Statistics of RNA Melting Kinetics

By

Manfred Tacker[a], Walter Fontana[b],
Peter F. Stadler[a,b] and Peter Schuster[a,b,c,*]

[a] Institut für Theoretische Chemie, Universität Wien
[b] Santa Fe Institute, Santa Fe
[c] Institut für Molekulare Biotechnologie, Jena

[*] Correspondence to:
Peter Schuster, Institut für Molekulare Biotechnologie
Beutenbergstraße 11, PF 100813, D-07708 Jena, Germany
Phone: ++49 (3641) 85 22 00
Fax: ++49 (3641) 33 13 25
Email: pks@mendel.imb-jena.de

Abstract

We present and study the behavior of a simple kinetic model for the melting of RNA secondary structures, given that those structures are known. The model is then used as a map that assigns structure dependent overall rate constants of melting (or refolding) to a sequence. This induces a "landscape" of reaction rates, or activation energies, over the space of sequences with fixed length. We study the distribution and the correlation structure of these activation energies.

1. Introduction

Single stranded RNA sequences fold into complex three-dimensional structures. A tractable, yet reasonable, model for the map from sequences to structures considers a more coarse grained level of resolution known as the secondary structure. The secondary structure is a list of base pairs such that no pairings occur between bases located in different loop regions. Algorithms based on empirical energy data have been developed to compute the minimum free energy secondary structure of an RNA sequence (Zuker and Stiegler 1981, Zuker and Sankoff 1984).

The structure influences a variety of biophysical quantities such as, for example, kinetic rate constants of melting. The situation is a composition of two mappings: a folding map that assigns a structure to a sequence, and a second map that assigns some biophysical property to the structure. By "landscape" we refer to the graph of this composite mapping that assigns to each sequence a scalar that quantifies some property based on the structure attained by the sequence.

Systematic studies on RNA landscapes (Fontana *et al.*1991, 1993a,b, Bonhoeffer *et al.*1993) were encouraged by the recent interest in statistical properties of randomly assembled RNA molecules which are used by several experimental techniques in applied molecular evolution (Horowitz *et al.*1989, Joyce 1989, Tuerk and Gold 1990, Ellington and Szostak 1990, Beaudry and Joyce 1992). The only class of value landscapes that is presently accessible to extensive computer explorations is derived from RNA secondary structures (Fontana *et al.* 1991, 1993a,b). In this case fairly reliable prediction algorithms are available (Jaeger *et al.* 1989, Zuker 1989). Computations of RNA tertiary structures or protein structures are both too expensive and not sufficiently accurate in order to allow systematic studies.

In recent work we have investigated essentially two kinds of landscapes: the scalar landscape of the free energy of folding, and the non-scalar landscape of the minimum free energy secondary structures themselves (Fontana, *et al.* 1991, 1993a,b). In this contribution we look into a further biophysical property that is mediated by the structure: the kinetics of melting. To this end we build a very simple kinetic model of melting (and, conversely, re-folding) of RNA secondary structures. The model is then used as a specification to compute for any given sequence overall rate constants of melting and structure formation, or, equivalently, to compute the corresponding activation energies. It is the landscape of these rates, more precisely: of the activation energies, that we investigate in this paper. It follows, in principle, the same generic concept mentioned recently in a short note (Fernández and Shakhnovich 1990), in which, however, no explicit sequence dependence was discussed.

We also consider three different alphabets: the binary **GC** alphabet, the biophysical **AUGC** alphabet and the synthetic **GCXK** alphabet. **A**,**U**,**G** and **C** denote the naturally occuring bases adenine, uracil, guanine and cytosine. **K** and **X** are abbreviations for 3-β-D-ribofuranosyl-(2,6-diaminopyrimidine) and xanthosine, a derivative of purine. The base pair between **K** and **X** has roughly the same energy as the base pair between **G** and **C** (Piccirilli *et al.* 1990). Due to the lack of experimental data we used the **GC** parameter set for **XK**. The synthetic **GCXK** alphabet represents an alphabet with two equally strong base pairs. It is of particular interest since it allows to study the effects of two base pairs versus one base pair (**GC**) without obscuring the results by the additional influence of different base pair strength and (**GU**) base pairing as in the natural **AUGC** alphabet.

2. RNA Melting Kinetics

2.1. A Model for Melting RNA Secondary Structures

An RNA secondary structure is a list of base pairs (i, j) with $i < j$ and satisfying two conditions: (1) each base is involved in at most one pairing interaction, and (2) if (i, j) and (k, l) are pairs, then $i < k < l < j$ or $k < i < j < l$ (Waterman and Smith 1978). Such a list can be visualized as a planar graph with two major structural elements: loops, or unpaired regions, and stacks of contiguous base pairs, also referred to as stacks. The condition (2) states that bases within a loop cannot pair with bases outside that loop. Secondary structures can be assigned free energies by summing up the (measured) energy contributions from their constituent loops and stacks. In the following the term "secondary structure" refers to the minimum free energy secondary structure.

Melting of RNA secondary structures is considered as an "all-or-none" process in the sense that only the completely folded minimum free energy secondary structure, **S**, and the open chain, **C**, are assumed to be present at measurable concentrations. According to the common stationarity assumption of chemical reaction kinetics the concentrations of all intermediates are considered as small and constant in time. The over-all process is then described by the following reaction equation:

$$\mathbf{C} \quad \underset{k_D}{\overset{k_R}{\rightleftharpoons}} \quad \mathbf{S} \tag{1}$$

Consistently with the current biophysical literature (Pörschke 1974) we use the terms "recombination" rate constant, k_R, and "dissociation" rate constant, k_D, when referring to the refolding and melting process, respectively. These terms were originally used in studies of double helix formation from separate RNA strands (Pörschke and Eigen 1971, Pörschke 1971).

Consider a minimum free energy secondary structure $\mathbf{S}$ consisting of n double helical stacks denoted by $\eta_1, \eta_2, \ldots, \eta_n$. We obtain $\mathbf{S}$ from a sequence by using a variant of the Zuker-Stiegler algorithm (Zuker and Stiegler 1981, Zuker and Sankoff 1984). We further assume that the stacks η_i, $i = 1, \ldots, n$, form and melt independently of one another, and that the only intermediates between $\mathbf{C}$ and $\mathbf{S}$ are the partially folded structures containing any of the stacks η_i. We do not consider intermediates consisting of other structural elements. Accordingly, all intermediates are encoded as a binary string $(s_1 s_2 \cdots s_n)$ of length n with $s_i = \mathbf{h}$ or $s_i = \mathbf{c}$ indicating whether the ith stacking region is in double helical form or in the molten state, respectively. The complete secondary structure $\mathbf{S}$ is expressed as $(\mathbf{hhh}\ldots\mathbf{h})$, and the open chain $\mathbf{C}$ is represented by $(\mathbf{ccc}\ldots\mathbf{c})$. The system, therefore, consists of 2^n different states: $\mathbf{S}$, $\mathbf{C}$ and the $2^n - 2$ intermediates encoded as

$$
\begin{array}{llll}
\mathbf{H}_1 = (\mathbf{hcc}\ldots\mathbf{c}) & \mathbf{H}_2 = (\mathbf{chc}\ldots\mathbf{c}) & \ldots & \mathbf{H}_n = (\mathbf{cc}\ldots\mathbf{ch}) \\
\mathbf{H}_{12} = (\mathbf{hhc}\ldots\mathbf{c}) & \mathbf{H}_{13} = (\mathbf{hch}c\ldots\mathbf{c}) & \ldots & \mathbf{H}_{n-1,n} = (\mathbf{c}\ldots\mathbf{chh}) \\
\vdots & & \ldots & \vdots \\
\mathbf{H}_{12\ldots n-1} = (\mathbf{hh}\ldots\mathbf{hc}) & \mathbf{H}_{13\ldots n} = (\mathbf{hch}\ldots\mathbf{h}) & \ldots & \mathbf{H}_{23\ldots n} = (\mathbf{chh}\ldots\mathbf{h})
\end{array}
$$

We retain the symbols $\mathbf{C} \equiv \mathbf{H}_0$ for the open chain and $\mathbf{S} \equiv \mathbf{H}_{12\ldots n}$ for the folded molecule. The individual structures can be thought of as occupying the corners of a Boolean hypercube of dimension n (Figure 1).

