J Stat Phys (2011) 142: 1337–1352
DOI 10.1007/s10955-011-0174-2

Arrhenius Lifetimes of RNA Structures from Free Energy Landscapes

Ben Sauerwine · Michael Widom

Received: 17 November 2010 / Accepted: 3 March 2011 / Published online: 18 March 2011

Abstract We develop a novel method to define an effective free energy barrier between various conformations available to an RNA sequence. Our approach utilizes the energy landscape computed using the `ViennaRNA` package. Replacing the energies with configurational free energies yields improved agreement with the Arrhenius law provided we account for the entropies of both the basin states and the barrier states. This improved agreement comes without altering the computational complexity class of the algorithm. We discuss the combinatorial explosion of high-energy states available to RNA, and the properties of Kawasaki and Metropolis transition models that affect the applicability our method.

Keywords RNA · Kinetic · Arrhenius · Lifetime

1 Introduction

1.1 Barrier Crossings in Biology

Energy barrier crossings play crucial roles throughout molecular biology [1]. They are particularly interesting in systems where a molecule possesses multiple conformations, each with distinct functions. Lifetimes of metastable states and the barrier-controlled rates of transitions between conformations, are often crucial to life. In fact, the Levinthal paradox [2] implies that the energy barriers in the conformational free energy landscape available to biopolymers must be naturally selected to facilitate desirable folding pathways. Chaperonin molecules [3] provide a typical example in the context of protein dynamics. ATP-assisted conformational transitions of the chaperonins are themselves used to facilitate the proper folding of other proteins. In other cases structural transitions may be deleterious, for example in diseases such as Creutzfeldt-Jakob, Alzheimer's, and Parkinson's, which are associated with transitions to misfolded protein states [4, 5]. In the context of nucleic acids, examples

B. Sauerwine · M. Widom (✉)
Carnegie Mellon University, Pittburg, USA
e-mail: widom@andrew.cmu.edu

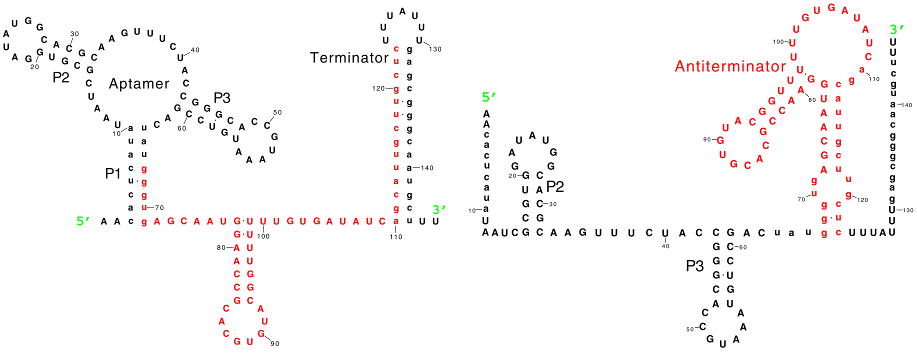

Fig. 1 (Color online) Alternate secondary structures of the guanine-binding *xpt* riboswitch in *Bacillus subtilis*. (*Left*) Aptamer in ligand-bound state forms three stems. The P_1 stem (*lowercase*) of the aptamer prevents formation of antiterminator (*red*), thereby allowing the terminator hairpin (*lowercase*) to form and halting transcription. We thus denote this switch as "bound off". (*Right*) Aptamer in unbound state allows formation of antiterminator, thereby preventing terminator hairpin formation and allowing transcription to proceed

include gene regulation by micro-RNA and riboswitches [6, 7]. Micro-RNA regulation depends the messenger RNA to open its folded structure so that the regulatory micro-RNA may bind to its target [8]. Riboswitches regulate gene expression by switching between alternate conformations upon binding or unbinding of a ligand molecule. Conformational changes in one region of the riboswitch catalyze structural changes in other regions.

This paper focuses on the folding of RNA because of its rich interplay of biological function with physical behavior (see references [9–21]). The simplification available by restricting attention to the *secondary* structure (the pattern of pairing between complementary nucleotides A-U and C-G) aids our study. We are motivated in particular by the dynamics of the riboswitch, because its function depends upon its ability to fold and refold on a time scale set by the rate of transcription.

A typical example, shown in Fig. 1, is the *xpt* riboswitch of *Bacillus subtilis* [7]. There are two distinct functional conformations ("on" and "off") and three relevant intermediate structures ("aptamer", "antiterminator", and "terminator"). Transcription proceeds from the 5′ end to the 3′ end at a rate of approximately 50 nt/s. When the appropriate ligand (in this case guanine) is present in solution, the ligand stabilizes the P1 stem of the aptamer. The stabilized aptamer resists formation of the antiterminator, in turn leaving the remaining nucleotides free to form the terminator hairpin structure. If the terminator hairpin structure has formed by the time the complete riboswitch has been transcribed, genetic transcription halts.

In the absence of guanine, the P1 stem competes with the antiterminator structure. Formation of the antiterminator structure prevents formation of the terminator hairpin and thus allows transcription to proceed. While the terminator hairpin conformation is more stable than the antiterminator conformation (even in the absence of guanine), a free energy barrier prevents the terminator hairpin structure from forming during the time required for transcription to proceed beyond the end of the riboswitch. Thus the lifetime of the metastable antiterminator state must exceed a threshold in order for the switch to function.

1.2 RNA Energy Landscape

Focusing on the secondary structure restricts the conformation space of any particular RNA sequence of finite length N to a finite discrete set. In contrast, a continuum of conformations characterize the *tertiary* structures that are crucial for understanding the folding of proteins. Empirical interaction energies [22] allow rapid calculation of the binding energy E_s of any secondary structure s. Actually these parameters include vibrational entropy contributions so they are a type of free energy, but we will refer to them as energies in order to distinguish them from the complete free energy of the ensemble of all possible secondary structures. Indeed, we consider a single RNA molecule as a complete one dimensional statistical mechanical system with associated thermodynamic properties. We define the partition function and free energy as

$$Z = \sum_s e^{-E_s/RT}, \qquad G = -RT \ln Z \tag{1}$$

where the sum is over all structures s in the ensemble.

