Folding rate dependence on the chain length for RNA-like heteropolymers

Oxana V Galzitskaya and Alexei V Finkelstein

Background: Computer experiments and analytical estimates have shown that protein and RNA chains can reach their most stable folds without an exhaustive search over all their possible conformations. Protein-like chain folding proceeds via a specific nucleus and under conditions optimal for the fastest folding of these chains the dependence of the folding time (t) on the chain length (L) is in accord with the power law $t \sim L^b$ (b is a constant).

Results: Using Monte-Carlo folding simulations for a simple model of RNA secondary structure formation, we estimate the RNA chain length dependence of the time necessary to reach the lowest energy fold. Our results are compatible with a relatively weak power dependence of the folding time on the chain length, $t \sim L^b$. Such dependencies have been observed for different folding conditions, both for random sequences (here, $b > 5$) and for sequences edited to stabilize their lowest energy folds (for extremely edited sequences, $b < 2$). Although folding transitions in RNA chains are not an all-or-none type in terms of thermodynamics, they proceed via a folding nucleus in terms of kinetics. The peculiarity (compared with protein folding) is that the RNA critical nucleus is big and non-specific.

Conclusions: We have obtained a general scaling for the dependence of the RNA secondary structure on the chain length. The obtained power dependence is very weak compared with an exponential dependence for an exhaustive sorting.

Address: Institute of Protein Research, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia.

Correspondence: Alexei V Finkelstein
E-mail: afinkel@sun.ipr.serpukhov.su

Key words: energy minimum, folding and stability, optimal folding conditions, random and edited sequences, RNA secondary structure

Received: **15 August 1997**
Revisions requested: **26 September 1997**
Revisions received: **07 November 1997**
Accepted: **19 November 1997**

Published: **26 January 1998**
http://biomednet.com/elecref/1359027800300069

Folding & Design 26 January 1998, **3**:69–78

Introduction

Computer experiments with protein-like and RNA-like heteropolymers have shown that these biopolymers can reach their lowest energy fold without an exhaustive search over all their conformations [1–13]. Moreover, for protein-like model chains it has been shown that, at least under conditions optimal for the fastest folding, their folding time t depends on their length L according to the power law, $t \sim L^b$, where b is a constant [6,13]. It has been shown that small proteins [14,15] and protein-like model chains [6,9–11,13] fold by a nucleation-and-growth mechanism in terms of kinetics and undergo an all-or-none transition in terms of thermodynamics [12,16]. On the contrary, it has been shown that tRNA molecules do not undergo an all-or-none transition in terms of thermodynamics [8,17]. In this work we consider a simple model of RNA secondary-structure formation to study the folding rate dependence on the chain length in a wide range of ambient conditions, such as temperature and solvent quality, and to elucidate the kinetic and thermodynamic peculiarities of RNA secondary structure folding (some preliminary results have been reported previously [5,7]).

The main difference between RNA and protein molecules is that a 'link' (an RNA fragment of ~10 nucleotides) in the RNA secondary structure usually interacts with only one partner at a time, whereas a link in the protein globule has many partners.

How does this change RNA folding compared with protein folding?

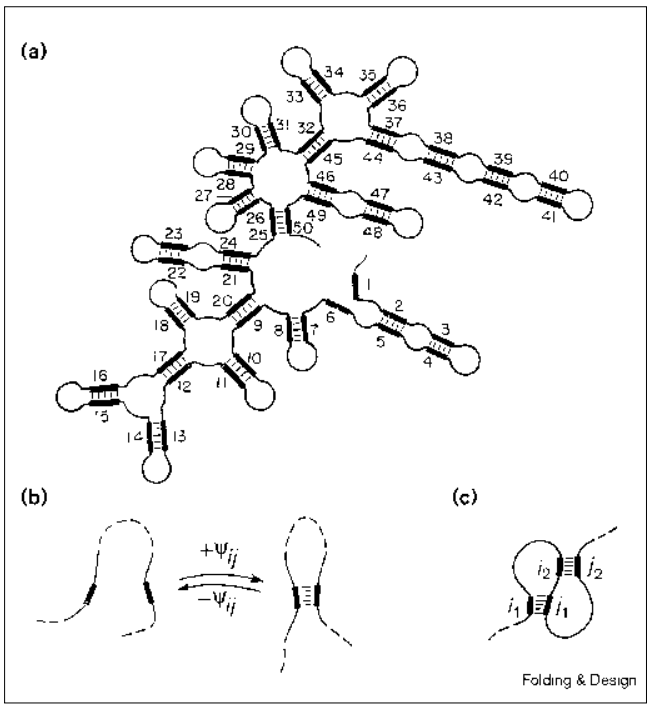
For RNA we shall only consider folding of the secondary structure. This is a reasonable simplification because, as a rule, the RNA secondary structure folds before packing into the tertiary structure and its folding can therefore be considered independently of the three dimensional structure formation [8,17–19].

The theoretical investigation of RNA secondary structure folding has an important technical advantage over studies of protein folding: here, we have conventional algorithms for quick calculations of the most stable structure [20,21] and the free energy for the chain [22]. Thermodynamics of RNA chains can therefore be investigated in a much wider range of chain lengths, sequences and folding conditions than for protein-like models.

Model and methods

We use a simplified model of RNA secondary structure [5]. The RNA chain is represented as a sequence of links that can stick together to form 'pairs' (Figure 1a), which model the double helices. The links stand for RNA fragments

Figure 1



RNA secondary structure. **(a)** One of the many possible secondary structures of an RNA chain. Bold lines show the 'links' (RNA fragments). The links form 'pairs' (RNA double helices; shown by hatched lines). In the secondary structure shown, all the links except 1 and 6 form pairs. **(b)** Elementary movements of a chain and the corresponding changes in free energy. Ψ_{ij} is the free energy of 'sticking' links i and j. **(c)** A forbidden structure: a pseudo-knot.

of ~10 nucleotides, so that their pairing is stabilized by a few complementary nucleotide pairs. This will allow us to use a statistical treatment of their interactions. It is assumed that these fragments do not overlap and that each link can interact with only one other link at a time. This simplification corresponds to a description of a long RNA secondary structure at a medium resolution (such a resolution usually reads as 'one double helix binds chain regions around nucleotides 60 and 840, another around 70 and 500 ... '). Ψ_{ij} is the free energy of pairing of links i and j (Figure 1b). The value Ψ_{ij} is assumed to be independent of the $\Psi_{i'j'}$ values for other pairings. This is also an approximation, because alternative folds in a true RNA can involve combinations between base pairings formed by adjacent chain fragments as well as by additional stacking of adjacent helical regions in a continuous helix. Another simplification is that we do not consider pseudo-knots (Figure 1c). Although they exist in real RNA structures, their number is always very small [23]. The advantage of neglecting them is that when the formation of pseudo-knots is not considered, one can strictly divide the secondary and the tertiary structure of RNA [24] and apply strict dynamic programming algorithms [20–22] for thermodynamic calculations of this RNA secondary structure model.

