*RNA* (2000), *6*:325–338. Cambridge University Press. Printed in the USA.


# RNA folding at elementary step resolution

**CHRISTOPH FLAMM,[1] WALTER FONTANA,[2,3] IVO L. HOFACKER,[1] and PETER SCHUSTER[1]**

[1]Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, A-1090 Wien, Austria
[2]Santa Fe Institute, Santa Fe, New Mexico 87501 USA

**ABSTRACT**

**We study the stochastic folding kinetics of RNA sequences into secondary structures with a new algorithm based on the formation, dissociation, and the shifting of individual base pairs. We discuss folding mechanisms and the correlation between the barrier structure of the conformational landscape and the folding kinetics for a number of examples based on artificial and natural sequences, including the influence of base modification in tRNAs.**

**Keywords: conformational spaces; foldability; RNA folding kinetics; RNA secondary structure**

## INTRODUCTION

The conformational diversity of nucleic acids or proteins is delimited by the loose random coil and the compact native state that is frequently the most stable or minimum free energy (mfe) conformation. Let us call a specific interaction between two segments of the chain a "contact." A random coil then is best characterized by the absence of contacts, whereas the mfe conformation maximizes their energetic contributions. Several different types of contacts are found in three-dimensional structures. Their energetics is not well understood, which makes the modeling of RNA folding from random coils into full structures too ill-defined to be tackled at present.

Fortunately, for single-stranded nucleic acid molecules, the simpler coarse-grained notion of secondary structure is accessible to mathematical analysis and computation. To a theorist the secondary structure is the topology of binary contacts that arises from specific base pairing (Watson–Crick and GU; see Figure 1 and the next section). It does not refer to a two- or three-dimensional geometry cast in terms of distances. Secondary structure formation is driven by the stacking between contiguous base pairs. However, any formation of an energetically favorable double-stranded region implies the simultaneous formation of an energetically unfavorable loop. This frustrated energetics leads to a vast combinatorics of stack and loop arrangements spanning the conformational repertoire of an individual RNA sequence at the secondary structure level.

The secondary structure is not only an abstract tool convenient for theorists. It also corresponds to an actual state that provides a geometric, kinetic, and thermodynamic scaffold for tertiary structure formation, and constitutes an intermediate on the folding path from random coil to full structure. With rising temperature, tertiary contacts usually disappear first and double helices melt later (Banerjee et al., 1993). The free energy of secondary structure formation accounts for a large fraction of the free energy of full structure formation. These roles put the secondary structure in correspondence with functional properties of the tertiary structure. Consequently, selection pressures become observable at the secondary structure level in terms of evolutionarily conserved base pairs (Gutell, 1993). Moreover, insights into the process of secondary structure formation can be extended to several types of tertiary contacts with roughly conserved local geometries, such as non-Watson–Crick base pairs, base triplets and quartets, or end-on-end stacking of double helices.

To provide a frame for our kinetic treatment of RNA folding, we give a short account of the formal issues surrounding conformational spaces, folding trajectories, and folding paths for RNA secondary structures. We then introduce the kinetic folding algorithm as a stochastic process in the conformation space of a sequence, and discuss applications to several selected problems that cannot be studied adequately with the thermodynamic approach alone.

Reprint request to: Christoph Flamm, Institut für Theoretische Chemie und Molekulare Strukturbiologie, Währingerstrasse 17, A-1090 Wien, Austria; e-mail: xtof@tbi.univie.ac.at.

[3]Present address: Institute for Advanced Study, Program in Theoretical Biology, 310 Olden Lane, Princeton, New Jersey 08540, USA.

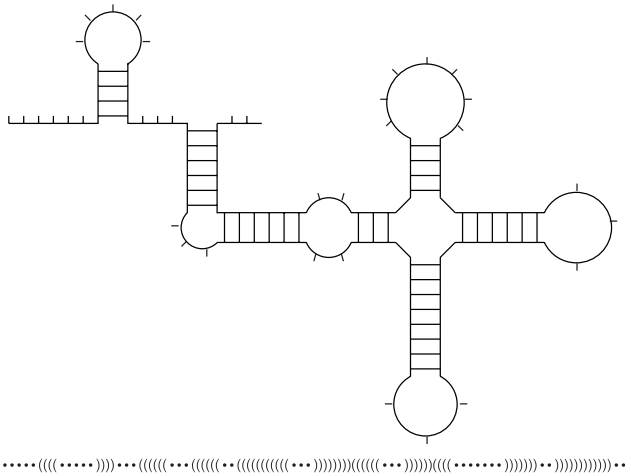

**FIGURE 1.** An RNA secondary structure graph. Unpaired positions are marked by ticks. They occur in loops, where they are enclosed by base pairs, and in free ends or links between independent structure modules, where they are called "external." The string of balanced parentheses below is an equivalent depiction of the secondary structure graph shown above: two matching parentheses represent a base pair between the corresponding positions (a left [right] parenthesis pairs downstream [upstream] along the sequence), and a dot stands for an unpaired base.

## CONFORMATION SPACES AND FOLDING PATHWAYS

We denote an RNA sequence by a string $I = (x_1 x_2 \dots x_n)$ of $n$ positions over the conventional nucleotide alphabet, $x_i \in \mathcal{A} = \{\mathrm{A,U,G,C}\}$. (If we need to distinguish between sequences $I_k$, we use superscripts, as in $x_i^{(k)}$, to denote the $i$th nucleotide of sequence $I_k$.) The bases $x_1$ and $x_n$ are the nucleotides at the 5′ end and the 3′ end of the sequence, respectively. A secondary structure $S$ can be conveniently discretized as a graph representing a pattern of contacts or base pairs (Fig. 1). The nodes of the graph correspond to bases $x_i$ at positions $i = 1, \dots, n$. The set of edges consists of two disjoint subsets. One subset is common to all secondary structure graphs, and represents the covalent backbone connecting the nodes $i$ and $i + 1$ for $i = 1, \dots, n - 1$. The other comprises the base pairs, denoted by $i \cdot j$, and constitutes the secondary structure proper. The base pairs form a set $\Pi$ with $j \neq \{i - 1, i, i + 1\}$ that must satisfy two conditions: (1) every edge in $\Pi$ connects a node to at most one other node, and (2) if both $i \cdot j$ and $k \cdot l$ are in $\Pi$, then $i < k < j$ implies $i < l < j$. Failure to meet condition (2) results in pseudoknots that are considered tertiary contacts.

Secondary structure graphs are formal combinatorial objects amenable to mathematical treatment. Of particular interest are secondary structures satisfying some extremal condition, such as minimizing the free energy (mfe structures). They can be computed by dynamic programming (Waterman & Smith, 1978; Nussinov & Jacobson, 1980; Zuker & Stiegler, 1981). We have recently extended the standard RNA thermodynamic folding algorithm to compute all conformations within some energy range above the mfe (Wuchty et al., 1999). This enables us to analyze the low-energy portion of the conformational landscape of individual sequences, and to put it in correspondence with their kinetic folding behavior derived from a computational model that we present below.

A sequence $I$ is called compatible with a secondary structure $S$, whenever positions that pair in the specification of $S$ ($i \cdot j \in \Pi(S)$) are occupied by nucleotides that can actually pair with one another:

$$i \cdot j \to [x_i\, x_j] \in \mathcal{B} = \{\mathrm{AU,UA,UG,GU,GC,CG}\},\ \forall\ i \cdot j \in \Pi(S).$$

A sequence $I$ specifies a set of structures with which it is compatible,

$$\mathcal{S}(I) = \{S_0, S_1, \dots, S_m\} \cup \{\mathbf{0}\},$$

where $S_0$ is the mfe conformation and $S_1 \dots S_m$ are suboptimal conformations ordered with respect to energy. **0** denotes the open chain conformation. The set $\mathcal{S}(I)$ and a metric still to be defined form the conformational space of the sequence $I$.

Secondary structure formation is described by a succession of elementary steps chosen according to some distribution from a pool of acceptable moves in conformation space. The result is a trajectory $\mathcal{T}(I)$ consisting of a time-ordered series of structures in $\mathcal{S}(I)$. A folding trajectory is defined as starting with the open chain **0** and ending with the mfe structure $S_0$:

$$\mathcal{T}(I) = \{\mathbf{0}, S(1), \dots, S(t-1), S(t), S(t+1), \dots, S_0;$$

$$S(j) \in \mathcal{S}(I)\}. \qquad (1)$$

Because the conformational space of secondary structures is always finite, every trajectory will reach $S_0$ after sufficiently long time. We define the "folding time" $\tau$ (associated with the trajectory) to be the first passage time, that is, the time elapsed until $S_0$ is encountered first. The folding time is a stochastic variable with a probability distribution $P_\tau(t) = \mathrm{Prob}\{\tau \leq t\}$. In practice $\tau$ may well be too long for a computer simulation. We therefore distinguish between trajectories that actually attain the ground-state structure within the limits of a simulation from those that are trapped in a thermodynamically suboptimal conformation (a long-lived metastable state).

