Rep. Prog. Phys. **69** (2006) 1419–1477 doi:10.1088/0034-4885/69/5/R04

# Prediction of RNA secondary structures: from theory to models and real molecules

**Peter Schuster**[1,2]

[1]Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, A-1090 Vienna, Austria
[2]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

E-mail: pks@tbi.univie.ac.at

Received 8 January 2006, accepted for publication 22 March 2006
Published 18 April 2006
Online at stacks.iop.org/RoPP/69/1419

## Abstract

RNA secondary structures are derived from RNA sequences, which are strings built form the natural four letter nucleotide alphabet, {**AUGC**}. These coarse-grained structures, in turn, are tantamount to constrained strings over a three letter alphabet. Hence, the secondary structures are discrete objects and the number of sequences always exceeds the number of structures. The sequences built from two letter alphabets form perfect structures when the nucleotides can form a base pair, as is the case with {**GC**} or {**AU**}, but the relation between the sequences and structures differs strongly from the four letter alphabet. A comprehensive theory of RNA structure is presented, which is based on the concepts of *sequence space* and *shape space*, being a space of structures. It sets the stage for modelling processes in ensembles of RNA molecules like evolutionary optimization or kinetic folding as dynamical phenomena guided by mappings between the two spaces.

The number of minimum free energy (mfe) structures is always smaller than the number of sequences, even for two letter alphabets. Folding of RNA molecules into mfe energy structures constitutes a non-invertible mapping from sequence space onto shape space. The preimage of a structure in sequence space is defined as its *neutral network*. Similarly the set of *suboptimal structures* is the preimage of a sequence in shape space. This set represents the *conformation space* of a given sequence. The evolutionary optimization of structures in populations is a process taking place in sequence space, whereas kinetic folding occurs in molecular ensembles that optimize free energy in conformation space. Efficient folding algorithms based on dynamic programming are available for the prediction of secondary structures for given sequences. The inverse problem, the computation of sequences for predefined structures, is an important tool for the design of RNA molecules with tailored properties. Simultaneous folding or *cofolding* of two or more RNA molecules can be modelled readily at the secondary structure level

This article was invited by Professor T J Newman.

and allows prediction of the most stable (mfe) conformations of complexes together with suboptimal states. Cofolding algorithms are important tools for efficient and highly specific primer design in the polymerase chain reaction (PCR) and help to explain the mechanisms of small interference RNA (si-RNA) molecules in gene regulation.

The evolutionary optimization of RNA structures is illustrated by the search for a target structure and mimics aptamer selection in evolutionary biotechnology. It occurs typically in steps consisting of short adaptive phases interrupted by long epochs of little or no obvious progress in optimization. During these quasi-stationary epochs the populations are essentially confined to neutral networks where they search for sequences that allow a continuation of the adaptive process. Modelling RNA evolution as a simultaneous process in sequence and shape space provides answers to questions of the optimal population size and mutation rates.

Kinetic folding is a stochastic process in conformation space. Exact solutions are derived by direct simulation in the form of trajectory sampling or by solving the master equation. The exact solutions can be approximated straightforwardly by Arrhenius kinetics on barrier trees, which represent simplified versions of conformational energy landscapes. The existence of at least one sequence forming any arbitrarily chosen pair of structures is granted by the *intersection theorem*. Folding kinetics is the key to understanding and designing multistable RNA molecules or *RNA switches*. These RNAs form two or more long lived conformations, and conformational changes occur either spontaneously or are induced through binding of small molecules or other biopolymers. RNA switches are found in nature where they act as elements in genetic and metabolic regulation.

The reliability of RNA secondary structure prediction is limited by the accuracy with which the empirical parameters can be determined and by principal deficiencies, for example by the lack of energy contributions resulting from tertiary interactions. In addition, native structures may be determined by folding kinetics rather than by thermodynamics. We address the first problem by considering base pair probabilities or base pairing entropies, which are derived from the partition function of conformations. A high base pair probability corresponding to a low pairing entropy is taken as an indicator of a high reliability of prediction. Pseudoknots are discussed as an example of a tertiary interaction that is highly important for RNA function. Moreover, pseudoknot formation is readily incorporated into structure prediction algorithms.

Some examples of experimental data on RNA secondary structures that are readily explained using the landscape concept are presented. They deal with (i) properties of RNA molecules with random sequences, (ii) RNA molecules from restricted alphabets, (iii) existence of neutral networks, (iv) shape space covering, (v) riboswitches and (vi) evolution of non-coding RNAs as an example of evolution restricted to neutral networks.

## Contents

## 1. Introduction

Natural ribonucleic acid (RNA) molecules are heteropolymers with a regular backbone built from four classes of monomers. The backbone is polar in the sense that it has two chemically different ends. The monomer unit consist of a purine base—**A**(denine) or **G**(uanine)—or a pyrimidine base—**U**(racil) or **C**(ytosine)—a ribose unit and a phosphate unit (figure 1). In the closely related deoxyribonucleic acid (DNA) molecules ribose is replaced by 2′-deoxyribose and **U**(racil) by **T**(hymine). Other purine and pyrimidine bases appear occasionally in natural nucleic acid molecules too. As shown in figure 1 a string containing the sequence of nucleotides, commonly called the *primary structure*, provides all the information for a reconstruction of the chemical formula of an RNA molecule since (i) the ribose-phosphate or 2′-deoxyribose-phosphate backbone is periodic and the same in all RNA or DNA molecules of the same chain length and (ii) the two different ends of nucleic acids are distinguished by the convention that the 5′-end coincides with the beginning (left-hand end) and the 3′-end with the end (right-hand end) of the string.

Nucleic acid structures are commonly classified as foldings of single molecules or cofoldings of two or more molecules. A duplex formed from complementary strands, the natural form of DNA, is the simplest and best known cofolded structure. The stability of the duplex structure is based on the perfect fit of $\mathbf{G}\equiv\mathbf{C}$ and $\mathbf{A}{=}\mathbf{U}$ (in RNA) or $\mathbf{A}{=}\ \mathbf{T}$ (in DNA) base pairs into the Watson–Crick double helix structure. Most of the RNA structures known at atomic resolution are single molecule folds. These structures contain Watson–Crick type helices formed by pairing complementary regions of the sequence running in opposite directions (figure 2). The repertoire of acceptable pairings between the natural nucleotide bases (**A**, **U**, **G** and **C**) is richer in single molecule folds of RNA than in DNA duplexes: the $\mathbf{G}{-}\mathbf{U}$ base pair is also admitted. The conventional representation of RNA secondary structures is a planar graph where the nodes are the individual nucleotides and the edges are the connections between neighbours from the backbone and the base pairs. In general, a secondary structure can be understood as a listing of base pairs, and this is illustrated by different representations.

RNA secondary structures are unique among biopolymer structures because they have a physical meaning as folding intermediates of RNA 3D structures [1, 2] and are accessible to mathematical analysis since secondary structure formation follows simple combinatorial rules. The discreteness of secondary structures is an advantage for the analysis of mappings between sequence and shape space. Methods for the prediction of RNA secondary structures from known sequences were developed in the 1980s, and it turned out that structures can be derived by means of dynamic programming at relatively low cost [3–5]. The first algorithm aiming at energy optimization encapsulated in the search for the structure of minimal free energy (mfe) followed soon afterwards [6]. Such computations are based on empirical free energy parameters determined from thermodynamic data of small RNA model compounds. Since the beginning of RNA secondary structure predictions the parameter sets were regularly updated as more reliable data on natural and synthesized RNA molecules became available (for the most frequently used data collections on RNA see [7–9]). Computation of single stranded DNA folding follows the same principles, and free energy parameters are available for DNA single strand folding as well as for DNA duplex formation [10, 11].

The inverse problem of RNA folding, the task to finding an RNA sequence that forms a given RNA secondary structure of mfe, has been solved in the 1990s using an iterative algorithm [12]: Starting from a random compatible sequence—this is a sequence that has

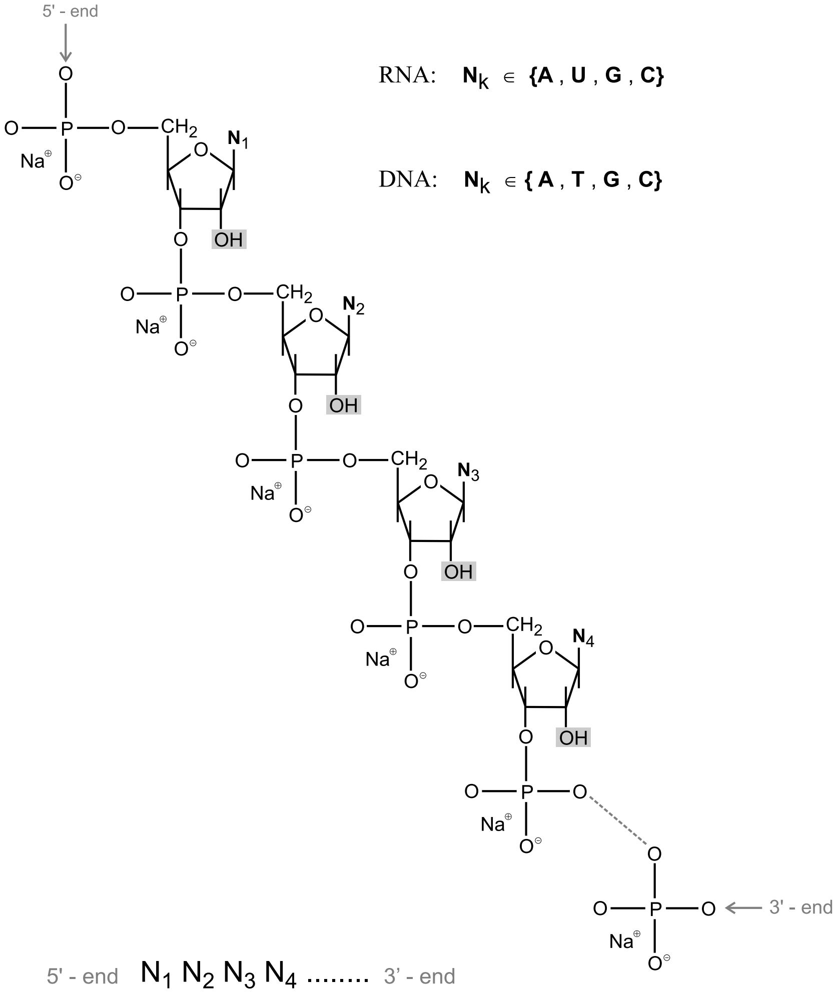

5' - end
RNA: N$_k$ ∈ {A , U , G , C}
DNA: N$_k$ ∈ { A , T , G , C}
O
O—P—O—CH$_2$
Na$^\oplus$
O$^\ominus$
N$_1$
N$_2$
N$_3$
N$_4$
O
OH
3' - end
5' - end N$_1$ N$_2$ N$_3$ N$_4$ ........ 3' - end

**Figure 1.** Chemical formula of nucleic acids. The molecule is a polymer with a regular ribose-phosphate (RNA) or 2′-deoxyribose-phosphate (DNA; the OH in the grey box is replaced by H) backbone. Nucleotide bases (**A**, **U** in RNA or **T** in DNA, **G** and **C**) are covalently attached to the 1′-position of the sugar unit. The molecule has two chemically different ends, 5′ and 3′. The sequence of nucleotides (sketched at the bottom of the figure) determines the molecule completely. The phosphate group of nucleic acids carries a negative charge at physiological pH, that is pH ≈ 7.4, and thus close to neutral pH. This charge is compensated by a counterion, commonly **Na**$^\oplus$. The counterions have a strong influence on nucleic acid structures; in particular, **Mg**$^{2\oplus}$ is often indispensable for spatial structures of RNA molecules.

pairable nucleotides[1] everywhere the structure demands a base pair- the solution is approached through coordinated sequence changes that reduce the difference between the current and the target mfe structure. Inverse folding revealed an important property of RNA structure landscapes: the mapping from sequences into structures is non-invertible since there are mfe structures that are formed by many RNA sequences. The main objective of this article is to derive and review formal concepts of RNA sequence and shape space and to apply them in

[1] These are two nucleotides that can form a base pair, in particular one combination out of the set $\mathcal{B} =$ {**AU**,**CG**,**GC**,**GU**,**UA**,**UG**} (see section 2).

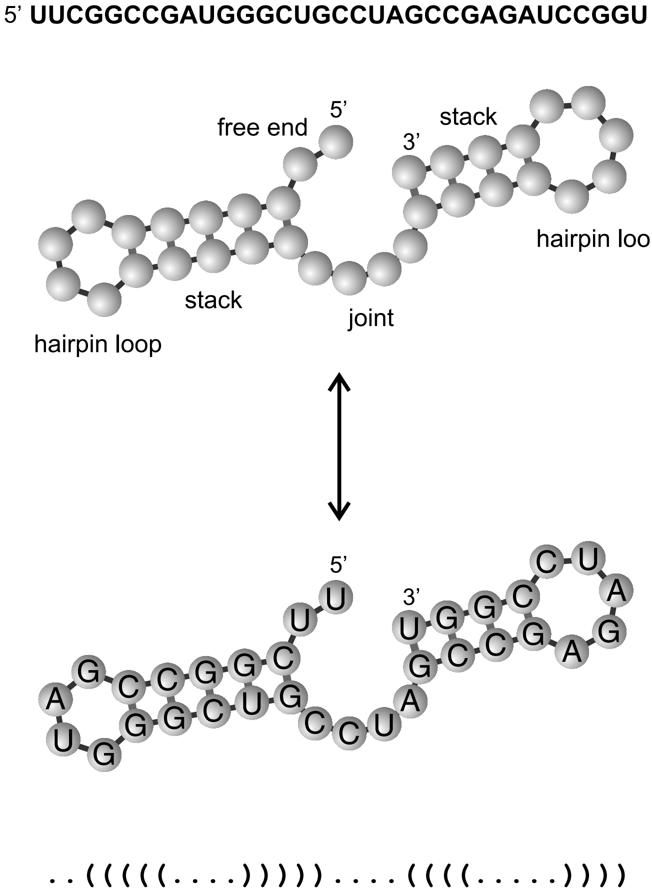


**Figure 2.** Secondary structures of RNA. The topmost part shows an RNA sequence of $n = 33$ nucleotides. Its secondary structure of minimal free energy (mfe) is shown in the middle. The structure contains two classes of nucleotides: (i) unpaired nucleotides and (ii) base pairs. The base pairs appear in *stacks* or double helical regions where the two paired strands run in opposite direction: $5' \to 3'$ and $3' \to 5'$. The nine base pairs defining the structure above are {3–16, 4–15, 5–14, 6–13, 7–12} in the left-hand stack and {21–33, 22–32, 23–31, 24–30} in the stack on the right-hand side. Thereby, the numbering of nucleotides follows a convention: No. 1 is assigned to the nucleotide at the $5'$-end and the nucleotide at the $3'$end gets no. $n$. Unpaired bases form various kinds of *loops*, hairpin loops in our example, *joints* connecting two substructures, and *free ends*. The sketch below shows an overlay of the sequence onto its mfe conformation. At the bottom of the figure we show a symbolic representation of the structure by means of three symbols: (i) 'dot' stands for a unpaired nucleotide, (ii) 'left-hand parenthesis' represents a base, which pairs with a downstream nucleotide, and (iii) 'right-hand parenthesis' completes a base pair by pairing with an upstream nucleotide. The no pseudoknot restriction (figure 3) guarantees that the mathematical rule of parentheses assignments is fulfilled in base pairing.

explanations and predictions of the properties of RNA molecules. RNA secondary structures, together with lattice protein models, are at present the only biological objects for which conformational landscapes and sequence-structure maps can be computed and analysed in sufficient detail. The concepts of mappings and landscapes turned out to be useful also for an understanding of dynamical processes in ensembles of RNA molecules like kinetic folding and evolutionary optimization.

## 2. RNA secondary structures, sequence and shape spaces

Two notions are important for the definition of secondary structures: (i) the nucleotide alphabet $\mathcal{A}$ being the set of nucleotides, $\mathcal{A} = \{\alpha_1, \ldots, \alpha_\kappa\} = \{\mathbf{A},\mathbf{U},\mathbf{G},\mathbf{U}\}$ in natural RNA molecules, and (ii) the set of accepted base pairs , $\mathcal{B} = \{\beta_1, \ldots, \beta_\varrho\}$, with $\beta_k = \alpha_i\alpha_j$ and $\mathcal{B} = \{\mathbf{AU},\mathbf{CG},\mathbf{GC},\mathbf{GU},\mathbf{UA},\mathbf{UG}\}$ being the set of allowed base pairs in RNA molecules occurring *in vivo*. We denote the size of the alphabet by $\kappa = |\mathcal{A}|$ and the number of accepted base pairs by $\varrho = |\mathcal{B}|$. As shown, we have $\kappa = 4$ and $\varrho = 6$ for natural RNA molecules. An RNA sequence is defined as a string of nucleotides chosen from an alphabet: $X = (x_1x_2 \ldots x_n)$; $x_j \in \mathcal{A}$.

### *2.1. Secondary structures*

A conventional RNA secondary structure[2] $S$ is a listing of base pairs that can be visualized by a planar graph. The nodes of the graph are nucleotides of the RNA molecule, $i \in \{1, 2, \ldots, n\}$ numbered consecutively along the chain (figure 3). The edges of the graph represent bonds between nodes which fall into two classes:
(i) the backbone, $\{i$—$(i + 1)\ \forall\, i = 1, \ldots, n - 1\}$, and (ii) the base pairs. The two ends of the sequence (5′- and 3′-ends) are chemically different. The backbone is completely defined for known $n$, and hence a secondary structure is completely determined by a listing of base pairs, $S$, where a pair between $i$ and $j$ will be denoted by $i$–$j$. For a conventional secondary structure the base pairs fulfil three conditions:

I. Binary interaction restriction. An individual nucleotide is either involved in one base pair or it is a single nucleotide forming no base pair.
II. No nearest neighbour pair restriction. Base pairs to nearest neighbours, $i$–$j$ with $j = i - 1$ or $j = i + 1$, are excluded.
III. No pseudoknot restriction. Two base pairs $i$–$j$ and $k$–$l$ with $i < j$, $i < k$ and $k < l$ are only accepted if either $i < k < l < j$ or $i < j < k < l$ is fulfilled—the second base pair is either enclosed by the first base pair or lies completely outside (figure 3).

Condition I forbids the formation of base triplets or higher interactions between nucleotides. Conditions II is required for steric reasons because stereochemistry does not allow pairing geometries between neighbouring nucleotides. As we shall see later on, the steric constraint is even more stringent in the sense that hairpin loops with fewer than three single nucleotides do not occur in real structures. Condition III is mainly a technical constraint, because the explicit consideration of pseudoknots impedes mathematical analysis of structures substantially and makes actual computations much more time consuming (see [13] and section 5.2).

The graph representation of secondary structures is fully equivalent to other representations that we shall not discuss here except two, the adjacency matrix[3]

$$A = \left\{ a_{ij} = a_{ji} = \begin{cases} 1 & \text{if and only if } i, j \in \Omega\,;\ i, j = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \right\}. \quad (1)$$

and the symbolic notation (figure 2). Throughout this review it will be convenient to identify a secondary structure by its set of base pairs Ω. More abstractly, we consider Ω as an arbitrary matching on $\{1, \ldots, n\}$. In other words we shall sometimes relax the conventional

[2] ‘Conventional’ means here that the structure is free of pseudoknots (condition III). Some other definitions include certain or all classes of pseudoknots.

[3] Here the backbone is excluded from the adjacency matrix, but it makes no difference when it is considered because the backbone does not change in superpositions of the structures discussed here.

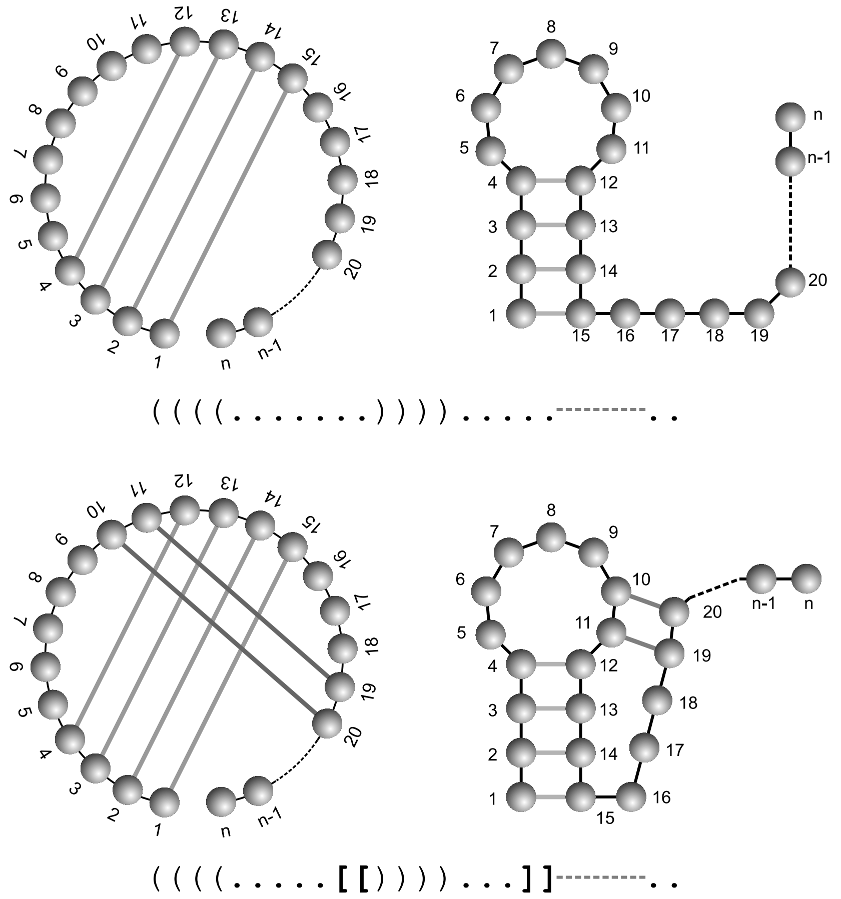


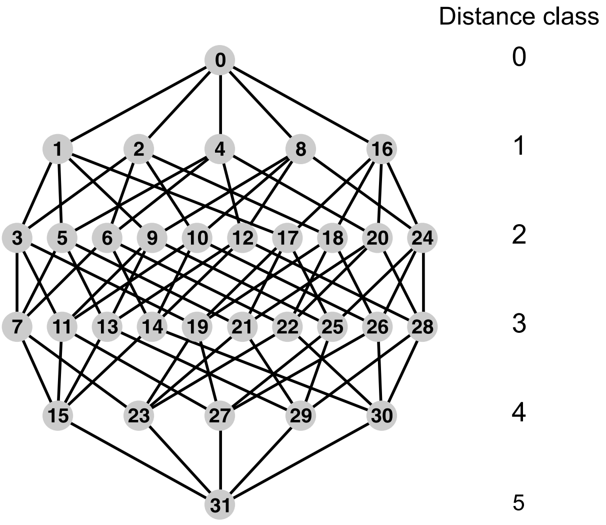
**Figure 3.** Representations of secondary structures and the no pseudoknot restriction. The upper part shows a simple hairpin loop in the circular representation corresponding to the conventional graph on the right-hand side: base pairs are represented in the circle by chords that are free of crossings by condition III. The symbolic notation is shown on the line below. The numbering of individual nucleotides and the string orientation define the two different ends uniquely: the nucleotide at the 5′-end always carries the number 1 and is positioned at the left-hand end of the string. The lower part of the figure shows an H-type pseudoknot ('H' stand for 'hairpin'): pseudoknots violate condition III and imply crossing of chords. In the symbolic notation pseudoknots require coloured parentheses in order to guarantee unique assignments.

no-pseudoknot condition III and insist only that each nucleotide takes part in at most one base pair (condition I).[4] Furthermore, let $\Upsilon$ be the set of unpaired bases, which is the subset of $\{1, \ldots, n\}$ that is not met by the matching $\Omega$. Each nucleotide of a sequence $X$ is either a single nucleotide or it takes part in a base pair and, accordingly, is uniquely assigned to one of the two sets such that the sequence fulfil: $X = \{(x_j \in \Omega \text{ Xor } x_j \in \Upsilon)\ \forall\ j = 1, \ldots, n\}$.

### 2.2. *Sequence space and shape space*

The sequence space $\mathcal{Q}_n^{(\mathcal{A})}$ is the set of all possible sequences $X_k$ of chain length $n$ over the alphabet $\mathcal{A}$. Connecting all nearest neighbours yields a simple object in $n$-dimensional space.

[4] Wherever confusion is possible we shall be precise and use '$S$' for conventional secondary structures and '$\Omega$' for the generalization.



Binary sequences are encoded by their decimal equivalents:

**C** = 0 and **G** = 1, for example,

"0" ≡ 00000 = **CCCCC,**

"14" ≡ 01110 = **CGGGC,**

"29" ≡ 11101 = **GGGCG**, etc.

**Figure 4.** Sequence space of binary sequences of chain length $n = 5$. The sequence spaces of binary sequences are hypercubes of dimension $n$. The sequences are encoded by their decimal equivalents. All sequences are equivalent in the sense that all points in sequence space are on the surface and have the same numbers of neighbours.

In the case of binary sequences, $\mathcal{A} = \{\mathbf{A},\mathbf{U}\}$ or $\mathcal{A} = \{\mathbf{G},\mathbf{C}\}$, this is a hypercube (figure 4); for three or four letters, $\mathcal{A} = \{\mathbf{A},\mathbf{U},\mathbf{G}\}$, $\mathcal{A} = \{\mathbf{U},\mathbf{G},\mathbf{C}\}$, and $\mathcal{A} = \{\mathbf{A},\mathbf{U},\mathbf{G},\mathbf{C}\}$, the sequence space is a straightforward generalization of hypercubes. The Hamming distance between two sequences $X_j$ and $X_k$, $d_H(X_j, X_k)$, defined as the number of positions in which two aligned[5] sequences differ, induces a metric in sequence space. It is commonly used in computer science and bioinformatics for the comparison of sequences. The Hamming distance is the natural distance measure for changes in sequences based on single point mutations as elementary steps or moves.[6] All points in sequence space are equivalent, in the sense that they have precisely the same numbers of neighbours (figure 4).

The shape space $\mathcal{S}_n$ is the set of all possible structures formed from sequences of chain length $n$, irrespective of whether there exists a sequence that can form the structure. The cardinality of shape space, i.e. the number of possible structures, will be evaluated by counting in section 2.3. The choice of a proper notion of distance in shape space is dependent on the move set for changes in structures. These moves have a meaning in physics when they are occurring as elementary steps in kinetic folding of RNA molecules. The simplest move set we shall choose uses base pair closing and base pair opening as the only elementary steps. Despite its simplicity the set is complete because each structure can be reached from each other structure by a series of base pair openings and closures. This simple move set corresponds to the base pair distance, $d_P(S_j, S_k)$, which counts the number of base pairs in which the two structures differ (figure 18).[7] Another definition of distance makes use of the symbolic notation of secondary structures and computes the Hamming distance between the two strings. This notion of distance corresponds to an extended move set containing base pair closing, base pair

[5] Optimal alignment of sequences is not trivial but a common and well understood problem in molecular genetics. We shall deal here only with the simple problem of end-to-end alignment of sequences with identical chain lengths.