Expressions for overall interconversion $\mathbf{C} \rightleftharpoons \mathbf{S}$ can be derived from the general reaction mechanism which allows for interconversion of any two species. The kinetic equations are given by $[\mathbf{C}] = c$, $[\mathbf{H}] = h$ and $[\mathbf{H}_\mu] = h_\mu$, $(\mu = 1, 2, \ldots, 23 \ldots n)$. Let $k_{\mu\rho}$ be the (first order) rate contant for the conversion of the intermediate μ into ρ

$$\begin{aligned}
\dot{c} &= \sum_\rho (k_{\rho\mathbf{C}} h_\rho - k_{\mathbf{C}\rho} c) \\
\dot{h}_\mu &= \sum_{\rho \neq \mathbf{C},\mathbf{S}} (k_{\rho\mu} h_\rho - k_{\mu\rho} h_\mu) - (k_{\mu\mathbf{C}} + k_{\mu\mathbf{S}}) h_\mu + k_{\mathbf{C}\mu} c + k_{\mathbf{S}\mu} s, \\
\dot{s} &= \sum_\rho (k_{\rho\mathbf{S}} h_\rho - k_{\mathbf{S}\rho} s)
\end{aligned} \tag{2}$$

with $k_{\mu\mu=0}$. This may be rewritten in matrix form as

$$\dot{\vec{h}} = (-A) \cdot \vec{h} + c \cdot \vec{p} + s \cdot \vec{q}, \tag{3}$$

where $\vec{h} = (h_\mu)$ is the concentration vector of the intermediates, $\vec{p}$ is the vector of the $k_{\mathbf{C}\mu}$, $\vec{p} = (k_{\mathbf{C}1}, \ldots, k_{\mathbf{C}12\ldots n-1})$ and $\vec{q} = (k_{\mathbf{S}1}, \ldots, k_{\mathbf{S}12\ldots n-1})$. The matrix entries of A are given by

$$a_{\mu\rho} = -k_{\rho\mu} + [\sum_\tau k_{\mu\tau} + k_{\mu\mathbf{C}} + k_{\mu\mathbf{S}}]\delta_{\rho\mu}, \qquad (4).$$

The steady state assumption for intermediates,

$$\dot{h}_\mu = 0 \quad \forall \mu \neq \mathbf{C}, \text{bf S} \tag{5}$$

allows to compute the stationary concentrations $\bar{h}_k$ of all partially folded structures from the system of linear equations:

$$A \cdot \bar{h} = c \cdot \vec{p} + s \cdot \vec{q} \tag{6}$$

If the system of monomolecular reactions is strongly connected, that is: if there is a reaction pathway with positive rate constants between any two species, then A is invertible, and the concentrations of the intermediates are uniquely determined by

$$\bar{h} = (A^{-1}\vec{p})c + (A^{-1}\vec{q})s. \tag{7}$$

This is also an immediate consequence of Feinberg's deficiency 0 theorem (Feinberg 1977). From the equation for c,

$$\dot{c} = k_D \cdot s - k_R \cdot c = \sum_{\mu \neq \mathbf{C},\mathbf{S}} k_{\mu\mathbf{C}} \bar{h}_\mu - \sum_{\mu \neq \mathbf{C},\mathbf{S}} k_{\mathbf{C}\mu} c + k_{\mathbf{SC}} s - k_{\mathbf{CS}} c,$$

we obtain the explicit expressions for the overall rate constants as

$$\begin{aligned} k_R &= \vec{1}\vec{p} + k_{\mathbf{CS}} - \vec{k}_c A^{-1} \vec{p} \\ k_D &= \vec{k}_c A^{-1} \vec{q} + k_{\mathbf{SC}} \end{aligned} \tag{8}$$

where $\vec{k}_c$ is the vector $k_{\mu\mathbf{C}}$ and $\vec{1}$ is the vector $(1, \ldots, 1)$.

2.2. Sequential Melting

The case of sequential folding and melting of the secondary structure $\mathbf{S}$ along an unbranched dominant path

$$\mathbf{C} \underset{k_{10}}{\overset{k_{01}}{\rightleftharpoons}} \mathbf{H}_1 \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} \mathbf{H}_{12} \underset{k_{32}}{\overset{k_{23}}{\rightleftharpoons}} \mathbf{H}_{123} \underset{k_{43}}{\overset{k_{34}}{\rightleftharpoons}} \ \ldots \ \underset{k_{m-1,m-2}}{\overset{k_{m-2,m-1}}{\rightleftharpoons}} \mathbf{H}_{12\ldots m-1} \underset{k_{m,m-1}}{\overset{k_{m-1,m}}{\rightleftharpoons}} \mathbf{S} \tag{9}$$

can be readily solved analytically. The intermediates are ordered along this path such that the stack η_1 forms most easily, η_2 forms next likely, etc. Formation of the stack η_m is least likely. Equivalently, the stack η_m melts most easily, etc. All other intermediates except the $m-1$ explicitly shown in the reaction scheme (9) are neglected.

Let us now denote the concentrations of the intermediates by: $[\mathbf{H}_1] = h_1$, $[\mathbf{H}_{12}] = h_2$, ... , $[\mathbf{H}_{12\ldots m-1}] = h_{m-1}$, $[\mathbf{C}] = c$ and $[\mathbf{S}] = s$. The kinetic equations are straightforward:

$$\begin{aligned} \dot{c}_1 &= -k_{01} c + k_{10} h_1 \\ \dot{h}_1 &= k_{01} c - (k_{10} + k_{12}) h_1 + k_{21} h_2 \\ \dot{h}_2 &= k_{12} h_1 - (k_{21} + k_{23}) h_2 + k_{32} h_3 \\ &\ \ \vdots \\ \dot{h}_{m-1} &= k_{m-2,m-1} h_{m-2} - (k_{m-1,m-2} + k_{m-1,m}) h_{m-1} + k_{m,m-1} st \\ \dot{s} &= -k_{m,m-1} s + k_{m-1,m} h_{m-1}. \end{aligned} \tag{10}$$

The matrix A in equation (3) becomes

$$A = \begin{pmatrix} k_{10}+k_{12} & -k_{21} & 0 & \dots & 0 \\ -k_{12} & k_{21}+k_{23} & -k_{32} & \dots & 0 \\ 0 & -k_{23} & k_{32}+k_{34} & \dots & 0 \\ 0 & 0 & -k_{34} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -k_{m-1,m-2} \\ 0 & 0 & 0 & \dots & k_{m-1,m-2}+k_{m-1,m} \end{pmatrix} \tag{11}$$

and the vectors $\vec{p}$ and $\vec{q}$ have only one non-zero component $p_1 = k_{\mathbf{C}1}$ and $q_m = k_{\mathbf{S},m-1}$. The overall rate constants result as

$$\begin{aligned} k_R &= \frac{k_{01}\cdot k_{12}\cdot\ldots\cdot k_{m-1,m}}{|A|}, \\ k_D &= \frac{k_{10}\cdot k_{21}\cdot\ldots\cdot k_{m,m-1}}{|A|}. \end{aligned} \tag{12}$$

As a consequence of the assumption of independent substructure formation the overall equilibrium constant factorizes into equilibrium constants for individual stacks:

$$K = \frac{k_R}{k_D} = \frac{k_{01}}{k_{10}}\cdot\frac{k_{12}}{k_{21}}\cdot\ldots\cdot\frac{k_{m-1,m}}{k_{m,m-1}} \tag{13}$$

We shall make use of this fact in the computation of individual rate constants.

