1

Modeling Conformational Flexibility and Evolution of Structure: RNA as an Example

P. Schuster and P.F. Stadler

In this chapter, RNA secondary structures are used as an appropriate toy model to illustrate an application of the landscape concept to understand the molecular basis of structure formation, optimization, adaptation, and evolution in simple systems. Two classes of landscapes are considered (1) conformational landscapes mapping RNA conformations into free energies of formation and (2) sequence–structure mappings assigning minimum free energy structures to sequences. Even without referring to suboptimal conformations, optimization of RNA structures by mutation and selection reveals interesting features on the population level that can be interpreted by means of sequence–structure maps. The full power of the RNA model unfolds when sequence–structure maps and conformational landscapes are merged into a more advanced mapping that assigns a whole spectrum of conformations to the individual sequence. The scenario is complicated further – but at the same time made more realistic – by considering kinetic effects that allow for the assignment of two or more long-lived conformations, together with their suboptimal folds, to a single sequence. In this case, molecules can be designed, which fulfil multiple functions by switching back and forth from one stable conformation to the other or by changing conformation through allosteric binding of effectors. The evolution of noncoding RNAs is presented as an example for the application of landscape-based concepts.

1.1 Definition and Computation of RNA Structures

RNA sequences form structures under appropriate conditions consisting of aqueous solution at sufficiently low temperatures, approximately neutral pH, and ionic strength. In most of the sufficiently well studied examples RNA folding occurs in two steps [1, 2] (1) the formation of a flexible so-called secondary structure requiring monovalent counterions and (2) the folding of the secondary structure into a rigid 3D-structure in the presence on divalent ions, especially $Mg^{2\oplus}$ [3] (for an exception see [4]). Experimental determination

of full spatial RNA structures is a hard task for crystallographers and NMR spectroscopists [5, 6]. Prediction of 3D-structures is also an enormously complex problem and at least as demanding as in the case of proteins [7]. RNA secondary structures, however, in contrast to protein secondary structures, have a physical meaning as folding intermediates and are useful tools in the interpretation and prediction of RNA function. In addition, conventional RNA secondary structures (Sect. 1.1.1) can be represented as (restricted) strings over a three-letter alphabet and they are accessible, therefore, to combinatorial analysis and other techniques of discrete mathematics [8–10]. The discreteness of secondary structures allows for straightforward comparisons of the spaces of sequences, structures, and conformations and provides the insights into flexibility and robustness of RNA molecules. Moreover, RNA secondary structures and lattice protein models are at present the only biological objects for which conformational landscapes and sequence–structure maps can be computed and analyzed in complete detail. Therefore, this contribution will be exclusively dealing with them.

1.1.1 RNA Secondary Structures

A conventional RNA secondary structure[1] 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 (Fig. 1.1). The edges of the graph represent bonds between, nodes which fall into two classes: (1) the backbone, $\{i — (i+1)\ \forall i = 1, \ldots, n-1\}$, and (2) the base pairs. The two ends of the sequence (5′- and 3′-end) 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 $i \mathbf{-} j$. For a conventional secondary structure, the base pairs fulfil three conditions:

1. *Binary interaction restriction.* An individual nucleotide is either involved in one base pair or it is a single nucleotide forming no base pair.
2. *No nearest neighbor pair restriction.* Base pairs to nearest neighbors, $i \mathbf{-} j$ with $j = i - 1$ or $j = i + 1$ are excluded.
3. *No pseudoknot restriction.* Two base pairs $i \mathbf{-} j$ and $k \mathbf{-} l$ with $i < j$, $i < k$ and $k < l$ are only accepted if either $i < k < l < j$ or $i < j < k < l$ are fulfilled – the second base pair is either enclosed by the first base pair or lies completely outside (Fig. 1.1).

Condition 1 forbids the formation of base triplets or higher interactions between nucleotides. Condition 2 is required for steric reasons because stereochemistry does not allow for pairing geometries between neighboring nucleotides. As we shall mention later, this condition is even more stringent in the

[1] "Conventional" means here that the structure is free of pseudoknots (Condition 3). Some other definitions include certain or all classes of pseudoknots.

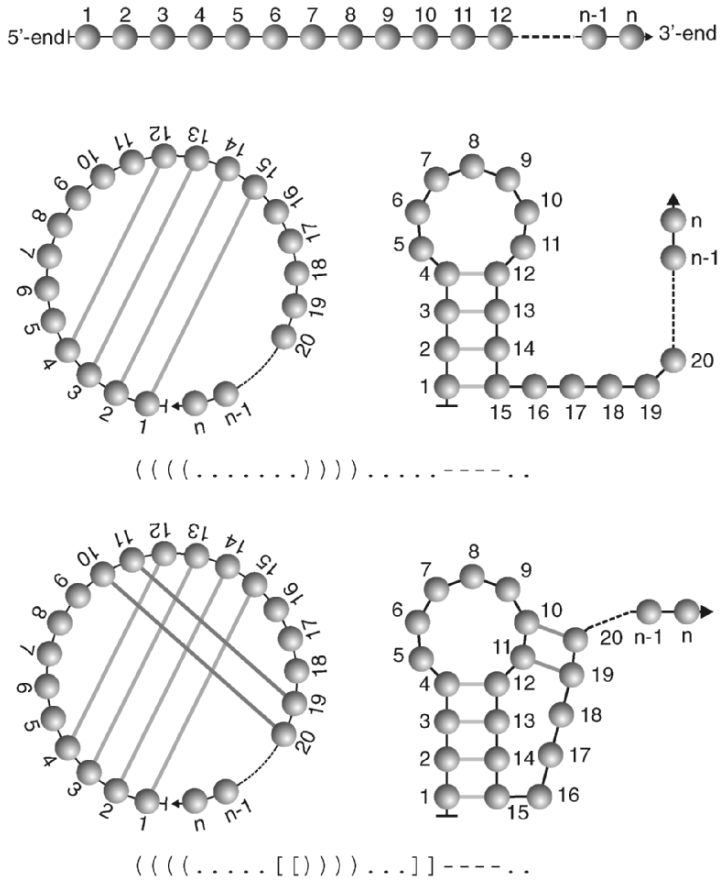

Fig. 1.1. Definition of RNA secondary structures. Each nucleotide inside the sequence forms two backbone bonds to its neighbors, the two nucleotides at the ends, 1 and n, are connected to one neighbor (topmost drawing: nucleotides are shown as *spheres*, the 3′-end is represented by an *arrow*). Each nucleotide can stay unpaired or form one (and only one) base pair to another nucleotide. In the circular representation of structures (left-hand side of the drawings in the *middle* and at the *bottom*), base pairs appear as *lines* crossing the circle. The upper secondary structures has no pseudoknot. The structure at the bottom contains a pseudoknot, which is easily recognized by crossings of *lines* in the circular representation. On the right-hand side of the two structures, we show the conventional drawings of secondary structures as they are used by biochemists and molecular biologists. Parentheses representations (see text) are shown below the two structures

sense that hairpin loops with less than three single nucleotides do not occur in real structures. Condition 3 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 [11].

Throughout this chapter, 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 no-pseudoknot Condition 3 and insist only that each nucleotide takes part in at most one base pairs (Condition 1).[2] Furthermore, let Υ be the set of unpaired bases, which is the subset of $\{1,\ldots,n\}$ that is not met by the matching Ω.

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

$$A = \left\{ a_{ij} = a_{ji} = \begin{cases} 1 & \text{if } i,j \in \Omega\,, \\ 0 & \text{otherwise}\,, \end{cases} \quad i,j = \{1,\ldots,n\} \right\}, \tag{1.1}$$

and the parentheses notation, which will be used later on to calculate base pair probabilities and compute distances between structures, respectively. In this notation, single nucleotides, $i \in \Upsilon$, are represented by dots and base pairs by parentheses (Fig. 1.1). Structures are strings of length n over the three-letter alphabet, $\{.,(,)\}$ with the restrictions that the number of left parentheses, "(," has to match exactly the number of right parentheses, ")," and no parenthesis must be closed before it had been opened. The no-pseudoknot restriction guarantees that left and right parentheses are assigned according to the rules of mathematics. Colored parentheses are required for the correct assignment in the presence of pseudoknots (bottom plot in Fig. 1.1).

Three classes of elements occur in structures (1) stacks, (2) various kinds of loops, and (3) external elements (Fig. 1.2). Stacks are arrays of consecutive base pairs in which the two strands run in opposite direction:

$$\begin{array}{lccccccccc} 5'\text{-end} \cdots & - & i & - & i+1 & - & i+2 & - & \cdots & 3'\text{-end} \\ & & \mid & & \mid & & \mid & & & \\ 3'\text{-end} \cdots & - & j & - & j-1 & - & j-2 & - & \cdots & 5'\text{-end}\,. \end{array}$$

Loops are commonly classified by the number of closing base pairs:[4]

(1) A loop of degree one has one closing base pair and is commonly called a hairpin loop.
(2) Loops of degree two are bulges or internal loops depending on the positioning of the two closing pairs. In bulges, the closing pairs are neighbors

[2] Wherever confusion is possible we shall be precise and use S for conventional secondary structures and Ω for the generalization.

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

[4] Each stack neighboring the loop ends in a pair is called a *closing pair* of the loop. The number of closing base pairs is easily determined: Imagine the loop as a circle and count all base pairs whose nucleotides are members of this circle.