While the inhomogeneity of the genetic sequence complicates the thermodynamic limit of large N, for *generic* sequences we find the energy and entropy densities per nucleotide approach finite limits. In particular, if we let $\bar{S}$ represent the mean entropy per nucleotide, then the ensemble for a length N sequence contains approximately $\exp(\bar{S}N/R)$ structures accessible at a given temperature. At high temperatures the entire ensemble consisting of N_s distinct structures will be sampled. We can place an upper bound on the value of N_s by noting that N nucleotides can form a maximum of $N/2$ pairs. There are C_p ways to arrange p pairs consistently with the requirement of no pseudoknots [23, 32], where $C_p = (2p)!/(p+1)!p!$ is the pth Catalan number. For each pairing there are $N!/(N-2p)!(2p)!$ ways to place the remaining unpaired nucleotides. Thus

$$N_{s,max} = \sum_{P=0}^{\lfloor N/2 \rfloor} C_P \binom{N}{2P} \sim 2^N/\sqrt{2\pi N} \tag{2}$$

gives the maximum number of structures available to the RNA strand of length N nucleotides.

The density of states $\Omega(E)$ is the number of structures per unit energy interval $(E, E + \mathrm{d}E)$. The number of states up to an energy threshold E_t is thus $N_s(E_t) = \int_0^{E_t} \Omega(E)\mathrm{d}E$, and we define $N_s = N_s(E_{max})$ as the total number of states in the full ensemble. The probability to be in a structure of energy of energy E represents a balance between the exponential decrease of the Boltzmann factor $\exp(-E/RT)$ and the variation of $\Omega(E)$, which often increases strongly with E (see, e.g. Fig. 2c), a phenomenon also known as entropy-enthalpy compensation [1]. For many realistic RNA sequences, the probability to be in the minimum energy structure can be negligibly small in thermal equilibrium. Remarkably, although the size of the ensemble grows exponentially in N, an algorithm exists [24] to find the minimum energy structure (ME) in $\mathcal{O}(N^3)$ time, and another [25] of the same complexity that calculates the partition function Z.

1.3 Arrhenius Laws

Analytic theories of barrier limited reaction kinetics have been developed previously in many contexts, always yielding rates of the generic form

$$r = ce^{-\Delta G/RT} \tag{3}$$

while differing in the precise form and interpretation of c and ΔG. The exponential term can be understood qualitatively in terms of a three state model of transition state theory,

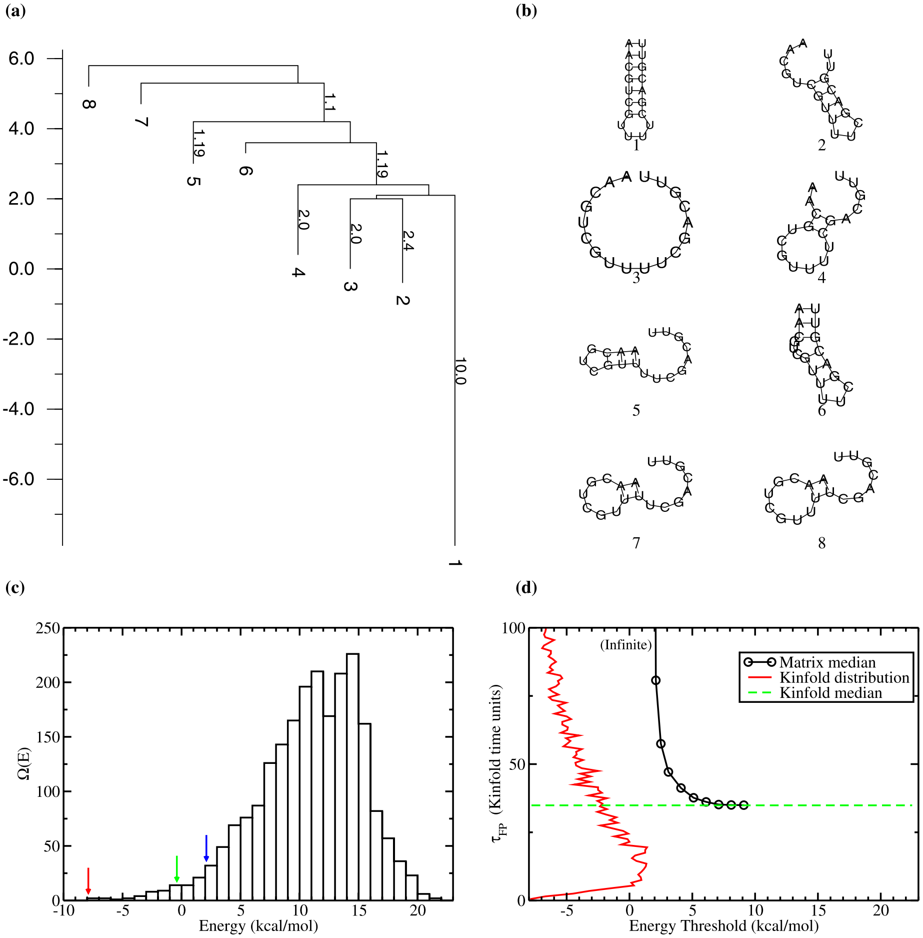

Fig. 2 (Color online) Ab-initio study of the sequence AACGUCGUUUUCGACGUU. (**a**) Basins and barriers calculated by the `barriers` program. Scale indicates energy E in kcal/mol. (**b**) Local energy minimum secondary structures of the eight energy basins. (**c**) Complete density of states of the 2201 legal secondary structures. *Red, green* and *blue* arrows label, respectively, E_f (state 1), E_i (state 2) and E_b. (**d**) Median first passage times obtained by matrix exponentiation compared with the distribution of Kinfold simulated first passage times. The horizontal *dashed line* indicates the median of the Kinfold distribution

consisting of an initial locally stable basin with free energy G_i (the "reactants"), an unstable barrier state with free energy G_b (the "activated complex"), and a final stable state with free energy G_f (the "reaction product"). Allowed transitions are $i \leftrightarrow b \rightarrow f$. The rate of transitions is limited by excitations from the initial basin i to the barrier b. Among systems that have not reached the final stable state, a quasi-equilibrium may be established between states i and b. A fraction of systems

$$\frac{e^{-G_b/RT}}{e^{-G_b/RT}+e^{-G_i/RT}} \approx e^{-\Delta G/RT} \tag{4}$$

(where $\Delta G = G_b - G_i \gg RT$) occupy the barrier state and have the potential to transition to the final stable state.