2.3. Hypercubic Melting

It is reasonable to assume that the only allowed reversible reactions are those that involve structures differing in the state of one stack. The total reaction network, then, is given by the edges of an n dimensional hypercube, that is: all $k_{\mu\rho}$ in equation (2) are zero except for intermediates whose binary encodings have Hamming distance $d_h = 1$. The explicit expressions for k_R and k_D become very unhandy with this mechanism. We restrict the analytical treatment to the most simple non-trivial case $n = 2$. For cases $n > 2$ we used equation (8) for numerical computation.

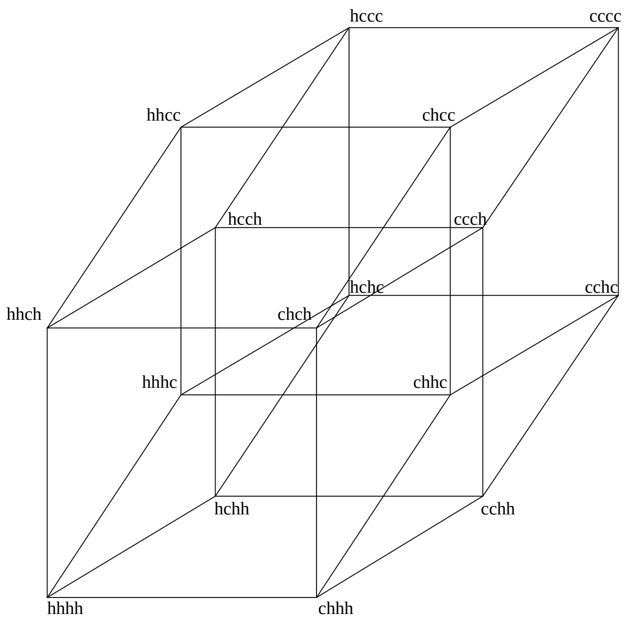


Figure 1: Stepwise formation and melting of an RNA secondary structure consisting of four helical regions. The unfolded molecule, **C**, the completed secondary structure, **S**, and all intermediates are represented by the corners of a hypercube. Edges correspond to reversible reactions involving the formation and the melting of a single stack.

Concentrations of intermediates are now denoted by $[\mathbf{H}_1] = [(\mathbf{ch})] = h_1$ and $[\mathbf{H}_2] = [(\mathbf{hc})] = h_2$, and the kinetic differential equations are given by

$$\begin{aligned}\dot{c} &= -(k_{01} + k_{02})c + k_{10}h_1 + k_{20}h_2 \\ \dot{h}_1 &= k_{01}c - (k_{10} + k_{13})h_1 + k_{31}s \\ \dot{h}_2 &= k_{02}c - (k_{20} + k_{23})h_2 + k_{32}s \\ \dot{s} &= k_{13}h_1 + k_{23}h_2 - (k_{31} + k_{32})s\end{aligned} \tag{14}$$

After elimination of h_1 and h_2 we find for the two over-all rate constants

$$\begin{aligned}k_R &= \frac{k_{01}k_{13}(k_{20} + k_{23}) + k_{02}k_{23}(k_{10} + k_{13})}{(k_{10} + k_{13})(k_{20} + k_{23})}, \\ k_D &= \frac{k_{10}k_{31}(k_{20} + k_{23}) + k_{20}k_{32}(k_{10} + k_{13})}{(k_{10} + k_{13})(k_{20} + k_{23})}\end{aligned} \tag{15}$$

which, of course, fulfill the equilibrium condition

$$K = \frac{k_R}{k_D} = \frac{k_{01}}{k_{10}} \cdot \frac{k_{13}}{k_{31}} = \frac{k_{02}}{k_{20}} \cdot \frac{k_{23}}{k_{32}} \tag{16}$$

3. Estimation of Rate Constants

Rate constants of a chemical reaction ξ are, in the simplest case, related to activation energies by means of an Arrhenius law

$$k_\xi = A_\xi \cdot \exp\left\{-\frac{\Delta G_\xi^\natural}{RT}\right\} \tag{17}$$

In order to actually compute the magnitude of the overall rate constants via equation (8), we need some assumptions concerning the individual rate constants $k_{\mu\rho}$.

Each reaction involving two partially folded structures must have an equilibrium constant

$$K_{\mu\rho} = \frac{k_{\mu\rho}}{k_{\rho\mu}} \tag{18}$$

The corresponding change in free energy, $\Delta G_{\mu\rho}$, is computed from the set of energy parameters used in the folding algorithm (Jaeger *et al.* 1989, Freier *et al.* 1986). For a calculation of the rate constants we partition the free energy into contributions resulting from both the folding and the melting process. This is done in accordance with experimental data reported by Pörschke (1974) for the folding kinetics of the oligoribonucleotide $\mathbf{A}_6\mathbf{C}_6\mathbf{U}_6$ into a hairpin helix. The individual contributions are split as follows.

Let μ and ρ be two structures which differ by a single stack which is present in ρ but molten in μ. Thus the relevant structural state in ρ consists of a stack of length s and the two flanking loops Λ_1 and Λ_2, of sizes λ_1 and λ_2, respectively. In μ this state turns into a large loop Λ of size $\lambda_1 + \lambda_2 + s$, according to the reaction path

$$\text{stack} \ + \ \Lambda_1 \ + \ \Lambda_2 \ \rightleftharpoons \ \Lambda_1 \ + \ \Lambda_2 \ \rightleftharpoons \ \Lambda, \tag{19}$$

with $\Lambda_1 \ + \ \Lambda_2$ as the transition state. Let the destabilizing (positive) energy of a loop l be $\Delta G(l)$. The contribution from eliminating loop Λ is calculated from the energy tables underlying the folding procedure as:

$$\Delta G_{\mu\rho}^{\text{loop}} = \Delta G(\Lambda_1) + \Delta G(\Lambda_2) - \Delta G(\Lambda) \tag{20}$$

(In case of a stack that does not close an internal loop both Λ_2 and Λ are joints or free ends, whose free energy contributions are set to zero.) The contribution from the stack formation is denoted by $\Delta G^{\rm stack}_{\mu\rho}$, and is directly read off the energy tables.