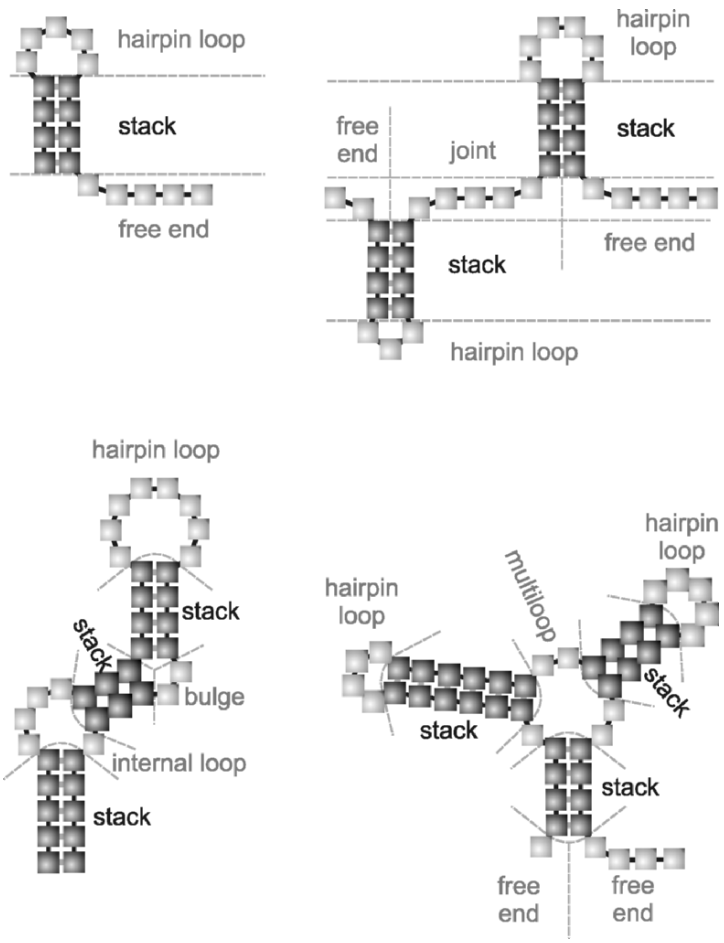

Fig. 1.2. Elements of RNA secondary structures. Three classes of structural elements are distinguished: (1) stacks (indicated by nucleotides in dark color), (2) loops, and (3) external elements being joints and free ends. Loops fall into several subclasses: Hairpin loops have one base pair, called the closing pair, in the loop. Bulges and internal loops have two closing pairs, and loops with three or more closing pairs are called multiloops

without a single nucleotide in between while they are separated by single bases on both sides in internal loops. Algorithmically, two stacked adjacent base pairs are treated as an interior loop without unpaired bases. Higher degree loops have three or more closing pairs and are called multiloops.

(3) Flexible substructures are free ends and parts of the nucleotide chain that join two modules of structure.

As indicated in Fig. 1.2 it is important for calculations of free energies that the individual substructures are independent in the sense that the free energy of a substructure is not changed by changes in the pairing pattern of another substructure.

It will turn out useful to introduce the notion of acceptable structures, which are a subset of the conventional structures [12]. Two restrictions are introduced that eliminate structures of high free energies, which are commonly well above the energy of the open chain (a) Condition 2 in the definition of secondary structures is made more stringent in the sense that base pairs to next nearest neighbors are also excluded, and hence the base pairs with the shortest distance along the sequence are $i - i + 3$, and (b) isolated base pairs are excluded implying that the shortest stacking regions consists of at least two base pairs formed by neighboring bases.

1.1.2 Compatibility of Sequences and Structures

A sequence $X = (x_1 x_2 \cdots x_n)$ over an alphabet $\mathcal{A}$ with κ letters is *compatible* with the matching Ω if $\{i - j\} \in \Omega$ implies that $x_i x_j$ is an allowed base pair. This situation is expressed by $x_i x_j \in \mathcal{B}$. For natural RNAs, we have $\mathcal{A} = \{\alpha_i\} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ (or $\{\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}\}$ for DNA) and $\mathcal{B} = \{\beta_{ij} = \alpha_i - \alpha_j\} = \{\mathrm{AU}, \mathrm{UA}, \mathrm{GC}, \mathrm{CG}, \mathrm{GU}, \mathrm{UG}\}$. We denote the set of all sequences that are compatible with a structure Ω by

$$\mathbf{C}[\Omega] = \left\{ X \middle| \{i - j\} \in \Omega \implies x_i x_j \in \mathcal{B} \right\} . \tag{1.2}$$

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 ϱ base pairs contained in $\mathcal{B}$. For a given structure we have, therefore,

$$|\mathbf{C}[\Omega]| = \kappa^{|\Upsilon|} \varrho^{|\Omega|} , \tag{1.3}$$

compatible sequences.

The problem has a relevant inverse too: How many structures are compatible with a given sequence X? The set of these structures comprises all possible conformations, i.e., the minimum free energy structure together with the suboptimal structures. The computation of this number is rather involved and has to use a recursion that has some similarity to the computation of the minimum free energy structure (Sect. 1.1.4). It can be also obtained as the partition function [13] in the limit of infinite temperature, $T \to \infty$ (Sect. 1.1.6). A simpler estimate is possible in terms of the stickiness of the sequence,

$$p(X) \;=\; 2 \sum_{\beta_{ij} \in \mathcal{B}} p_i(X) p_j(X) \quad \text{with} \quad p_i(X) = \frac{n_i(X)}{n} \text{ and } p_j(X) = \frac{n_j(X)}{n} , \tag{1.4}$$

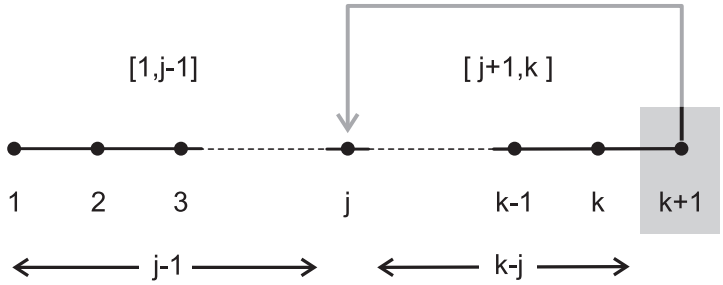

Fig. 1.3. Basic principle of recursions for secondary structures. The property of a sequence with chain length n is built up recursively from the properties of smaller segments under the assumption that the contributions are additive: The property for the segment $[1, k+1]$ is identical with that of the segment $[1, k]$ if the nucleotide x_{k+1} forms no base pair. If it forms a base pair with the nucleotide x_j the segment $[1, k+1]$ is bisected into two smaller fragments $[1, j-1]$ and $[j+1, k]$. The solution of a problem can be found by starting from the smallest segments and progressing successively to larger segments. This procedure leads either to a recursion formula (1.6, 1.7) or it can be converted into a dynamic programming algorithm as in the case of minimum free energy structure determination

where $n_i(X)$ and $n_j(X)$ are the numbers of nucleotides α_i and α_j in the sequence X, respectively, and $n = \sum_{\alpha_i \in \mathcal{A}} n_i(X)$, the chain length of the molecule.

On the basis of the assumption of additive contributions from structure elements, the properties associated with secondary structures can be computed in recursive manner from smaller to larger segments (Fig. 1.3). It is straightforward to enumerate, for example, all possible secondary structures for a given chain length n, s_n, by means of a recursion [14,15]. For a minimal length for hairpin loops, $n_{\mathrm{lp}} \geq \lambda$, one finds [12,16]:

$$s_{m+1} = s_m + \sum_{j=1}^{m-\lambda} s_{j-1} \cdot s_{m-j} = s_m + \sum_{j=\lambda}^{m-1} s_j s_{m-j-1}$$
$$\text{with} \quad s_0 = s_1 = \cdots = s_\lambda = 1\,. \tag{1.5}$$

For a (random) sequence X with nucleotide composition $(p_1, \ldots, p_\kappa)$, the probability that two nucleotides form a base pair is given by the stickiness $p(X)$. Insertion into the recursion leads to [17]:

$$s_{m+1}(p) = s_m(p) + p \sum_{j=1}^{m-\lambda} s_{j-1}(p) \cdot s_{m-j}(p)$$
$$\text{with} \quad s_0(p) = s_1(p) = \cdots = s_\lambda(p) = 1\,, \tag{1.6}$$

and $s_n(p)$ yields a rough 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, $n_{\mathrm{st}} \geq \sigma$ [12]:

$$\begin{aligned} s_{m+1}(p) &= \Xi_{m+1}(p) + \phi_{m-1}(p)\,, \\ \Xi_{m+1}(p) &= s_m(p) + \sum_{k=\lambda+2\sigma-2}^{m-2} \phi_k(p) \cdot s_{m-k-1}(p) \\ \phi_{m+1}(p) &= p \sum_{k=\sigma-1}^{\lfloor (m-\lambda+1)/2 \rfloor} \Xi_{m-2k+1}(p) \cdot p^k \end{aligned} \tag{1.7}$$

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 with a rough 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 1.1. The numbers of suboptimal structures become very large at moderate chain length n already. 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 the recursion (1.7) 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.

Table 1.1. Estimates on the numbers of suboptimal structures, $s_n(p)$ with $\lambda = 3$ and $\sigma = 1$ and $p(X)$ being the stickiness of sequence X

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

1.1.3 Sequence Space, Shape Space, and Conformation Space

The analysis of relations between sequences and structures is facilitated by means of three formal discrete spaces (1) the sequence space being the space of all sequences of chain length n, (2) the shape space meant here as the space of all secondary structures that can be formed by sequences of chain length n, and (3) a conformation space containing all structures that can be formed by one particular sequence of chain length n.

Sequence Space

The sequence space is a metric space of cardinality κ^n with κ being the size of the alphabet. In addition to natural molecules built from the four-letter alphabet, $\{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}$ for DNA and $\{\mathrm{A},\mathrm{U},\mathrm{G},\mathrm{C}\}$ for RNA, sequences over three-letter, $\{\mathrm{A},\mathrm{U},\mathrm{G}\}$ [18] and two letter, $\{\mathrm{D},\mathrm{U}\}$[5] [19], alphabets were found to form perfect catalytic RNA molecules. Accordingly, we shall discuss also non-natural alphabets. The Hamming distance $d_{\mathrm{H}}(X_1, X_2)$, defined as the number of positions in which two aligned sequences differ,[6] fulfills the three requirements of a metric on sequence space:

$$d_{\mathrm{H}}(X_1, X_1) = 0 \,, \tag{1.8a}$$

$$d_{\mathrm{H}}(X_1, X_2) = d_{\mathrm{H}}(X_2, X_1) \,, \quad \text{and} \tag{1.8b}$$

$$d_{\mathrm{H}}(X_1, X_3) \leq d_{\mathrm{H}}(X_1, X_2) + d_{\mathrm{H}}(X_2, X_3) \,. \tag{1.8c}$$

The Hamming metric corresponds to choosing the single point mutation as the elementary move in sequence space.

