A new entropy model for RNA: part II. Persistence-related entropic contributions to RNA secondary structure free energy calculations

Wayne Dawson,[1] Kenji Yamamoto,[2] Kentaro Shimizu,[1] Gota Kawai[3]

[1]Bioinformation Engineering Laboratory, Department of Biotechnology, Graduate School of Agriculture and Life Sciences, The University of Tokyo; [2]Research Institute, National Center for Global Health and Medicine, Tokyo; [3]Chiba Institute of Technology, Chiba, Japan

Abstract

In previous work, we have shown that the entropy of a folded RNA molecule can be divided into local and global contributions using the cross-linking entropy (CLE) model, where, in the case of RNA, the cross-links are the base-pair stacking interactions. The local contribution to the CLE is revealed in the Kuhn length (a measure of the stiffness of the RNA). The Kuhn length acts as a scaling parameter. When the size of the system is rescaled, the relationship between local and global free energy must be renormalized to reflect this rescaling. In this renormalization process, the Kuhn length increases, the local entropy also increases due to freezing out of the local conformational degrees of freedom. At the same time, as the number of degrees of freedom decrease, there is a significant reduction in the global entropy. Here we present a method, based on the concepts of renormalization theory, to quantitatively estimate the size of the contribution from the local entropy as a function of the Kuhn length. The local entropy correction is used to predict the current empirically derived constant in the Jacobson-Stockmayer equation. The variation in the Kuhn length is shown to be largely influenced by the length of the double-stranded RNA stems formed in the secondary structure of folded RNA. This result is used to test the resulting entropy under a variable Kuhn length in stem-loop structures. Comparisons between a variable Kuhn length and a static Kuhn length on a short stem-loop of RNA are also examined. The model is quite general and is also directly applicable to protein structure and folding problems.

Correspondence: Wayne Dawson, Department of Computational Biology, Faculty of Frontier Sciences, The University of Tokyo, AIST Tokyo Waterfront BIO-IT Research Building, 2-4-7 Aomi, Moto-Ku, Tokyo, 135-0064, 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, Kuhn length, thermodynamics, bioinformatics.

Contributions: WD, wrote the manuscript and did the primary research; KY, advice, guidance and support; KS, advice, guidance and support and assistance in preparing the manuscript; GK, advice, guidance and support and assistance in preparing the manuscript.

Acknowledgments: this work was supported in part from grants from Japan International Science and Technology Exchange Center (JISTEC) and Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Global Center for Excellence (GCOE) program. We thank the staff at AIST and Prof Kiyoshi Asai at Univ of Tokyo (UofT) and Computational Biology Research Center (CBRC) for providing assistance and facilities for the final manuscript preparation. Special thanks to Dr. Shingo Nakamura (Catalent Pharma Solutions) for his many science related questions. We also thank Profs Shugo Nakamura, Tohru Terada, and Kazuya Sumikoshi (UofT), YZ and CS for their ongoing encouragement.

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

Received for publication: 30 May 2011.
Revision received: 18 January 2012.
Accepted for publication: 2 February 2012.

Journal of Nucleic Acids Investigation 2013; 4:e2
doi:10.4081/jnai.2013.e2

Introduction

In a recent work,[1] we explained how to unify the concept of conformational entropy between lattice models, the Gaussian polymer chain (GPC) model, the worm-like chain (WLC) model and the contact order (CO) model. Entropy in a polymer is a measure of the disorder of the polymer chain and a function of the number of conformational degrees of freedom. The entropy model was divided between local and global contributions to the free energy (FE), where the local FE involves the direct interactions between neighboring monomers (mers) along the polymer chain and the global FE involves long range correlation between distance mers. In the ideal polymer, where the mers consist of non-interacting beads on a chain, this FE is entropic in character. A significant parameter associated with this entropy is the Kuhn length (ξ) or, in other parlance the persistence length.[2,3] [The persistence length is derived from the WLC model and is about 1/2 the size of the Kuhn length.[3] Studies of double-stranded DNA and RNA often are expressed in terms of the persistence length]. The local entropy becomes more negative with increasing ξ due to the local freezing out of the degrees of freedom of the neighboring mers. The global entropy becomes more negative whenever the ideal polymer is distorted from its equilibrium position, either by stretching or compressing (folding).

In previous work,[1,4-6] we have introduced and developed the cross linking entropy (CLE) model in which we focused on the concept of Kuhn length in the context of the *global* entropy. We argued that the Kuhn length is usually longer than the monomer-to-monomer (mer-to-mer) separation distance, and, for single-stranded RNA (ssRNA), typically varying between 3 to 10 nucleotides (nt) for RNA. Therefore the modeling must be made coarse-grained by definition and coarse-grained at a scale larger than the natural mer-to-mer distance. However, we have not examined how the Kuhn length influences the local entropy in the coarse-grained approach, other than to write down some of the relevant equations.

Pedagogically, it has proved highly instructive to train students to conceptualize problems at the monomer (mer) level (or a smaller atomic level). There is little information on how we must fix our calculation strategy to account for the fact that the monomers (amino acids, nucleic acids, etc.) interact locally as a group in Kuhn-length sized units. Likewise, though some general concepts of renormalization exist,[7,8] neither is there much information on how to scale these local corrections quantitatively. As a result, calculations are typically done in a mer-by-mer or atom-by-atom fashion in biophysical problems.

The Kuhn length introduces a straightening effect that requires constraints to account for it. These constraints reduce the number of degrees of freedom. In effect, the mer-by-mer unit approach must be replaced by a group of monomers: more of a group-by-group approach. If a polymer of N monomers has complete freedom of motion at the position of each monomer, then the polymer has N degrees of freedom. Most polymers have a Kuhn length greater than the mer-to-mer separation distance. Roughly speaking this means that, for a given polymer with Kuhn length ξ and N monomers, the polymer has approximately N/ξ degrees of freedom. We must account for this change in the number of degrees of freedom by incorporating constraints on a local scale and correcting the overall entropy of these frozen out degrees of freedom.

The Kuhn length (and persistence length) has been a subject of study in double-stranded RNA (dsRNA) and double-stranded DNA (dsDNA). By far, the majority of the studies that recognize the importance of persistence length are directed to dsDNA. Experiments note that the persistence length of dsDNA is typically at least 20 times longer than the persistence length of ssRNA or single-stranded DNA (ssDNA).[9-15] Most of these studies have focused on dsDNA melting where the base pair (bp) separation forms symmetric internal loops (I-loops), also known as *bubbles* because of the apparent bulging in the loop region. Experimental work has also focused on the effects of ionic strength.[16-20] With the development of atomic force spectroscopy (AFM) and related deposition experiments, the persistence length has been measured based on the curvature of the DNA on a 2D planar surface under various conditions.[21-23] Numerous studies of dsDNA using optical tweezers in force extension experiments report the persistence length.[12-15,24-26] Theoretical studies generally recognized that the ssDNA in the I-loop has a very different persistence length than in the dsDNA region. However, studies have been largely directed to finding the melting temperature (T_m) and the average bubble size of long dsDNA sequences,[27-32] or evaluating the structure of dsDNA near the melting transition.[30-32] There are only a limited number of experiments directed to measuring the length dependence of the Kuhn length itself as a function of sequence.[18] There have also been a few studies of stacking of ssDNA in which there are some indications that ξ may depend on the sequence context.[33-36]

For folded ssRNA, Felsenfeld's group attempted to measure the Kuhn length of various unfolded ssRNA sequences using sedimentation techniques in the late 60s.[37-39] Later, in the mid 80s, Hagerman's group used transient electric birefringence (TEB) to measure the persistence length of folded ssRNA and dsRNA for specific RNA structures.[40-44] More recently, AFM was used to evaluate the persistence length of dsRNA.[45] Optical tweezers experiments have been directed to folded ssRNA measured under stretching conditions where Tinoco's group has studied a considerable number of RNA molecules.[46-54]

Changes in persistence length due to RNA folding have been done based on the evaluating the radius of gyration.[55,56] However, the majority of these studies were done using TEB by Hagerman's group,[40-43] where it was also shown that the RNA stiffens due to the formation of base pairs (stacking) rather than due to electrostatic effects.[57]

Although there have been this handful of dedicated experimental and theoretical studies of DNA and RNA, the Kuhn length (or persistence length) has not been applied to the prediction and folding of ssRNA structures except in our work.

