) $\diamond$ and (iii) $\triangle$

The mean loop size converges to a constant value for all alphabets and parameters. Mean loop sizes for structures without pseudoknots are particularly for GC sequences much smaller than in all other cases. The biggest mean loop sizes are produced by parameter set (ii). Except GC parameter set (iii) all mean loop sizes increase with chain length.

6.2.6 Frequencies of Structures

As we have shown in section 6.2 the number of possible bi-secondary structures is much smaller than the number of sequences for any sequence length n. Our estimation counts the number of syntactically admissible structures irrespective of their stability. Since many sequences must fold into identical structures, the question arises how these relatively few structures are distributed over sequences. To determine the frequency distribution of secondary structures one can fold large pools of random sequences, sort the resulting structures by frequency and plot the rank of each structure versus its frequency. Using the secondary structure at full resolution this can be done only for very short (≤ 40) sequences since for longer sequences one will not find any identical structures in the sample. For longer sequences we can only study the frequencies of more coarse grained structures. We used the so called loop structure. It is obtained by denoting a stack by a single vertex and omitting the unpaired bases. If we compare the frequency distribution for sequences of different lengths even at different levels of coarse graining, remarkably similar results show up, following roughly the generalized form of Zipf's law given by Mandelbrot [47, 46].

$$f(r) = a(1 + r/b)^{-c} \tag{16}$$

where r is the rank (by frequency) of the structure S and f(r) is the fraction of occurrences of S in the sample. Zipf's law was originally derived from the analysis of the frequency of words in literary texts [107] and has since been found in a variety of contexts [38]. It states that "if one takes the words making up an extended body of text and rank them by frequency of occurrence, then the rank multiplied by it's frequency of occurrence f(r) will be approximately a constant". The form given above can be derived analytically for simple models of random text [43, 11]. Zipf's law suggests that most sequences fold into few very common structures while most structures are extremely rare. In the above parameterization of Zipf's Law the exponent c describes the distribution of rare sequences, the constant b is a rough measure for the number of frequent structures, while a gives the frequency of the most common structures. The parameters b and c depend strongly on the chain length. The parameter c describing the scaling of the power law tail of the distribution decreases with chain length, indicating that a larger fraction of sequences folds into rare structures for longer chains.

Our intention was to investigate if energy functions dealing with bi-secondary structures follow the some law. We calculated frequency distributions at full resolution (n=30) and coarse grained structure (n=70,100) for GCAU and GC alphabets. Again parameter sets (i)-realistic parameters, (ii)-nonrealistic parameters and (iii)-no pseudo knots, were compared.