Shape Space

The shape space comprises all possible secondary structures of chain length n. The number of structures is given by recursion (1.6) with $p = 1$, or the recursion (1.7) with $p = 1$, in case physically meaningful restrictions are applied to the lengths of hairpin loops (n_{lp}) or stacks (n_{st}). It is also straightforward to define a distance between structures. Several choices are possible (Sect. 1.2.1), we shall make use of two of them because they correspond to move sets that are important in kinetic folding of RNA (1) the base pair distance, $d_{\mathrm{P}}(S_j, S_j)$, and (2) the Hamming distance between the parentheses notations of structures, $d_{\mathrm{H}}(S_j, S_j)$, (Fig. 1.4). The Hamming distance between structures is simply the number of positions in which the two strings representing the secondary structures differ whereas the base pair distance is twice the minimal number

[5] Because of weak bonding in the **A – U** pair adenine has been replaced by **D** being 2,6-diamino-purine in these studies.

[6] Unless stated otherwise we shall consider here binary end-to-end alignments of sequences with equal lengths.

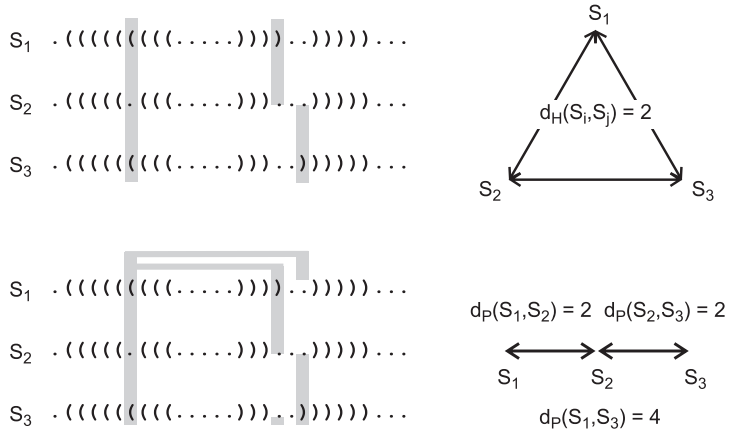

Fig. 1.4. Two measures of distances between secondary structures. The Hamming distance between parentheses notations of secondary structures is shown in the upper plot. Base pair opening and base pair closure contribute $d_{\mathrm{H}} = 2$, but simultaneous opening and closing, corresponding to a shift of one or more nucleotides, leads also to the same distance and the three structures are equidistant in shape space with Hamming metric. If we use the base pair distance instead, we find also $d_{\mathrm{P}} = 2$ for opening or closing of a base pair, but now the shift move is not in the move set and the two contributions for opening and closing add up to $d_{\mathrm{P}} = 4$

of base pairs that have to be erased and formed to convert one structure into the other.[7] Figure 1.4 shows the difference between the two distances in a sequence of two consecutive steps (1) a base pair is removed in going from S_1 to S_2, and (2) a base pair is closed, which involves one of the two nucleotides that formed the pair in S_1, in the step from S_2 to S_3. In base pair distance, we have $d_{\mathrm{P}}(S_1, S_3) = d_{\mathrm{P}}(S_1, S_2) + d_{\mathrm{P}}(S_2, S_3) = 4$, but in Hamming distance we find $d_{\mathrm{H}}(S_1, S_3) = d_{\mathrm{H}}(S_1, S_2) = d_{\mathrm{H}}(S_2, S_3) = 2$. The interpretation is straightforward: The base pair distance corresponds to a set of two moves, base pair opening and base pair closure, whereas the Hamming distance corresponds to a larger move set that involves, in addition to single base pair operations, (synchronous) shifts of one or more base pairs resulting in the migration of a bulge, internal loop or other structural element.

Conformation Space

The conformational space refers to a single sequence (X) and contains all structures that are compatible with X. Accordingly, it is a subspace of shape space:

$$\mathbf{C}[X] = \{\Omega \,|X \in \mathbf{C}[\Omega_i]\} \ . \tag{1.9}$$

[7] To make the two measures of distance comparable base pair distances are multiplied by factor 2 since base pair involves two nucleotides.

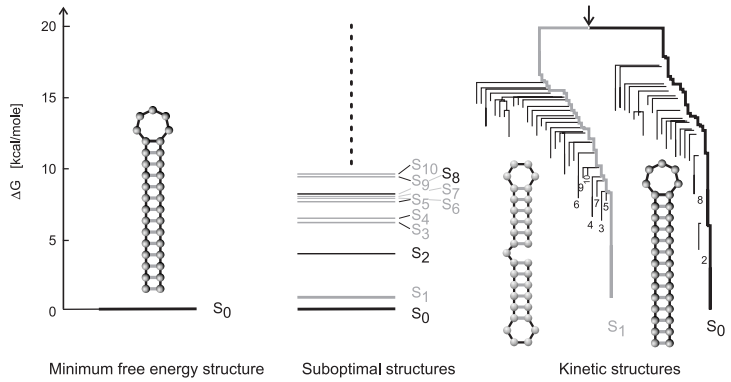

Fig. 1.5. Three notions of structures. The mfe-structure is shown as the only relevant conformation on the left-hand side corresponding in a formal sense to the zero temperature limit ($\lim T \to 0$). In the middle, we show the set of suboptimal structures as it is considered at equilibrium and temperature T in 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

The conformation space is of particular importance for kinetic folding of RNA. In addition, it represents the structural diversity of conformations that is accessible from the ground state Ω_0 on excitation. The two move sets discussed in the context of a measure of distance on shape space are also relevant for conformational space since are tantamount to elementary moves in kinetic folding of RNA [20–22]. In Fig. 1.5, we show by means of a real example how the notion of RNA structure is extended to account for suboptimal foldings and kinetic effects. Conventional RNA folding assigns the minimum free energy (mfe) structure to the sequence. As we have seen above many suboptimal structures accompany the mfe-structure 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. At finite time (Fig. 1.5; energy diagram on the right-hand side showing an RNA switch) the situation might be different and the RNA molecule may have one or more long-lived metastable conformation in addition to the mfe-structure. Then the actual molecular structure depends also on initial conditions and on the time window of the observation. The transitions between long-lived states are determined by the activation energies, which are shown in the construct of a barrier tree.[8]

[8] The barrier tree is a simplification of the conformational energy landscape and will be discussed in Sect. 1.2.2.

1.1.4 Computation of RNA Secondary Structures

Computation of secondary structures with minimum free energies [23] is based on the same principle as shown for counting the numbers of structures (Fig. 1.3). First, the free energies of the smallest possible substructures are taken or computed from a list of parameters, 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 reveals the structure. The conventional approach is empirical and uses the free energies and enthalpies of RNA model compounds to derive the parameters for the individual structural elements. These elements correspond to the substructures shown in Fig. 1.2 at sufficiently high resolution for sequence specific contributions.

As an example, we show the free stacking energy of a cluster of **GC**-pairs in Fig. 1.6, which is obtained from three free stacking energy parameters for the **GC**-pairs interacting at different geometries. On total, 21 different free stacking energy parameters are required for the six base pairs. 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 nucleotides stacking on top of regular stacks, especially stable configurations, for example tetraloops[9] with specific sequences, end-on-end stacking of stacks, etc. Stacks are (almost) the only structure stabilizing elements, because base pair stacking is a contribution with substantial negative free energy. Further structure stabilization comes from single bases stacking on stacks called "dangling ends" and some

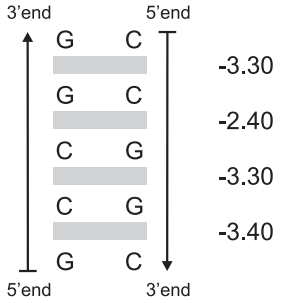

Fig. 1.6. The stacking parameters for the interaction between **GC** base pairs. Free energies of stacking are given for the three different interaction geometries (the first and the third paired pairs are identical). Values are given in kcal mol^{-1}. Additivity is assumed and therefore, we obtain a free energy of interaction of $\Delta G = -12.40$ kcal mol^{-1} for the stack of five pairs

[9] It is common to indicate the size of small hairpin loops by special wording: "triloops" are hairpin loops with three single nucleotides in the loop, "tetraloops" have four, and "pentaloops" five singles bases.

other sequence specific contributions. Loops are almost always destabilizing because of the entropic effect of the ring closure that freezes degrees of internal rotation.

Listings of parameters, which are updated every few years, can be found in the literature [24–27]. These parameters enter 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 [28, 29]).