Such reactions were modeled as a one dimensional diffusing Brownian particle by Kramers [26] who derived a prefactor c depending upon the viscosity and a characteristic molecular frequency. Langer [27] addressed the decay of metastable states and applied his theory to droplet nucleation and growth. He incorporated the full vibrational spectrum at the metastable basin and barrier, thereby treating ΔG as a vibrational free energy difference, while the prefactor c includes the unstable eigenvalue of perturbations around the saddle point of the transition surface.

The dominant factor in these theories is an Arrhenius-like exponential dependence of rate on a free energy. Various theories differ on the precise definition of the free energy ΔG and the value of the prefactor c. Indeed, no precise decomposition into barrier and prefactor is possible because an entropic contribution to the free energy, expressed as $\Delta G = \Delta H - T\Delta S$ transforms the Arrhenius law of (3) into

$$ce^{-\Delta G/RT} = (ce^{S/R})e^{-\Delta H/RT}. \tag{5}$$

Realistic systems relevant to biology (RNA structures in particular) usually contain more than three states. Instead, a network of many basins is linked though multiple barriers. Despite this complexity, the network may reduce to an *effective* three state system, in which disjoint subsets of the ensemble of states coalesce to form the effective initial and final basins and the effective barrier. In this case it may still be possible to define an Arrhenius law such as (3), but ΔG will clearly include entropic contributions, and the prefactor c may lack universality.

1.4 Outline of Paper

We begin in Sect. 2 by discussing ab-initio methods for determining the lifetimes of metastable states of RNA, then we move in Sect. 3 to heuristic approaches that trade accuracy for greater physical insight. Our ab-initio methods start with the kinetic Monte Carlo program `Kinfold` [28], belonging to the ViennaRNA package [29], that we use both to define the microscopic dynamics and to calculate "correct" metastable state lifetimes for comparison with heuristic methods. An alternative ab-initio method exploits our ability to fully enumerate every possible secondary structure in the configurational ensemble, utilizing the ViennaRNA program `RNAsubopt`. We construct the complete rate matrix underlying the dynamics of `Kinfold` and solve the Fokker-Planck equation for the evolution of configuration space probabilities.

Moving on to heuristic methods, we first review the ViennaRNA program `barriers` that creates an energy landscape by assigning secondary structures to local energy basins separated by energy barriers. We show that the predicted energy barriers ΔE correlate with the `Kinfold`-simulated lifetimes but with poor quality of fit to the Arrhenius law. Finally we propose a new analysis of the barrier structure that replaces the energy landscape with a free energy landscape. The predicted free energy barriers show improved quality of fit to the Arrhenius law.

None of the other methods we describe competes with `Kinfold` in speed, and indeed weighted ensemble [30, 31] and other methods exist that could determine mean first-passage time even more efficiently. However, the methods we describe may yield deeper understanding of the underlying physics than is possible with a direct simulation.

2 Ab-initio Methods

Our approaches to metastable state lifetime calculation exploit our ability to define a discrete dynamics on the space of secondary structures (the set of bound pairs of nucleotides) of RNA molecules independent of other conformational parameters. For reasons of computational efficiency, pseudoknots [32] are not considered. Nonetheless, all methods described in Sects. 2 and 3 could be applied to any system with a notion of adjacency where it is possible to enumerate states below a particular energy threshold, thus pseudoknots could be included in principle. We begin our discussion of ab-initio methods with a discussion of the `ViennaRNA` package and the `Kinfold` move set. We let the elementary steps of `Kinfold` define the dynamics that are employed throughout the remainder of this work. Indeed, we take lifetimes calculated by `Kinfold` as the standard for comparison of all other methods.

2.1 The `ViennaRNA` Package

The `ViennaRNA` package is a suite of computer programs centered on RNA secondary structure that assign energies based on empirically determined interactions among nucleotides. We make use of three programs. First, we use `Kinfold` [28], which performs Monte Carlo simulations on the space of RNA secondary structures. Both Kawasaki and Metropolis rules are available for selecting moves subject to the constraints of detailed balance. All of our `Kinfold` runs are performed with default settings except as specified. Next, we use `RNAsubopt` [33], a tool which enumerates all of the RNA secondary structures of a sequence below a given energy threshold E_t. Its algorithm makes use of a recursion using the partition function of a structure which can be generated in $\mathcal{O}(N^3)$ time, where N is the length in nucleotides of the RNA chain under consideration. In order to report these structures, however, one must still iterate through all of them which requires $\mathcal{O}(N_s)$ time (i.e. exponential time) since the number of RNA secondary structure available grows exponentially in the length of the chain. Finally, we use the `barriers` program [34], which finds local energy minima, basins of excited states related to these local minima, and barriers between adjacent basins.

2.2 The `Kinfold` Kinetic Monte Carlo Program

Central to a model of secondary structure dynamics is a model of adjacency in configuration space. In particular, we consider the set of allowed states (secondary structures) $\{s_i, i = 1, \dots, N_s\}$ and a symmetric adjacency matrix $\mathbf{A}$ with matrix elements $A_{ji} = 1$ if s_j is reachable in exactly one move from s_i, and $A_{ji} = 0$ otherwise. Given any two conformations, `Kinfold` considers legal moves to be either the opening or closing of a single base pair, or for a single base to change the nucleotide it's bound to. `Kinfold` returns a trajectory of states visited along with a simulated time in arbitrary units. The calibration of these arbitrary units to experiment depends on the move set, whether Metropolis or Kawasaki dynamics are chosen for the Monte Carlo simulation, temperature and energy parameters. We do not require calibration for the purposes of this paper, but the value would be of order microseconds [35]. In contrast to conventional molecular dynamics, whose simulation times are limited to picoseconds or nanoseconds, Kinfold can routinely reach time scales of order seconds.