All the features neglected in our model (tertiary interactions, pseudo-knotting, overlapping of secondary structure fragments, etc.) are those that make a true RNA more similar to a protein globule. By having two extreme models (that of the RNA secondary structure and that of the protein globule), a stereoscopic view on folding events can be obtained.

A general form for the free energy parameters

All kinetic and thermodynamic properties of the RNA secondary structure in our model are determined by the values of Ψ_{ij}/RT (for $1 \le i < j \le L$, where L is the number of chain links), where Ψ_{ij} is the free energy of formation of pair (i,j), R is the gas constant, and T is the temperature. The term Ψ includes the free energy of 'sticking' (i.e. contact between links; the solvent effects are included) as well as the entropic effects connected with loop formation. It is convenient to represent the values Ψ_{ij}/RT in a general form [5]:

$$\Psi_{ij}/RT = \varepsilon_{ij}/RT + \varphi \qquad 1 \le i < j \le L \qquad (1)$$

where $\langle \varepsilon_{ij} \rangle \equiv 0$, φ $(= \langle \Psi_{ij} \rangle / RT)$ is the mean sticking strength, and $\langle\ \rangle$ indicates an averaging over all the pairs (i,j). Such a form of presentation of the Ψ_{ij} values is convenient for distinguishing the effects connected with the chemical heterogeneity of the monomers from those connected with the mean sticking strength φ (determined by the solvent quality and flexibility of the chain) and with the temperature T. This allows us to investigate the range of parameters φ and T for a given set of specific interaction energies ε_{ij} determined by the RNA sequence.

Statistical mechanics of RNA-like chains

The free energy of an unfolded chain (i.e. of the chain without any secondary structure) is taken as zero. The free energy of a chain with secondary structure σ is:

$$F_\sigma\{\Psi\} = \sum_{(i<j)} \delta_{ij}^\sigma \Psi_{ij} \qquad (2)$$

Here, the sum is taken over all the possible pairs (i,j); $\delta_{ij}^\sigma = 1$ if the pair (i,j) is present in the secondary structure σ and, if the pair is not, $\delta_{ij}^\sigma = 0$; each link can enter only one pair; pseudo-knots are forbidden, hence $\delta_{ij}^\sigma \cdot \delta_{i'j'}^\sigma \equiv 0$ when $i < i' < j < j'$ or $i' < i < j' < j$ in any allowed structure σ (Figure 1c). The partition function for the RNA chain is:

$$Z = \sum_\sigma \exp\left(-F_\sigma / RT\right) = \sum_\sigma \prod_{(i<j)} \exp\left(-\left(\Psi_{ij} / RT\right) \cdot \delta_{ij}^\sigma\right) \qquad (3)$$

The sum is taken over all the allowed (i.e. without pseudo-knots) secondary structures, and the product is taken over all the pairs (i,j), where $1 \le i < j \le L$). Such a partition function can be calculated recursively for any set of Ψ_{ij} within the time proportional to L^3 [22].

We take the 'native' structure as the secondary structure with the lowest free energy among those with the maximal possible number of pairs ($N_\sigma = L/2$). The free energy of the native structure is:

$$F_N\{\Psi\} = \min_{\substack{\sigma \\ N_\sigma = L/2}} (F_\sigma\{\Psi\}) \quad (4)$$

The algorithm to find this structure [5] is based on the ideas of Nussinov and Jacobson [20].

The thermodynamic probability of the native structure is calculated according to the Boltzmann equation:

$$w_N = \exp(-F_N/RT)/Z \quad (5)$$

The native structure is thermodynamically stable when $w_N > 0.5$.

The average energy of the RNA secondary structure and the average number of pairs in this structure are calculated as:

$$E = \sum_{(i<j)} p_{ij}\,\varepsilon_{ij} \quad (6)$$

$$N = \sum_{(i<j)} p_{ij} \quad (7)$$

the average secondary structure content is:

$$\theta = N/(L/2) \quad (8)$$

Here, the sums are taken over all the residue pairs ($1 \le i < j \le L$); ε_{ij} is the energy of pair (i,j); p_{ij} is the probability that residues i and j form pair (i,j). This probability is computed as:

$$p_{ij} = \exp(-\Psi_{ij}/RT)\, Z_{i+1,j-1}\, Z^*_{j+1,i-1}/Z \quad (9)$$

where $Z_{i+1,j-1}$ is the partition function of secondary structures formed in the chain region between the links i and j; $Z^*_{j+1,i-1}$ is the partition function of secondary structures formed in the chain regions outside this pair; and Z (which is equal to $Z_{1,N}$) is the total chain partition function (see Equation 3). The Z_{ij} values are computed recursively ($Z_{i+1,i}$ and $Z_{i,i}$ are equal to unity) at propagation from short-range to long-range (i,j) pairs and stored in computer memory; the Z^*_{ji} values are also computed recursively [22,25] at back propagation.

Generation of random and edited chains

To investigate the dependence of folding time on the chain length, we considered chains with 4–100 links. Every 'nucleotide sequence' is presented as a matrix of specific interaction energies ε_{ij} ($1 \le i < j \le L$, $\varepsilon_{ij} = \varepsilon_{ji}$).

For a random sequence, the values ε_{ij} are generated with a Gaussian distribution having the mean characteristics $\langle\varepsilon_{ij}\rangle = 0$ and $\langle\varepsilon_{ij}^2\rangle^{1/2} = 1$, where 1 is the energy unit. In this study the same unit is also the temperature (or, more correctly, RT) unit.

The Gaussian distribution for pairing energies is certainly a simplification. But if a random pairing of fragments with 10 nucleotides gives 2–3 complementary base pairings on average, ~50% of the fragment pairings are bound by two or three AT or GC pairs, 20% by one or four such pairs, etc. (i.e. the energy distribution will have a roughly a quasi-Gaussian form).

Furthermore, we edited the random sequences to make their native folds more stable. To this end we singled out the native structure of a random sequence and added some negative energy Δ to all the ε_{ij} values that contribute to the native structure energy. We did 'weakly', 'moderately' and 'strongly' edited offsprings (sequences where Δ is added to each ε_{ij} value contributing to the native structure energy) of a random sequence using $\Delta = -0.5$, -1.0 and -2.0 energy units, respectively. Such editing can be interpreted as a stabilization of the native double helices by introducing additional complementary base pairs. It is known that native RNA double helices include a greater number of complementary nucleotide pairs than expected by chance [26]. $-\Delta$ is a quantitative measure of the degree of editing, which changes from zero for random chains to infinity for ideally designed sequences; the ideally designed sequences correspond to the Go model [27], where $\varepsilon_{ij} = 1$ for the native contacts and $\varepsilon_{ij} = 0$ for all the other contacts.

Investigation of folding kinetics

Folding simulations were done as in [5], using the Metropolis scheme [28] of the Monte-Carlo method. The elementary movements include only the formation of a new pair and the decay of the existing pair. The Metropolis criterion [28] is used to accept or reject the movements [5]. The movements are repeated until the native structure is reached or the computation time limit (10^7 Monte-Carlo steps) is exceeded. A folding simulation always begins from the unfolded chain.