The key assumption in estimating the rate constants is that along the reaction path (19) the absolute value of the stacking energy enters the rate constant of melting as the free energy of activation, $\Delta G^{\natural}_D$:

$$k_{\rho\mu} = A_- \cdot \exp\left(-\frac{|\Delta G^{\rm stack}_{\rho-\mu}|}{RT}\right), \tag{21}$$

while the activation energy of recombination, $\Delta G^{\natural}_R$, is given by the change in loop energy:

$$k_{\mu\rho} = A_+ \cdot \exp\left(-\frac{\Delta G^{\rm loop}_{\rho-\mu}}{RT}\right). \tag{22}$$

The rate constants determined by Pörschke for a loop of six bases at room temperature, $k_D = 2.6 \times 10^4\,\mathrm{sec}^{-1}$ and $k_R = 2.4 \times 10^4\,\mathrm{sec}^{-1}$ ($T = 24.2\,^\circ$ C), serve as reference. The energy parameter set used here (Jaeger *et al.* 1989, Freier *et al.* 1986) yields $\Delta G^{\rm stack} = -5.2$kcal/mol and $\Delta G^{\rm loop} = 5.1$kcal/mol, and thus

$$A_+ = 1.35 \cdot 10^8 \qquad \text{and} \qquad A_- = 1.75 \cdot 10^8, \tag{23}$$

which agree within experimental error. We set $A = A_+ = A_-$. The effective overall activation energies then are simply given by

$$\Delta G^{\natural}_R = -RT \ln \frac{k_R}{A} \qquad \text{and} \qquad \Delta G^{\natural}_D = -RT \ln \frac{k_D}{A}. \tag{24}$$

Our calculations will refer to $\Delta G^{\natural}_R$ and $\Delta G^{\natural}_D$, since the over-all rate constants vary by more than 10 orders of magnitude (see also the appendix).

4. Computational Results

4.1. Dynamics

In this section we briefly discuss the melting dynamics by numerically integrating equations (2) for the case of $\mathrm{tRNA}^{\mathrm{Phe}}$ from *E.coli*. The secondary structure of $\mathrm{tRNA}^{\mathrm{Phe}}$ is shown in Figure 2a. It contains four stacks, labelled from 1 to 4 clockwise, with stack 4 being the acceptor stack. A configuration is a binary string of length four, with the leftmost position referring to stack 1 and the rightmost position to stack 4. Configuration **chch**, for example, indicates the presence of stacks 2 and 4, and the absence of stacks 1 and 3.

Figure 2b shows the dynamics at 37°C starting from the linear form. The two structures each consisting of one stack that closes a smallest hairpin loop, **hccc** and **chcc**, build up fastest. Formation of hairpin 3, **cchc**, is slightly slower, while **ccch**, leading to the largest loop, is negligible. Once a hairpin structure has formed, other hairpin components are added easier than the closing of stack 4, since it now involves the formation of a more destabilizing internal loop. Hence from **hccc**, **chcc** and **cchc** the folding process leads to **hhcc**, **chhc** and **hchc**. The closing of stack 4 is now even more delayed, since it would involve formation of a multiloop whose free energy of formation is less favorable than other loop structures. Hence **hhcc**, **chhc** and **hchc**, all contribute to the build-up of **hhhc**. Finally the fully developed secondary structure, **hhhh**, is formed by closing the multiloop with stack 4.

Figure 2c shows the melting curve of $\mathrm{tRNA}^{\mathrm{Phe}}$ as the temperature dependence of the specific heat C, where $C = dH/dT$ and $H = kT^2 \partial \ln Q / \partial T$ (McCaskill 1990). The partition function Q has been obtained using McCaskill's generalization of the dynamic programming algorithm for secondary structures (McCaskill 1990). The shoulder at about 45°C is due to the opening of the multiloop closing stack 4. The rest of the structure melts at 68°C. The kinetics of melting is shown in Figure 2d for a temperature of 80°C. The melting process begins with stack 4, and essentially reverses the sequence of the folding events. With rising temperature the stacking energy of stack 4 drops the most relative to the other stacks: it therefore opens first, even if it is the longest paired region. The temperature at which the linear structure is dominant for the first time, 71.5°C, agrees fairly well with the major peak at 68°C in the computed melting curve.

4.2. Comparison of Sequential and Hypercubic Mechanism

The hypercubic melting model, equation (3), is solved numerically to calculate k_R and k_D. The vectors $(A^{-1}\vec{p})$ and $(A^{-1}\vec{q})$ are obtained from the linear system $A\vec{h} = \vec{p}$ and $A\vec{h} = \vec{q}$ by a Gaussian elimination scheme based on Crout's algorithm for the LU decomposition (Press *et al.* 1988, p. 43).

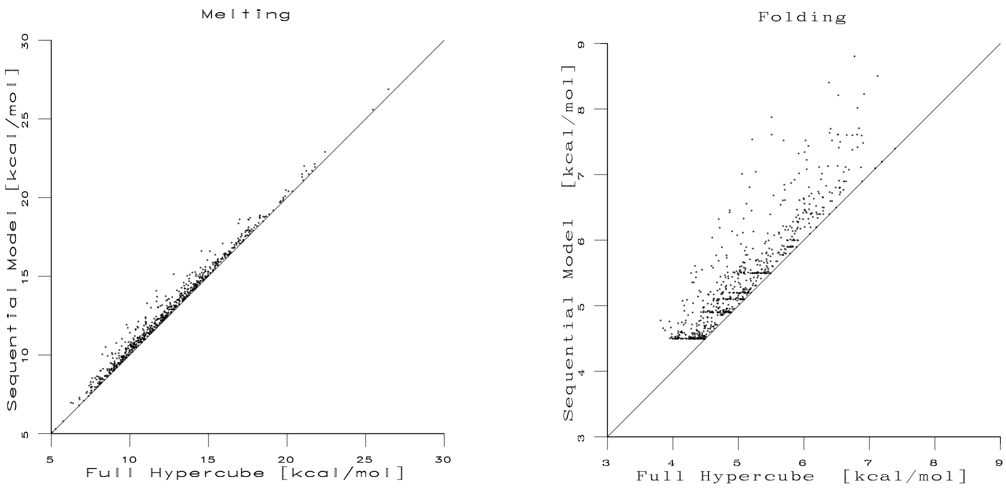
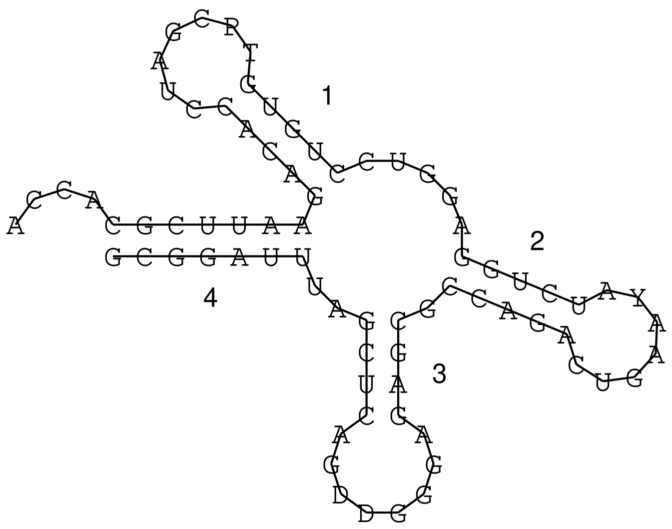


Figure 3: Comparison of effective activation energies $\Delta G_D^{\natural}$ (r.h.s.) and $\Delta G_R^{\natural}$ (l.h.s), as obtained from the sequential and the hypercubic melting mechanism.