Given an initial state s_i and the set of states s_j reachable in a single move, `Kinfold` determines the a matrix $\mathbf{R}$ of transition rates between the initial states and its neighboring states using either the Metropolis (M) or Kawasaki (K) rules:

$$R^M_{ji} = \begin{cases} \min(1, e^{-\Delta E/RT})\ A_{ji} & (j \neq i) \\ -\sum_{k\neq i} R^M_{ki} & (j = i) \end{cases} \tag{6}$$

$$R^K_{ji} = \begin{cases} e^{-\Delta E/2RT}\ A_{ji} & (j \neq i) \\ -\sum_{k\neq i} R^K_{ki} & (j = i) \end{cases} \tag{7}$$

In (6) and (7), the energies E_i are provided for each state s_i by the ViennaRNA energy model, and $\Delta E = \Delta E_{ji} = E_j - E_i$.

Given these transition rates, `Kinfold` increments the system time from an exponential distribution of decay rates from some state s_i by

$$\Delta t = \frac{\ln(1/X)}{-R_{ii}} \tag{8}$$

where X is a uniform random variable in (0, 1]. Finally, the system transitions out of state s_i to an adjacent state s_j $(j \neq i)$ chosen with probability: $p_{i\to j} = R_{ji}/|R_{ii}|$ This process repeats until the simulation reaches a user-defined final structure s_f. Every `Kinfold` step results in a transition, provided the state s_i has neighbors. We treat the ending condition as a trapping state with no neighbors by modifying the transition rate matrix so that $R_{jf} = 0$ and $R_{ff} = 0$.

It is instructive to contrast the Metropolis and Kawasaki dynamics. Notice, first, that for our simple three state system consisting of initial, barrier and final state, the effective barrier for transitions from initial to barrier state matches the expectation based on the Arrhenius law for Metropolis dynamics but is only half as large for Kawasaki. On the other hand, Metropolis dynamics makes no distinction in the rates for downhill (energy decreasing) transitions, while Kawasaki favors pathways of more rapid energy decrease and thus captures the qualitative effect of a driving force deriving from the gradient of the energy.

2.3 Exact Calculation of Arrhenius Barrier

For short sequences the space of secondary structures is of sufficiently small size N_s that we can explicitly construct and manipulate the full rate matrix $\mathbf{R}$. For intermediate length sequences the exponential growth of N_s prevents construction of the complete matrix $\mathbf{R}$ but still allows us to construct its restriction to low and moderate energy structures up to an energy threshold E_t that include the barriers between low energy basins and many structures in the neighborhoods of those barriers.

`Kinfold` samples configuration space utilizing a Markov chain based on elements of the rate matrix R_{ji} to mimic a continuous time stochastic process. We can formally treat the associated Fokker-Planck equation for the evolutions of the state occupation probabilities $\boldsymbol{\theta} \equiv \{\theta_i(t)\}$.

$$\frac{\partial \boldsymbol{\theta}}{\partial t} = \mathbf{R}\boldsymbol{\theta} \tag{9}$$

whose solution is simply the exponential

$$\boldsymbol{\theta}(t) = e^{\mathbf{R}t}\boldsymbol{\theta}(0). \tag{10}$$

Typically we begin in some specific initial state s_i. The probability to be in the final state s_f at time t is then $\theta_f(t)$. The rate to arrive at s_f at time t is $\mathrm{d}\theta_f/\mathrm{d}t = [\mathbf{R}\exp(\mathbf{R}t)]_{fi}$ and the mean first passage time is thus

$$\tau_{mean} = \int_0^\infty t\ [\mathbf{R}e^{\mathbf{R}t}]_{fi}\ \mathrm{d}t \tag{11}$$

Given τ_{mean} we define the Arrhenius barrier as $\Delta G = -RT \ln \tau_{mean}$. While formally exact, these expressions are of limited practical utility because exponentiating the matrix $\mathbf{R}$ requires diagonalization, an $\mathcal{O}(N_s^3)$ process under the best of circumstances, which is further aggravated by the ill-conditioning of $\mathbf{R}$.

As an alternative to exactly evaluating the matrix exponential, we approximate $\exp(\mathbf{R}\Delta t) \approx \mathbf{I} + \mathbf{R}\Delta t$ for small times Δt, then extend to longer times by matrix multiplication since $\exp(\mathbf{R}t_1)\exp(\mathbf{R}t_2) = \exp(\mathbf{R}(t_1 + t_2))$. We calculate the *median* first passage time τ_{median} by solving for $[\exp(\mathbf{R}\tau_{median})]_{fi} = 1/2$, which can be achieved with $\mathcal{O}(\ln(\tau_{median}/\Delta t))$ multiplications as outlined in Algorithm 1 in Appendix A. Although the individual matrix multiplications remain $\mathcal{O}(N_s^3)$ operations, this eliminates the difficulty associated with the ill–conditioning of $\mathbf{R}$.

As an example, we compare the median first passage time from an exact calculation to the median first passage time obtained from Kinfold for a short RNA sequence (see Fig. 2). We choose the nucleotide sequence "AACGUCGUUUUCGACGUU", as a typical example in which a metastable state (s_i, labeled 2) transforms to a stable final state (s_f, labeled 1) through a barrier. In the barrier state (s_b, indicated by the horizontal line joining 2 to 1 in Fig. 2a, but not illustrated in Fig. 2b) all nucleotides are unpaired except the final G-U pair at the end of the stem in s_i. No pathway reaches s_f from s_i without encountering a configuration of at least this energy, hence this particular barrier state represents a saddle point. The associated energies are $E_i = -0.4$, $E_b = +2.1$ and $E_f = -7.9$ in units of kcal/mol.

Figure 2c shows the complete density of states $\Omega(E)$, 2201 states in total, including many of energy significantly higher than E_b. This energy landscape is reflected in the matrix-calculated median first passage times shown in Fig. 2d. These are obtained by evaluating the Kawasaki transition rate matrix $\mathbf{R}^K$ in configurations subspaces of varying energy threshold E_t. Also shown is a histogram of Kinfold-simulated first passage times. With a threshold below E_b no paths reach s_f from s_i and thus the matrix-calculated time diverges. Precisely at threshold energy $E_t = E_b$ transitions are possible passing directly over the saddle. At this threshold the matrix-calculated τ_{median} greatly exceeds the `Kinfold`-simulated value. However, as E_t increases the calculated median decreases monotonically and converges to the simulated value. Thus the majority of actual transitions follow pathways whose maximum energy significantly exceeds E_b.

