A new entropy model for RNA: part III. Is the folding free energy landscape of RNA funnel shaped?

Wayne Dawson,[1] Toshikuni Takai,[2] Nobuharu Ito,[2] Kentaro Shimizu,[1] Gota Kawai[2]

[1]Department of Biotechnology, Graduate School of Agriculture and Life Sciences, The University of Tokyo; [2]Chiba Institute of Technology, Japan

Abstract

The concept of a free energy (FE) landscape, in which the conformations of a polymer progressively take on the structure of the native state while spiraling down a FE surface that resembles the shape of a funnel, has long been viewed as the reason why a complex protein structure forms so rapidly compared to the number of conformations available to it. On the other hand, this landscape picture is less clear with RNA due to the multiplicity of conformations and the uncertainties in the current thermodynamics. It is therefore sometimes proposed that within the ensemble of suboptimal states of the RNA molecule, the vast majority of those states all closely resemble the native state and therefore simply overwhelm the few states that represent the global minimum FE. However, calculations of the free energy of observed structures often suggest that the most frequently observed cluster of structures are far from the minimum FE, particularly in the case of long sequences. If so, then such a FE surface is unlikely to be funnel shaped. We have been developing a version of *vsfold* that can evaluate the suboptimal structures of the FE surface (through a modified version called *vs_subopt*). Here we show that the ensemble of suboptimal structures for a number of known RNA structures can actually be both close to the minimum FE and also be the dominant observed structure when a proper Kuhn length is selected. Two state aptamers known as riboswitches can show neighboring FE states in the suboptimal structures that match the observed structures and their relative difference in FE is well within the range of the binding free energy of the metabolite. For the riboswitches and other short RNA sequences (less than 250 nt), the flow of the suboptimal structures (including pseudoknots) tended to resemble a rock rolling down a hill along the reaction coordinate axis. An important insight yielded by the cross-linking entropy (CLE) model is that the global entropy limits the size of domains. Hence, based on the CLE model, Levinthal's paradox is overcome by the funnel shape in the FE, by a reduction in the number of degrees of freedom due to Kuhn length, and by limits on the size of the domains that can form. These concepts are also applicable to calculating transition rates between different suboptimal structures.

Introduction

In Part I of this series, we showed that the global aspects of the cross-linking entropy (CLE) model satisfy the fundamental requirements of a heat engine in reversible processes. In Part II, we showed the role of the Kuhn length of a polymer in influencing the local entropy in structure predictions, the local contribution of the CLE. Also, in a recent report,[1] we showed that the CLE model is able to unify the lattice model, the Gaussian polymer chain model and the contact order model,[2,3] under the same framework. Likewise, in previous work, we have also shown how the application program we are continuing to develop (*vsfold5*) is able to successfully predict RNA secondary structure[4-6] and pseudoknots,[1,7] at least when a good choice of Kuhn length is employed (Part II, Section 7).

What has not been explored so far is the folding free energy (FE) landscape itself when using the CLE model. Is the landscape funnel shaped? In other words, is the optimal folded structure the final point of a process which is characterized by a series of suboptimal structures that are similar to the optimal structure and gradually descending through a funnel in the free energy landscape toward the global minimum. A lot of work has been devoted to studying the folding landscape of biopolymers, particularly for proteins,[8-17] where some have attempted to construct a FE model that includes the Jacobson-Stockmayer (JS) model discussed in Part I.[18] Some work has also been done for RNA,[1,19-24] which often defers to the standard JS-model.[20-23] However, the RNA problem has been less encouraging with the JS-model because known and observed RNA structures tend to be scattered within a mountain of suboptimal structures like a needle in a haystack, and there is no easily discernible pattern in the thermodynamic distribution of key suboptimal structures either.[6] With some RNA structures having passed through 3.5 billion years of natural selection to remove inefficiencies and instabilities in the RNA structures, relegating these *known and observed* RNA structures to positions deep within a list of suboptimal structures is rather incongruous. It may be possible that long RNA sequences require chaperones to fold correctly; however, the majority of shorter RNA sequences (like tRNA and riboswitches) should have solved this problem through natural selection. At least, it would be far more satisfying if a funnel shaped landscape could be shown for some representative

Correspondence: Wayne Dawson, Department of Biotechnology, Graduate School of Agriculture and Life Sciences, The University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan.
Tel.: +81.3.5841.5449 - Fax: +81.3.5481.8002.
E-mail: dawson@bi.a.u-tokyo.ac.jp, vsfold@gmail.com

Key words: entropy, RNA structure, suboptimal structure, thermodynamics, bioinformatics.

Contributions: WD wrote the manuscript and did the primary research and development of vs_subopt software; TT provided independent research results on SAM-riboswitch studies; NI provided research results on the tRNA structure studies of Thermus thermophilus HB8; KS, advice, guidance and support and assistance in preparing the manuscript; GK, advice, guidance and support in the software development of vsfold5 and vs_subopt.

Conflict of interests: the authors declare no potential conflict of interests.

Acknowledgments: this work was supported in part by a grant from Ministry of Education, Culture, Sports, Science and Technology (MEXT). The authors would like to thank Kouji Nakamura (Tsukuba University) for discussions and insights into the SAM riboswitch. We thank Dr. Shingo Nakamura (Catalent Pharma Solutions) for his many science related questions. We also remember the students at CIT who worked with vsfold5 and particularly Amiu Shino, Misaki Imai and Kenta Kondo. We also thank Profs Shugo Nakamura, Tohru Terada, and Kazuya Sumikoshi (UofT), Dr. YZ, RJ, and CS for their encouragement. Post Production Comment: We found over time that that best parameters for folding times were $w_{trans} \sim 100$ and $\tau_{chem} = 1e\text{-}5$ s.

Received for publication: 30 May 2011.
Revision received: 18 January 2012.
Accepted for publication: 6 February 2012.

Licensee PAGEPress, Italy
Journal of Nucleic Acids Investigation 2014; 5:2652
doi:10.4081/jnai.2014.2652

cases or the exceptions explained.

It has long been thought that the folding landscape for both RNA and protein folding is funnel shaped.[9,10,25] Here we added the functionality to calculate suboptimal structures into the *vsfold* program we have been developing (through a modified version called *vs_subopt*). Using the CLE model, we show that the folding landscape of RNA is indeed funnel shaped, at least for the representative and important RNA molecules we have tested.

It is also well known that some RNA molecules are used by organisms as two state switches that bind metabolites in order to signal other transcription processes to stop or start, depending on the particular system and its purpose. Assuming the FE landscape is funnel shaped, it is our hypothesis that both states should be relatively stable, where trapping of the metabolite should require only modest changes in the total FE, since binding affinities are likely to be modest in magnitude given the metabolite should be easily released when the somatic conditions are changed. This suggests that when the structure is unbound, it is oscillating between the two states (for some riboswitches, the structure is quite different in appearance) at a sufficient rate that is amenable to capturing the metabolite on a reasonable time scale for biology (certainly less than seconds for many processes and probably more like milliseconds). We show in this study that typical distributions of suboptimal structures of several known riboswitches do exist in largely two distinct states near the minimum FE and with a relatively small FE difference but a significant activation barrier. They therefore are oscillating between these largely distinct patterns with few competing alternative intermediates and therefore likely possess the expected properties of a two state switching device.

Due to the simplicity of the theory in the CLE model, we also show that domain sizes in RNA structures can be estimated based on the average base composition of the sequences. This is the only FE based model that predicts this based on a theoretical framework. The same framework also permits estimating the folding times between different suboptimal structures.

This is a theoretical model. This report is limited to a few *hand-picked* example structures not because these structures are the only cases that succeed; rather these examples represent what is generally the case and what is generally expected based on theoretical grounds.

Materials and Methods

Determination of suboptimal structures using *vs_subopt*

The general design and procedure used to calculate secondary structure and pseudoknots is explained in detail in Dawson *et al.*[7] Therefore, we will focus on the method for evaluating suboptimal structures. Let i and j represent the identity of two bases that are numbered in a sequence such that $i<j$ and each successive base is counted sequentially from the 5' to the 3' end of the sequence. As *vsfold* calculates the secondary structure and pseudoknot structure, the information about the best (optimal) structure is retained at each position i and j, which we denote by the ordered pair $(i, j)_o$. This residual structural information is used to help find the suboptimal structures.

The program searches in terms of junction points or levels in the hierarchy of secondary/pseudoknot structure. In Figure 1, a complex RNA structure including a pseudoknot is shown. Stems are defined in terms of a tail and a head of the structure, where the tail is the 5' and 3' most positions of the RNA sequence, Figure 1 (denoted in several places in Figure 1). Hence, if (i_h, j_h) is the closing bp at the head of stem and (i_t, j_t) at the tail, then $j_h - i_h < j_t - i_t$.

The zeroth level defines the arrangement of domains of secondary structure (or pseudoknots) and corresponds to the structures that extend off of level 0 in Figure 1. These are indicated by the labels *domain* 1, 2 (the pseudoknot) and 3 in Figure 1. In the Figure, the closing stem of domain 1 has the label [0] and we express the boundaries of this domain in this work by the notation $(i_t, j_t)[0]$, the tail of the closing stem for [0]. Level 1 of stem [0] begins at $(i_h, j_h)[0]$, in the region between $i_h < i, j < j_h$. For stem [0], the next level has stems labeled [0,0], [0,1] and [0,2], where the first index identifies the stem at level zero and the next index identifies the stem at level 1. Hence, [0,1,0,2] identifies a stem in domain 1 at level 3, connected to stem 0 of level 2, connected to stem 1 of level 1, connected to stem 0 of level 0. Likewise, the label [1] indicates the pseudoknot (PK), where its boundaries are $(i_t, j_t)[1]$. In the case of the PK, in the *vsfold5* program, pointers are used to register which stem is assigned [1,0] and which [1,1] and from there, the hierarchy is the same. A level 1 search will have the label of the form (i_h, j_h)[-,-], level 2 (i_h, j_h)[-,-,-], level 3 (i_h, j_h)[-,-,-,-], and so on.