In part, this is because experimental techniques typically only extract average values for the Kuhn length (or persistence length).[22,23,55,56] Yet, in terms of the appearance of known RNA structures, it is largely understandable that scaffolding typically is very stiff and therefore involves long stems whereas moving segments of an RNA chain should be more flexible and therefore involve short stems or free strand regions. Recognition regions could be either stiff or flexible, depending on the binding context and the type of cognate structure involved. Usually, this is discussed in x-ray and NMR structure under the category of flexibility.[58] Since flexibility is, in essence, the inverse of the Kuhn length, the concept of a variable Kuhn length already has support from experimental data. Likewise, recent studies into DNA have begun to ask questions about the nature of this persistence length.[59]

Neglecting these constraints can be both significant and misleading in structure prediction. For example, in our first version of CLE model,[60,61] we only considered the global contribution and filtered other RNA secondary structure predictions obtained independently. When the *global* CLE model was set up to work on its own without constraints on the flexibility, a new problem emerged where the structures tended to crinkle up. In Figure 1A, the correct prediction for a sequence of C-U-G repeats is shown and is predicted by all current methods including our own. Figure 1B shows what happens when we only considered the global entropy and we fail to constrain the number of degrees of freedom. The structure crinkles up because there are too many degrees of freedom and the flexibility of the structure is overestimated. When we accounted for these straightening effects caused by the Kuhn length,

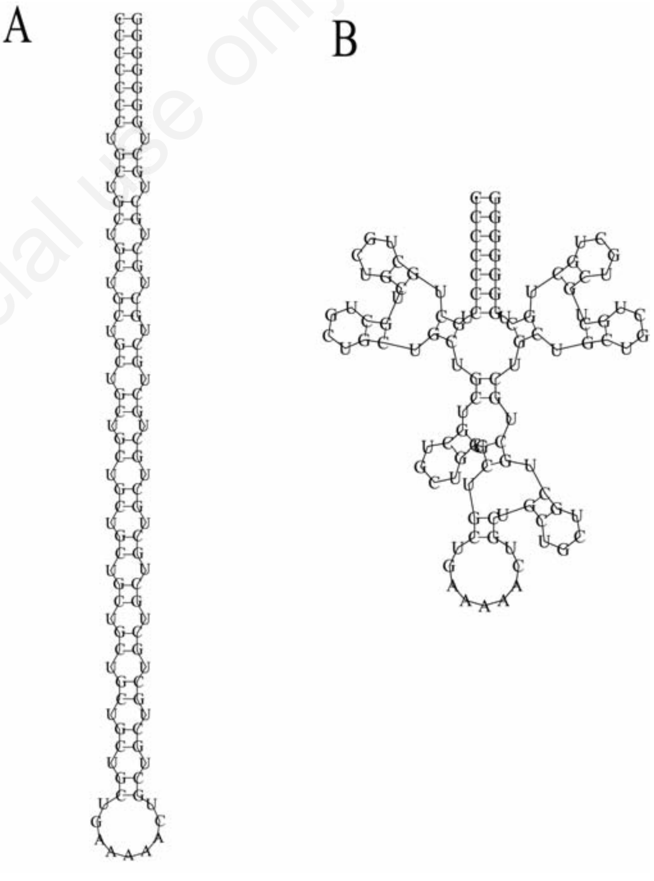

Figure 1. A) The calculated secondary structure of a CUG repeat sequence using a standard genre of RNA structure prediction programs. This structure prediction is essentially correct. It is also predicted correctly using the current versions of vsfold4 and vsfold5 (with option *-cug 14*). B) The calculated secondary structure of a CUG repeat sequence when only the global entropy is used without correcting with constraints (straightening effects caused by the Kuhn length) and accounting for the reduction in the number of degrees of freedom. This structure prediction is incorrect and changing the Kuhn length to 10 nt or more does not change this result even though this should mean that the structure will tend to straighten over at least the distance of 10 nt in the double helical regions.

applying constraints and excluding the excess degrees of freedom, we arrived at the correct structure in Figure 1A.

Although traditional methods for predicting RNA structure have neglect the Kuhn length and are able to obtain correct structures without ever considering the Kuhn length, a Kuhn length-blind approach leaves the users in the dark about the flexibility of the structure. We can neither deduce the mechanical action of various parts of the RNA nor can we discern the true thermodynamics of transition states, even if the correct structure is fortuitously obtained as a byproduct of a Kuhn length blind approach. Therefore, neglecting the Kuhn length limits our understanding of RNA structure.

In previous studies, we have emphasized the effects of the global entropy.[1,4-6,60,61] Whereas some of the conclusions on the local entropy have been published before, the derivation and the full perspective have not been explained. In this study, we aim to introduce the full methodology that is used to compute the local entropy with the Kuhn length and the effects of changing the number of degrees of freedom when the Kuhn length is changed. It is important to see this local entropy as a kind of constraint that changes the overall stiffness of the RNA. After developing the basic methodology for calculating the local entropy, we show the relationship between the local entropy and the constant term found in the Jacobson-Stockmayer (JS) equation (A_{JS}) from first principles (where the original derivation is explained in Appendix A of Part I of this series). From there, we consider how a variable Kuhn length influences the FE calculations. In the final Section, we discuss these findings in the context of computational approaches. The concepts discussed here are directed to folded ssRNA; however, the concepts are applicable to all polymers with appropriate modifications.

Summary of the global cross-linking entropy model

The global entropy in the CLE model is explained in considerable detail in several papers,[1,4-6,60,61] some of which are public access. Therefore, we simply write these equations with little further explanation.

Let *N* be the number of mers and *b* the distance between consecutive mers on the RNA polymer chain. Let *i* and *j* represent the indices of a pair of mers subject to $1 \leq i < j \leq N$. Let the Kuhn length (ξ) be defined in units of mers such that the distance $b' = \xi b$. Hence, when $\xi = 1$, $b' = b$ and the separation between monomers is of unit length in such a case.

In RNA, ξ is always longer than the monomer-to-monomer (mer-to-mer) separation distance. This freezing out of the degrees of freedom of the individual monomers results in the formation of *effective mers*. Therefore, we must apply renormalization theory to correct the FE to reflect these changes.[7,8] The essential concept behind renormalization theory as used in the context of folded ssRNA structure in this work is explained in the Appendix.

Suppose that we can somehow turn off the complex interactions between the mers in an RNA molecule (even better than a denaturing solvent). This would represent the conditions of an ideal polymer. In such conditions, the root-mean-square separation distance between mers *i* and *j* (ij-rmsd) is

$$\langle r^2 \rangle_{ij}^{1/2} = \kappa \left| j - i + 1 \right|^{\nu} b \text{, or } , \langle r^2 \rangle_{ij}^{1/2} = \xi^{1-\nu} N_{ij}^{\nu} b \qquad (1)$$

where N_{ij} is the number of residues separating *i* and *j* ($N_{ij} = j - i + 1$), $\kappa = \xi^{1-\nu}$ and ν is a parameter expressing the excluded volume. From the central limit theorem,[62] we find the variance $\langle r^2 \rangle_{ij} = \xi N_{ij} b^2$: *i.e.*, the ij-rmsd for $\nu = 1/2$. For a Gaussian polymer chain (GPC), $\nu \equiv 1/2$, $\kappa = \xi^{1/2}$ and $\langle r^2 \rangle_{ij} = \xi N_{ij} b^2$. The parameter ν can range between $1/3 < \nu < 3/5$, where $\nu < 1/3$ expresses a collapsed polymer and $\nu = 3/5$ expresses a swelled polymer.[2,63]

When $i=1$ and $j=N$, $\langle r^2 \rangle_{1N} = \langle r^2 \rangle$ expresses the end-to-end mean-square distance separating the ends of the RNA and is a measurable parameter based on the radius of gyration.[2,3,63] If one were to cut this sequence at *i* and *j* such that the new sequence is length $N_{ij} = j - i + 1$, then the ij-rmsd would be the same as the end-to-end rmsd.

Now, let r_{ij} represent some experimentally observed distance between mers *i* and *j*, not necessarily equivalent to $\langle r^2 \rangle_{ij}^{1/2}$. For $\xi > 1$, the global contribution to the entropy for the interaction between mers *i* and *j* is

$$S(r_{ij}) = \frac{k_{\mathrm{B}}}{\xi} \left\{ \ln(A_{\delta\gamma} C_{ij\xi}^{\gamma\delta}) + \delta\gamma \ln\left(\frac{r_{ij}}{b}\right) - \vartheta_{ij\xi} \left(\frac{r_{ij}}{b}\right)^{\delta} \right\} \qquad (2)$$

where ξ scales the entropy contribution due to stem formation by a corresponding reduction in degrees of freedom because the length scale is based on *effective mers* rather than mers (Appendix), δ is a finite positive constant and γ (>0) is a weight that corrects for the fact that real polymer chains cannot have more than one mer occupying the same space at the same time, where the common value used in RNA calculations is $\gamma = 1.75$ compared to Gaussian statistics ($\gamma \equiv 1$).[1] This is known as *self-avoidance* and differs from the excluded volume associated with the pamameter ν. Of the other parameters, $A_{\delta\gamma}$ is the spherically symmetric contribution to the volume term,

$$A_{\delta\gamma} = \frac{\delta \pi^{\gamma + 1/\delta}}{\Gamma(\gamma + 1/\delta)} \qquad (3)$$

where $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$ is the Gamma-function; $\vartheta_{ij\xi}$ weights the variance ($\langle r^2 \rangle_{ij}$) in Equation (2),

$$\vartheta_{ij\xi} = \left(\frac{\Gamma(\gamma + 3/\delta)}{\Gamma(\gamma + 1/\delta)} \frac{b^2}{\langle r^2 \rangle_{ij}} \right)^{\delta/2} = \zeta(\gamma, \delta) \left(\frac{b^2}{\langle r^2 \rangle_{ij}} \right)^{\delta/2} = \frac{\zeta(\gamma, \delta)}{\xi^{\delta(1-\nu)} N_{ij}^{\delta\nu}} \qquad (4)$$

with

$$\zeta(\gamma, \delta) = \left[\Gamma(\gamma + 3/\delta) / \Gamma(\gamma + 1/\delta) \right]^{\delta/2} \qquad (5)$$

and $C_{ij\xi}^{\gamma\delta}$ is a normalization constant

$$C_{ij\xi}^{\gamma\delta} = \frac{\delta \vartheta_{ij\xi}^{\gamma + 1/\delta}}{A_{\delta\gamma} \Gamma(\gamma + 1/\delta)} \qquad (6)$$

with $\langle r^2 \rangle_{ij}$ defined in Equation (1).[63] In the case of the GPC (where $\delta \equiv 2$, $\gamma \equiv 1$ and $\nu \equiv 1/2$): $A_{\delta\gamma} = 4\pi$, $\langle r^2 \rangle_{ij} = \xi N_{ij} b^2$, $\vartheta_{ij\xi} = 3/(2\xi N_{ij})$ and $C_{ij\xi} = [3/(2\pi\xi N_{ij})]^{3/2}$.[1,3]

The global change in entropy is measured by considering two stable states of the system: the denatured structure, where $r_{ij} = \langle r^2 \rangle_{ij}^{1/2} = \kappa N_{ij}^{\nu} b$, and the native state, where $r_{ij} = \lambda b$ for RNA. Since the polymer involves very crude approximations of shape, the value of λb is not equivalent to the chemical bond length between nucleic acids in the double helix, but the distance between effective mers.