3 Empirical Methods

We now introduce empirical approaches to lifetime calculations. While these are not strictly needed, given our ability to accurately simulate lifetimes using `Kinfold`, we use our ability to fully enumerate the configuration space or its low energy subsets to gain insight into the factors that determine metastable state lifetimes. Specifically we contrast predictions based on local basin minimum and saddle point energies as obtained from the `barriers` program with a new variant we have developed based on basin and barrier *free* energies.

We are interested in the competition of incompatible hairpins, in analogy with the competition of metastable antiterminators with stable terminators common to riboswitches. Hence

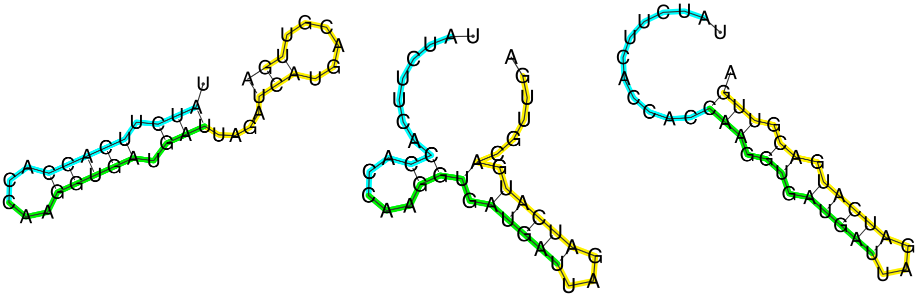

Fig. 3 (Color online) Example competing hairpin structures for sequence "UAUCUUCACCACCAAGGUGAUGAUUAGAUCAUGACGUUGA". (*Left*) Initial structure, $E = -8.1$ kcal/mol. (*Center*) barrier, $E = -1.4$ kcal/mol. (*Right*) final structure, $E = -8.5$ kcal/mol

we generated a set of random sequences of the form $SLS^\star L'S$, where $S^\star$ is the reverse complement of S so as to form hairpin stems with S, and L and L' are hairpin loop sequences. We then introduced random mutations so as to introduce the possibility of bulges within the hairpin stems (see Fig 3). We began with an initial set of 200 possible sequences, all of length $N = 40$ of which we selected by hand 84 that matched our competing basin criteria and exhibited barriers in the range 5–8 kcal/mol, in the range of what we observe for real riboswitches. An alternate sample set consists of completely random length $N = 40$ sequences, with s_f and s_i defined as the lowest and second-lowest basins respectively. Out of an initial pool of 100 such random sequences we employ all 91 for which the barrier calculation proved tractable. The energy barriers of random sequences ranged from 1–8 kcal/mol.

We assess the performance of the alternate calculational methods by fitting the calculated Arrhenius barriers to the `Kinfold`-simulated mean first passage time, in the form $\tau_{mean} = c_0 e^{-(\Delta G/RT)}$ where ΔG is either ΔE from the `barriers` program, or else the free energy difference obtained from our own approach. We set $RT = 0.6163$ kcal/mol appropriate for $T = 37°\mathrm{C} = 310$ K and allow the coefficient c_0 to vary. The fit yields a "coefficient of determination" R^2, which is the mean square deviation the fitted data, normalized to the mean square deviation of the actual data. We convert this to a normalized mean square error $\bar{\sigma}^2 = 1 - R^2$ which represents the mean square difference of the fit from the data, normalized to the mean square deviation of the data.

3.1 The `barriers` Program

The `barriers` program takes a list of suboptimal states from the `RNAsubopt` program and employs the `Kinfold` move set in order to find saddle points separating basins surrounding local minima in the free energy landscape. The program examines each secondary structure available to a sequence from lowest to highest energy. If a structure is found that is not adjacent to any other structure, it is reported as a local minimum and will define a configurational basin. If a structure is found adjacent to a local minimum or to other structures that belong to some local minimum's basin, then the structure also belongs to that basin. Finally, the lowest energy structure that is adjacent to multiple basins is reported as a saddle point connecting these basins. This method examines every structure and compares it to every previously examined structure, yielding an $\mathcal{O}(N_s^2(E_t))$ scaling of computational effort. The calculation generally becomes prohibitive for sequences of length 75 or more.

Results of applying the `barriers` calculation are shown in Fig. 4 for our competing hairpin model sequences (see black circles). Fits for both the competing hairpins and our

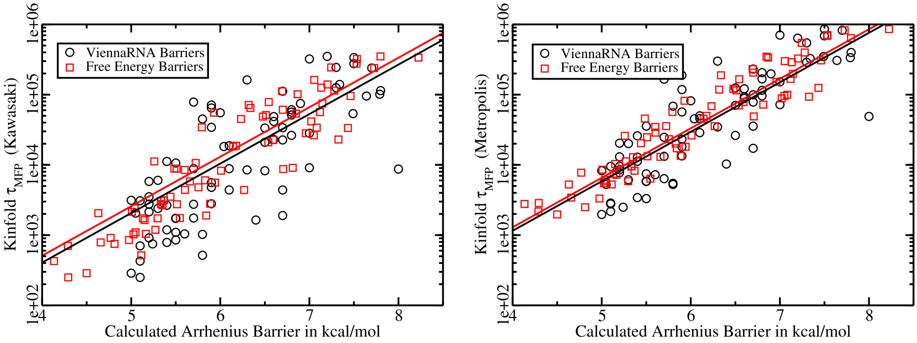

Fig. 4 (Color online) Plots of (*left*) Kawasaki-rule simulated and (*right*) Metropolis-rule simulated lifetime vs. the calculated barrier in dual-hairpin topology short sequences 40 nucleotides long. *The lines* shown for comparison are fits to the function $c_0 \exp(\Delta G/RT)$ for $T = 37°$. For numerical data on quality of fit, see Table 1