At each level, the region enclosed by the corresponding $i_h < i, j < j_h$ is scanned for alternative structures. The resulting suboptimal structures are then grafted onto the calculated stem. For intermediate levels, the best zeroth level is selected first, then the best first level and so on. If

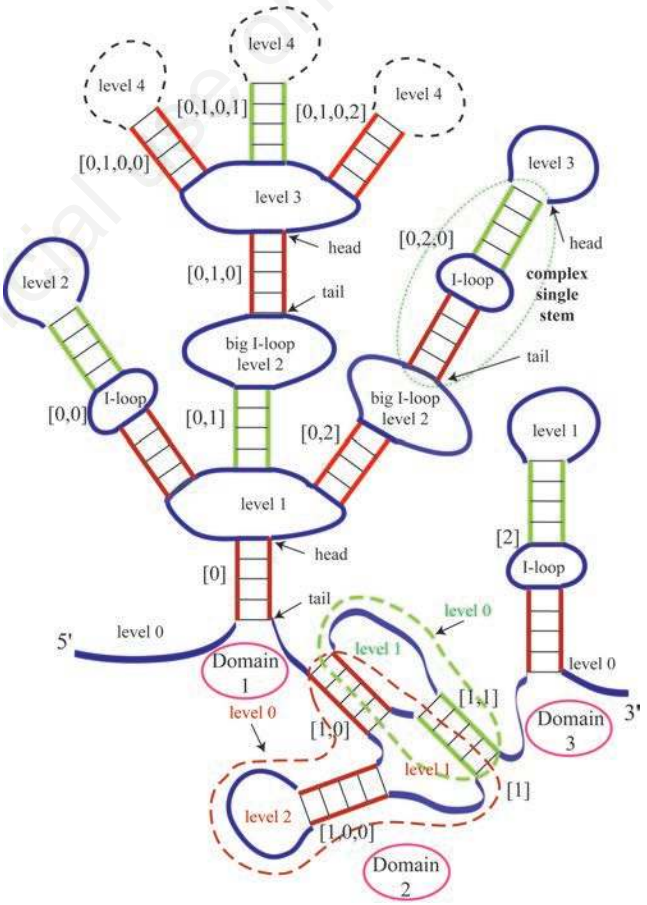

Figure 1. A schematic of the concept of levels in secondary structure and levels of the search. Level 0 represents the base domain of the structure. These are labeled domain 1, 2 and 3. The corresponding stems that close the domains are labeled [0]..[2]. All higher levels represent subdomains. These stems are expressed with progressively more indices, depending on the level. For example, stem [0,2,0] is at level 2 and [1,0] is at level 1. All searches except level zero begin from the head of the stem, level zero scans directly from the 5' to 3' ends of the sequence and begins with the optimal domain structure followed by successively suboptimal structures of the 5' to 3' sequence.

the list is exhausted, then the program switches to next best structure from level 0, and so on. It is also possible to specify constraints in order to search a particular region of a specified structure as long as the search region is defined in terms of an actual stem.

In the examples of stems, some stems are shown with small interior loops (I-loops) that are ignored whereas other larger I-loops are treated as new levels. Short I-loops are often considered part of a stem in *vsfold5*. The precise rules for how *vsfold* defines these composite stems are explained in the distributed manual that comes with the *vsfold5* distribution. In general, if the I-loop is short like a 1×1 or 2×2 I-loop, it is skipped in the evaluation of suboptimal structure because *vsfold* treats it as a bona fide stem.

The default setting of *vs_subopt* assumes the zeroth level as the desired suboptimal structures. This provides complete structures that are optimal for the sequence with a domain located between base *i* and base *j*. A fixed number of suboptimal structures are searched for until either no more structures are found, or the total number has been found within the default energy range (10.0 kcal/mol). The user can override the default settings by a variety of command line flags; *-so_level n* selects the level of interest (n=0,1,2..), *-so_max N* adjusts the number of suboptimal structures to be scanned (default, N=20) and *-FEspan E* sets the energy range (from the optimal structure) to be scanned (default, 10 kcal/mol).

Estimation of activation barriers and transition times

In general, a chemical reaction involving bond formation or Van der Waals interactions happen on a time scan of ps or shorter. On the other hand, folding of RNA structures involves time scales greater than μs, typically ms and sometimes minutes.[26-31] This is because different parts of a RNA sequence must diffuse together against the force of the chain entropy to form a specific base pairing interaction (explained in the model in part I, Section 5 and also in Dawson *et al.* and Cheng *et al.*).[1,6,16]

The formation of contiguous bps into a stem, which is known as the stacking process,[32] involves local diffusion processes and chemical interactions. From molecular dynamics simulations, one can observe a single bp that frays from the stem has a lifetime of approximately 5 to 10 ps. The process involves both the mechanical motion of the bp due to the local chain motions in solvent combined with chemical reaction when the stack is actually formed,[32,33] a process of Van der Waals coupling between adjacent bps and exclusion of water between the layers.[33] Hence the RNA chain folding interactions often happen on a much longer time scale than the stacking interactions but both are largely diffusion limited.

It follows that in the binding of base pair (bp) *i* and *j*, the dominant time scale in the rate will usually be the diffusion of different parts of the RNA chain together and local diffusion of the bases into stacks. For a diffusion rate k_{diff} and chemical reaction rate k_{chem}, the total rate (k) of a diffusion limited formation of chemical species (*i.e.*, an RNA bp)[34] is approximated by the following expression

$$\frac{1}{k} = \frac{1}{k_{diff}} + \frac{1}{k_{chem}} \approx \frac{1}{k_{diff}} \qquad (1)$$

where k_{chem} should be seen as a combined local diffusion of water during the stacking process, rearrangement of the RNA chain, and chemical reaction rate.

In the concept of the CLE model, the statistical mechanics entity is the stem. Stems in the structures should independently disassemble and reassemble as blocks of stems with a global entropy weight. Evidence suggesting this can be found in the force extension experiments where a stem unzips or re-zips as a unit,[35-37] and in differential melting experiments where one can assign the melting transitions to particular stems.[38-44] The global entropy places a strong weight on the folding time of a particular block.[6] Because these stems separate independently, in this scenario, the maximum activation barrier (and the rate determining step), is dependent on the stem that exhibits the maximum entropy loss in the process of joining the two different parts of the single-strand RNA chain together. Therefore, in a model where the unit is a stem, estimates of the transition energies should be a function of the Kuhn length.

The rate can be estimated by viewing the process as the sum of the mutually dependent folding times of the individual stems τ_s

$$\tau_{net} = \tau_1 + \tau_2 + \ldots \tau_n, \text{ with } \quad \tau_s = \frac{1}{k^s_{diff}} + \frac{1}{k^s_{chem}} \qquad (2)$$

where n is the total number of stems, the dominant process is the global CLE, the reference is the time it takes for the structure to fold from a denatured state to the native stem state and the longest folding time will be a function of the maximum length of the subsequence separating bp (i, j); $\max\{N_{ij} = |j - i|+1\}$.[6]

The folding rates are sometimes handled by considering all the RNA folding pathways using some approach based on the Morgan-Higgs algorithm.[20,21,45] Here, we are more interested in the overall time scale of the stem formation (formation of contiguous bps) in transforming between two different structures. For example, two general structures can be generated from the following toy example:

```
GGGGAAAACCCCUUUU
((((....)))).....   structure A
.....((((....))))   structure B
```

Structures A and B are mutually exclusive stems-loops that share no common base pairs (bps). One stem has to unfold for the other stem to form. The activation barrier is then found by computing the change in free energy required to un-pair stem A (deletion) and to form stem B (insertion); *i.e.,* the minimum number of editing steps that permits transition between two different structures of the same sequence. It should contain the total path because the transitions between structure A and B are in thermodynamic equilibrium and the one stem must unfold before the other can form. Therefore, to model the above simple process, we imagine the folding time to be the sum of the individual folding times for structures A (τ^{A}_{net}) and B (τ^{B}_{net}); *i.e.*, the sum of the two processes ($\tau^{A\leftrightarrow B}_{net} = \tau^{A}_{net} + \tau^{B}_{net}$). However, in general, we should expect some stems to be common to both structure A and structure B. Following Eq (2), let A∩B express those stems that are common to both structure A and structure B. Then the folding time for A↔B is expressed as the sum of the independently folded structures A and B minus the time for folding A∩B,

$$\tau^{A\leftrightarrow B}_{net} = \frac{1}{2}\left(\tau^{A}_{net} + \tau^{B}_{net}\right) - \tau^{A\cap B}_{net} = \frac{1}{2}\left(\sum_{s}^{n_A}\tau^{A}_{s} + \sum_{s}^{n_B}\tau^{B}_{s}\right) - \sum_{s}^{n(A\cap B)}\tau_s(A\cap B) \qquad (3)$$

where τ^{X}_{s} is the transition time for a given stem s of structure X, and τ_s(A∩B) is the folding time for a particular stems common to both structures A and B. Since the above toy-example contained no common stems, τ_s(A∩B)=0 and ($\tau^{A\leftrightarrow B}_{net} = \tau^{A}_{net} + \tau^{B}_{net}$). In general, however, we must identify the common parts of the stems.

Therefore, to obtain the folding time, we first need to find the minimum number of stems that must change; *i.e.,* the minimum number of stem-editing steps. This is done by comparing the similarity of the two suboptimal structures. We then assume a sequential set of steps in which the structure becomes partially unfolded in order to transition to the alternative structure. In this case, the folding rate is dependent on the total time for each of the stems (or parts of a stem) to come apart.

The diffusion limited contribution to the transition rate for a single stem can be estimated to be:[34,46]

$$k_{stem} = \omega_{trans}\frac{k_B T}{h}\exp\left(-\frac{\Delta G_{stem}}{k_B T}\right) \qquad (4)$$

where k_B is the Boltzmann constant, h is the Planck constant, T is the temperature, ω_{trans} is the transmission coefficient (related to activities) and ΔG_{stem} is the FE (including the CLE) to remove a stem. The transmission coefficient here is likely to be a function of various unknown activity coefficients and associated concentration.[34,47] For a polymer in solution, ω_{trans} is probably at least a few orders of magnitude smaller than 1. For RNA, ω_{trans} has been lumped together with k_BT/h and called the pre-exponential factor $\alpha=\omega_{trans}\ k_BT/h$.[48,49] At present, there are no actual values for ω_{trans}. However, given the maximum folding rate of proteins (and perhaps RNA)[50] is on the order of microseconds,[51] this can be tuned for a single stem. For stems, because ΔG_{stem} is much larger than the bp FE, the correction is on the order of $\omega_{trans} \approx 10^{-2}$ to 10^{-3} compared to 10^{-6} for a single bp.

Since the transition time is found in the inverse rate, we write

$$\frac{1}{k_{net}} = \frac{1}{k_1} + \frac{1}{k_2} + \cdots + \frac{1}{k_n} = \frac{1}{\omega_{trans}} \sum_{s=1}^{n} \frac{h}{k_B T} \exp\left(\frac{\Delta G_{stem}^{(s)}}{k_B T}\right) \tag{5}$$

In general, this would require a detailed evaluation of all the stems and determination of the changes in FE for each case. However, if the Kuhn length is essentially constant for a region of the sequence, we can make a simpler estimate. In part II, we showed that the Kuhn length and the stem length should be essentially the same value. In such a case, all the stems are assumed to be of similar length and the total number of stems that change can be estimated from the total number of base pairs that must be removed divided by the Kuhn length,

$$\Delta n_{stem} = \Delta N_{bp}/\xi \tag{6}$$

where ΔN_{bp} is the minimum number of editing operations that are required on the base-pairs (bp) to achieve the transition and ξ is the Kuhn length. Likewise, the average energy of each stem is just the sum of the FE from all the bps that are different, or the FE of stems that change, divided by Δn_{stem},

$$\left\langle \Delta G_{stem} \right\rangle^{\overline{A\cap B}} = \frac{\sum_{\Delta(bp)} \Delta G_{bp}^{\overline{A\cap B}}}{\Delta n_{stem}} = \frac{\xi \sum_{\Delta(bp)} \Delta G_{bp}^{\overline{A\cap B}}}{\Delta N_{bp}} \tag{7}$$

where $\overline{A\cap B}$ specifies those stems in structures A and B that are not shared in common (the complement of $A\cap B$), $\Delta(bp)$ refers to the associated bps in $\overline{A\cap B}$ ($\Delta(bp)\in\overline{A\cap B}$), and $\Delta G_{bp}^{\overline{A\cap B}}$ refers to the FE of $\Delta(bp)$, where both stacking and the global CLE are included.