[6] Single point mutations are sequence changes at precisely one position.

[7] In order to allow straightforward comparisons based on single nucleotide exchanges the number of different base pairs is multiplied by two.

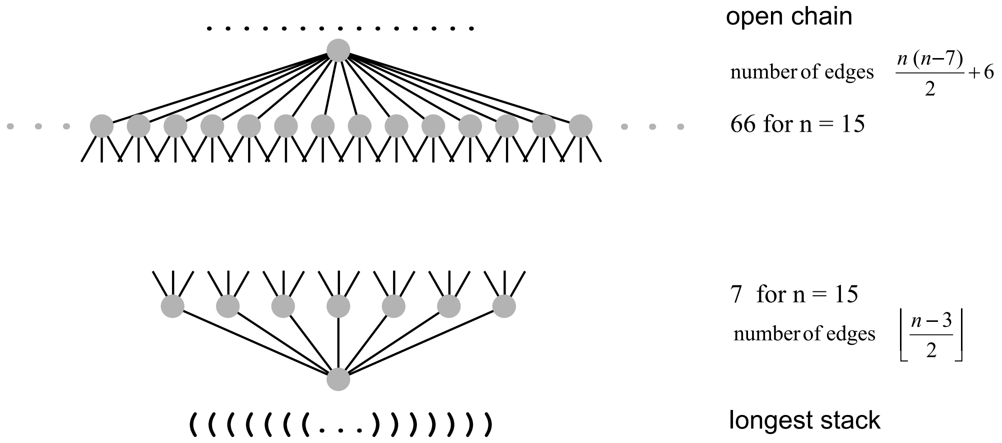

open chain
number of edges $\frac{n\,(n-7)}{2}+6$
66 for n = 15
7 for n = 15
number of edges $\left\lfloor \frac{n-3}{2} \right\rfloor$
(((((((...)))))))
longest stack

**Figure 5.** Shape space of secondary structures with $n = 15$. We show the neighbourhood of two selected structures: (i) the open chain and (ii) the longest stable hairpin. As outlined in section 2.3 the minimal size for a hairpin loop is three nucleotides, and this leads to $n(n-7)/2+6$ structures that can be formed from the open chain by closing one base pair. The longest possible stack contains $\lfloor (n-3)/2 \rfloor$ base pairs; each of them can be opened to yield a structure at one move distance, and hence the numbers of nearest neighbours for the two structures is different.

opening, and base pair shifts. In contrast to sequence space the points in shape space are not equivalent, as shown by means of an example in figure 5.

### *2.3. Counting structures*

RNA structures are composed of structural elements that are assumed to contribute additively to the energetic and other extensive properties of the molecules. Examples for the partitioning of structures are shown in figure 6. Based on the assumption of additivity the molecular properties derived from the secondary structures' properties can be computed recursively from smaller to larger and larger segments (figure 7).

In order to provide a realistic basis for these enumerations the notion of a physically acceptable structure is introduced. Some classes of secondary structures, which would be allowed according to the definition, are excluded. Because of steric strain, hairpin loops with one or two single nucleotides have such high free energies that they are never formed. The size of hairpin loops is restricted to three or more nucleotides without eliminating relevant structures ($n_{\mathrm{lp}} \geqslant \lambda = 3$). It is straightforward to calculate, for example, all possible secondary structures for a given chain length $n$, $s_n^{(\lambda)}$, by means of a recursion [4,14]. For a minimal length for hairpin loops, $n_{\mathrm{lp}} \geqslant \lambda$, one finds [15,16]

$$s_{m+1}^{(\lambda)} = s^{(}_{m}\lambda) + \sum_{j=1}^{m-\lambda} s_{j-1}^{(\lambda)} \cdot s_{m-j}^{(\lambda)} = s_m^{(\lambda)} + \sum_{j=\lambda}^{m-1} s_j^{(\lambda)} s_{m-j-1}^{(\lambda)} \quad \text{with } s_0^{(\lambda)} = s_1^{(\lambda)} = \cdots = s_\lambda^{(\lambda)} = 1. \quad (2)$$

In table 1 we see a comparison of the numbers of structures calculated for $\lambda = 1$ (condition I: no nearest neighbour pair restriction) and $\lambda = 3$ according to the loop strain energies. Less stringent but nevertheless straightforward is the exclusion of structures with isolated base pairs as candidates for mfe structures because the major part of the stabilization of RNA structures results from stacking of base pairs ($n_{\mathrm{st}} \geqslant \sigma = 2$). In order to be able to account for minimal stack lengths the recursion has to be extended to an enumeration with a restriction in the

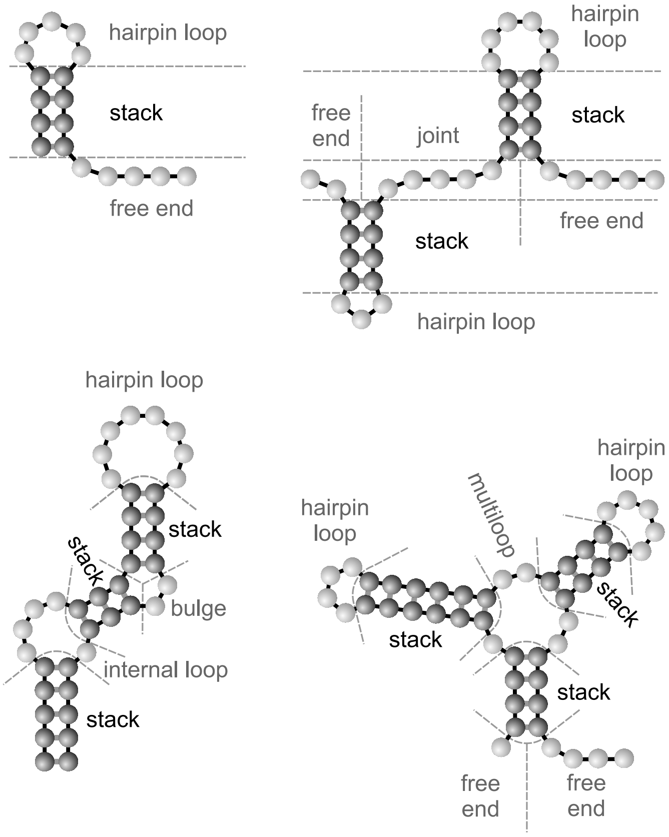


**Figure 6.** Elements of RNA secondary structures. Three classes of structural elements are distinguished: (i) stacks (indicated by nucleotides in dark colour), (ii) loops, and (iii) external elements, namely joints and free ends. Loops fall into several subclasses: Hairpin loops have one base pair, called the closing pair, in the loop. Bulges and internal loops have two closing pairs, and loops with three or more closing pairs are called multiloops. The number of closing pairs is denoted as the *degree* of the loop.

length of stacks, $n_{\mathrm{st}} \geqslant \sigma$ [15]. In this case the most convenient recursion makes use of three arrays:

$$\begin{aligned}
s_{m+1}^{(\lambda,\sigma)} &= \Xi_{m+1}^{(\lambda,\sigma)} + \Phi_{m-1}^{(\lambda,\sigma)}, \\
\Xi_{m+1}^{(\lambda,\sigma)} &= s_{m}^{(\lambda,\sigma)} + \sum_{k=\lambda+2\sigma-2}^{m-2} \Phi_{k}^{(\lambda,\sigma)} \cdot s_{m-k-1}^{(\lambda,\sigma)}, \\
\Phi_{m+1}^{(\lambda,\sigma)} &= \sum_{k=\sigma-1}^{\lfloor (m-\lambda+1)/2 \rfloor} \Xi_{m-2k+1}^{(\lambda,\sigma)},
\end{aligned} \tag{3}$$

with $s_0^{(\lambda,\sigma)} = s_1^{(\lambda,\sigma)} = \cdots = s_{\lambda+2\sigma-1}^{(\lambda,\sigma)} = 1$, $\Phi_0^{(\lambda,\sigma)} = \Phi_1^{(\lambda,\sigma)} = \cdots = \Psi_{\lambda+2\sigma-3}^{(\lambda,\sigma)} = 0$, and $\Xi_0^{(\lambda,\sigma)} = \Xi_1^{(\lambda,\sigma)} = \cdots = \Xi_{\lambda+2\sigma-1}^{(\lambda,\sigma)} = 1$. Performing the recursion up to $m + 1 = n$ provides us with the numbers of secondary structures, $s_n^{(\lambda,\sigma)}$, as presented in tables 1 and 3.

For long RNA sequences the recursions approach asymptotic expressions of the type [15]

$$s_n^{(\lambda,\sigma)} \approx s_{\lim}^{(\lambda,\sigma)}(n) = A_{\lambda,\sigma} \times n^{-3/2} \left(B_{\lambda,\sigma}\right)^n \text{ for large } n. \tag{4}$$

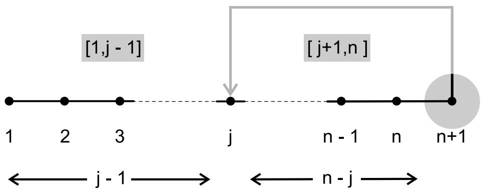


**Figure 7.** The build-up principle of RNA secondary structures. The concept sketched in the figure is basic to counting structures as well as to the computation of minimal free energy (mfe) structures by means of dynamic programming. The recursion proceeds from $n$ to $n+1$ by considering all possible cases: (i) the added nucleotide $(n+1)$ does not form a base pair and the property under consideration is the same as for the structure with $n$ nucleotides, $[n]$, plus the contribution for the single nucleotide, $n + 1$, or (ii) the added nucleotide does form a base pair with the nucleotide in position $j$. Then the structure is partitioned into three parts, one segment $[1, j - 1]$, the base pair $(j, n + 1)$, and the segment $[j + 1, n]$. The recursion is completed by consideration of all possible partitions with $j = 1, \ldots, n$. The necessary condition for the applicability of the recursion is additivity of the three individual contributions in the calculation of the property under consideration.

**Table 1.** Comparison of the numbers of RNA sequences and structures as a function of the chain length $n$. Given are numbers computed from recursions (2) and (3) for different values of $\lambda$ and $\sigma$ as well as the parameters $A$ and $B$ for the asymptotic expressions. In the two rightmost columns the values of the recursion are compared with those computed from the asymptotic formula, $s^{(3,2)}(n)$.

| | Number of sequences | | Number of structures | | | |
|---|---|---|---|---|---|---|
| $n$ | $2^n$ | $4^n$ | $s_n^{(1,1)}$ | $s_n^{(3,1)}$ | $s_n^{(3,2)}$ | $s_{\text{lim}}^{(3,2)}(n)$ |
| 10 | 1024 | $1.049 \times 10^6$ | 423 | 65 | 14 | 21.92 |
| 20 | $1.049 \times 10^6$ | $1.100 \times 10^{12}$ | $2.516 \times 10^6$ | $1.066 \times 10^5$ | 2741 | 3618 |
| 100 | $1.268 \times 10^{30}$ | $1.607 \times 10^{60}$ | $6.764 \times 10^{38}$ | $6.320 \times 10^{32}$ | $8.478 \times 10^{36}$ | $8.816 \times 10^{36}$ |
| 200 | $1.607 \times 10^{60}$ | $2.582 \times 10^{120}$ | $1.518 \times 10^{80}$ | $2.072 \times 10^{68}$ | $1.233 \times 10^{50}$ | $1.270 \times 10^{50}$ |
| $\lim_{n\to\infty}$ | | | $A = 1.1044$<br>$B = 2.618$ | $A = 0.7131$<br>$B = 2.289$ | $A = 1.4848$<br>$B = 1.849$ | |

Thus, all recursions increase exponentially with chain length $n$. Some values of $A$ and $B$ are given in table 1. For the physically meaningful case, $\lambda = 3$ and $\sigma = 2$, the asymptotic expression becomes $s_{\text{lim}}^{(3,2)}(n) = 1.4848 \times n^{-3/2}(1.84892)^n$. As shown in table 1 the asymptotic values are somewhat larger than the values from the recursion. For $n = 200$ the error is around 3%.

The general conclusions to be drawn from table 1 are (i) the two free energy based restrictions applied to the secondary structures reduce their numbers drastically, whereby the effect of the minimal stack length $\sigma = 2$ is somewhat more effective than the minimal hairpin, loop length $\lambda = 3$, and (ii) for the four letter alphabet, $\mathcal{A} = \{\mathbf{A},\mathbf{U},\mathbf{G},\mathbf{C}\}$, we have always more sequences than structures, whereas only the application of both restrictions (hairpin loop length and stack length restriction) leads to more sequences than structures for the two letter alphabets, $\mathcal{A} = \{\mathbf{A}, \mathbf{U}\}$ and $\mathcal{A} = \{\mathbf{G}, \mathbf{C}\}$. Indeed we compute more structures than sequences for sufficiently long two letter sequences with $\sigma = 1$ and $\lambda = 1, 3$.

### *2.4. Compatibility of sequences and structures*

A sequence $X = (x_1x_2 \ldots x_n)$ over an alphabet $\mathcal{A}$ with $\kappa$ letters is *compatible* with the structure or the matching $\Omega$ if $\{i–j\} \in \Omega$ implies that $x_ix_j$ is an allowed base pair. This situation is expressed by $x_ix_j \in \mathcal{B}$. We denote the set of all sequences that are compatible with a structure $\Omega$ by

$$\mathbf{C}^{(\mathcal{A})}[\Omega] = \left\{ X \,|\, \{i\text{–}j\} \in \Omega \implies x_i x_j \in \mathcal{B} \right\}. \tag{5}$$

Clearly, for each $i \in \Upsilon$ we may choose an arbitrary letter from the nucleic acid alphabet $\mathcal{A}$, while for each pair we may choose any of the $\varrho$ base pairs contained in $\mathcal{B}$. For a given structure we have, therefore,

$$|\mathbf{C}^{(\mathcal{A})}[\Omega]| = \kappa^{|\Upsilon|} \varrho^{|\Omega|} \tag{6}$$

compatible sequences.

The problem has a relevant inverse too: a structure is compatible with a sequence when it fulfils precisely the compatibility relation defined above. The set of all structures which are compatible with a sequence $X$ over an alphabet $\mathcal{A}$ is given by

$$\mathbf{C}[X^{(\mathcal{A})}] = \left\{ \Omega \middle| \{i\text{–}j\} \in \Omega \implies x_i x_j \in \mathcal{B} \right\} = \{\Omega \,|\, X \in \mathbf{C}[\Omega_i]\}. \tag{7}$$

This compatible set comprises all possible structures consisting of the mfe structure together with all suboptimal conformations. The two notions of compatibility assign, based on the same condition, a set of sequences to a given structure and conversely a set of structures to a predefined sequence. The two relations are somewhat complementary in sequence space and shape space.

How many structures are compatible with a given sequence $X$? The calculation of this number is rather involved, and apart from computation and exhaustive enumeration of all (suboptimal) structures, which is possible for small RNA molecules only, we have to rely here on an estimate that is based on recursions analogous to (2) and (3).[8] The estimate makes use of the *stickiness* of an RNA sequence, $p(X)$, expressing the probability that two arbitrarily chosen nucleotides can form a base pair,

$$p(X) = 2 \sum_{\alpha_i \alpha_j \in \mathcal{B}} p_i(X) p_j(X) \text{ with } p_k(X) = \frac{n_k(X)}{n}, \; k = i, j, \tag{8}$$

where $n_i(X)$ and $n_j(X)$ are the numbers of nucleotides $\alpha_i$ and $\alpha_j$ in the sequence $X$, respectively, $\sum_{i=1}^{\kappa} p_k(X) = 1$, and $n = \sum_{\alpha_i \in \mathcal{A}} n_i(X)$ is the chain length of the molecule. For a (random) sequence $X$ with defined nucleotide composition $(p_1, \ldots, p_\kappa)$ and stickiness $p(X)$ the recursion can be extended according to [18]

$$\begin{aligned} s_{m+1}(p) &= s_m(p) + p \sum_{j=1}^{m-\lambda} s_{j-1}(p) \cdot s_{m-j}(p), \\ \text{with } & s_0(p) = s_1(p) = \cdots = s_\lambda(p) = 1, \end{aligned} \tag{9}$$

The quantity $s_n(p)$ yields an estimate of the number of structures that are compatible with the sequence $X$. The recursion and the estimate can be extended to a restriction of the length of stacks, $n_{\text{st}} \geqslant \sigma$ [15]:

$$\begin{aligned} s_{m+1}(p) &= \Xi_{m+1}(p) + \Phi_{m-1}(p), \\ \Xi_{m+1}(p) &= s_m(p) + \sum_{k=\lambda+2\sigma-2}^{m-2} \Phi_k(p) \cdot s_{m-k-1}(p), \\ \Phi_{m+1}(p) &= p \sum_{k=\sigma-1}^{\lfloor (m-\lambda+1)/2 \rfloor} \Xi_{m-2k+1}(p) \cdot p^k, \end{aligned} \tag{10}$$

[8] The number of all compatible structures can also be obtained from the partition function [17] in the limit of infinite temperature, $T \to \infty$ (section 4.2). This limit, however, is not easy to compute because of numerical problems.

**Table 2.** Estimates on the numbers of suboptimal structures, $s_n(p)$, with $\lambda = 3$, $\sigma = 1$, $p(X)$ being the stickiness of sequence $X$.

| Chain length, $n$ | Stickiness, $p(X)$ | | | |
|---|---|---|---|---|
| | 1.0 | 0.5 | 0.375 | 0.25 |
| 10 | 65 | 21.4 | 14.3 | 8.6 |
| 20 | $1.07 \times 10^{5}$ | 7 403 | 2 778 | 787.8 |
| 50 | $1.82 \times 10^{15}$ | $1.27 \times 10^{12}$ | $8.52 \times 10^{10}$ | $2.57 \times 10^{9}$ |
| 100 | $6.32 \times 10^{32}$ | $2.09 \times 10^{26}$ | $8.05 \times 10^{23}$ | $5.81 \times 10^{20}$ |
| 200 | $2.07 \times 10^{68}$ | $1.55 \times 10^{55}$ | $1.95 \times 10^{50}$ | $8.06 \times 10^{43}$ |

with $s_0 = s_1 = \cdots = s_{\lambda+2\sigma-1} = 1$, $\Phi_0 = \Phi_1 = \cdots = \Psi_{\lambda+2\sigma-3} = 0$, and $\Xi_0 = \Xi_1 = \cdots = \Xi_{\lambda+2\sigma-1} = 1$. Performing the recursion up to $m + 1 = n$ provides us again with an estimate for the numbers of secondary structures.

Physically acceptable suboptimal structures exclude hairpin loops with one or two single nucleotides and hence $\lambda = 3$. Since suboptimal conformations need not fulfil the criterion of negative free energies, no restriction on stack lengths is appropriate. For a minimum hairpin loop length of $\lambda = 3$ and $\sigma = 1$ we find the numbers collected in table 2. The numbers of suboptimal structures become very large at moderate chain length $n$ itself. The expressions given here become asymptotically correct for long sequences. In order to provide a test for smaller chain lengths we refer to one particular case where the number of suboptimal structures has been determined by exhaustive enumeration: the sequence

**AAAGGGCACAGGGUGAUUUCAAUAAUUUUA**

with $n = 30$ and $p = 0.4067$ has 1,416,661 configurations, and the estimate by means of recursion (10) yields a value $s_{30}(0.4067) = 1.17 \times 10^6$ for $\lambda = 3$ and $\sigma = 1$ that is fairly close to the exact number.

### 2.5. *Computation of mfe structures*

Secondary structures of RNA molecules with minimal free energies are modelled in terms of a mapping from sequence space onto shape space,

$$\psi : \{\mathcal{Q}_n^{(\mathcal{A})}; d_H(X_i, X_j)\} \overset{\text{mfe}}{\Longrightarrow} \{\mathcal{S}_n; d_S(S_i, S_j)\} \text{ or } S_k = \psi(X_k). \tag{11}$$

Thereby we make the implicit assumption that mfe structures can be uniquely assigned to sequences. This assumption is essentially correct for not too small sequences and a sufficiently high resolution in the free energy parameters; the exceptions are only molecules that have degenerate ground state structures because of symmetry.[9]

Computation of secondary structures with minimum free energies [6] is based on the same principle as counting the numbers of structures (figure 7). First, the free energies of the smallest possible substructures are taken or computed from a list of parameters, and then a dynamic programming table of free energies is progressively completed by proceeding from smaller to larger segments until the minimum free energy of the whole molecule is obtained. Backtracking reconstructs the structure. The conventional approach is empirical and uses the free energies and enthalpies of RNA model compounds to derive the parameters

[9] We remark that almost all RNA folding routines return a single mfe structure for a sequence input no matter whether the ground state is degenerate, because backtracking commonly accepts the first solution it finds. Ground state degeneracies are found in calculations of all suboptimal structures (for example by means of the routine `RNAsubopt` in the Vienna RNA package; see also section 4.2).

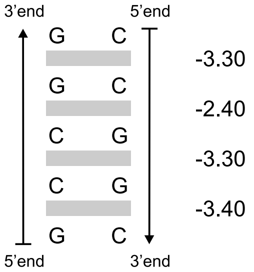


**Figure 8.** Stacking parameters for the interaction between **GC** base pairs. Free energies of stacking are given for the three different interaction geometries (the first and the third pair of pairs are identical). Values are given in kcal mol$^{-1}$. Additivity is assumed, and summation yields $\Delta G_0 = -12.40$ kcal mol$^{-1}$ for the free energy of interaction in the stack of five pairs.

for the individual structural elements. These elements correspond to the substructures shown in figure 6 which are partitioned further to allow for base pair specific contributions. As an example we show a computation of the stacking free energy for a cluster of **GC**-pairs in figure 8, which is obtained from three free stacking energy parameters for the **GC**-pairs interacting at different geometries. In total 21 different free stacking energy parameters are required for the six base pairs. In order to be able to compute the temperature dependence, 21 stacking enthalpy parameters are required in addition. Loops are taken into account with loop size dependent parameters, and hairpin loops, bulges, internal loops, and multiloops are treated differently. Other parameters consider nucleotide stacking on top of regular stacks, especially stable configurations, for example tetraloops[10] with specific sequences and end-on-end stacking of stacks. Stacks are (almost) the only structure stabilizing elements because only base pair stacking is a contribution with substantial negative free energy. Further structure stabilization comes from stacking of single bases called 'dangling ends' upon stacks, and there are other sequence specific contributions. Loops are almost always destabilizing because of the entropic effect of the ring closure that freezes internal degrees of freedom.

Listings of parameters, which are updated every few years, can be found in the literature [7–9,19]. These parameters define an energy function $E(X;\Omega)$ that assigns a unique free energy value to every substructure and provides the tool for completing the entries in the dynamic programming table. Several software packages are available, and web servers make secondary structure calculations easily accessible for everybody (see, for example, the Vienna RNA package and the Vienna RNA server [12,20]).

The calculated numbers of structures can now be compared with the numbers of mfe structures actually obtained by folding sequences of chain lengths $n$ over an alphabet $\mathcal{A}$ (table 3). The number of structures formed by **AU** and **AUG** sequences is substantially smaller than those obtained through folding sequences from **AUGC**, **UGC**, and **GC** alphabets. The explanation is straightforward: since the hydrogen bonding and stacking free energies of **G–C** pairs are substantially larger than those of **A–U** pairs sequences lacking **C** can only form weak pairs and, accordingly, stable stacks have to be longer. The instability of shorter stacks in the alphabets {**A**,**U**} and {**A**,**U**,**G**} implies that structures with these stacks cannot be obtained as mfe structures. In other words, a realistic estimate of the numbers of structures formed by **AU** and **AUG** sequences requires a value of $\sigma > 2$. Although the restriction of hairpin loops and stacks to $\lambda = 3$ and $\sigma = 2$ for **G** and **C** containing alphabets comes close to the results

[10] It is common to denote the size of small hairpin loops by special words: 'triloops' are hairpin loops with three single nucleotides in the loop, 'tetraloops' have four single nucleotides, and 'pentaloops' five single bases.

**Table 3.** Comparison of exhaustively folded sequence spaces. The values are derived through exhaustive folding of all sequences of chain length $n$ from a given alphabet. The numbers refer to actually occurring minimum free energy structures (open chain included) without isolated base pairs and are directly comparable to the total numbers of acceptable structures, $s_n^{(3,2)}$, with $\lambda = 3$ and $\sigma = 2$ as computed from the recursion in equation (3) [15]. The parameters are taken from [7][a].

| Chain length $n$ | Number of sequences | | Number of structures | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $2^n$ | $4^n$ | $s_n^{(3,2)}$ | GC | UGC | AUGC | AUG | AU |
| 7 | 128 | $1.64 \times 10^4$ | 2 | 1 | 1 | 1 | 1 | 1 |
| 8 | 256 | $6.55 \times 10^4$ | 4 | 3 | 3 | 3 | 2 | 1 |
| 9 | 512 | $2.62 \times 10^5$ | 8 | 7 | 7 | 7 | 3 | 1 |
| 10 | 1024 | $1.05 \times 10^6$ | 14 | 13 | 13 | 13 | 5 | 3 |
| 12 | 4096 | $1.68 \times 10^7$ | 37 | 35 | 35 | 36 | 14 | 8 |
| 14 | $1.64 \times 10^4$ | $2.68 \times 10^7$ | 101 | 83 | 89 | 93 | 31 | 20 |
| 16 | $6.55 \times 10^4$ | $4.29 \times 10^9$ | 304 | 214 | 246 | 260 | 72 | 44 |
| 18 | $2.62 \times 10^5$ | $6.87 \times 10^{10}$ | 919 | 582 | 735 | | 180 | 96 |
| 20 | $1.05 \times 10^6$ | $1.10 \times 10^{12}$ | 2 741 | 1 599 | 2 146 | | 504 | 232 |
| 25 | $3.36 \times 10^7$ | $1.13 \times 10^{15}$ | 44 695 | 18 400 | | | | 1 471 |
| 30 | $1.07 \times 10^9$ | $1.15 \times 10^{18}$ | 760 983 | 218 318 | | | | 21 315 |

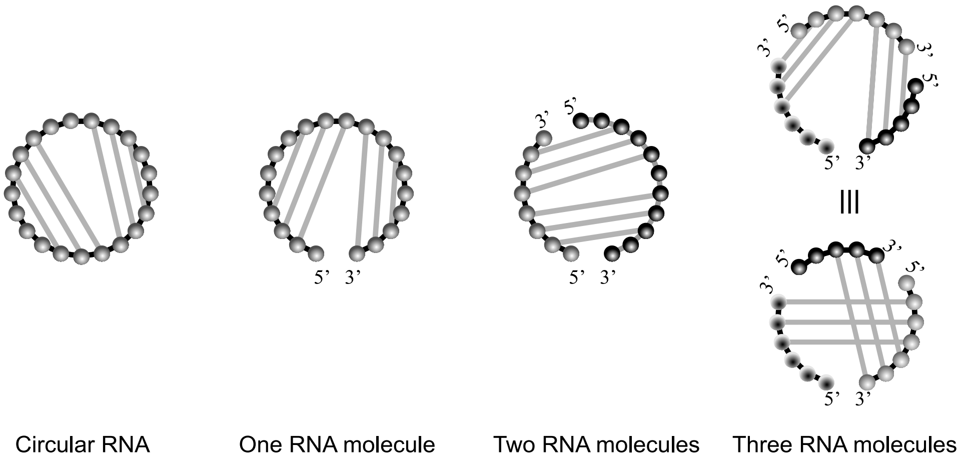
[a] We remark that later updates of parameters, e.g. [8], yield smaller numbers of mfe structures because the triloops are more strongly disfavoured in more recent parameter sets (see also figure 9).

```
1  ((...))..        1  ((...))...     8  .((....)).
2  .((...)).        2  .((...))..     9  ..((....))
3  ..((...))        3  ..((...)).    10  (((....)))
4  (((...)))        4  ...((...))    11  ((.....)).
5  ((....)).        5  (((...))).    12  .((.....))
6  .((....))        6  .(((...)))    13  ((......))
7  ((.....))        7  ((....))..    14  ..........
8  .........
```

**Figure 9.** All secondary structures of sequences with $n = 9$ and 10, $\lambda = 3$ and $\sigma = 2$. The sketch shows all eight structures with hairpin loop sizes $n_{\text{hp}} \geqslant 3$ and stack sizes $n_{\text{st}} \geqslant 2$ for $n = 9$, and all 14 structures for $n = 10$. Structures not realized as mfe structures within the natural **AUGC** alphabet (with the parameters set of [7]) are indicated in grey colour. The structures not realized have a hairpin loop of size 3 enclosed by a stack of two base pairs with no 3′-dangling end. In computations with the more recent parameter set [8] all structures with triloops enclosed by two stacked base pairs are missing. These are structures 1–3 for sequences of chain length $n = 9$ and structures 1–4 for sequences with $n = 10$.

of exhaustive folding and enumeration, we do still find fewer structures formed by **AUGC** sequences than predicted.