By finding the end-to-end separation distance, we have a way to describe the denatured state of the RNA. The entropy-loss due to bp formation as the structure folds from the denatured state to the native state has the general form

$$\Delta S_{bp}(N_{ij}) = S(\lambda b) - S\left(\langle r^2 \rangle_{ij}^{1/2} \right) =$$

$$-\frac{k_{\mathrm{B}}}{\xi} \left\{ \delta\gamma \ln\left(\frac{\langle r^2 \rangle_{ij}^{1/2}}{\lambda b} \right) - \vartheta_{ij\xi} \left[\left(\frac{\langle r^2 \rangle_{ij}^{1/2}}{b} \right)^{\delta} - \lambda^{\delta} \right] \right\}. \qquad (7)$$

Substituting Equations (1) and (4) into Equation (7), we obtain

$$\Delta S_{bp}(N_{ij}) = -\frac{k_B}{\xi}\left\{\nu\delta\gamma\ln\left(\Psi_{\nu\xi}N_{ij}\right) - \zeta(\gamma,\delta)\left(1 - 1/(\Psi_{\nu\xi}N_{ij})^{\delta\nu}\right)\right\} \quad (8)$$

where $\Psi_{\nu\xi}=\xi^{-1}(\xi/\lambda)^{1/\nu}$.

For the GPC, Equation (8) reduces to

$$\Delta S_{bp}(N_{ij}) = -\frac{k_B}{\xi}\left\{\ln\left(\Psi_{1/2\xi}N_{ij}\right) - \frac{3}{2}\left(1 - \frac{1}{\Psi_{1/2\xi}N_{ij}}\right)\right\}. \quad (9)$$

Now, based on Equation (1), the outlined derivation of renormalization in the context of folded ssRNA (Appendix), and using Equation (A6) with the appropriate renormalization weight $1/\xi$, we integrate the force to derive the entropy.[1] The total entropy-loss is the sum of the local correction (renormalization constant) and the global contribution caused by stem formation[1]

$$\Delta S_{cle} = \Delta S_{\xi\gamma\delta} + \sum_{bp(ij)} \Delta S_{bp}(N_{ij}), \quad (10)$$

where $\Delta S_{bp}(N_{ij})$ is the global contribution given in Equation (8) and the derivation of the local entropy term ($\Delta S_{\xi\gamma\delta}$) will be explained in the following Sections. From Equations (7) to (9), as ξ increases, the influence of the *global entropy* associated with base pairing ($\Delta S_{bp}(N_{ij})$) reduces by $1/\xi$ as a result of the renormalization. The summation of $\Delta S_{bp}(N_{ij})$ in Equation (10) has been derived from first principles in numerous independent ways and can be understood as an integration of the base pairing entropy.[1,60,61]

Derivation of the local entropy loss corrections

When the Kuhn length of a polymer changes, the number of effective mers changes. This changes the number of degrees of freedom and therefore the entropy.

The Kuhn length is a parameter that expresses the length scale where weak coupling approximations become valid.[64] Several monomers are grouped together into a single link due to strong mutual coupling. Increasing the Kuhn length introduces local order and strong coupling between monomers over a similar length scale as the Kuhn length. This results in a negative increase in the local entropy that is a function of ξ. As the local structure couples and hardens, the global contribution to the CLE will decrease due to the $1/\xi$ scaling of the number of effective mers [Equations (8) and (10)]: *i.e.*, decreased long range coupling. In accordance with the renormalization approach, the FE is redistributed between the local and global contributions such that the FE remains essentially the same.

We should not view the Kuhn length as though the polymer chain were a mere group of straight rigid rods on hinges; rather the Kuhn length represents substructures in a more or less frozen conformation over a length scale of ξ.[43] It is a statistical quantity that describes the effective scale of the local structure of the polymer. Of the 3N degrees of freedom in the polymer, each link (of length ξ) has ξ degrees of freedom that are frozen out of its total of 3ξ degrees of freedom. In RNA, this is due to lack of free rotation of the residues due to clashes with other residues in the same chain, base stacking and the stabilizing effects of monovalent and divalent cations.[17,42,65-67]

To describe the segmentation in the polymer chain relative to some independent reference, we introduce a fundamental reference Kuhn length segment (ξ_o: measured in units of mers relative to N). Initially, since the reference here is the monomer, the reference Kuhn segment length is $\xi_o\equiv1$ mer. To describe relative changes in the Kuhn length, this reference is important. In this Section, we will assume that ξ_o=1 mer. Cases where $\xi_o\neq1$ will be considered in the next Section.

When the Kuhn length increases from 1 mer to ξ mers, these ξ mers are absorbed into a single effective mer. For a given polymer chain of length N, we associate ξ monomers as a single unit and count $\tilde{N}(=N/\xi)$ independent units (effective mers or links) in the biopolymer. The renormalized polymer chain now contains $\tilde{N}$ links (each containing ξ mers) where at each joint of the polymer chain, the connecting links are free to move in any orientation relative to one another (with self-avoidance accounted for in the γ term).

Now we must consider how to compute this renormalization. In the previous Section, when the global FE is evaluated by way of the force (Appendix, Equation A6), the response is entirely entropic. The local entropy is more complex. Base pairing involves both enthalpy and entropy, $\Delta G_{bp}=\Delta H_{bp}$-$T\Delta S_{bp}$. In the experimentally obtained base pairing parameters, the entropic term is approximately -25 cal/molK (about -22 cal/molK for AU and -27 cal/molK for GC) and this is true for both RNA and DNA.[68-70] This suggests that ΔS_{bp} expresses a local freezing out on a scale smaller than 1 mer (our basic unit of size in these problems). Interestingly, an early calculation by Tinoco's group found a value in this range for the bp interaction.[71] Whether the value of ΔS_{bp} changes due to different dsRNA/dsDNA lengths is not clear, but the magnitude appears to be limited to freezing out at a mer length scale. Therefore, since the length scale of ΔG_{bp} is smaller than ξ and this FE is applied in a additive fashion independent of any other local or global issues, we can treat ΔG_{bp} as independent of the configurations of length scale ξ>1 mer. For this reason, the entropy of base pairing can be neglected here. In the final Section, we will come back to this point in an actual stem-loop calculation.

Let $\Delta s_{\xi_o(\equiv1)\longrightarrow\xi}$ and $\Delta g_{\xi_o(\equiv1)\longrightarrow\xi}$ represent the change in entropy and FE (respectively) of these effective mers (of initial length 1 mer) due to a change 1→ξ. With this change in ξ (currently *treated as though* it were identical throughout the sequence), let $\Delta S_{1\rightarrow\xi}=\tilde{N}\Delta s_{1\rightarrow\xi}$ represent the Kuhn length correction of the entropy for the entire biopolymer chain. Then, $\partial(\Delta S_{1\rightarrow\xi})/\partial\tilde{N}=\Delta s_{1\rightarrow\xi}$. This implies that the effect we are looking at is associated with the chemical potential μ, where the correction to the FE becomes $\Delta G_{1\rightarrow\xi}=\tilde{N}\mu$. The Kuhn length corrections are far more local in character than the global CLE, which ranges over the entire length of the biopolymer (*e.g.*, RNA: 5' and 3'; proteins N-terminus and C-terminus).

The next step is to express the change in the number of effective mers. For a segment of the chain of length $n=\xi$ mers, initially we have ξ_o=1 mer and this increases to ξ. The total number of effective mers has decreased from N to $\tilde{N}=N/\xi$. This means the number of effective mers has changed from a count $n_i=\xi$ [mers] to a count of n_f=1 [effective mer] over the length of the coarse-grained sequence. Since we are interested in the entropic cost of changing the course-grained size of the link, we compute the change in the number of effective mers as

$$\Delta n = n_f - n_i = 1 - \xi = \xi\left(\frac{1}{\xi} - 1\right) \quad (11)$$

For a system of constant end-to-end separation distance r [from Equation (2)] and T (temperature), the Kuhn length correction FE of a single link is

$$\left(\Delta g_{1\rightarrow\xi}\right)_{r,T} = \mu\Delta n = -T\left(\Delta s_{1\rightarrow\xi}\right)_{r,U} \quad (12)$$

where U is the internal energy and, for a system of r and T, $(\partial U/\partial V)_{T,r}$ can be neglected in these problems because we are not evaluating vulcanized rubber (where the volume does change) but a single polymer chain of RNA. Moreover, $(\partial U/\partial r)_{T,V}$=0 for an ideal polymer. Hence, the FE is entirely entropic in character for an ideal polymer (ΔU~0),[63] as also discussed in Part I. Equation (2) contains a scaling factor ($1/\xi$) because the interactions considered involve individual mers. Here, we consider a collective group of mers of length ξ. Hence, we do not need to consider this scaling factor. Therefore, to compute Δ*s*, Equation (2) is used without the 1/ξ renormalization weight because the computation is *of the local interactions on a length scale of ξ*

$$S(\xi,r) = k_{\mathrm{B}}\left\{\ln(A_{\delta\gamma}C_{\xi}^{\gamma\delta}) + \delta\gamma\ln\left(\frac{r}{b}\right) - \vartheta_{\xi}\left(\frac{r}{b}\right)^{\delta}\right\}.$$

where $r=\xi^{\nu}b$ because the region of interest is a sub-sequence (*of length* ξ) that is stretched to a length $\xi^{\nu}b$ in some part of the polymer chain. The entropy of the sub-sequence is