1.1.5 Mapping Sequences into Structures

The numbers of physically accessible structures obtained from the recursion (1.7) are compared in Table 1.2 with the actual numbers of minimum free energy structures computed by means of a folding routine. To this end, all sequences of a chain length n were folded, grouped with respect to structures, and enumerated. The numbers refer to structures without single base pairs. Exhaustive folding of entire sequence spaces was performed for five different alphabets: **GC**, **UGC**, **AUGC**, **AUG**, and **AU**. As follows directly from the table, the mapping $\Omega = f(X)$ is many-to-one in all five alphabets. The set of sequences that form a given matching Ω, the preimage of Ω in sequence space

$$\mathbf{G}[\Omega] \;=\; f^{-1}(\Omega) \;\doteq\; \{X|f(X)=\Omega\}\;, \tag{1.10}$$

is turned into a graph, the neutral network G, by connecting all pairs of nodes with Hamming distance one by an edge. Global properties of neutral networks are derived by means of random graph theory [30]. The characteristic quantity for a neutral network is the degree of neutrality $\bar{\lambda}$, which is obtained by averaging the fraction of Hamming distance one neighbors that form the same minimum free energy structure, $\lambda_X = N^{(1)}_{\rm ntr}/\big(n\cdot(\kappa-1)\big)$ with $N^{(1)}_{\rm ntr}$ being the number of neutral one-error neighbors, over the whole network, $\mathbf{G}[\Omega]$:

$$\bar{\lambda}[\Omega] = \;\frac{1}{|\mathbf{G}(\Omega)|}\sum_{X\in\mathbf{G}[\Omega]}\lambda_X\;. \tag{1.11}$$

Connectedness of neutral networks is, among other properties, determined by the degree of neutrality [31]:

$$\text{With probability one network is}\begin{cases}\text{connected} & \text{if } \bar{\lambda}>\lambda_{\rm cr}\\ \text{not connected} & \text{if } \bar{\lambda}<\lambda_{\rm cr}\,,\end{cases} \tag{1.12}$$

$$\text{where } \lambda_{\rm cr} = 1-\kappa^{-1/(\kappa-1)}\,.$$

Computations yield $\lambda_{\rm cr} = 0.5$, 0.423, and 0.370 for the critical value in two-, three-, and four-letter alphabets, respectively.

Table 1.2. Comparison of exhaustively folded sequence spaces

Chain length	Number of sequences			Number of structures				
(n)	2^n	4^n	$s_n(1)$	GC	UGC	AUGC	AUG	AU
7	128	1.64×10^4	2	1	1	1	1	1
8	256	6.55×10^4	4	3	3	3	2	1
9	512	2.62×10^5	8	7	7	7	3	1
10	1,024	1.05×10^6	14	13	13	13	5	3
12	4,096	1.68×10^7	37	35	35	36	14	8
14	1.64×10^4	2.68×10^7	101	83	89	93	31	20
16	6.55×10^4	4.29×10^9	304	214	246	260	72	44
18	2.62×10^5	6.87×10^{10}	919	582	735		180	96
20	1.05×10^6	1.10×10^{12}	2,741	1,599	2,146		504	232
25	3.36×10^7	1.13×10^{15}	44,695	18,400				1,471
30	1.07×10^9	1.15×10^{18}	760,983	218,318				21,315

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 $s_n(1)$ with $\lambda = 3$ and $\sigma = 2$ as computed from the recursion (1.7) [12]. The parameters are taken from [25]

Random graph theory predicts a single largest component for nonconnected networks, i.e. networks below 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 nonuniform distribution of the sequences belonging to $\mathbf{G}[S_k]$ in sequence space caused by specific structural properties of S_k [32,33]. 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: $x_{\mathrm{G}} = x_{\mathrm{C}}$ and $x_{\mathrm{A}} = x_{\mathrm{U}}$. In case of **GC**-sequences, the neutral network is then depleted from sequences in the middle of sequence space and we find two largest components, one at excess **G** and one at excess **C**.

In Table 1.3 we show, as an example, computed values of the degree of neutrality, $\bar{\lambda}[S]$ in neutral networks derived from tRNA-like cloverleaf structures with different stack lengths of the hairpin loops. The most striking feature of the data is the weak structure dependence of $\bar{\lambda}[S]$ with a family: For a given alphabet the cloverleafs S_1, S_2, S_3, and S_4, have almost the same $\bar{\lambda}$ values irrespective of the stability of the corresponding folds. Because of the shorter stack lengths in S_1, S_2 and S_3 and the weakness of the **AU** pair no

Table 1.3. Degree of neutrality in different nucleotide alphabets

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

The values for the degree of neutrality, $\bar{\lambda}$, were obtained by sampling 1,000 random sequences folding into the four cloverleaf structures with different stack sizes [a] using the inverse folding routine [28]. [a] The following cloverleaf structures were used:

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

AU-sequences forming these structures were obtained by inverse folding. The same was found for S_1 in case of **AUG**-sequences. Considering the fact that λ_{cr} decreases from two to four-letter alphabets, we see that neutral networks in two-letter sequence spaces ($\bar{\lambda} \approx 0.06$ and $\lambda_{\mathrm{cr}} = 0.5$) and four-letter sequence spaces ($\bar{\lambda} \approx 0.3$ and $\lambda_{\mathrm{cr}} = 0.37$) must have very different extensions, the former being certainly non connected and whereas the latter come close to threshold.

The extension of neutral networks can be visualized also by evaluating the lengths of neutral path. A neutral path connects pairs of neighboring neutral sequences of Hamming distance $d_{\mathrm{H}} = 1$ for single nucleotide exchanges or $d_{\mathrm{H}} = 1, 2$ 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 neighbors that are closer to the reference sequence. Table 1.4 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 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. 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, and provide an explanation why the (almost unconstrained) ribozymes of Schultes and Bartel [35] stay functional along very long neutral paths whereas functional tRNAs, which have to fulfil multiple constraints, tolerate only very limited variability in their sequences.

Table 1.4. The lengths of neutral paths through sequence space

Molecule	Alphabet	Degree of neutrality ($\bar{\lambda}$)	Neutral path length $\bar{d}_{\mathrm{H}}(X_0, X_{\mathrm{f}})$
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

The degree of neutrality, $\bar{\lambda}$, and the mean lengths of neutral paths through sequence space, $\bar{d}_{\mathrm{H}}(X_0, X_{\mathrm{f}})$ (with X_0 being the initial and X_{f} the last sequence), is compared for three examples (1) folding of (stand alone) **AUGC** sequences of chain lengths $n = 100$, (2) cofolding of **AUGC** sequences of chain lengths $n = 100$ with a single fixed sequence, and (3) cofolding of **AUGC** sequences of chain lengths $n = 100$ with two single fixed sequences. The values represent averages over samples of 1,200 random sequences. The value for the path length in **GC** sequence space with $n = 100$ is an estimate from Fig. 10 in [34].

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 (e.g., [36, 37]). Several theoretical investigations were also dealing with random pools of RNA sequences [38–41] and showed, for example, that natural RNA molecules have lower free folding energies than the average of random energies thus demonstrating the effect of evolutionary selection for stable structures.

1.1.6 Suboptimal Structures and Partition Functions

Algorithms for the computation of suboptimal conformations have been developed and two of them are frequently used [42, 43]. 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 latter of the two algorithms [43] has been designed for the calculation of all conformations within a given energy band above the mfe and adopts a technique originally proposed for suboptimal alignments of sequences [44]. 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 Fig. 1.5, the set of structures, mfe and suboptimal conformations $\{S_0, S_1, S_2, \ldots\}$, is ordered since their free energies, $\{\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots\}$ fulfill the relation $\varepsilon_0 \leq \varepsilon_1 \leq \varepsilon_2 \ldots$.

At equilibrium and temperature T, the individual conformations form a Boltzmann ensemble that contains a structure S_j with the Boltzmann weight

$\gamma_j = g_j \exp\bigl(-(\varepsilon_j - \varepsilon_0)/RT\bigr)/Q(T)$, where R is the Boltzmann constant for one mole, $R = N_{\mathrm{L}} \cdot k_{\mathrm{B}}$, and $Q(T)$ is the partition function[10]

$$Q(T) \;=\; \sum_i g_i \exp\bigl(-(\varepsilon_i - \varepsilon_0)/RT\bigr)\,. \tag{1.13}$$

Instead of having a structure with a set of 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

$$P(X,T) \;=\; \sum_k \gamma_k\, A(S_k) \;\;\text{or}\;\; p_{ij}(X,T) \;=\; \sum_k \gamma_k\, a_{ij}(S_k)\,, \tag{1.14}$$

which is a Boltzmann weighted superposition of the adjacency matrices (1.1) of the individual structures with the following properties: In the limit $T \to 0$, the base pairing probabilities converge to the base pairing pattern of S_0 (for a nondegenerate ground state, $\varepsilon_0 < \varepsilon_1$) as described by the adjacency matrix $A(S_0)$ and in the limit $T \to \infty$ 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 [13]. It has been incorporated into the Vienna RNA package [28].

1.2 Design of RNA Structures

The design of RNA molecules boils down to finding sequences that fold into molecules with predefined structures and properties. Consequently, an algorithm is needed that computes sequences that fold into predefined mfe structures. The required procedure thus corresponds to an inversion of the conventional folding procedure.

1.2.1 Inverse Folding

Given a sequence X, the *folding problem* consists in finding a matching Ω that minimizes an energy function $E(X;\Omega)$ and (if desired) satisfies other constraints, such as the no-pseudoknot condition. In Sect. 1.1.4, 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 Ω and search for a sequence

[10] Sometimes different microstates S_i with the same free energy ε_j are lumped together to form one "mesoscopic" state in the partition function and then the factor g_j accounts for this degeneracy.

X that has Ω as an optimal structure. We denote the set of solutions of the inverse folding problem by $f^{-1}(\Omega)$. Note that $f^{-1}(\Omega)$ may be empty, since there are logically possible secondary structures that are not formed as minimum energy structures of any sequence.

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 as a combinatorial optimization problem. To this end, we consider a measure $D(\Omega_1, \Omega_2)$ for the structural dissimilarity of two RNA secondary structures Ω_1, Ω_2. A variety of such distance measures have been described in the literature [28, 45–48]. Since we will be interested here only in the sequences of equal length, we may simply use the cardinality of the symmetric difference of Ω_1 or in Ω_2:

$$D(\Omega_1, \Omega_2) = \big|(\Omega_1 \cup \Omega_2) \setminus (\Omega_1 \cap \Omega_2)\big|. \tag{1.15}$$

Clearly, sequence X folds into structure Ω, if and only if $\Xi(X) = D(\Omega, f(X)) = 0$. 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:

$$f^{-1}(\Omega) \subseteq \mathbf{C}[\Omega]. \tag{1.16}$$

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 [28]: Let $E(X;\Omega)$ be the energy of structure Ω for sequence X, and let $G(X)$ be the ensemble free energy of sequence X, which can be computed by McCaskill's algorithm [13]. Sequences with the desired property minimize

$$\Xi(X) = E(X;\Omega) - G(X) = -RT \ln \gamma_X(\Omega)\,, \tag{1.17}$$

where $\gamma_X(\Omega)$ is the probability of structure Ω in the Boltzmann ensemble of sequence X.