Folding trajectories may contain loops, in the sense that certain suboptimal conformations are visited more than once: $S(t) = S(t + \ell)$, where $\ell$ is the length of the loop. We call a trajectory from which all loops have been eliminated a "folding path." Clearly, no structure appears twice in a folding path.

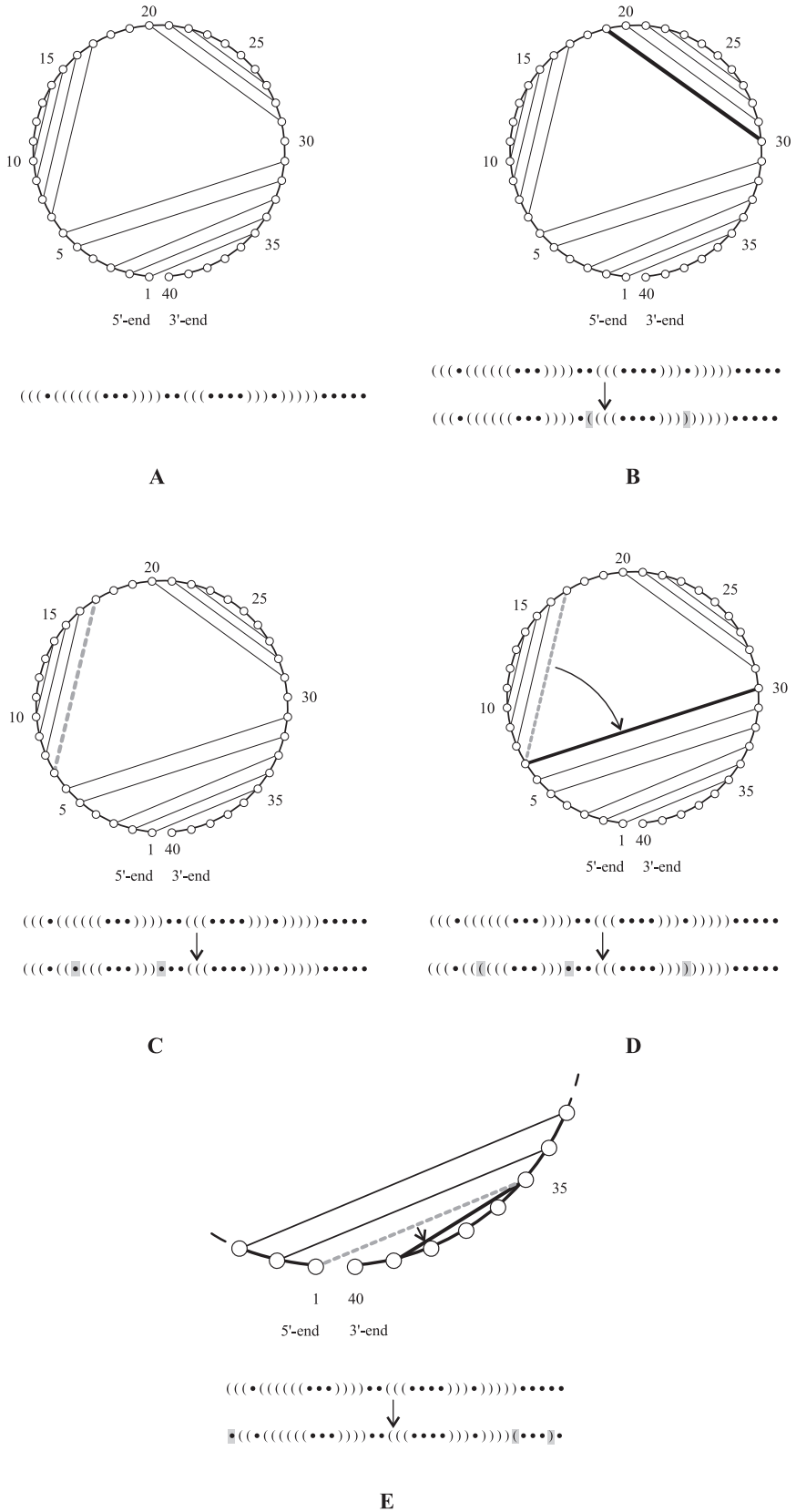


**FIGURE 2.** Elementary moves in the RNA folding algorithm. Secondary structures are shown in circle and parenthesis representation. The structure (**A**) is changed by the formation (**B**) or the removal (**C**) of a base pair. A shift move of a base pair can occur either within the structure (**D**) or by flipping over the gap between the 3′ and the 5′ end (**E**). The base pair after a move is shown in bold, the one being changed is shown by a gray dotted line. For details see text.

## MOVE SETS AND THE FOLDING ALGORITHM

The set of structures $\mathcal{S}(I)$ compatible with a sequence $I$ is organized into a space by defining a relation that specifies whether two structures are accessible from each other by an elementary event or "move" that is physically reasonable. The simplest conceivable modification of a secondary structure is the removal of a single base pair contact or its addition in compliance with the no-pseudoknot restriction. This is most easily visualized with the circle representation of RNA structures in Figure 2 (moves **B** and **C**). A new base pair adds a chord to the diagram that is not allowed to intersect any existing chord. The two moves **B** and **C** are a complete set in the sense that they are sufficient for constructing a path connecting any pair of structures. The metric induced by this simple move set on the conformation space is known as the "base pair distance."

Although sufficient, this simple move set fails to capture "defect diffusion," a mechanism believed to be important in the dynamics of RNA folding (Pörschke, 1974). Since helices nucleate statistically along the RNA chain, intermediate formation of helices with incomplete base pairing is expected to occur. Such mismatched regions can anneal by a fast chain sliding process. For instance, the loop position of a bulge in a helix may move (if the nucleotide composition permits) by a rapid process of base pair formation and dissociation (Fig. 3, top). Defect diffusion seems to be some orders of magnitude faster than zippering (Pörschke, 1974). If a bulge loop forms at one end of a double-stranded region and disappears at the other, the opposing strands shift by the size of the loop (Fig. 3, top).

To facilitate chain sliding, the simple move set must be extended by a further event that we call a "shift." As shown in Figure 2, the shift is a combination of a base pair removal and a base pair addition during which one position remains invariant. In the circle plot, the shift appears as the displacement of a chord with one end fixed, maintaining compliance with the noncrossing rule. It certainly is physically possible as an elementary event, and our results suggest that its actual occurrence is quite likely. We distinguish two cases of shifts, depending on whether the pairing orientation (upstream versus downstream) of the constant position changes (**D** and **E** in Fig. 2). In case **E** the change in pairing orientation is caused by crossing the 5′/3′ gap. The extended set of three moves also induces a metric on the space of conformation, but it is much harder to cast it into a simple expression.

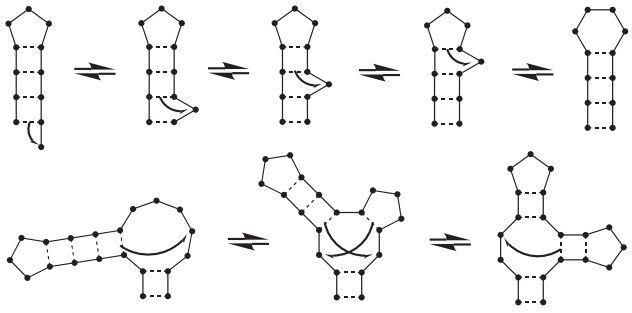

**FIGURE 3.** Defect diffusion and helix morphing. The shift move (Fig. 2, **D**) facilitates the diffusion of loops through double-stranded regions, as well as the interconversion of helices.

Note that the extended move set also facilitates the morphing of double-stranded regions into one another, particularly if the two regions are located inside a multiloop (Fig. 3, bottom). Such a process would be energetically unfavorable with the simple move set.