Hence, the transition requires a larger energy than a single base pair, which is certainly too small and would suggest far too rapid a rate than a stem requires. Likewise, the transition energy is typically much smaller than the total change in energy of all the editing steps, which is often far too large and would require unreasonably long time intervals. It follows that the total time for a transition from structure A to structure B will simplify to

$$\begin{aligned} \left\langle \tau_{net} \right\rangle^{A\leftrightarrow B} &= w_{trans} \frac{\Delta n_{stem} h}{k_B T} \exp\left(\frac{\left\langle \Delta G_{stem} \right\rangle}{k_B T}\right) + \Delta N_{bp} \tau_{chem} \\ &= \Delta N_{bp} \left\{ w_{trans} \frac{h}{\xi k_B T} \exp\left(\frac{\xi \sum_{bp} \Delta G_{bp}}{\Delta N_{bp} k_B T}\right) + \tau_{chem} \right\} \end{aligned} \tag{8}$$

where τ_{chem} is $1/k_{che}$ (Eq 1) and $w_{trans}=1/\omega_{trans}$ (Eq 4). Currently, we make the crude estimate that τ_{chem} is about 10 ns (based on estimates in Pocschke *et al*. and might be as large as μs for a whole stem),[33,52] and w_{trans} is 100, based very roughly on activities.

The rate of the reaction will be the inverse

$$\left\langle k_{net} \right\rangle^{A\leftrightarrow B} = \frac{1}{\left\langle \tau_{net} \right\rangle^{A\leftrightarrow B}} \tag{9}$$

Hence, the transition energy of a single stem changing is

$$\left\langle \tau \right\rangle_{stem}^{A\leftrightarrow B} = \frac{h}{\omega_{trans} k_B T} \exp\left(\frac{\left\langle \Delta G_{stem} \right\rangle}{k_B T}\right) + \Delta N_{bp} \tau_{chem} \tag{10}$$

Since base pair energies range from –0.5 to –2.5 kcal/mol, for a ξ=5 nt, $\langle \Delta G_{stem} \rangle$ is approximately of the order of –2.5 to –12.5 kcal/mol. Using Eq (8), this computes to a transition time ranging between 10 ps (with τ_{chem}=0) to 10 μs at 310 K. For ξ=10 nt, $\langle \Delta G_{stem} \rangle$ (including the CLE contribution) is approximately of the order of –5 to –25 kcal/mol, which corresponds to a transition time of ranging between 500 ps to 14 h (at 310 K). It seems that 14 h is rate limiting for most biological processes.

The model used here is consistent with the view that the transitions tend to be cooperative in RNA. Co-operativity is well known in the melting of proteins.[53] For RNA structures, this is seen in the peaks in differential melting curves where particular stems can be assigned to specific melting temperatures in small RNA molecules like tRNA,[38,39] and pseudoknots.[40-43] Likewise, it can be seen in the force-extension experiments using optical-tweezers, where the stems suddenly unzip or re-zip (refold).[36,37,54] These experiments all suggest that RNA unzips and refolds by stems, not by bps. This can be understood as a direct consequence of the Kuhn length, which specifies that the bases act collectively as a group.

To estimate transition time between different suboptimal structures, we will use this method of estimation. The main economy of this strategy is i) that we don't have to calculate all the folding pathways or even the specific stems of a given pair of suboptimal structures, and ii) that we can establish a baseline for the folding times that avoids any systematic issues of any particular algorithm that models the folding pathways. The main deficiencies in this estimation strategy are i) that the CLE model currently only uses one Kuhn length for a whole sequence, where many of the active RNA structures of interest clearly have variable stem lengths and therefore different Kuhn lengths, and ii) that the calculated value is the average stem FE, not the specific FE of any particular stem. The method proposed here, therefore, should be seen as a concept in which an average Kuhn length is applied over the entire sequence. This is only proposed as an estimation technique. Future work on *vsfold* and *vs_subopt* will attempt to address variable Kuhn lengths in the computation of RNA structure and are intended to adhere closer to the precise concept of Eq (3) through (5). Nevertheless, this estimation approach should provide a ball-park approximation of these transition times.

Results and Discussion

Figure 2A shows the result of a calculation of suboptimal structures within 10 kcal/mol of the minimum free energy using ξ=5 nt and level 0 search for yeast tRNA(Phe) based on the unmodified sequence of tRNA(Phe).[55-58] The results are sorted by the predicted free energy (FE) with the minimum free energy shown at the top. Figure 2B groups the folding landscape according to similar folding structural intermediates. The optimal structure (minimum FE) is the familiar cloverleaf pattern of tRNA (Figure 2B, right most structure). The D-stem (Figure 2A, purple) and anticodon stem (Figure 2A, green) are already present in the structure even 10 kcal/mol away from the native state. They are preserved throughout the list including the optimal structure (the observed tertiary structure) at the top. The persistence or conservation of similar stems throughout the list (*i.e*., the structural homology) of the T-stem (Figure 2B, blue) is apparent in many of the scans.

The dominant progression of homologous structures as a function of energy is shown on the top side of Figure 2B with the red arrows and

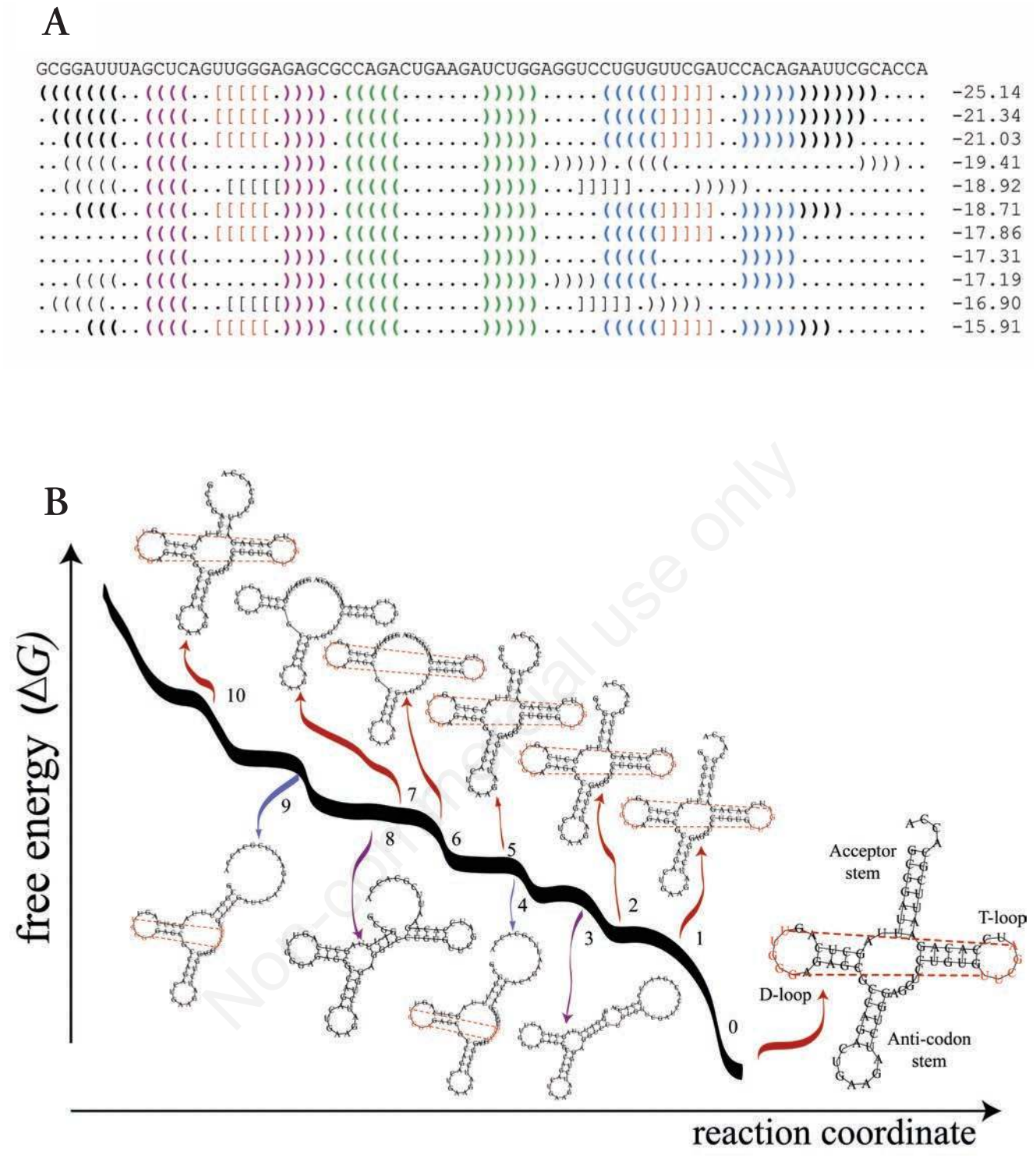

Figure 2. Calculations of the suboptimal structures including pseudoknots for tRNA(phe). A) The list of the predicted suboptimal structures (including the free energies) that are within 10 kcal/mol of the optimal structure listed in the order of the predicted energies. The RNA structures are listed in the Fontana-Schuster tree-notation plus a bracket notation [[..]] for the pseudoknot notation. The bold colors represent known parts of the RNA structure and are labeled on the optimal tRNA structure: (purple) the D-stem (closing the D-loop), (green) the anticodon stem, (blue) the T-stem (closing the T-loop), and (black) the Acceptor stem. The red stem is known to occur in the 3D structure of tRNA. B) A rough schematic depiction of different secondary structures in the order that they appear in the calculated suboptimal structure list in A), where the right most labeled tRNA structure represents the optimal predicted structure. The structures on the top (red arrows) represent the corresponding structure's position and free energy on the approximate reaction coordinate. The purple and magenta arrows represent structures not directly on the reaction coordinate. The arrows propose some possible points where the improper stem unfolds and the structures refold and join the structures along the reaction coordinate.

numbers (10>7>6>5>2>1>0) corresponding to the order in the list of suboptimal structures in Figure 2A. The bottom side of Figure 2B shows the progress of two alternative pathways: 9>4 and 8>3. The majority of structures on the top side follow a natural progression along a gradient down to the native state and minimum free energy, corresponding to a natural energy-to-structure progression. The exception is structure 10, which has a higher FE because the acceptor stem is considerably shorter than the Kuhn length (ξ=5 nt) leading to local instability (Part II of this series). The alternative pathways (bottom of Figure 2B) are far fewer and those that appear have several features in common with the native state. Considering the folding times, these alternative pathways most likely connect somewhere with the general progression or become suppressed somewhere within the last 5 kcal/mol of the minimum FE.