$$\left(\Delta s_{1\to\xi}\right)_{r,U} = \left\{s(\xi)_{r,U} - s(1)_{r,U}\right\}$$
$$= k_{\mathrm{B}}\left\{\ln\left(\frac{C_{\xi}^{\gamma\delta}}{C_{\xi_o}^{\gamma\delta}}\right) - \left(\vartheta_{\xi} - \vartheta_{\xi_o}\right)\left(\frac{(\xi^{\nu}b)}{b}\right)^{\delta}\right\}$$

where $s(\xi)_{r,U}$ is the entropy expressed as a function of the Kuhn length for a single segment of length ξ. Inserting $\vartheta_{\xi} = \zeta(\gamma,\delta)/(\xi^{\delta(1-\nu)}N^{\delta\nu})$, the first term become $\delta(1-\nu)(\gamma+1/\delta)\ln(\xi_o/\xi)$. In ϑ_{ξ}, *N*=1 because we began with $\xi_o=1$ (corresponding to *one* mer) and we ended with ξ (corresponding to *one* effective mer). Then

$$\left(\Delta s_{1\to\xi}\right)_{r,U} =$$
$$k_{\mathrm{B}}\left\{(1-\nu)\delta\left(\gamma+\frac{1}{\delta}\right)\ln\left(\frac{\xi_o}{\xi}\right) - \zeta(\gamma,\delta)\left(\frac{1}{\xi^{\delta(1-\nu)}} - \frac{1}{\xi_o^{\delta(1-\nu)}}\right)(\xi)^{\nu\delta}\right\} \quad (13)$$

for the case where δ=2 and ν=1/2, ζ(γ,δ)=(γ+1/2). Explicitly using $\xi_o=1$,

$$\left(\Delta s_{1\to\xi}\right)_{r,T} = k_{\mathrm{B}}\left\{(\gamma+1/2)\ln\left(\frac{1}{\xi}\right) - (\gamma+1/2)\xi\left(\frac{1}{\xi}-1\right)\right\} \quad (14)$$

Hence, we obtain

$$\left(\Delta g_{1\to\xi}\right)_{r,T} = k_{\mathrm{B}}T\left\{\left(\gamma+\frac{1}{2}\right)\ln(\xi) + \alpha(\xi,1)\left(\frac{1}{\xi}-1\right)\right\} \quad (15)$$

where α(ξ,1)=ζξ is a weight that must be determined from the boundary conditions of Equation (15). Interestingly, Equation (15) resembles an expression for the change in free volume in Flory-Huggins theory.[72]

The biopolymer under study is frozen locally along only one axial direction, yet ξ in Equation (15) is one dimensional and uniaxial. Experimentally, long chains whose lengths are of similar order to ξ are not simple straight rod-shaped structures.[43] Therefore, we must scale this value by the dimensionality of the system (*D*) to properly express the frozen degrees of freedom in the link (where we assume *D*≡3 dimensions).

Using Equations (11) through (12), and solving for the chemical potential for ξ>1 yields

$$\mu = -T\left(\frac{\Delta s}{\Delta n}\right)_{r,U} = \frac{k_{\mathrm{B}}T}{D}\left\{\frac{(\gamma+1/2)\ln(\xi)}{\xi(1/\xi-1)} + \frac{\alpha(\xi,1)}{\xi}\right\} \quad (16)$$

where we have used Equations (12) and (14) to express the differential. The properties of μ require that $\lim_{\xi\to1^+}\mu=0$. Consequently, for δ≡2, α(ξ,1)=(γ+1/2)ξ.

Equation (16) simplifies to α/ξ→(γ+1/2). To obtain the total FE of a link of length ξ, Equation (16) must be integrated over the path 1→ξ,

$$\mu = -T\int_1^{\xi}\left(\frac{\Delta s}{\Delta n}\right)d\xi = \frac{(\gamma+1/2)}{D}k_{\mathrm{B}}T\int_1^{\xi}\left(\frac{\ln(x)}{(1-x)}+1\right)dx. \quad (17)$$

The change in the FE due to variation in ξ becomes

$$\Delta G_{1\to\xi} = \tilde{N}\mu = \frac{(\gamma+1/2)Nk_{\mathrm{B}}T}{D\xi}\int_1^{\xi}\left(\frac{\ln(x)}{(1-x)}+1\right)dx \quad (18)$$

where *D*≈3. The integral

$$f(\xi) = \int_1^{\xi}\left(\frac{\ln(x)}{(1-x)}+1\right)dx \quad (19)$$

can only be solved numerically for arbitrary ξ. Values for Equation (19) are tabulated in Table 1. From Table 1 or inspecting the limits of Equation (19), it can be seen that Equation (18) approaches Δ*G*→(γ+1/2)*Nk*$_{\mathrm{B}}$*T*/3 asymptotically for very large ξ, and Δ*G*=0 for ξ=1 (as expected). Landau and Lifshitz approximate the local entropy loss as a linear function of the Kuhn length.[73] Equation (18) can also be found empirically by calculating the FE of a structure with a good value for ξ (including the global entropy contribution to the FE), fixing the baseline of the total FE to a known experimental value and gradually increasing ξ and evaluating the FE using the global entropy. As ξ becomes large, the global contribution tends toward zero leaving only the local contribution. Hence $\lim_{\xi\to\infty} f(\xi)/\xi=1$: (*N*-1)/*N* in the limit is 1.

For the case where δ≠2 and ξ>1 nt, similar steps used to obtain Equation (18) yield a general expression (for δ>0, γ>0, and 0<ν<1)

$$-T\Delta S_{\xi\gamma\delta} = \Delta G_{\xi\gamma\delta}$$
$$= \left(\frac{N}{\xi}\right)\frac{k_{\mathrm{B}}T}{D}\int_{+1}^{\xi}\left\{\frac{(1-\nu)(\delta\gamma+1)\ln(x)}{1-x} + \varpi\zeta(\gamma,\delta)x^{\delta(2\nu-1)}\frac{1-x^{\delta(1-\nu)}}{1-x}\right\}dx \quad (20)$$

where, ϖ=(γ+1/δ)/ζ(γ,δ) is a stretching weight on the gamma function, and in general, we assume *D*=3. When δ≡2, then ϖ≡1 and Equation (20) reduces to Equation (18).

As shown in recent work,[1] the generalized case of ζ(γ,δ) need not correspond to a gamma function. In such cases, ζ takes the form of a weight. In that work, it was shown that the stretching component and

Table 1. A compilation of standard values for f (ξ) in Equation (19). It can be seen that this function approaches ξ asymptotically.

ξ	f (ξ)	f (ξ)/ξ
1	0.000	0.000
2	0.178	0.089
3	0.563	0.188
4	1.061	0.265
5	1.630	0.326
6	2.251	0.375
7	2.910	0.416
8	3.600	0.450
9	4.314	0.479
10	5.049	0.505
15	8.942	0.596
20	13.071	0.654
50	39.798	0.796
100	86.799	0.868
200	183.334	0.917
500	478.016	0.956
1000	973.419	0.973
2000	1968.300	0.984
5000	4960.654	0.992
10000	9954.077	0.995

the compression component are decoupled, at least within the framework of this class of approximations.

In Reference 3, Flory provides an example of polycatena sulfur in which the Kuhn length is shorter than the mer-to-mer distance (b) (see pp 157-159). Therefore, for completeness, we also should consider the case where $1>\xi$: the number of effective mers increases from 1 mer to $1/\xi$ effective mers.

Using Equation (11), Δn becomes $\Delta n = n_f - n_i = 1/\xi - 1 = (1)(1/\xi - 1)$. Consequently, the general expression for Δn is

$$\Delta n = \max\{\xi, 1\}\left(\frac{1}{\xi} - 1\right) \tag{21}$$

where $\max\{\cdots\}$ is the maximum value of the arguments. For biopolymers, $\xi_o \equiv 1 < \xi$, and Equation (21) expresses the *loss* in the number of mers; $\Delta n = \xi(1/\xi - 1) < 0$. If $\xi_o \equiv 1 > \xi$, then Equation (21) expresses the *gain* in the number of mers; $\Delta n = \xi(1/\xi - 1) > 0$.

Using similar steps, the case where $\xi<1$, Equation (14) becomes

$$\left(\Delta s_{1\to\xi}\right)_{r,T} = k_B\left\{(\gamma+1/2)\ln\left(\frac{1}{\xi}\right) - (\gamma+1/2)\left(\frac{1}{\xi}-1\right)\right\}$$

Following the remaining steps, yields the following chemical potential

$$\mu = -T\left(\frac{\Delta s_{1\to\xi}}{\Delta n}\right)_{r,U} = \frac{(\gamma+1/2)k_B T}{D}\int_1^{\xi}\left\{\frac{x\ln(x)}{(1-x)} + 1\right\}dx \tag{22}$$

and the following FE for the sequence as a whole

$$\Delta G_{\xi\to 1} = \tilde{N}\mu = -\frac{(\gamma+1/2)Nk_B T}{3}\int_{\xi}^{1}\left(\frac{x\ln(x)}{(1-x)} + 1\right)dx \tag{23}$$

with $0<\xi\le 1$. Equation (23) can occur when internal attractive forces are significant;[3] however, this is not applicable to RNA or proteins.

Variable Kuhn Length for heterogeneous monomers and block copolymers

So far, we have handled the problem with a presumed static Kuhn length. For many systems, this is a reasonable assumption and restriction. However, functional RNA sequences are likely to have a distribution of Kuhn lengths rather than a single Kuhn length, because such RNA contains recognition regions, scaffolding, etc. We seek to generalize the expressions of the previous Section to reflect this variable Kuhn length.