Table 1 Arrhenius prefactors c_0 and normalized mean-square errors, $\bar{\sigma}^2_{B,F} = 1 - R^2$ (R^2 is the coefficient of determination), for barriers (B) and our free energy methods (F)

Topology	Rule	c_0^B	$\bar{\sigma_B}^2$	c_0^F	$\bar{\sigma_F}^2$
Dual-Hairpin	Kawasaki	0.62	0.238	0.77	0.084
Dual-Hairpin	Metropolis	1.75	0.160	1.96	0.055
Random	Kawasaki	0.18	0.028	0.17	0.021
Random	Metropolis	0.65	0.022	0.61	0.018

random sequences are presented in Table 1. While there is a qualitative trend of increasing Kinfold-simulated τ_{MFP} with increasing calculated barrier, the scatter of the data is considerable. Notice that the times are larger, on average, for Metropolis dynamics than for Kawasaki. This is expected because as we noted the factor of $1/2$ in the Kawasaki exponent (see (7)) increases the rate to escape from a local minimum, and additionally the Kawasaki rules favor steeper downhill transitions.

Notice also that the mean square errors are smaller for random structure than for our competing hairpin structures. This is because the random topologies tend to have less complex transition pathways than the dual hairpin topologies.

3.2 Arrhenius Barriers Using the Free Energy Landscape

The key difference between our algorithm and `barriers` is that we consider the entropies of the barriers and basins in addition to their energies. The algorithm itself is separated into four separate phases. First, we use `RNAsubopt` in order to gather a list of metastable structures below a chosen energy threshold E_t, sorted ascending by energy. Second, we label each state in this list by the free energy basin or basins that it belongs to (see Algorithm 2 in Appendix). Third, we calculate the partition function for structures in each subset of basins (see Algorithm 3). Finally, we use the partition functions from the third step in order to calculate the rate-limiting Arrhenius barrier between an initial and final state (see Algorithm 4).

A simple example of our method for discovering this rate-limiting barrier is illustrated in Fig. 5. In part (a), we show the initial system under consideration. Three basins, the initial, final and one other basin are labeled. Separating the basins are “border” states, so-named because of the topological nature of their definition, and to clarify the distinction with the

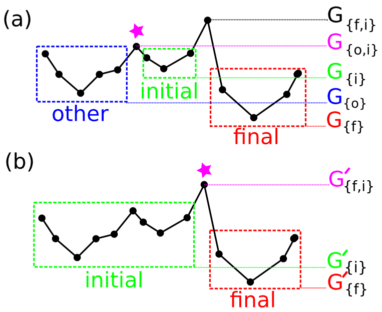

Fig. 5 (Color online) Simplified one-dimensional model of our free-energy based barrier finding program. *Black line* segments indicate allowed transitions. Free energies G increase vertically upwards. Part (**a**) shows the initial landscape as obtained from the `barriers` program. Part (**b**) shows the merger of two metastable basins

free energy barrier ΔG. The border states between these are considered members of all basins they neighbor for the purpose of merging, and are considered members of neither for the purpose of calculating basin free energy. Although not apparent in this simple one-dimensional example, in general the border between basins consists of a large number of states.

The set notation in the subscripts of free energy G indicates the basins that all states included in G belong to. For example, $G_{\{i\}}$ is the free energy of all states belonging to basin i, and $G_{\{o,i\}}$ is the free energy of all states (in this illustration just a single state) that border basin i and o. A set with more than one element is part of a border, and a set with exactly one element is part of a basin.

In the first iteration of our algorithm, applied to this example, $G_{\{o,i\}}$ (starred in Fig. 5a) is the lowest border state between basins. The free energy barrier $\Delta G = G_{\{o,i\}} - G_{\{i\}}$ is recorded because the highest of these discovered among all iterations will represent the rate-limiting barrier ΔG for eventual $i \to f$ transitions. During the merge step, all states belonging to the state to be merged (o in this example) become members of the expanded initial basin i. Notice that $G'_{\{i\}} < G_{\{i\}}$ owing to the merging of basins. Next, the lowest available border state between the merged initial state and another neighbor is starred. The free energy difference between the merged initial state and the border state is higher than the previously recorded rate-limiting step, so $\Delta G' = G'_{\{f,i\}} - G'_{\{i\}}$ from part (b) becomes the calculated Arrhenius barrier energy between the start and end state.

If two borders were of equal free energy, a situation not described in our pseudocode Algorithm 4, then one must be selected for merging by some tie-breaker method. For our purposes, we select the neighboring basin with the lowest free energy. In the rare situation where the neighboring basins have the same free energy we select the lexicographically first local minimum structure, a situation that did not arise in our tests.

Results of applying the our calculation are shown in Fig. 4 for our competing hairpin model sequences (see red squares). Fits for both the competing hairpins and our random sequences are presented in Table 1. While there is still some scatter relative to the Arrhenius law, the mean square error is significantly reduced relative to the original `barriers` calculation.

4 Conclusions

It is interesting to ask what the relative contributions of basin and barrier entropy are to the improved agreement. Figure 6 reproduces the Metropolis data previously shown in Fig. 4, now with two additional data sets. The set labeled “Rate-limiting barrier” includes the border

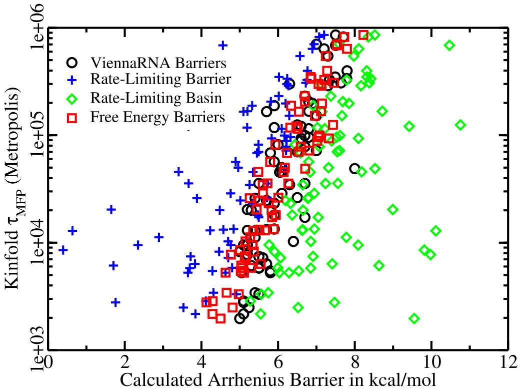

Fig. 6 (Color online) Comparison of basin and barrier entropy contributions to Arrhenius barrier

entropy in G_{border} but includes only the minimum energy of the initial basin, leading to a barrier $\Delta G = G_{border} - E_i$. Since the entropy reduces the free energy of the barrier state, the calculated barrier heights are systematically reduced. However, the scatter of the fit is greatly increased. The set labeled "Rate-limiting Basin" includes the basin entropy in G_i but only includes the minimum energy of the border, a single saddle configuration, E_{border}. This leads to a barrier $\Delta G = E_{border} - G_i$. Since the entropy reduces the free energy of the basin but not the barrier, the calculated barrier heights are systematically increased. However, the scatter of the fit also increases. Thus the improved fit to the Arrhenius law produced by our free energy method results from combined and largely offsetting contributions of both basin and barrier entropy. The improved fit results from the entropy *difference* between barrier and basin.