In order to present an example for why some structures cannot be formed as mfe structures we consider the completely folded sequence spaces with $n = 9$ and $n = 10$ (figure 9). The non-realized mfe structures are those missing an unpaired nucleotide at the 3′-end of the stack with two base pairs. The extra stabilization of the stack by the 3′-dangling end is indispensable for the formation of a stable structure. This is also supported by choosing slightly changed triloop parameters: the marginal stability of small triloops with the shortest possible stacks is



Circular RNA

One RNA molecule

Two RNA molecules

Three RNA molecules

**Figure 10.** Cofolding of RNA molecules. The cofolding of two or more, in general $N$, RNA molecules follows the same principles as the folding of a single open chain molecule, or even folding a circular RNA molecule, provided three features are properly accounted for: (i) The pairs of free ends (zero for circular RNAs, one for ordinary folding of a single strand, two and more for cofolding) constitute an exterior loop that does not contribute energetically, (ii) for three or more molecules each of the $(N-1)!$ different permutations of the molecules along the circle has to be taken into account, and (iii) complex formation is concentration dependent and proper accounting of the thermodynamic equilibrium is indispensable. As shown in the rightmost example, a conformation that is forbidden by the no-pseudoknot condition (III) in one permutation, say $1 \rightarrow 3 \rightarrow 2$, is legitimate in another permutation, here $1 \rightarrow 2 \rightarrow 3$.

turned into instability by using the parameter set [8]. With these parameters only triloops with three adjacent base pairs are stable elements in small RNA molecules (see figure 9).

### 2.6. *Cofolding of RNA molecules*

The RNA folding algorithm can be generalized in a straightforward way to compute structures resulting from simultaneous folding of two or more RNA molecules [12,21,22]. As illustrated in figure 10, the search for the most stable hybrid of $N$ molecules starts by concatenation and circular of the $N$ molecules. On a circle only $(N-1)!$ permutations are different[11], and all of them have to be computed. To obtain the energetically optimal hybrid structure of the $N$ molecules we choose the complex with the lowest mfe out of the $(N-1)!$ different solutions. Accordingly, for cofolding two molecules there is only one arrangement, for three molecules we have to consider two different arrangements on the circle, six for four molecules, etc. Whenever one of the sequence cuts is bridged by a base pair an exterior loop is formed, which does not contribute to the free energy of the complex.

Cofolding depends on the concentrations since the complex formation is a bimolecular chemical reaction [21, 23]. The two monomeric species $\mathsf{A}$ and $\mathsf{B}$ can form three different complexes, $\mathsf{AA}$, $\mathsf{AB}$ and $\mathsf{BB}$, which by mass action fulfill three equilibrium relations,

$$\mathsf{R} + \mathsf{S} \overset{K_{\mathsf{RS}}}{\rightleftharpoons} \mathsf{RS} \quad \text{with} \quad \mathsf{R} = \mathsf{A}, \mathsf{B};\ \mathsf{S} = \mathsf{A}, \mathsf{B}. \tag{12}$$

[11] In total we have $N!$ permutations. On the circle it makes no difference with which molecule we start counting and therefore only $N!/N = (N-1)!$ arrangements are different.

The equilibrium constants can be obtained from the computed partition functions

$$\begin{aligned} K_{AA} &= \frac{1}{2}\left(\frac{Q_{AA}}{(Q_A)^2} - 1\right) e^{-\Theta_I/RT}, \\ K_{BB} &= \frac{1}{2}\left(\frac{Q_{BB}}{(Q_B)^2} - 1\right) e^{-\Theta_I/RT}, \\ K_{AB} &= \left(\frac{Q_{AB}}{Q_A\,Q_B} - 1\right) e^{-\Theta_I/RT}, \end{aligned} \tag{13}$$

wherein $Q$ denotes the partition functions of monomers and dimers as obtained by the folding routine and $\Theta_I$ is the free energy difference composed with some reference state commonly called the initiation energy (section 4.2). The free concentrations of monomers, $a = [\mathsf{A}]$ and $b = [\mathsf{B}]$, are obtained from mass conservation and the total monomer concentrations $a_0$ and $b_0$, respectively, as solutions of the two nonlinear equations

$$\begin{aligned} a + 2\,a^2\,K_{\mathsf{AA}} + a\,b\,K_{\mathsf{AB}} \; - \; a_0 &= 0, \\ b + 2\,b^2\,K_{\mathsf{BB}} + a\,b\,K_{\mathsf{AB}} \; - \; b_0 &= 0. \end{aligned} \tag{14}$$

The dimer concentrations result then from equations (12):

$$[\mathsf{AB}] = K_{\mathsf{AB}}\,ab \qquad [\mathsf{AA}] = K_{\mathsf{AA}}\,a^2, \text{ and} \qquad [\mathsf{BB}] = K_{\mathsf{BB}}\,b^2. \tag{15}$$

Stable hybridization obviously requires $[\mathsf{AB}] \gg \{a, b, [\mathsf{AA}], [\mathsf{BB}]\}$.

## 3. Evolution and design of RNA structures

RNA structures and properties can be optimized through mutation and selection (for reviews see [24–26] and the collection of papers [27]). We shall focus here on the influence of RNA secondary structures on the evolutionary process. Towards this goal we shall discuss first the inverse problem of RNA folding [12], which is the basis of designing RNA-structures, then investigate the mapping of sequences into structures in more detail [28], and eventually discuss computer simulations of RNA optimization [29].

### *3.1. Inverse folding*

Given a sequence $X$, the folding problem consists of finding a structure $S$ that minimizes an energy function $E(X; S)$ and satisfies other constraints, such as the no-pseudoknot condition III. In section 2.5 we have seen that the folding problem for pseudoknot-free secondary structures is easily solved by means of dynamic programming. In the *inverse folding problem* we have the same energy function $E$ and the same constraints, but we are given the structure $S$ and search for a sequence $X$ that has $S$ as an optimal structure. We denote the set of solutions of the inverse folding problem by $\psi^{-1}(S) \doteq \{X|\psi(X) = S\}$. Note that $\psi^{-1}(S)$ may be empty, since there are secondary structures that are not formed as minimum energy structures of any sequence (section 2.5).

Just as the folding problem can be regarded as an optimization problem on the energy landscape of a given sequence, we can also rephrase the inverse folding problem and turn it into a combinatorial optimization problem. To this end, we consider a measure $d(S_1, S_2)$ for the structural dissimilarity of two RNA secondary structures $S_1, S_2$. A variety of such distance measures have been described in the literature [12,30–33], and two of them have been mentioned already (section 2.2). Since we are interested here only in sequences of equal length, we may simply use the cardinality of the symmetric difference of $S_1$ and $S_2$:

$$d(S_1, S_2) = \big|(S_1 \cup S_2) \setminus (S_1 \cap S_2)\big|. \tag{16}$$

Clearly, sequence $X$ folds into structure $S$ if and only if $\Xi(X) = d(S, \psi(X)) = 0$. Hence, inverse folding translates into minimizing $d$ over all sequences. We know *a priori* that solutions to the inverse folding problem must be compatible with the structure:

$$\psi^{-1}(S) \subseteq \mathbf{C}[S]. \tag{17}$$

It is straightforward to modify this approach to search, for instance, for sequences in which the ground state is much more stable than any structural alternative [12]: let $E(X; S)$ be the energy of structure $S$ for sequence $X$, and let $\Delta G(X)$ be the ensemble free energy of sequence $X$, which can be computed by McCaskill's algorithm [17] (See also 4.2). Then, sequences with the desired property minimize

$$\Xi(X) = E(X; S) - G(X) = -RT \ln \gamma_X(S), \tag{18}$$

where $\gamma_X(S)$ is the weight of structure $S$ in the Boltzmann ensemble of sequence $X$.

It has been found empirically [12] that this combinatorial optimization problem is easily solvable by means of adaptive walks. Starting from a randomly chosen initial sequence $X_0$ we produce mutants by exchanging a nucleotide at the unpaired positions $\Upsilon$ or by replacing one of the six pairing combinations by another one in a pair in $S$. A mutant is accepted if the cost function $\Xi(X)$ decreases. In a more sophisticated version, implemented in the program `RNAinverse`, a significant speed-up is achieved by optimizing parts of the structure individually. This reduces the number of evaluations of the folding procedure for long sequences. A more sophisticated stochastic local search algorithm is used in the `RNA-SSD` software [34].

### *3.2. Neutral networks*

Inverse folding cannot provide a unique answer for every structure since we have many more sequences than structures. As follows directly from table 3 the mapping $S = \psi(X)$ is many-to-one in all five alphabets.

The set of sequences that form a given mfe structure $S$, the pre-image of $S$ in sequence space or the neutral set

$$\mathbf{G}[S] = \psi^{-1}(S) \doteq \{X|\psi(X) = S\}, \tag{19}$$

is a subset of the compatible set of structure $S$: $\mathbf{G}[S] \subset \mathbf{C}[S]$. The neutral set is turned into a graph, the *neutral network*, by connecting all pairs of nodes with Hamming distance 1 by an edge.

The global properties of neutral networks may be derived using random graph theory [35]. The characteristic quantity for a neutral network is the degree of neutrality, $\bar{\lambda}$, which is obtained by averaging the fraction of Hamming distance 1 neighbours that form the same mfe structure, $\lambda_X = n_{\mathrm{ntr}}^{(1)}/\big(n \cdot (\kappa - 1)\big)$, with $n_{\mathrm{ntr}}^{(1)}$ being the number of neutral one-error neighbours, over the whole network, $\mathbf{G}[S]$:

$$\bar{\lambda}[S] = \frac{1}{|\mathbf{G}(S)|} \sum_{X \in \mathbf{G}[S]} \lambda_X. \tag{20}$$

The connectedness of neutral networks is, among other properties, determined by the degree of neutrality [36]:

$$\text{with probability 1 a network is } \begin{cases} \text{connected} & \text{if } \bar{\lambda} > \lambda_{\mathrm{cr}} \\ \text{not connected} & \text{if } \bar{\lambda} < \lambda_{\mathrm{cr}} \end{cases}, \tag{21}$$

where $\lambda_{\mathrm{cr}} = 1 - \kappa^{-\frac{1}{\kappa-1}}$. Computations yield $\lambda_{\mathrm{cr}} = 0.5, 0.423$ and $0.370$ for the critical value in two, three and four letter alphabets, respectively. Random graph theory predicts a single largest component for non-connected networks, i.e. networks below the threshold, that is commonly called the ‘giant component’. Real neutral networks derived from RNA secondary structures may deviate from the prediction of random graph theory in the sense that they have two or four equally sized largest components. This deviation is readily explained by the non-uniform distribution of the sequences belonging to $\mathbf{G}[S_k]$ in sequence space, which is caused by the specific structural properties of $S_k$ [37, 38]. In particular, sequences that fold into structures which allow for closure of additional base pairs at the ends of the stacks are more probable to be formed by sequences that have an excess of one of the two bases forming a base pair than by those with the uniform distribution $x_{\mathsf{G}} = x_{\mathsf{C}}$ and $x_{\mathsf{A}} = x_{\mathsf{U}}$. In the case of **GC**-sequences the neutral network of such a structure is then depleted from sequences in the middle of the sequence space and we find two largest components, one at excess **G** and one at excess **C**.

The union of the one-error neighbourhoods of all sequences belonging to a neutral network $\mathbf{G}[S_k]$ is characterized as the *shadow* of structure $S_k$ in sequence space. Since single nucleotide exchanges are the most frequent mutations the shadows determine the role of structures in evolutionary processes. Here we consider the shadows of a few typical examples of RNA structures.

The first example is a series of hairpin structures of chain length $n = 33$ with one long stack (table 4). We compare the longest possible hairpin with a triloop and a stack of $n_{\mathrm{st}} = 15$ base pairs ($S_1^{(\mathrm{hp33})}$) with shorter stacks, $n_{\mathrm{st}} = 13$ and 11, and larger hairpin loops, $n_{\mathrm{lp}} = 7$ and 11 ($S_2^{(\mathrm{hp33})}$ and $S_3^{(\mathrm{hp33})}$), or longer free ends ($S_4^{(\mathrm{hp33})}$ and $S_5^{(\mathrm{hp33})}$). First the two letter alphabets, **GC** and **AU**, show always substantially lower degrees of neutrality than all alphabets with three or four letters. In order to provide a reference we compute the maximum degree of neutrality, $\lambda_{\max}(S)$, for the five different alphabets by assuming that mutation of an unpaired nucleotide does not change the mfe structure, whereas mutation of a nucleotide in a base pair will change the structure with a certain probability (figure 11) and find

$$\lambda_{\max}(S) = \frac{1}{n}\left(\gamma_{\mathrm{sn}} \cdot |\Upsilon| + \gamma_{\mathrm{bp}} \cdot 2\,|\Omega|\right) . \tag{22}$$

Herein, $\gamma_{\mathrm{sn}}$ and $\gamma_{\mathrm{bp}}$ are the probabilities of staying within the set of compatible sequences when an unpaired nucleotide or a nucleotide of a base pair is mutated, respectively. Clearly, $\gamma_{\mathrm{sn}} = 1$ and for the base pairs we compute by averaging: $\gamma_{\mathrm{bp}} = 0$ for **AU** and **GC**, $\gamma_{\mathrm{bp}} = 1/4$ for **AUG** and **UGC**, and $\gamma_{\mathrm{bp}} = 2/9$ for **AUGC**. Actually, $\lambda_{\max}(S)$ is the probability of staying within the compatible set $\mathbf{C}[S]$ after a point mutation at an arbitrary position. the maximal degrees of neutrality for the hairpins in table 4 are summarized in table 5: the maximum degrees of neutrality in the two letter alphabets are smaller than the values for the three and four letter alphabets. As a matter of fact the $\lambda_{\max}$ values of three and four letter alphabets are quite close, with somewhat higher values for three letters. The data in table 4 essentially confirm the estimate of equation (22), except that that in all examples, except that the simple hairpin sequences over the **AUGC** alphabet show the highest degree of neutrality. The explanation is straightforward: neutral networks contain sequences forming the same mfe structure and not sequences that are compatible with the structure. Mutated unpaired nucleotides can pair with other nucleotides and the probability for the occurrence of such an event reduces $\gamma_{\mathrm{sn}}$. Neglecting all stereochemical and other restrictions, the probability of remaining unpaired is $\gamma_{\mathrm{sn}} = 1/3 = 0.3333$ for **AUG** and **UGC** but $\gamma_{\mathrm{sn}} = 19/36 = 0.5278$ for **AUGC**. Despite the approximate nature of this estimate it is definitely suitable for explaining the observed trends in the degrees of neutrality.

It is interesting to note that alphabets without **C** have increasingly higher degrees of neutrality with increasing loop size ($S_1^{(\mathrm{hp33})} \rightarrow S_2^{(\mathrm{hp33})} \rightarrow S_3^{(\mathrm{hp33})}$). An interpretation of this observation can be given in terms of base pair stability: **GC** pairs and **GC** pair stacking is much

**Table 4.** Degree of neutrality in different nucleotide alphabets (Part I). The values for the degree of neutrality, $\bar{\lambda}[S]$, were obtained by sampling 10,000 random sequences folding into the structures $S$ using the inverse folding routine [12] and computing their complete one-error neighbourhoods in sequence space. Four types of mfe structures are considered: (i) hairpin loops of chain length $n = 33$, $S_k^{(\mathrm{hp33})}$ ($k = 1, \ldots, 5$), (ii) a structure of chain length $n = 33$ with two hairpin loops, $S^{(\mathrm{dhp33})}$, (iii) a Y-shaped structure of chain length $n = 50$, $S^{(\mathrm{y50})}$, and (iv) a two component structure of chain length $n = 135$, $S^{(\mathrm{dcp135})}$. Clover-leaf structures of chain length $n = 76$ will be presented in table 6. The values for the degree of neutrality, $\bar{\lambda}[S]$, were obtained by sampling 10 000 random sequences folding into the mfe structures. In some cases several smaller samples were used because of CPU economy reasons.

| Structure[a] | Nucleotide alphabet | | | | |
|---|---|---|---|---|---|
| | **GC** | **UGC** | **AUGC** | **AUG** | **AU** |
| $S_1^{(\mathrm{hp33})}$ | 0.09 ± 0.00 | 0.23 ± 0.02 | 0.24 ± 0.01 | 0.24 ± 0.02 | 0.09 ± 0.00 |
| $S_2^{(\mathrm{hp33})}$ | 0.10 ± 0.02 | 0.29 ± 0.03 | 0.31 ± 0.02 | 0.32 ± 0.03 | 0.15 ± 0.00 |
| $S_3^{(\mathrm{hp33})}$ | 0.11 ± 0.04 | 0.30 ± 0.05 | 0.33 ± 0.04 | 0.36 ± 0.05 | 0.21 ± 0.04 |
| $S_4^{(\mathrm{hp33})}$ | 0.21 ± 0.01 | 0.33 ± 0.02 | 0.34 ± 0.01 | 0.35 ± 0.02 | 0.21 ± 0.00 |
| $S_5^{(\mathrm{hp33})}$ | 0.21 ± 0.03 | 0.37 ± 0.05 | 0.40 ± 0.02 | 0.41 ± 0.03 | 0.27 ± 0.01 |
| $S^{(\mathrm{dhp33})}$ | 0.09 ± 0.05 | 0.27 ± 0.08 | 0.34 ± 0.08 | – | – |
| $S^{(\mathrm{y50})}$ | 0.06 ± 0.03 | 0.24 ± 0.07 | 0.29 ± 0.06 | 0.21 ± 0.07 | 0.08 ± 0.04 |
| $S^{(\mathrm{dcp135})}$ | 0.04 ± 0.02 | 0.21 ± 0.06 | 0.26 ± 0.05 | 0.19 ± 0.06 | 0.05 ± 0.03 |

[a] The following structures were used:

$S_1^{(\mathrm{hp33})}$: `(((((((((((((((...)))))))))))))))`

$S_2^{(\mathrm{hp33})}$: `(((((((((((((.......)))))))))))))`

$S_3^{(\mathrm{hp33})}$: `(((((((((((...........)))))))))))`

$S_4^{(\mathrm{hp33})}$: `....(((((((((((((...)))))))))))))`

$S_5^{(\mathrm{hp33})}$: `....(((((((((((.......)))))))))))`

$S^{(\mathrm{dhp33})}$: `..(((((....)))))....((((.....))))`

$S^{(\mathrm{y50})}$: `.(((((...(((((.....)))))...(((((.....)))))...)))))`

$S^{(\mathrm{dcp135})}$: `..((((((.((((((.....))))))..((((((......))))))...((((((....))))))..))))))`
`.......((((((.((((((.....))))))....((((((......))))))..)))))).`

more favoured energetically than **AU** or **GU** pairing. The larger the free energy gain for a new base pair created by mutation is, the more likely it will show up in the mfe structure, and **GC**-rich sequences are more likely to change structure on mutation therefore.

The double hairpin structure, $S^{(\mathrm{dhp33})}$, in figure 4 is perfectly stable in **GC** containing alphabets but unstable in the **AU** and **AUG** alphabets. In particular, the stack with only four base pairs is unable to stabilize the pentaloop. We shall find a similar problem in the case of the clover-leaf structures in table 6. A stack length of 4 is not enough to stabilize a pentaloop, but five base pairs in structure $S^{(\mathrm{y50})}$ are sufficient to stabilize the pentaloop and the multiloop of the 'Y' conformation. The last structure considered in table 4 presents a more complex structure of chain length $n = 135$ consisting of two joined structural motifs, a clover-leaf and a Y-element. The degrees of neutrality in the various alphabets are about 10% to 20% lower than in the simpler Y-structure, but the distribution over the alphabets,

$$\bar{\lambda}^{(\mathsf{AUGC})}(S) > \bar{\lambda}^{(\mathsf{UGC})}(S) > \bar{\lambda}^{(\mathsf{AUG})}(S) \gg \bar{\lambda}^{(\mathsf{AU})}(S) \geqslant \bar{\lambda}^{(\mathsf{GC})}(S),$$

is characteristic for the majority of generic RNA structures.[12]

[12] 'Generic' stands here for typical in the sense that constructed conformations like single long hairpins may have deviant distributions.

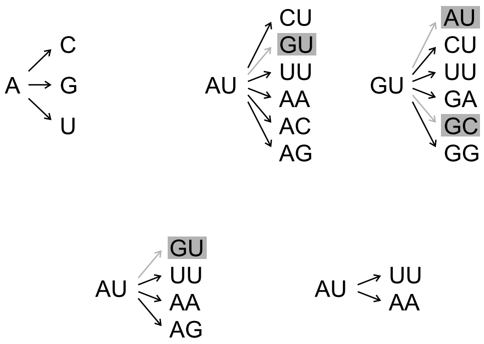


**Figure 11.** Compatibility of sequences and single point mutations. The upper part of the sketch shows mutations in unpaired nucleotides and base pairs in the natural four letter alphabet $\mathcal{A} = \{\textbf{A},\textbf{U},\textbf{G},\textbf{C}\}$. Four base pairs, **AU**, **CG**, **GC**, and **UA**, remain compatible with probability $\gamma_{\text{bp}} = 1/6$ and two base pairs with probability $\gamma_{\text{bp}} = 1/3$, yielding an average probability of $\bar{\gamma}_{\text{bp}} = 2/9$. The two lower diagrams show base pair mutation in the three letter alphabets $\mathcal{A} = \{\textbf{A},\textbf{U},\textbf{G}\}$ and $\mathcal{A} = \{\textbf{U},\textbf{G},\textbf{C}\}$ and the two letter alphabets $\mathcal{A} = \{\textbf{A},\textbf{U}\}$ and $\mathcal{A} = \{\textbf{G},\textbf{C}\}$, yielding $\gamma_{\text{bp}} = 1/4$ and $\gamma_{\text{bp}} = 0$, respectively.

**Table 5.** Maximum degree of neutrality $\lambda_{\max}(S)$ calculated using equation (22).

| Structure | Nucleotide alphabets | | |
|---|---|---|---|
| | **AU**, **GC** | **AUG**, **UGC** | **AUGC** |
| $S_1^{(\text{hp33})}$ | $\frac{1}{11} = 0.090\,909$ | $\frac{7}{22} = 0.318\,182$ | $\frac{29}{99} = 0.292\,929$ |
| $S_2^{(\text{hp33})}$ | $\frac{7}{33} = 0.212\,121$ | $\frac{9}{22} = 0.409\,091$ | $\frac{115}{297} = 0.387\,205$ |
| $S_3^{(\text{hp33})}$ | $\frac{1}{3} = 0.333\,333$ | $\frac{1}{2} = 0.5$ | $\frac{13}{27} = 0.481\,482$ |
| $S_4^{(\text{hp33})}$ | $\frac{7}{33} = 0.212\,121$ | $\frac{9}{22} = 0.409\,091$ | $\frac{115}{297} = 0.387\,205$ |
| $S_5^{(\text{hp33})}$ | $\frac{1}{3} = 0.333\,333$ | $\frac{1}{2} = 0.5$ | $\frac{13}{27} = 0.481\,482$ |

**Table 6.** Degree of neutrality in different nucleotide alphabets (Part II). The values for the degree of neutrality, $\bar{\lambda}[S]$, were obtained by sampling 1000 random sequences folding into the four clover-leaf structures with different stack sizes[a] using the inverse folding routine [12].

| Structure[a] | Nucleotide alphabet | | | | |
|---|---|---|---|---|---|
| | **GC** | **UGC** | **AUGC** | **AUG** | **AU** |
| $S_1$ | $0.05 \pm 0.03$ | $0.26 \pm 0.07$ | $0.28 \pm 0.06$ | – | – |
| $S_2$ | $0.06 \pm 0.03$ | $0.26 \pm 0.07$ | $0.28 \pm 0.06$ | $0.22 \pm 0.05$ | – |
| $S_3$ | $0.06 \pm 0.03$ | $0.25 \pm 0.07$ | $0.29 \pm 0.06$ | $0.21 \pm 0.06$ | – |
| $S_4$ | $0.07 \pm 0.03$ | $0.25 \pm 0.06$ | $0.31 \pm 0.06$ | $0.20 \pm 0.06$ | $0.07 \pm 0.03$ |

[a] The following clover-leaf structures were used:

$S_1$: `((((((...((((........)))).(((((.......))))).....(((((.......))))).))))))....`
$S_2$: `((((((...(((((......))))).(((((.......))))).....(((((.......))))).))))))....`
$S_3$: `((((((...(((((......))))).(((((.......))))).....((((((.....)))))).))))))....`
$S_4$: `((((((...((((((....)))))).((((((.....)))))).....((((((.....)))))).))))))....`

In table 6 we show, as a last example, computed values of the degree of neutrality, $\bar{\lambda}[S]$, in neutral networks derived from tRNA-like clover-leaf structures with different stack lengths of the hairpin loops. The most striking feature of the data is the weak structure dependence of $\bar{\lambda}[S]$ within a family: for a given alphabet the clover-leaves $S_1$, $S_2$, $S_3$ and $S_4$, have almost the

**Table 7.** The lengths of neutral paths through sequence space. The degree of neutrality, $\bar{\lambda}$, and the mean lengths of neutral paths through sequence space, $\bar{d}_H(X_0, X_f)$ (with $X_0$ being the initial and $X_f$ the last sequence), are compared for three examples: (i) folding of (stand alone) **AUGC** sequences of chain length $n = 100$, (ii) cofolding of **AUGC** sequences of chain length $n = 100$ with a single fixed sequence, and (iii) cofolding of **AUGC** sequences of chain length $n = 100$ with two single fixed sequences. The values represent averages over samples of 1200 random sequences. The value for the path length in **GC** sequence space with $n = 100$ is an estimate from figure 10 in [40, 39].

| Molecule | Alphabet | Degree of neutrality, $\bar{\lambda}$ | Neutral path length, $\bar{d}_H(X_0, X_f)$ |
|---|---|---|---|
| Single fold | **GC** | 0.08 | ≈45 |
| Single fold | **AUGC** | 0.33 | >95 |
| Cofold with one sequence | **AUGC** | 0.32 | 75 |
| Cofold with two sequences | **AUGC** | 0.18 | 40 |

same $\bar{\lambda}$ values irrespective of the stability of the corresponding fold. Because of the shorter stack lengths in $S_1$, $S_2$ and $S_3$ and the weakness of the **AU** pair, no **AU**-sequences forming these structures were obtained by inverse folding. The same was found for $S_1$ in the case of **AUG**-sequences. Considering the fact that $\lambda_{\text{cr}}$ decreases from two to four letter alphabets, we see that neutral networks in two letter sequence spaces ($\bar{\lambda} \approx 0.06$ and $\lambda_{\text{cr}} = 0.5$) and four letter sequence spaces ($\bar{\lambda} \approx 0.3$ and $\lambda_{\text{cr}} = 0.37$) must have rather vast extensions, the former being certainly non-connected, whereas the latter approach the connectivity threshold.