If the Kuhn length is variable over the length of the sequence, then we suppose that the sum of the individual Kuhn lengths (ξ_i) should equal the total number of monomers, *i.e.*,

$$N = \sum_i \xi_i \tag{24}$$

This suggests that we can write Equation (20) as a summation

$$\Delta G_{net} = -T\Delta S_{net} = \sum_i \Delta g_i(\xi_i, \gamma, \delta)$$

$$= \sum_i \frac{k_B T}{D}\int_{+1}^{\xi_i}\left\{\frac{(1-\nu)(\delta\gamma+1)\ln(x)}{1-x} + \varpi\zeta(\gamma,\delta)x^{\delta(2\nu-1)}\frac{1-x^{\delta(1-\nu)}}{1-x}\right\}dx \tag{25}$$

Simplifying Equation (25) to a form like Equation (18) yields

$$\Delta G_{net} = \sum_i \frac{(\gamma+1/2)k_B T}{D}\int_1^{\xi_i}\left(\frac{\ln(x)}{(1-x)} + 1\right)dx \tag{26}$$

For the limiting case of large x_i, the local FE approaches

$$\Delta g_i(\xi_i, \gamma, \delta) \to \frac{(\gamma+1/2)\xi_i k_B T}{D} \tag{27}$$

and using Equation (24), it is clear that the total free energy caused by this entropy change is again approximately $(\gamma+1/2)Nk_BT/D$. Hence, the model easily adapts to a more general expression with little difficulty.

This reasoning, where we break down the problem into smaller units of $\Delta g_i(\xi_i, \gamma, \delta)$, can also be applied to polymers with heterogeneous monomers of different size mer-to-mer distance (b). This would permit treatment of block copolymers.

In the first part, we assumed $\xi_o \equiv 1$. For a homogeneous polymer, this is a valid assumption. However, there are situations where the polymer consists of a heterogeneous mixture of monomers with different lengths in the chain: for example, the cap region of messenger RNA (mRNA). In such situations, the N monomers should add such that

$$L = \sum_i b_i = b\sum_i \xi_{oi} = N\bar{b} \quad \text{with} \quad \bar{b} = \frac{1}{N}\sum_i b_i \tag{28}$$

where b_i is the mer-to-mer distance between mer i and mer $i+1$, $\bar{b}$ is the mean mer-to-mer separation distance, $\xi_{oi} \ge 1$ (given at least one i satisfies $\min\{\xi_{oi}\}=1$) and b must be defined as the minimum length $\min\{b_i\}=b$.

Let $\Delta s_{\xi_{oi}\to\xi_i}$ and $\Delta g_{\xi_{oi}\to\xi_i}$ represent the change in entropy and FE (respectively) of these effective mers (of initial length ξ_{oi}) due to a change $\xi_{oi}\to\xi_i$. For a system of constant r and T, the Kuhn length correction FE of a single link is

$$\left(\Delta g_{\xi_{oi}\to\xi_i}\right)_{r,T} = \mu_i\Delta n = -T\left(\Delta s_{\xi_{oi}\to\xi_i}\right)_{r,U} \tag{29}$$

Following the same procedure as outlined in the previous Section, we substitute ξ_{oi} and Equation (2) into Equation (29) and solve for $\Delta g_{\xi oi\to\xi i}$ with $\delta \equiv 2$. This yields

$$\left(\Delta g_{\xi_{oi}\to\xi_i}\right)_{r,T} = k_B T\left\{\left(\gamma+\frac{1}{2}\right)\ln\left(\frac{\xi_i}{\xi_{oi}}\right) + \alpha(\xi_i, \xi_{oi})\left(\frac{1}{\xi_i} - \frac{1}{\xi_{oi}}\right)\right\} \tag{30}$$

Now, generalizing Equation (12) for the change in the number of effective mers

$$\Delta n \sim \max\{\xi_i, \xi_{oi}\}\left(\frac{1}{\xi_i} - \frac{1}{\xi_{oi}}\right) \tag{31}$$

leads to

$$\mu_i = -T\int_{\xi_o}^{\xi_i}\left(\frac{\Delta s}{\Delta n}\right)d\xi = \frac{(\gamma+1/2)k_B T\xi_{oi}}{D}\int_1^{\xi_i/\xi_{oi}}\left(\frac{\ln(x)}{(1-x)} + 1\right)dx \quad . \tag{32}$$

Similarly, for $\xi_i<\xi_{oi}$

$$\Delta g(\xi_i, \xi_{oi}) = \mu_i = -\frac{(\gamma+1/2)k_B T\xi_{oi}}{D}\int_{\xi_i/\xi_{oi}}^{1}\left(\frac{x\ln(x)}{(1-x)} + 1\right)dx\,. \tag{33}$$

From here, the problem can be broken up into a heterogeneous mixture of mers of different mer-to-mer separation distance (b). On an immediate level, the main application of this methodology is in computing heterogeneous systems such as the cap region of mRNA, or mixing of strands of RNA with DNA or the mixing of proteins with sugar chains in glycoprotein structures. With a diverse network of complex side chains, the treatment can expand to more heterogeneous monomers than nucleic acids (which are at least of similar size and chemical behavior).

Stem binding and destabilization free energy

The rough linear dependence of large ξ on the local entropy accounts for the entropy corrections in the model so far presented. However, it does not, in of itself, treat the problem of structures crinkling up as in

Figure 1B. Here we explain how we made vsfold handle these issues and present a theoretical justification for why these corrections are a physical manifestation of the Kuhn length and the local entropy presented in the previous two Sections.

The global entropy generally tends to correct for cases like Figure 2A, where $L_{stem}>>\xi$. In such stem-loops, the cumulative weight of the global entropy renders the formation of such stems ($L_{stem}>>\xi$) unfavorable when the compensating base pairing energies are insufficient. Therefore, if the stem length (L_{stem}, in units of mer separation distance b) is as long (or longer) than ξ ($L_{stem}\geq\xi$), Figure 2A), then the discussion in the previous two Sections can be applied with only a little modification. However, when ($L_{stem}<\xi$, Figure 2B), this means the region should be straighter and more inflexible than expected and forces are needed to enforce this state condition. The two interacting strands that form the stem should easily tear apart because the unbound regions ($\Delta x=\xi-L_{stem}$, Figure 2B) have nothing to hold them down. Inasmuch as polymers form bramble-like patterns, these twisted structures must reflect some sensible function of the polymer's Kuhn length: an experimentally discernible quantity for which an average value can be measured from the radius of gyration of the polymer. The accounting methods in the previous two Sections do not exact a cost for the formation of brambles, like in Figure 1B, that are inconsistent with the actual Kuhn length: stem structures should be roughly as long (or longer) than the Kuhn length.

It is known that dsRNA (*i.e.*, without the folded ssRNA looping) and dsDNA can achieve Kuhn lengths on the order of 50 nm (about 150 bps).[18,21,29,44,57] However, well below the melting temperature, dsRNA consists of a single contiguous stem and does not contain large regions of loops where disordered structure is likely. Near the melting temperature, this disorder is modeled in dsDNA with a loop entropy and a much smaller Kuhn length.[21,29] Folded ssRNA typically consists of contiguous stems that are much shorter than 150 bps (commonly 5 to 10 bps).[37-39,43] It therefore makes no sense in folded ssRNA problems to claim that stem regions have a Kuhn length of 150 bps, when the longest coherent stem in the structure is only 5 bps or even 10 bps. Kuhn lengths should be properly accounted for in the FE; particularly because the stem lengths and single strand lengths are often of similar order. One way to do this is to propose an average Kuhn length (as is done in vsfold currently) and hope that the FE functions are forgiving enough to compensate for small discrepancies. This approach appears to be fairly successful for many types of RNA structures. However, surely some local information about the flexibility will be lost. Therefore, we propose that a reasonable method of accounting for the FE due to changes in flexibility is one where, in the free strand regions, the Kuhn length is about 3 nt (as some models for dsDNA melting propose for the *bubble* regions)[9,10,27,28,31,32,59,74] and, in the stem regions, we propose that the Kuhn length should be proportional to the length of the stem (L_{stem}, Figure 2C). There are a plethora of issues associated with defining a stem that we cannot afford to delve into here. Nevertheless, it should be reasonable, we think, to say that something that looks reasonably contiguous probably is, and something that looks like a junction, probably is not contiguous and therefore not part of a stem.

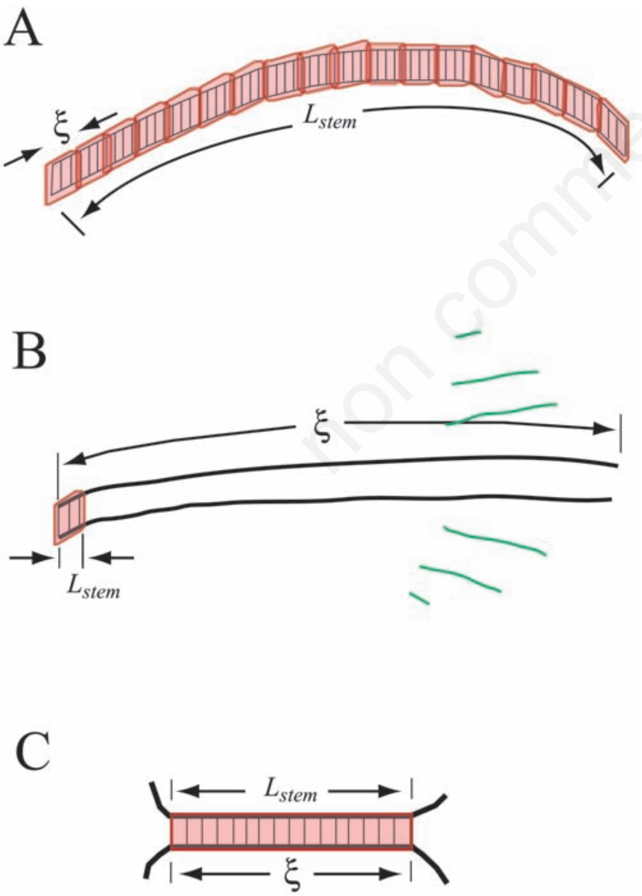

Figure 2. Examples of different cases for the Kuhn length (ξ) and stem length (L_{stem}). A) Case where $L_{stem}>>\xi$. B) Case where $L_{stem}<<\xi$. C) Case where $L_{stem}=\xi$.