The sequential model strictly underestimates the rate constants. A comparison of the effective activation energies (equation (23)), $\Delta G_D^{\natural}$ and $\Delta G_R^{\natural}$, calculated for a sample of random sequences with both the sequential and the hypercubic mechanism is shown in figure 3. Although the differences are



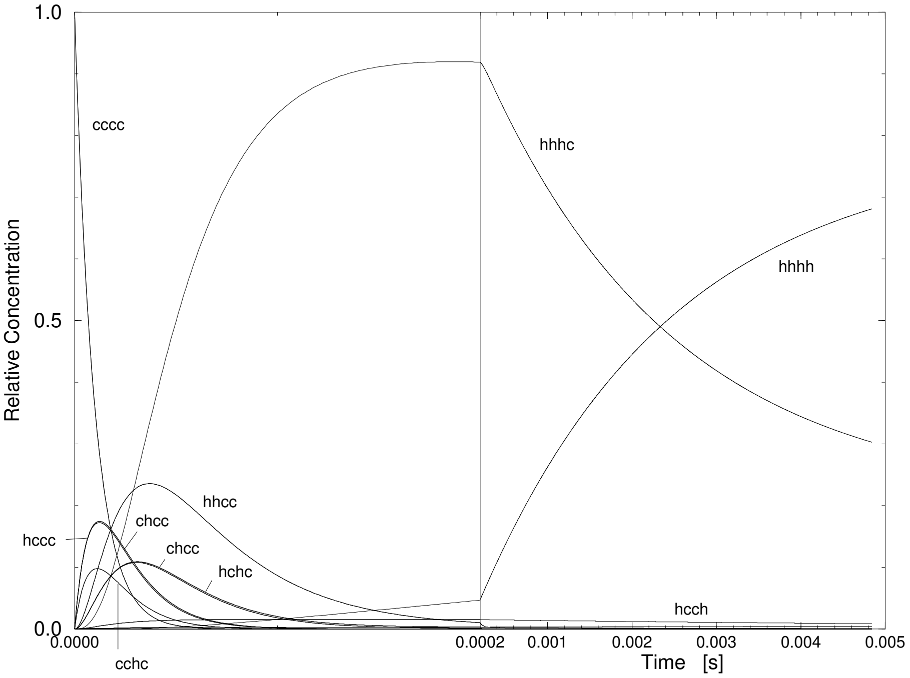


Figure 2: Dynamics of refolding and melting of tRNA$^{\mathrm{Phe}}$ from *E.coli*. (a) Minimum free energy structure. Each number indicates the position that refers to the corresponding stack in the binary encoding of the individual configurations. The leftmost position is 1, the rightmost is 4. (b) Refolding dynamics at 37°C. (c) Melting curve computed via the partition function. (d) Melting dynamics at 80°C. See text for details.

small, we chose the full hypercubic mechanism for all subsequent computations.

In the remaining sections we study the statistical structure of the landscapes that results from assigning to each sequence a $\Delta G^{\natural}_{D}$ and a $\Delta G^{\natural}_{R}$ at 37°C according to our model.

4.3. Distribution of Activation Energies

Figure 4 shows that the average of $\Delta G^{\natural}_{D}$ increases linearly with the chain length n. The single most important contribution to k_D is likely to come from the most stable stack. We know that the average size of a stack becomes rapidly a constant as sequences grow longer, while the average number of stacks increases linearly with n (Fontana *et al.* 1992b). Hence the linear increase in $\Delta G^{\natural}_{D}$.

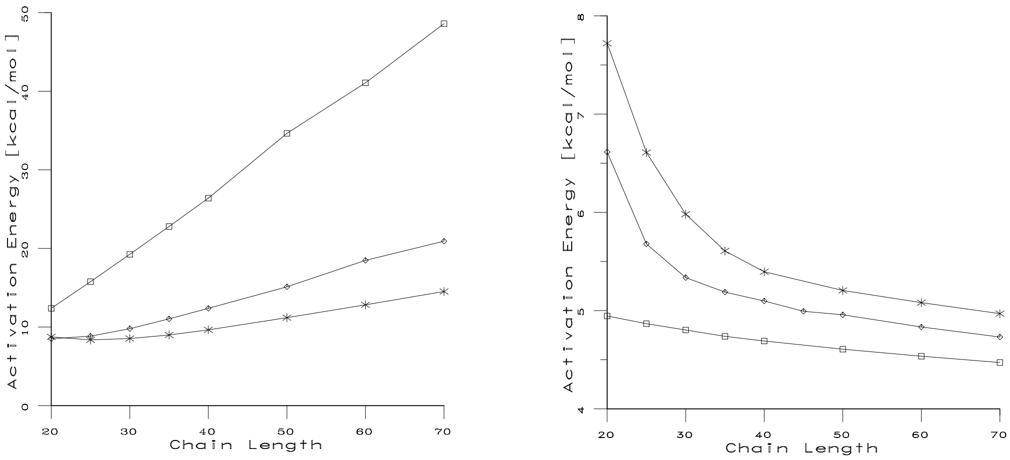


Figure 4: Average activation energies for the dissociation (l.h.s.) and the recombination reaction (r.h.s.) as a function of the chain length. □ **GC**, ⋆ **AUCG**, ♢ **GCXK**.

For a given length n, **GC** sequences form the most stable structures, **GCXK** sequences are less stable and **AUGC** sequences are the most unstable. This fact is reflected in the dependence of $\Delta G^{\natural}_{D}$ on the chosen alphabet.

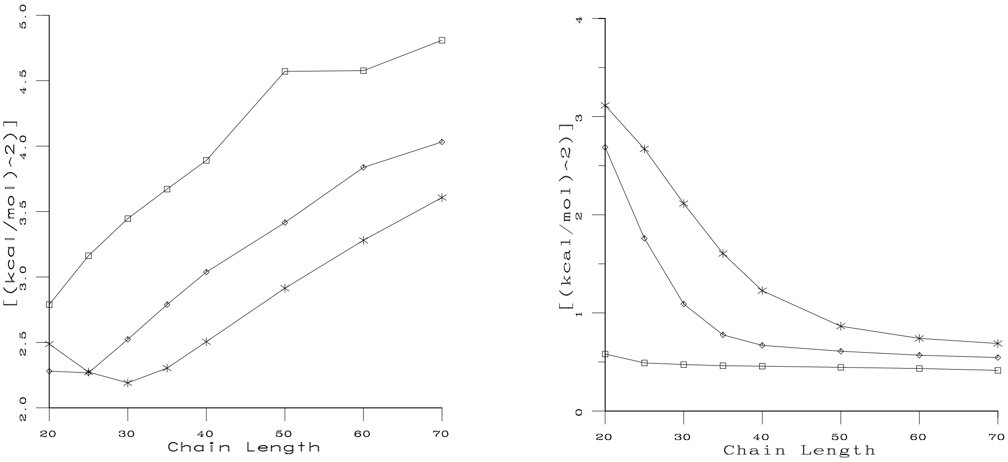


Figure 5: Variances of the activation energies □ **GC**, ⋆ **AUCG**, ♢ **GCXK**. L.h.s.: melting reaction, r.h.s.: recombination reaction.