It has been found empirically [28] 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 Υ or by replacing one of the six pairing combinations by another one in a pair in Ω. A mutant is accepted if the cost function $\Xi(X)$ decreases. In a more sophisticated version, implemented in the program `RNAinverse`, a significant speedup 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 [49].

1.2.2 Multiconformational RNAs

Figure 1.5 indicates 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. Stable alternative conformations have been observed experimentally for a variety of RNA molecules [50–53].

Alternative conformations of the same RNA sometimes determine completely different functions [54, 55]. SV11, for instance, is a relatively small molecule that is replicated by Qβ replicase [56, 57]. It exists in two major conformations, a metastable multicomponent structure and a rod-like conformation, constituting the stable state, separated by a huge energy barrier. While the metastable conformation is a template for Qβ 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 [58]. Another, 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 [35].

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

$$\mathbf{C}[\Omega_1, \Omega_2, \ldots, \Omega_k] = \bigcap_{j=1}^{k} \mathbf{C}[\Omega_j] \tag{1.18}$$

and if so, what is the size of this intersection of sets of compatible sequences. To answer this question, it is useful to consider the graph Ψ 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

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

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

$$\big|\mathbf{C}[\Omega_1, \Omega_2, \ldots, \Omega_k]\big| = \prod_{\text{components } \psi \text{ of } \Psi} F(\psi)\,, \tag{1.19}$$

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

(3) For the biophysical alphabet holds: $\bigcap_j \mathbf{C}[\Omega_j] \neq \emptyset$ *if and only if* Ψ 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

$$F(P_k) = 2\big(\mathrm{Fib}(k) + \mathrm{Fib}(k+1)\big) = 2\mathrm{Fib}(n+2) \tag{1.20}$$

$$F(C_k) = 2\big(\mathrm{Fib}(k-1) + \mathrm{Fib}(k+1)\big)\,, \tag{1.21}$$

where P_k and C_k are path and cycle components of Ψ with k vertices.

For a proof of these propositions see [31, 59]. Interestingly, for two structures there is always a nonempty intersection $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]$. In contrast, the chance that the intersection of three randomly chosen structures in nonempty decreases exponentially with sequence length [60]. Recently, an alternative attempt has been made to extend the design aspect of the intersection theorem to three or more sequences [61].

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 Ω_1 and Ω_2 here. For example, one might be interested in sequences that have two prescribed structures Ω_1 and Ω_2 as stable local energy minima with roughly equal energy, and for which the energy barrier between these two minima is roughly ΔE. It is not hard to design a cost function $\Xi(X)$ for this problem. In [59], the following ansatz has been used successfully:

$$\begin{aligned}\Xi(X) = E(X,\Omega_1) + E(X,\Omega_2) - 2G(X) + \xi\,\big(E(X,\Omega_1) - E(X,\Omega_2)\big)^2 \\ +\zeta\,\big(B(X,\Omega_1,\Omega_2) - \Delta E\big)^2\,.\end{aligned} \tag{1.22}$$

Here, $B(X,\Omega_1,\Omega_2)$ is the energy barrier between the two conformations Ω_1, Ω_2, which can be readily computed from the barrier tree of the sequence X.

1.2.3 Riboswitches

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 in-active conformation of the transcript [62]. 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 [63]. 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 behavior are the riboswitches [64]. These are 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 [65], 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 with artificial aptamers obtained by in vitro directed evolution experiments. Riboswitches regulate several key metabolic pathways [66, 67] 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 ongoing, e.g., [68, 69]. Riboswitches and engineered allosteric ribozymes [70, 71] demonstrate impressively that RNA is indeed capable of maintaining and regulating a complex metabolic state without the help of proteins.

1.3 Processes in Conformation, Sequence, and Shape Space

Kinetic folding and evolutionary optimization of RNA molecules are considered as stochastic processes, in particular as constrained walks in conformation and sequence and/or shape space. We present a brief overview of the basic concepts and then consider the evolution of noncoding RNA molecules as one actual and particular interesting example.

1.3.1 Kinetic Folding

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

$$\Omega_0 \to \Omega_1 \to \Omega_2 \to \cdots \to \Omega_T \ , \tag{1.23}$$

where initial and target structures, Ω_0 and Ω_T, may be chosen at will. Commonly, $\Omega_0 = \mathbf{O}$ and $\Omega_T = S_0$ are used corresponding to the open chain and the mfe-structure, respectively. Individual trajectories (1.23) 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 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, it

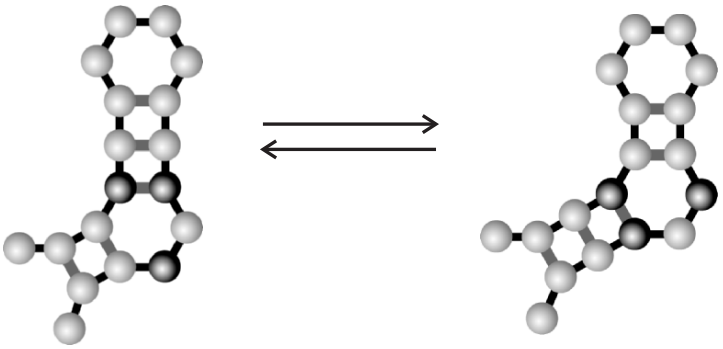

Fig. 1.7. The shift move in kinetic RNA folding. The shift move is a combination of base pair opening and base pair closure that occurs simultaneously. The requirement for an allowed shift move is that it takes place within one substructure element, bulge, internal loop or multiloop. Shifts involving free ends are also considered legitimate

contains base pair closure and base pair opening according to the conventional secondary structure rules (Conditions 1–3). Such a move set corresponds to the base pair distance, d_{P}, as metric in shape space (Fig. 1.4). It turned out to be important to introduce also a shift move (Fig. 1.7) since the trajectories approach the target much faster then [20]. If the move set is extended to simultaneous shifts of as many nucleotides as possible within a given substructure element, the set has the Hamming metric between parentheses notation of structures, $d_{\mathrm{H}}(S_i, S_j)$ (Fig. 1.4), as proper measure of distance.

The stochastic process (1.23) can also be described by a master equation for the probabilities of the ensemble: $P_k(t)$ is the probability to observe the conformation S_k at time t. The time derivatives fulfil the equation

$$\begin{aligned}\frac{\mathrm{d}P_k}{\mathrm{d}t} &= \sum_{i=0}^{m+1}\left(P_{ik}(t) - P_{ki}(t)\right) = \sum_{i=0}^{m+1} k_{ik}P_i - P_k\sum_{i=0}^{m+1} k_{ik} \\ \text{with} \quad & k = 0, 1, \ldots, m+1 \quad \text{and} \quad i \to k \in \text{move set}\,, \end{aligned} \tag{1.24}$$

where we assume that the open chain conformation **O** is not part of the suboptimal conformations, $S_1, \ldots, S_m$. The transition probabilities are computed from the free energies of the conformations

$$P_{ik}(t) = k_{ik}\,P_i(t) = P_i(t)\,\mathrm{e}^{-(g_k-g_i)/(2RT)}/\Sigma_i\ , \tag{1.25}$$

$$P_{ki}(t) = k_{ki}\,P_k(t) = P_k(t)\,\mathrm{e}^{-(g_i-g_k)/(2RT)}/\Sigma_k\ , \tag{1.26}$$

$$\text{with} \quad \Sigma_j = \sum_{i=0, i\neq j}^{m+1} \exp\left(-(g_j-g_i)/(2RT)\right)\ .$$

To avoid the necessity of additional parameters the free energies are taken from the suboptimal foldings. Calibration of the time scale occurs through

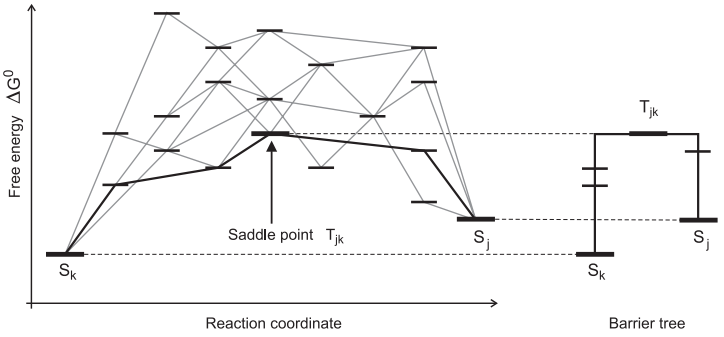

Fig. 1.8. Construction of barrier trees. The set of suboptimal conformations is related by a move set as shown in the left-hand part of the sketch. The barrier tree is derived from the set of suboptimal structures by eliminating all conformations except local minima of the free energy surface and minima 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 on physically meaningful alterations of the move set

adjusting the folding kinetics of a model system to the experimental data. Although it is straightforward to solve the master equation (1.24) by means of an eigenvalue problem, practical difficulties arise from the enormously high number of suboptimal conformations determining the dimensionality of the system [72].

A simplification of full kinetic folding is introduced in the form of "barrier trees" (Fig. 1.8). All suboptimal conformations that do 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 (1.24). Cases of less satisfactory agreement can be predicted [72].