In Equation (18), in the limit of long ξ,

$$\Delta G \to (\gamma+1/2)k_B TN/D \qquad (34)$$

i.e., the FE approaches that of a set of N free particles. For γ=1, $\Delta G\approx(1/2)Nk_BT$ and for the standard value (γ=1.75) used in the JS-model, $\Delta G\approx(3/4)Nk_BT$.

[Note that, since the total translational kinetic energy of N free particles is $(3/2)Nk_BT$ and ΔG (when γ=1.75) is half this value, the local entropy (corresponding to ΔG) is essentially expressing a tethered system in which half is free and half is constrained (where the translational motion is neglected in the free dimensions). Since these are simply beads on a chain, they are effectively particles. The equal partition of the energy means that the kinetic energy of the *free particle* is $(3/2)k_BT$. Hence, the calculated value with γ=1.75 has a physical basis and the free motion of the tethered system is fractal and not completely 2D as would be the case were this a pure Gaussian type chain (γ=1)].

We are concerned with stem segments that are shorter than the Kuhn length $L_{stem}<\xi$. We define the difference

$$\Delta\xi = \begin{cases} \xi - L_{stem}, & \xi > L_{stem} \\ 0, & \xi \leq L_{stem} \end{cases} \qquad (35)$$

which expresses the additional artificially constrained stem length. Since the stems come together and interact independently, the interaction of the unbound regions is the sum of all possible configurations (Figure 2B). Based on the simple linear relationship for ξ in Equations (27) and (34), this suggests that we should integrate, which yields

$$\Delta G_{bend} = \begin{cases} (\gamma+1/2)k_BT\int_0^{\Delta\xi} z\,dz = \frac{1}{2}(\gamma+1/2)k_BT(\Delta\xi)^2, & L_{stem}<\xi \\ 0, & L_{stem}\geq\xi \end{cases} \qquad (36)$$

where the weight $1/D$ is omitted because the entropic contribution of bps to the FE is not added to this FE. This freezing out cost is usually paid by contributions from bps formation; as noted in previously, the average is around -25 cal/molK for both RNA and DNA.[69,70] These highly local bp formation costs are generally applied as a constant for some set of dinucleotide bps. However, for the structure in Figure 2B, these costs must be born solely by the fictitious stem and therefore, the best accounting we can do is to restrict all 3$\Delta\xi$ degrees of freedom.

The quantity ΔG_{bend} is defined as the stem bending/destabilization FE. The property is only invoked when the stems are shorter than ξ and the property tends to be the main factor enforcing a straighter structure within the RNA and protein structure calculations.[6] The exact form of this interaction is unknown; however, based on computer experiments using vsfold, we found that a quadratic function in $\Delta\xi$ was effective in rooting out distortions such as those seen in Figure 1B.

It also follows that, because ξ is finite in dsRNA, there is some maximum (ξ_{max}) for the stem in folded single-stranded RNA such that

$$\xi = \begin{cases} L_{stem}, & L_{stem} < \xi_{\max} \\ \xi_{\max}, & L_{stem} \geq \xi_{\max} \end{cases}, \tag{37}$$

where the best value should be $\xi \approx L_{stem}$.

For example, when $L_{stem} >> \xi$ (Figure 2A), the global entropy exacts a large cost. Likewise, if $L_{stem} << \xi$ (Figure 2B), then the local entropy exacts a large cost. If ξ and L_{stem} are nearly equal (Figure 2C), the weight of this contribution is simply the local contribution. Therefore, a proper value for ξ in the stem region is $\xi \approx L_{stem}$ and the minimum local free energy for the segment occurs when $\Delta G(\xi)=\Delta G(L_{stem})$. The way to estimate ξ_{max} will have to be discussed elsewhere, but ξ_{max} can reach lengths in excess of 200 bps in dsRNA.

Equation (36) can also be understood from the worm like chain model. Landau and Lifshitz (LL)[73,75] describe the deviation of a polymer chain from a straight rod as the inner product of two unit vectors $\mathbf{t}_a$ and $\mathbf{t}_b$ that run along the thread of the polymer chain, separated by a distance $\Delta n = \Delta\xi$ (in units of mers), Figure 3. From this, LL obtain a relation for the average deviation of the inner product angle

$$\overline{\mathbf{t}_a \bullet \mathbf{t}_b} = \overline{\cos\theta} = \exp\left(-\Delta n \cdot k_B T / \kappa\right) \tag{38}$$

where κ is the bending force constant (in units of energy) for small angle deviation θ (in units of radians). At high temperature, the mean square distance between the ends of the chain is $\langle r^2 \rangle = 2N\kappa b^2/k_B T$. From Equation (1), $\langle r^2 \rangle = \xi N b^2$, hence $\xi = 2\kappa/(k_B T)$. In essence, $\kappa \propto \xi$. In this important respect, the Kuhn length can be treated largely in the same way as the persistence length and is equal to roughly twice the persistence length.

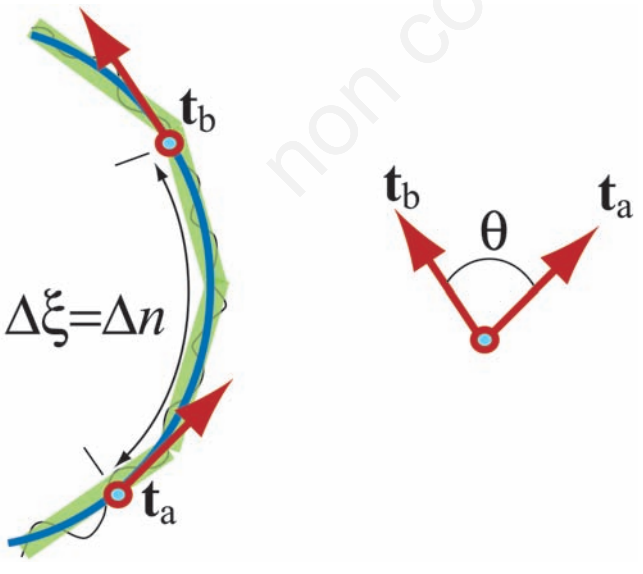

Figure 3. Cartoon describing the vector notation in the Landau and Lifshitz derivation of the end-to-end distance based on the worm-like chain model. The vectors $\mathbf{t}_a$ and $\mathbf{t}_b$ are along the contour of the polymer (Δn), and the angle between them (θ) is based on the inner product of the two vectors.

Since the joining of the two chains into a single chain requires both chains to be independently straight over their joining contour lengths, the entropy cost that both segments will accumulate, to maintain the additional mutually binding segment ($\Delta n = \Delta\xi$) and to satisfy $\xi = L_{stem} + \Delta\xi$, is

$$\Delta S = k_B \xi \int_{x=0}^{\Delta\xi} \ln\left\{\overline{\cos\left(\theta(x)\right)}\right\} dx = -k_B \xi \int_{x=0}^{\Delta\xi} \frac{k_B T x}{\kappa} dx = -k_B (\Delta\xi)^2$$

which is a similar form to Equation (36).

Note that $\xi = 2\kappa/k_B T$ is only valid for an isolated polymer, at best. It should not be assumed that this can be blindly applied to real polymer systems where other materials in the system (*e.g.*, other proteins in a cell) exert complex forces on that polymer.

The cross-linking entropy model and the Jacobson-Stockmayer equation

We have shown a general description of the CLE model. Now we show how we can reduce the CLE model to the terms found in the Jacobson-Stockmayer (JS) equation that is used in current RNA structure prediction schemes.[67,76]

As explained in Part I (second Section), the JS equation was derived from theoretical considerations;[77] however, the currently used Jacobson-Stockmayer equation is an empirical expression

$$-T\Delta S(n_L) = T\left(A_{JS} + \gamma k_B \ln(n_L)\right), \tag{39}$$

where n_L is the number of bases in a hairpin loop and A_{JS} is a constant obtained by fitting many sequences and finding the best fit.[67] Specifically, for loops less than 30 nt in length (and particularly so for loops less than 8 nt), Equation (39) is substituted with constant values based on these fits. For lengths greater than 9 nt, Equation (39) is used.

Figure 4 shows the FE contribution to loop formation at 37°C for several RNA data sets for the hairpin loop penalties plotted as a function of n_L and fitted using Equation (39); from the legend (magenta circles) mfold 3.0 (black circles)[67] mfold 2.3[78,79] – data obtained from the Wisconsin package (GCG) e99 parameter set – and (blue circles) the GCG e98 parameter set. The JS-equation is based on simplifications of the Gaussian polymer chain: two implicit ($\delta \equiv 2$ and $\nu \equiv 1/2$) and one explicit ($\gamma \equiv 1.75$). Fitting the e99 and e98 data sets using Equation (39) with variables γ and $A_{JS}(T_{37}=310.15\text{K}$, *i.e.*, 37°C), yields the following: $A_{JS}T_{37}=3.01\pm0.09$ kcal/mole and $\gamma=1.65\pm0.06$ for e99, and $A_{JS}T_{37}=3.3\pm0.1$ kcal/mole and $\gamma=1.8\pm0.1$ for e98.

The mfold 2.3 and 3.0 sets already have a fixed value for $n_L>9$ nt. For mfold 3.0, $A_{JS}T_{37}=4.0$ kcal/mol. Hence, the parameter γ agrees closely to $\gamma=1.75$, but A_{JS} differs between the three fits. Nevertheless, all data sets fit reasonably well to a logarithmic curve (of course, particularly for $n_L>9$).