In all the kinetics experiments we are interested in the 'first passage time', the time spent to reach the native structure for the first time. The characteristic Monte-Carlo first passage time, $t_{1/2}$, for a given chain and given conditions (temperature T and mean 'sticking' strength φ) is determined as the number of Monte-Carlo steps required to complete 50% of Monte-Carlo runs [5]. Thus, $t_{1/2}$ can be determined even when 50% of runs fail to converge within the computation time limit. The characteristic $t_{1/2}$ has a meaning and can be determined computationally independently of the thermodynamic stability of the native structure. When the native structure is thermodynamically stable, $t_{1/2}$ coincides with the experimentally observable native state folding time.

Statistical analysis of the results

To estimate both the characteristic $t_{1/2}$ and an error in this estimate for a given chain and given conditions (φ,T), we

Table 1

Characteristic first passage times and the native state Boltzmann probabilities for a strongly edited (Δ = –2) 40-link chain under different temperatures and mean sticking strengths.

	T^{-1}									
	1.5		2		2.5		3		3.5	
φ	Pt	Bp	Pt	Bp	Pt	Bp	Pt	Bp	Pt	Bp
5	10^7	10^{-7}	27000	10^{-2}	920	2	840	20	**1620**	**50**
4.5	$>10^7$	10^{-5}	3500	0.1	740	7	880	36	**1980**	**67**
4	260000	10^{-3}	1000	1	800	18	**1500**	**50**	**4000**	**80**
3.5	11500	10^{-2}	700*	5	1000	33	**25000**	**66**	**6000**	**86**
3	2500	0.3	760	14	**1200**†	**50**	**4000**	**80**	**10500**	**90**
2.5	1500	2	900	30	**2000**	**65**	**7000**	**85**	**20000**	**94**
2	800	8	1500	45	**2500**	**76**	**8000**	**90**	**28000**	**96**
1.5	1000	20	**2000**	**60**	**5500**	**85**	**17000**	**90**	**45500**	**98**

T, temperature (in energy units); φ, mean sticking strength (in energy units); Pt, passage time (in Monte-Carlo steps); Bp, Boltzmann probabilities (%). *The minimum of $t_{1/2}^{opt}$; †the minimum of $t_{1/2}^{stab}$. The region where the native state thermodynamic probability exceeds 50% is shown in bold.

performed two sets of 25 Monte-Carlo runs for each chain and (φ,T) point. For each set of runs, $t_{1/2}$ was determined as the number of Monte-Carlo steps sufficient to come to the native state in 13 of 25 runs. Having $t_{1/2}'$ and $t_{1/2}''$ for these two sets, we obtain the estimate of the characteristic folding time and the error in this estimate as:

$$t_{1/2}^{0} \pm \delta t_{1/2} = (t_{1/2}' + t_{1/2}'')/2 \pm |t_{1/2}' - t_{1/2}''|/2 \qquad (10)$$

Having four sequences ($n = 4$) of each chain length and degree of editing, and two characteristic times ($t_{1/2}'$ and $t_{1/2}''$; $i = 2$) for each sequence at a given (φ,T) point, we can calculate $\overline{f(t)}$ — the average value of each folding time-dependent function f in the (φ,T) point — as well as the error in this average estimate, $|\Delta f(t)|$, as:

$$\overline{f(t)} = \frac{1}{8} \sum_{n=1}^{4} \sum_{i=1}^{2} f_{i,n}(t)$$

$$|\Delta f(t)| = \left[\frac{1}{7} \sum_{n=1}^{4} \sum_{i=1}^{2} (f_{i,n}(t) - \overline{f(t)})^2\right]^{1/2} \qquad (11)$$

Because the logarithm of the reaction [29], rather than the reaction time itself, is usually important in kinetic analysis, in this work we consider $f(t) = \ln(t)$ and also $f(t) = \ln[\ln(t)]$.

To analyze the first passage time on the chain length dependence, we made a linear interpolation of experimental points in different coordinate axes. To obtain the best approximation of experimental points $y_r (r = 1,2 ..., \nu)$ by a theoretical dependence $y(x_r) = a + bx_r$ (where a and b values are the fitting coefficients), the value:

$$\chi^2 = \sum_{r=1}^{\nu} [(a + bx_r - y_r)^2/\Delta y_r^2] \qquad (12)$$

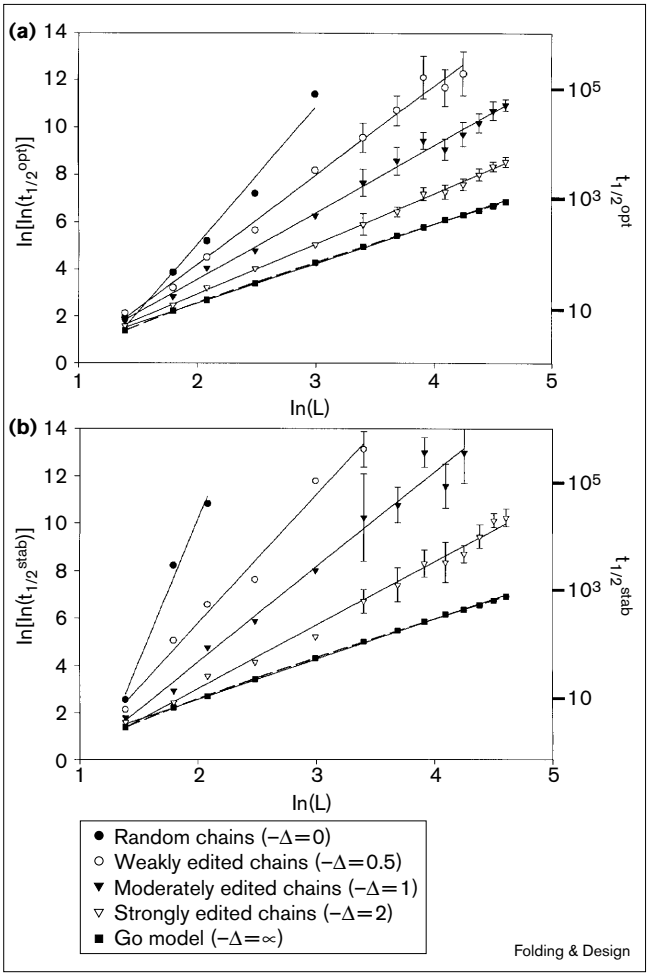
(where Δy_r is an error in the estimate of y_r) must be minimized over a and b [30]. In this way the coefficients a and b are obtained and, furthermore, the quality of approximation is estimated from a comparison of the obtained χ^2 value with the one tabulated for a random distribution of χ^2 values [30]. The tabulated χ^2 values depend on the number of degrees of freedom, which is $\nu - 2$ because we have ν experimental points and two adjustable parameters.

Results and discussion

In the computer experiments we investigate the dependence of the first passage time on the chain length; this is done under conditions of the fastest achievement of the native state (the corresponding value is called $t_{1/2}^{opt}$) and under conditions of the most rapid folding to the thermodynamically stable native state (this value is called $t_{1/2}^{stab}$). Using our preliminary estimates of (φ,T) conditions appropriate for the fast folding and for the native state stability [5,7], we have done a further scan over (φ,T) coordinates to determine $t_{1/2}^{opt}$ and $t_{1/2}^{stab}$ for each chain. Table 1 illustrates this search for one of the chains.