Given our microscopic (extended) move set, we model RNA folding as a Markov process in conformation space. The time-dependent random variable $\mathcal{X}(t)$ describes individual folding trajectories, for example, $\mathcal{T}(I)$ in equation (1). We understand it as the index $j$ of the conformation observed at time $t$:

$$\mathcal{X}(t) = i \Rightarrow S(t) = S_i;\ i \in \{0,1,\ldots,m+1\},$$

where $m+1$ is the index of the open chain **0**. The probability of observing conformation $S_i$ at time $t$ as the secondary structure of $I$ is given by $P_i(t) = \mathrm{Prob}\{\mathcal{X}(t) = i\}$. Following conventional stochastic kinetics of chemical reactions (Gardiner, 1985), folding is described by the master equation

$$\frac{dP_i(t)}{dt} = \sum_{j=0}^{m+1} \left(P_j(t)k_{ji} - P_i(t)k_{ij}\right). \qquad (2)$$

All elements of the transition matrix $\mathbf{k} = \{k_{ij}\}$ that do not correspond to single moves of the chosen set are assumed to be zero. The non-zero transition elements have to be consistent with the free energies differences of the conformations involved. They must satisfy

$$\frac{k_{ij}}{k_{ji}} = \exp(-\Delta G_{ij}/RT) = \exp\left(-(\Delta G_j^0 - \Delta G_i^0)/RT\right),$$

where $\Delta G_i^0$ and $\Delta G_j^0$ are the free energies as obtained from folding the sequence $I$ into the conformations $S_i$ and $S_j$, respectively. We tried two definitions for the individual transition frequencies: (1) the Metropolis rule (Metropolis et al., 1953)

$$k_{ij} = \begin{cases} \exp(-\Delta G_{ij}/RT) & \text{if } \Delta G_j^0 > \Delta G_i^0, \\ 1 & \text{if } \Delta G_j^0 < \Delta G_i^0, \end{cases}$$

and (2) a symmetric rule introduced by Kawasaki (1996)

$$k_{ij} = \exp(-\Delta G_{ij}/2RT).$$

Computer simulations with both rules showed that the second assumption leads to substantial improvement

in folding performance without changing the character of folding paths. In the remainder of this paper we use the Kawasaki dynamics.

Previous attempts at modeling the RNA folding process begin with generating a list of possible helices, and mainly differ in the criteria used to decide which helix to incorporate (or destroy) next (Breton et al., 1997; Galzitskaya & Finkelstein, 1996; Gultyaev et al., 1995; Mironov & Lebedev, 1993; Schmitz & Steger, 1996; Suvernev & Frantsuzov, 1995). The physical relevance of such moves seems debatable, because they cause large structural changes per time step. This makes them inadequate for resolving folding trajectories. (The concept itself might even lose its physical meaning.) Moreover, rather *ad hoc* assumptions about the overall rates of helix formation and disruption have to be made.

Transition probabilities defined by means of a move set based on individual base pairs (rather than entire stacks) are sufficiently flexible to allow for diverse pathways. For example, the formation of a single hairpin exhibits a variety of intermediate energies depending on the actual trajectory (an illustration is given in Table 1). The two folding paths of Table 1 build the double helix in the same contiguous fashion, but along opposite directions, one starting from the innermost base pair outwards and the second proceeding from the outermost pair inwards. The free activation energies are 3.6 and 4.82 kcal/mol, respectively.

Computing the transition matrix using only free energies of the involved conformations is less rigorous than a treatment based on a stochastic theory of the activated complex (Jacob et al., 1997a, 1997b), but makes it easy to take into account new regularities of RNA structure as they are discovered. It is straightforward to extend the folding analysis to include tertiary interactions for which sufficient experimental data become available. Examples are H-type pseudoknots, coaxial continuation of stacks, extension of double helices by non-Watson–Crick base pairs (commonly purine–purine pairings), U-U pairs in interior loops, and base triplets.

**TABLE 1**. The cumulative free energies along two folding pathways of the model sequence $A_6C_6U_6$. Differences in the energy values are caused by the size dependence of the loop energies and by the energy contributions of dangling ends.

| Base pair | Free energy kcal/mol* | Base pair | Free energy kcal/mol* |
|---|---|---|---|
| — | **0.0** | — | **0.0** |
| 6–13 | 3.6 | 1–18 | 4.82 |
| 5–14 | 2.7 | 2–17 | 3.78 |
| 4–15 | 1.8 | 3–16 | 2.71 |
| 3–16 | 0.9 | 4–15 | 1.61 |
| 2–17 | 0.0 | 5–14 | 0.5 |
| 1–18 | **–0.5** | 6–13 | **–0.5** |

*Bold numbers indicate the free energies of the initial and final conformations.

Despite the relative simplicity of the master equation (2), analytic solutions are available only for further drastic restrictions on allowed transitions and equal values of their probabilities ($k_{j,j+1} = k_{\rightarrow}$, $k_{j,j-1} = k_{\leftarrow}$). Here we rely instead on numerical simulations to study the stochastic process as defined above. To this end we use a variant of a Monte Carlo algorithm developed in the 1970s by Gillespie (1976, 1977) to study stochastic kinetics in chemical reaction networks. Gillespie's method is based on the same assumption as the derivation of the master equation: individual elementary steps are uncorrelated and the occurrence of an event follows a Poisson process on a proper time scale. Probability distributions, expectation values, variances, and other ensemble properties are obtained through sampling sufficiently many trajectories with identical initial conditions. The computer program we implemented is freely available upon request from Christoph Flamm.

## APPLICATIONS TO SELECTED PROBLEMS

Five problems are chosen to illustrate our kinetic RNA folding scheme. Three molecules are constructed on paper and two are naturally occurring examples, a tRNA and SV-11. The latter is a small variant RNA found in the Q$\beta$ replication assay. They illustrate different aspects of folding and also demonstrate that most of the issues typically arising in the context of long natural sequences appear at much shorter chain lengths as well.

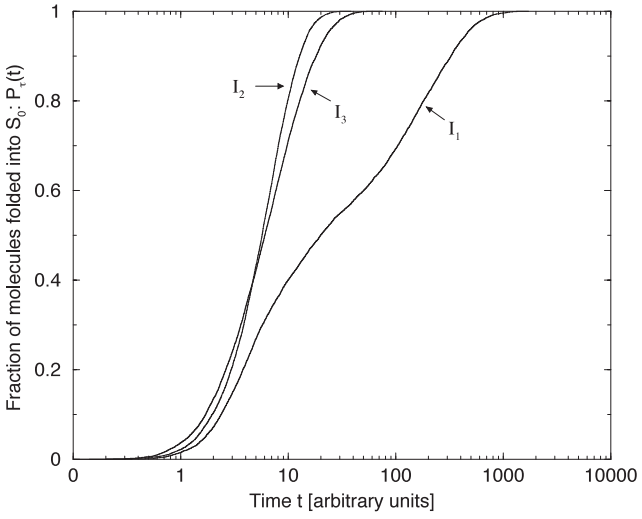


**FIGURE 4.** Distribution of folding times. The graph shows the kinetics of three sequences folding into a small hairpin. (For a definition of the structure and the sequences see Fig. 5.) The fraction of folding trajectories that reached the mfe structure at times $\tau \leq t$ is plotted on a logarithmic time scale. Sequence $I_1$ is an inefficient folder; approximately 50% of the folding trajectories lead to the mfe structure on a direct route, whereas the rest passes first through a local minimum. $I_2$ and $I_3$ fold efficiently.

### A small hairpin loop and ground state degeneracy

In our first example we consider the structure $S_0$ = [. . ( ( ( ( . . . . ) ) ) ) .] consisting of a tetraloop closed by a stack of four base pairs. The stack has two free ends of lengths 2 and 1. A random sequence $I_1$ = (ACUGAUCGUAGUCAC) with $S_0$ as the minimum free energy structure is obtained by inverse folding (Hofacker et al., 1994). The folding behavior is characterized by the distribution of folding times ($\tau_f$) in Figure 4. We easily recognize two folding mechanisms, a fast and a slow one, of almost equal probability of occurrence. In the slow regime the mfe conformation is reached only after the trajectories have first spent some time in one or more local minima.