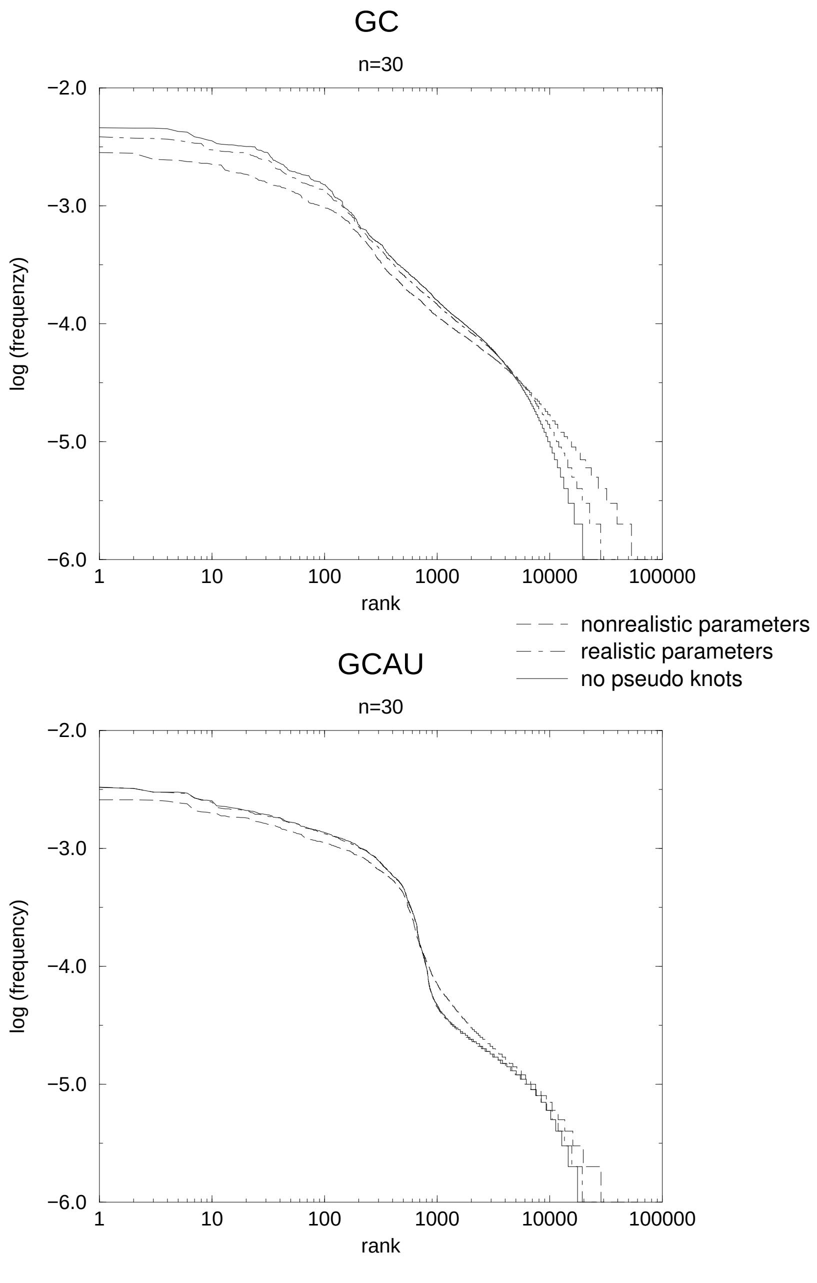


Figure 25: Zipf plot n=30

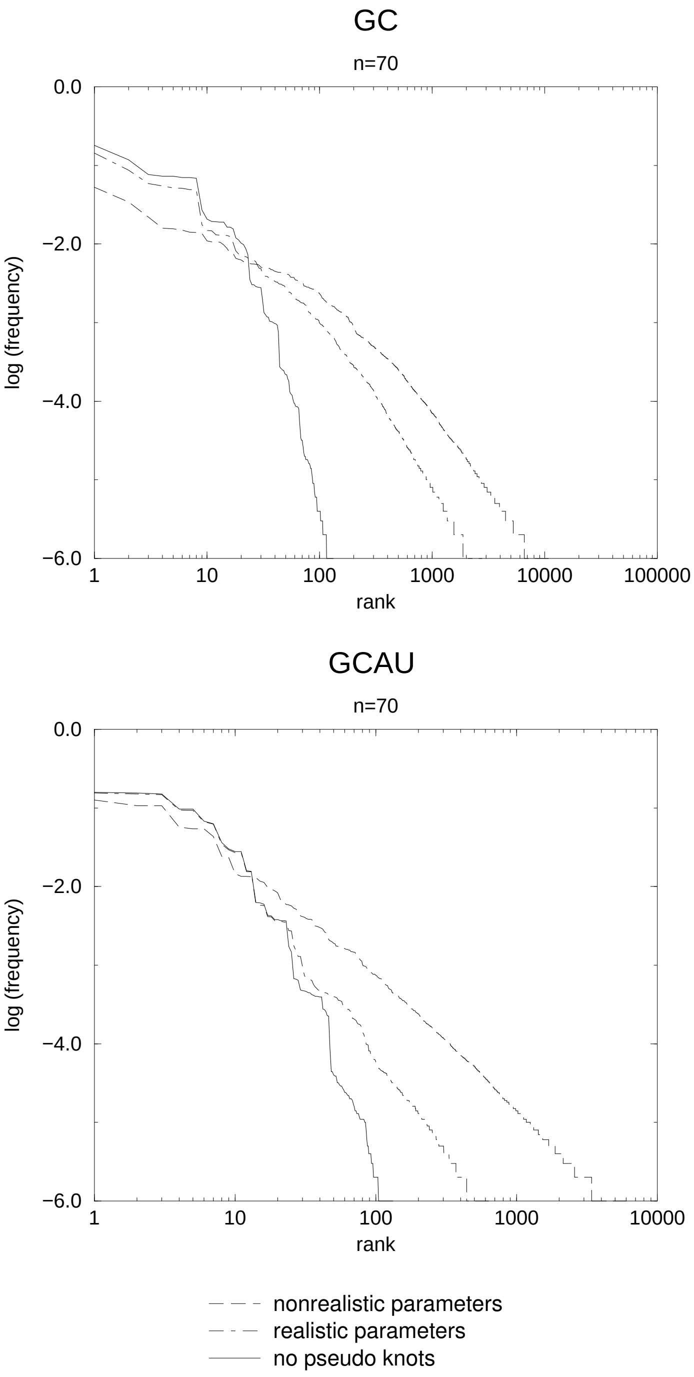