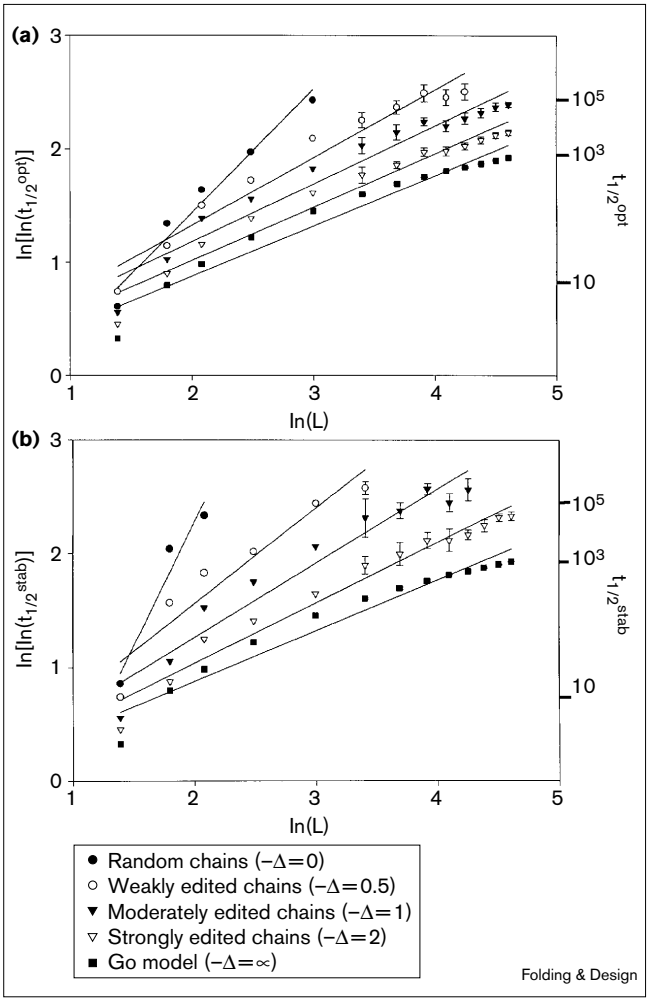
The longer the chains, the longer the folding times and the less edited the chain, the longer it takes to fold. Thus, a reliable estimate of $t_{1/2}^{opt}$ for random sequences was obtained only for short chains (L = 4–20) and a reliable estimate of $t_{1/2}^{stab}$ was obtained only for L = 4–12. For the same reason, the values $t_{1/2}^{opt}$ and $t_{1/2}^{stab}$ were not obtained for very long chains with weakly and moderately edited sequences. A body of computer simulation results can be found in Tables 2–5 of [7]. The times $t_{1/2}^{opt}$ and $t_{1/2}^{stab}$ coincide for the extremely edited chains: their fastest folding occurs in a wide (φ,T) region, where the native folds are thermodynamically stable.

Figure 2



Dependence of the characteristic first passage time $t_{1/2}$ (measured as the number of Monte-Carlo steps) on the number of chain links L in the coordinates $\ln(t_{1/2})$ versus ln(L). In these coordinates the dependence $t = AL^b$ is a straight line: $\ln(t) = \ln(A) + b \cdot \ln(L)$. Errors are shown by vertical bars (when larger than symbol size) for $L \geq 30$ only, to avoid overloading the figure. **(a)** First passage time $t_{1/2}^{opt}$, corresponding to the fastest folding conditions. **(b)** First passage time $t_{1/2}^{stab}$, corresponding to conditions of the most rapid folding of the thermodynamically stable native structure. For the Go model, the dashed curve for the lowest set of data points represent the best fit of a dependence $\ln(t) = \ln(A) + b \cdot \ln(L) + c/\ln(L)$. See the text for more details.

Figure 3



Dependence of the characteristic first passage time $t_{1/2}$ on the number of chain links L in the coordinates $\ln[\ln(t_{1/2})]$ versus ln(L). The dependence $t = \exp(AL^b)$ is a straight line; $\ln[\ln(t)] = \ln(A) + b \cdot \ln(L)$.

In this study we examine two main possible functional dependencies of the folding time $t_{1/2}$ on the chain length L: $t_{1/2} = AL^b$ and $t_{1/2} = \exp(AL^b)$, where A and b are constants. The aim is to elucidate the quality of these approximations. The dependence $t_{1/2} = A \cdot \exp(bL)$ has been also investigated in [7] and it has been shown that it does not fit the experimental data. The results obtained are presented in Figures 2 and 3 in coordinates linearizing these functional dependencies: $\ln(t_{1/2})$ versus ln(L) and $\ln[\ln(t_{1/2})]$ versus ln(L). The best fitted parameters of the linear interpolations — the values of ln(A) and b — are presented in Table 2 together with the correlation coefficients C and the error functions χ^2 for each of the tested functional dependencies.

One can see in Figures 2 and 3 that the experimental points obtained in computer simulations fit rather well by the straight lines in all the examined cases. It is noteworthy that the approximation $t_{1/2} = \exp(AL^b)$, which we suggested earlier, [5] becomes worse than the approximation $t_{1/2} = AL^b$ only when the examined chain lengths L varied by more than an order of magnitude. The correlation coefficients are very high for both the tested dependencies: on average they are 0.974 for the dependence $t = \exp(AL^b)$ and 0.996 for the dependence $t = AL^b$. Thus, from Figures 2 and 3 or the correlation coefficients given in Table 2 it can be concluded that both examined dependencies are good but that shown in Figure 2 is better. To rule out one

Table 2

Parameters of the best fitted approximations presented in Figures 2 and 3.

Approximation		$t_{1/2}^{opt}$					$t_{1/2}^{stab}$				
		$-\Delta=0$	$-\Delta=0.5$	$-\Delta=1$	$-\Delta=2$	$-\Delta=\infty$	$-\Delta=0$	$\Delta=0.5$	$-\Delta=1$	$-\Delta=2$	$-\Delta=\infty$
$t=AL^b$	ln(A)	−5.51	−3.19	−2.12	−1.36	−0.77	−14.72	−5.40	−4.15	−2.33	−0.77
	b	5.23	3.72	2.84	2.14	1.69	12.48	5.45	4.12	2.63	1.69
	C	0.990	0.997	0.997	0.999	0.999	0.997	0.998	0.993	0.994	0.999
	χ^2	1.13	1.96	3.52	1.65	51.97*	0.06	0.58	5.51	14.35	51.97*
	ν	5	10	13	13	13	3	6	10	13	13
		$\Sigma\chi^2=8.3$, $\Sigma(\nu-2)=33$, $p=1$				$p=3\times10^{-7}$	$\Sigma\chi^2=20.5$, $\Sigma(\nu-2)=24$, $p=0.7$				$p=3\times10^{-7}$
$t=\exp(AL^b)$	ln(A)	−0.36	0.38	0.61	0.33	0.01	−1.96	−0.18	0.06	0.14	0.01
	b	0.94	0.53	0.39	0.40	0.44	2.11	0.82	0.62	0.48	0.44
	C	0.988	0.976	0.958	0.977	0.968	0.975	0.978	0.964	0.987	0.968
	χ^2	1.56	14.44	22.87	30.15	3645	0.53	5.43	14.32	23.91	3645
	ν	5	10	13	13	13	3	6	10	13	13
		$\Sigma\chi^2=69.0$, $\Sigma(\nu-2)=33$, $p=2\times10^{-4}$				$p=0$	$\Sigma\chi^2=44.2$, $\Sigma(\nu-2)=24$, $p=7\times10^{-3}$				$p=0$

C, the standard correlation coefficient between the experimental and the fitted values; χ^2, the sum of quadratic deviations of the fitted approximations from the experimental values; ν, the number of experimental points; ν − 2, the number of degrees of freedom; p, the probability that the given χ^2 value is due to a random deviation.