To understand the folding behavior in more detail, we compute the "barrier tree" of the conformational landscape. The leaves of the tree are the local minima (with respect to energy) of the landscape. The barrier state connecting two local minima $S_i$ and $S_j$ is the minimum of the maxima (lowest saddle point) along all paths between $S_i$ and $S_j$ (Fig. 5). The barrier tree provides a picture of which local minima are aggregated into basins and how these basins are hierarchically linked with one another. Stated in terms of a flooding metaphor, if the energy abscissa were to measure altitude and the landscape was flooded up to a given height, a vertical

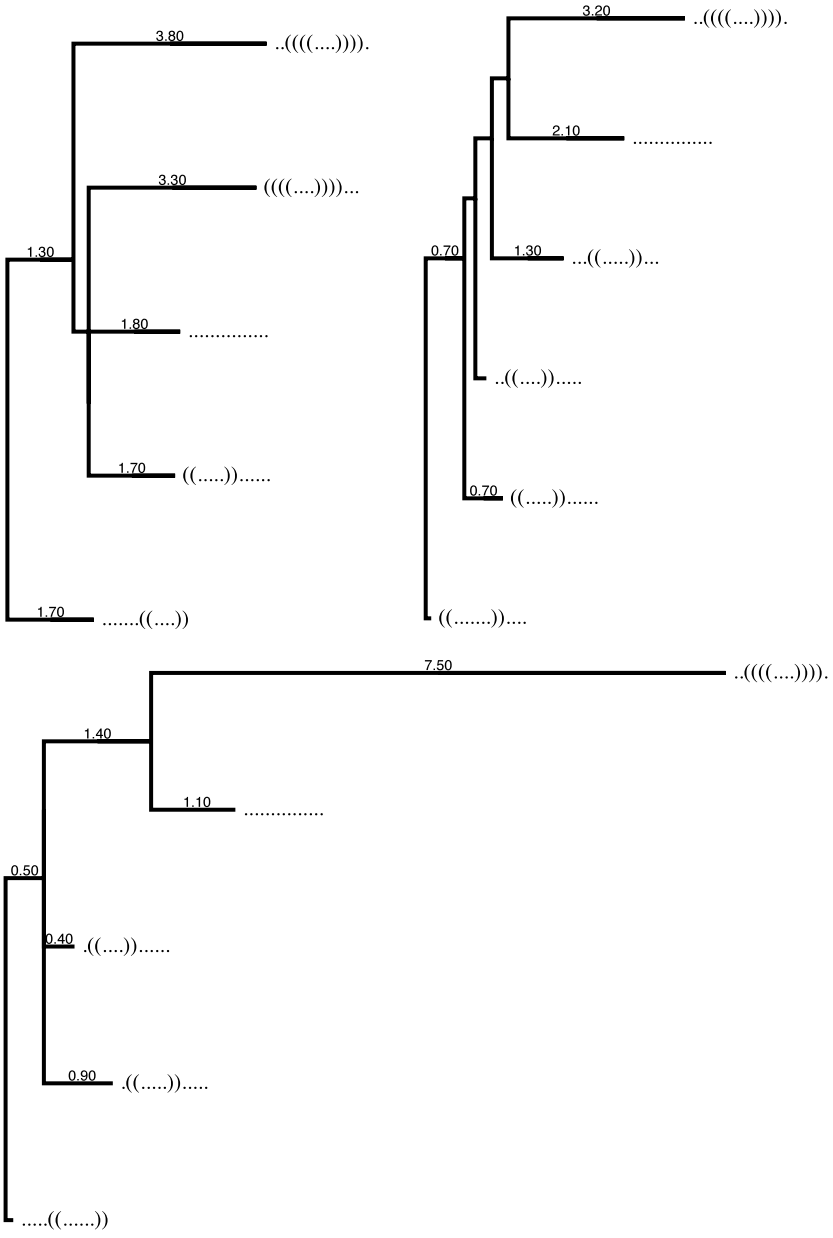


**FIGURE 5.** Barrier tree of the conformational landscape. The trees for three different sequences folding into the same mfe structure, $S$ = [. . ( ( ( ( . . . . ) ) ) ) .] are shown. Numbers labeling the branches are free energy differences in kilocalories/mole. The sequence $I_1$ = (ACUGAUCGUAGUCAC) folds inefficiently (upper left). The open chain and the mfe structure are located in different folding funnels. Evolutionary optimization of folding behavior generated another sequence, $I_2$ = (AUUGAG CAUAUUCAC) folding into the ground state $S$ with high efficiency. This sequence, however, has a degenerate ground state. A further sequence that folds well is $I_3$ = (CGGGCUAUUUAGCUG). It has a nondegenerate ground state. The barrier tree for the efficient folders (upper right diagram for $I_2$, and lower half for $I_3$) contain both the open chain and the mfe structure in the same funnel.

cut through the graph tells which states are joined under water (on the right), and which basins were to merge next as the water level is raised.

The folding path is almost immediate from the barrier tree. Its reconstruction starts from the open chain, **0** (which is a local minimum), and proceeds upward step by step until a saddle is reached that connects downward to the mfe conformation. In our particular example $I_1$, two suboptimal conformations can be accessed descending from the first saddle point located at a free energy of 1.80 kcal/mol. One of them, $S_1$ = [( ( ( ( . . . . ) ) ) ) . . .], lies only 0.2 kcal/mol above the mfe. This conformation acts as a trap, as it takes a relatively long time to undo it. The escape route passes through the saddle point [. . . . . ( . . . . ) . . . .] located at 2.10 kcal/mol from the open chain. From there, a downward path leads to the mfe conformation. In contrast, the fast mechanism corresponds to direct folding by visiting the two saddles one after the other.

We evolved a sequence with a better folding behavior, but the same ground-state structure, through mutation and replication in a simulated flow reactor that has been developed to study the optimization of RNA properties (Fontana & Schuster, 1987). The sequence $I_2$ = (AUUGAGCAUAUUCAC) was obtained from such an optimization process. It folds with high efficiency, as shown by the distribution of folding times (Fig. 4). The barrier tree (Fig. 5) provides an immediate explanation: both conformations, the open chain **0** and the mfe structure $S_0$, are in the same basin or folding funnel. The folding path reaches the target through a single saddle point with no traps in between. This also accounts for the narrow distribution of folding times.

The sequence $I_2$ has a degenerate ground state: $S_1$ = [. . ( ( ( . . . . . ) ) ) .] has the same free energy, $-1.1$ kcal/mol, as $S_0$. This is the simplest case of an ensemble of mfe structures. The folding algorithm only determines the first passage time from the open chain to some structure arbitrarily chosen from the mfe ensemble. The stationary probability distribution $\bar{P}_i$ within a set of $\ell$ mfe configurations ($i = 0, \ldots, \ell - 1$) is given by:

$$\bar{P}_i = \frac{e^{-\Delta G_{\text{mfe}}/kT}}{\sum_{k=0}^{m+1} e^{-\Delta G_k/kT}} = \frac{1}{\ell + \sum_{k=\ell}^{m+1} e^{-\Delta g_k/kT}},$$

where $\Delta g_k$ is the gap energy $\Delta G_k - \Delta G_{\text{mfe}}$. For $T = 0$ K this is the uniform distribution $\bar{P}_i = 1/\ell$.

It is not hard to find a sequence that folds efficiently into a nondegenerate ground state *S*. An example is given by the sequence $I_3$ = (CGGGCUAUUUAGCUG). We obtain again a barrier tree that contains **0** and the mfe structure *S* in the same folding funnel. The overall kinetic behavior of $I_3$ is very similar to that of $I_2$. It is worth pointing out, however, that the mfe of $I_3$ is much lower than the mfe for the other two sequences (because of the larger number of GC base pairs), and yet the folding times of $I_2$ are a little shorter and their distribution is narrower. Evidently, the folding behavior of a structure does not reflect its thermodynamic stability.

The main lesson of this simple example comes from comparing the folding behavior with the barrier structure of the conformational landscape. Folding efficiency seems to be a consequence of the multiplicity of folding paths, and does not depend on the minimum free energy or the energy gap between mfe and the first suboptimal conformation. The number of conformations representing local minima of the free energy surface (Fig. 5) is not particularly useful for predicting the folding efficiency. What actually matters is the number of saddle points at which a folding trajectory can split into paths leading to basins that do not contain the ground state. A folding mechanism that passes through a single saddle point cannot bifurcate and yields the fastest kinetics.

### Direct folding and escape pathway

The second example deals with the escape path from a conformational trap. The sequence *I* = (GGGAUUU CUCGCUAUUCCAGUGGGA) forms the mfe structure $S_0$ = [. . . . . ( ( ( ( ( ( ( . . . . . ) ) ) ) ) ) )] and a lowest suboptimal structure, $S_1$ = [( ( ( . . . . ) ) ) . . . . . ( ( ( . . . . ) ) )] with almost the same free energy. Figure 6 shows the sequence of structures on an escape path leading from $S_1$ to $S_0$, as well as that route's free energy profile. The profile illustrates the effect of the shift move: the path computed without the shift move passes two additional saddle points between conformations 7 and 9. The figure also shows the direct path from the open chain to $S_0$, which, after a first activation step, is a classic base pair zipper (similar to the example shown in Table 1). The folding behavior reflects the barrier tree of *I*'s conformational landscape (not shown): it contains two major branches leading to $S_0$ and $S_1$, as well as five minor branches.

The detection of distinct folding mechanisms (ensembles of related paths) is made easier by a modified probability density of folding times, the "folding characteristic":

$$\chi(t) = t \cdot \frac{d \log P_\tau(t)}{dt} = \frac{t}{P_\tau(t)} \cdot \frac{dP_\tau(t)}{dt}. \qquad (3)$$

The major humps in the folding characteristic of sequence *I* (Fig. 7) correspond to the two predominant folding paths discussed previously, the direct zipper (Fig. 6) and the alternative route visiting the $S_1$ trap. Details of the curve, such as the shoulder on the right flank of the faster folding hump, indicate the presence of minor mechanisms.