The extension of neutral networks can be visualized best by evaluating the lengths of neutral paths [28]. A neutral path connects pairs of neighbouring neutral sequences of Hamming distance $d_H = 1$ for single nucleotide exchanges and $d_H = 1$ or 2 for base pair exchange with the condition that the Hamming distance from a reference sequence increases monotonously along the path. The path ends when it reaches a sequence, which has only neutral neighbours that are closer to the reference sequence. Table 7 compares the degree of neutrality and the length of neutral path for **GC** and **AUGC** sequences of chain length $n = 100$ with the expected result: networks in **AUGC** space extend through whole sequence space, whereas **GC** networks sustain a neutral path of roughly only half of this length. The table also contains comparisons with constrained molecules that were cofolded with one or two fixed sequences [39]. The three values demonstrate the influence of multiple constraints on neutrality, which lead to a decrease in both degree of neutrality and length of neutral path.

*Shape space covering* is a consequence of the existence of neutral networks and the widespread almost random distribution of neutral sequences in sequence space. Sequences forming common shapes are distributed (almost) randomly in sequence space. Accordingly, one need not search the entire sequence space in order to find a sequence that folds into a given common shape. One can indeed show that it is sufficient to screen a (high-dimensional) sphere around an arbitrarily chosen reference sequence in order to find (with probability one) at least one sequence for every common shape [28]. The radius of this shape space covering sphere, $r_{\text{cov}}(n)$, can be estimated straightforwardly [40, 41]:

$$r_{\text{cov}}(n) = \min\left\{h = 1, 2, \ldots, n \mid B_h(n, \kappa) \geqslant \frac{\kappa^n}{S_n^{(3,2)}}\right\},$$

where $B_h$ is the number of sequences contained in a ball of radius $h$ and can be easily obtained from the recursion

$$B_h(n,\kappa) = \sum_{i=1}^{h} b_i(n,\kappa); \qquad b_i = b_{i-1}\cdot\frac{(\kappa-1)(n+1-i)}{i}; \qquad b_0 = 1.$$

The covering radius for common shapes is much smaller than the radius of sequence space ($n/2$). For example, it amounts to $r_{\rm cov} = 15$ for **AUGC**-sequences of chain length $n = 100$ and thus one has to search only a small fraction of sequence space $\mathcal{Q}_{100}^{(\mathsf{AUGC})}$ that contains $2.21\times10^{-38}\times|\mathcal{Q}_{100}^{(\mathsf{AUGC})}| = 0.355\times10^{23}$ sequences in order to find at least one sequence for each of the common shapes.

### *3.3. Evolutionary optimization*

The evolution of RNA molecules based on replication, mutation, and selection in a constant environment can be described by an ODE [43]:[13]

$$\begin{aligned} \frac{dx_i}{dt} &= \sum_{k=1}^{M} f_k\, Q_{ki}\, x_k - x_i\,\Phi(t)\,,\ i = 1,\ldots,M,\ \text{with} \\ \Phi(t) &= \sum_{k=1}^{M} f_k\, x_k(t)\ \text{ and } \sum_{1=1}^{M} x_i = 1. \end{aligned} \tag{23}$$

Herein the relative concentrations of the $M$ individual RNA sequences are denoted by $x_i = [X_i]$, and $Q_{ij}$ are the elements of a mutation matrix Q. These elements, in the simplest case of the uniform error rate assumption, can be expressed by an (average) error rate $p$ per site and replication:

$$Q_{ij} = p^{d_H(X_i,X_j)}\cdot(1-p)^{n-d_H(X_i,X_j)}. \tag{24}$$

The mutation probability thus is only a function of the error rate and the Hamming distance, $d_H(X_i, X_j)$, between the two sequences involved. The results of the analysis of replication-mutation kinetics have been presented and discussed extensively [44–47]. We refrain here from repeating them in detail but mention the error threshold phenomenon that confines the possibility of evolutionary optimization to mutation rates below a critical value. For constant chain length $n$ one obtains

$$p\ <\ p_{\rm crit} = 1-\sigma_m^{-\frac{1}{n}}\ \text{ with }\ \sigma_m = \frac{f_m}{\bar{f}_{-m}}\ \text{ and }\ \bar{f}_{-m} = \frac{\sum_{j=1,j\neq m}^{M} f_j x_j}{1-x_m}, \tag{25}$$

where $X_m$ is the sequence with the highest replication rate, $f_m = \max\{f_1, f_2, \ldots, f_M\}$, called the *master sequence*, $\sigma_m$ is called the *superiority* of the master sequence, and $\bar{f}_{-m}$ is the mean replication rate of all sequences except the master sequence. Simple replication rate parameter landscapes, for example the single-peak landscape with $f_m = \alpha$ and $f_j = \beta\ \forall\ j = 1,\ldots,M,\ j \neq m$ and hence $\bar{f}_{-m} = \beta$, have been used frequently in studies of equation (23) and its solutions [46, 47]. More realistic landscapes are the basis of the computer simulations described in this section. At error rates below the error threshold the population approaches an ordered stationary distribution in sequence space that has been called a *quasispecies*. In mathematical terms the quasispecies is the largest eigenvector of the matrix $\mathrm{W} = \{W_{ij} = f_j\, Q_{ij},\ i, j = 1, 2, \ldots, M\}$. If, however, the error rate increases above the critical value the ordered population structure changes abruptly into the uniform distribution

[13] We remark that the use of ODEs to describe the kinetics of mutation and selection implies that the population size, $N$, is sufficiently large for it to have no influence on the results.

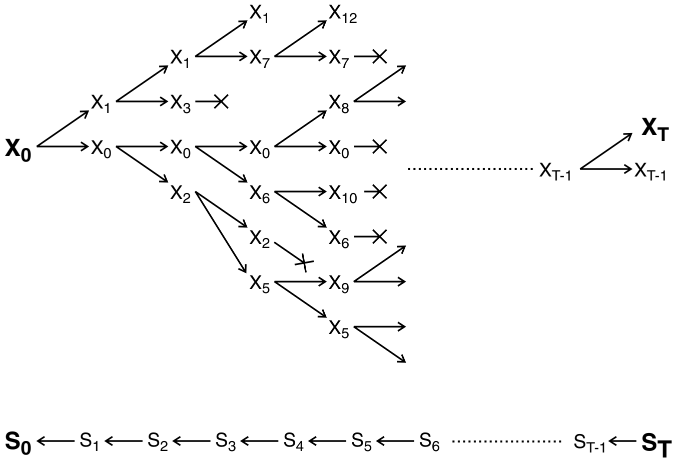


**Figure 12.** Evolutionary optimization as a multitype branching process. The sketch in the upper part shows only replication acts that lead to mutation. A full genealogy is a time ordered series, which records all individual replication acts, for example $X_0, \cdots, X_0, X_a, \cdots, X_a, X_b, \cdots, \ldots, X_{T-1}, X_T$, leading to the target. The population size is either constant (Moran model [42]) or it fluctuates around a constant value (flow reactor: $N \pm \sqrt{N}$), and hence every replication act has to be compensated by the elimination of one molecule that is tantamount to the end of some trajectory in the system. The sketch at the bottom illustrates the reconstruction of the optimization run by means of a 'relay series'.

$x_1 = x_2 = \cdots = x_M = 1/M$. In reality, no stationary population with a uniform distribution can exist since the size of the sequence space, $|\mathcal{Q}_n^{(\mathrm{AUGC})}| = 4^n$, exceeds any population size, $N$, by many orders of magnitude. As a consequence, the population can never cover whole sequence space and drifts randomly. Depending on the mutation rates and landscape structure it will drift as a whole or separate into clones of related sequences [48]. The behaviour of populations in the neighbourhood of the critical error rate depends strongly on the distribution of the replication rate parameters around the master sequence: single-peak and other steep landscapes are characterized by sharp error thresholds, whereas flat landscapes give rise to a gradual transition from quasispecies to drifting populations (see, e.g. [47, 49]).

In the ODE approach, the population dynamics is considered as a process taking place exclusively in sequence space. As in population genetics, the structures and properties of phenotypes appear in the model only as parameters. Additionally, as mentioned above, kinetic differential equations refer to an infinite population size. Accordingly, a different description is required for the study of finite size effects on evolutionary optimization. Replication and mutation of RNA molecules leading to selection in confined populations have indeed been studied in finite populations also. The best suited stochastic methods for modelling the system are multitype branching processes [50]. Simplified versions of the branching trajectories in replication and mutation are shown in figure 12. As expected, the mean values of the stochastic variables coincide with the deterministic solution [51]. Standard deviations, however, can be enormous, as we shall see in the numerical data shown below.

In order to simulate the interplay between mutation acting on the RNA sequence and selection operating on phenotypes, here the RNA structure, the sequence-structure map has to be an integral part of the model [29, 53, 54]. The simulation tool starts from a population of RNA molecules and simulates chemical reactions corresponding to replication and mutation in a continuous stirred flow reactor (CSTR) using Gillespie's algorithm [55, 56]. In target

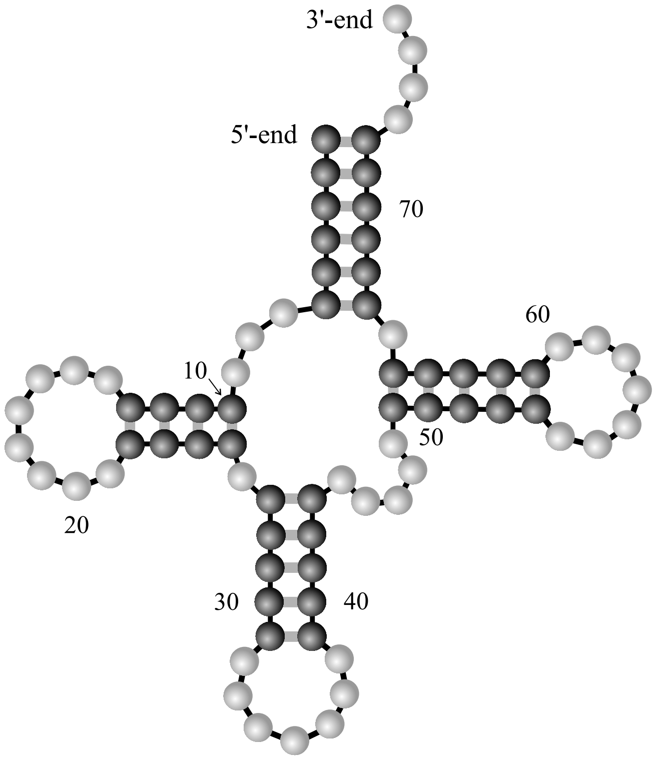


**Figure 13.** The secondary structure of a typical transfer RNA. Shown is the phenylalanyl-transfer RNA (tRNA$^{\text{phe}}$) from the yeast *Saccharomyces cerevisiae*.

search problems the replication rate of a sequence $X_k$, representing its fitness $f_k$, is chosen to be a function of the Hamming distance between the mfe structure formed by the sequence, $S_k = f(X_k)$, and the target structure $S_T$,

$$f_k(S_k, S_T) = \frac{1}{\alpha + d_H(S_k, S_T)/n}\,, \tag{26}$$

which increases when $S_k$ approaches the target ($\alpha$ is an adjustable parameter that is commonly chosen to be 0.1). A trajectory is completed when the population reaches a sequence that folds into the target structure. Accordingly, the simulated stochastic process has two absorbing barriers, the target and the state of extinction. For sufficiently large populations ($N > 30$ molecules) the probability of extinction is very small, for population sizes reported here, $N \geqslant 1000$, extinction has never been observed.

A typical trajectory is shown in figure 14. In this simulation a homogenous population consisting of $N$ molecules with the same random sequence and the corresponding structure is chosen as the initial condition. The target structure is the well-known secondary structure of phenylalanyl-transfer RNA (tRNA$^{\text{phe}}$) shown in figure 13. The mean distance to the target of the population decreases in steps until the target is reached [29, 48, 54]. Individual (short) adaptive phases are interrupted by long quasi-stationary epochs. In order to reconstruct the optimization dynamics, a time ordered series of structures was determined that leads from an initial structure, $S_I$, to the target structure, $S_T$. This series, called the *relay series*, is a uniquely defined and uninterrupted sequence of shapes. It is retrieved through backtracking, that is in the opposite direction from the final structure to the initial shape (see the lower

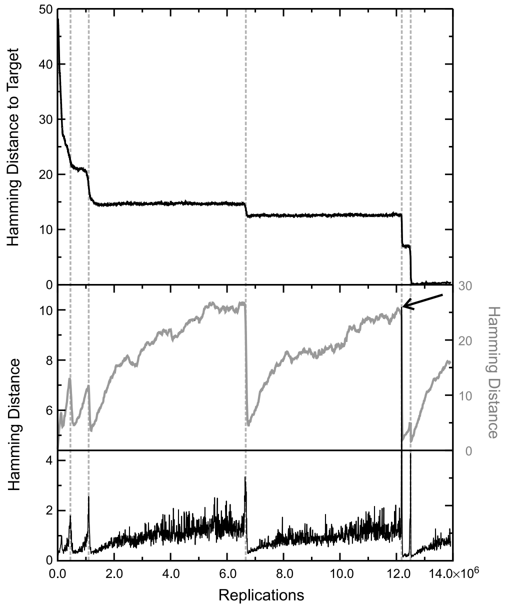


**Figure 14.** A trajectory of evolutionary optimization. The topmost plot presents the mean distance to the target structure of a population of 1000 molecules. The plot in the middle shows the width of the population in Hamming distance between sequences, and the plot at the bottom is a measure of the velocity with which the centre of the population migrates through sequence space. Diffusion on neutral networks causes spreading on the population in the sense of neutral evolution [52]. A remarkable synchronization is observed: at the end of each quasi-stationary plateau a new adaptive phase in the approach towards the target is initiated, which is accompanied by a drastic reduction in the population width and a jump in the population centre (the top of the peak at the end of the second long plateau is marked by a black arrow). A mutation rate of $p = 0.001$ was chosen, the replication rate parameter is defined in equation (26), and the initial and target structures are shown below table 8.

part of figure 14). The procedure starts by highlighting the final structure and traces it back during its uninterrupted presence in the flow reactor until the time of its first appearance. At this point we search for the parent shape from which it descended by mutation. Now we record the time and structure, highlight the parent shape, and repeat the procedure. Recording further backwards yields a series of shapes and times of first appearance which ultimately ends in the initial population.[14] Usage of the relay series and its theoretical background allows classification of transitions [29, 58]. Inspection of the relay series together with the sequence record on the quasi-stationary plateaus provides hints for the distinction of two scenarios:

(i) The structure is constant and we observe neutral evolution in the sense of Kimura's theory of neutral evolution [59]. In particular, the number of neutral mutations accumulated is proportional to the number of replications in the population, and the evolution of the population can be understood as a diffusion process on the corresponding neutral network [52] (see also figure 14).

[14] It is important to stress two facts about relay series: (i) The same shape may appear two or more times in a given relay series. Then, it is extinct between two consecutive appearances. (ii) In contrast to a genealogy, which is the complete recording of parent–offspring relations in the form of a time-ordered series of genotypes, the relay series lists only changes in shapes.

**Table 8.** Statistics of the optimization trajectories. The table shows the results of sampled evolutionary trajectories leading from a random initial structure, $S_I$, to the structure of tRNA$^{\text{phe}}$, $S_T$, as the target[a]. Simulations were performed with an algorithm introduced by Gillespie [55–57]. The time unit is here undefined. A mutation rate of $p = 0.001$ per site and replication were used. The mean and standard deviation were calculated under the assumption of a log-normal distribution that fits well the data of the simulations.

| Alphabet | Population size, $N$ | Number of runs, $n_R$ | Real time from start to target | | Number of replications [$10^7$] | |
|---|---|---|---|---|---|---|
| | | | Mean value | $\sigma$ | Mean value | $\sigma$ |
| **AUGC** | 1 000 | 120 | 900 | +1380 −542 | 1.2 | +3.1 −0.9 |
| | 2 000 | 120 | 530 | +880 −330 | 1.4 | +3.6 −1.0 |
| | 3 000 | 1199 | 400 | +670 −250 | 1.6 | +4.4 −1.2 |
| | 10 000 | 120 | 190 | +230 −100 | 2.3 | +5.3 −1.6 |
| | 30 000 | 63 | 110 | +97 −52 | 3.6 | +6.7 −2.3 |
| | 100 000 | 18 | 62 | +50 −28 | – | – |
| **GC** | 1 000 | 46 | 5160 | +15700 −3890 | – | – |
| | 3 000 | 278 | 1910 | +5180 −1460 | 7.4 | +35.8 −6.1 |
| | 10 000 | 40 | 560 | +1620 −420 | – | – |

[a] The structures $S_I$ and $S_T$ were used in the optimization:

```
S_I:  ((.(((((((((((((............(((....)))......)))))).))))))).))...(((......)))
S_T:  ((((((...((((........)))).(((((.......))))).....(((((.......))))).))))))....
```

(ii) The process during the stationary epoch involves several structures with identical replication rates, and the relay series reveals a kind of random walk in the space of these neutral structures.

The diffusion of the population on the neutral network is illustrated by the plot in the middle of figure 14 that shows the width of the population as a function of time [48, 60]. The population width increases during the quasi-stationary epoch and sharpens almost instantaneously after a sequence had been created, that allows the start of a new adaptive phase in the optimization process. The scenario at the end of the plateau corresponds to a *bottleneck* of evolution. The lower part of the figure shows a plot of the migration rate or drift of the population centre and confirms this interpretation: the drift is almost always very slow unless the population centre 'jumps' from one point in sequence space to another point in sequence space where the molecule initiating the new adaptive phase is located. A closer look at the figure reveals the coincidence of the three events: (i) beginning of a new adaptive phase, (ii) collapse-like narrowing of the population spread, and (iii) jump-like migration of the population centre.

Table 8 collects some numerical data obtained from repeated evolutionary trajectories under identical conditions.[15] Individual trajectories show enormous scatter in the time or the number of replications required to reach the target. The mean values and the standard deviations were obtained from the statistics of the trajectories under the assumption of a log–normal distribution. Despite the scatter, three features are unambiguously detectable:

(i) The search in **GC** sequence space takes about five time as long as the corresponding process in **AUGC** sequence space, in agreement with the difference in neutral network structure discussed above.
(ii) The time to target decreases with increasing population size.
(iii) The number of replications required to reach the target increases with population size.

[15] 'Identical' means here that everything was kept constant except the seeds for the random number generators.

Combining items (ii) and (iii) allows a clear conclusion concerning the time and material requirements of the optimization process: fast optimization requires large populations, whereas economic use of material suggests working with small population sizes just sufficiently large to avoid extinction.

Systematic studies on the parameter dependence of RNA evolution were reported in a recent simulation [61]. An increase in the mutation rate leads to an error threshold phenomenon that is close to one observed with quasispecies on a single-peak landscape as described above [47]. Evolutionary optimization becomes more efficient[16] with increasing error rate until the error threshold is reached. A further increase in the error rate leads to an abrupt breakdown of the optimization process. As expected, the distribution of replication rates or fitness values, $f_k$, in sequence space is highly relevant too: a steep decrease in fitness with the distance to the master structure represented by the target that has the highest fitness value leads to the sharp threshold behaviour as observed with single-peak landscapes, whereas flat landscapes show a broad maximum of optimization efficiency without an indication of a threshold-like behaviour.

## 4. Kinetic folding of RNA

In the neutral network concept and in evolutionary optimization we assigned the mfe structure, a single conformation, to every sequence and were interested in the changes in structure resulting from changes in the sequence. Every ground state or conformation of lowest free energy is accompanied by a large set of suboptimal states, and often these suboptimal states contribute to the molecular properties at biologically relevant temperatures ($273\,\mathrm{K} < T < 373\,\mathrm{K}$). For the properties of RNA molecules not only the existence of suboptimal conformations is important but also their interconnections on the time scale: which conformations can be reached from given initial states and how long does it take for them to appear? RNA switches represent a class of molecules which take part in genetic and metabolic regulation by means of conformational changes (see section 6).

In figure15 three different notion of structure are illustrated by means of energy level diagrams: (i) conventional RNA folding assigns the mfe structure to the sequence (leftmost diagram); (ii) suboptimal conformations accompany the mfe-structure (middle diagram) and contribute to the molecular properties in the sense of a Boltzmann ensemble, the partition function is the proper description of the RNA molecule at thermodynamic equilibrium or in the limit of infinite time; and (iii) at finite time the situation is different, since a single molecule may have one or more long-lived metastable conformations in addition to the mfe structure (rightmost diagram showing the energy levels and saddle point of an 'RNA switch'). Then the observed molecular structure depends also on the initial conditions and on the time window of the experiment. The relation between the energy levels of the suboptimal structures and their role in folding kinetics is introduced by means of the transitions states relating them. We see structures that are very close on the energy scale, for example $S_0$ and $S_1$ or $S_7$ and $S_8$ in figure 15, but difficult to reach from each other because they are separated by a high free energy barrier. To go from $S_2$ to $S_0$, on the other hand, is much easier, because they belong the the same *basin* of conformations. In this section we shall illustrate the problem of kinetic folding using the concept of the *conformation space* of RNA secondary structures. The lifetime of conformations in a landscape with multiple local minima depends on the barriers between them. For RNA secondary structures these barriers can be readily computed [62, 63],

[16] The efficiency of evolutionary optimization is measured by the average and best fitness values obtained in populations after a predefined number of generations.

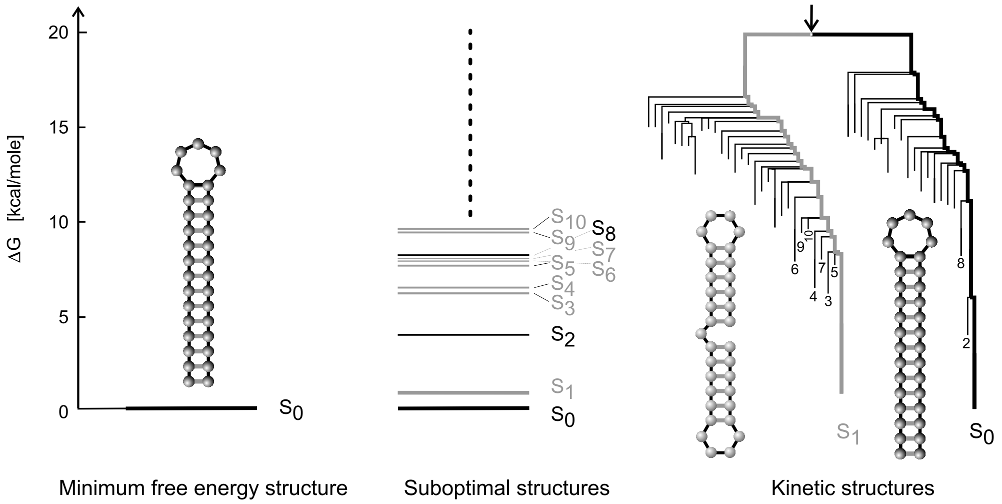


**Figure 15.** Three different notions of structure. The mfe structure is shown as the only relevant conformation on the left-hand side corresponding in a formal sense to the zero temperature limit ($\lim T \to 0$). In the middle we show the set of suboptimal structures as it is considered at equilibrium and temperature $T$ in the form of the partition function. The notion of the equilibrium structure implies the limit of infinite time ($\lim t \to \infty$). On the right-hand side we show the barrier-tree of a molecule which exemplifies a situation that is encountered, for example, in RNA switches. At finite time we may find one or more long-lived conformations in addition to the mfe structure.

and Arrhenius kinetics on the barrier landscape provides a reliable tool for estimating these lifetimes [64].

The kinetic aspects of folding RNA sequences into secondary structures were already considered in the mid-1980s [65]. A large number of algorithms aiming at a computation of folding kinetics were developed later [63,66–75]. Most of them treat whole stacks as single unities and this is justified for long stacks because of the well known cooperativity of stack formation [76]. In the secondary structures of RNA molecules we have long as well as small stacks and occasionally even single base pairs in native structures (for an example see the 5S RNA in figure 27). Only a few approaches to kinetic folding resolve the process to elementary steps that involve single base pairs or single nucleotides [63, 74, 75]. We shall be concerned with an algorithm at single nucleotide resolution here in subsection 4.3.