Currently, our methods can access barrier crossings in subunits of functional RNA but not very large RNA structures such as riboswitches. The primary limitation is our requirement of fully enumerating the configurational ensemble. Hope for further progress on large molecules depends on introducing some sort of sampling procedure or coarse graining [36]. Presently the most efficient method for actually calculating metastable state lifetimes is direct Kinfold simulation. If the lifetimes prove too large to directly simulate, methods such as "weighted ensemble" path sampling [30, 31] can improve simulation efficiency. The chief contribution of our study is an understanding of the relative contributions of basin and barrier entropy, and methods to calculate these in principle.

In conclusion, understanding the lifetimes of RNA structures is a challenging problem, and to computationally study the longer timescales on the order of seconds relevant to competition between various conformations currently requires secondary structure simulations such as those performed by `Kinfold`. We have produced an algorithm of comparable computational cost to `barriers` and tested it on short, 40 nucleotide structures using both Metropolis and Kawasaki rules for kinetic moves and compared it to an Arrhenius Law for transitions at 37°C. We found that for the purpose of predicting transitions between conformations in these structures our algorithm provides a more reliable prediction than `barriers`. This is not surprising as our recipe is aware of the entire free energy landscape instead of only local minima and saddle points. Since there is a combinatorial explosion in the density of states of RNA structures at higher energies, the use of these high energy structures for the purpose of transitions between local free energy basins leads to greater accuracy. Further, our algorithm is quite versatile: it can be applied to any system with a notion of adjacency provided that discrete energy states up to a chosen threshold can be enumerated.

Acknowledgements The authors gratefully acknowledge financial support of this research by the DSF Charitable Foundation. We also acknowledge useful discussions with Dan Zuckerman.

Appendix A: Pseudocode for algorithms

Algorithm 1 Determine the median first passage time from an initial distribution to a final state indexed f using the rate matrix $\mathbf{R}^{exact \to f}$. Below, I indicates the identity matrix.

```
Require: Matrix $\mathbf{R}$
Require: Initial occupation of states $\boldsymbol{\theta}(0)$
Require: Target state $f$
Require: Time increment $\Delta t$
  $\mathbf{M}_1 \leftarrow (\mathbf{I} + \mathbf{R}\Delta t)$
  $c \leftarrow 1$
  while $[\mathbf{M}_c \boldsymbol{\theta}(0)]_f < 0.5$ do
    $c \leftarrow c + 1$
    $\mathbf{M}_c \leftarrow \mathbf{M}_{c-1} \cdot \mathbf{M}_{c-1}$
  end while
  $\tau \leftarrow 0$
  $\mathbf{S} \leftarrow \mathbf{I}$
  for $b = c$ to 1 do
    if $[\mathbf{S} \cdot \mathbf{M}_b \boldsymbol{\theta}(0)]_f < 0.5$ then
      $\tau \leftarrow \tau + 2^b \Delta t$
      $\mathbf{S} \leftarrow \mathbf{S} \cdot \mathbf{M}_b$
    end if
  end for
  return Median first passage time $\tau$
```

Algorithm 2 Determine the basin or basins that each state from RNAsubopt is in. Note that for any list S having states of the same energy, the states assigned to the border by this program can depend on the ordering of these degeneracies. All states in S that are adjacent to multiple basins will be members of each basin, and these constitute the border states.

```
Require: List of states $S$ from RNAsubopt
  $Visited \leftarrow \emptyset$
  $Count \leftarrow 0$
  for all ($s \in S$) in order of ascending energy do
    $basins_s \leftarrow \emptyset$
    for all $v \in Visited$ do
      if $s$ is adjacent to $v$ by a single Kinfold move then
        $basins_s \leftarrow basins_s \cup \{\min(basins_v)\}$
      end if
    end for
    if $basins_s = \emptyset$ then
      $Count \leftarrow Count + 1$
      $basins_s \leftarrow \{Count\}$
    end if
    $Visited \leftarrow Visited \cup \{s\}$
  end for
  return Basin membership for each state $s$ is $basins_s$
```

Algorithm 3 Determine the partition functions for each basin and border state. Note that while the number of possible distinct basin states grows exponentially in the number of local minima, the results of this program are well suited for sparse storage since there cannot possibly be more local minima than states and so the program runs in linear time in the number of states if the list is initially sorted by basin membership.

```
Require: List of states $S$ labeled with their energies and basin membership from Algorithm 2.
Require: List of basin memberships $basins$ for each state $s \in S$.
Require: Temperature $T$, Boltzmann constant $k$
  while $S \neq \emptyset$ do
    $B \leftarrow basins_s$ for an arbitrary $s \in S$
    $pf_B \leftarrow 0$
    for all $s \in S$ with $basins_s = B$ do
      $pf_B \leftarrow pf_B + \exp(-s_{energy}/kT)$
      $S \leftarrow S \setminus s$
    end for
  end while
  return Partition function for each basin or border $B$ is $pf_B$.
```