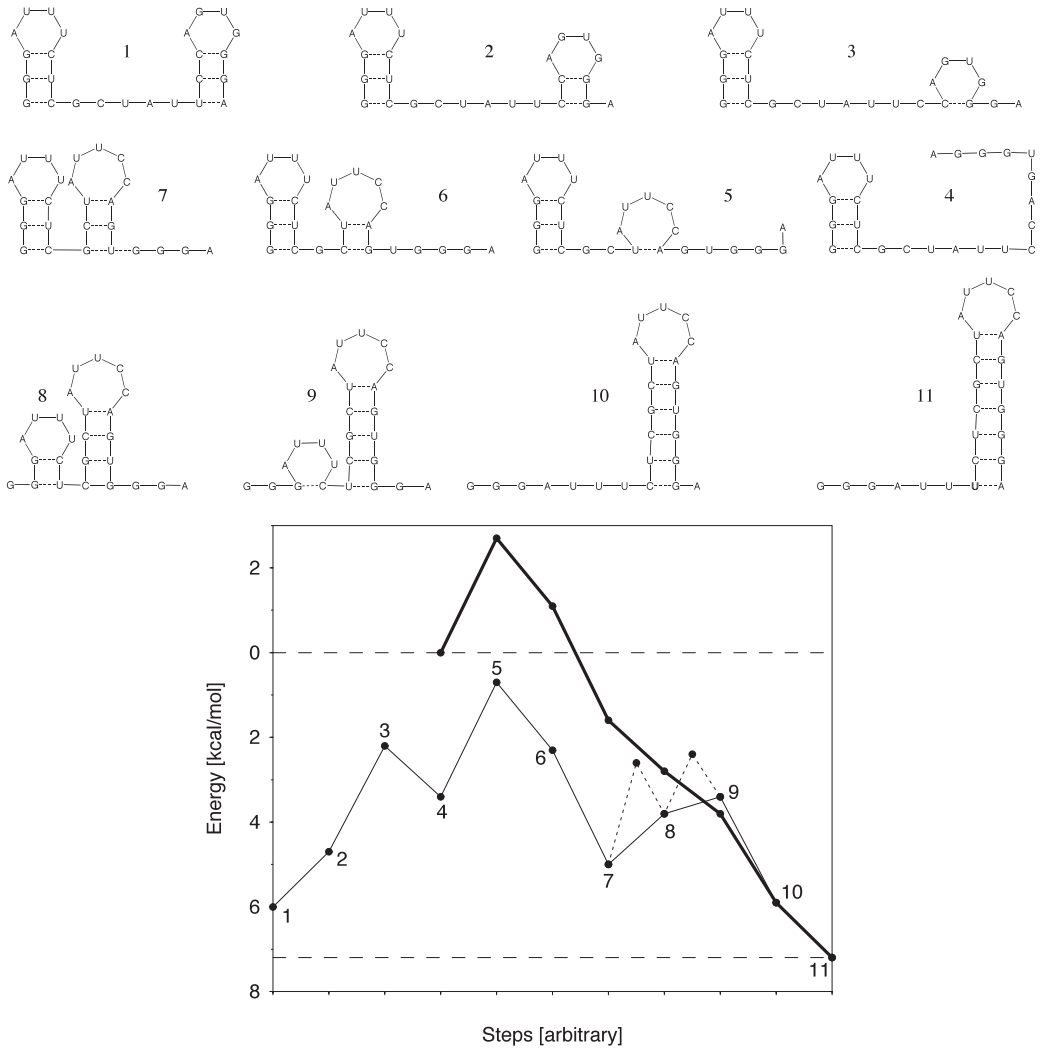


**FIGURE 6.** The escape from a conformational trap. The sequence of structures along an escape path from a folding trap is shown together with its free energy profile. Dashed lines indicate energy barriers in the absence of shift moves. The bold line corresponds to the fast folding path following a simple zipper from the open chain to $S_0$.

### A switching molecule

The third example is a molecule designed to have two almost equally stable, yet sufficiently distinct conformations of low energy. This construction tells us something about the role of nucleation centers in the folding process. The sequence

$$I = (\text{GGCCCCUUUGGGGGCCAGACCC}$$

$$\text{CUAAAGGGGUC}),$$

folds into two highly stable conformations, the mfe structure consisting of a long hairpin with 14 base pairs,

$$S_0 = [((((((((((((((.....))))))))))))))],$$

and a first suboptimal conformation with two hairpins of six base pairs each,

$$S_1 = [((((((....)))))).((((((....))))))].$$

The complete barrier structure of $I$'s landscape (Fig. 8) consists of two neatly separated folding funnels. The bottom of the $S_1$ basin is energetically sufficiently deep to prevent molecules that have fallen into it from refolding into the mfe structure within the time spans of our computer simulations. The conformation $S_1$ is a true long-lived metastable state.

In Figure 9 we plot the fraction of folded molecules as a function of time. The ratio of the conformations $S_0$ and $S_1$ is close to 1:2. This can be rationalized by observing that folding starts with the nucleation of a

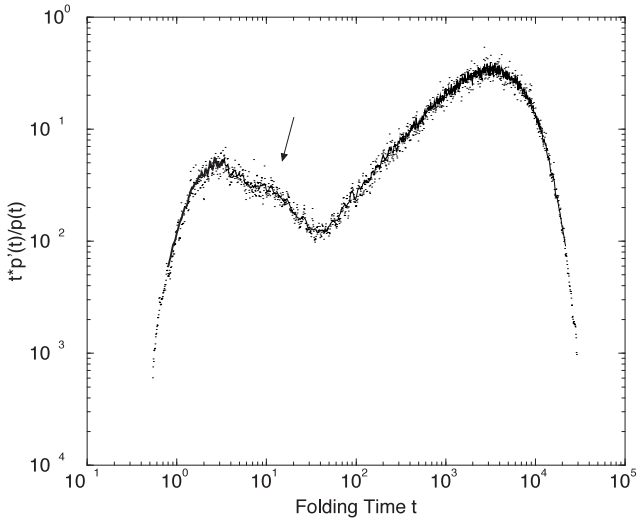


**FIGURE 7.** The folding characteristic of a small RNA. The folding characteristic $\chi(t)$ of the sequence $I$ = (GGGAUUUCUCGCUAUUC CAGUGGGA) is plotted on a logarithmic time scale. The points correspond to individual trajectories and show some scatter that has been smoothed by a moving average in the solid line. The curve shows two distinct humps corresponding to different folding paths (direct or via $S_1$). A less prominent pathway appears as a shoulder on the right flank of the first hump (arrow).

double helical region. The approximate 1:2 ratio arises because $S_1$ has two stacks, and hence two nucleation centers, while $S_0$ has only one. Considering similar cases we find that the number of nucleation centers determines the frequency of dominant conformations.

## Modified bases and tRNA folding kinetics

In this section we discuss how base modification influences the folding kinetics of tRNA molecules. Base modification is taken into account by excluding such bases from pairing, but not from single-base interactions, such as the stacking of unpaired bases onto adjacent double helices (terminal mismatches). The role of base modification on tRNA stability has been recently discussed (Wuchty et al., 1999) in terms of the free energy gap between the mfe structure and the first suboptimal configuration ($\Delta\varepsilon = \Delta G(S_1) - \Delta G(S_0)$). Because many of the modified bases are unable to form regular base pairs, several suboptimal conformations that would otherwise be present cannot be formed, thereby increasing the energy gap $\Delta\varepsilon$. When the unmodified sequence does not form the correct clover