Figure 26: Zipf plot n=70

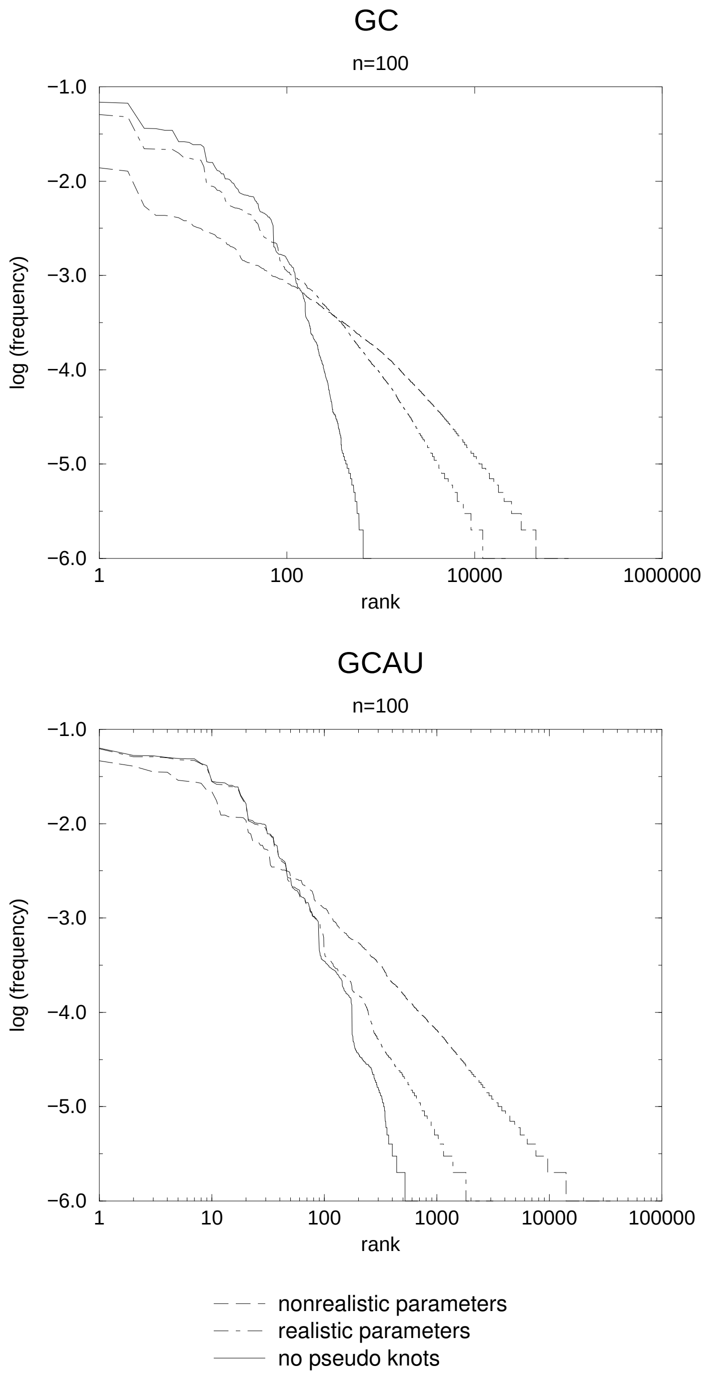
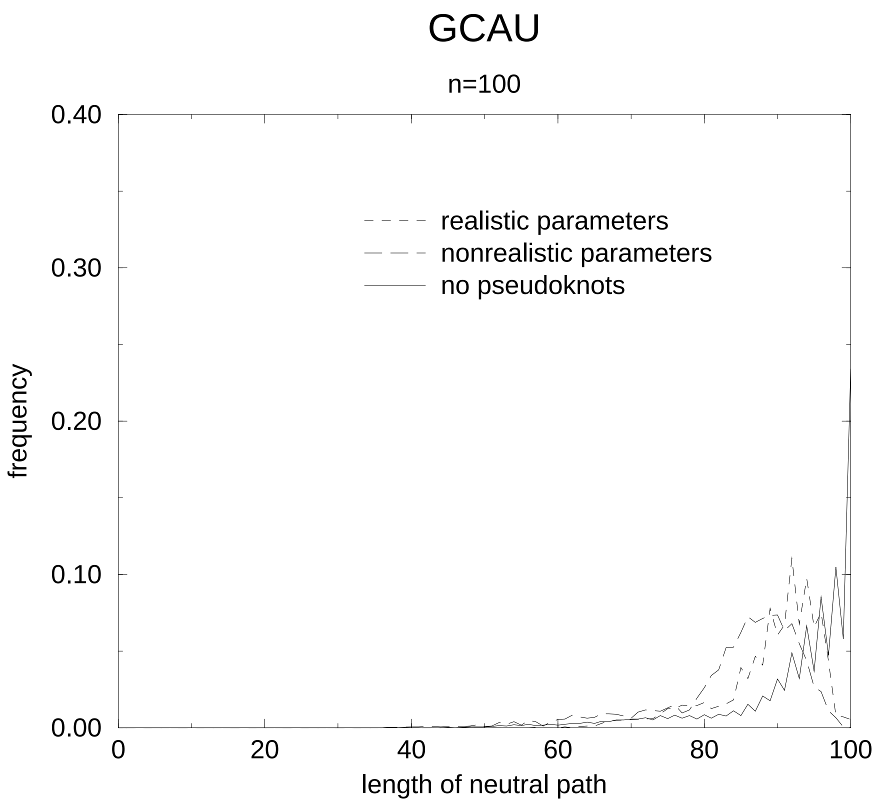