Whereas *vsfold5* works in a 5' to 3' folding, *vs_subopt* folding looks more akin to a denature/refolding experiment where all parts of the sequence fold and compete with each other. The predicted transition rates from the misfolded structures to a neighboring structure along the path of the funnel (Figure 2B) are all on the order of μs for this RNA; *e.g.*, structures 3 and 4 ending up at structure 2 and structures 8 and 9 ending up at structure 7. This may be a little fast, but even 1000 fold slower rate would still render these transitions on the order of ms. The results of experiments on tRNA(phe) are consistent with the assumption that any intermediates that might form during the folding process do not contributed significantly to slowing down the folding rate. These experiments also assign the first two stems-loops (D-loop and anticodon stem) to the most stable structures.[38,39] The fact that these two stem-loops appear in every example is consistent with this observation.

The perspective of the CLE model, and therefore *vs_subopt*, is that the experimentally observed structure should also be a structure that is at the minimum FE. Trapping in structural intermediates is certainly possible. However, the experimentally observed RNA is the minimum FE and the intermediate should be at most an obstruction along the path. Even for tRNAs, *vsfold* can fail to find the cloverleaf structure at the minimum FE; however, the reasons are more likely to be issues like non-Watson-Crick pairing in the RNA stems.

The progression sorted in terms of energy follows a dominant pathway to the bottom of the free energy landscape and has a dominant structural homology that follows a natural folding progression. Even considered with multiple pathways, the dominant progression follows a reaction coordinate that resembles a rock rolling down a hill. Therefore, considered in terms of a dominant structural homology, there is a clear reaction coordinate that can be discerned from these calculations. What is important is not that a set of structures are sorted in terms of their respective free energies, rather the sequence of suboptimal structures follow a dominant pattern that, in effect, points to the native structure.

Figure 3 shows the results for rodent tRNA(Ala),[59] for which the optimal structure has been reported in previous work.[1,4,7] Again, the optimal structure is the tRNA structure and within 10 kcal/mol of the minimum free energy, the majority of the structures correspond to the expected cloverleaf structure: the D-stem (purple), anticodon-stem (green), T-loop (blue), and acceptor-stem (black, bold face).

Contrary to the notion of having to search an ensemble of structural intermediates existing in non-equilibrium states of the molecule far from the minimum free energy using very sophisticated bioinformatics tools to find the most frequently encounter cluster of structures, the native fold with the largest set of related intermediates simply falls in line showing a regular progression. In Figure 2B, four structures are shown below the curve depicting the FE along the rough schematic depiction of the 1D reaction coordinate; two with purple arrows (structures 4 and 9) and two with magenta arrows. The arrows (pointing from positions along the suggested reaction coordinate) are meant to propose some possible points where the improper stem unfolds and the structures refold into one of the neighboring structures along the suggested 1D depiction of the reaction coordinate. These structures, which do not closely resemble the native state, may persist for some time as *blind alleys*. Hence, the predictions suggest some potential trapped intermediates that may form and slow down the folding process. However, in the last three structures where the acceptor stem forms (upper side, red arrows), the structures all have the same major features as the native state. Moreover, the key structural features of the tRNA are largely conserved throughout the set of structures within 10 kcal/mol of the minimum. Melting studies of the tRNA do not show sig-

Figure 3. Calculations of the suboptimal structures including pseudoknots for tRNA(Ala) displayed in Fontana-Schuster notation. The optimal structure is shown in at the top of the list. The bold face markings indicate structures associated with the observed tRNA structure. The pseudoknot is not predicted because there are too many non-Watson-Crick pairs in the structure to make a clear distinction with current thermodynamic parameters. The coloring for the stems is the same as used in Figure 2.

nificant trapping in intermediate states for these simple RNA molecules,[38,39,60-63] though perhaps there is a very small fraction that was missed.[64] In Figure 3, the tRNA(Ala) shows two major structural variations, including an incorrect anticodon-stem, T-stem and acceptor stem. However, after the cloverleaf structure begins to form, the frequency of these alternative structures is greatly suppressed in favor of the correct anticodon- and T-stem. Therefore, with a good model for the entropy and suitable thermodynamic parameters such as Kuhn length, the structures largely fall into place. This shows that RNA is every bit as capable at finding the native state as proteins.

Moreover, this tendency is not unique to these two specific examples. We previously used *vsfold* to analyze all the tRNA sequences found in the genome of *Thermus thermophilus* HB8, and found that 80% of those structures were predicted with the expected cloverleaf structure. Failure was typically the result of neglecting non-Watson-Crick (non-WC) pairing, for which there are no established thermodynamic parameters at present. Using *vs_subopt*, given there is nothing particularly unusual about the sequence, we can often obtain the cloverleaf and a strongly reinforced homology resembling the native state all the way through the folding landscape and dominating most of the thermodynamically stable structures.

It is notable that there are only a handful of structures that lie with the 10 kcal/mol range and the majority of them point toward the native structure. There are two reasons for this. First, the analysis is considering domains of RNA structure in this search (*i.e.,* a level 0 search). We specified level 0 because our interest was in the primary domains that form, not the detailed fluctuations that will happen inside of a domain in level 1 or higher. If we request level 1 instead, the cloverleaf structures would show greater variations; however, we already see this variation at level 0 in the structures that are 10 kcal/mol away from the minimum FE. The tRNA(Phe) has very little variation in the stems of the cloverleaf pattern whereas tRNA(Ala) has at least two types of structures that fluctuate in the suboptimal structure list. Therefore, examining the variations in the structures that happen at level 0 can also reveal the overall variation in the resulting structures. Second, we must consider the implications from Part II and from discussions in previous work about the Kuhn length.[1] As shown in Dawson *et al.*,[1] the number of conformations is a function of the Kuhn length (ξ) and is of order (N/ξ); the conformation space decreases factorially with increasing ξ. This also applies to simple base-pairing combinatorics (*i.e.*, paired or unpaired), where $2^{(N-1)}$ combinations of bps reduced to $2^{(N-1)/\xi}$ combinations of stems.

Furthermore, the CLE model also limits the size of the domains that can form.[5,6] If the number of degrees of freedom were equal to the number of monomers and the entire search space must be searched, it is easy to show using Levinthal's paradox,[65] that even a fundamental RNA structure like ribosomal RNA 16S (approximately 1500 nt) would take longer than the lifetime of the universe for a search of all the folding conformations.[17] Yet even given the large reduction (order N/ξ) in the number of degrees of freedom, scaling by N/ξ only changes the length scale. For example, we can simply propose a sequence with a structure that is ξ times longer in length and we will return to the same dilemma. Let p represent the ratio of the number of bps (N_{bp}) divided by the sequence length (N), $p=N_{bp}/N$, where $0\le p \le 1/2$. Based on Dawson *et al.*,[1] the global entropy can be approximated as:

$\Delta G(N) \approx pNk_BT\{\gamma(\ln(\Psi N)-1) - (\gamma+1/2)\} = pNk_BT\hat{g}(N)$

where $\hat{g}(N)$ is the dimensionless function in the brackets and $\Psi=\xi/\lambda^2$ with $\lambda\approx 2$ the chain-chain separation distance between the bps.[1] Let $\Delta H_{bp} \approx -pN\overline{h}$ where $\overline{h}$ the absolute value of the average enthalpy of the base pairs in the sequence and let $\hat{h}(T)=\overline{h}/(k_BT)$ represent the dimensionless form of this average enthalpy. The local entropy correction to the FE for a free strand structure with a fixed Kuhn length (Part II) is

$$\Delta G_{l,fs}(\xi_{fs}) = (\gamma+1/2)\, Nk_BTf(\xi_{fs})/(D\xi_{fs}),$$

where $\xi_{fs}=3$ nt approximates the Kuhn length in the free strand regions. The correction for stems in the structure (from Part II, Section 8, Eq 43) is

$$\Delta G_{l,bp}(\xi_{ds}, p) = (2pN)(\gamma+1/2)\, k_BT\,(f(\xi_{ds})/\xi_{ds}) - f(\xi_{fs})/\xi_{fs})/(D\xi_{ds})$$

where ξ_{ds} nt is the Kuhn length in the free strand regions. Hence, substituting $\xi_{fs}=3$ and $\xi_{ds}=\xi$, the local correction is

$$\begin{aligned}\Delta G_l &= \Delta G_{l,fs}(\xi_{fs}) + \Delta G_{l,bp}(\xi_{ds}, p)\\ &= \frac{(\gamma+1/2)Nk_BT}{D}\left\{\frac{f(3)}{3}(1-2p) + \frac{2pf(\xi)}{\xi}\right\}\\ &= (\gamma+1/2)Nk_BT\hat{g}_l / D\end{aligned} \tag{11}$$

where, by inspection, if $p=1/2$, then Eq (11) becomes $\hat{g}_l=f(\xi)/\xi$, and if $p=0$, then $\hat{g}_l=f(3)/3$. The approximate FE of a domain of length N is therefore

$$\begin{aligned}\Delta G &= \Delta H_{bp} + \Delta G_{bp}(N) + \Delta G_l\\ &= Nk_BT\left\{-p\hat{h}(T) + \frac{1}{\xi}\left[p\left(\gamma\left(\ln(\Psi N)-1\right) - \left(\gamma+1/2\right)\right) + \frac{\xi\left(\gamma+1/2\right)\hat{g}_l}{D}\right]\right\}\end{aligned}$$

Evaluating the derivative with respect to N, a stationary point is found at

$$N = \frac{4}{\xi}\exp\left\{\frac{\xi\hat{h}}{\gamma} - \frac{\left(\gamma+1/2\right)}{\gamma}\left(\frac{\hat{g}_l}{pD} - \frac{1}{\xi}\right)\right\} \tag{12}$$

where N will become too large to sustain further growth.

Maximum domain sizes based on Eq (12) are tabulated in Table 1 for 37°C and $p=0.25$, where we assume that half the potential bps form stems in any random selection of RNA sequences. Typical values of $\overline{h}=k_BT\hat{h}(T)$ should range on the order of 2 kcal/mol (corresponding to similar fractions of A, C, G and U with AU, GC and GU pairing) and the

Table 1. A list of estimates for the maximum domain size of RNA (in units of nucleotides [nt]) given an average base-pair binding free-energy (C_2: estimated from averaging the Turner rules for the stems in a particular sequence) and the Kuhn length (ξ) of the stems in the domain.

Base pairing	Average bp weight (C_2) [kcal/mol]	Kuhn length (ξ) [nt] 3.0	5.0	7.0	9.0
AU rich	1.5	94 [nt]	669	6530	7.33E+04
AU rich	2.0	374	6700	1.64E+05	4.64E+06
AU rich	2.5	1491	6.71E+04	4.14E+06	2.94E+08
GC rich	3.0	5940	6.72E+05	1.04E+08	1.86E+10

average Kuhn length is around 5 bps. Hence, Table 1 indicates that most RNA has a maximum domain size that can range from about 300 nt to 6000 nt, where a reasonable value is likely around 500 to 1000 nt. When $p \rightarrow 0$, Eq (12) vanishes. Around $p=0.25$, the predicted domain size does not increase ($p>0.25$) or decrease ($p<0.25$) dramatically, so there is little gain or loss by changing this quantity around $p=0.25$. This is due to the global FE costs of bp formation. It is also clear from Table 1 that it might be possible to make very large domains using GC rich sequences. Perhaps a relatively equal amount of ACGU is favored because it maximizes the randomness and therefore the amount of information that can be stored or perhaps a synergy in coupling with different biological processes requires time scales that favor smaller domains than the maximum conceivable.