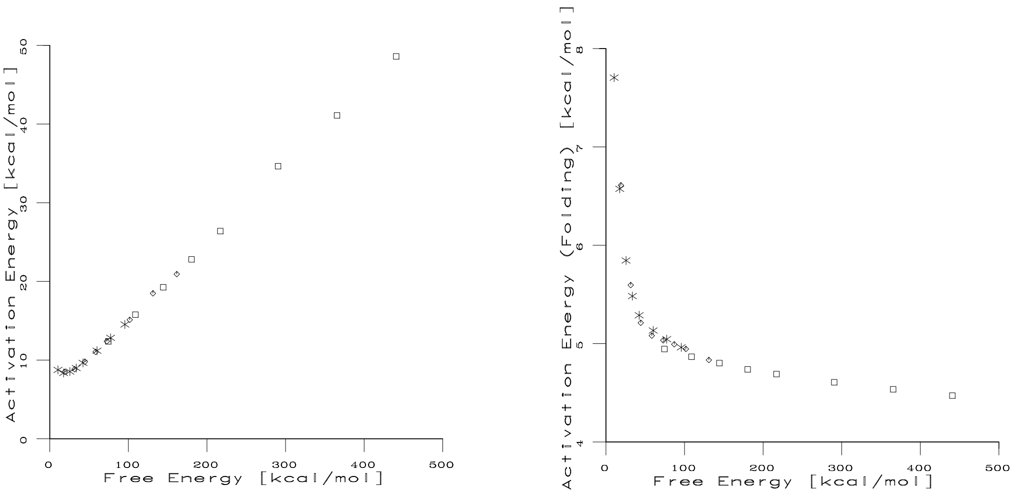
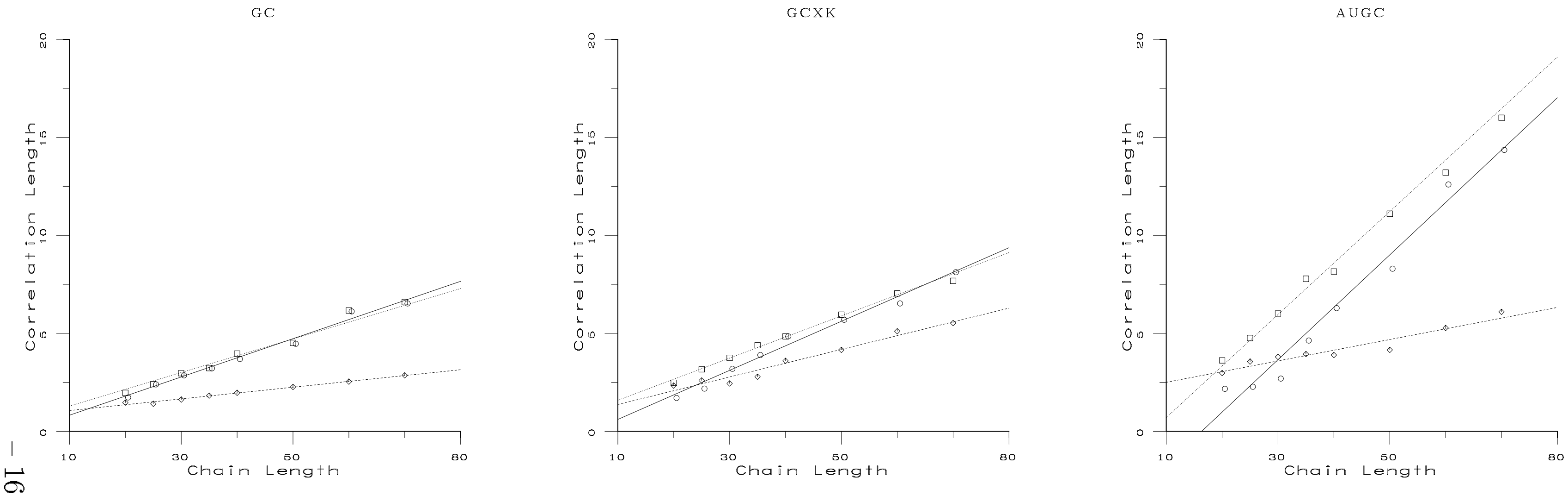


Figure 6: Activation energies as function of the free energy of folding for the three different alphabets □ **GC**, ⋆ **AUCG**, ♢ **GCXK**. L.h.s.: melting reaction, r.h.s.: recombination reaction.

For the **GC** alphabet the dissociation activation energies of longer chains are so large that the structures practically never open totally.

The average of $\Delta G^{\natural}_{R}$ (Figure 4) levels off towards a constant for longer chains. This suggests that the rate limiting step for structure formation is



GC **GCXK** **AUGC**

Figure 7: Chain length dependence of correlation length.
Free Energy ΔG°: $\square$ and *dotted line.*
Effective activation energy of melting $\Delta G^{\natural}_{D}$: $\circ$ and *full line.*
Effective activation energy of recombination $\Delta G^{\natural}_{R}$: $\diamondsuit$ and *dashed line.*

the closing of the first hairpin loop (and there may be several parallel and equivalent paths for this). The loop formation energies enter $\Delta G_R^\natural$, and the average loop size becomes constant for long chains (Fontana *et al.* 1992b). The alphabet dependence for the overall activation energy of recombination is reversed with respect to the melting (dissociation) case.

The distribution of activation energies is consistent with a Gaussian distribution for long chains (not shown). This behavior is to be expected since the individual structure elements contribute independently to the activation energies and the number of structure elements increases (linearly) with chain length (Fontana *et al.* 1992b). For all alphabets the variance of $\Delta G_R^\natural$ is decreasing with n, whereas the variance of $\Delta G_D^\natural$ is growing approximately linearly (figure 5). The distribution of free energies of activation for the recombination landscape, therefore, sharpens with increasing n.

The alphabet and chain length dependencies are exclusively due to the differences in the average structure stability they induce. This is seen in figure 6, where average activation energies of both recombination and dissociation are represented as a function of the average free energy of folding: the curves are independent of alphabet and chain length. The same holds for the variances (not shown).

4.4 Landscape correlation lengths

Landscapes that arise by assigning some property $f(x)$ to a configuration x can be characterized by their ruggedness (Kauffman and Levin 1987, Kauffman *et al.* 1988, Macken and Perelson 1989, Eigen *et al.* 1989, Weinberger 1990, 1991, Schuster 1991, Fontana *et al.* 1991, 1992a,b, Weinberger and Stadler 1992). This ruggedness can be quantified conveniently by means of a landscape correlation function

$$\rho(d) = \frac{\langle f(x)f(y)\rangle_{d(x,y)=d} - \langle f\rangle^2}{\langle f(p)f(q)\rangle_{\mathrm{random}} - \langle f\rangle^2}. \tag{25}$$

The averages $\langle\,.\,\rangle_{d(x,y)=d}$ refer to pairs of sequences with given Hamming distance d in sequence space, and $\langle\,.\,\rangle_{\mathrm{random}}$ refers to a pair of sequences which are chosen independently at random. It has proven useful (Fontana *et al.* 1991) to characterize the autocorrelation function by a "correlation length" ℓ defined by

$$\rho(\ell) = 1/e \tag{26}$$

although $\rho(d)$ is usually not a single decaying exponential.

The correlation lengths computed for both activation energy landscapes increase linearly with chain length n (see figure 7). The same behavior is observed for the correlation length of the free energy of structure formation (Fontana *et al.* 1992b). For all three alphabets the rate of increase is considerably smaller for the activation energy of recombination.