1.3.2 Evolutionary Optimization

Evolution of RNA molecules based on replication, mutation, and selection in constant environment can be described by an ODE [73]:

$$\begin{aligned}\frac{\mathrm{d}x_i}{\mathrm{d}t} &= \sum_{k=1}^{m} f_k\, Q_{ki}\, x_k - x_i\, \phi(t)\,, \quad i = 1, \ldots, m\ , \\ \phi(t) &= \sum_{k=1}^{m} f_k\, x_k(t)\,. \end{aligned} \tag{1.27}$$

Herein the concentrations of individual RNA sequences are denoted by $x_i = [X_i]$ and Q_{ij} are the elements of a mutation matrix whose elements, in the

simplest case of the uniform error rate assumption, can be expressed by an (average) error rate p per site and replication.

$$Q_{ij} \;=\; p^{d_{\mathrm{H}}(X_i,X_j)} \cdot (1-p)^{n-d_{\mathrm{H}}(X_i,X_j)} \;. \tag{1.28}$$

The mutation probability thus is only a function of the error rate and the Hamming distance $d_{\mathrm{H}}(X_i, X_j)$ between the two sequences involved. The results of the analysis of replication–mutation kinetics have been presented and discussed extensively [74–77] and we dispense here from repeating them. Kinetic differential equations refer to infinite population size and accordingly, a different description is required for the study of finite size effects on evolutionary optimization. In addition, population dynamics is considered as a process taking place exclusively in sequence space and structural properties enter the model as parameters only.

Replication and mutation of RNA molecules leading to selection in confined populations have indeed been studied also in finite populations. The best-suited stochastic methods for modeling the system are multitype branching processes [78]. A simplified version of the branching trajectories in replication and mutation is shown in Fig. 1.9. As expected, the mean value of the stochastic process coincides with the deterministic solution [80]. The standard deviation, however, can be enormous as we shall see in detail later.

To simulate the interplay between mutation acting on the RNA sequence and selection operating on phenotypes, here RNA structures, the sequence–structure map has to be an integral part of the model [81–83]. 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) by using Gillespie's algorithm [84, 85]. In target search problems, the replication rate of a sequence X_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,

$$f_k(S_k, S_T) \;=\; \frac{1}{\alpha \,+\, d_{\mathrm{H}}(S_k, S_T)/n} \,, \tag{1.29}$$

which increases when S_k approaches the target (α is an adjustable parameter that was 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, $N \geq 1,000$ it has been never observed.

A typical trajectory is shown in Fig. 1.10. The mean distance to target of the population decreases in steps until the target is reached [82,83,86]. Individual (short) adaptive phases are interrupted by long quasi-stationary epochs. 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

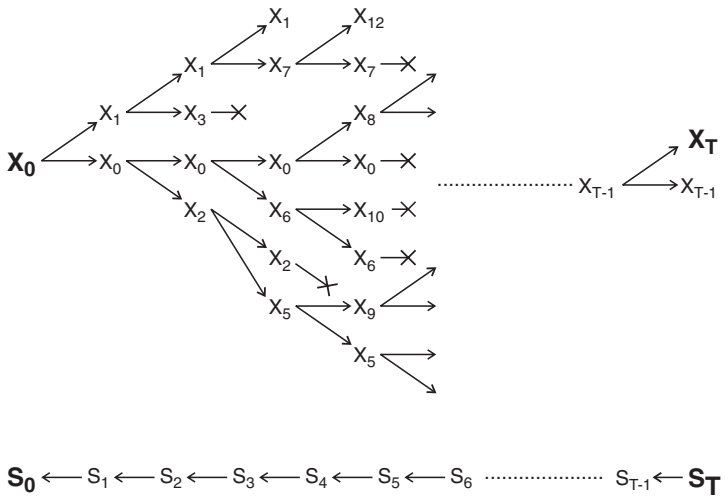

Fig. 1.9. Evolutionary optimization as a multitype branching process. The sketch in the upper part shows only replication acts that lead to mutation. A full genealogy is a time ordered series, which records all individual replication acts, for example $X_0, \ldots, X_0, X_a, \ldots, X_a, X_b, \ldots, \ldots, X_{T-1}, X_T$ leading to target. The population size is either constant (Moran model [79]) 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 molecules that is tantamount to the end of some trajectory in the system. The sketch on the *bottom* illustrates the reconstruction of the optimization run by means of a "relay series"

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 opposite direction from the final structure to the initial shape (see the lower part of Fig. 1.9). 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 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.[11] Usage of the relay series and its theoretical background allows for classification of transitions [83, 87]. Inspection of the relay series on the quasistationary plateaus allows for a distinction of two scenarios:

(1) The structure is constant and we observe neutral evolution in the sense of Kimura's theory of neutral evolution [88]. In particular, the number of

[11] It is important to stress two facts about relay series (1) the same shape may appear two or more times in a given relay series. Then, it was extinct between two consecutive appearances. (2) A relay series is not a genealogy, which is the full recording of parent–offspring relations a time-ordered series of genotypes.

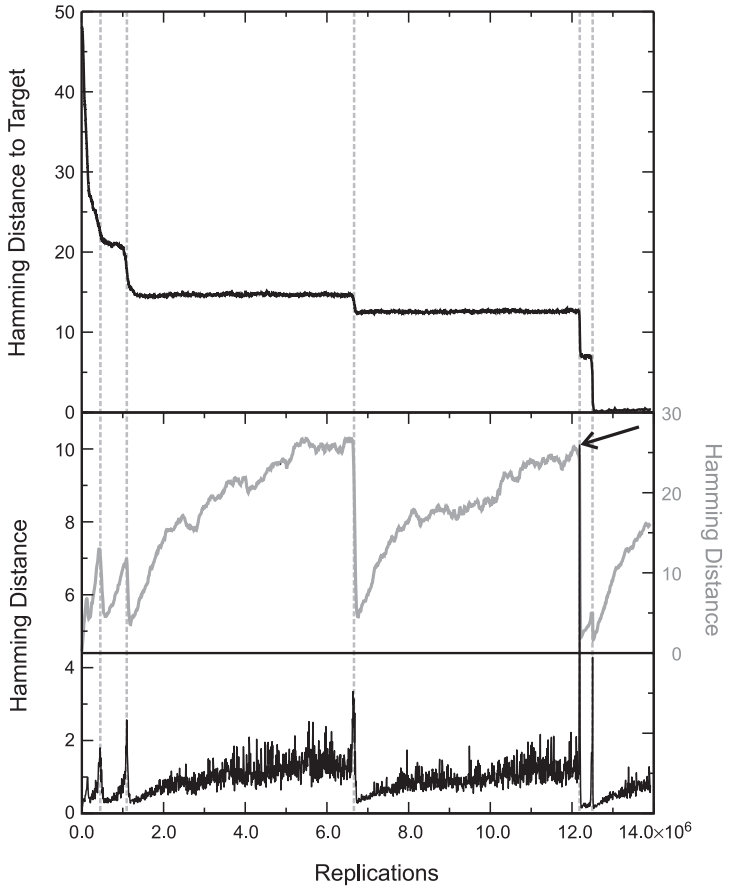

Fig. 1.10. A trajectory of evolutionary optimization. The *topmost* plot presents the mean distance to the target structure of a population of 1,000 molecules. The plot in the middle shows the width of the population in Hamming distance between sequences and the plot at the *bottom* is a measure of the velocity with which the center of the population migrates through sequence space. A remarkable synchronization is observed: At the end of a quasi-stationary plateau an adaptive phase of the migration to target is initiated that is accompanied by a drastic shrinking of the population width and a jump in the population center. A mutation rate of $p = 0.001$ was chosen, the replication rate parameter is defined in (1.29), and initial as well as target structure is shown in Table 1.5

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 [89].

(2) The process during the stationary epoch involves several structures with identical replication rates and the relay series reveal 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 Fig. 1.10 that shows the width of the population as a function of time [86, 90]. The population width increases during the quasi-stationary epoch and sharpens almost instantaneously after a sequence had been formed that allows for a continuation of the optimization process. The scenario at the end of the plateau corresponds to a *bottle neck* of evolution. The lower part of the figure shows a plot of the migration rate or drift of the population center and confirms this interpretation: The drift is almost always very slow unless the population center "jumps" from one point in sequence space to the other point where the sequence initiating the new adaptive phase had appeared. A closer look at the figure reveals the coincidence of the three events (1) beginning of a new adaptive phase, (2) collapse-like narrowing of the population spread, and (3) jump-like migration of the population center.

Table 1.5 collects some numerical data obtained from repeated evolutionary trajectories under identical conditions.[12] Individual trajectories show enormous scatter in the time or the number of replications required to reach

Table 1.5. Statistics of the optimization trajectories

Alphabet	Population size	Number of runs	Real time from start to target		Number of replications (10^7)	
	(N)	(n_R)	Mean value	σ	Mean value	σ
AUGC	1,000	120	900	$+1,380-542$	1.2	$+3.1-0.9$
	2,000	120	530	$+880-330$	1.4	$+3.6-1.0$
	3,000	1,199	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	5,160	$+15,700-3,890$	–	–
	3,000	278	1,910	$+5,180-1,460$	7.4	$+35.8-6.1$
	10,000	40	560	$+1,620-420$	–	–

The table shows the results of sampled evolutionary trajectories leading from a random initial structure S_{I} to the structure of $\mathrm{tRNA}^{\mathsf{phe}}$, S_{T} as target. [a] Simulations were performed with an algorithm introduced by Gillespie [84,85,91]. The time unit is here undefined. A mutation rate of $p = 0.001$ per site and replication was used. The mean and standard deviation were calculated under the assumption of a log-normal distribution that fits well the data of the simulations [a] The structures S_{I} and S_{T} were used in the optimization:

S_{I}: ((.((((((((((((((.............(((....)))......))))))).)))))))..))...(((......)))
S_{T}: ((((((...((((........)))).(((((.......)))))).....(((((.......)))))).)))))).....

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

the target. The mean values and the standard deviation were obtained from statistics of trajectories under the assumption of a log-normal distribution. Despite the scatter three features are unambiguously detectable:

(1) 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.
(2) The time to target decreases with increasing population size.
(3) The number of replications required to reach the target increases with population size.

Combining items (2) and (3) allows for a clear conclusion concerning time and material requirements of the optimization process: Fast optimization requires large populations whereas economic use of material suggests to work with small population sizes.