*See the text for fitting of dependence $t_{1/2}=AL^b\cdot\exp(c/\ln(L))$, which gives $\chi^2=13.56$, $p=0.3$ and $b=1.53$. Random ($-\Delta=0$), weakly ($-\Delta=0.5$), moderately ($-\Delta=1$), strongly ($-\Delta=2$) and ideally ($-\Delta=\infty$) edited sequences.

of the two dependencies, a measure of accuracy of the tested dependencies more sensitive than the correlation coefficient must be used.

To this end, we employed the χ^2 criterion (see Statistical analysis of the results), which is commonly used [30] to test the probabilities of hypotheses. The χ^2 values obtained for the random, weakly, moderately and strongly edited sequences (Table 2) are quite compatible with a hypothesis that the observed deviation of the experimental points from the dependence $t_{1/2}=AL^b$ is obtained by chance. The corresponding probabilities are very high, 1.0 for $t_{1/2}^{opt}$ and 0.7 for $t_{1/2}^{stab}$. At the same time, the χ^2 deviations from the dependence $t_{1/2}=\exp(AL^b)$ are significant for these sequences: the observed χ^2 values can be obtained by chance with a probability of only 2×10^{-4} for $t_{1/2}^{opt}$ and 7×10^{-3} for $t_{1/2}^{stab}$. Thus, the $t_{1/2}=\exp(AL^b)$ functional dependence can be ruled out and the dependence $t_{1/2}=AL^b$ is basically valid.

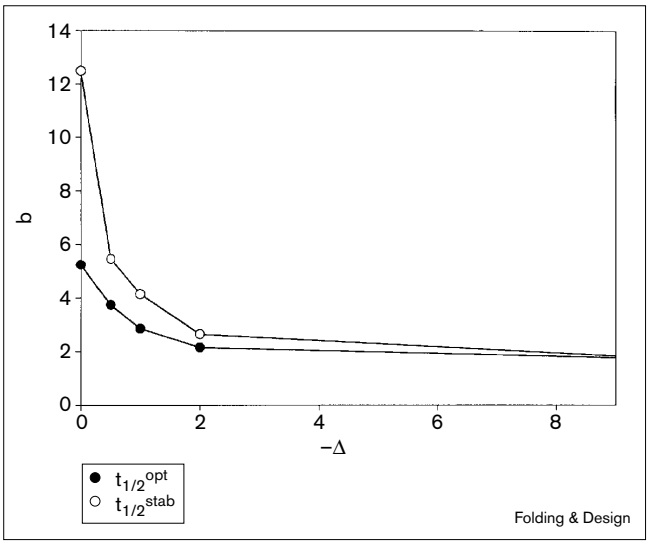
Certainly, the results of folding simulations presented in Figures 2 and 3 cannot rule out a composed scaling dependence like $t_{1/2}=AL^b\cdot\exp(\alpha L^\beta)$ where α and/or β are sufficiently small. In particular, no scaling over L when $L<10^3$ can rule out a multiplier $\exp(0.001L)$, although this multiplier will dominate when $L>>10^3$. This is a usual problem with the scaling of experimental results, including those of computer simulations (cf. [6]). Actually, one can only test some *a priori* hypotheses to see if they fit the experiment, rather than withdraw a scaling law directly from the experiment. The presented results, however, show that the algebraic term L^b is the main term in the RNA secondary structure folding time scaling, as long as the RNA length L does not exceed hundreds of links.

The above consideration is illustrated by the results obtained for the extremely edited sequences (the 'Go model'). Here, the observed χ^2 deviations are big because the errors in the folding time estimates are very small; the small errors allow estimates of the dependence of $t_{1/2}$ on L to be more precise. The observed χ^2 deviations (Table 2) rule out the dependence $t_{1/2}=\exp(AL^b)$ completely, but a significant χ^2 value observed for the tested dependence $t_{1/2}=AL^b$ shows that this dependence is also only approximately valid.

The dependence $t_{1/2}=AL^b\cdot\exp[c/\ln(L)]$, where c is a constant (we tried a more or less arbitrary form with the same algebraic asymptotics at $L\to\infty$ because a rough estimate of the Go model shows that the power b is between 1 and 2 at $L\to\infty$), gives a reasonable fit at $\ln(A)=0.09$, $b=1.53$ and $c=-1.15$: the resulting $\chi^2=13.56$ (for the same 13 points). In other words, the probability of a random deviation is reasonably high, $p=0.3$. It is noteworthy, however, that this complication of the $t_{1/2}$ scaling dependence on L results in a very small modification of the corresponding curve in Figure 2.

Thus, we see that the dependence $t_{1/2}=AL^b$ gives a good (although not absolutely precise) description of the first passage time dependence on the chain length, both for the most rapid folding ($t_{1/2}^{opt}$) and for the fastest folding to

Figure 4



The dependence of the power-law exponent b on the degree of editing, −Δ. −Δ = 0 corresponds to the random chains and −Δ = ∞ corresponds to the ideal editing of chains (the Go model). For this model, b = 1.7 both for $t_{1/2}^{opt}$ and $t_{1/2}^{stab}$.