### *4.1. Conformation space*

The set of all structures that are compatible with a single sequence $X$, defined in equation (7), is a commonly high dimensional subspace of shape space. This subspace is the conformation space of the sequence unless a restriction is introduced in the form of neglecting all high energy conformations that are not required as saddle points in transitions between (meta)stable conformations. The conformation space represents the structural diversity of conformations that are accessible from the ground state, $S_0$, on thermal excitation. A free energy landscape on conformation space is a useful tool for understanding and modelling the dynamics of conformational changes. A metastable structure is tantamount to a local minimum of this free energy surface. The two move sets discussed in the context of a measure of distance on shape space (section 2.2) are also relevant for conformation space since they are tantamount to elementary

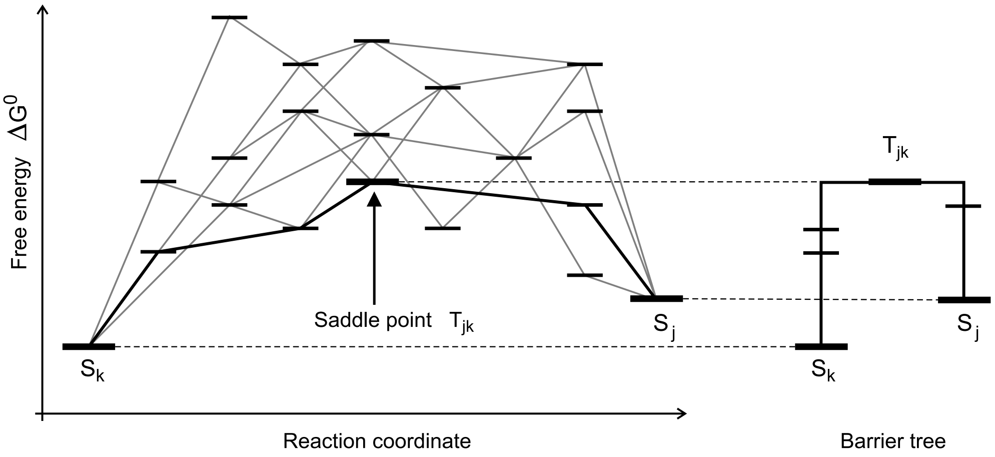


**Figure 16.** Construction of barrier trees. The set of suboptimal conformations is related by a move set as shown in the left-hand part of the sketch. The barrier tree is derived from the set of suboptimal structures by eliminating all conformations except local minima of the free energy surface and minima connecting saddle points of lowest free energy. We remark that the set of local minima depends on the choice of the move set, although important local minima are very unlikely to be changed for physically meaningful alterations of the move set.

moves in kinetic folding of RNA (figure 18, see also [63,74,75]). It is important to note that the set of local minima in conformation space that is assigned to an RNA sequence by kinetic folding depends on the move set applied. The two move sets sketched in figure 18, however, agree in all major local minima corresponding to sufficiently long lived metastable conformations.

Assigning a free energy value to every point in conformation space yields the conformational energy landscape that can be used as a metaphor or as a tool to calculate kinetic folding trajectories. Such a free energy landscape of a biopolymer is a highly complicated object since the energy depends on nonlocal interactions and the carrier of the landscape, and the conformation space is high-dimensional. For most purposes a simplification of the conformational energy landscape called the *barrier tree* is sufficient [63]. A barrier tree is constructed from the set of suboptimal conformations and their free energies. For pairs of local minima the lowest path connecting them is selected and only three points are retained: the two minima and the saddle (figure 16). Applying the construction principle to entire conformation space yields the barrier tree of the molecule (see section 4.3).

### *4.2. Suboptimal structures and partition functions*

Algorithms for the computation of suboptimal conformations have been developed, and two of them are frequently used [77,78]. As we have already seen from our estimate, the numbers of suboptimal states are very large and, moreover, they increase exponentially with chain length, $n$. The first algorithm [77] is convenient and efficient but misses certain classes of conformations. The latter of the two algorithms [78] has been designed for calculation of all conformations within a given energy band above the mfe and adopts a technique originally proposed for suboptimal alignments of sequences [79]. The algorithm starts from the same dynamic programming table as the conventional mfe conformation but considers all backtracking results within the mentioned energy band. As indicated in figure 15 the set of structures, mfe and suboptimal conformations $\{S_0, S_1, S_2, \ldots\}$, is ordered since their free energies, $\{\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots\}$ fulfil the relation $\varepsilon_0 \leqslant \varepsilon_1 \leqslant \varepsilon_2 \ldots$.

At equilibrium and temperature $T$ the conformations form a Boltzmann ensemble that contains $S_j$ with the Boltzmann weight $\gamma_j(T) = g_j \, \exp\big(-(\varepsilon_j - \varepsilon_0)/RT\big)/Q(T)$, where $R$ is the Boltzmann constant for 1 mole, $R = N_L \cdot k_B$, and $Q(T)$ is the partition function[17]

$$Q(T) = \sum_i g_i \exp\Big(-(\varepsilon_i - \varepsilon_0)/RT\Big). \tag{27}$$

Instead of having an mfe structure with defined base pairs, the ground state is now described by a temperature dependent linear combination of states where the weighted superposition of base pairs gives rise to base pairing probabilities $p_{ij}(X, T)$ which are the elements of the matrix

$$P(X, T) = \sum_k \gamma_k(T)\, A(S_k) \quad \text{or} \quad p_{ij}(X, T) = \sum_k \gamma_k(T)\, a_{ij}(S_k)\,, \tag{28}$$

which is a Boltzmann weighted superposition of the adjacency matrices (1) of the individual structures with the following properties: in the limit $T \rightarrow 0$ the base pairing probabilities converge to the base pairing pattern of $S_0$ (for a non-degenerate ground state, $\varepsilon_0 < \varepsilon_1$) as described by the adjacency matrix $A(S_0)$, and in the limit $T \rightarrow \infty$ all (micro)states have equal weights and the partition function converges to the total number of all conformations of the sequence $X$. An elegant algorithm that computes the partition function $Q(T)$ directly by dynamic programming is found in [17]. It has been incorporated into the Vienna RNA package [12].

Sequences folding into the same mfe structure, which are sequences belonging to one neutral network, differ strongly with respect to the values of the mfes as well as the numbers of suboptimal structures. For the purpose of illustration we choose again the simple double-hairpin structure of chain length $n = 33$ (figure 2): the distribution of mfes fulfils $\overline{\Delta G_0} = -5.74 \pm 2.37\,\text{kcal mol}^{-1}$. The spread of the distribution is best illustrated by the smallest and the largest values: $(\Delta G_0)_{\min} = -15.80$ and $(\Delta G_0)_{\max} = -0.40\,\text{kcal mol}^{-1}$. The numbers of suboptimal suboptimal structures in an interval of $10\,\text{kcal mol}^{-1}$ even more broadly distributed: $\bar{n}_{\text{subopt}} = 12\,000 \pm 10\,600$ with a smallest and a largest value of $(\bar{n}_{\text{subopt}})_{\min} = 353$ and $(\bar{n}_{\text{subopt}})_{\max} = 92{,}406$, respectively. Although there is no perfect correlation, molecules with lower free energies have fewer suboptimal conformations within a given energy band above the mfe. In table 9 we show the two extreme cases, $X_1^{(\text{dhp33})}$ and $X_3^{(\text{dhp33})}$, together with two more sequence examples, $X_2^{(\text{dhp33})}$ and $X_4^{(\text{dhp33})}$, lying close to the extreme cases. On the other hand, the sequences with the lowest and highest free energies are not extreme with respect to the numbers of states and thus the mfe and the number of suboptimals are neither independent nor fully correlated properties:

| | | |
|---|---|---|
| GAUCGGGGUGGUUUGAAGAAGAGUAGUGAACUU: | $\Delta G_0 = -0.40\,\text{kcal mol}^{-1}$ | $\bar{n}_{\text{subopt}} = 51\,610$ |
| CUAAUAGCAUCCUAUUCCCCGAGACAGUAUCUU: | $\Delta G_0 = -0.40\,\text{kcal mol}^{-1}$ | $\bar{n}_{\text{subopt}} = 14\,962$ |
| GGGCAUAGGCGUGUGUGAUUCGAGCAUCUUUCG: | $\Delta G_0 = -2.30\,\text{kcal mol}^{-1}$ | $\bar{n}_{\text{subopt}} = 92\,406$ |
| CCUAGGAGGGAUCUUGUAUGCUCGGCGCUUGAG: | $\Delta G_0 = -2.30\,\text{kcal mol}^{-1}$ | $\bar{n}_{\text{subopt}} = 87\,792$ |
| UUCGGCCGAUGGGCUGCCUAGCCGAGAUCCGGU: | $\Delta G_0 = -10.60\,\text{kcal mol}^{-1}$ | $\bar{n}_{\text{subopt}} = 4\,406$ |
| ACGCGUUUCCAAACGCAAAUGCCCAGGAAGGGC: | $\Delta G_0 = -11.50\,\text{kcal mol}^{-1}$ | $\bar{n}_{\text{subopt}} = 399$ |
| CAGAGUGGUGCCGCUCGAAGCCCCAAUACGGGG: | $\Delta G_0 = -13.60\,\text{kcal mol}^{-1}$ | $\bar{n}_{\text{subopt}} = 353$ |
| AAGGGCGGCGACGCCCACUCGGCGCGAAACGCU: | $\Delta G_0 = -15.80\,\text{kcal mol}^{-1}$ | $\bar{n}_{\text{subopt}} = 529$ |

Without going into details we remark that other often-postulated relations are anything but perfect: the energy gap between the mfe and the free energy of the first suboptimal conformation $(S_1)$, $\Delta\varepsilon_{0\rightarrow 1}$, correlates neither well with the value of the mfe nor with the weight of the mfe

[17] Sometimes different microstates $S_i$ with the same free energy $\varepsilon_j$ are lumped together to form one 'mesoscopic' state in the partition function and then the factor $g_j$ accounts for this degeneracy.

**Table 9.** Suboptimal structures and partition functions $Q(T)$. Compared are four sequences folding into the same mfe structure, the double hairpin $S^{(\mathrm{dhp33})}$ with very small ($X_1^{(\mathrm{dhp33})}$, $X_2^{(\mathrm{dhp33})}$) and very large numbers of suboptimal structures in a free energy interval of 10 kcal mol$^{-1}$ above the mfe. Apart from the mfe values ($\Delta G_0$) we present also the free energies computed from the partition function ($T = 37°$C). The last sequence, $X^{(\mathrm{swt33})}$, has a single hairpin mfe structure (section 4.4). All energy values in the table are given in kcal mol$^{-1}$. The table presents further the Boltzmann weight of the mfe structure in the partition function.

| Sequence | Free energy | | Fraction of mfe structure in $Q(T)$ | Number of suboptimal structures | |
|---|---|---|---|---|---|
| | $\Delta G_0^{(310)}$ | $\Delta G_{0,Q}^{(310)}$ | | mfe ↔ mfe + 10.0 | mfe ↔ 0.0 |
| $X_1^{(\mathrm{dhp33})}$ | −13.60 | −13.69 | 0.8633 | 353 | 2 501 |
| $X_2^{(\mathrm{dhp33})}$ | −11.50 | −11.76 | 0.6556 | 399 | 948 |
| $X_3^{(\mathrm{dhp33})}$ | −2.30 | −4.06 | 0.0577 | 92 406 | 131 |
| $X_4^{(\mathrm{dhp33})}$ | −2.30 | −4.30 | 0.0391 | 87 792 | 172 |
| $X_5^{(\mathrm{dhp33})}$ | −10.60 | −11.26 | 0.3440 | 4 406 | 6 268 |
| $X^{(\mathrm{swt33})}$ | −26.30 | −26.50 | 0.7196 | 343 | 253 970 |

[a] The following sequences, all folding into the structure $S^{(\mathrm{dhp33})}$, were used:

$X_1^{(\mathrm{dhp33})}$: CAGAGUGGUGCCGCUCGAAGCCCCAAUACGGGG
$X_2^{(\mathrm{dhp33})}$: ACGCGUUUCCAAACGCAAAUGCCCAGGAAGGGC
$X_3^{(\mathrm{dhp33})}$: GGGCAUAGGCGUGUGUGAUUCGAGCAUCUUUCG
$X_4^{(\mathrm{dhp33})}$: CCUAGGAGGGAUCUUGUAUGCUCGGCGCUUGAG
$X_5^{(\mathrm{dhp33})}$: UUCGGCCGAUGGGCUGCCUAGCCGAGAUCCGGU
$X^{(\mathrm{swt33})}$: GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC

$S^{(\mathrm{dhp33})}$: ..(((((....)))))....((((.....))))

structure in the partition function. Table 9 finally shows the number of suboptimal states up to zero energy, the free energy of the open chain. As expected, more stable molecules tend to have more suboptimal conformations with negative energies.

Two examples of calculations of partition functions and their evaluation in dot-plots are shown in figure 17. We chose one example with a very large Boltzmann weight of the mfe structure ($X_1^{(\mathrm{dhp33})}$, upper diagram) and, indeed, the dot-plot of the partition function shows almost exclusively the squares of the most stable conformation. The second example is one with a rather low weight, and here we can see directly the contributions of many other suboptimal states to the base pairing probabilities $p_{ij}(X, T)$.

### 4.3. Folding kinetics

Kinetic folding of RNA molecules can be understood and modelled as a stochastic process in RNA conformation space. The process corresponds to a time ordered series of secondary structures, a trajectory

$$\Omega_0 \rightarrow \Omega_1 \rightarrow \Omega_2 \rightarrow \cdots \rightarrow \Omega_{\mathrm{T}}, \tag{29}$$

where the initial and target structures, $\Omega_0$ and $\Omega_{\mathrm{T}}$, may be chosen at will. Commonly, $\Omega_0 = \mathbf{O}$ and $\Omega_{\mathrm{T}} = S_0$ are used, corresponding to the open chain and the mfe structure, respectively. Individual trajectories (29) may contain loops, i.e. the same structure may be visited two or more times. In general, it is of advantage to define the target conformation as an absorbing state. Leaving the target state unconstrained causes the trajectory to approach a thermodynamic ensemble in the sense that it visits the individual conformations with frequencies according to the Boltzmann weights. For practical purposes the time required to fulfil the condition of

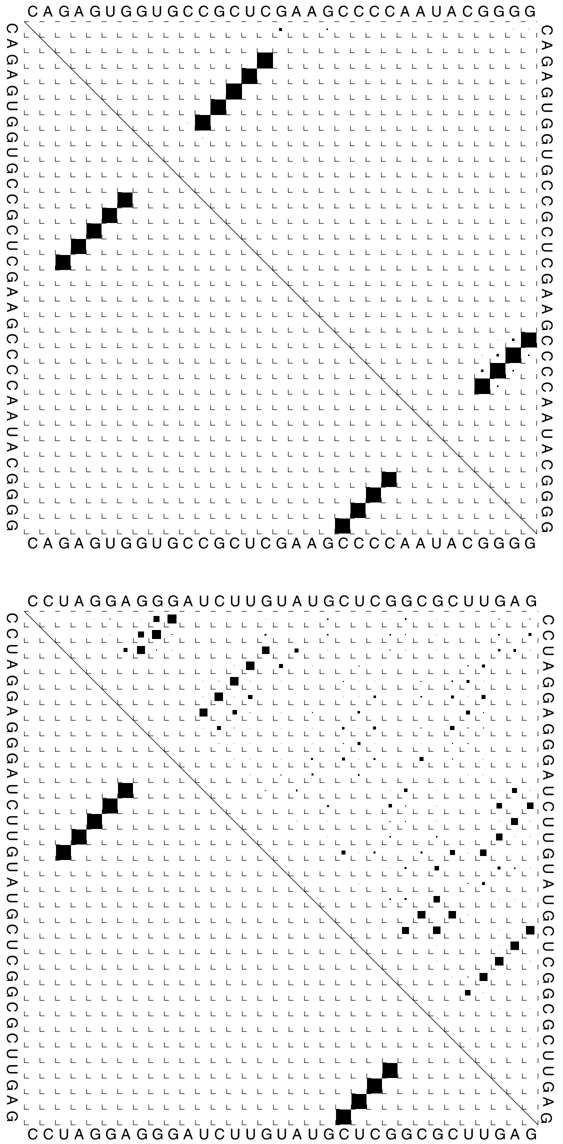

C A G A G U G G U G C C G C U C G A A G C C C C A A U A C G G G G
C C U A G G A G G G A U C U U G U A U G C U C G G C G C U U G A G

**Figure 17.** Partition functions as 'dot-plots'. The two diagrams show the partition functions of two sequences, $X_1^{(\mathrm{dhp}33)}$ and $X_4^{(\mathrm{dhp}33)}$ (see caption of table 9), both forming the structure $S^{(\mathrm{dhp}33)}$ as the mfe. The lower left triangle shows the mfe structure as a dot-plot: each black square is tantamount to a '1′ in the corresponding position of the adjacency matrix. The upper right triangle shows the base pairing probabilities, where the size of the square corresponds to the value of $p_{ij}(X, T)$ in equation (28).

ergodicity, however, is prohibitively long. Basic to the stochastic process is a set of moves that defines the allowed transitions between conformations. In the simplest case the move set contains base pair closure and base pair opening according to the conventional secondary structure rules (conditions I to III). Such a move set corresponds to the base pair distance, $d_P$, as a metric in shape space. It turned out to be relevant to introduce also a shift move (figure 18), since the trajectories approach the target much faster then [63] and there is experimental

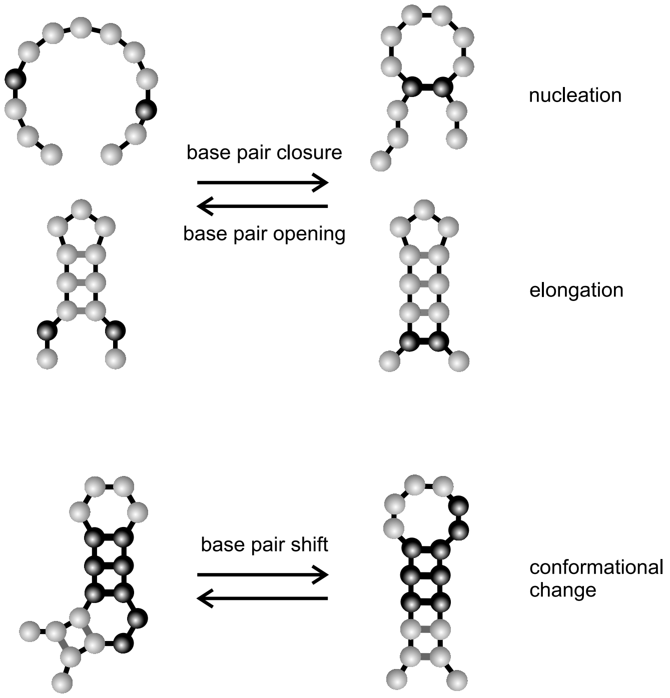


**Figure 18.** Move sets used in kinetic folding of RNA. Move set I considers only base pair closure and base pair opening. The difference between nucleation and elongation is reflected by the difference in the free energy values. Move set II contains, in addition, a shift move. For a different goal—simulation of RNA folding and measure of distance in shape and conformation space—one may restrict the shift to single base pairs or allow for simultaneous shifts of groups of stacked pairs (three base pairs shifting together in the lower part of the figure).

evidence for shifting groups of base pairs [76, 80]. If the move set is extended to simultaneous shifts of as many nucleotides as possible within a given substructure element, the Hamming metric between the symbolic or parentheses notations of structures, $d_H(S_i, S_j)$, turns out to be the proper measure of distance.

The stochastic process (29) is described by a master equation for the ensemble probabilities: $P_k(t)$ is the probability observing the conformation $S_k$ at time $t$. The time derivatives fulfil the equation

$$\frac{dP_k}{dt} = \sum_{i=0}^{m+1} (P_{ik}(t) - P_{ki}(t)) = \sum_{i=0}^{m+1} k_{ik} P_i - P_k \sum_{i=0}^{m+1} k_{ik}, \qquad (30)$$
$$\text{with } k = 0, 1, \ldots, m+1 \text{ and } i \to k \in \text{themove set.}$$

We assume that the open chain conformation $\mathbf{O} = S_{m+1}$ is part of the suboptimal conformations, $S_1, \ldots, S_m$. The transition probabilities are computed from the free energies of the conformations:

$$P_{ik}(t) = k_{ik}\, P_i(t) = P_i(t)\, e^{-(\Delta G_k - \Delta G_i)/(2RT)} / \Sigma_i, \qquad (31)$$
$$P_{ki}(t) = k_{ki}\, P_k(t) = P_k(t)\, e^{-(\Delta G_i - \Delta G_k)/(2RT)} / \Sigma_k, \qquad (32)$$
$$\text{with } \Sigma_j = \sum_{i=0, i\neq j}^{m+1} \exp\big(-(\Delta G_j - \Delta G_i)/(2RT)\big).$$

In order to avoid the necessity of additional parameters the free energies are taken from the suboptimal foldings. Calibration of the time scale occurs through adjusting the folding kinetics of a model system to the experimental data [81]. Although it is straightforward to solve the master equation (30) by means of an eigenvalue problem, practical difficulties arise from the enormously high number of suboptimal conformations determining the dimensionality of the system [64].

A simplification of full kinetic folding is introduced in the form of 'barrier trees' (figure 16). All suboptimal conformations that neither represent a local minimum of the conformational energy landscape nor a lowest energy transition state between two local minima are neglected. The remaining barrier tree can be used to simulate kinetic folding by means of conventional Arrhenius kinetics. The results are often in astonishingly good agreement with the exact computations based on equation (30). Cases of less satisfactory agreement can be predicted [64].

Again we illustrate kinetic folding of RNA molecules by means of the same example as used before, the double hairpin structure $S^{(\mathrm{dhp33})}$. Later we shall consider an RNA switch, an especially designed molecule that sustains two different long lived conformations. The sequence of the double hairpin molecules was chosen arbitrarily, $X_5^{(\mathrm{dhp33})}$= UUCGGCCGAUGGGCUGCCUAGCCGAGAUCCGGU. We begin by considering the partition function shown in figure 19. The Boltzmann weight of the mfe structure is 0.3440 (table 9) and, in addition to the two hairpins of the ground state conformation, $S_0$,[18] we recognize traces of longer single hairpins. Next we compute and consider the barrier tree of the molecule (figure 20). The restriction to local minima and saddle points reduces the 6268 conformations with non-positive free energies (table 9) to the mfe structure plus 120 local minima including the open chain, $S_{120}$. In addition to an appreciable number of smaller basins we recognize four major folding families with free energies 50% of the mfe or lower: $\{S_9\}$, $\{S_1\ ,\ S_2$; including $S_{16}\}$, $\{S_3\ ,\ S_4\}$, and $\{S_0\ ,\ S_7$; including $S_{15}\}$.

The folding kinetics from the open chain into the mfe structure of the double hairpin molecule with sequence $X_4^{(\mathrm{dhp33})}$ is shown in figure 21. The computation yields the relative concentrations, $x_k(t)$, for all 121 conformations ($k = 0, \ldots, 120$) as a function of time. Most of the concentrations are so small that the curves coincide with the $t$-axis for all practical purposes. Therefore we show only the 12 states that reach higher concentration values. The time order in which the first six intermediates appear corresponds precisely to the sequence of the first six branching saddles in the barrier tree: $S_{76} < S_{106} < S_9 < S_{16} < S_7$. In addition, we see that states belonging to the same basin disappear together: $\{S_3\ ,\ S_4\}$ or $\{S_1\ ,\ S_2\}$. The good agreement between the Arrhenius kinetics and the stochastic simulation is remarkable. The double hairpin structure with sequence $X_4^{(\mathrm{dhp33})}$ is a typical inefficient folder: only about 50% of the molecules fold into the mfe directly, whereas the second half of the ensemble stays for a relatively long time in one of the states $\{S_1\ ,\ S_2\ ,\ S_3\ ,\ S_4\ ,\ S_9\}$. Eventually we consider the structures of the 12 conformations (figure 22) in order to provide an explanation for the nature of the basins. Conformations in the same basin share major structural features: $S_0$ and $S_7$, for example, share the (lhs) tetraloop of the double hairpin structure and this is apparently the structure whose nucleation is more difficult than that of the (rhs) pentaloop, $S_1$ and $S_2$ share the inner loop structure with five base pairs, $S_3$ and $S_4$ have the outer six base pairs in common, and $S_{16}$ has the same four inner base pairs as $S_1$ and $S_2$.

As the last example of this subsection we present the kinetics of the conformational change between the mfe structure, $S_0$, and the first suboptimal conformation, $S_1$, of an RNA

[18] For the sake of simplicity we shall omit the superscript '$^{(\mathrm{dhp33})}$' in the forthcoming discussion. The numbered structures in figure 20 correspond to suboptimal conformations, which are recorded and appreciably populated as folding intermediates in the process from the open chain $S_{120}$ to the mfe structure $S_0$.

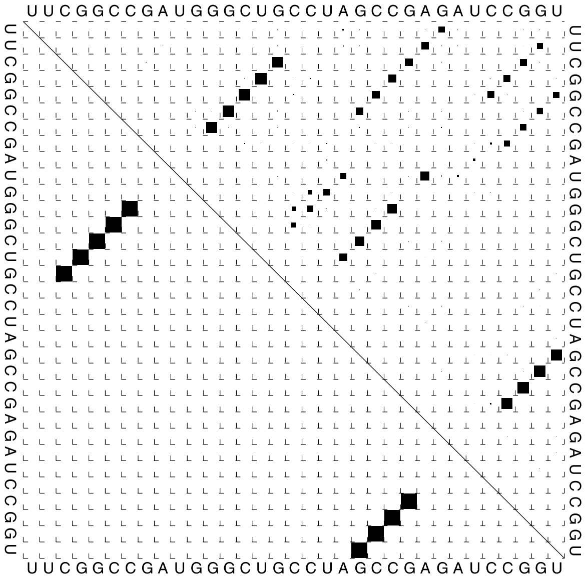


**Figure 19.** Partition function of sequence $X_5^{(\mathrm{dhp33})}$ as a dot-plot. In addition to the two small hairpins of the mfe structure, conformations with single long hairpins can be readily recognized.

switch (figure 24). This RNA switch is a molecule of chain length $n = 33$ with the sequence $X^{\mathrm{swt33}}$=(GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC) that has been designed to have in essence only two conformations, a long hairpin as the mfe structure, $S_0$, and a metastable double hairpin $S_1$. Indeed, the partition function contains only contributions from the two conformations $S_0$ and $S_1$ (figure 23). In table 9 we see that the Boltzmann weights are about 0.72 and 0.28, respectively. The kinetic curves shown in figure 24 show a pure two state transition in both directions. From the difference in free energies, $\Delta G_0(S_0) = -26.30\,\mathrm{kcal\,mol^{-1}}$ and $\Delta G_0(S_1) = -25.30\,\mathrm{kcal\,mol^{-1}}$, we compute a difference in the transition times of approximately 5 that fits very well the results of the Arrhenius kinetics. Comparing the dot-plots and the barriers trees of the two systems, the randomly chosen double hairpin structure and the designed switch, we see that the major effect of the design was to eliminate minor basins and conformations, which are unfavourable for the transition.

### *4.4. RNA molecules with multiple structures*

The barrier trees considered in the previous (section 4.3) indicate that the energy surface of a typical RNA sequence has a large number of local minima with often high energy barriers separating different basins of attraction. Thus non-native conformations can have energies comparable to the ground state, and they can be separated from the native state by very high energy barriers.