Figure 27: Zipf plot n=100

The results can be interpreted if we consider the different ratios of structures to sequences. If we allow pseudoknot formation to occur very easily, the ratio becomes bigger and the maximum rank grows. The same holds true if we use a two letter alphabet instead of a four letter alphabet. In the first case the accessible structure space grows (because of the pseudoknots) in the second case the sequence space shrinks. In all cases the notion of Zipf's law obtained for secondary structures is maintained.

6.2.7 Neutral Networks

A question related to the relative frequencies of structures is their special distribution over sequence space. RNA folding maps show a characteristic feature called neutrality which has been studied in detail for secondary structure.[26, 65] Neutrality means that there are extended nets of connected sequences perlocating the sequence space which are folding into the same common structure. A sequence is connected to another sequences if the Hamming distance between them is one (point mutation) ore two (compensatory mutation of a base pair). Based on neutral networks a model of evolutionary adaptation can be proposed. Because of the conserved secondary structure on the neutral net, fitness values do not change. Therefore random walks along the neutral net can be performed until a point is reached where a better secondary structure can be obtained within a few mutations. To investigate how far neutral nets extend in sequence space we implemented the following algorithm: Starting from a random initial sequence I_0 we constructed a monotonously diverging "neutral path " by mutating our test sequence I_n, accepting the mutated sequence I_{n+1} if the mutation is neutral $S(I) + S(I_o)$ and the Hamming distance does not decrease $d(I(_{n+1}), I_o) \geq d(I_n, I_o)$. As mutations we again allow the exchange of a single unpaired base or to exchange two bases paired in the reference structure. The length L of a path is the Hamming distance between the reference sequence and the last sequence, and hence lower bound on the diameter of the connected neutral network. Clearly, a neutral path cannot be longer than the chain length, $L \leq n$.



As we can see the length of the neutral path decreases if pseudoknots are allowed. This is due to the fact that much more sequences are accessible. Therefore mutations can lead easier to new structures as it is in the case for secondary structures.

7 Conclusion

Secondary structures form a particular class of contact structures. In this work we have considered a natural generalization of this class. Indeed, most known RNA structures with pseudoknots are bi-secondary structures (which do not involve nested pseudoknots). Bi-secondary structures correspond to planar graphs while secondary structures form the sub-class of outerplanar graphs.

The inconsistency graph introduced in section 3.2 is a useful construction capturing most of the geometrical features of nucleic acid structure. Its chromatic number may serve as a measure of structural complexity. It seems possible that an analogous construction will be useful for classifying and comparing protein structures as well.

The analysis of graph-theoretical properties of classes of contact structures is also useful for designing energy models. In order to understand the sequence-structure mapping of a class of biopolymers it is necessary to have bounds on the number of structures that can possibly be formed for a given set of sequences. While the number of possible contact structures grows faster than exponentially with the length of the molecules we find exponential upper bounds when the structural complexity is limited. In particular, there are not more than some 4.7^n possible bi-secondary structures. If we enforce in addition the sterical (loop-length at least 3) and thermodynamic (no isolated base pairs) constraints of natural RNA sequences, then this bound drops to 3.42^n. Exhaustive enumeration indicates that the actual number of bi-secondary structures with biophysical constraints grows roughly as 2.35^n. Therefore the number of sequences, 4^n , exceeds by far the number of possible bi-secondary structures.