The Kuhn length tends to drastically reduce the number of degrees of freedom but the global entropy (a function of the Kuhn length) sets limits on the size of the *relevant* search space. In combination, it becomes possible to estimate folding times that are consistent with the observed biologically relevant folding time-scales.[17] Levinthal's paradox is overcome by i) the funnel shaped FE landscape, ii) reduction of the number of degrees of freedom, and iii) the limits on the domain size due to faster growth of the global entropy compared to the base-pair free energy (Turner rules).[65] The last point is not predicted by any other method.

Riboswitches are an example of a type of RNA that does not fold into a single type of molecule, but must exist in two different states. To have a reasonable likelihood of capturing a metabolite, the molecule likely spends some time in a configuration close to its cognate structure. Further, the riboswitch should be able to release that metabolite when somatic conditions change. Hence, both states should be present in the observed two-state system, where the molecule hops back and forth through a FE barrier. In such a case, the folding landscape is likely to have both structures coexisting near the minimum FE with an activation barrier that prevents the structure from spending too much time in the alternative state, but with enough time that a metabolite can find the structure in a desirable configuration (perhaps like many biological systems, a rate of about ms^{-1} to μs^{-1}). Therefore, a good test of the CLE model is to see if both states of the riboswitch are close together in the list of suboptimal structures and that the transition times are within a few ms at most.

Figure 4 shows two states of the *Vibrio vulnificus add* Adenine riboswitch.[66,67] Figure 4A is the optimal structure (#0) and is similar to the structure proposed for the bound metabolite (Adenine). However, the P1 stem is only partial. Figure 4B is the first suboptimal structure (#1) and represents the unbound structure where stem P1 is lost,[68] with $\xi=6$ nt. The two structures have rather similar FEs ($|\Delta\Delta G|=0.1$ kcal/mol). Figure 4C,D compare the optimal structure (#0) and the second suboptimal structure (#2) for $\xi=7$ nt. For the $\xi=7$ nt case, the full P1 stem is found in the minimum FE structure and lost in the #2 suboptimal structure. Both Figure 4B and D show the same structure and contain the translation repression stem (right hand side of the structure). In Figure 4C,D, the FE difference between the structures #0 and #2 is $|\Delta\Delta G|=6.0$ kcal/mol, which is a little high, but part of this is because the Kuhn length is too long for the structure on the left hand side of Figure 4A. The CLE model is the only approach that even considers the stiffness of the RNA in the FE calculation, and thus it provides more information on the nature of the RNA.

From Eq (10), the predicted transition time is about 1 μs, with $w_{trans}=100$ and $\tau_{chem}=10$ ns for the two different structures. The alternative structures within 2 kcal/mol all tend to have a slower formation time by a factor of 4. With such a rate of fluctuation, diffusion of the metabolite (about μs to ms within the volume of a cell)[34,69-71] and binding energy difference (more than 2 kcal/mol) render this a system sufficiently rapid in fluctuation to uptake stray metabolites and slow enough to stall unwanted translation. In the determined tertiary structure, the pseudoknot in Figure 4A was found. At present, it is not known if the pseudoknot in Figure 4C exists.

Figure 5 shows the suboptimal structures of the *Bacillus subtilis xpt* Guanine riboswitch.[67,72,73] Figure 5A shows the optimal structure (#0) with $\xi=5$ nt, which resembles the observed bound state. Figure 5B is suboptimal structure (#3) with $\xi=5$ nt, which resembles a major part of the unbound form. With $\xi=5$ nt, the anti-terminator stem and shift between terminator and anti-terminator is not observed. The energy difference between the unbound structure (minimum free energy) and bound metabolite structure is on the order of 3 kcal/mol; within a necessary range of energy that a binding metabolite can successfully utilize. For $\xi=8$ nt, Figure 5C shows the optimal structure (#0) with the very stable terminator stem. Figure 5D shows the corresponding suboptimal structure (#11) with the anti-terminator stem for the same $\xi=8$ nt. The energy difference is about 11 kcal/mol, which is rather large. Whereas the pseudoknot structure appears in several suboptimal structures of lower energy difference from the optimal and at various Kuhn lengths between 5 and 8 nt, the full anti-terminator is not so easily obtained and appears as structure #11 on the list (in a level 0 search). The estimated transition time is about 1 ms in these calculations (with $w_{trans}=100$ and $\tau_{chem}=10$ ns at 37°C). The X-ray structural studies of the Adenine riboswitch and Guanine riboswitch do not contain the regulating region having the translational initiation signal or the transcriptional termination signal.[67,72] We therefore cannot determine to what extent, if any, the pseudoknot interaction is present in Figure 4C or Figure 5D. According to Garst *et al.*,[74] the restructuring around the metabolite causes downstream changes in the structure which result in the formation of a the transcriptional termination signal (A-riboswitch) or the antiterminator stem (G-riboswitch). In Figure 5E, the lowest five out of six structures ($\xi=5$ nt) have the P1 stem and the #2 structure does not. The structures containing the P1 stem resemble the bound structure and #2 resembles the structure in the unbound state, with the exception of an additional stem.[68]

We use the term *bound* and *unbound* because in the case of the Adenine riboswitch the bound state is in the on position and for the Guanine riboswitch, it is in the off position.

The results of a study on the *Bacillus subtilis ydhL* Adenine riboswitch are similar and will be reported in a future study. Both the Guanine- and Adenine riboswitch provide examples of quantitative estimates of the transition energies and highly stable structures with a number of persistent native folds within 10 kcal/mol of the minimum FE.

Figure 6A shows the first eight suboptimal structures for the S-adenosylmethionine binding riboswitch, the SAM riboswitch[75,76] from the *Bacillus subtilis yusC* gene (or *yusCBA* operon).[77] In this case, structure #1 contains the terminator stem and structures #4-#7 contain the anti-terminator stem, with the FE difference in the two structures in reasonable proximity between structure #1 and structure #4 ($|\Delta\Delta G|=1.7$ kcal/mol), Figure 6B,C. The *vs_subopt* application (and therefore the CLE model) is largely partitioning these structures into distinct states, something we would expect for a two state system. The SAM riboswitch is more difficult to evaluate with a fixed Kuhn length because, unlike tRNA or the Adenine- and Guanine riboswitch structures, the stem lengths of the SAM riboswitch vary even more extensively. To set a value for the Kuhn length for the SAM riboswitch, we tried five combinations of Kuhn length (ξ) and minimum stem length (L_S) with $\xi/L_S=\{6/3, 8/3, 8/4, 9/3, 9/4\}$ (the command line options -xi ξ and -xi_min L_S), and found that 9/3 yielded the best results. As explained in Part II, the Kuhn length is a function of this stem length. Therefore, the Kuhn length should vary with the stems that are present. Nevertheless, as also shown in Part II, even a poor choice of Kuhn length can still sometimes yield success in a robust model.

The estimated transition time between the two structures in Figure 6B,C is about 30 ms, which may be rather fast based on current available information.[74] However, the folding rate is very sensitive to Kuhn length, which is a function of the stem length (as shown in Part II). A 20%

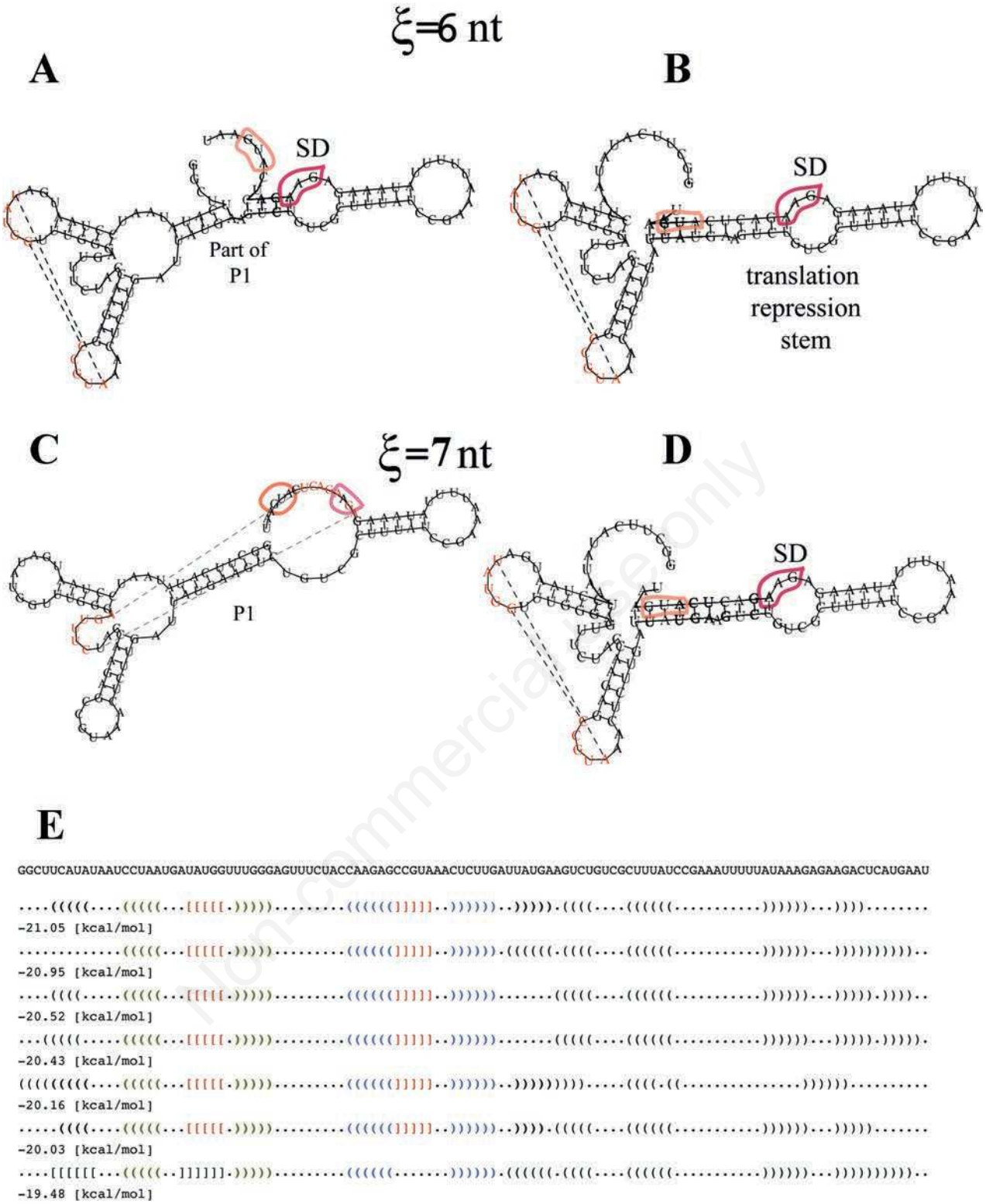

Figure 4. Suboptimal structure results for the Vibrio vulnificus add Adenine riboswitch. A) The optimal structure (#0) for ξ=6 nt, where part of P1 is present and the structure resembles most of the features of the unbound structure. B) The first suboptimal structure (#1) for ξ=6 nt, where the P1 is removed and the structure resembles the bound form. The purple circled region represents the Shine-Dalgarno (SD) GAA sequence and the orange circled region represents the initiation codon. C) The optimal structure (#0) for ξ=7 nt, where P1 is complete and the secondary structure matches the unbound riboswitch. Here, a pseudoknot closes off the SD region but leaves the initiation codon free. D) The same suboptimal structure as in B) with ξ=7 nt. E) A list of the suboptimal structures and free energies within 2 kcal/mol of the minimum FE for this riboswitch (with ξ=6 nt), where the top most structure is the minimum FE and the FE is sorted in increasing order. The two states of this riboswitch are represented by the first and second suboptimal structure (energy difference 0.10 kcal/mol).