In order to deal with multiple conformations, we consider a collection of structures (matchings) $\Omega_1, \Omega_2, \ldots, \Omega_k$ on the same sequence $X$. The fundamental question in this context is whether there is a sequence in

$$\mathbf{C}[\Omega_1, \Omega_2, \ldots, \Omega_k] = \bigcap_{j=1}^{k} \mathbf{C}[\Omega_j] \tag{33}$$

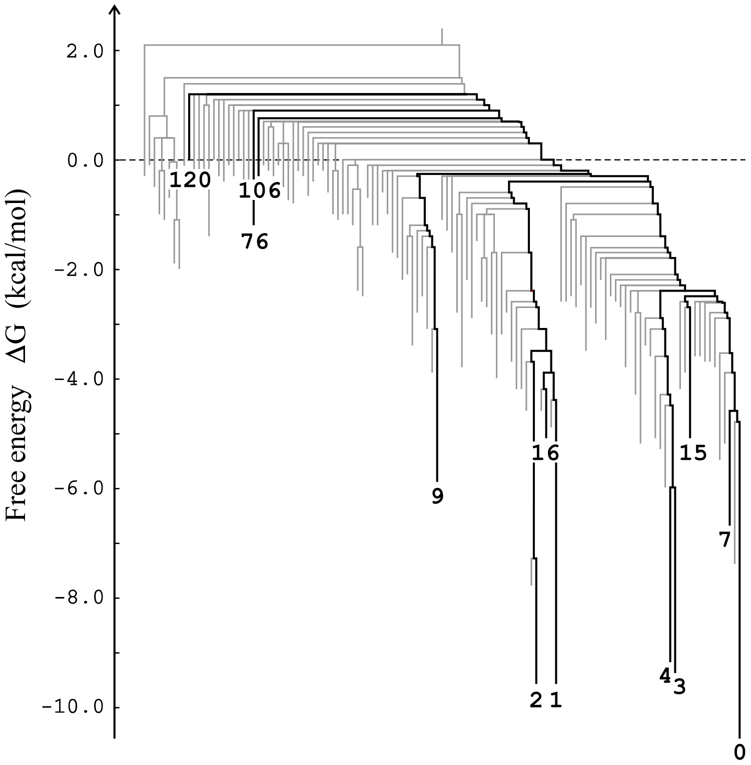


**Figure 20.** Barrier tree of the double hairpin molecule with sequence $X_5^{(\mathrm{dhp33})}$. The numbered suboptimal conformations correspond to intermediates that are sufficiently populated to be recognized in the recording of the folding kinetics (figure 21).

and if so, what the size of this *intersection* of sets of compatible sequence is. To answer this question, it is useful to consider the graph $\Psi$ with vertex set $\{1, \ldots, n\}$ and edge set $\bigcup_{j=1}^{k} \Omega_j$.

Generalized Intersection Theorem. Suppose $\mathcal{B} \subseteq \mathcal{A} \times \mathcal{A}$ contains at least one symmetric pair, i.e. $\mathsf{XY} \in \mathcal{B}$ implies $\mathsf{YX} \in \mathcal{B}$. Then

(i) $\mathbf{C}[\Omega_1, \ldots, \Omega_k] \neq \emptyset$ if $\Psi$ is bipartite.
For $k = 2$, $\Psi$ is a disjoint union of paths and cycles with even length and hence always bipartite.

(ii) The number of sequences that are compatible with all structures can be written in the form

$$\left|\mathbf{C}[\Omega_1, \Omega_2, \ldots, \Omega_k]\right| = \prod_{\text{components } \psi \text{ of } \Psi} F(\psi), \tag{34}$$

where $F(\psi)$ is the number of sequences that are compatible with the connected component $\psi$.

(iii) For the biophysical alphabet $\bigcap_j \mathbf{C}[\Omega_j] \neq \emptyset$ holds if and only if $\Psi$ is a bipartite graph.
In particular, for the case of bistable sequences, $k = 2$, we can express the size of the intersection explicitly in terms of Fibonacci numbers,

$$\begin{aligned} F(P_k) &= 2\Big(\mathrm{Fib}(k) + \mathrm{Fib}(k+1)\Big) = 2\mathrm{Fib}(n+2), \\ F(C_k) &= 2\Big(\mathrm{Fib}(k-1) + \mathrm{Fib}(k+1)\Big), \end{aligned} \tag{35}$$

where $P_k$ and $C_k$ are the path and cycle components of $\Psi$ with $k$ vertices.

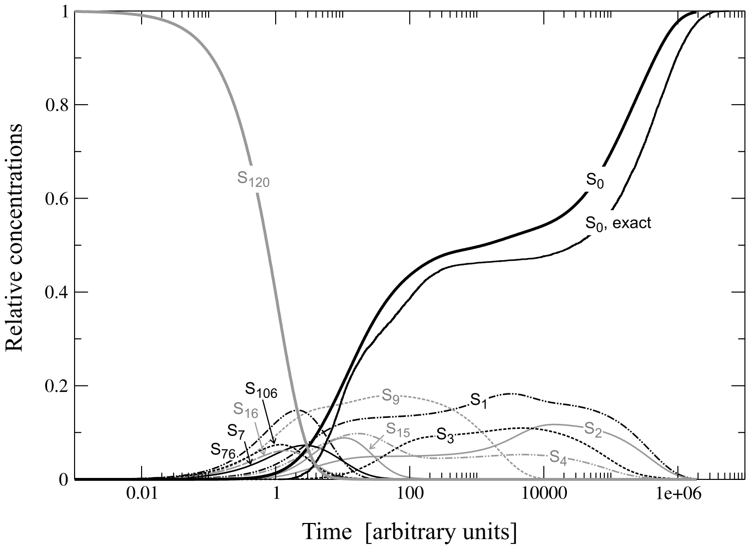


**Figure 21.** Folding kinetics of the double hairpin molecules with sequence $X_5^{(\mathrm{dhp33})}$. Shown are the 12 relative concentrations of several intermediates that contribute appreciably to the total concentration. All curves are obtained from an Arrhenius-type kinetics as described in detail in [64] except the curve marked '$S_0$, exact', which was calculated through sampling trajectories of the folding process according to [63]. In order to make the Arrhenius kinetics and the simulation of the stochastic process comparable we defined the process leading to the mfe structure $S_0$ to be irreversible, corresponding to $S_0$ being an absorbing state.

| | | |
|---|---|---|
| $X_4^{(\mathrm{dhp33})}$ | UUCGGCCGAUGGGCUGCCUAGCCGAGAUCCGGU | |
| $S_{120}^{(\mathrm{dhp33})}$ | ................................. | 0.00 |
| $S_{106}^{(\mathrm{dhp33})}$ | ..........((....))............... | - 0.30 |
| $S_{76}^{(\mathrm{dhp33})}$ | ...(.((...)).)................... | - 1.20 |
| $S_{16}^{(\mathrm{dhp33})}$ | ...........((((....)))).......... | - 5.10 |
| $S_{15}^{(\mathrm{dhp33})}$ | ...(.((...)).)......((((.....)))) | - 5.10 |
| $S_{9}^{(\mathrm{dhp33})}$ | .(((((((.((((...))))..)).)..)))). | - 5.90 |
| $S_{7}^{(\mathrm{dhp33})}$ | ..(((((....))))).................. | - 6.70 |
| $S_{4}^{(\mathrm{dhp33})}$ | ((((((....((....))..)))))))....... | - 9.20 |
| $S_{3}^{(\mathrm{dhp33})}$ | ((((((...((((...)))))))))))....... | - 9.40 |
| $S_{2}^{(\mathrm{dhp33})}$ | ....((((.(.((((....)))).)....)))) | - 9.60 |
| $S_{1}^{(\mathrm{dhp33})}$ | .((((....(.((((....)))).)...)))). | - 9.60 |
| $S_{0}^{(\mathrm{dhp33})}$ | ..(((((....)))))....((((.....)))) | - 10.60 |

**Figure 22.** Suboptimal conformations in folding the double hairpin structure $S^{(\mathrm{dhp33})}$. Conformations in the same basin are related in structure. Examples are $\{S_0$ and $S_7\}$, $\{S_1$, $S_2$, and $S_{16}\}$, $\{S_3$ and $S_4\}$. Conformation $S_{15}$ is related to $S_7$ and $S_0$, but has the more-difficult-to-nucleate left hairpin misfolded (triloop instead of tetraloop). Free energies in kcal mol$^{-1}$.

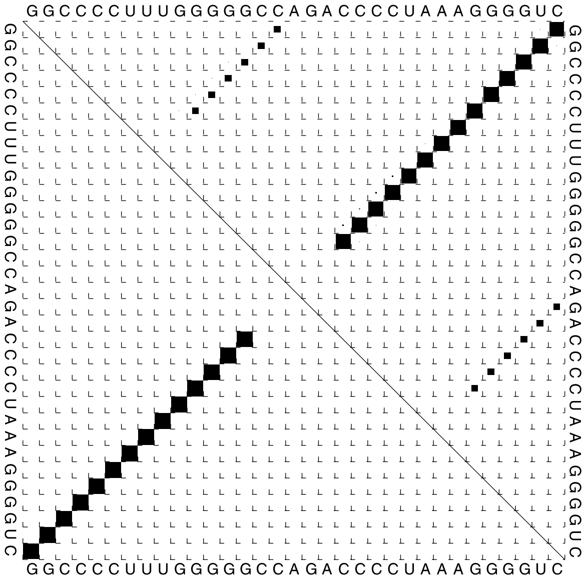


**Figure 23.** Partition function of sequence $X^{(\mathrm{swt33})}$ as a dot-plot. The molecule has essentially two conformations, the long hairpin, $S_0$, consisting of a stack of 14 base pairs closed by a pentaloop, and the metastable conformation $S_1$ consisting of two hairpin loops, each with six base pairs closed by a tetraloop, that are joined by one single nucleotide. The dot-plot indicates dominance of the mfe structure accompanied by the double hairpin conformation and no other structure present at a detectable concentration at equilibrium.

For a proof of these propositions see [36, 82]. Interestingly, for two structures there is always a non-empty intersection $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]$. In contrast, the chance that the intersection of three randomly chosen structures is non-empty decreases exponentially with sequence length [83]. Recently, an alternative attempt has been made to extend the design aspect of the intersection theorem to three or more sequences [84].

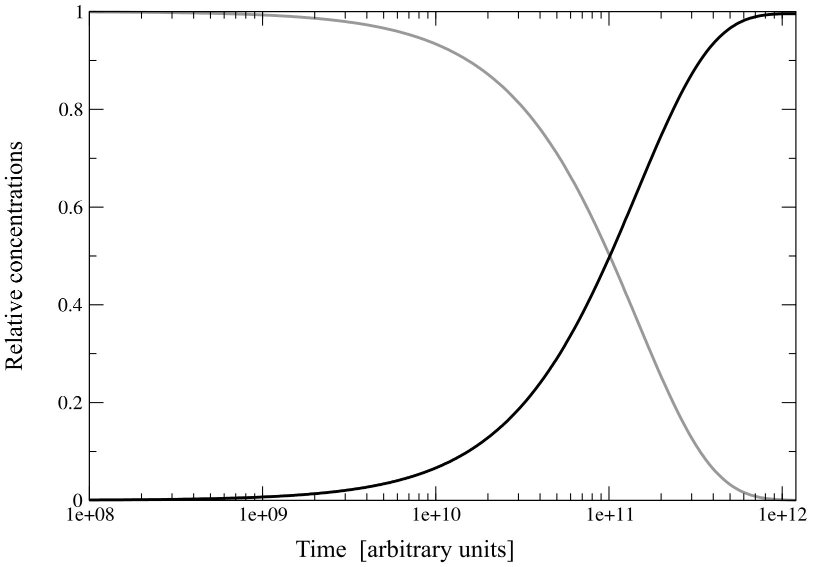
Given a collection of alternative secondary structures, we can again ask the *inverse folding* or *sequence design* question. For simplicity we restrict ourselves to two structures $\Omega_1$ and $\Omega_2$ here. For example, one might be interested in sequences that have two prescribed structures $\Omega_1$ and $\Omega_2$ as stable local energy minima with roughly equal energy, for which the energy barrier between these two minima is roughly $\Delta E$. It is not hard to design a cost function $\Xi(X)$ for this problem. In [82], the following ansatz has been used successfully:

$$\begin{aligned}\Xi(X) = {} & E(X, \Omega_1) + E(X, \Omega_2) - 2G(X) + \\ & + \xi\,\big(E(X, \Omega_1) - E(X, \Omega_2)\big)^2 + \zeta\,\big(B(X, \Omega_1, \Omega_2) - \Delta E\big)^2 . \end{aligned} \tag{36}$$

Here, $B(X, \Omega_1, \Omega_2)$ is the energy barrier between the two conformations $\Omega_1$, $\Omega_2$ which can be readily computed from the barrier tree of the sequence $X$.

## 5. Confronting RNA secondary structure prediction with reality

A comprehensive discussion of the state of the art in RNA structure prediction would justify a review in its own right with many different facets. Here we shall only consider two possibilities to estimate or improve secondary structure prediction of RNA molecules: (i) estimates of



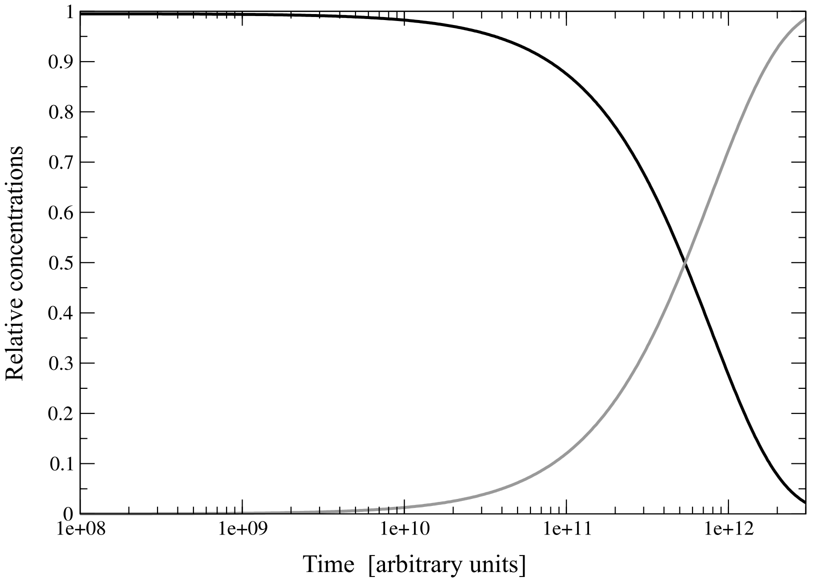


**Figure 24.** Transitions between long-lived conformations in the RNA molecule with the sequence $X^{\text{swt33}}$. The calculations of the conformational changes are carried out by means of Arrhenius-type kinetics on the barrier tree shown in figure 15. The upper part shows the transition from the first suboptimal conformation ($S_1$, double hairpin; grey curve) into the mfe structure ($S_0$, single hairpin; black curve). The lower plot presents the kinetics of the transition in the opposite direction, $S_0 \rightarrow S_1$. Because of the energy difference, $\Delta\varepsilon_{0\rightarrow 1} = 1.0\,\text{kcal mol}^{-1}$, the transition $S_0 \rightarrow S_1$ occurs about a factor 5 times slower than $S_1 \rightarrow S_0$.

the reliability of secondary structure prediction and (ii) the inclusion of certain classes of tertiary interactions like pseudoknots, end-on-end stacking and others that introduce structural constraints on the stereochemically unintegrated double-helical regions of the secondary structure.

### *5.1. Reliability of secondary structure prediction*

Medium size and large RNA molecules have suboptimal conformations of energies that are close to the mfe. As we have seen, such a situation may occur also for small RNA molecules. Low-lying suboptimal states provide a severe obstacle to the prediction of molecular structures for several reasons:

(i) The parameters for RNA and DNA secondary structures are of limited accuracy because model compounds are rare or even lacking for certain classes of base and base pair interactions.

(ii) Tertiary interactions are neglected that may lead to changes in the energetic sequence of conformations (see next section 5.2).

(iii) Kinetic folding rather than thermodynamics may determine the structures found in nature [70, 71].

In this subsection we shall consider the effect of low-lying suboptimal states accounted for by the partition function on the reliability of predicted structures, modules or segments of structures. The basis of the reliability measures is the well-justified assumption that errors in the empirical parameters will change the ordering of conformations on the energy axis but not the nature of the suboptimal structures. Base pairs that occur in most conformations have a high probability and are predicted reliably, therefore. Two different measures will be used for this goal: (i) the base pairing probability, $p_{ij}(X, T)$, as directly obtained from the partition function and (ii) the pairing entropy of a given nucleotide $x_i(X)$,

$$s_i(X, T) = -\sum_{j=1}^{n} p_{ij}(X, T) \ln\, p_{ij}(X, T), \tag{37}$$

where we have absorbed the contribution of remaining unpaired, $p_i^{(u)}$, in the diagonal element,[19]

$$p_i^{(u)}(X, T) = p_{ii}(X, T) = 1 - \sum_{j=1, j\neq i}^{n} p_{ij}(X, T). \tag{38}$$

The base pairing probability, $p_{ij}$, refers to an individual pair $i$–$j$ in a given structure, commonly an mfe structure, and hence the two nucleotides in a pair have always the same probability of being predicted correctly. A pairing probability close to 1 implies a high reliability of prediction. For an unpaired base the value in the base pairing probability plot is the probability of remaining unpaired (38). In the case of the pairing entropy, $s_i$, the estimate refers to a single nucleotide, and $s_i$ and $s_j$ may be different although the two nucleotides form a pair. An entropy value close to zero implies complete determinism and high reliability of the prediction. High entropy values correspond to high uncertainty. An upper value of the pairing entropy can be estimated from the uniform distribution of pairing probabilities, $p_{ij} = 1/n$:

$$(s_i)_{\max} = -\sum_{i=1}^{n} \frac{1}{n} \ln \frac{1}{n} = \sum_{i=1}^{n} \frac{1}{n} \ln\, n = \ln\, n.$$

Although the two reliability measures may differ in detail, they have the same reference in the sense that the unique assignment of bases to base pairs is given by probability 1, $p_{ij} = 1$, corresponding to zero entropy, $s_i = s_j = 0$, and indeed they yield very similar results in actual applications. Three examples are discussed here: (i) a randomly chosen small molecule with a large variety of suboptimal conformations ($X_4^{(\mathrm{dhp33})}$), (ii) phenylalanyl-transfer RNA as an example of a molecule with an evolutionarily optimized rigid structure, and (iii) 5S ribosomal

[19] This assumption converts the matrix $P(X, T) = \{p_{ij}(X, T)\}$ into a bistochastic matrix.

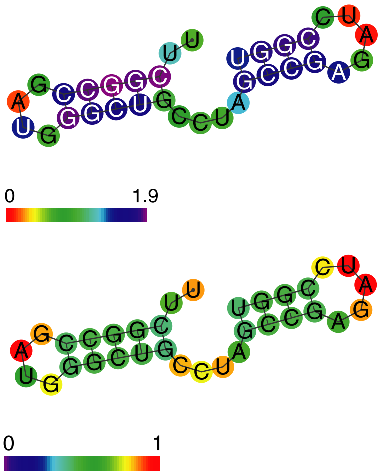
RNA as an example of a molecule consisting of a rigid and easy-to-predict part and a flexible module with low-lying suboptimal conformations.

In order to illustrate the two measures proposed for reliability estimates of secondary structures, we consider first the molecules $X_4^{(\mathrm{dhp33})}$ that have been used to illustrate kinetic folding with intermediates (figure 25). Only three positions (9, 27, and 28, marked in red) are unpaired in (almost) all[20] conformations and can be predicted with certainty therefore. Because of the single-hairpin structure of the two lowest suboptimal conformations, all base pairs have relatively small probabilities or high entropies. With this particulary flexible small RNA molecule itself we encounter a problem that cannot be answered by purely thermodynamic reliability measures: whether folding ends up with a double-hairpin or a single-hairpin structure is a question of folding kinetics (see section 4.3).

The phenylalanyl-tRNA secondary structure has been chosen as the second example because it represents a rigid RNA molecule and allows us to study the effect of base modifications on the structure: three guanine residues are modified by methylation (M) or conversion into the Y-base (Y), and four uracil residues are converted into dihydro-uracil (D), thymidine (T) or pseudouridine (P) residues. Among other functions, base modifications are believed to stabilize the structure and facilitate folding. The effect on structure stabilization is easily recognized in figure 26. In the molecule without modification all stacks except the terminal stack are uncertain and only a few unpaired nucleotides are predicted with high reliability (yellow, orange or red). This uncertainty is also reflected, for example, by the frequency of the mfe in the Boltzmann ensemble of conformations: as shown in table 10 this frequency is about seven times larger in the case of the sequence with modified bases. The spectrum of conformations, eventually, provides the final piece of the puzzle through coarse graining secondary structures and looking only for clover-leaves. The sequence without modified bases has no clover-leaf structure as the mfe and shows clover-leaf structures only at positions 2, 4 and 15–19 in an energy ordered set of suboptimal conformations. In the molecule with modified nucleotides the situation is entirely different: the first seven conformations are clover-leaves, and they are accompanied by further clover-leaves at positions 9, 12–14 and 16–19. Thus modification of bases reinforces the clover-leaf structure and, moreover, in the specific way its done in the tRNA$^{\mathrm{phe}}$ sequence it canalizes folding into the specific structure of this molecule. It is also worth mentioning that base modification in tRNA$^{\mathrm{phe}}$ has a strong effect on kinetic folding: the process is not only speeded up, it is also made more efficient in the sense that a very high percentage of the molecules fold directly in the clover-leaf [63]. In the unmodified molecule approximately 50% fold fast and the rest reach the mfe structure in a very slow process via long-lived folding intermediates (or folding traps).

The third and last example is an RNA molecule that is somewhat larger than a tRNA, the ribosomal 5S RNA from the archaebacterium *Methanospirillum hungatei* with $n = 126$ nucleotides. It has been chosen because it illustrates very well the usefulness of the reliability concept for structure predictions. The molecule as shown in figure 27 falls into two parts: (i) the stack in the middle and the hairpin on the right-hand side of the structure are predicted with a probability very close to 1 and the prediction matches perfectly the experimental (native) structure; and (ii) the multiloop on the left-hand side has a low pairing probability and is different from the native structure that contains a hairpin loop interrupted by an internal loop. Accordingly, the prediction of the left part of the molecule (figure 27) is unreliable and indeed we observe substantial disagreement between the predicted mfe structure and the native structure of the molecule. It is also interesting to evaluate the position of the native

[20] All 19 lowest conformations shown in figure 25 have an unpaired base at these positions.



```
                                 ΔG0 [kcal/mol]
..(((((....)))))....((((.....))))  -10.60
.((((....(.((((....)))).)...)))).   -9.60
....((((.(.((((....)))).)....))))   -9.60
((((((...((((...)))))))))).......   -9.40
((((((....((....))..)))))).......   -9.20
.((((....(.(((......))).)...)))).   -9.20
..(((....(.((((....)))).)...)))..   -9.20
...((((....)))).....((((.....))))   -9.20
....((((.(.(((......))).)....))))   -9.20
((((((...(((.....))))))))).......   -9.10
..(((((....))))).....(((.....))).   -9.10
.(((((...((((...)))))))))........   -9.00
.(((((....((....))..)))))........   -8.80
..((((......))))....((((.....))))   -8.80
..(((....(.(((......))).)...)))..   -8.80
.(((((...(((.....))))))))........   -8.70
..(((((....)))))....(((.......)))   -8.40
((((((....(((...))).)))))).......   -8.10
.....(((.(.((((....)))).)....))).   -8.00
```

**Figure 25.** Illustration of the reliability of secondary structure prediction. We consider a small RNA molecule that is derived from a random sequence $X_4^{(\mathrm{dhp33})}$. The upper drawing on the left-hand side shows the colour coded pairing entropy, and the base pairing probability is shown in the drawing in the lower part. The table on the right-hand side shows structures and energies of the mfe-conformation together with the 18 suboptimal configurations of lowest energy. Marked in red are three unpaired bases, which occur in all 19 structures and are predicted correctly (almost) with probability 1. Two positions with relatively high uncertainty are marked in violet and blue. Colour code: red→orange→yellow→green→blue→violet . . ., decreasing reliability. The range for the calculated base pairing entropies is $0 \leqslant s_i < 1.9$. Hence, the largest actually observed value is well below the maximal possible entropy, $\ln 33 = 3.5$.

structure within the spectrum of suboptimal conformations of the molecule: the native structure is conformation #329 936 and lies 8.39 kcal mol$^{-1}$ above the mfe. For this particular molecule we counted 1,814,405 conformations within an interval of of 10 kcal mol$^{-1}$ above the ground state. In addition, computed the degree of neutrality of this not very frequent structure and obtained $\bar{\lambda} = 0.20 \pm 0.04$, which is somewhat less than the corresponding quantity in the tRNAs (table 6). It is also interesting that two nucleotides were highly conserved on the neutral network: **C**$_{68}$ and position 69 with 96% **A** and 4% **G**. The native sequence has **C**$_{68}$ and **G**$_{69}$, respectively.

Low base pair probabilities or high pairing entropies are a useful indication of low prediction reliability. As said already, close energy values of conformations can easily lead to sequence inversion on the energy scale for small changes in the parameters. Low base pair probabilities, however, can result from other effects as well: two long-lived conformations, for example $S_0$ and $S_1$ of $X^{\mathrm{swt33}}$ (see also the sketch in figure 15), give a similar superposition result as two conformations that are readily converted into each other because they are separated by a small barrier only. For the high barrier molecule $X^{\mathrm{swt33}}$, both structures are well defined on time scales shorter than the time constant of the conversion, which is about $10^{11}$ time units compared to the folding of $X^{\mathrm{dhp33}}$, where folding is completed in $10^6$ time units. If the time of experimental observation is shorter than the time of interconversion of the conformations, the superposition is a spurious result of the partition function, and Boltzmann ensembles restricted to the basin of $S_0$ or $S_1$ corresponding to either of the two base pairing patterns detectable in figure 23 (dominant or less probable) would be appropriate.