The energy model for bi-secondary structures introduced in section 4 tries to incorporate sterical considerations. RNA stacks are viewed as stiff rods whereas unpaired regions are assumed to be very flexible. Three parameters were used to quantify pseudoknot producing loop formation. The exact parameter values were adjusted with the help of known bi-secondary structures. It turned out, that our energy model is capable to predict short range pseudoknots (H-type) in a quite satisfying way. With long range pseudoknots the situation is different, here the structure prediction is by far insufficient. This may be due to sterical interactions with proteins or other parts of the RNA molecule. It has to be considered that a relatively simple kinetic folding algorithm was used. The prediction accuracy could be improved if the energy function is implemented in a branch&bound search algorithm. This would in turn give a better parameter adjustment. The statistic of bi-secondary structure elements shows generic features which resemble those obtained from secondary structures.

- The number of base pairs, loops as well as stacks (including pseudoknots) scale linearly with the chain length.
- Mean stack size and mean loop size become a constant for large chain length.

The frequency of bi-secondary structures also follow the generalized Zipf low. Although the possibility of pseudoknot formation increases the maximum rank significantly, because the shape space gets bigger. The enlarged shape space also causes the neutral paths to be shorter compared to pure secondary structure space. From the existence of such neutral networks one can expect far reaching consequences for evolutionary optimization where the fitness depends structure. Given a suitable error frequency an evolving population should perform a random walk along the neutral net, until it reaches a point where a better secondary structure can be reached within a few mutations (i.e. a neutral net with higher fitness comes sufficiently near). During the time where the population diffuses on the neutral net, only the phenotype is conserved while genotypic information is unstable.

Open Questions and Outlook

It remains to be investigated if the enlarged shape space and the subsequently shortened neutral paths, changes significantly our notion of evolutionary optimization on neutral networks. Therefore it would be necessary to develop an "inverse folding" algorithm. With the help of this algorithm the so called shape space covering could be studied. The shape space covering radius is the radius a ball in sequence space must have to contain the most common structures. Another important task for the future is the implementation of a more sophisticated folding algorithm. As pointed out above, a suitable branch&bound algorithm would also improve the parameter adjustment of our energy function.

References

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

[2] S. Altman, L. Kirsebom, and S. Talbot. Recent studies of ribonuclease P. *FASEB J.*, 7:7–14, 1993.

[3] V. P. Antao and I. Tinoco. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetralooops. *Nucl.Acid.Res*, 20(4):819–824, 1992.

[4] S. Bonhoeffer, J. S. McCaskill, P. F. Stadler, and P. Schuster. RNA multistructure landscapes. a study based on temperature dependent partition functions. *Eur.Biophys.J.*, 22:13–24, 1993.

[5] I. Brierley, N. J. Rolley, A. J. Jenner, and S. C. Inglis. Mutational analysis of the rna pseudoknot component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol*, 229:889–902, 1991.

[6] J. W. Brown. Structure and Evolution of Ribonuclease P RNA. *Biochemie*, 73:689–697, 1991.

[7] J. W. Brown. The ribonuclease p database. *Nucleic Acids Research*, 25:263–264, 1997.

[8] M. Chamorro, N. Parkin, and H. E. Varmus. An RNA Pseudoknot and an Optimal Heptameric Shift Site Are Required for Highly Efficient Ribosomal Frameshifting on a Retroviral Messenger RNA. *Proc Natl Acad*, 89:713–717, 1992.

[9] G. Chartrand and F. Harary. Planar permutation graphs. *Ann. Inst. Henri Poincarè B*, 3:433–438, 1967.

[10] A. Chauhan and D.Apirion. The gene for a small stable rna (10sa rna) of eschericia coli. *Mol. Microbiol*, 3:1481–1485, 1989.

[11] Y. Chen. Zipf's law in text modeling. *Int. J. General Systems*, 15:232, 1989.

[12] Y. Colin de Verdière. Sur un novel invariant des graphes et un critère de planarité. *J. Comb. Theory B*, 50:11–21, 1990.