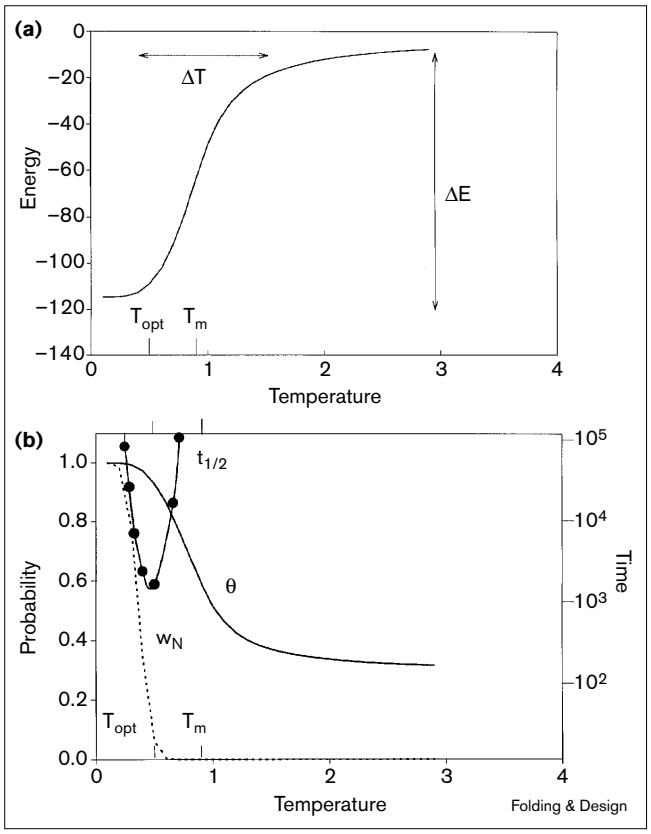
the thermodynamically stable native state ($t_{1/2}^{stab}$). An additional significant property of the dependence $t_{1/2} \sim L^b$ is that all the interpolated lines intersect the axis $\ln(t_{1/2}) = 0$ in the region where ln(L) is close to unity. Thus, the found dependence appears as $t_{1/2} \approx (L/2)^b$, which corresponds to a natural requirement that a chain of two links must fold in one step.

Figure 4 shows the dependence of the power b on the editing parameter Δ. It is noteworthy that b < 3 for strongly edited sequences. Thus, the kinetic search of the native state is done faster for these chains than by dynamic programming algorithms.

It is also noteworthy that the power values b observed for the random and for the extremely edited RNA-like chains in the region of their fastest folding (5.5 and 1.7, respectively) are approximately one unit less than the power values observed in the same region for the random and for the extremely edited protein-like chains (6 and 2.7, respectively [6]).

Thus, computer simulations show that the RNA-like and the protein-like [6] model chains have similar power scaling laws ($t \sim L^b$) for the folding time dependence on the chain length in the region of the most rapid folding of these chains. Outside these regions, however, the scaling laws for the protein-like and the RNA-like chains may be different. In particular, for the border of the native structure stability we obtained $t_{1/2}^{stab} \sim L^b$ for the RNA-like

Figure 5



Temperature (T) dependencies for a 70-link chain with a strongly edited (−Δ = 2) sequence at φ = 3 (this φ value corresponds to the most rapid folding time of this chain to its native structure). The temperature dependence of **(a)** the average secondary structure energy (E) and **(b)** the average secondary structure content θ, the native state thermodynamic probability w_N, and the characteristic first passage time $t_{1/2}$. The values w_N, θ and E are calculated using Equations 5–8. ΔT is the width of the transition, ΔE is the energy difference between the native and the unfolded states, T_m is the midpoint of secondary structure melting, T_{opt} is the optimal temperature for the fast kinetics.

chains, although an analytical estimate of the protein folding time [9] is proportional to $\exp(L^{2/3})$ and for other ambient conditions it can also scale as $\exp(L^{1/2})$ [8,13,15]. Moreover, thermodynamics of folding transitions are quite different for RNA-like and protein-like chains. The all-or-none transitions with a very low content of intermediate states are typical for globular proteins [16] and their models [12], but not for the RNA-like heteropolymers [25]. The temperature dependence of the RNA secondary structure energy (Figure 5a) does not satisfy the Van't Hoff criterion [29] of an all-or-none transition, $\Delta E \cdot \Delta T = 4RT_m^2$, where T_m is the transition temperature, ΔT is the transition width and ΔE is the energy change: the observed ΔT is much greater than $4RT_m^2/\Delta E$, which means that there is a high population of folding intermediates [16]. Figure 5b shows that the decrease in the native

state probability w_N is much faster than the overall decrease in the secondary structure content θ. It is also noteworthy that the fastest achievement of the native state corresponds to a temperature where w_N is rather low, while θ is close to unity.

Thus, folding thermodynamics are absolutely different for the RNA-like and protein-like chains, although their phenomenological folding kinetics (at least, in the region of the most rapid folding) are rather similar. Does the similar folding kinetics mean a similarity of their folding mechanisms? In particular, is the 'nucleation-and-growth' mechanism, which is typical for proteins [11,14,31], valid for RNA secondary structure folding?

Actually, we did not expect to find any nucleation phenomena in the RNA secondary structure formation. In proteins and protein models, each link can have many simultaneous interactions, which leads to the surface tension and the consequent determination of the nucleus. In the examined RNA model, each link can interact with only one other link at a time; hence, here one pairing has no particular effect on pairing of adjacent links. To our surprise, however, we found a kind of critical folding nucleus in RNA folding.

The critical folding nucleus is a structure corresponding to the transition state (i.e. to the free energy maximum) on the folding pathway. If the transition state exists on the folding pathway, and if the molecule is in the transition state, then the molecule can take one of two pathways, both going downhill in free energy but in the opposite directions (cf. [11]). One pathway leads to the native state; following this downhill route, the molecule rapidly achieves its native structure. The other pathway leads to the denatured state(s); if the molecule takes this route, it comes to the denatured state and can then come to the native structure only after a long time, in the same way as a molecule starting from the unfolded state. In addition, a molecule can spend some time wandering at the top of the free energy barrier. One will find the critical folding nucleus if one chooses a structure distinguished by such bifurcated kinetics or a by a very broad spectrum of the folding times.

Assuming that the transition state exists, one can determine its overall characteristics (mean energy $E_\#$ and mean