1.3.3 Evolution of Noncoding RNAs

In recent year, there has been mounting evidence that noncoding RNAs in fact dominate the regulatory networks of the cell (see, e.g., [92–96] for reviews). Unlike protein coding genes, noncoding 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 noncoding RNA genes has remained elusive, see e.g., [97]. Most classes of the currently known noncoding RNAs, however, are characterized by a common, evolutionarily very well conserved, secondary structure, while at the same time their sequence is rather variable. This feature can be understood as a consequence of stabilizing selection acting (predominantly) on the secondary structure, while the sequence remains (mostly) free to diffuse on the neutral network.

Diffusion in sequence space, i.e., Kimura's *neutral theory* [88], in fact, forms the conceptual basis of phylogenetic inference. It is important to notice, however, that substitution rates differ dramatically between unpair regions and base-paired regions, since sequence positions that form conserved base pairs are highly correlated. This effectively restricts the diffusion process to the neutral network [89]. Corresponding stochastic models of sequence evolution are described, e.g., in [98–101]. The `phase` package [102,103] implements such a model and is specifically designed to infer phylogenies from RNAs that have a conserved secondary structure, including rRNAs.

Structural conservation in the presence of sequence variation is also the basis of recent comparative genomics approaches toward RNA gene finding. The first tool of this type, `qrna` [104] is based upon an SCFG approach to asses the probability that a pair of aligned sequences evolved under a constraint for preserving a secondary structure. The program `RNAz` [105] uses two independent criteria for classification: a z-score measuring thermodynamic stability of

individual sequences, and a *structure conservation index* obtained by comparing folding energies of the individual sequences with the predicted consensus folding. Both quantities measure different aspects of stabilizing selection for RNA structure.

In the remainder of this section, we give a brief overview of the evolutionary patterns of the most prominent RNA families. For a recent, much more detailed review, we refer to [106]. Similar to protein-coding genes, most ncRNAs appear in multiple paralogous copies in the genome. Unlike protein coding genes, however, some classes of ncRNAs appear to be associated with a large number of pseudogenes, this is in particular true for tRNAs and small nuclear RNAs.

Ribosomal RNA sequences are probably the most widely used source of data in molecular phylogenetics: rRNAs are abundant, very well conserved, and therefore easy to access experimentally. Because of concerted evolution, usually, there are no divergent paralogues despite the fact that rRNA genes, in higher eukaryotes at least, typically are arranged in large tandem-repeated clusters. It may not come as a surprise, however, that divergent paralogues of both SSU [107, 108] and LSU [109] do occur in some lineages.

Multiple copies of functional tRNA genes, the existence of numerous pseudogenes, and tRNA-derived repeats are general characteristics of tRNA evolution [110]. Comparative sequence analysis of transfer RNA by means of statistical geometry provides strong evidence that transfer RNA sequences diverged long before the divergence of archaea and eubacteria [111]. Indeed, in a sample of tRNAs for very diverse organisms, those with the same anticodon rather than those from the same organism form coherent subtrees. Models for the origin of tRNA from even simpler components are discussed, e.g., in [112–114].

Like rRNAs and tRNAs, there are typically multiple genomic copies of the spliceosomal snRNAs. Surprisingly, the copy numbers in the genome vary significantly between even closely related species. The mechanism generating this pattern remains unclear at present.

The absence of small nucleolar RNAs (snoRNAs) from bacterial genomes suggests that snoRNPs arose in the archaeal and eukaryotic branch after the divergence of the bacteria. SnoRNAs fall into two structurally distinct classes, box C/D and H/ACA snoRNAs, that guide two different types of chemical modifications of rRNAs and some other ncRNAs, see e.g., [115] for a review. The numerous box C/D and H/ACA snoRNAs of Archaea and Eukarya are likely to have arisen through duplication and variation of the guide sequence [116]. A recent case study of the evolution of the vertebrate U17/E1, E2, and E3 snoRNAs [106] shows that divergent paralogues of snoRNAs have been produced throughout vertebrate evolution. Most vertebrate snoRNAs are encoded in introns. Interestingly, paralogues often reside in adjacent introns of the same gene. In some cases at least, these copies appear to be subject to concerted evolution.

MicroRNA evolution follows a pattern on its own. The mature microRNA is only about 22nt long. It is processed from a thermodynamically very stable stem-loop structure of about 70–80nt in length. Frequently, tandem duplications seem to lead to poly-cistronic transcripts [117]. In contrast to rRNA, tRNAs, and snRNAs, divergent paralogues appear to be the rule rather than the exception for microRNAs. Consequently, most microRNAs that can be traced back to the vertebrate ancestor are present in 2–4 paralogues copies that are remnants of the vertebrate-specific genome duplications. Interestingly, it has been found that tandem-duplications typically predate the non-local duplication events [118]. The origin of microRNAs remains unknown. As yet, no microRNA with homologues in both animals and plants has been described so far, although the microRNA processing machinery in animals and plants is clearly homologous. In [119] it has been argued that microRNA could easily arise *de novo* since stem-loop structures resembling pre-miRNAs are very abundant secondary structures in genomic sequences. A recent study on the evolution of animal miRNAs showed that a large number of novel microRNAs appeared in early vertebrates and in placental mammals, while the rate of annotation is otherwise much lower.

Acknowledgments

This work has been supported financially by the Austrian "Fonds zur Förderung der wissenschaftlichen Forschung" (FWF), Project Nos. P-13093, P-13887, and P-14898. Part of the work has been carried out during a visit at the Santa Fe Institute within the External Faculty Program. The support is gratefully acknowledged.

References

1. D. Thirumalai, Proc. Natl. Acad. Sci. **95**, 11506 (1998)
2. D. Thirumalai, N. Lee, S.A. Woodson, D.K. Klimov, Annu. Rev. Phys. Chem. **52**, 751 (2001)
3. D.E. Draper, RNA **10**, 335 (2004)
4. M. Wu, I. Tinoco, Jr., Proc. Natl. Acad. Sci. USA **95**, 11555 (1998)
5. S.R. Holbrook, Curr. Opt. Struct. Biol. **15**, 302 (2005)
6. G. Varani, I. Tinoco, Jr., Q. Rev. Biophys. **24**, 479 (1991)
7. S. Louise-May, P. Auffinger, E. Westhof, Curr. Opin. Struct. Biol. **6**, 289 (1996)
8. P.F. Stadler, J. Math. Chem. **20**, 1 (1996)
9. P. Schuster, P.F. Stadler, in *Discrete Models of Biopolymers*. ed. by M.J.C. Crabbe, M. Drew, A. Konopka. Handbook of Computational Chemistry (Marcel Dekker, New York, 2004) pp. 187–222
10. C.M. Reidys, P.F. Stadler, SIAM Rev. **44**, 3 (2002)
11. E. Rivas, S.R. Eddy, J. Mol. Biol. **285**, 2053 (1999)
12. I.L. Hofacker, P. Schuster, P.F. Stadler, Discr. Appl. Math. **89**, 177 (1998)
13. J.S. McCaskill, Biopolymers **29**, 1105 (1990)

14. M.S. Waterman, *Introduction to Computational Biology: Maps Sequences and Genomes* (Chapman and Hall/CRC, London/Boca Raton, 2000)
15. M.S. Waterman, T.F. Smith, Math. Biosci. **42**, 257 (1978)
16. M. Tacker, P.F. Stadler, E.G. Bornberg-Bauer, I.L. Hofacker, P. Schuster, Eur. Biophys. J. **25**, 115 (1996)
17. M. Zuker, D. Sankoff, Bull. Math. Biol. **46**, 591 (1984)
18. J. Rogers, G. Joyce, Nature **402**, 323 (1999)
19. J.S. Reader, G.F. Joyce, Nature **420**, 841 (2002)
20. C. Flamm, W. Fontana, I.L. Hofacker, P. Schuster, RNA **6**, 325 (1999)
21. W. Zhang, S.J. Chen, J. Chem. Phys. **118**, 3413 (2003)
22. W. Zhang, S.J. Chen, J. Chem. Phys. **119**, 8716 (2003)
23. M. Zuker, P. Stiegler, Nucleic Acids Res. **9**, 133 (1981)
24. D.H. Turner, N. Sugimoto, Annu. Rev. Biophys. Chem. **17**, 167 (1988)
25. A.E. Walter, D.H. Turner, J. Kim, M.H. Lyttle, P. Müller, D.H. Mathews, M. Zuker, Proc. Natl. Acad. Sci. USA **91**, 9218 (1994)
26. D.H. Mathews, J. Sabina, M. Zuker, D.H. Turner, J. Mol. Biol. **288**, 911 (1999)
27. D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, D.H. Turner, Proc. Natl. Acad. Sci. USA **101**, 7287 (2004)
28. I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster, Mh. Chemie **125**, 167 (1994)
29. I.L. Hofacker, Nucleic Acids Res. **31**, 3429 (2003)
30. B. Bollobás, *Random Graphs* (Academic, London, 1985)
31. C. Reidys, P.F. Stadler, P. Schuster, Bull. Math. Biol. **59**, 339 (1997)
32. W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P. Schuster, Mh. Chemie **127**, 355 (1996)
33. W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P. Schuster, Mh. Chemie **127**, 375 (1996)
34. P. Schuster, J. Biotechnol. **41**, 239 (1995)
35. E. Schultes, D. Bartel, Science **289**, 448 (2000)
36. D.M. Held, S.T. Greathouse, A. Agrawal, D.H. Burke, J. Mol. Evol. **57**, 299 (2003)
37. Z. Huang, J.W. Szostak, RNA **9**, 1456 (2003)
38. W. Fontana, D.A.M. Konings, P.F. Stadler, P. Schuster, Biopolymers **33**, 1389 (1993)
39. P.G. Higgs, J. Phys. I (France) **3**, 43 (1993)
40. J. Gevertz, H.H. Gan, T. Schlick, RNA **11**, 853 (2005)
41. P. Clote, F. Ferré, E. Kranakis, D. Krizanc, RNA **11**, 578 (2005)
42. M. Zuker, Science **244**, 48 (1989)
43. S. Wuchty, W. Fontana, I.L. Hofacker, P. Schuster, Biopolymers **49**, 145 (1999)
44. M.S. Waterman, T.H. Byers, Math. Biosci. **77**, 179 (1985)
45. B.A. Shapiro, K. Zhang, Comput. Appl. Biosci. **6**, 309 (1990)
46. C. Reidys, P.F. Stadler, Comput. Chem. **20**, 85 (1996)
47. V. Moulton, M. Zuker, M. Steel, R. Pointon, D. Penny, J. Comput. Biol. **7**, 277 (2000)
48. M. Höchsmann, T. Töller, R. Giegerich, S. Kurtz, *Proceedings of the Computational Systems Bioinformatics Conference*, vol. 159 (Stanford, CA, CSB 2003)
49. M. Andronescu, A.P. Fejes, F. Hutter, H.H. Hoos, A. Condon, J. Mol. Biol. **336**, 607 (2004)