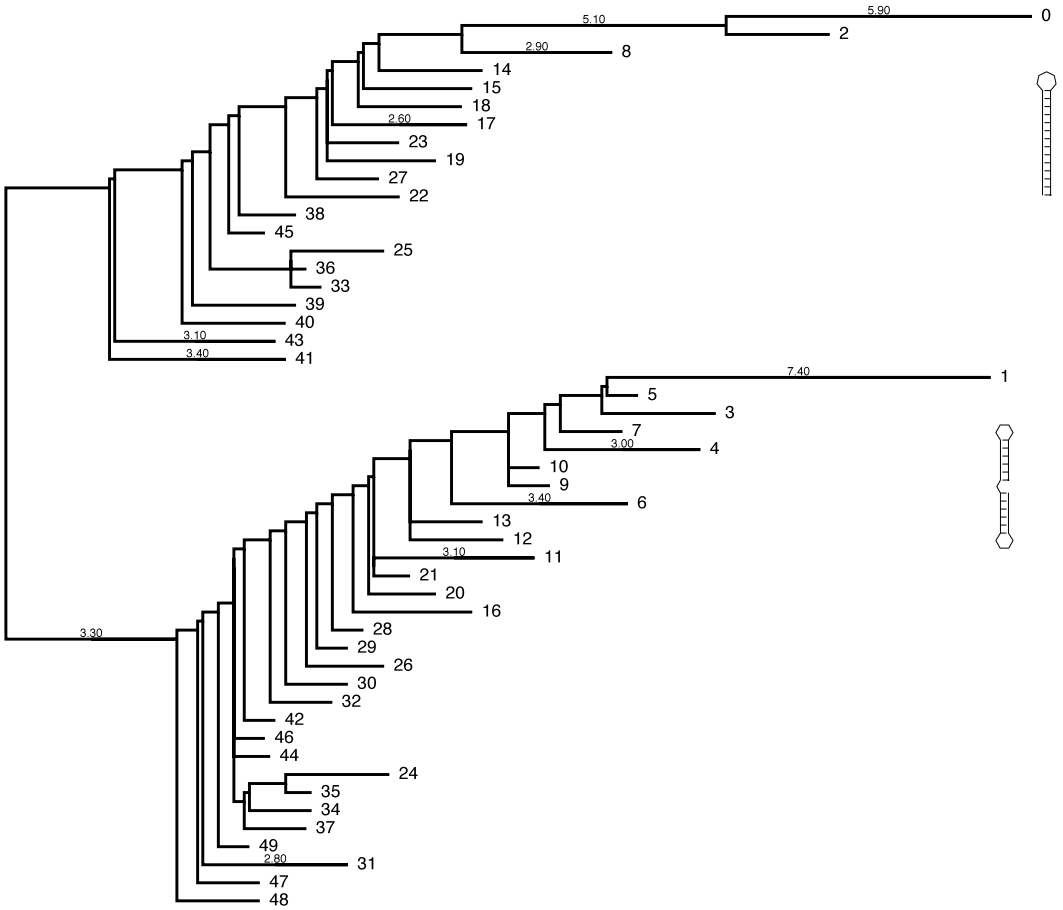


**FIGURE 8.** The barrier tree for a sequence with two dominant conformations. The conformational space is partitioned into two folding funnels, which enables an estimate of folding frequencies.

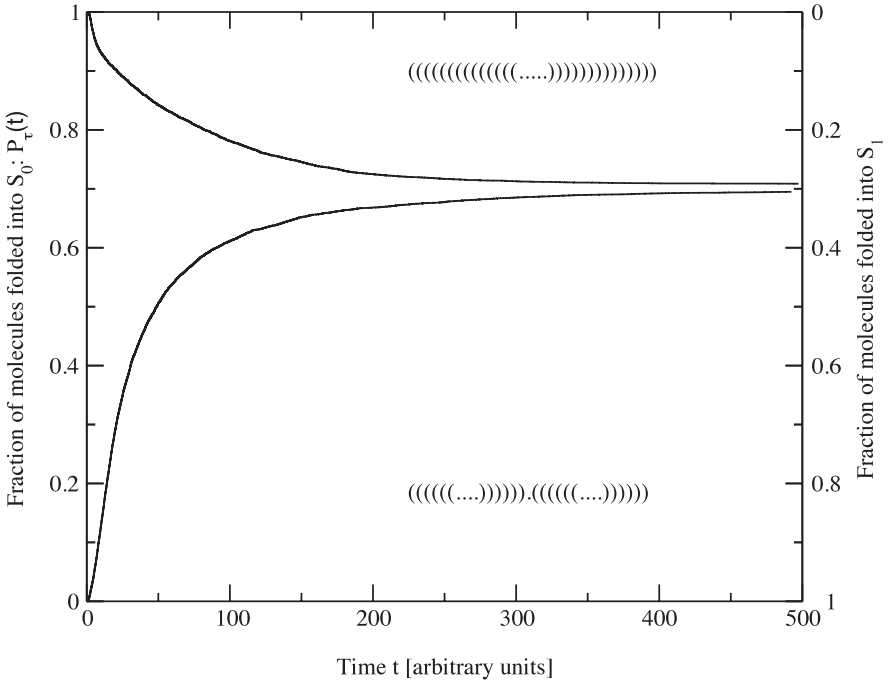


**FIGURE 9.** The fraction of folded molecules. The two curves show the abundance of conformations, $S_0$ (left ordinate), and $S_1$ (right ordinate, upside down). As time progresses, the two conformations of low free energy, $S_0$ and $S_1$, increase in frequency at the expense of all other conformations. The final ratio of $S_0/S_1$ is close to 1:2 in agreement with the number of nucleation centers in the two conformations.

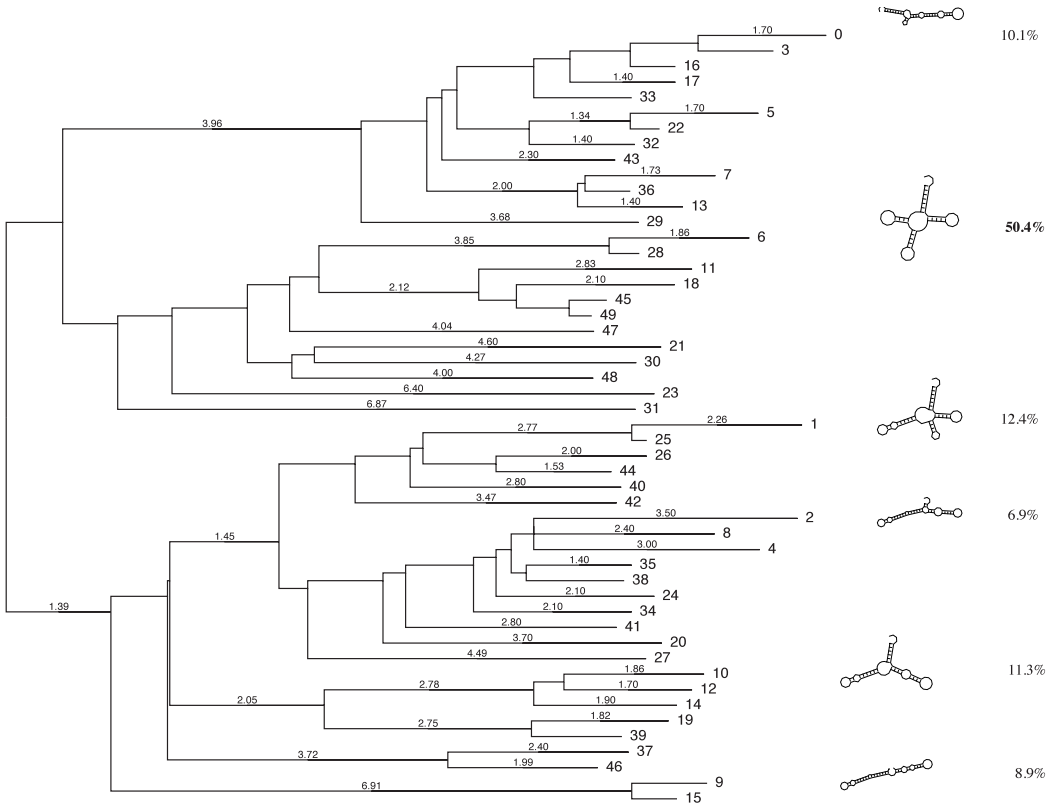


**FIGURE 10.** The barrier tree of $tRNA^{phe}$. The tree refers to the 50 lowest local minima (obtained with the suboptimal thermodynamic folding procedure (Wuchty et al., 1999)) of the unmodified sequence. Six folding funnels leading to the conformations $S_0$, $S_1$, $S_2$, $S_6$, $S_9$, and $S_{10}$ can be distinguished. These partition the conformation space into six basins in addition to a tiny folding funnel comprising the conformations $S_{37}$ and $S_{46}$. The mfe structure $S_0$ is not the naturally occurring conformation: the correct clover leaf appears in the tree as conformation $S_6$. The numbers at the right indicate the fraction of folding trajectories ending in the corresponding basins.

leaf structure, base modification can change the ground state. Here we study the impact of base modification on the barrier structure of the conformation space and on the kinetics of folding.

Figure 10 shows the barrier tree for the low-energy portion of the conformation space of the unmodified sequence. Six folding funnels can be distinguished that are dominated by the conformations $S_0$, $S_1$, $S_2$, $S_6$, $S_9$, and $S_{10}$. The correctly folded clover leaf is not the mfe structure. It appears as conformation $S_6$ with a free energy of about 1 kcal/mol above the ground state. Base modification changes the kinetic connectedness of conformations and turns the clover leaf structure into the ground state.