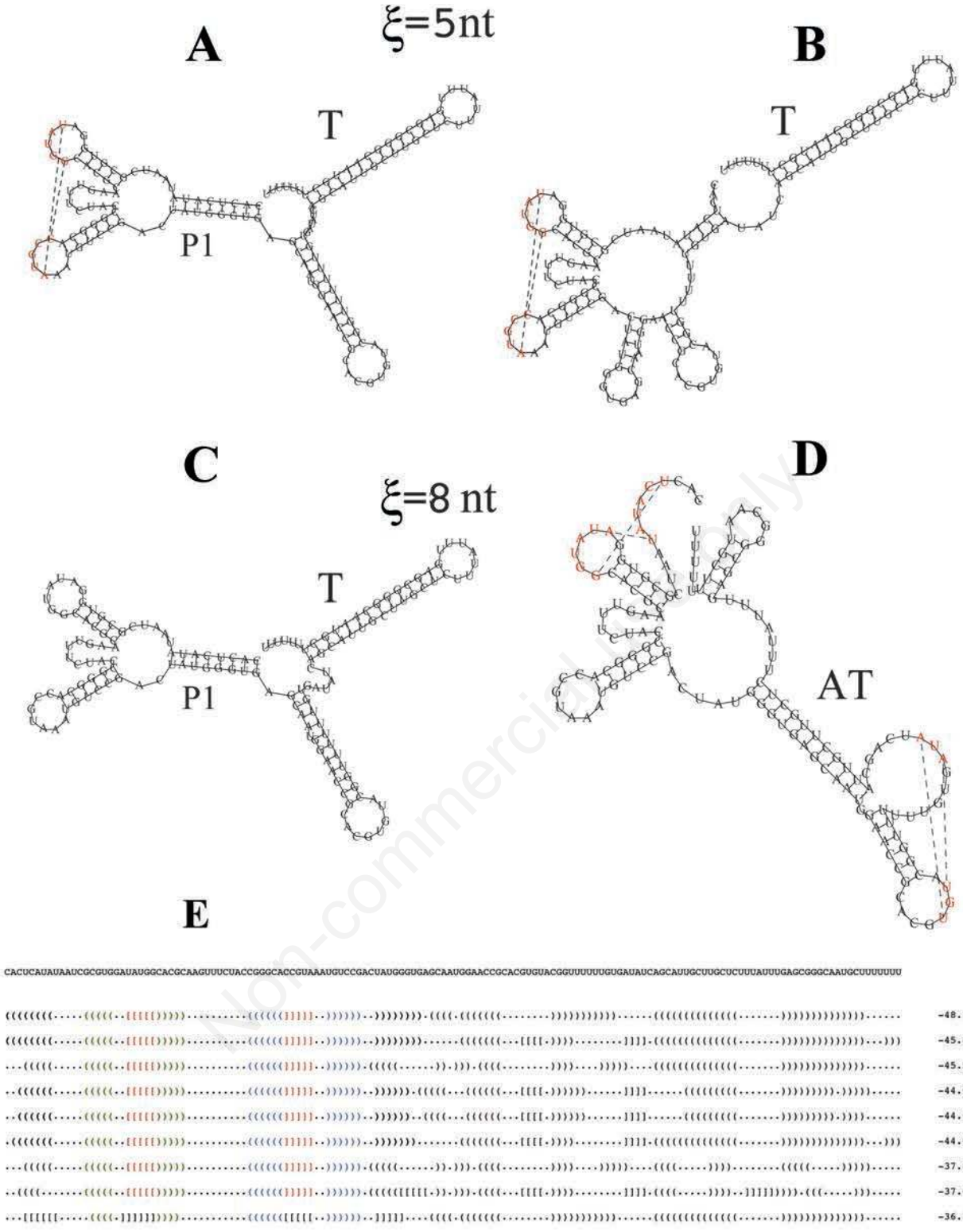

Figure 5. Suboptimal structure results for the Bacillus subtilis xpt Guanine riboswitch. A) Optimal structure (#0) at ξ=5 nt where the switch is in the bound form, where the label T is for the terminator stem. B) The second suboptimal structure (#2) at ξ=5 nt where the switch approximates the unbound form without the P1 stem. Here, the terminator stem is not removed. C) Optimal structure (#0) at ξ=8 nt, which is similar to A) except for the pseudoknot and represents the bound form. D) Suboptimal structure (#11) at ξ=8 nt, which resembles most of the features of the unbound form including the anti-terminator stem (AT). E) The suboptimal structures and free energies that are within 10 kcal/mol of the minimum free energy for this riboswitch for ξ=5 nt, where the top most FE is the minimum and the FE (right most) increases down the list. P1=paired stem 1, etc., T=terminator and AT=anti-terminator. The difference to remove the P1 stem is represented by the #0 and #2 structures (energy difference 3.24 kcal/mol with ξ=5 nt). To completely break up the terminator stem and remove the P1 stem, a transition energy of 11 kcal/mol is needed.

increase in the Kuhn length of the terminator stem would produce a transition time on the order of seconds. Furthermore, the transmission weight is currently unknown and may be much larger than we have estimated.

A total of 10 SAM-ribowitches were tested and similar observations were obtained. The results will be reported elsewhere.

SAM riboswitch structures similar to those shown in Figure 6A can also be found by simply recalculating using a different Kuhn length. Since binding a metabolite is likely to change the stiffness of the structure (the Kuhn length), this means changes in the Kuhn length also should be expected. Figure 7A shows the structure of the *yusC* riboswitch with the terminator stem (structure #0) and closely resembling the bound structure, and Figure 7(b) represents a neighboring suboptimal structure (#4) that approximately corresponds to the proposed unbound metabolite structure and has the antiterminator stem. In these calculations, the Kuhn length of ξ=10 nt, minimum stem length 3 and the *-Mg* option was necessary to make the precise yusC structure with all stems correctly matched. A further adjustment was the option *-cc_dist 5* that extends the internal-loops of contiguous stems to a maximum of 5×5 (where the default is 4×4 for ξ=10 nt). Figure 7C is the prediction (#0) when the Kuhn length is changed from ξ=10 to ξ=8 nt. The key features in the structure strongly resemble the unbound structure #5 found in Figure 6.

The precise extent that the Kuhn length changes in the presence of a metabolite is not known and similar suboptimal structures can also be found using ξ=9 nt. Changing the Kuhn length has some of the same effects as calculating the suboptimal structures. The merit in checking suboptimal structures with the same Kuhn length is that all other

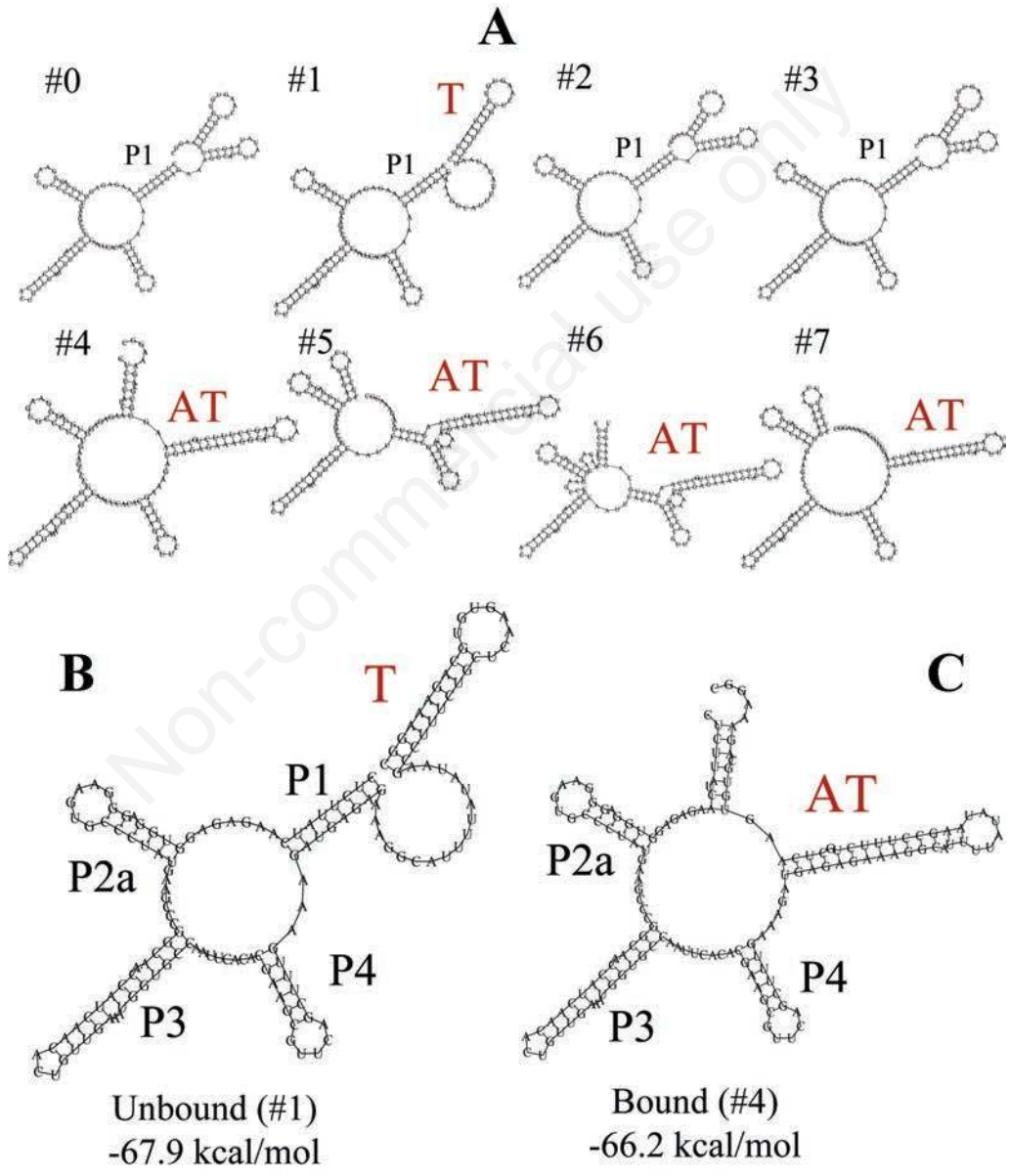

Figure 6. The results of the yusC SAM riboswitch. A) The first 8 structures predicted by vs_subopt for the yusCBA operon, where structure #1 represents the bound form and structures #5 through #8 represent the unbound form of the riboswitch. B) The bound form of the SAM riboswitch. C) The energetically nearby structure proposed for the unbound form of the SAM riboswitch. P1=paired stem 1, etc., T=terminator and AT=anti-terminator.

parameters are left unchanged. Therefore, it minimizes the systematic issues of estimating the differences in energy between structures using different Kuhn lengths. Nevertheless, comparing Figure 7A and C, it turns out that the free energy favors Figure 7C by roughly 10 kcal/mol. In the absence of the metabolite, the unbound structure should probably be favored.