All the data sets show considerable scatter for $n_L<9$ nt, where mfold 3.0 is the largest but also has the most data to support it.[68,80] The case of $n_L=4$ nt is consistent with the fact that the loop size (tetraloop) is of similar length scale to the Kuhn length of the free strand regions. Special corrections for unusually stable tetraloops are also used.[81,82] Likewise, for $n_L=3$ nt, the increase in FE mainly accounts for the triloop length, and specific loop sequences have their own specific corrections.[83]

Since A_{JS} is a single value for all hairpins, we suppose that Equation (39) expresses the properties of a generic stem-loop structure and that using these characteristic parameters in the CLE model will generate Equation (39) and A_{JS} from first principles.

First, from Equation (8), assuming $\delta \equiv 2$ and $\nu \equiv 1/2$, the global contribution to the entropy-loss due to the formation of a stem-loop of $\overline{n}_s$ bps becomes

$$\Delta S_L = -\frac{k_B}{\xi} \sum_{\{ij\}}^{\overline{n}_s} \left\{ \gamma \ln(\Psi_{1/2} N_{ij}) - \left(\gamma + \frac{1}{2}\right)\left(1 - \frac{1}{\Psi_{1/2} N_{ij}}\right) \right\}, \qquad (40)$$

where $\Psi_{1/2} = \xi / \lambda^2$ [Equation (8)], $\{ij\}$ is the set of base pairs comprising the stem and $\overline{n}_s$ is the *average stem length* of our generic RNA stem-loop structure. Certainly for large N_{ij}, the leading contribution to the entropy-loss comes from the first term inside Equation (40) and $1/(\Psi_{1/2} N_{ij})$ can be neglected. Since Equation (40) ratchets up with each additional base pair, if we form a stem of length $\overline{n}_s = \xi$, the FE-contribution from the entropy-loss (due to stem formation) is approximately $\xi T \Delta S(\overline{N}_{ij})$, where $\overline{N}_{ij}$ is the midpoint of the stem. This is essentially what happens in the examples we show in the Appendix (Figures A1 and A2). This yields the second (variable) term in Equation (39), viz.

$$-T\Delta S(\overline{N}_{ij}) = \gamma k_B T \ln(\Psi_{1/2} \overline{N}_{ij}). \qquad (41)$$

Note, however, that Equation (39) refers to the free strand in the hairpin loop ($n_L = j-i-1$) whereas Equations (40) and (41) refer to the cross-link: respectively, $N_{ij} = j-i+1$ and $\overline{N}_{ij} = \overline{j} - \overline{i} + 1$ (with and denoting average positions of i and j). Hence, based on Part I of this series and the treatment here, a better approximation might be expected to come from evaluating this contribution from the midpoint of the *stem*.

Second, we incorporate the concepts of the local CLE introduced earlier. To do this, we must also consider that stems are typically stiff and loop regions are typically flexible. Reference 1 shows how to incorporate a variable Kuhn length into the global entropy. When the sequence is subdivided into stiffer and looser regions, the CLE can evaluate these variations independently. This fact allows us to suppose two distinct Kuhn lengths: one for loops (ξ_L) and one for stems (ξ_s).

Now we suppose that the constant (A_{JS}) in the JS expression comes from the renormalization constant ($\Delta G_\xi / \tilde{N}$): introduced in Equation (18) and Equation (20) as applied to a generic stem-loop structure

$$T_{37} A_{JS}^{cle} = 2\Delta G_\xi(\overline{n}_s, \xi_s) + \Delta G_\xi(\overline{n}_L, \xi_L) \qquad (42)$$

where A_{JS}^{cle} is the derived JS constant (A_{JS}), $\overline{n}_s$ becomes the *average stem length* and $\overline{n}_L$ is the *average loop length* of the generic RNA stem-loop structure. The factor of two in Equation (42) is from the fact that the two independent single-stranded parts of the chain must join together to form the double-stranded helix of the RNA stem and each segment achieves this new Kuhn length independently.

To compare the CLE model with JS-model in Equation (39), we use all the same parameters as the JS model. We fit the various versions of the JS parameters to

$$-T_{37}\Delta S_{cle} \approx -T_{37}\Delta S_L + T_{37} A_{JS}^{cle} \qquad (43)$$

using the adjustable parameters $\overline{n}_L$, $\overline{n}_s$, ξ_L, and ξ_s. Note that Equation (42) should generally use n_s and n_L, yielding a more general expression $\Delta S_{cle} = \Delta S_L + \Delta S_{local}$. Hence, A_{JS}^{cle} is understood to express an averaged local correction ($-\overline{\Delta S}_{local}$).

In Table 2, the JS-model for a hairpin loop of RNA (using mfold 3.0 parameters) is compared with the CLE model for the same hairpin loop calculated at T_{37}. The stem length is fixed at 8 bps and the loop length is permitted to vary between 4 nt and 30 nt. The best fit turned out to be $\overline{n}_L = 8$ nt (with ξ_L=2.5 nt) and $\overline{n}_s = 8$ bp (with ξ_s=8 nt).

With these settings, the value of A_{JS}^{cle} is close to A_{JS} using mfold 3.0 parameters. This differs from the mfold 2.3 parameters where the best fit turned out to be $\overline{n}_L = 5$ nt (with ξ_L=3.0 nt) and $\overline{n}_s = 7$ bp (with ξ_s=7 nt), yielding $A_{JS}T_{37}$=3.30 kcal/mol. For the GCG e98 parameter set, we found that a good fit was $\overline{n}_s = 7$ [bp], $\overline{n}_L = 5$ [nt] with ξ_s=5.0 [bp] and ξ_L=10.0 [nt], yielding $A_{JS}T_{37}$=3.06 kcal/mol, which is rather skewed compared to the new parameters. The findings in this Section show exactly why the mfold 2.3 and 3.0 JS-parameters are significantly improved.

Recall that A_{JS} was originally determined by fitting the parameter (A_{JS}) to a large training set of sequences with known RNA secondary structures. Therefore, it is implicitly a mean value. Since a typical catalogue of loops and stems would likely contain many loops of length 3 to 8 nt and many stems of length 5 to 10 bps, these parameter setting are a very reasonable estimate for a generic stem-loop. Table 2 shows that the constant A_{JS} comprises the renormalization contribution for a *generic stem-loop* of RNA with stem length 8 bp (ξ_s=8 nt) and loop length 8 nt (ξ_L=2.5 nt). The CLE model generates the JS-model from first principles.

The JS equation also originates from the same Gaussian polymer chain model discussed from the second Section to this point. (A derivation of it can be found in Part I of this series in Appendix A). From this perspective, we see that the Jacobson-Stockmayer equation is an impressively simple (and often effective) way to model *rather short* regions of generic stem-loop RNA structures.

Finally, as mentioned in Appendix A of Part I, the empirically derived value for A_{JS} (that fits the training data set) cannot be generated from theory by using the JS equation in its original form.[77] It can only be obtained by evaluating Equation (42), or a similar expression, to find A_{JS}^{cle}. In Part I (Appendix A), it was shown that the JS equation yields a constant of the form

$$A_{JS} = k_B \left[(3/2)\ln\left(2\pi\xi/3\right) - \ln(v_s / b^3) \right] \qquad (44)$$

Since the size of the bound structure should be on the order of ξb, this means that $v_s \approx (\xi b)^3$. Substituting into Equation (44), we find

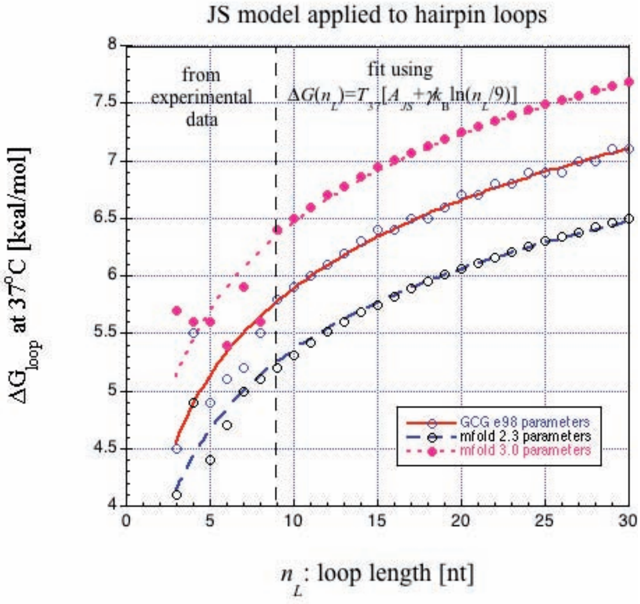

Figure 4. Plot of the hairpin-loop free energy data (at 37°C) for different loop lengths (n_L) and for 3 different parameter sets: (magenta circles) mfold 3.0 parameters,[68] (open blue circles) mfold 2.3 parameters from GCG e99 table,[78,79] (open black circles) parameters from the GCG e98 table.

$$A_{JS} \approx \frac{3k_B}{2}\ln\left(\frac{2\pi}{3\xi}\right) \qquad (45)$$

For a typical value such as ξ=5 mer, A_{JS} is negative. Hence, even aiming to fit this with realistic values for ξ (which is not even in the original formulation), A_{JS}>0 simply cannot be satisfied. This is why A_{JS} has historically been treated as an experimental parameter.

It should also be noted for the record, that in the original derivation of the JS equation, the existence of a Kuhn length (stiffness) was acknowledged.[77] However, issues related to the Kuhn length (and particularly the nature and size of υ_s) were rarely considered explicitly in subsequent usage of the JS expression.

Errors generated by invoking a static Kuhn length

As justified in previous work,[1] in the previous Section, we used a model with a variable Kuhn length where the stems were assigned ξ=7 nt and the loop ξ=3 nt. However, vsfold currently calculates the FE under the assumption that the Kuhn length is constant over the entire RNA sequence. Here we consider how such errors may affect the calculation.