When comparing the computed distributions of folding times for the unmodified and the modified sequence (Fig. 11), it becomes apparent that base modification leads to a remarkable improvement of the folding behavior. Practically all trajectories lead to the clover leaf. The folding characteristic of the modified sequence (Fig. 11) consists of one hump, indicating a single folding mechanism. This is consistent with a recent analysis of experimental data (Thirumalai & Woodson, 1996) suggesting a direct pathway to the native state of tRNA$^{phe}$.

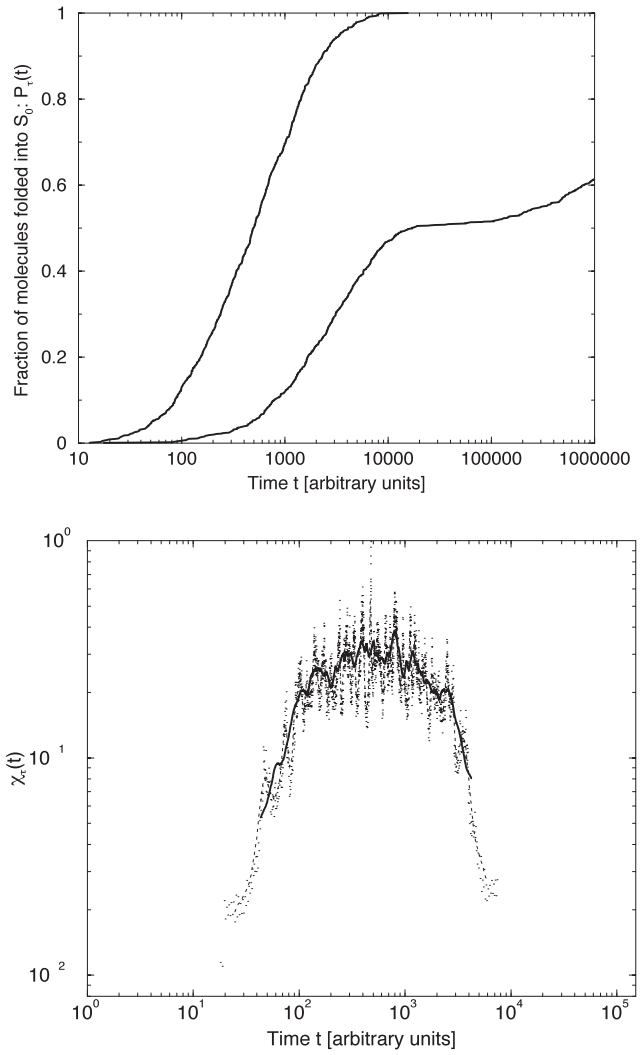


**FIGURE 11.** The folding kinetics of tRNA$^{phe}$. The upper part of the figure compares the distribution of folding times for the modified (upper curve) and the unmodified (lower curve) sequence of tRNA$^{phe}$. The folding characteristic of the sequence with modified bases (lower graph) indicates excellent folding behavior. As in Figure 7, the solid curve was obtained by a moving average over individual points.

It is worth pointing out that for the unmodified sequence the clover leaf is neither the structure with the lowest energy nor the one with the largest folding funnel (expressed in terms of numbers of local minima belonging to the basin), and yet it is formed directly in about 50% of the folding trajectories (see Fig. 10). All other stable conformations are reached by less than 12.5% or one-eighth of all folding simulations. As in the previous example, the high frequency of trajectories ending up in the clover leaf can be explained by the larger number of independent nucleation centers, as compared to the competing conformations.

The tRNA$^{phe}$ case confirms once more that there is no relation between thermodynamic well-definition of the ground state and the capacity to access it kinetically. We also designed special sequences with the correct clover leaf as their mfe structure and could not find a correlation between energy gap $\Delta\varepsilon$ and folding behavior.

### The Q$\beta$ variant SV-11

The Q$\beta$ variant SV-11 is an RNA sequence 110 nt long that was derived from the natural phage by means of serial transfer experiments. It provides an example of how dramatically the thermodynamic picture can differ from the kinetic one. The mfe structure is a long hairpin interrupted by five bulges and internal loops (Fig. 12). The base pair probabilities computed from the thermodynamic partition function (lower left box in Fig. 12) give the impression that there are no serious alternative structures to the ground state. However, the probabilities accumulated from kinetic folding paths yield a rather different dominant structure (Fig. 12), located 25 kcal/mol above the minimum free energy. The ground state is reached by only 16% of the folding trajectories. Figure 13 shows the probability with which local minima of a given energy are visited by a folding path. Most paths are trapped in a large basin with a fairly flat bottom consisting of many states that are structurally similar to the metastable state shown in Figure 12. Previous models (Gultyaev et al., 1995; Morgan & Higgs, 1996) either failed to predict the metastable conformation or reproduced it only when folding was done in conjunction with chain growth. Our results suggest that chain growth is not necessary to obtain this structure. The relevance of the metastable SV-11 conformation is to function as a template for replication by the Q$\beta$ replicase. The mfe hairpin is completely inactive in this respect.

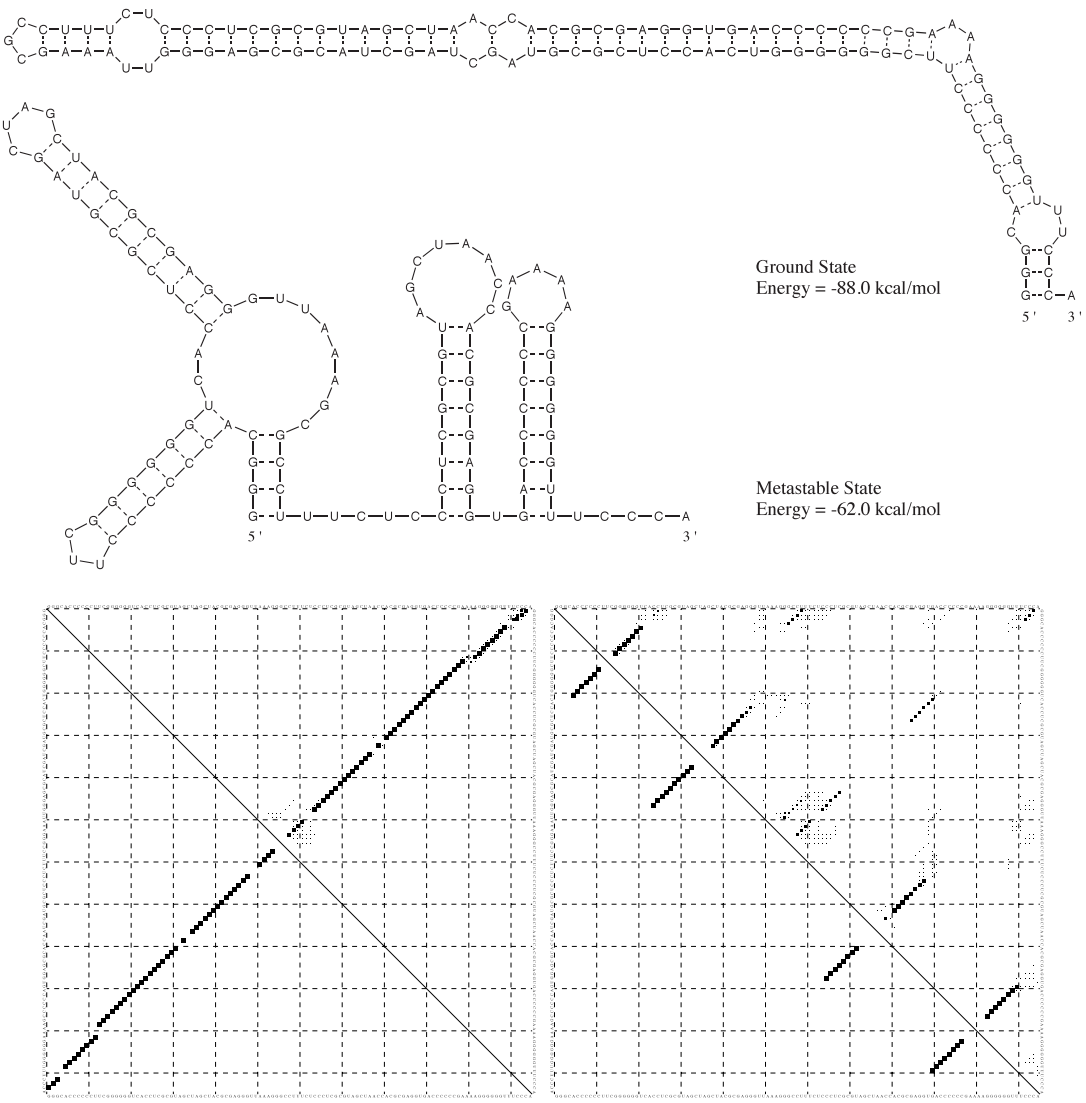


**FIGURE 12.** Structures and base pairing density plots for the mfe structure and the metastable conformation of the Q$\beta$ variant SV11. The secondary structures and their free energies are shown in the upper part. In the lower half we show the matrix of base pair probabilities as obtained from the thermodynamic partition function (McCaskill, 1990; Hofacker et al., 1994) (left) and from kinetic trajectories (right).