[13] E. B. T. Dam, C. W. A. Pleij, and L. Bosch. RNA Pseudoknots and Translational Frameshifting on Retroviral, Coronaviral and Luteoviral RNAs. *Virus Genes*, 4:121–136, 1990.

[14] S. C. Darr, J. W. Brown, and N. R. Pace. The varieties of ribonuclease P. *Trends Biochem. Sci.*, 17:178–182, 1992.

[15] J. D. Dinman, T. Icho, and R. B. Wickner. A -1 Ribosomal Frameshifting in a Double-stranded RNA Virus of Yeast Forms a Gag-Pol Fusion Protein. *Proc Natl Acad Sci U S A*, 88:174–178, 1991.

[16] G. A. Dirac. A property of 4-chromatic graphs and some remarks on critical graphs. *J. London Math. Soc*, 27:85–92, 1952.

[17] B. Felden, H. Himeno, A. Muto, J. McCutcheon, J. Atkins, and R. Gesteland. Probing the structure of the Escherichia coli 10Sa RNA (tmRNA). *RNA*, 3:89–103, 1997.

[18] W. Fontana, T. Griesmacher, W. Schnabl, P. F. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Mh. Chem.*, 122:795–819, 1991.

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

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

[21] A. C. Forster and S. Altman. Similar Cage-shaped Structures for the RNA Component of All Ribonuclease P and Ribonuclease MRP Enzymes. *Cell*, 62:407–409, 1990.

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

[23] D. R. Gallie, J. N. Feder, R. T. Schmike, and V. Walbot. Functional Analysis of the Tobacco Mosaic Virus tRNA-like Structure in Cytoplasmic Gene Regulation. *Nucleic Acids*, 19:5031–5036, 1991.

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

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

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

[27] A. P. Gultyaev. The computer simulation of RNA folding involving pseudoknot formation. *Nucl. Acid. Res.*, 19(9):2489 – 2494, 1991.

[28] R. Gutell. Comparative studies of RNA: Inferring higher order structure from patterns of sequence variation. *Current Opinion in Structural Biology*, 3:313, 1993.

[29] E. S. Haas, D. P. Morse, J. W. Brown, J. F. Schmidt, and N. R. Pace. Long-range Structure in Ribonuclease P RNA. *Science*, 254:853–856, 1991.

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

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

[32] I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 1996. submitted, SFI preprint 94-04-026.

[33] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucleic acids research*, 12:67–74, 1984.

[34] D. A. Holton and J. Sheehan. *The Petersen Graph*, volume 7 of *Australian Mathematical Society Lecture Series*. Cambridge University Press, Cambridge, UK, 1993.

[35] W. N. Hsieh. Proportions of irreducible diagrams. *Studies in Appl. Math.*, 52:277–283, 1973.

[36] M. A. Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43:165–169, 1996.

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

[38] H. Ishii. The distribution of duplicate books in university libraries and its relationship to Zipf's Law. *Toshokan Gakki nenpo (Annals of Japan Society)*, 36(3):97, 1990.

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

[40] D. Kleitman. Proportions of irreducible diagrams. *Studies in Appl. Math.*, 49:297–299, 1970.

[41] R. Kole, M. F. Baer, B. C. Stark, and S. Altman. E. coli RNase P has a required RNA component in vivo. *Cell*, 19:881–887, 1980.

[42] K. Kuratowski. Sur le problème des courbes gauches en topologie. *Fund. Math.*, 15:271–283, 1930.

[43] W. Li. Random texts exhibit Zipf's-law-like word frequency distribution. Technical Report 91-03-016, Santa Fe Institute, 1991.

[44] L. Lovász and A. Schrijver. The Colin de Verdière number of linklessly embeddable graphs. preprint, 1996.

[45] J. M. M. Wu and D. H. Turner. A periodic table of symmetric tandem mismatches in rna. *Biochemistry*, volume 34::2304–11, 1995.

[46] B. B. Mandelbrot. An information theory of the statistical structure of language. In *Proceedings of the Symposium on Applications of Communications Theory*, London, 1953. Butterworths.

[47] B. B. Mandelbrot. *The Fractal Geometry of Nature*. Freeman & Co., New York, 1983.

[48] R. Mans, C. Pleij, and L. Bosch. Transfer RNA-like Structures: Structure, Function and Evolutionarz Significance. *Eur J Biochem*, 201:303–324, 1991.