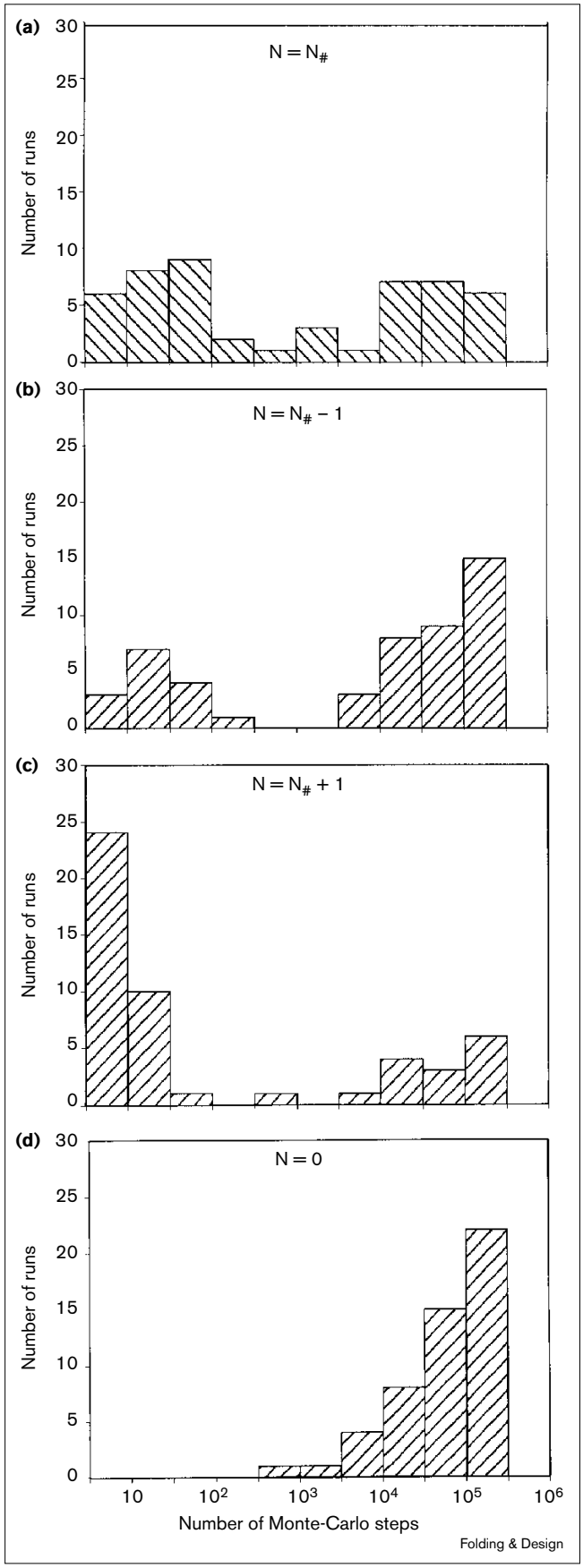


Figure 6

The distribution of the folding times (in Monte-Carlo steps) obtained in 50 independent runs at the point of the fastest folding ($\varphi = -0.5$ and $T^{-1} = 1.5$) for the 20-link chain with a random sequence. The simulations begin with four different starting conformations: **(a)** the nucleus with $N_\# = 8$ native pairings; **(b)** the same nucleus minus one pairing; **(c)** the same nucleus plus one native pairing; and **(d)** the unfolded chain.

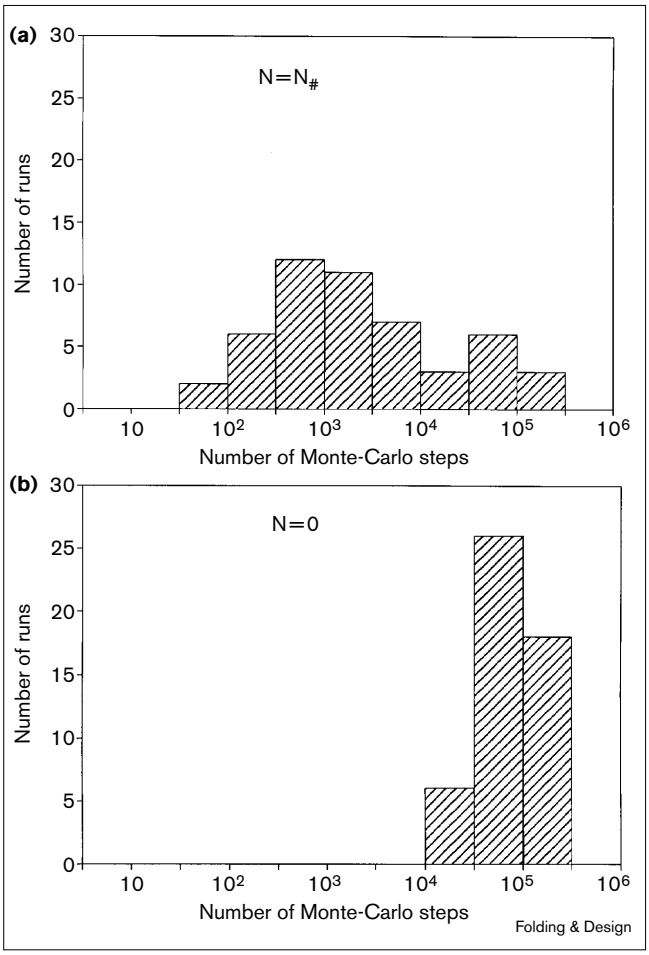
number of pairings $N_{\#}$) in the (φ,T) point of the fastest folding: these characteristics coincide with the equilibrium characteristics (E_D and N_D) of the denatured state [4,32]. Within the frame of the transition state theory [29], the folding time ($t_{1/2}$) is proportional to $\exp[(F_{\#} - F_D)/RT]$, where $F_{\#} = E_{\#} + N_{\#}\varphi T - TS_{\#}$ is the free energy of the transition state and $F_D = E_D + N_D\varphi T - TS_D$ is the free energy of the denatured state ($S_{\#}$ and S_D are conformational entropies of these states). Because $\partial t_{1/2}/\partial T \sim t_{1/2}(E_D - E_{\#})/T^2$ and $\partial t_{1/2}/\partial\varphi \sim t_{1/2}(N_{\#} - N_D)$, $E_{\#} = E_D$ and $N_{\#} = N_D$ at the (φ,T) point where $t_{1/2}$ reaches its minimum (where $t_{1/2} = t_{1/2}^{opt}$ [4]). The native state probability w_N is very small at this point (see Figure 5b), so the denatured state is predominant here and the E_D and N_D values virtually coincide with the E and N values calculated (see Equation 6–7) for the chain at this point.

In this way the values $E_{\#}$ and $N_{\#}$ corresponding to the $t_{1/2} = t_{1/2}^{opt}$ point have been computed for each of the investigated chains. If the folding nucleus exists, they must characterize its energy and the secondary structure content. In all the examined cases, the $E_{\#}$ and $N_{\#}$ values turned out to be rather close to the E_N and $N_N = L/2$ values characterizing the native state (see Figure 5). This means that the critical nucleus (if it exists) is very big: its secondary structure content is ≥ 80% and this value grows with the chain length and the degree of editing (results not shown). At the same time the native state probability w_N decreases with the chain length, but grows with the degree of editing [5–7].

Now one had to prove that such a big fragment of the native RNA structure indeed works like a critical nucleus, in other words the folding kinetics starting from this fragment are indeed bifurcated.

To find a candidate structure for the role of the critical nucleus, we determined the $N_{\#}$ and $E_{\#}$ values for the given chains as described above and examined all the possible structures with $N_{\#}$ native pairs whose summary energy is close to $E_{\#}$. We did 50 folding simulations starting from each candidate structure at the (φ,T) point corresponding to the fastest folding of the chain. Typical distributions of these folding times (Figures 6a and 7a) are very broad and bifurcated: they vary from a few steps up to the times typical for folding from the completely unfolded state (cf. Figures 6d and 7b). Such a distribution of the folding times confirms the nucleation mechanism of the native secondary structure folding in the RNA-like chains, although the 'nuclei' in this case are much bigger than thought before and, in particular, than those observed in the protein-like chains [11]. A bifurcated picture is more pronounced for random sequences (Figure 6) than for edited ones (Figure 7); this pattern virtually does not depend on the chain length. When the folding simulation starts with the critical nucleus minus

Figure 7



The distribution of the folding times (in Monte-Carlo steps) obtained in 50 independent runs at the point of the fastest folding ($\varphi = 0.5$ and $T^{-1} = 2$) for the for 70-link chain with a weakly edited sequence. The simulations begin with: **(a)** the nucleus with $N_{\#} = 29$ native pairings and **(b)** the unfolded state. When folding starts from the nucleus, the folding time spectrum is 300 times wider than that corresponding to folding from the unfolded state.