## CONCLUSIONS

The kinetic folding algorithm for RNA secondary structures presented in this article is, to our knowledge, the first successful attempt to model formation of polynucleotide structure at the level of single base-pairing events. A natural and obvious move set contains two elementary events, the making and breaking of individual base pairs. To accommodate the empirical observation of defect diffusion, we introduced a base-pair shift as an additional elementary event. Our definition of transition probabilities was motivated by the desire to construct a procedure that is not restricted to the present definition of RNA secondary structures. This led us to use quantities that are determined independently of such a definition, and free energies of conformations are ideally suited for this purpose. Using only free energy differences allows us, for example, to extend the kinetic procedure to any kind of tertiary interaction for which sufficient empirical knowledge becomes available. Free energy differences, however, determine only the ratio of the transitions probabilities ($k_{ij}/k_{ji}$), and any common factor would be consistent with it. The choice of Kawasaki's dynamics (Kawasaki, 1966) rather than the usual Metropolis assumption (Metropolis et al., 1953) for individual transition probabilities was motivated by the greater efficiency of the former. The Kawasaki assumption favors downhill steps with larger free energy gain, yielding shorter folding times. We stress, however, that we were unable to detect any qualitative difference between the (loop-free) folding trajectories generated by either dynamics.

An important question relates to the reliability of the predicted results. In a previous study we investigated

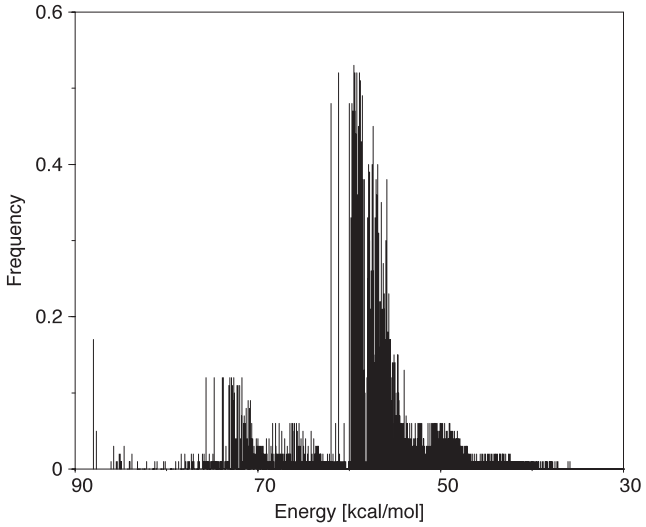


**FIGURE 13.** Fraction of folding paths visiting local minima in the Q$\beta$ variant SV11. The majority of paths visits the local minima in the basin of the metastable structure where the paths get trapped. Only about 16% reach the ground state.

the robustness of most currently available algorithms in predictions on RNA structures (Tacker et al., 1996) and found that certain statistical properties of the sequence-to-structure map are reproduced quite well, although predictions for individual sequences may be poor (see also Huynen et al., 1997). Similarly, in the context of the present study we expect that generic features are reproduced correctly by our kinetic simulations. Such features pertain to the relationship between major kinetic properties and the barrier structure of conformational spaces. Examples considered here emphasized the partition of conformation space into basins of attraction for several dominant structures, and, consequently multiple folding mechanisms and folding time scales. We also found the existence of dominant metastable states with large basins in the presence of mfe structures that are kinetically far less accessible despite their thermodynamic dominance.

The often made conjecture that a large free energy gap between the ground state and the first suboptimal conformation of a biopolymer is indicative of good folding properties was shown to be incorrect, at least for RNA secondary structures. We designed several counterexamples. What actually determines the folding behavior is the number of nucleation centers for double helical regions, as well as the numbers and the heights of the saddle points that have to be passed along a trajectory from the open chain to the folded conformation. In other words, what matters is the multiplicity of trajectories that are roughly equivalent with respect to their overall energy profile. The barrier trees that organize the local minima in a hierarchical fashion turned out to be an excellent tool for studying folding pathways. Their most serious limitation, however, consists in not providing any information about the entropy of paths, that is, the multiplicity of pathways whose highest saddlepoint is slightly worse than the barrier.

We have considered folding so far as a process leading from the open chain to the mfe conformation or a metastable state. Future work aims at developing extensions of our algorithm to enable the study of kinetic RNA/RNA cofolding, that is, the hybridization of two RNA molecules into a joint secondary structure, as well as kinetic RNA folding in conjunction with chain growth. The latter is particularly important when understanding the folds of very long sequences.

## ACKNOWLEDGMENTS

We thank Paul Higgs for discussions that proved useful in the early stages of this project. Financial support for this work was provided by the Austrian Science Foundation, FWF, (Projects P-11065 and P-13093), by the Commission of the European Union (Project PL-970-189), and by core grants to the Santa Fe Institute from the John D. and Catherine T. MacArthur Foundation, the National Science Foundation, and the U.S. Department of Energy. W. F. and his research program are supported by Michael A. Grantham.

*Manuscript accepted without revision November 23, 1999*

## REFERENCES

Banerjee AR, Jaeger JA, Turner DH. 1993. Thermal unfolding of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry 32*:153–163.

Breton N, Jacob C, Daegelen P. 1997. Prediction of sequentially optimal RNA secondary structures. *J Biomol Struct Dyn 14*:727–740.

Fontana W, Schuster P. 1987. A computer model of evolutionary optimization. *Biophys Chem 26*:123–147.

Galzitskaya OV, Finkelstein AV. 1996. Computer simulation of secondary structure folding of random and "edited" RNA chains. *J Chem Phys 105*:319–325.

Gardiner CW. 1985. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, 2nd ed. Berlin: Springer-Verlag.

Gillespie DT. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comp Phys 22*:403–434.

Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem 81*:2340–2361.

Gultyaev AP, van Batenburg FHD, Pleij CWA. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J Mol Biol 250*:37–51.

Gutell RR. 1993. Evolutionary characteristics of RNA: Inferring higher order structure from patterns of sequence variation. *Curr Opin Struct Biol 3*:313–322.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh Chem 125*:167–188.

Huynen M, Gutell R, Konings D. 1997. Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol 267*:1104–1112.

Jacob C, Breton N, Daegelen P. 1997a. Stochastic theories of the activated complex and the activated collision: The RNA example. *J Chem Phys 107*:2903–2912.

Jacob C, Breton N, Daegelen P, Peccoud J. 1997b. Probability distribution of the chemical states of a closed system and thermodynamic law of mass action from kinetics: The RNA example. *J Chem Phys 107*:2913–2919.

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

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers 29*:1105–1119.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys 21*:1087–1092.
Mironov A, Lebedev VF. 1993. A kinetic model of RNA folding. *BioSystems 30*:49–56.
Morgan SR, Higgs PG. 1996. Evidence for kinetic effects in the folding of large RNA molecules. *J Chem Phys 105*:7152–7157.
Nussinov R, Jacobson AB. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci USA 77*:6309–6313.
Pörschke D. 1974. Model calculations on the kinetics of oligonucleotide double helix coil transitions. Evidence for a fast chain sliding reaction. *Biophys Chem 2*:83–96.
Schmitz M, Steger G. 1996. Description of RNA folding by simulated annealing. *J Mol Biol 225*:254–266.
Suvernev AA, Frantsuzov PA. 1995. Statistical description of nucleic acid secondary structure folding. *J Biomol Struct Dyn 13*:135–144.
Tacker M, Stadler PF, Bornberg-Bauer EG, Hofacker IL, Schuster P. 1996. Algorithm independent properties of RNA secondary structure predictions. *Eur Biophys J 25*:115–130.
Thirumalai D, Woodson SA. 1996. Kinetics of folding of proteins and RNA. *Acc Chem Res 29*:433–439.
Waterman MS, Smith TF. 1978. RNA secondary structure: A complete mathematical analysis. *Math Biosci 42*:257–266.
Wuchty S, Fontana W, Hofacker IL, Schuster P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers 49*:145–165.
Zuker M, Stiegler P. 1981. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res 9*:133–148.