Figure 7 also demonstrates another aspect of the CLE model; namely, the robustness of the predictions. The structures in Figure 7 were fitted with additional 5' sequence from the yusCBA operon, yet very similar cloverleaf structures are found for the bound state. This shows yet again that *vsfold* is valuable for doing RNA-biology research because the researcher does not have to manicure sequences to help the model find the domains. Rather, the CLE model is robust enough to find those domains on its own power. For the small expense of recalculating *the same sequence* using a different Kuhn length, the researcher learns far more information on the domain character and stiffness of a particular RNA sequence under study. A structure prediction model should aid researchers in understanding the physics of structure.

We have shown for representative structures of RNA that a funnel-shaped energy landscape with the optimal structure being the observed structure and the suboptimal structures largely pointing in the direction of the optimal structure was largely achievable using the CLE model. We also observed in several important riboswitches that the free-energy difference between the bound metabolite state and the unbound state of

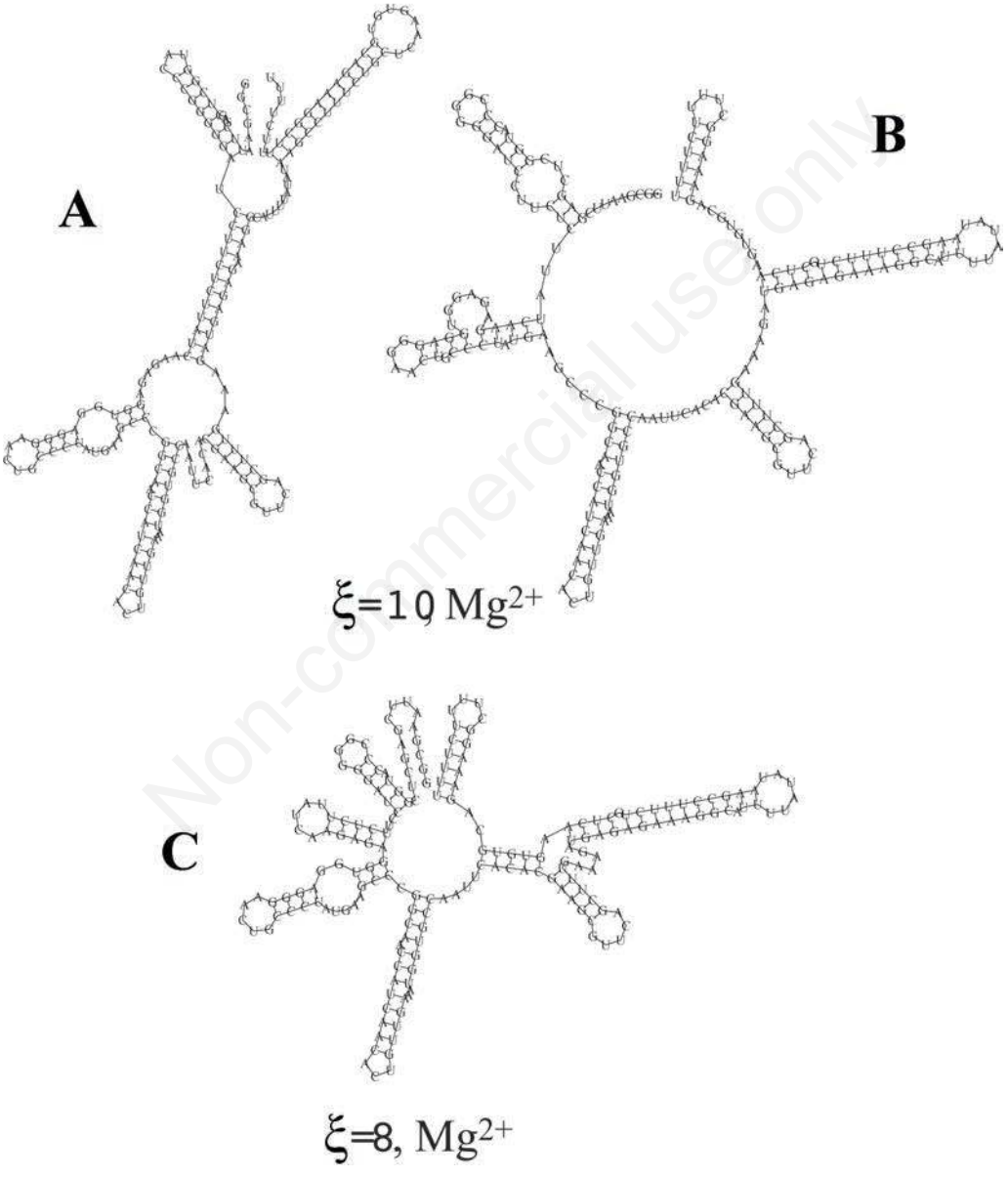

Figure 7. Results of the yusC SAM riboswitch for different Kuhn lengths. A) The switch is in the bound form (using ξ=10 nt and including Magnesium ion interactions), structure #0. B) The switch is in the unbound form of the structure #4. C) Calculation of the yusC SAM riboswitch using ξ=8 nt and including Mg^{2+} interactions showing the unbound form of the structure (#0). The difference in energy between A) and C) is approximately 10 kcal/mol in favor of C), the unbound structure. All calculations use the -Mg option and -cc_dist 5 to expand the contiguous stem length.

the structure were well within the range expected for binding a metabolite and that these structures were partitioned between the two states and in close proximity, consistent with a two state mechanism where the neighboring state should be in close proximity in energy. The two states of the riboswitch are predicted to be close enough to one another in free energy that it is relatively easy for the metabolite to bind and stabilize the alternative structure. The CLE method is not only predicting structures, it is predicting reasonable energy differences.

A significant improvement here is that the structures are often found in a clear and straightforward order. Key native state structural elements form early and these structures often persist well before the expected native state structure is reached. Different Kuhn lengths generate different results. However, this gives the user some sense of the particular stiffness of the RNA under study. Such concepts are not even thought about by other approaches currently. In the case studies presented here, the optimal structures are certainly the observed structures in NMR and X-ray analysis and the structures ordered in terms of energy with the native state structures forming early and persisting largely throughout the energy landscape. This contrasts with the arrangement of suboptimal structure using conventional techniques, where the correct structure is often just one of many suboptimal structures bearing no particularly close free energy relationship to the optimal structure.[6] This is often justified on the basis that the dominant structural features occur at a much higher frequency. However, it is generally thought that crystals of these structures are in their optimal structure and therefore at the minimum FE. A crystal composed of suboptimal structures would contain a high degree of disorder, especially if the difference between the observed structure and the structure that corresponds to the minimum FE exceeds 20 kcal/mol at typical temperatures.[7,8] Granted, there might be a lot of states that are close together for the observed structure, but it would make more sense if 3.5 billion years of natural selection had already tuned the FE to have the maximum number of states all clustered around the minimum FE. Moreover, even if the commonly observed structure were so far from equilibrium, we should still observe the true optimal structures under at least some experimental conditions.

The distribution and type of species predicted by the CLE model appears to largely overcome these issues, even with the current severe limitations of using one single Kuhn length per calculation attempt. With the current model, the Kuhn length must be decided by the user. Based on the concepts outlined in Part II, stem lengths and Kuhn lengths are strongly correlated. Hence, the Kuhn length can often be discerned from the nature of the RNA itself, *e.g.*, tRNA has short stems and should have a small Kuhn length (ξ=5 nt) whereas the SAM riboswitch (for example) tends to have long stems and therefore a longer Kuhn length (ξ=9,10 nt). We therefore think that further development of the model, particularly in the area of a variable Kuhn length (Part II and Dawson *et al.*[1]), is likely to yield even more accurate and instructive insights, especially since the Kuhn length could also change during folding. In short, what we have been able to show is an approximation, and, given more flexibility in the parameterization (particularly ξ), better insights are likely to follow. At this point, the experimental data is far too unclear to address these matters further.

Conclusions

In this work, we have shown that the CLE model easily generates a folding landscape that is essentially funnel shaped for small RNA structures. For a good choice of Kuhn length, the observed structure turns out to be that predicted by the CLE model to lie at the bottom of the well. Substructures of the native state seem to be grouped consecutively in free energy in an understandable fashion and components of the native state often persist throughout the majority of the suboptimal structures and, therefore, a large part of the folding process.

We also observed that the free energy of the two different structures in a riboswitch can (at least) be relatively close in energy, where one state is the bound state and the other unbound. In general, riboswitches also typically tend to have structural distributions consistent with a two state system. The activation barrier appears to be accessible on a time scales of ms or less, enough time for a metabolite to diffuse to the location and bind. In the case of these two state systems, the folding would fall into either state and then oscillate between them. For such systems, it would have to have two wells on the funnel. Nevertheless, the structures fall into one or the other well and then hop through an activation barrier to the other well.

The observed tendency of the folding landscape to be funnel shaped is consistent with natural selection where, given sufficient time, the thermodynamics of the most essential biomolecules will surely be tuned to optimize the ensemble of suboptimal states (local minima) for folding efficiency, or use the predictable folding process to regulate this rate by mechanisms like trapping, or tune it to differentiate between a finite set of specific states for switching or recognition.

The CLE model is able to add unique insights into how we overcome Levinthal's paradox. By limiting domain size, the maximum theoretically possible search space for RNA structure is finite for a given base composition. Natural selection may go even further in that the average base composition can, to some extent, bias the size of this cutoff. This also means that, computationally, there will be some cutoff where the computation can be done in linear time. Though this cutoff may still be somewhat prohibitive at this time, a cutoff would allow the application of parallel calculations at least over maximum-domain-size length scales.

The CLE model is robust, providing the researcher with more information than other available approaches about the stiffness of the RNA and providing a far deeper understanding about the stability of functional domains of RNA structure.

Software

A binary version of *vs_subopt* is available upon request to the corresponding author and upon written consent to the license agreement. Available formats are 64 bit Linux (x86_64), or 32 bit Linux, Window XP, and Mac OSX 10.5 and above.