[49] R. Mans, M. H. V. Steeg, P. Verlaan, C. Pleij, and L. Bosch. Mutational Analysis of the Pseudoknot in the tRNA-like Structure of Turnip Yellow Mosaic Virus RNA. Aminoacylation Efficiency and RNA Pseudoknot Stability. *J Mol Biol*, 223:221–232, 1992.

[50] H. M. Martinez. An RNA folding rule. *Nucl. Acid. Res.*, 12:323–335, 1984.

[51] F. Michel and E.Westhof. Modelling of the Three-dimmensional Architecture of Group I Catalytic Introns Based on Comparative Sequence Anaysis. *J Mol Biol*, 216:585–610, 1990.

[52] D. Moazed and H. F. Noller. Transfer RNA Shields Specific Nucleotides in 16S Ribosomal RNA from Attack by Chemical Probes. *Proc Natl Acad Sci U S A*, 47:985–994, 1986.

[53] S. Morse and D. E. Draper. Purine-purine mismatches in rna helices: evidence for protona ted ga pairs and next-nearest neighbor effects. *Nucleic Acids Res.*, 23::302–6, 1995.

[54] H. F. Noller, W. Hoffarth, and L. Zimiak. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science*, 256:1416–1419, 1992.

[55] C. Papanicolau, M. Gouy, and J. Ninio. An energy model that predicts the correct folding of the tRNA and the 5S RNA molecules. *Nucl. Acid. Res.*, 12:31–44, 1984.

[56] R. C. Penner and M. S. Waterman. Spaces of RNA secondary structures. *Adv. Math.*, 101:31–49, 1993.

[57] A. E. Peritz, R. Kierzek, N. Sugimoto, and D. H. Turner. Thermodynamic study of internal loops in oligonucleotides: Symmetric loops are more stable than assymetric loops. *Biochemistry*, 30:6428–36, 1991.

[58] C. Philippe, C. Portier, M. Mougel, M. Grunberg-Manago, J. P. Ebel, B. Ehresmann, and C. Ehresmann. Target site of escherichia coli ribosomal protein S15 on its messenger RNA. *J Mol Biol*, 211:415–426, 1990.

[59] C. W. A. Pleij. Pseudoknots a New Motiv in the RNA Game. *Trends Biochem Sci*, 15:143–147, 1990.

[60] T. Powers and H. F. Noller. A Functional Pseudoknot in 16S Ribosomal RNA. *EMBO*, 10:2203–2214, 1991.

[61] J. D. Puglisi, J. R. Wyatt, and I. Tinocco. RNA Pseudoknots. *Acc Chem Res*, 24:152–158, 1991.

[62] A. L. N. Rao, T. W. Dreher, L. E. Marsch, and T. C. Hall. Telomeric Function of the tRNA-like Structure of Brome Mosaic Virus RNA. *Proc Natl Acad Sci*, 86:5335–5339, 1989.

[63] R. E. Reed, M. F. Baer, C. Guerrier-Takada, H. Donis-Keller, and S. Altman. Nucleotide sequence of the gene encoding the RNA subunit (M1 RNA) of ribonuclease P from Escherichia coli. *Cell*, 30:627–636, 1982.

[64] C. Reich, K. J. Gardiner, G. J. Olsen, B. Pace, T. L. Marsh, and N. R. Pace. The RNA component of the Bacillus subtilis RNase P: sequence, activity, and partial secondary structure. *J. Biol. Chem.*, 261:7888–7893, 1986.

[65] C. Reidys and P. F. Stadler. Bio-molecular shapes and algebraic structures. *Comp. & Chem.*, 20:85–94, 1996. SFI preprint 95-10-098.

[66] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[67] N. Robertson, P. Seymore, and R. Thomas. Petersen family minors. *J. Comb. Theory B*, 64:155–184, 1995.

[68] N. Robertson, P. Seymore, and R. Thomas. Sachs' linkless embedding conjecture. *J. Comb. Theory B*, 64:185–227, 1995.

[69] N. R.Pace and D. Smith. Ribonuclease P: function and variation. *J. Biol. Chem.*, 265:3587–3590, 1990.

[70] T. B. S. Ebel and A. N. Lane. Thermodynamic stability and solution conformation of tandem g a mismatches in rna and rna.dna hybrid complexes. *Eur. J. Biochem.*, 220::703–15, 1994.

[71] W. Saenger. *Principles of Nucleic-Acid Structure.* Springer-Verlag, New York, first edition, 1984.

[72] W. Salser. Globin messenger RNA sequences - analysis of base-pairing and evolutionary implications. *Cold Spring Harbour Symp. Quant. Biol.*, 42:985, 1977.