The **AUGC** alphabet has the highest correlation lengths for both activation energy landscapes; the **GC** alphabet has the smallest. This reflects the alphabet dependency in the stability of the secondary structures towards mutations in the underlying sequences. Independently of alphabet the correlation lengths of the $\Delta G_R^{\natural}$ landscapes are extremely short, indicating an almost uncorrelated landscape.

In agreement with figure 6 the $\Delta G_R^{\natural}$ landscape has a very weak dependence on chain length, suggesting once more that structure formation according to this model depends mostly on the first stack formation event.

5. Conclusions

We have presented a simple model for the kinetic melting (and refolding) behavior of RNA secondary structures. The model does not formalize a kinetic process of folding, or melting, that occurs on the space of all possible secondary structure configurations that are accessible to a given sequence. It rather assumes that the secondary structure is known beforehand by an independent calculation. It then considers only those processes that involve those structural elements that constitute the minimum free energy secondary structure. Nevertheless, in the case of $tRNA^{Phe}$ the melting point, as determined by the kinetic model, agrees remarkably well with an independently calculated melting curve that takes into account the entire equilibrium ensemble of structures for the sequence (Computations according to McCaskill 1990 and Bonhoeffer *et al.*, 1993).

The average values and variances of the activation energies of both melting and refolding are seen to depend only on the average energy of the structures, independently of the alphabet used to build the sequences and independently of their chain length.

The average overall activation energy of refolding saturates to a constant, and its distribution sharpens for longer chains. This indicates that the rate limiting step is the formation of the first stack from the linear structure.

The correlation lengths for both activation energy landscapes scale linearly with chain length for all alphabets. The detailed scaling behavior of the "melting landscape" closely resembles the landscape of the free energy of structure formation investigated in earlier work. The correlation structure of the "recombination landscape" varies weakly with chain length, in agreement with the previous result indicating a nucleation event.

Acknowledgements

Useful discussions with Dipl. Phys. Ivo Hofacker and Dipl Chem. Erich Bornberg-Bauer are gratefully acknowledged. This work was supported financially by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung, Project Nos. S5305-PHY and P8526-MOB.

Appendix

In this appendix we motivate formally why we consider the landscape of activation energies rather than directly the landscape of rate constants. The informal reason is obvious: due to the exponential relationship between activation energy and rate constants, the landscape of the latter will appear practically uncorrelated.

For the sake of simplicity we will assume, that the pre-exponential factor A_ξ in equation (17) does not depend on the RNA sequence. We further assume that the $\Delta G^\natural$ landscape has a Gaussian distribution with correlation coefficients

$$\rho(x,y) = \frac{\langle \Delta G^\natural(x) \Delta G^\natural(y) \rangle - \langle \Delta G^\natural \rangle^2}{\langle (\Delta G^\natural)^2 \rangle - \langle \Delta G^\natural \rangle^2} \qquad (A1)$$

where x and y denote fixed-length sequences. In view of our results in section 4.3 the Gaussian assumption is justified. In order to facilitate the subsequent calculations we will use the dimensionless variables

$$g_x = \frac{\Delta G^\natural(x) - \langle \Delta G^\natural \rangle}{\sqrt{\langle (\Delta G^\natural)^2 \rangle - \langle \Delta G^\natural \rangle^2}} \qquad (A2)$$

The correlation coefficient remains unchanged by this linear transformation. From the assumption of a Gaussian distribution we obtain for the joint distribution

$$p(g_x, g_y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \cdot \left(g_x^2 - 2\rho g_x g_y - g_y^2 \right) \right\} \qquad (A3)$$

for any two sequences x and y (using the abbreviation $\rho = \rho(x,y)$), and

$$p(g) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} g^2). \qquad (A4)$$

The correlation coefficient $\hat{\rho}(x,y)$ of the corresponding rate constants is given by

$$\hat{\rho}(x,y) = \frac{\langle k_x k_y \rangle - \langle k \rangle^2}{\langle k^2 \rangle - \langle k \rangle^2} \qquad (A5)$$

This expression is evaluated by considering that the rate constant can be expressed as

$$
\begin{aligned}
k_x &= B \cdot e^{q \cdot g_x} \qquad \text{with} \\
B &= A \cdot e^{-\frac{\langle \Delta G^\natural \rangle}{RT}}, \\
q &= -\frac{\sqrt{\langle (\Delta G^\natural)^2 \rangle - \langle \Delta G^\natural \rangle^2}}{RT}
\end{aligned}
\tag{A6}
$$

From the Gaussian integrals

$$
\begin{aligned}
\langle k \rangle &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} g^2) \cdot B e^{qg} dg = B e^{q^2/2} = B\sqrt{\alpha} \\
\langle k^2 \rangle &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} g^2) \cdot B^2 (e^{qg})^2 dg = B^2 e^{2q^2} = B^2 \alpha^2 (A7) \\
\langle k_x k_y \rangle &= \frac{B^2}{2\pi\sqrt{1-\rho^2}} \int\int_{-\infty}^{+\infty} \exp\left\{ -\frac{g_x^2 - 2\rho g_x g_y - g_y^2}{2(1-\rho^2)} \right\} e^{q(g_x+g_y)} dg_x dg_y \\
&= B^2 e^{q^2(1+\rho)} = B^2 \alpha^{1+\rho(x,y)}
\end{aligned}
$$

and with the abbreviation

$$
\alpha = e^{q^2} = \exp\left\{ \frac{\langle (\Delta G^\natural)^2 \rangle - \langle \Delta G^\natural \rangle^2}{(RT)^2} \right\} \tag{A8}
$$

we obtain

$$
\hat{\rho}_\alpha(x,y) = \frac{\alpha^{\rho(x,y)} - 1}{\alpha - 1} \tag{A9}
$$

The function $f(t) = \frac{\alpha^t - 1}{a-1}$ is concave for all t and all $\alpha > 1$, i.e. all $T > 0$, and furthermore $f(0) = 0$ and $f(1) = 1$. It follows immediately that

$$
|\rho(x,y)| > |\hat{\rho}_\alpha(x,y)| \tag{A10}
$$

for all $T > 0$ whenever $\rho(x,y) \neq \hat{\rho}(x,y) = 0$ or 1.

Let us assume that $\rho(x,y)$ is only a function of the Hamming distance $d = d_H(x,y)$ between sequences x and y. Suppose $\rho(d)$ has a representation

$$
\rho(d) = 1 - \frac{1}{\ell} d + o(d) \tag{A11}
$$

where ℓ may be interpreted as a characteristic length scale. A simple calculation then shows that $\hat{\rho}(d)$ may be represented as

$$\hat{\rho}_\alpha(d) = 1 - \frac{1}{\hat{\ell}_\alpha} d + o(d) \tag{A12}$$

with the characteristic length scale on the landscape of rate constants given by

$$\hat{\ell}_\alpha = \frac{\alpha - 1}{\alpha \log \alpha} \cdot \ell \tag{A13}$$

For large n the parameter α increases exponentially because

$$\langle (\Delta G^\natural)^2 \rangle - \langle \Delta G^\natural \rangle^2 \ \sim\ s^2 \cdot n \tag{A14}$$

and thus $\frac{\alpha-1}{\alpha} \to 1$. We finally obtain

$$\hat{\ell}_\alpha \sim \left(\frac{RT}{s} \right)^2 \cdot \frac{\ell}{n} \tag{A15}$$

In view of the results of section 4.4, the correlation lengths for the activation energy landscapes considered here certainly obey the relation

$$\ell \leq \mathcal{O}(n) \tag{A16}$$

and, therefore, we expect that

$$\hat{\ell}_\alpha \leq \mathcal{O}(1) \tag{A17}$$

for any positive temperature T. We conclude that landscapes of reaction rate constants become essentially uncorrelated for long sequences.