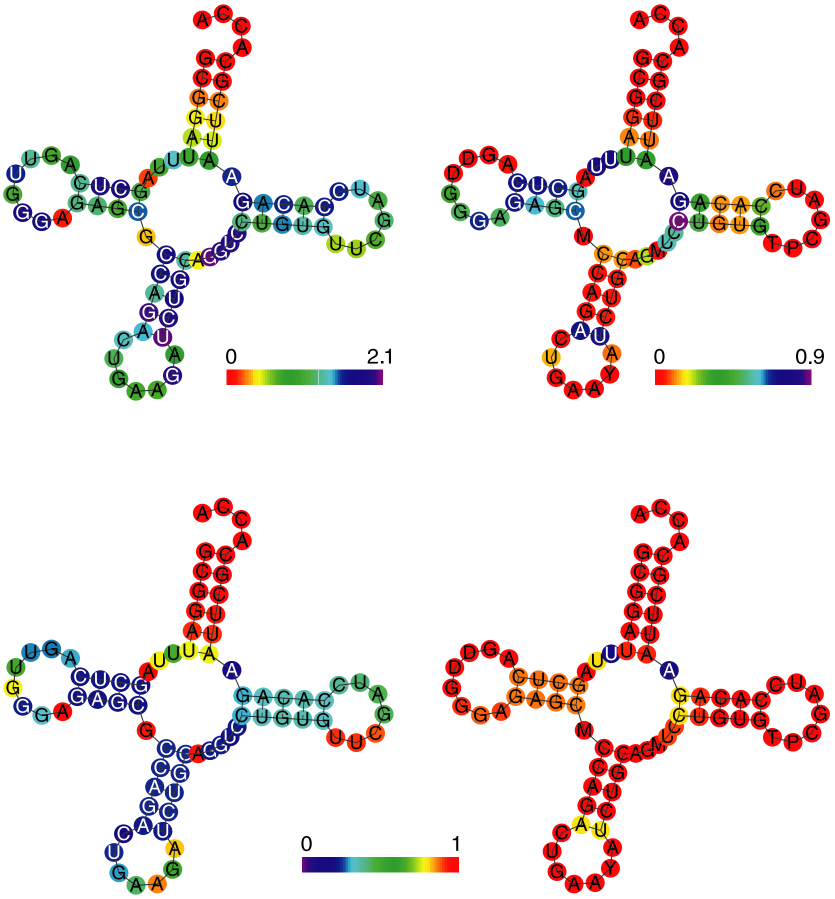


**Figure 26.** Reliability of secondary structure prediction of the tRNA$^{\mathrm{phe}}$ from yeast (*Saccharomyces cerevisiae*). The figure shows colour coded base paring entropies (upper part) and base pairing probabilities (lower part). On the left-hand side we show the results for the pure four letter sequence without modifications. It is important to note that the entropy scale is different in the two drawings: for the unmodified molecule we have $0 < s < 2.1$ compared to $0 < s < 0.9$ in the case of the modified bases. The drawings on the right-hand side refer to the sequence with base modifications (for sequence, structure, and base modifications see [85–87]). Since sufficient empirical data are not available for modified bases, we generally excluded them from Watson–Crick base pairing. Colour code: see caption of figure 25.

## *5.2. Pseudoknots and other tertiary interactions*

Tertiary interactions are not considered in our definition of secondary structures mainly for two reasons: (i) the parameters available for most of these interactions are much less reliable than those for the conventional secondary structure elements because only few experimental examples are available, and (ii) the inclusion of some of these elements is in conflict with the efficient dynamic programming algorithm for thermodynamic folding. Several tertiary interactions were already observed in the 3D-structure of tRNA$^{\mathrm{phe}}$ [85]: examples are a pseudoknot of ‘kissing loop type’ with strand inversion ($\mathbf{G}_{18}$–$\boldsymbol{\Psi}_{55}$, $\mathbf{G}_{19}$–$\mathrm{C}_{56}$), several non Watson–Crick base pairs, three base triplets ($\mathbf{G}_{22}$–$\mathrm{C}_{13}$+$\mathbf{M}_{46}$, $\mathbf{A}_{23}$–$\mathrm{U}_{12}$+$\mathbf{A}_{9}$, $\mathbf{C}_{25}$–$\mathrm{G}_{10}$+$\mathbf{G}_{45}$), and end-on-end stacking of RNA double helices. Consideration of

**Table 10.** Comparison of tRNA$^{\mathrm{phe}}$ secondary structure and suboptimal conformations without and with modified nucleotides. 'clvlf' stands for clover-leaf, a structure with a multiloop that is surrounded by four stacks (see figure13), 'frequency' refers to the statistical weight of the mfe structure in the Boltzmann ensemble, and the conformation number is the number index of the native structure in the energy ordered spectrum of suboptimal structures (the mfe structure is no. 1 in this list). All energies in kcal mol$^{-1}$.

| Sequence | Mfe structure | | | Native structure | | No. of structures ⟨10 kcal mol$^{-1}$⟩ |
|---|---|---|---|---|---|---|
| | Energy | Structure | Frequency | Energy | Conf. No. | |
| Unmodified | −23.80 | no clvlf | 0.064 | −22.40 | 19 | 398,180 |
| Modified | −21.50 | clvlf | 0.437 | −20.50 | 3 | 26,512 |

end-on-end stacking provided a straightforward explanation for the L-shape of tRNAs, which was first surprising because the combination of two times two stacks from the clover-leaf secondary structure in order to form longer double helices was completely unexpected. Later on, this list has been extended by many other transferable types of interactions between nucleotides (for a more recent comprehensive classification of base pairs see [89]).[21]

A detailed discussion of tertiary interactions would justify a review in its own right and cannot be given here. Instead we shall present a few key references on pseudoknots since they came recently into the focus of interest because of their importance in ribozyme structure and function. In addition they are easier to incorporate into standard RNA prediction routines than most of the other tertiary interactions. The first algorithm computing pseudoknots are about 25 years old [66,90]. The next important step in secondary structure prediction of RNA including pseudoknots was an elegant dynamic programming algorithm [13] which, however, suffered from a rather prohibitive $\mathcal{O}(n^6)$ complexity in time and $\mathcal{O}(n^4)$ in storage requirements. A faster and less demanding ($\mathcal{O}(n^5)$) dynamic programming algorithm allow the computation of the partition function including pseudoknots [91]. A database for RNA pseudoknots has been installed [92] and provides a solid empirical basis for further improvement of pseudoknot prediction. Kinetic folding algorithms based on stochastic base pair or base stack formation and cleavage (subsection 4.3) are not restricted to pseudoknot free structures and are indeed used for structure predictions including pseudoknots and knots [63, 73, 93, 94]. Recently, systematic work was done on the classification of pseudoknot topologies because of their decisive role in RNA function [95, 96]. We mention also a recent prediction heuristic that outperforms the dynamic programming based algorithms [97]. Finally, there are new prediction methods for structures with pseudoknots, which are based on several aligned sequences that are known to form identical structures [98–100]. Alignment based methods, in general, reach higher reliability in their predictions, because they combine the information from several sequences [101, 102].

## 6. Evolutionary and rational design of RNA molecules *in vitro*

The concepts based on applying sequence to structure mappings and their inversions to RNA secondary structures have been tested and used in evolution biotechnology and rational design of RNA molecules. Only a short account of some experimental results from selection and

[21] 'Transferable' indicates that the same type of base interaction has been found several times in different RNA molecules.

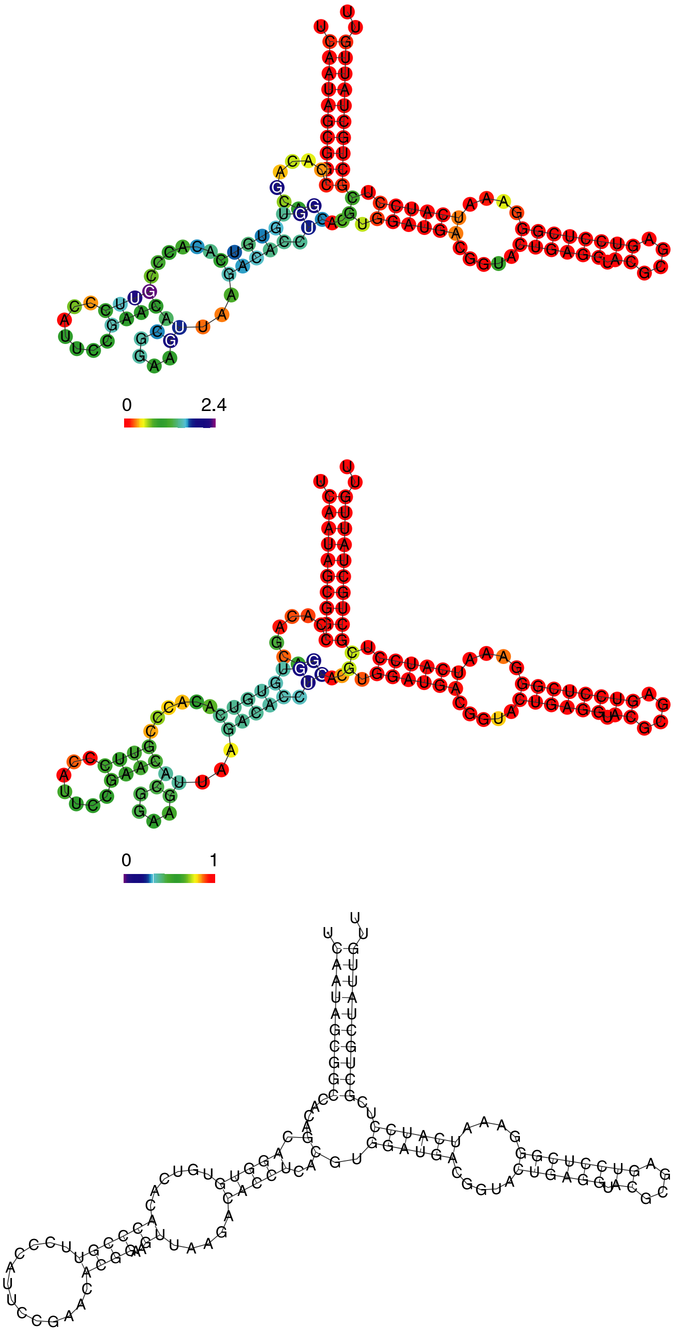


**Figure 27.** Reliability of secondary structure prediction of the 5S RNA from *Methanospirillum hungatei*. The topmost drawing shows the colour coded pairing entropy, the drawing in the middle presents the base pairing probability, and the structure at the bottom is the native structure [88]. Colour code: see caption of figure 25.

evolution of *aptamers*[22] and *ribozymes*[23] can be given here. We order them by the topics of this review.

*Properties of RNA molecules with random and natural sequences.* A few theoretical investigations dealt with random pools of RNA sequences [108–111]. In contrast to proteins and other biopolymers, almost all random RNA sequences of sufficient chain length ($n > 20$ for {**AUGC**}) form stable structures. A typical random structure has a characteristic stack length [108], has many internal loops and bulges, and looks irregular. Natural RNA molecules, on the other hand, are the result of many millions of years of selection processes and differ in many aspects from random RNA molecules. Natural RNAs, for example, have lower free folding energies, than the average of random energies, thus demonstrating the effect of evolutionary selection for stable structures. Natural RNAs fulfil multiple purposes and they are optimized according to several criteria in the sense of a process heading for a point on the Pareto front. The notion of barrier trees, in particular those of local minima, saddles and basins, can be extended to landscapes built upon partially ordered sets (posets) [112].

*RNA molecules from restricted alphabets.* Attempts to evolve RNA ligation catalyzing ribozymes from restricted alphabets have been successful [113, 114]. The first example uses **AUG**, the alphabet that lacks cytidine and was motivated by prebiotic reasoning since cytosine is less stable than the other three nucleotide bases and might have been very rare under primordial conditions. As predicted from secondary structure statistics [108], longer stacks are required for stability and the isolated ribozyme without **C** [113] contains indeed only longer stacks than the parent {**AUGC**}-molecule from which it was derived by directed evolution. The attempt to make an RNA ligase that is free of **G** and **C** has not been successful. Replacing **A** by 2,6-diamino-purine (**D**), which makes a stronger base pair than **A**, however, allowed is to produce functional ribozymes [114]. Again this molecule has substantially longer stacks than the parent ligase. The results obtained with the weakly binding restricted alphabets, {**AUG**} and {**AU**} or {**DU**}, fully agree with the predictions of nonexisting stable structures in tables 4 and 6.

*Neutral networks.* The existence of neutral networks and neutral paths in real RNA molecules has been demonstrated by several experimental studies on selection of RNA molecules with predefined properties, in particular aptamers and ribozymes (examples in [115–118]). The search for a multipurpose ribozyme by Schultes and Bartel [115] revealed a long neutral path through sequence space along which the secondary structures stayed unchanged and the catalytic efficiencies of the two ribozymes, a ligase and a cleavage enzyme, remained constant at values as high as the reference molecules. An obvious question is, Why show the artificial RNA molecules and the computer calculations such a high degree of neutrality whereas functional tRNA molecules tolerate only very limited sequences variability? An answer is provided by the data in table 7: cofolding with other RNA molecules can be regarded as a model for multiple constraints, and the values in the table show the reduction in the degree of neutrality with more binding partners. Indeed, the artificial ribozymes are almost unconstrained and the tRNAs

[22] Aptamers are RNA molecules that bind specifically other molecules. They are commonly prepared using the *SELEX* technique through variation and selection [26, 103, 104].

[23] ‘Ribozyme’ is a new word created through merging ribo(nucleic acid en)zyme. It characterizes a catalytically active RNA molecule. The first ribozymes were found in nature [105–107]. Later on a great variety of ribozymes with diverse catalytic functions ranging from organic catalysis—for example the Diels–Alder reaction—to RNA replication has been prepared by means of variation and selection.

are typical multiple task fulfilling molecules. It is worth mentioning that the occurrence of neutral networks is not restricted to RNA molecules; they were also found with lattice protein models [119] and with full protein structures [120–123].

*Shape space covering.* Folding complete sequence spaces followed by enumeration of structures has shown that a relatively small number of common structures is opposed by a large number of rare structures [28, 37]. In particular, the frequency-rank ordered distribution of structures shows a modified power law that levels off at the high frequency end.[24] Only common sequences are relevant for evolutionary biotechnology and natural evolution because it is very unlikely to find a rare structure by random searches or evolutionary strategies. Sequences folding into common structures are distributed all over sequence space[25], and therefore it is not necessary or sufficient to search the whole sequence space in order to find a given secondary structure. With probability 1 a sequence folding into the structure is contained in a (high dimensional) sphere of computable radius in sequence space around every arbitrarily chosen reference. Evidence for shape space covering has been found in the search for aptamers and ribozymes [117, 118].

*Riboswitches.* Stable alternative conformations have been observed experimentally and reported for a variety of natural RNA molecules [126–131]. Alternative conformations of the same RNA molecule may determine completely different functions [132, 133]. Another example is a relatively small molecule, SV11, that is replicated by Q$\beta$ replicase [134, 135]. It exists in two major conformations, a metastable multicomponent structure and a rod-like hairpin conformation, constituting the mfe structure separated from the metastable native state by a huge energy barrier. While the metastable conformation is a template for Q$\beta$ replicase, the ground state is not. By melting and rapid quenching the molecule can be reverted from the inactive stable to the active metastable form [136]. Small switching RNA molecules ($25 \leqslant n \leqslant 100$) were designed, synthesized and investigated [115, 125, 137] (figure 28). NMR spectroscopy turned out to be a very suitable tool (see, for example, [138]. A particularly impressive example is a designed sequence that can satisfy the base-pairing requirements of both the hepatitis delta virus self-cleaving ribozyme and an artificially selected self-ligating ribozyme, which have no base pairs in common. This 'intersection sequence' displays catalytic activity for both cleavage and ligation reactions [115]. A recent publication presents several examples of drastic changes in catalytic functions and structures of ribozymes induced by a few point mutations [118].

The capability of RNA molecules to form multiple (meta)-stable conformations with different function is used in nature to implement so-called *molecular switches* that regulate and control the flow of a number of biological processes. Gene expression, for example, can be regulated when the two mutually exclusive structural alternatives correspond to an active and inactive conformation of the transcript [139]. Mechanistically, one fold of the mRNA, the repressing conformation, contains a terminator hairpin or some other structural element which conceals the translation initiation site, whereas in the alternative conformation the gene can be expressed [140]. The switching between two competing RNA conformations can be triggered by molecular events such as the binding of a target metabolite. The best known example of such a behaviour is provided by riboswitches [141]. These are

[24] The power law applies for the rare structures, whereas the frequencies of the common structures are closer than a power law distribution would predict (an operational definition of 'common' is presented in [40]).

[25] Although this distribution is not (completely) random and has structure specific features [124], it is sufficiently close to uniformity to allow the shape space covering conjecture.

autonomous structural elements primarily found within the 5′-UTRs of bacterial mRNAs, which, upon direct binding of small organic molecules, can trigger conformational changes, leading to an alteration of the expression for the downstream located gene. Their general architecture shows two modular units [142], a 'sensor' for a small metabolite and a unit which 'interprets' the signal from the sensor unit and interfaces to those RNA elements involved in gene expression regulation. The size of the sensor unit ranges typically from 70 to 170 nucleotides, which is unexpectedly large compared to artificial aptamers obtained by *in vitro* directed evolutionary experiments. Riboswitches regulate several key metabolic pathways [143, 144] in bacteria including those leading to coenzyme $B_{12}$, thiamine, pyrophosphate, flavin monophosphate $S$-adenosylmethionine and a couple of important amino acids. The search for additional elements is continuing, e.g. [145, 146]. Riboswitches and engineered allosteric ribozymes [147, 148] demonstrate impressively that RNA is indeed capable of maintaining and regulating a complex metabolic state without the help of proteins.

Algorithms for the design of RNA switches have been developed [82, 149]. As examples for switching RNAs, two small molecules each with two conformations are shown in figure 28. The subtle balance between stacking energies and loop strain allows fine tuning of thermodynamic and kinetic parameters. The example in the figure makes use of extra stabilization by means of an especially stable tetraloop.

*Evolution of non-coding RNA molecules.* In recent years there has been mounting evidence that non-coding RNAs play a dominant role in the regulatory networks of the cell (see, e.g. [150–154] for reviews). Unlike protein coding genes, non-coding RNA (ncRNA) gene sequences do not exhibit a strong *common* statistical signal that separates them from their genomic context. Consequently, a reliable general purpose computational gene-finder for non-coding RNA genes has remained elusive (see, for example [155]). Most classes of the currently known non-coding RNAs, however, are characterized by a common, evolutionarily very well conserved, secondary structure, while at the same time their sequence is rather variable. This observation can be explained as the consequence of stabilizing selection acting (predominantly) on the secondary structure in order to conserve RNA function, whereas the sequences remain almost completely unconstrained and diffuse freely on the neutral network. Diffusion in sequence space in the sense of Motoo Kimura's *neutral theory* [59] forms indeed the conceptual basis of phylogenetic inference. It is important to note, however, that substitution rates differ dramatically between unpaired regions and base-paired regions, since sequence positions that form conserved base pairs are highly correlated. This fact restricts the diffusion process to the neutral network [52]. Corresponding stochastic models of sequence evolution are described, for example [156–159]. The `phase` package [160, 161] implements such a model and is specifically designed to infer phylogenies from RNAs, including ribosomal RNAs, which have a conserved secondary structure.

Structural conservation in the presence of sequence variation is also the basis of recent comparative genomics approaches towards RNA gene finding. The first tool of this type, `qrna` [162], is based upon an approach which assesses the probability that a pair of aligned sequences evolved under the constraint for preserving secondary structure. The program `RNAz` [163] uses two independent criteria for classification: a $z$-score measuring the thermodynamic stability of individual sequences and a *structure conservation index* obtained by comparing the folding energies of the individual sequences with the predicted consensus folding. Both quantities measure different aspects of stabilizing selection for RNA structure.

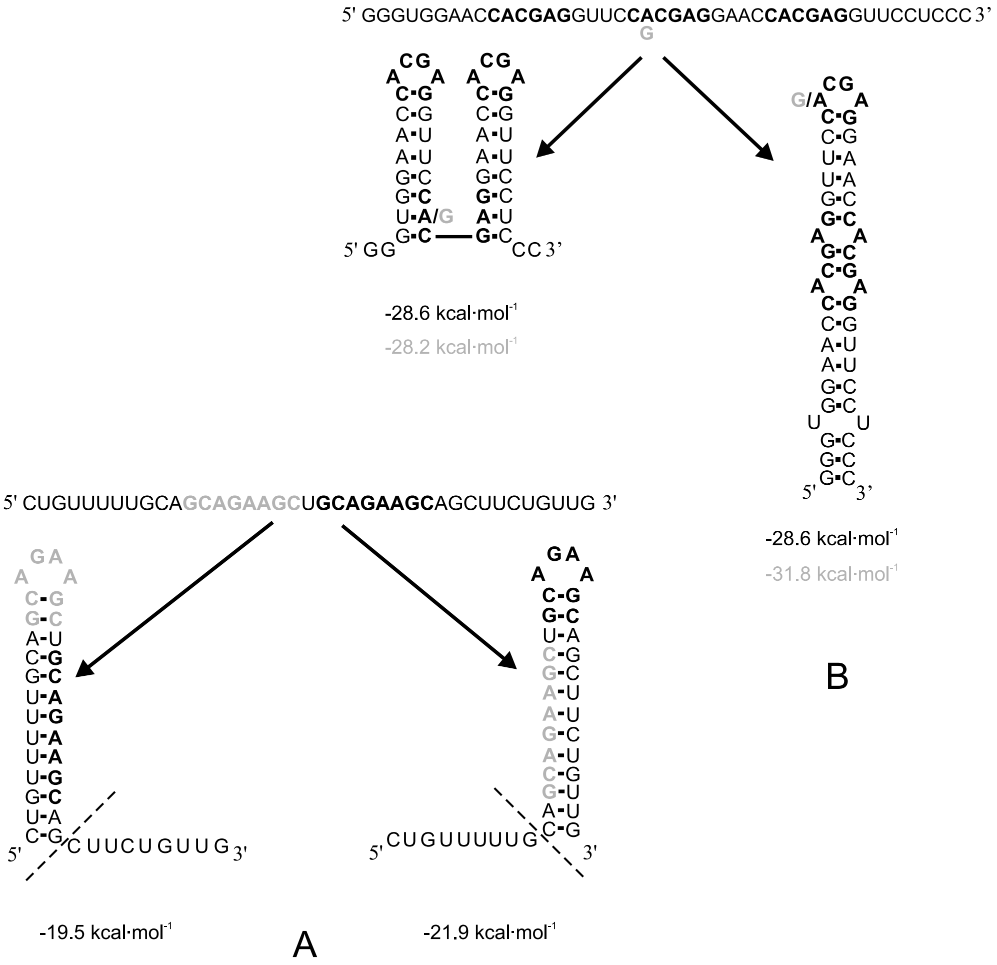


**Figure 28.** Two examples of self-induced small RNA switches [125]. The numbers below the structures represent free energies relative to the unfolded conformation. Both cases demonstrate how the stability of folds can be engineered by a proper choice of parts of sequences. A: the conformers differ in the sequence of base pairs in the middle part of the stacking region, 3 **A–U**+2 **G–U**+1 **G–C** ↔ 3 **A–U**+1 **G–U**+2 **G–C**. The replacement of a **G–U** by a **G–C**, apart from other minor differences in the orientation of base pairs, makes the stack in the structure on the right-hand side more stable. B: Here we show how the influence of the mutation **A**→**G** in position 21 of the sequence destabilizes the double hairpin-structure on replacing an **A–U** pair by a **G–U** pair. At the same time the single hairpin structure becomes more stable because the mutation leads to an especially stable tetraloop of the **GNRA** class.

## 7. Perspectives of the RNA landscape concept

In this review the landscape paradigm has been applied to RNA secondary structures in two distinct ways. Evolution of RNA through optimization in populations of molecules was seen as a process guided by (i) sequence-structure mappings and (ii) kinetic folding of RNA molecules. Table 11 presents an attempt to compare the two processes and to point at common features and differences. Both processes involve ensembles of RNA molecules and an optimization criterion that may be natural, maximal fitness or minimal

**Table 11.** Comparison of RNA evolution and kinetic folding as processes in sequence and conformation space.

| | Evolutionary optimization | Kinetic folding |
|---|---|---|
| Optimization | In sequence space | In conformation space |
| Compatible set | Set of sequences compatible with a given structure | Set of structures compatible with a given sequence |
| Restriction | Compatible set criterion: $\Downarrow$ mfe struct. | Compatible set crit.: $\Downarrow$ $\Delta\varepsilon_{0\to k} \leqslant \Delta\varepsilon_{\max}$ |
| | Neutral network | Relevant conformations |
| Invariance | None (structure for random drift on neutral network) | Sequence |
| Move set | Single point mutation | Base pair closure, opening (and shift) |
| Detrimental noise | Too high mutation rate, $p > p_{\max}$ | Too high thermal energy, $T > T_m$ |
| Initial state | Population at $t = 0$ (arbitrary) | Open chain, start conformation or ensemble |
| Final state | Target structure or property | Predefined conformation or thermodynamic ensemble |
| Trajectory | A genealogy is a time series of sequences | A folding path is a time series of structures |
| Process | RNA evolution is the summation over genealogies at the population level | RNA folding is the summation over trajectories at the ensemble level |
| Optimization criterion | Maximal fitness (or artificially predefined) | Minimal free energy (or artificially predefined) |

free energy, or predefined by the experimenter. Initial states can be chosen at will, provided they can be prepared in an proper experimental setup. Final states can be either pure structures or ensembles. Both processes are driven by noise inducing population changes, mutations or thermal fluctuations, respectively, and they share sensitivity to too much noise destroying inheritance in the form of the error threshold [43, 44, 47] or thermal energy melting the secondary structures. One important difference between the two processes, however, is the existence of an additional invariant property in kinetic folding: the RNA sequence remains unchanged during the entire process, whereas the distribution of structures changes in evolution. The relay series, for example, is a simplified documentation of the migration of the population in shape space (figure 14). Restriction of the evolutionary process to a single neutral network leads to a diffusion-like process of neutral evolution conserving structure [52], as is suggested for non-coding RNAs, and then the mfe structure is invariant.

The next logical step is an extension of the landscape concept combining sequence and conformation space as indicated in figure 29: the set of suboptimal structures is added to the mfe, opening a new dimension with the free energy as the ordering criterion. The object, which is optimized by selection, is the distribution of structures defined by the sequence. Introduction of a folding timescale into the concept in the sense of kinetic folding is straightforward. The distribution of structures can be replaced, for example, by barrier trees, and then the optimization process concerns the folding behaviour together with the structure. The optimization occurs according to two or more criteria and will lead to Pareto optimal sets. The computational tools needed for studying the mappings underlying this extended evolutionary process are still to be developed and comprise, for example, an inverse kinetic folding routine that allows the computation of RNA sequences that give rise to a predefined folding kinetics. At the same time such a software tool would be suitable for designing molecules with given kinetic properties at the secondary structure level. Combining kinetic folding and evolution will eventually provide the answers to one of the open questions concerning biopolymer structures and properties: how does the simultaneous

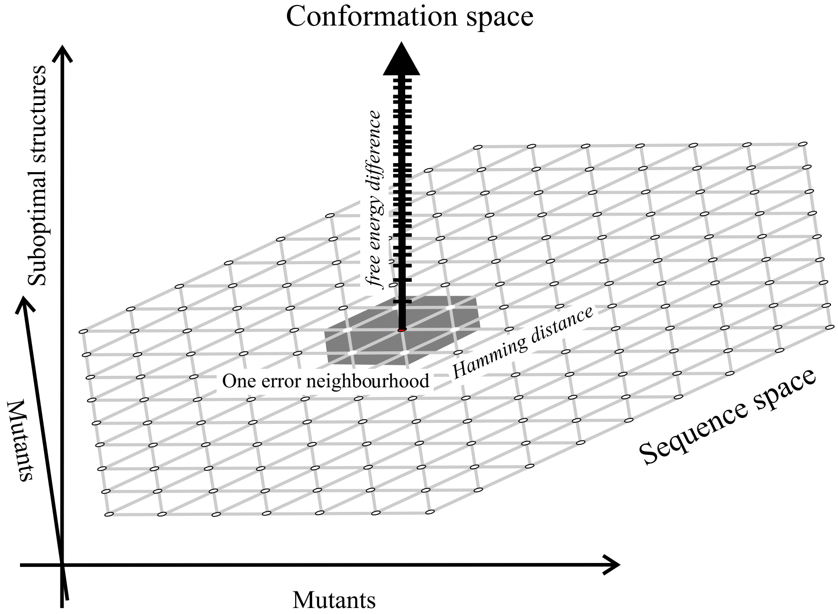


**Figure 29.** Combined sequence and conformation space. For modelling design and evolution of RNA kinetic folding the notions of sequence space and conformation space have to be combined. The object to be optimized is no longer a single RNA structure but the whole set of suboptimal structures and their interconversions or, in a simplified version, the barrier tree.

optimization of thermodynamic stability and efficient folding behaviour operate in nature?