Algorithm 4 Determine the Arrhenius barrier between an initial and final state. A system in the initial state will gradually spread to more basins until it reaches the full ensemble. Starting with an initial local minimum basin, find the highest free-energy barrier crossed before the final state's basin is reached. In order to find this barrier, we start with the local minimum basin of the starting state and repeatedly cross the lowest available free energy border to another local minimum basin, then merge all of the states in the new basin with the starting basin and repeat. We record the highest free energy barrier crossed in this procedure. A simple example of this procedure is diagrammed in Fig. 5.

```
Require: List of partition functions $pf$ indexed by the set of basins they represent, from
  Algorithm 3.
Require: $pf_s = 0$ if not otherwise specified by Algorithm 3.
Require: The set of sets of basins the partition functions $pf$ represent is $S$, or for each
  $pf_s \neq 0, s \in S$.
Require: Temperature $T$, Boltzmann constant $k$, initial basin $i$ and final basin $f$.
  $highest \leftarrow -\infty$
  while $\{f\} \in S$ do
    $pf_{lowest} \leftarrow 0$
    $pf_{basin} \leftarrow pf_{\{i\}}$.
    for all basins $b \in \bigcup_{j \in S} j$ do
      $pf_{border} \leftarrow 0$
      for all partition functions $q = pf_s$ where $s \ni i$ and $s \ni b$ do
        $pf_{border} \leftarrow pf_{border} + q$
      end for
      if $pf_{border} > lowest$ then
        $pf_{lowest} \leftarrow pf_{border}$
        $merge \leftarrow b$
      end if
    end for
    if $pf_{lowest} = 0$ then
      return Undefined barrier. No states connecting $i$ and $f$ exist.
    end if
    if $highest < -kT \ln(pf_{border}) + kT \ln(pf_{basin})$ then
      $highest \leftarrow -kT \ln(pf_{border}) + kT \ln(pf_{basin})$
    end if
    for all sets of basins $s \in S$ do
      if $merge \in s$ then
        $pf_{(s \setminus \{merge\}) \cup \{i\}} \leftarrow pf_{(s \setminus \{merge\}) \cup \{i\}} + pf_s$
        $pf_s \leftarrow 0$
        $S \leftarrow S \setminus s$
        $S \leftarrow S \cup \{((s \setminus \{merge\}) \cup \{i\})\}$
      end if
    end for
  end while
  return $highest$
```

References

1. Wales, D.J.: Energy Landscapes. Cambridge University Press, Cambridge (2003)
2. Levinthal, C.: J. Chim. Phys. **65**, 44 (1968)
3. Horwich, A.L., Fenton, W.A., Chapman, E., Farr, G.W.: Annu. Rev. Cell Dev. Biol. **23**, 115 (2007)
4. Prusiner, S.B.: Proc. Natl. Acad. Sci. USA **95**(23), 13363 (1998)
5. Selkoe, D.J.: Nat. Cell. Biol. **6**, 1054 (2004)
6. Bartel, D.P.: Cell **116**, 281 (2004)
7. Tucker, B.J., Breaker, R.R.: Curr. Opin. Struct. Biol. **15**(3), 342 (2005)
8. Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., Bartel, D.P.: Mol. Cell **27**, 91 (2007)
9. Schwartz, A., Rahmouuni, A.R., Boudvillain, M.: EMBO J. **22**, 3385 (2003)
10. Wickiser, J., Winkler, W., Breaker, R., Crothers, D.: Mol. Cell **18**, 49 (2005)
11. Mandal, M., Breaker, R.R.B.: Nat. Struct. Mol. Biol. **11**, 29 (2003)
12. Wilson, K.S., von Hippel, P.H.: Proc. Natl. Acad. Sci. USA **92**(19), 8793 (1995)
13. Wong, T.N., Sosnick, T.R., Pan, T.: Proc. Natl. Acad. Sci. USA **104**(46), 17995 (2007)
14. Sudarsan, N., Wickiser, J.K., Nakamura, S., Ebert, M.S., Breaker, R.R.: Genes Dev. **17**(21), 2688 (2003)
15. Mansilla, M.C., Albanesi, D., de Mendoza, D.: J. Bacteriol. **182**(20), 5885 (2000)
16. Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E., Breaker, R.R.: Nat. Struct. Biol. **10**, 701 (2003)
17. Batey, R.T., Gilbert, S.D., Montange, R.K.: Nature **432**, 411 (2004)
18. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S.: J. Biol. Chem. **277**(50), 48949 (2002)
19. Mandal, M., Lee, M., Barrick, J.E., Weinberg, Z., Emilsson, G.M., Ruzzo, W.L., Breaker, R.R.: Science **306**(5694), 275 (2004)
20. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S.: Nucleic Acids Res. **31**(23), 6748 (2003)
21. Roth, A., Winkler, W.C., Regulski, E.E., Lee, B.W.K., Lim, J., Jona, I., Barrick, J.E., Ritwik, A., Kim, J.N., Welz, R., Iwata-Reuyl, D., Breaker, R.R.: Nat. Struct. Mol. Biol. **14**, 308 (2007)
22. Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H.: J. Mol. Biol. **288**, 9218 (1999)
23. Schmitt, W.R., Waterman, M.S.: Discrete Appl. Math. **51**, 317 (1994)
24. Zuker, M., Stiegler, P.: Nucleic Acids Res. **9**(1), 133 (1981)
25. McCaskill, J.S.: Biopolymers **29**, 1105 (1990)
26. Kramers, H.A.: Physica **7**, 284 (1940)
27. Langer, J.S.: Ann. Phys. **54**, 258 (1969)
28. Flamm, C., Fontana, W., Hofacker, I.L., Schuster, P.: RNA **6**, 325 (2000)
29. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P.: Monatsh. Chem. **125**, 167 (1994)
30. Huber, G.A., Kim, S.: Biophys. J. **70**(1), 97 (1996)
31. Zhang, B., Jasnow, D., Zuckerman, D.M.: J. Chem. Phys. **132**, 054107 (2010)
32. Staple, D.W., Butcher, S.E.: PLoS Biol. **3**, e213 (2005)
33. Wuchty, S., Fontana, W., Hofacker, I., Schuster, P.: Biopolymers **49**, 145 (1999)
34. Flamm, C., Hofacker, I., Stadler, P., Wolfinger, M.: Z. Phys. Chem. **216**, 155 (2002)
35. Bonnet, G., Krichevsky, O., Libchaber, A.: Proc. Natl. Acad. Sci. USA **95**(15), 8602 (1998)
36. Lorenz, R., Flamm, C., Hofacker, I.L.: In: GCB 2009. Lecture Notes in Informatik, vol. 157, pp. 11–20 (2009)