References

[1] Beaudry A.A. and G.F. Joyce. Directed Evolution of an RNA Enzyme. *Science* **257**, 635-641.

[2] Bonhoeffer S., J.S. McCaskill, P.F. Stadler and P. Schuster. Temperature Dependent RNA Landscapes, A Study based on Partition Functions. *Eur. Biophys. J.* **22**, 13-24.

[3] Eigen M. Selforganization of Matter and the Evolution of Macromolecules. *Naturwissenschaften* **10**, 465-523 (1971).

[4] Eigen M., J.S. McCaskill and P. Schuster. The Molecular Quasispecies. *Adv. Chem. Phys.* **75**, 149-263 (1989).

[5] Ellington A.D. and J.W. Szostak. *In vitro* selection of RNA molecules that bind specific ligands.*Nature* **346**,818-822 (1990).

[6] Feinberg M. Mathematical Aspects of Mass Action Kinetics. In: Lapidus L and Amundson N.R. (eds): *Chemical Reactor Theory*, Eaglewood Cliffs, NJ, 1-78 (1977)

[7] Fernández A. and E.I. Shakhnovich. Activation-energy landscape for metastable RNA folding. *Phys.Rev.A* **42**, 3657-3659 (1990)

[8] Fontana W. and P. Schuster. A Computer Model of Evolutionary Optimization. *Biophys. Chem.* **26**, 123-147 (1987).

[9] Fontana W., W. Schnabl and P. Schuster. Physical aspects of evolutionary optimization and adaptation. *Phys. Rev. A* **40**, 3301-3321 (1989).

[10] Fontana W., T. Griesmacher, W. Schnabl, P.F. Stadler and Peter Schuster. Statistics of Landscapes Based on Free Energies Replication and Degradation Rate Constants of RNA Secondary Structures. *Mh. Chem.* **122**, 795-819 (1991).

[11] Fontana W., D.A.M. Konings, P.F. Stadler and P. Schuster. Statistics of RNA Secondary Structures. *SFI preprint* 92-02-007, submitted to *Biopolymers* 1992a.

[12] Fontana W., P.F. Stadler, E.G. Bornberg-Bauer, T. Griesmacher, I.L. Hofacker, M. Tacker, P. Tarazona, E.D. Weinberger and P. Schuster. *Phys. Rev. E* **47**, 2083-2099 (1993b).

[13] Freier S.M., R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson and D.H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Biochemistry* **83**, 9373-9377 (1986).

[14] Hamming R.W. *Coding and Information Theory.* (2nd ed.) Prentice Hall, Englewood Cliffs, NJ, 1986.

[15] Horowitz M.S.Z., D.K. Dube and L.A. Loeb. Selection of new biological activities from random nucleotide sequences *Genome* **31**, 112-117 (1989)

[16] Jaeger J.A., D.H. Turner and M. Zuker. Improved predictions of secondary structures for RNA. *Biochemistry* **86**, 7706-7710 (1989).

[17] Joyce G.F. Amplification, mutation and selection of catalytic RNA. *Gene* **82**,83-87 (1989)

[18] Kauffman, S.A. and S. Levin. Towards a General Theory of Adaptive Walks on Rugged Landscapes, *J. theor. Biol.* **128**, 11-45 (1987).

[19] Kauffman, S.A., E.D. Weinberger, and A.S. Perelson. Maturation of the Immune Response Via Adaptive Walks On Affinity Landscapes, *Theoretical Immunology, Part I* Santa Fe Institute *Studies in the Sciences of Complexity*, A.S. Perelson (ed.), Addison-Wesley, Reading, Ma. 1988.

[20] Macken, C.A. and A.S. Perelson. Protein Evolution on Rugged Landscapes, *Proceedings of the National Academy of Sciences* **86**, 6191-6195 (1989)

[21] McCaskill J.S. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers* **29**, 1105-1119 (1990)

[22] Maynard-Smith J. Natural Selection and the Concept of a Protein Space. *Nature* **225**, 563-564 (1970).

[23] Peritz A.E., R. Kierzek, N. Sugimoto and D.H. Turner. Thermodynamic Study of Internal Loops in Oligoribonucleotides: Symmetric Loops Are More Stable Than Asymmetric Loops. *Biochemistry* **30**, 6428-6436 (1991).

[24] Piccirrilli J.A., T. Krauch, S.E. Moroney and S.A. Brenner. Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **343**, 33-37 (1990).

[25] Pörschke D. Cooperative nonenzymic base recognition II. Thermodynamics of the helix-coil transition of oligoadenylic + oligouridylic acids. *Biopolymers* **10**, 1989-2013 (1971).

[26] Pörschke D. and M. Eigen. Co-operative non-enzymic base recognition III. Kinetics of the oligoribouridylic-oligoriboadenylic acid system and of oligoriboadenylic acid alone at acidic pH. *J.Mol.Biol.* **62**, 361-381 (1971)

[27] Pörschke D. Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix. *Biopys.Chem.* **1**, 381-386 (1974).

[28] Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. *Numerical Recipes in C.* Cambridge Univ. Press, New York 1988.

[29] Schuster P. Complex Optimization in an Artificial RNA World. In: D. Farmer, C. Langton S. Rasmussen and C. Taylor: *Artificial Life II*, SFI Studies in the Science of Complexity XII Addison-Wesley, Reading, Mass. 1991.

[30] Swetina J. and P. Schuster. Self-Replication with Errors. A Model for Polynucleotide Replication. *Biophys. Chem.* **16**, 329-345 (1982).

[31] Stadler P.F. and R. Happel. Correlation Structure of the Landscape of the Graph-Bipartitioning-Problem. *J. Phys. A: Math. Gen.* **25**, 3103-3110 (1992).

[32] Stadler P.F. and W. Schnabl. The Landscape of the Travelling Salesman Problem. *Phys. Lett. A* **161** 337-344 (1992).

[33] T C.Tuerk C., L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.*Science***249**, 505-510 (1990)

[34] Weinberger E.D. Correlated and Uncorrelated Fitness Landscapes and How to Tell the Difference. *Biol.Cybern.* **63**, 325-336 (1990).

[35] Weinberger, E.D. Local Properties the N-k Model, a Tuneably Rugged Energy Landscape. *Phys. Rev. A* **44**, 6399-6413 (1991).

[36] Weinberger E.D. and P.F. Stadler. Why Some Fitness Landscapes are Fractal. submitted to *J. Theor. Biol.* 1992.

[37] Wright S. The role of mutation, inbreeding, crossbreeding, and selection in evolution. In: *Proceedings of the 6th International Congress on Genetics* Vol.1, pp. 356-366. (1932).

[38] Zuker M., and P. Stiegler. Optimal Computer Folding of Large RNA Sequences Using Thermodynamic and Auxilary Information. *Nucl. Acid Res.* **9**, 133-148 (1981).

[39] Zuker M., and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.* **46**, 591-621 (1984).