References

1. Dawson W, Kawai G. Modeling the chain entropy of biopolymers: unifying two different random walk models under one framework. J Comput Sci Syst Biol 2009;2:1-23.
2. Ivankov DN, Garbuzynskiy SO, Alm E, et al. Contact order revisited: influence of protein size on the folding rate. Protein Sci 2003;12:2057-62.
3. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985-94.
4. Dawson W, Fujiwara K, Kawai G, Futamura Y, Yamamoto K. A method for finding optimal RNA secondary structures using a new entropy model (vsfold). Nucleotides Nucleosides Nucleic Acids 2006;25:171-89.
5. Dawson W, Suzuki K, Yamamoto K. A physical origin for functional domain structure in nucleic acids as evidenced by cross linking entropy: part I. J Theor Biol 2001;213:359-86.
6. Dawson W, Suzuki K, Yamamoto K. A physical origin for functional domain structure in nucleic acids as evidenced by cross linking entropy: part II. J Theor Biol 2001;213:387-412.
7. Dawson W, Fujiwara K, Kawai G. Prediction of RNA pseudoknots

using heuristic modeling with mapping and sequential folding. PLoS One 2007;2:e905.
8. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167-95.
9. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84:7524-8.
10. Wales DJ, ed. Energy Landscapes: with applications to clusters, biomolecules and glasses. Cambridge: Cambridge University Press; 2003.
11. Dill KA, Stigter D. Modeling protein stability as heteropolymer collapse. Adv Protein Chem 1995;46:59-104.
12. Onuchic JN, Nymeyer H, Garcia AE, et al. The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. Adv Protein Chem 2000;53:87-152.
13. Go N. The consistency principle revisited. In: Kuwajima K, Arai M, eds. Amsterdam: Elsevier Science; 1999.
14. Rouget JB, Schroer MA, Jeworrek C, et al. Unique features of the folding landscape of a repeat protein revealed by pressure perturbation. Biophys J 2010;98:2712-21.
15. Sosnick TR, Barrick D The folding of single domain proteins: have we reached a consensus? Curr Opin Struct Biol 2011;21:12-24.
16. Cheng RR, Uzawa T, Plaxco KW, Makarov DE. Universality in the timescales of internal loop formation in unfolded proteins and single-stranded oligonucleotides. Biophys J 2010;99:3959-68.
17. Dawson W, Kawai G, Yamamoto K. Modeling the long range entropy of biopolymers: a focus on protein structure prediction and folding. Rec Res Dev Exp Theor Biol 2005;1:57-92.
18. Alm E, Baker D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc Natl Acad Sci USA 1999;96:11305-10.
19. Xayaphoummine A, Viasnoff V, Harlepp S, Isambert H. Encoding folding paths of RNA switches. Nucleic Acids Res 2007;35:614-22.
20. Dotu I, Lorenz WA, Van Hentenryck P, Clote P. Computing folding pathways between RNA secondary structures. Nucleic Acids Res 2010;38:1711-22.
21. Geis M, Flamm C, Wolfinger MT, et al. Folding kinetics of large RNAs. J Mol Biol 2008;379:160-73.
22. Hofacker IL, Flamm C, Heine C, et al. BarMap: RNA folding on dynamic energy landscapes. RNA 2010;16:1308-16.
23. Chen SJ. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Ann Rev Biophys 2008;37:197-214.
24. Woodson SA. Taming free energy landscapes with RNA chaperones. RNA Biol 2010;7:677-86.
25. Frauenfelder H, Leeson DT. The energy landscape in non-biological and biological molecules. Nat Struct Biol 1998;5:757-9.
26. Pan T, Sosnick TR. Intermediates and kinetic traps in the folding of a large ribozyme revealed by circular dichroism and UV absorbance spectroscopies and catalytic activity. Nat Struct Biol 1997;4:931-8.
27 Pan J, Woodson SA. The effect of long-range loop-loop interactions on folding of the tetrahymena self-splicing RNA. J Mol Biol 1999;294:955-65.
28. Fang XW, Thiyagarajan P, Sosnick TR, Pan T. The rate-limiting step in the folding of a large ribozyme without kinetic traps. Proc Natl Acad Sci USA 2002;99:8518-23.
29. Sosnick TR, Pan T. Reduced contact order and RNA folding rates. J Mol Biol 2004;342:1359-65.
30. Wan Y, Suh H, Russell R, Herschlag D Multiple unfolding events during native folding of the Tetrahymena group I ribozyme. J Mol Biol 2010;400:1067-77.
31. Furtig B, Wenter P, Pitsch S, Schwalbe H. Probing mechanism and transition state of RNA refolding. ACS Chem Biol 2010;5:753-65.
32. Turner DH, Sugimoto N, Freier SM. RNA Structure prediction. Ann Rev Biophys Biophys Chem 1988;17:167-92.
33. Porschke D, Eggers F. Thermodynamics and kinetics of base-stacking interactions. Eur J Biochem 1972;26:490-8.
34. Levine IN. Physical chemistry. Singapore: Mc Graw-Hill; 2003. p 966.
35. Collin D, Ritort F, Jarzynski C, et al. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. Nature 2005;437:231-4.
36. Li PT, Tinoco I, Jr. Mechanical unfolding of two DIS RNA kissing complexes from HIV-1. J Mol Biol 2009;386:1343-56.
37. Liphardt J, Onoa B, Smith SB, et al. Reversible unfolding of single RNA molecules by mechanical force. Science 2001;292:733-7.
38. Hinz HJ, Filimonov VV, Privalov PL. Calorimetric studies on melting of tRNA Phe (yeast). Eur J Biochem 1977;72:79-86.
39. Privalov PL, Filimonov VV. Thermodynamic analysis of transfer RNA unfolding. J Mol Biol 1978;122:447-64.
40. Laing LG, Gluick TC, Draper DE. Stabilization of RNA structure by Mg ions. Specific and non-specific effects. J Mol Biol 1994;237:577-87.
41. Qiu H, Kaluarachchi K, Du Z, et al. Thermodynamics of folding of the RNA pseudoknot of the T4 gene 32 autoregulatory messenger RNA. Biochemistry 1996;35:4176-86.
42. Gluick TC, Draper DE. Thermodynamics of folding a pseudoknotted mRNA fragment. J Mol Biol 1994;241:246-62.
43. Laing LG, Draper DE. Thermodynamics of RNA folding in a conserved ribosomal RNA domain. J Mol Biol 1994;237:560-76.
44. Coutts SM, Gangloff J, Dirheimer G. Conformational transitions in tRNA Asp (brewer's yeast). Thermodynamic, kinetic, and enzymatic measurements on oligonucleotide fragments and the intact molecule. Biochemistry 1974;13:3938-48.
45. Morgan S, Higgs PG. Barrier heights between ground states in a model of RNA secondary structure. J Phys A: Math Gen 1998;31:3153-70.
46. Fersht A. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. New York: W.H. Freeman; 1997. p 631.
47. Vander Meulen KA, Butcher SE Characterization of the kinetic and thermodynamic landscape of RNA folding using a novel application of isothermal titration calorimetry. Nucleic Acids Res 2012;40:2140-51.
48. Bartley LE, Zhuang Ξ, Das R, et al. Exploration of the transition state for tertiary structure formation between an RNA helix and a large structured RNA. J Mol Biol 2003;328:1011-26.
49. Silverman SK, Cech TR. An early transition state for folding of the P4-P6 RNA domain. RNA 2001;7:161-6.
50. Deng ML, Zhu WQ. Stochastic dynamics and denaturation of thermalized DNA. Phys Rev E Stat Nonlin Soft Matter Phys 2008;77:021918.
51. Hagen SJ, Hofrichter J, Szabo A, Eaton WA. Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding. Proc Natl Acad Sci USA 1996;93:11615-7.
52. Fernandez A, Cendra H. In vitro RNA folding: the principle of sequential minimization of entropy loss at work. Biophys Chem 1996;58:335-9.
53. Grosberg AY, Khokhlov AR. Statistical physics of macromolecules. New York: AIP Press; 1994.
54. Chen G, Chang KY, Chou MY, et al. Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of -1 ribosomal frameshifting. Proc Natl Acad Sci USA 2009;106:12706-11.
55. Hagerman PJ. Flexibility of RNA. Annu Rev Biophys Biomol Struct 1997;26:139-56.
56. Holbrook SR, Sussman JL, Warrant RW, Kim SH. Crystal structure of yeast phenylalanine transfer RNA. II. Structural features and functional implications. J Mol Biol 1978;123:631-60.
57. Sussman JL, Holbrook SR, Warrant RW, et al. Crystal structure of yeast phenylalanine transfer RNA. I. Crystallographic refinement. J Mol Biol 1978;123:607-30.
58. Hingerty B, Brown RS, Jack A. Further refinement of the structure of yeast tRNAPhe. J Mol Biol 1978;124:523-34.

59. Rozhdestvensky TS, Kopylov AM, Brosius J, Huttenhofer A. Neuronal BC1 RNA structure: evolutionary conversion of a tRNA(Ala) domain into an extended stem-loop structure. RNA 2001;7:722-30.
60. Frazer-Abel AA, Hagerman PJ. Determination of the Angle between the Acceptor and Anticodon stems of a truncated mitochondrial tRNA. J Mol Biol 1999;285:581-93.
61. Friederich MW, Gast FU, Vacano E, Hagerman PJ. Determination of the angle between the anticodon and aminoacyl acceptor stems of yeast phenylalanyl tRNA in solution. Proc Natl Acad Sci USA 1995;92:4803-7.
62. Friederich MW, Hagerman PJ. The angle between the anticodon and aminoacyl acceptor stems of yeast tRNA(Phe) is strongly modulated by magnesium ions. Biochemistry 1997;36:6090-9.
63. Friederich MW, Vacano E, Hagerman PJ. Global flexibility of tertiary structure in RNA: yeast tRNAPhe as a model system. Proc Natl Acad Sci USA 1998;95:3572-7.
64 Serebrov V, Clarke RJ, Gross HJ, Kisselev L. Mg2+-induced tRNA folding. Biochemistry 2001;40:6688-98.
65. Levinthal C. How to fold graciously: mossbauer spectroscopy in biologcal systems. In: DeBrunner JTP, Munck E, eds. Monticello: University of Illinois Press; 1969. pp 22-24.
66. Mandal M, Breaker RR. Adenine riboswitches and gene activation by disruption of a transcription terminator. Nat Struct Mol Biol 2004;11:29-35.
67. Serganov A, Yuan YR, Pikovskaya O, et al. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. Chem Biol 2004;11:1729-41.
68. Wakeman CA, Winkler WC, Dann III CE. Structural features of metabolite-sensing riboswitches. Trends Biochem Sci 2007;32:415-24.
69. Selivanov VA, Krause S, Roca J, Cascante M. Modeling of spatial metabolite distributions in the cardiac sarcomere. Biophys J 2007;92:3492-500.
70. Barros LF, Martinez C. An enquiry into metabolite domains. Biophys J 2007;92:3878-84.
71. Koopman WJ, Distelmaier F, Hink MA, et al. Inherited complex I deficiency is associated with faster protein diffusion in the matrix of moving mitochondria. Am J Physiol Cell Physiol 2008;294:C1124-32.
72. Batey RT, Gilbert SD, Montange RK. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature 2004;432:411-5.
73. Mandal M, Boese B, Barrick JE, et al. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. Cell 2003;113:577-86.
74. Garst AD, Batey RT. A switch in time: detailing the life of a riboswitch. Biochim Biophys Acta 2009;1789:584591.
75. Montange RK, Batey RT. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. Nature 2006;441:1172-5.
76. Winkler WC, Nahvi A, Sudarsan N, et al. An mRNA structure that controls gene expression by binding S-adenosylmethionine. Nat Struct Biol 2003;10:701-7.
77. Nakamura K. Information and discussions on the yusC SAM riboswitch. Tesis Dissertation. Tsukuba University.
78. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 2004;5:105.