[73] P. Schimmel. RNA Pseudoknots that Interact with Components of the Translation Apparatus. *Cell*, 58:9–12, 1989.

[74] W. R. Schmitt and M. S. Waterman. Linear trees and RNA secondary structure. *Discr. Appl. Math.*, 12:412–427, 1994.

[75] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *Journal of Biotechnology*, 41:239–257, 1995.

[76] 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.

[77] M. J. Serra, T. J. Axenson, and D. H. Turner. A model for the stabilities of rna hairpins based on a study on the sequence dependence of stability for hairpins of six nucleotides. *Biochemystry*, 33::14289–965., 1994.

[78] M. J. Serra, M. H. Lyttle, T. J. Axenson, C. A. Schadt, and D. H. Turner. Rna hairpin loop stability depends on the closing base pair. *Nucleic Acids Res.*, 21::3845–9, 1993.

[79] F. Söler and K. Jankowski. Modeling RNA secondary structures I. Mathematical structural model of predicting RNA secondary structures. *Math. Biosc.*, 105:167–190, 1991.

[80] P. R. Stein. On a class of linked diagrams, I. Enumeration. *J. Comb. Theory A*, 24:357–366, 1978.

[81] P. R. Stein and C. J. Everett. On a class of linked diagrams. II. Asymptotics. *Disc. Math.*, 22:309–318, 1978.

[82] P. R. Stein and M. S. Waterman. On some new sequences generalizing the Catalan and Motzkin numbers. *Disc. Math.*, 26:261–272, 1978.

[83] W. Stockmayer and H. Jacobson. *J. Chem. Phys.*, 18:1600–1606, 1950.

[84] R. H. Symons. Ribozymes. *Curr. Opin. Struct. Biol.*, 4:322–330, 1994.

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

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

[87] C. K. Tang and D. E. Draper. An unusual mRNA pseudoknot structure is recognized by a protein translation repressor. *Cell*, 57:531–536, 1989.

[88] C. K. Tang and D. E. Draper. Evidence for allosteric coupling between the ribosome and repressor binding sites of a translationally regulated mRNA. *Biochemistry*, 29:4434–4439, 1990.

[89] E. ten Dam, I. Brierly, S. Inglis, and C. Pleij. Identification and analysis od the pseudoknot-containing *gag-pro* ribosomal frameshift signal of simian retrovirus-1. *Nucl. Acids Res.*, 22:2304–2310, 1994.

[90] J. Touchard. Sur une problème de configurations et sur les fractions continues. *Canad. J. Math.*, 4:2–25, 1952.

[91] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 17:167–192, 1988.

[92] T. H. Tzeng, C. L. Tu, and J. A. Bruenn. Ribosomal Frameshifting Requires a Pseudoknot in the Saccharomyces cerevisiae Double-stranded RNA Virus. *J Virus*, 66:999–1006, 1992.

[93] V. V. Vlassov, G. Zuber, B. Felden, J. P. Behr, and R. Griege. Cleavage of trna with imidazole and spermine imidazole constructs: A new approach for probing rna structures. *Nucleic Acids Res*, 23:3161–3167, 1995.

[94] K. Wagner and R. Bodendiek. *Graphentheorie II.* B.I. Verlag, Mannheim, Germany, 1990.

[95] A. E. Walter, D. H. Turner, J. Kim, M. Lyttle, P. Muller, D. H. Mathews, and M. Zucker. Coaxial stacking of helices enhances binding of oligoribunucleotides and improves prediction of rna folding. *Proc. Natl Acad Sci.*, 91::9218–22, 1994.

[96] A. E. Walter, M. Wu, and D. Turner. The stability and structure of tandem g-a mismatches in rna depends on closing base pairs. *Biochemistry*, 33::9218–22, 1994.

[97] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Studies*, 1:167 – 212, 1978.

[98] M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences, and Genomes.* Chapman & Hall, London, 1995.

[99] M. S. Waterman and T. F. Smith. Combinatorics of RNA hairpins and cloverleaves. *Studies Appl. Math.*, 60:91–96, 1978.

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

[101] J. Weber. *Dynamics on Neutral Evolution.* PhD thesis, Friedrich Schiller University, Jena, January 1997.

[102] A. M. Weiner and N. Maizels. tRNA-like Structures Tag the 3' ends of Genomic RNA Molecules for Replication: Implications for the Origin of Protein Sybthesis. *Proc Natl Acad Sci*, 84:7383–7387, 1987.