(Figure 6b) or plus (Figure 6c) one native pairing, the two-phase distribution of folding times remains essentially the same, with some decrease or increase, respectively, in the number of fast achievements of the native structure. When we did kinetic simulations starting from the randomly chosen candidates with $N_{\#}$ native pairings whose energy was different from $E_{\#}$, we observed essentially the same two-phase distribution of folding times as seen in Figures 6 and 7. This shows that folding nuclei in the RNA-like chains are non-specific, and big, unlike those in the protein-like chains, which are specific, and small [11].

Analysing the results of simulations, we can suggest that the main reason for fast folding starting from the nucleus is that the nucleus divides the remaining RNA chain into

short independent branches, which achieve the native state independently of each other. Thus, one big and complicated problem converts into many small and simple ones, which are solved rapidly. The nucleation mechanism in kinetics is usually conjugated with the first-order (or all-or-none) transition in thermodynamics [33], but in the RNA-like chains nucleation kinetics exists in the absence of any all-or-none thermodynamics.

Conclusions

Protein-like and RNA-like models are rather different. The protein-like models consider many-particle interactions, whereas RNA-like models consider only pairings of links. The consequence is that the thermodynamics of folding are different for the protein-like and the RNA-like chains. Their folding kinetics have many common features, however: for any chain there are temperature and solvent conditions under which the chain can fold to its native (the lowest energy) fold much faster than by an exhaustive search over all its conformations [2,4–6]; editing of the random sequences, which makes their native folds more stable, results in an acceleration of achievement of their native folds [4–6]; strong long-range native contacts speed up folding [34,35]; the chains have the same power scaling law $t \sim L^b$ for their folding time dependence on the chain length, at least at the point of the fastest folding; and the chains fold via a folding nucleus, although it is small and specific for the protein-like chains and large and non-specific for the RNA-like chains.

Acknowledgements

This work was supported in part by grant No. 96-04-50605 of the Russian Foundation for Basic Research and by an International Research Scholar's Award No. 75195-544702 from the Howard Hughes Medical Institute.

References

1. Šali, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of protein folding. A lattice model study of requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614-1636.
2. Socci, N.D. & Onuchic, J.N. (1994). Folding kinetics of protein heteropolymers. *J. Chem. Phys.* **101**, 1519-1528.
3. Galzitskaya, O.V., Reva, B.A. & Finkelstein, A.V. (1994). Protein chain can achieve the energy minimum without exhaustive sorting of all its conformations: computer modeling and phenomenological theory. *Mol. Biol. (Russia)* **28**, 1412-1427.
4. Galzitskaya, O.V. & Finkelstein, A.V. (1995). Folding of chains with random and edited sequences: similarities and differences. *Protein Eng.* **8**, 883-892.
5. Galzitskaya, O.V. & Finkelstein, A.V. (1996). Computer simulation of secondary structure folding of random and 'edited' RNA chains. *J. Chem. Phys.* **105**, 319-325.
6. Gutin, A.M., Abkevich, V.I. & Shakhnovich, E.I. (1996). Chain length scaling of protein folding time. *Phys. Rev. Lett.* **77**, 5433-5436.
7. Galzitskaya, O.V. & Finkelstein, A.V. (1997). Theoretical investigation of folding rate dependence on the chain length for the RNA-like heteropolymers. *Mol. Biol. (Russia)* **31**, 488-491.
8. Thirumalai, D. & Woodson, S.A. (1996). Kinetics of folding of proteins and RNA. *Acc. Chem. Res. B* **29**, 433-439.
9. Finkelstein, A.V. & Badretdinov, A.Y. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold. Des.* **2**, 115-121.
10. Thirumalai, D. (1995). From minimal models to real proteins: time scales for protein folding kinetics. *J. Phys.* **5**, 1457-1467.
11. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.
12. Shakhnovich, E.I. & Gutin, A.M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA* **90**, 7195-7199.
13. Wolynes, P.G. (1997). Folding funnels and energy landscapes of larger protein within the capillarity approximation. *Proc. Natl Acad. Sci. USA* **94**, 6170-6175.
14. Fersht, A.R. (1995). Characterizing transition-states in protein folding - an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79-84.
15. Silow, M. & Oliveberg, M. (1997). High-energy channeling in protein folding. *Biochemistry* **36**, 7633-7637.
16. Privalov, P.L. (1979). Stability of proteins. Small globular proteins. *Adv. Protein Chem.* **33**, 167-241.
17. Filimonov, V.V., Privalov, P.L., Glangloff, J. & Dirheimer, G. (1978). A calorimetric investigation of melting of $tRNA^{Asp}$ from brewer's yeast. *Biochim. Biophys. Acta.* **521**, 209-216.
18. Zarrinkar, P.P. & Williamson, J.R. (1996). The kinetic folding pathway of the Tetrahymena ribozyme reveals possible similarities between RNA and protein folding. *Nat. Struct. Biol.* **3**, 432-438.
19. Laing, L.G. & Draper, D.E. (1994). Thermodynamics of RNA folding in a conserved ribosomal RNA domain. *J. Mol. Biol.* **237**, 560-576.
20. Nussinov, R. & Jacobson, A.B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. USA* **77**, 6309-6313.
21. Zuker, M. (1989). The use of dynamic programming algorithms in RNA secondary structure prediction. In *Mathematical methods for DNA sequencing.* (Waterman, M.S., ed.), pp.159-184, CRC Press, Boca Raton, USA.
22. McCaskill, J.S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**, 1105-1119.
23. de Peer, Y.V., Chapelle, S. & de Wachter, R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res.* **24**, 3381-3391.
24. Gluick, T.C. & Draper, D.E. (1994). Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.* **241**, 246-262.
25. Gutin, A.M. & Galzitskaya, O.V. (1993). Helix-coil transition in a simple model of large RNAs. II. Consideration of non-specific interactions. *Biofizika* **38**, 93-98.
26. Higgs, P.G. (1993). RNA secondary structure: a comparison of real and random sequences. *J. Phys.* **3**, 43-59.
27. Taketomi, H., Ueda, Y. & Go, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. *Int. J. Pept. Protein Res.* **7**, 445-449.
28. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1092.
29. Cantor, C.R. & Shimmel, P.R. (1985). In *Biophysical Chemistry, Part III.* W. H. Freeman and Company, San Francisco.
30. Hudson, D.J. (1964). *Statistics. Lectures of elementary statistics and probability.* CERN, Geneva.
31. Fersht, A.R. (1997). Nucleation mechanism in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.
32. Finkelstein, A.V. & Galzitskaya, O.V. (1996). Folding rate and stability of native structure in random and edited chains. *Mol. Biol. (Russia)* **30**, 91-96.
33. Lifshits, E.M. & Pitaevskii, L.P. (1981). *Physical Kinetics.* Pergamon, Oxford.
34. Galzitskaya, O.V. (1997). Geometrical factor and physical reasons for its influence on the kinetic and thermodynamic properties of RNA-like heteropolymers. *Fold. Des.* **2**, 193-201.
35. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460-471.