50. J.R. Fresco, A. Adains, R. Ascione, D. Henley, T. Lindahl, Cold Spring Harb. Symp. Quant. Biol. **31**, 527 (1966)
51. E.R. Hawkins, S.H. Chang, W.L. Mattice, Biopolymers **16**, 1557 (1977)
52. V.L. Emerick, S.A. Woodson, Biochemistry **32**, 14062 (1993)
53. R. Micura, C. Höbartner, Chembiochem **4**, 984 (2003)
54. T. Baumstark, A.R. Schroder, D. Riesner, EMBO J. **16**, 599 (1997)
55. A.T. Perrotta, M.D. Been, J. Mol. Biol. **279**, 361 (1998)
56. C.K. Biebricher, S. Diekmann, R. Luce, J. Mol. Biol. **154**, 629 (1982)
57. C.K. Biebricher, R. Luce, EMBO J. **11**, 5129 (1992)
58. H. Zamora, R. Luce, C.K. Biebricher, Biochemistry **34**, 1261 (1995)
59. C. Flamm, I.L. Hofacker, S. Maurer-Stroh, P.F. Stadler, M. Zehl, RNA **7**, 254 (2000)
60. I. Abfalter, C. Flamm, P.F. Stadler, in *Design of Multistable Nucleic Acid Sequences*. ed. by H.W. Mewes, V. Heun, D. Frishman, S. Kramer. Proceedings of the German Conference on Bioinformatics (GCB 2003), vol. 1 (Belleville Verlag Michael Farin, München, 2003) pp.1–7
61. P. Clote, L. Gaşieniec, R. Kolpakov, E. Kranakis, D. Krizanc, J. Theor. Biol. **236**, 216 (2005)
62. E. Merino, C. Yanofsky, in *Regulation by Termination-Antitermination: A Genomic Approach*. ed. by A.L. Sonenshein, J.A. Hoch, R. Losick. *Bacillus subtilis* and its Closest Relatives: From Genes to Cells (ASM, Washington, DC, 2002) pp. 323–336
63. T.M. Henkin, C. Yanofsky, Bioessays **24**, 700 (2002)
64. A.G. Vitreschak, D.A. Rodionov, A.A. Mironov, M.S. Gelfand, Trends Genet. **20**, 44 (2004)
65. W.C. Winkler, R.R. Breaker, Chembiochem **4**, 1024 (2003)
66. S. Brantl, Trends Microbiol. **12**, 473 (2004)
67. E. Nudler, A.S. Mironov, Trends Biochem. Sci. **29**, 11 (2004)
68. J.E. Barrick, K.A. Corbino, W.C. Winkler, A. Nahvi, M. Mandal, J. Collins, M. Lee, A. Roth, N. Sudarsan, I. Jona, J.K. Wickiser, R.R. Breaker, Proc. Natl. Acad. Sci. USA **101**, 6421 (2004)
69. E.A. Lesnik, G.B. Fogel, D. Weekes, T.J. Henderson, H.B. Levene, R. Sampath, D.J. Ecker, Biosystems **80**, 145 (2005)
70. R.R. Breaker, Curr. Opin. Biotechnol. **13**, 31 (2002)
71. S.K. Silverman, RNA **9**, 377 (2003)
72. M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, P.F. Stadler, J. Phys. A: Math. Gen. **37**, 4731 (2004)
73. M. Eigen, Naturwissenschaften **58**, 465 (1971)
74. M. Eigen, P. Schuster, Naturwissenschaften **64**, 541 (1977)
75. M. Eigen, P. Schuster, Naturwissenschaften **65**, 7 (1978)
76. J. Swetina, P. Schuster, Biophys. Chem. **16**, 329 (1982)
77. M. Eigen, J. McCaskill, P. Schuster, Adv. Chem. Phys. **75**, 149 (1989)
78. P. Jagers, *Branching Processes with Biological Applications* (Wiley, London, 1975)
79. P.A.P. Moran, *The Statistical Processes of Evolutionary Theory* (Clarendon, Oxford, UK, 1962)
80. L. Demetrius, P. Schuster, K. Sigmund, Bull. Math. Biol. **47**, 239 (1985)
81. W. Fontana, P. Schuster, Biophys. Chem. **26**, 123 (1987)
82. W. Fontana, W. Schnabl, P. Schuster, Phys. Rev. A **40**, 3301 (1989)

83. W. Fontana, P. Schuster, Science **280**, 1451 (1998)
84. D.T. Gillespie, J. Comput. Phys. **22**, 403 (1976)
85. D.T. Gillespie, J. Phys. Chem. **81**, 2340 (1977)
86. P. Schuster, in *Molecular Insight into the Evolution of Phenotypes*. ed. by J.P. Crutchfield, P. Schuster. Evolutionary Dynamics: Exploring the Interplay of Accident, Selection, Neutrality, and Function (Oxford University Press, New York, 2003) pp. 163–215
87. B.R.M. Stadler, P.F. Stadler, G.P. Wagner, W. Fontana, J. Theor. Biol. **213**, 241 (2001)
88. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, UK, 1983)
89. M.A. Huynen, P.F. Stadler, W. Fontana, Proc. Natl. Acad. Sci. USA **93**, 397 (1996)
90. K. Grünberger, U. Langhammer, A. Wernitznig, P. Schuster, *RNA evolution in Silico* (Technical Report, Institut für Theoretische Chemie, Universität Wien, 2005)
91. D.T. Gillespie, J. Stat. Phys. **16**, 311 (1977)
92. D.P. Bartel, C.Z. Chen, Nat. Genet. **5**, 396 (2004)
93. O. Hobert, Trends Biochem. Sci. **29**, 462 (2004)
94. J.S. Mattick, Bioessays **25**, 930 (2003)
95. J.S. Mattick, Nat. Genet. **5**, 316 (2004)
96. M. Szymański, M.Z. Barciszewska, M. Żywicki, J. Barciszewski, J. Appl. Genet. **44**, 1 (2003)
97. S.R. Eddy, Nat. Genet. **2**, 919 (2001)
98. M. Schöninger, A. von Haeseler, J. Mol. Evol. **49**, 691 (1999)
99. B. Knudsen, J.J. Hein, Bioinformatics **15**, 446 (1999)
100. N.J. Savill, D.C. Hoyle, P.G. Higgs, Genetics **157**, 399 (2001)
101. J. Otsuka, N. Sugaya, J. Theor. Biol. **222**, 447 (2003)
102. H. Jow, C. Hudelot, M. Rattray, P.G. Higgs, Mol. Biol. Evol. **19**, 1591 (2002)
103. C. Hudelot, V. Gowri-Shankar, H. Jow, M. Rattray, P.G. Higgs, Mol. Phylogenet. Evol. **28**, 241 (2003)
104. E. Rivas, R.J. Klein, T.A. Jones, S.R. Eddy, Curr. Biol. **11**, 1369 (2001)
105. S. Washietl, I.L. Hofacker, P.F. Stadler, Proc. Natl. Acad. Sci. USA **102**, 2454 (2005)
106. A.F. Bompfünewerer, C. Flamm, C. Fried, G. Fritzsch, I.L. Hofacker, J. Lehmann, K. Missal, A. Mosig, B. Müller, S.J. Prohaska, B.M.R. Stadler, P.F. Stadler, A. Tanzer, S. Washietl, C. Witwer, Theor. Biosci. **123**, 301 (2005)
107. S. Carranza, J. Baguñà, M. Riutort, J. Mol. Evol. **49**, 250 (1999)
108. A.P. Rooney, Mol. Biol. Evol. **21**, 1704 (2004)
109. M.J. Telford, P.W.H. Holland, J. Mol. Evol. **44**, 135 (1997)
110. F.E. Frenkel, M.B. Chaley, E.V. Korotkov, K.G. Skryabin, Gene **335**, 57 (2004)
111. M. Eigen, B.F. Lindemann, M. Tietze, R. Winkler-Oswatitsch, A.W.M. Dress, A. von Haeseler, Science **244**, 673 (1989)
112. M. Eigen, R. Winkler-Oswatitsch, Naturwissenschaften **68**, 282 (1981)
113. S. Rodin, S. Ohno, A. Rodin, Proc. Natl. Acad. Sci. USA **90**, 4723 (1993)
114. M. Di Giulio, J. Theor. Biol. **226**, 89 (2004)
115. M.P. Terns, R.M. Terns, Gene Expr. **10**, 17 (2002)

116. D. Lafontaine, D. Tollervey, Trends Biochem. Sci. **23**, 383 (2002)
117. Y. Lee, K. Jeon, J.T. Lee, S. Kim, V.N. Kim, EMBO J. **21**, 4663 (2002)
118. J. Hertel, M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I.L. Hofacker, P.F Stadler, The Students of Bioinformatics Computer Labs 2004 and 2005. BMC Genomics **7**, 25 (2006)
119. A. Tanzer, P.F. Stadler, J. Mol. Biol. **339**, 327 (2004)