## Acknowledgments

The research of the Vienna group reported here is joint work with Professors Peter Stadler (University of Leipzig), Walter Fontana (Harvard Medical School), Ivo Hofacker and Christoph Flamm (University of Vienna). We are grateful to Ivo Hofacker for intense discussions and help with calculations of base pairing entropies. Dr. Michael Wolfinger kindly helped with the calculations on kinetic folding and Dr. Stephan Bernhart provided unpublished work on RNA cofolding. Financial support by the Austrian 'Fonds zur Förderung der wissenschaftlichen Forschung' (FWF), Projects No. P-13887 and No. P-14898, by the Austrian Gen-AU Bioinformatics Integration Network (BIN) sponsored by the Austrian Bundesministerium für Bildung, Wissenschaft und Kunst (BMBWK) is gratefully acknowledged. Part of the work has been carried out during a visit to the Santa Fe Institute within the External Faculty Program.

## References

[1] Thirumalai D 1998 Native secondary structure formation in RNA may be a slave to tertiary folding *Proc. Natl Acad. Sci. USA* **95** 11506–8

[2] Thirumalai D, Lee N, Woodson S A and Klimov D K 2001 Early events in RNA folding *Annu. Rev. Phys. Chem.* **52** 751–62

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

[4] Waterman M S and Smith T F 1978 RNA secondary structure: a complete mathematical analysis *Math. Biosci.* **42** 257–66

[5] Nussinov R and Jacobson A B 1980 Fast algorithm for predicting the secondary structure of single-stranded RNA *Proc. Natl Acad. Sci. USA* **77** 6309–13

[6] Zuker M and Stiegler P 1981 Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information *Nucleic Acids Res.* **9** 133–48

[7] Walter A E, Turner D H, Kim J, Lyttle M H, Müller P, Mathews D H and Zuker M 1994 Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding *Proc. Natl Acad. Sci. USA* **91** 9218–22

[8] Mathews D H, Sabina J, Zuker M and Turner D H 1999 Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure *J. Mol. Biol.* **288** 911–40

[9] Mathews D H, Disney M D, Childs J L, Schroeder S J, Zuker M and Turner D H 2004 Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure *Proc. Natl Acad. Sci. USA* **101** 7287–92

[10] SantaLucia Jr J, Allawi H L and Seneviratne P 1996 Improved nearest-neighbor parameters for predicting DNA duplex stability *Biochemistry* **35** 3555–62

[11] SantaLucia Jr J 1998 A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics *Proc. Natl Acad. Sci. USA* **95** 1460–5

[12] Hofacker I L, Fontana W, Stadler P F, Bonhoeffer L S, Tacker M and Schuster P 1994 Fast folding and comparison of RNA secondary structures *Mh. Chemie* **125** 167–88

[13] Rivas E and Eddy S R 1999 A dynamic programming algorithm for RNA structure prediction including pseudoknots *J. Mol. Biol.* **285** 2053–68

[14] Waterman M S 2000 *Introduction to Computytional Biology. Maps Seqeunces and Genomes* (Boca Raton, FL: Chapman & Hall/CRC Press)

[15] Hofacker I L, Schuster P and Stadler P F 1998 Combinatorics of RNA secondary structures *Discr. Appl. Math.* **89** 177–207

[16] Tacker M, Stadler P F, Bornberg-Bauer E G, Hofacker I L and Schuster P 1996 Algorithm independent properties of RNA secondary structure predictions *Eur. Biophys. J.* **25** 115–30

[17] McCaskill J S 1990 The equilibrium partition function and base pair binding probabilities for RNA secondary structure *Biopolymers* **29** 1105–19

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

[19] Turner D H and Sugimoto N 1988 RNA structure prediction *Annu. Rev. Biophys. Biophys. Chem.* **17** 167–92

[20] Hofacker I L 2003 Vienna RNA secondary structure server *Nucleic Acids Res.* **31** 3429–31

[21] Dimitrov R A and Zuker M 2004 Prediction of hybridization and melting for double-stranded nucleic acids *Biophys. J.* **87** 215–26

[22] Rehmsmeier M, Steffen P, Höchsmann M and Giegerich R 2004 Fast and efficient prediction of microRNA/target duplexes *RNA* **10** 1507–17

[23] Bernhart S H, Tafer H, Mückstein U, Flamm C, Stadler P F and Hofacker I 2006 Partition function and base pairing probabilities of RNA heterodimers *Algorithms Mol. Biol.* **1** 3

[24] Gold L, Tuerk C, Allen P, Binkley J, Brown D, Green L, MacDougal S, Schneider D, Tasset D and Eddy S R 1993 RNA: The shape of things to come *The RNA World* ed R F Gesteland and J F Atkins (Plainview, NY: Cold Spring Harbor Laboratory Press) pp 497–509

[25] Szostak J W and Ellington A D 1993 *In vitro* selection of functional RNA sequences *The RNA World* ed R F Gesteland and J F Atkins (Plainview, NY: Cold Spring Harbor Laboratory Press) pp 511–33

[26] Marshall K A and Ellington A D 2000 *In vitro* selection of RNA apatmers *Methods Enzymol.* **318** 193–214

[27] Watts A and Schwarz G (ed) Evolutionary biotechnology—from theory to experiment *Biophysical Chemistry* vol 66/2-3 (Amsterdam: Elesvier) pp 67–284

[28] Schuster P, Fontana W, Stadler P F and Hofacker I L 1994 From sequences to shapes and back: a case study in RNA secondary structures *Proc. R. Soc. Lond.* B **255** 279–84

[29] Fontana W and Schuster P 1998 Continuity in evolution. On the nature of transitions *Science* **280** 1451–5

[30] Shapiro B A and Zhang K 1990 Comparing multiple RNA secondary structures using tree comparisons *CABIOS* **6** 309–18

[31] Reidys C and Stadler P F 1996 Bio-molecular shapes and algebraic structures *Computers Chem.* **20** 85–94

[32] Moulton V, Zuker M, Steel M, Pointon R and Penny D 2000 Metrics on RNA secondary structures *J. Comp. Biol.* **7** 277–92

[33] Höchsmann M, Töller T, Giegerich R and Kurtz S 2003 Local similarity in RNA secondary structures *Proc. Computational Systems Bioinformatics Conf. (Stanford, CA, August 2003 (CSB 2003))* pp 159–68

[34] Andronescu M, Fejes A P, Hutter F, Hoos H H and Condon A 2004 A new algorithm for RNA secondary structure design *J. Mol. Biol.* **336** 607–24

[35] Bollobás B 1985 *Random Graphs* (London: Academic)

[36] Reidys C, Stadler P F and Schuster P 1997 Generic properties of combinatory maps. Neutral networks of RNA secondary structure *Bull. Math. Biol.* **59** 339–97

[37] Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker I L and Schuster P 1996 Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks *Mh. Chemie* **127** 355–74

[38] Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker I L and Schuster P 1996 Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering *Mh. Chemie* **127** 375–89

[39] Stephan-Otto Attolini C and Stadler P F 2005 Neutral networks of interacting RNA secondary structures *Adv. Complex Syst.* **8** 275–84

[40] Schuster P 1995 How to search for RNA structures. Theoretical concepts in evolutionary biotechnology *J. Biotechnol.* **41** 239–57

[41] Schuster P 1997 Landscapes and molecular evolution *Physica* D **107** 351–65

[42] Moran P A P 1962 *The Statistical Processes of Evolutionary Theory* (Oxford: Clarendon)

[43] Eigen M 1971 Selforganization of matter and the evolution of biological macromolecules *Naturwissenschaften* **58** 465–523

[44] Eigen M and Schuster P 1977 The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle *Naturwissenschaften* **64** 541–65

[45] Eigen M and Schuster P 1978 The hypercycle. A principle of natural self-organization. Part B: The abstract hypercycle *Naturwissenschaften* **65** 7–41

[46] Swetina J and Schuster P 1982 Self-replication with errors—a model for polynucleotide replication *Biophys. Chem.* **16** 329–45

[47] Eigen M, McCaskill J and Schuster P 1989 The molecular quasispecies *Adv. Chem. Phys.* **75** 149–263

[48] Schuster P 2003 Molecular insight into the evolution of phenotypes *Evolutionary Dynamics—Exploring the Interplay of Accident, Selection, Neutrality, and Function* ed J P Crutchfield and P Schuster (New York: Oxford University Press) pp 163–215

[49] Tarazona P 1992 Error threshold for molecular quasispecies as phase transitions: from simple landscapes to spin-glass models *Phys. Rev.* A **45** 6038–50

[50] Jagers P 1975 *Branching Processes with Biological Applications* (London: Wiley)

[51] Demetrius L, Schuster P and Sigmund K 1985 Polynucleotide evolution and branching processes *Bull. Math. Biol.* **47** 239–62

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

[53] Fontana W and Schuster P 1987 A computer model of evolutionary optimization *Biophys. Chem.* **26** 123–47

[54] Fontana W, Schnabl W and Schuster P 1989 Physical aspects of evolutionary optimization and adaptation *Phys. Rev.* A **40** 3301–21

[55] Gillespie D T 1976 A general method for numerically simulating the stochastic time evolution of coupled chemical reactions *J. Comp. Phys.* **22** 403–34

[56] Gillespie D T 1977 Exact stochastic simulation of coupled chemical reactions *J. Phys. Chem.* **81** 2340–61

[57] Gillespie D T 1977 Concerning the validity of the stochastic approach to chemical kinetics *J. Stat. Phys.* **16** 311–18

[58] Stadler B R M, Stadler P F, Wagner G P and Fontana W 2001 The topology of the possible: formal spaces underlying petterns of evolutionary change *J. Theor. Biol.* **213** 241–74

[59] Kimura M 1983 *The Neutral Theory of Molecular Evolution* (Cambridge, UK: Cambridge University Press)

[60] Grünberger K, Langhammer U, Wernitznig A and Schuster P 2005 RNA evolution *in silico Technical Report* Institut für Theoretische Chemie, Universität Wien

[61] Kupczok A and Dittrich P 2006 Determinants of simulated RNA evolution *J. Theor. Biol.* **238** 726–35

[62] Morgan S R and Higgs P G 1998 Barrier heights between ground states in a model of RNA secondary structure *J. Phys. A: Math. Gen.* **31** 3153–70

[63] Flamm C, Fontana W, Hofacker I L and Schuster P 1999 Elementary step dynamics of RNA folding *RNA* **6** 325–38

[64] Wolfinger M T, Svrcek-Seiler W A, Flamm C, Hofacker I L and Stadler P F 2004 Efficient computation of RNA folding dynamics *J. Phys. A: Math. Gen.* **37** 4731–41

[65] Martinez H M 1984 An RNA folding rule *Nucleic Acids Res.* **12** 323–34

[66] Abrahams J P, van den Berg M, van Batenburg E and Pleij C 1990 Prediction of RNA secondary structure, including pseudoknotting, by computer simulation *Nucleic Acids Res.* **18** 3035–44

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

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

[69] Gultyaev A P, van Batenburg F H D and Pleij C W A 1995 The influence of a metastable structure in plasmid primer RNA on antisense RNA-binding kinetics *Nucleic Acids Res.* **23** 3718–25

[70] Higgs P G and Morgan S R 1995 Thermodynamics of RNA folding. When is an RNA molecule in equilibrium? *Lecture Notes in Artificial Intelligence* **929** 852–61

[71] Morgan S R and Higgs P G 1996 Evidence for kinetic effects in the folding of large RNA molecules *J. Chem. Phys.* **105** 7152–7

[72] Gultyaev A P, van Batenburg F H D and Pleij C W A 1998 RNA folding dynamics: computer simulations by a genetic algorithm *ACS Symp. Ser.* **682** 229–45

[73] Isambert H and Siggia E D 2000 Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme *Proc. Natl Acad. Sci. USA* **97** 6151–520

[74] Zhang W and Chen S-J 2003 Master equation approach to finding the rate-limiting steps in biopolymer folding *J. Chem. Phys.* **118** 3413–20

[75] Zhang W and Chen S-J 2003 Analyzing the biolpolymer folding rates and pathways using kinetic cluster method *J. Chem. Phys.* **119** 8716–29

[76] Pörschke D 1977 Elementary steps of base recognition and helix-coli transitions in nucleic acids *Chemical Relexation in Molecular Biology* ed I Pecht and R Rigler (Berlin: Springer) pp 191–218

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

[78] Wuchty S, Fontana W, Hofacker I L and Schuster P 1999 Complete suboptimal folding of RNA and the stability of secondary structures *Biopolymers* **49** 145–65

[79] Waterman M S and Byers T H 1985 A dynamic programming algorithm to find all solutions in a neighborhood of the optimum *Math. Biosci.* **77** 179–88

[80] Pörschke D and Eigen M 1971 Co-operative non-enzymic base recognition. III. Kinetics of the helix–coli transition of the oligoribouridylic–oligoriboadenylic acid systems and of oligoriboadenylic acid alone at acidic pH *J. Mol. Biol.* **62** 361–81

[81] Pörschke D 1974 Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix *Biophys. Chem.* **1** 381–6

[82] Flamm C, Hofacker I L, Maurer-Stroh S, Stadler P F and Zehl M 2001 Design of multi-stable RNA molecules *RNA* **7** 254–65

[83] Abfalter I, Flamm C and Stadler P F 2003 Design of multi-stable nucleic acid sequences *Proc. German Conf. on Bioinformatics. GCB 2003* vol 1, ed H-W Mewes *et al* (Munich: Belleville Verlag Michael Farin) pp 1–7

[84] Clote P, Gąsieniec L, Kolpakov R, Kranakis E and Krizanc D 2005 On realizing shapes in the theory of RNA neutral networks *J. Theor. Biol.* **236** 216–27

[85] Rich A and RajBhandary U L 1976 Transfer RNA: molecular structure, sequence, and properties *Annu. Rev. Biochem.* **45** 805–60

[86] Sprinzl M, Grueter F, Spelzhaus A and Gauss D H 1980 Compilation of tRNA sequences *Nuleic Acids Res.* **8** R1–131

[87] Sprinzl M, Horn C, Brown M, Ioudovitch A and Steinberg S 1998 Compilation of tRNA sequences and sequences of tRNA genes *Nuleic Acids Res.* **26** 148–53

[88] Specht T, Wolters J and Erdmann V A 1990 Compilation of 5S RNA and 5S RNA gene sequences *Nucleic Acids Res.* **18** (suppl.) 2215–30

[89] Leontis N B and Westhof E 2001 Geometric nomenclature and classification of RNA base pairs *RNA* pp 499–12

[90] Gultyaev A P 1991 The computer simulation of RNA folding involving pseudoknot formation *Nucleic Acids Res.* **19** 2489–94

[91] Dirks R M and Pierce N A 2003 A partition function algorithm for nucleic acid secondary structure including pseudoknots *J. Comput. Chem.* **24** 1664–77

[92] van Batenburg F H D, Gultyaev A P, Pleij C W A, Ng J and Oliehoek J 2000 PseudoBase: a database with of RNA pseudoknots *Nucleic Acids Res.* **28** 201–4

[93] Xayaphoummine A, Buchner T, Thalmann F and Isambert H 2003 Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations *Proc. Natl Acad. Sci. USA* **100** 15310–15

[94] Xayaphoummine A, Buchner T and Isambert H 2005 Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots *Nucleic Acids Res.* **33** W605–10

[95] Pillsbury M, Orland H and Zee A 2005 Steepest descent calculations of RNA pseudoknots *Phys. Rev.* E **72** 011911

[96] Pasquali S, Gan H H and Schlick T 2005 Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs *Nucleic Acids Res.* **33** 1384–98

[97] Ren J, Rastegari B and Hoos A C H H 2005 HotKnots: heuristic prediction of RNA secondary structures including pseudoknots *RNA* **11** 1494–504

[98] Gardner P P and Giegerich R 2004 A comprehensive comparison of comparative RNA structure prediction approaches *BMC Bioinf.* **5** 140

[99] Witwer C, Hofacker I L and Stadler P F 2004 Prediction of consensus RNA secondary structures including pseudoknots *IEEE/ACM Trans. Comp. Biol. Bioinf.* **1** 66–77

[100] Tahi F, Stefan E and Régnier M 2005 P-DCFOLD or how to predict all kinds of pseudoknots in RNA secondary structure *Int. J. Artficial Intell. Tools* **14** 703–16

[101] Hofacker I L, Fekete M and Stadler P F 2002 Secondary structure prediction for aligned RNA sequences *J. Mol. Biol.* **319** 1059–66

[102] Tahi F, Gouy M and Régnier M 2002 Automatic RNA secondary structure prediction with a comparative approach *Computers Chem.* **26** 521–30

[103] Ellington A D and Szostak J W 1990 *In vitro* selection of RNA molecules that bind specific ligands *Nature* **346** 818–22

[104] Tuerk C and Gold L 1990 Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase *Science* **249** 505–10

[105] Guerrier-Takada C, Gardiner K, Marsh T, Pace N and Altman S 1983 The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme *Cell* **35** 849–57

[106] Kruger K, Grabowski P J, Zaug A J, Sands J, Gottschling D E and Cech T R 1982 Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena *Cell* **31** 147–57

[107] Zaug A J and Cech T R 1986 The intervening sequence RNA of *tetrahymena* is an enzyme *Science* **231** 470–5

[108] Fontana W, Konings D A M, Stadler P F and Schuster P 1993 Statistics of RNA secondary structures *Biopolymers* **33** 1389–404

[109] Higgs P G 1993 RNA secondary structure: a comparsion of real and random sequences *J. Phys. I France* **3** 43–59

[110] Gevertz J, Gan H H and Schlick T 2005 *In vitro* RNA random pools are not structurally diverse: a computational analysis *RNA* **11** 853–63

[111] Clote P, Ferré F, Kranakis E and Krizanc D 2005 Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency *RNA* **11** 578–91

[112] Stadler P F and Flamm C 2003 Barrier trees on poset-velued landscapes *Genet. Prog. Evol. Mach.* **4** 7–20

[113] Rogers J and Joyce G 1999 A ribozyme that lacks cytidine *Nature* **402** 323–5

[114] Reader J S and Joyce G F 2002 A ribozyme composed of only two different nucleotides *Nature* **420** 841–4

[115] Schultes E and Bartel D 2000 One sequence, two ribozymes: implications for the emergence of new ribozyme folds *Science* **289** 448–52

[116] Held D M, Greathouse S T, Agrawal A and Burke D H 2003 Evolutionary landscapes for the acquisition of new ligand recognition by RNA aptamers *J. Mol. Evol.* **57** 299–308

[117] Huang Z and Szostak J W 2003 Evolution of aptamers with a new specificity and new scondary structures from an ATP aptamer *RNA* **9** 1456–63

[118] Curtis E A and Bartel D P 2005 New catalytic structures from an existing ribozyme *Nat. Struct. Mol. Biol.* **12** 994–1000

[119] Li H, Helling R, Tang C and Wingreen N 1996 Emergence of preferred structures in a simple model of protein folding *Science* **273** 666–9

[120] Govindarajan S and Goldstein R A 1996 Why are some protein structures so common? *Proc. Natl Acad. Sci. USA* **93** 3341–5

[121] Govindarajan S, Recabarren R and Goldstein R A 1999 Estimating the total number of protein folds *Proteins* **35** 408–14

[122] Aita T, Ota M and Husimi Y 2003 An *in silico* exploration of the neutral network in protein sequence space *J. Theor. Biol.* **221** 599–613

[123] Bastolla U, Porto M, Roman H E and Vendruscolo M 2003 Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution *J. Mol. Evol.* **56** 243–54

[124] Schuster P and Stadler P F 2004 Discrete models of biopolymers *Handbook of Computational Chemistry* ed M J C Crabbe *et al* (New York: Marcel Dekker) chapter 5, pp 187–222

[125] Nagel J H A, Flamm C, Hofacker I L, Franke K, de Smit M H, Schuster P and Pleij C W A 2006 Structural parameters affecting the kinetic competition of RNA hairpin formation *Nucleic Acids Res.* **34** at press

[126] Fresco J R, Adains A, Ascione R, Henley D and Lindahl T 1966 Tertiary structure in transfer ribonucleic acids *Cold Spring Habor Symp. Quant. Biol.* **31** 527–39

[127] Hawkins E R, Chang S H and Mattice W L 1977 Kinetics of the renaturation of yeast tRNA$_3^{Leu}$ *Biopolymers* **16** 1557–66

[128] LeCuyer K A and Crothers D M 1993 The *Leptomonas collosoma* spliced leader RNA can switch between two alternate structural forms *Biochemistry* **32** 5301–11

[129] Emerick V L and Woodson S A 1993 Selfsplicing of the Tetrahymena pre-rRNA is decreased by misfolding during transcription *Biochemistry* **32** 14062–7

[130] Nagel J H A, Gultyaev A P, Gerdes K and Pleij C 1999 Metastable structures and refolding kinetics in hok mRNA of plasmid R1 *RNA* **5** 1408–18

[131] Nagel J H A and Pleij C 2002 Self-induced structural switches in RNA *Biochimie* **84** 913–23

[132] Baumstark T, Schroder A R and Riesner D 1997 Viroid processing: switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation *EMBO J.* **16** 599–610

[133] Perrotta A T and Been M D 1998 A toggle duplex in hepatitis delta virus self-cleaving RNA that stabilizes an inactive and a salt-dependent pro-active ribozyme conformation *J. Mol. Biol.* **279** 361–73

[134] Biebricher C K, Diekmann S and Luce R 1982 Structural analysis of self-replicating RNA synthesized by Q$\beta$ replicase *J. Mol. Biol.* **154** 629–48

[135] Biebricher C K and Luce R 1992 *In vitro* recombination and terminal elongation of RNA by Q$\beta$ replicase *EMBO J.* **11** 5129–35

[136] Zamora H, Luce R and Biebricher C K 1995 Design of artificial short-chained RNA species that are replicated by Q$\beta$ replicase *Biochemistry* **34** 1261–6

[137] Micura R and Höbartner C 2003 On secondary structue rearrangements and equilibria of small RNAs *ChemBioChem* **4** 984–90

[138] Höbartner C and Micura R 2003 Bistable secondary structures of small RNAs and their structural probing by comparative imino proton NMR spectroscopy *J. Mol. Biol.* **325** 421–31

[139] Merino E and Yanofsky C 2002 Regulation by termination-antitermination: a genomic approach Bacillus Subtilis *and Its Closest Relatives: From Genes to Cells* ed A L Sonenshein *et al* (Washington, DC: ASM Press) pp 323–36

[140] Henkin T M and Yanofsky C 2002 Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcribtion termination/antitermination decision *BioEssays* **24** 700–7

[141] Vitreschak A G, Rodionov D A, Mironov A A and Gelfand M S 2004 Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Gen.* **20** 44–50

[142] Winkler W C and Breaker R R 2003 Genetic control by metabolite-binding riboswitches *Chembiochem* **4** 1024–32

[143] Brantl S 2004 Bacterial gene regulation: from transcription attenuation to riboswitches and ribozymes *Trends Microbiol.* **12** 473–5

[144] Nudler E and Mironov A S 2004 The riboswitch control of bacterial metabolism *Trends Biochem. Sci.* **29** 11–17

[145] Barrick J E, Corbino K A, Winkler W C, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser J K and Breaker R R 2004 New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control *Proc. Natl Acad. Sci. USA* **101** 6421–6

[146] Lesnik E A, Fogel G B, Weekes D, Henderson T J, Levene H B, Sampath R and Ecker D J 2005 Identification of conserved regulatory RNA structures in prokaryotic metabolic pathway genes *Biosystems* **80** 145–54

[147] Breaker R R 2002 Engineered allosteric ribozymes as biosensores components *Curr. Opin. Biotechnol.* **13** 31–39

[148] Silverman S K 2003 Rube goldberg goes (ribo)nuclear? molecular switches and sensors made from RNA *RNA* **9** 377–83

[149] Voss B, Meyer C and Giegerich R 2004 Evaluating the predictability of conformational switching in RNA *Bioinformatics* **20** 1573–82

[150] Bartel D P and Chen C-Z 2004 Micromanagers of gene expression: the potentially wide-spread influence of metazoan microRNAs *Nat. Genet.* **5** 396–400

[151] Hobert O 2004 Common logic of transcription factor and microRNA action *Trends Biochem. Sci.* **29** 462–8

[152] Mattick J S 2003 Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms *Bioessays* **25** 930–9

[153] Mattick J S 2004 RNA regulation: a new genetics? *Nat. Genet.* **5** 316–23

[154] Szymański M, Barciszewska M Z, Zywicki M and Barciszewski J 2003 Noncoding RNA transcripts *J. Appl. Genet.* **44** 1–19

[155] Eddy S R 2001 Non-coding RNA genes and the modern RNA world *Nat. Genet.* **2** 919–29

[156] Schöninger M and von Haeseler A 1999 Towards assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models *J. Mol. Evol.* **49** 691–8

[157] Knudsen B and Hein J J 1999 Using stochastic context free grammars and molecular evolution to predict RNA secondary structure *Bioinformatics* **15** 446–54

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

[159] Otsuka J and Sugaya N 2003 Advanced formulation of base pair changes in the stem regions of ribosomal RNAs; its application to mitochondrial rRNAs for resolving the phylogeny of animals *J. Theor. Biol.* **222** 447–60

[160] Jow H, Hudelot C, Rattray M and Higgs P G 2002 Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution *Mol. Biol. Evol.* **19** 1591–601

[161] Hudelot C, Gowri-Shankar V, Jow H, Rattray M and Higgs P G 2003 RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences *Mol. Phylogenet. Evol.* **28** 241–52

[162] Rivas E, Klein R J, Jones T A and Eddy S R 2001 Computational identification of noncoding RNAs in *E. coli* by comparative genomics *Curr. Biol.* **11** 1369–73

[163] Washietl S, Hofacker I L and Stadler P F 2005 Fast and reliable prediction of noncoding rnas *Proc. Natl Acad. Sci. USA* **102** 2454–9