In methods that assume a static ξ (like vsfold), a typical error is to overestimate the Kuhn length in the loop regions (ξ_L) where there is greater flexibility and constraining stem lengths to some predetermined fixed stem length (ξ_s). As mentioned in the Section on stem destabilization, vsfold will usually reject stems much shorter than the Kuhn length because the stabilization costs render such stems highly unfavorable. Therefore, it is important to understand the contributions to the error introduced by using a Kuhn length larger than the true value as often occurs when fitting with a static Kuhn length.

Figures 5A and 6A compare the deviation in FE when a static ξ is used on the whole structure (ΔG(sum,S)) with the case where a variable Kuhn length (ΔG(sum,V)) is used on the same structure. In both Figures, the same structure as the previous Section is used: a structure with stem length of 7 nt and a variable loop length. For the variable Kuhn length structure in Figures 5 and 6, the fit uses ξ_s=7 nt for the stem and ξ_L=3 nt for the loop region. For the static Kuhn length, ξ=5 nt in Figure 5 and ξ=7 nt in Figure 6. The static Kuhn length of ξ=5 nt is used because it is the average of 3 and 7 nt. Similarly, the static Kuhn length of ξ=7 nt is tested here because the most common problem faced in fitting structures using vsfold (with a the static Kuhn length) is an over estimate of the entropy corrections in the loop regions. Unlike the previous Section, here we use the actual stem lengths and loop lengths to compute the local entropy correction

$$\Delta G_{local} = 2\Delta G_{\xi}(n_s,\xi_s) + \Delta G_{\xi}(n_L,\xi_L) \qquad (46)$$

i.e., not just the mean loop length as in Equation (42).

Figure 5B shows the deviation in the FE due to using a static Kuhn length of 5 nt, where the total deviation $\Delta\Delta G$(sum,V-S) expresses the difference between ΔG(sum,V) and ΔG(sum,S) in Figure 5A (red data points with green dashed line). The static Kuhn length has a total error [$\Delta\Delta G$(sum,V-S)] that is typically less than ±0.2 kcal/mol. Similarly, Figure 6B shows the same deviation in the FE due to using the greatly oversized static Kuhn length of 7 nt. Clearly, there is far greater deviation when the Kuhn length is much larger than it should be (3 nt *vs* 7 nt in the loop region).

Figures 5B and 6B also compare the deviation in the global entropy contribution [indicated in the legend by $\Delta\Delta G$(global,V-S)] and the deviation in local entropy contribution [indicated by $\Delta\Delta G$(local,V-S)] due to the use of static Kuhn lengths 5 and 7 nt, respectively. Errors in the global entropy correction to the FE have a positive slope in both Figures because the true Kuhn length in the loop region is 3 nt and using 5 nt and 7 nt for the stat-

Table 2. A comparison between the JS-model (used in typical RNA secondary structure calculations) and the CLE model (discussed here) on a hairpin loop of length n_L and stem of length 8 bps. The temperature is set to the standard T_{37} value. For the CLE model, the Kuhn length of the stem region is set to 8 nt and the loop region 2.5 nt. The parameters used in Eqns (8) and (20) are δ≡2, γ≡1.75, and ν≡1/2 (exactly the same parameters as used implicitly and explicitly in the JS-model). Stem-loop lengths are discrete in the CLE model. For the average stem-loop, we used $\bar{n}_s = 8$ [bp], $\bar{n}_L = 8$ [nt]. Column 1 indicates the loop length, Column 2 is the global CLE contribution (Eqn (8)), Column 3 is the local CLE (Eqn (20)), Column 4 is the sum of Columns 2 and 3, Column 5 indicates the tabulated mfold 3.0 Jacobson-Stockmayer parameters,[68,80] and the last column is the difference between Columns 5 and 4. This shows that $-T_{37}\Delta S_{local} \approx A_{JS}T_{37}$.

Loop	Cross linking entropy model			JS-model	Difference
length	Loop	Local	Sum		
n_L	$-T_{37}\Delta S_L$	$-T_{37}A_{JS}\approx$	$-T_{37}\Delta S_{cle}=$	$-T_{37}\Delta S_{JS}(n_L)$	$-T_{37}\Delta(\Delta S)=$
		$-T_{37}\Delta S_{local}$	$-T_{37}(\Delta S_{local}+\Delta S_L)$		$-T_{37}[\Delta S_{JS}(n_L)-\Delta S_{cle}]$
[nt]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]	[kcal/mol]
4	2.113	3.873	5.986	5.600	-0.386
5	2.226	3.873	6.099	5.600	-0.499
6	2.331	3.873	6.204	5.400	-0.804
7	2.428	3.873	6.301	5.900	-0.401
8	2.519	3.873	6.392	5.600	-0.792
9	2.604	3.873	6.477	6.400	-0.077
10	2.685	3.873	6.557	6.500	-0.057
15	3.026	3.873	6.899	6.940	0.041
20	3.299	3.873	7.171	7.250	0.079
25	3.526	3.873	7.398	7.490	0.092
30	3.720	3.873	7.593	7.690	0.097

ic case underestimates the cost of forcing a looser chain with far more degrees of freedom to close. Errors in the local entropy correction to the FE have a negative slope in both figures because the local entropy contribution in the loop region (for ξ_L=3 nt) should be small. As a result, the local entropy correction with a static Kuhn length *over-scales* the free energy correction that was developed in the Sections on the derivation of the local CLE. The global error in Figure 5B starts out negative because the entropy contribution from the stem is underestimated. A crossover can be seen when the loop length reaches 10 nt. In this respect, slightly *over-scaling* the loop region and *under-scaling* the stem regions produced a nice balance for the static Kuhn length of 5 nt. In Figures 5B and 6B, the errors from the global and local contributions are largely self-canceling. In Figure 6, because an overall *over-scaled* value of 7 nt is selected for the static Kuhn length, the errors gradually grow significant. However, the total sum still exhibits a relatively slow deviation.

It should be clear that finding a good average value for ξ is certainly best. Nevertheless, the error introduced by a moderately poor choice is largely ameliorated by the self-correcting effects of the local and global contributions to the total entropy. Hence, there is only a gradual change in the baseline for improper values of ξ. Moreover, the error is an overestimate of the entropy contribution. Hence, when using a static Kuhn length, our best results will come only when we chose a good average value in general. For long sequences of highly variable ξ, it is unlikely that such a value for ξ can be found. Moreover, it should be remembered that these errors are cumulative, rendering calculations with a static Kuhn length all the more problematical. Therefore, a method for calculating with a variable Kuhn length is an important area of development for future versions of vsfold.

Strategy for calculating a variable Kuhn length

From Equations (25) and (46), we showed how to include the FE corrections with a variable Kuhn length for a stem-loop. For the simple stem-loop structure in Equation (46), it is not difficult to carry out such a calculation. However, when many calculations must be done for many different structures, this becomes a cumbersome issue. To simplify this

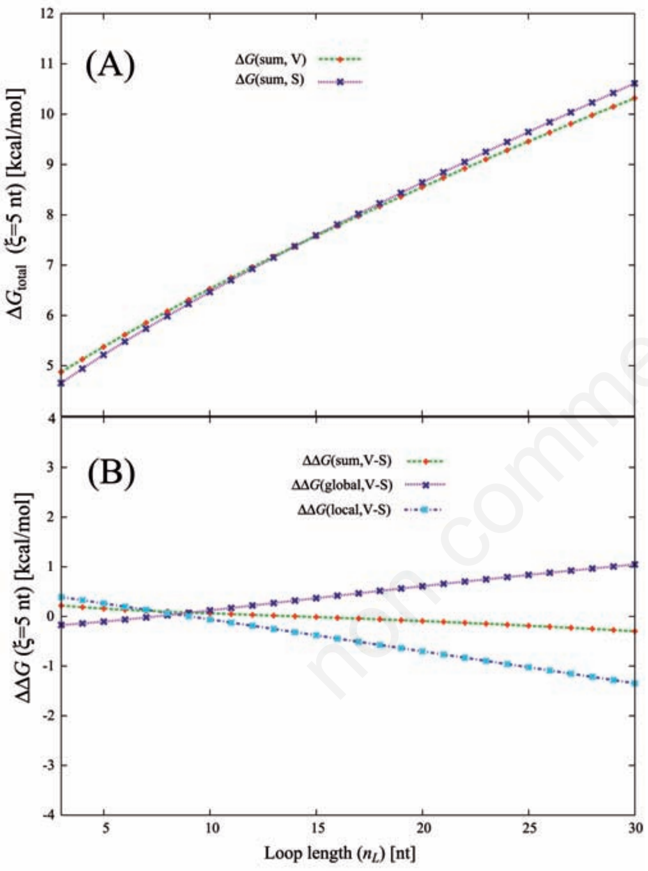

Figure 5. The error caused by using a static Kuhn length ξ=5 nt compared to the variable Kuhn length, for a structure with a stem of length 7 bp and a variable loop length (3 to 30 nt). The variable Kuhn length uses 7 nt for the stem and 3 nt for the loop. The static Kuhn length uses ξ=5 nt throughout. (A) The CLE contribution to the free energy (entropy loss) for the static Kuhn length (*G*(sum,S)) and the variable Kuhn length (*G*(sum,V)). (B) The differences in the CLE contribution to the free energy, where *G*(sum,V-S) = G(sum,V) – G(sum,S). These differences are further divided into the global *G*(global,V-S) and the local *G*(local,V-S) contributions. The negative slope in *G*(sum,V-S) indicates that the FE obtained by using the static Kuhn length tends to overestimate the entropy loss. ξ_S=7[bp] and ξ_L=3.0 [nt].

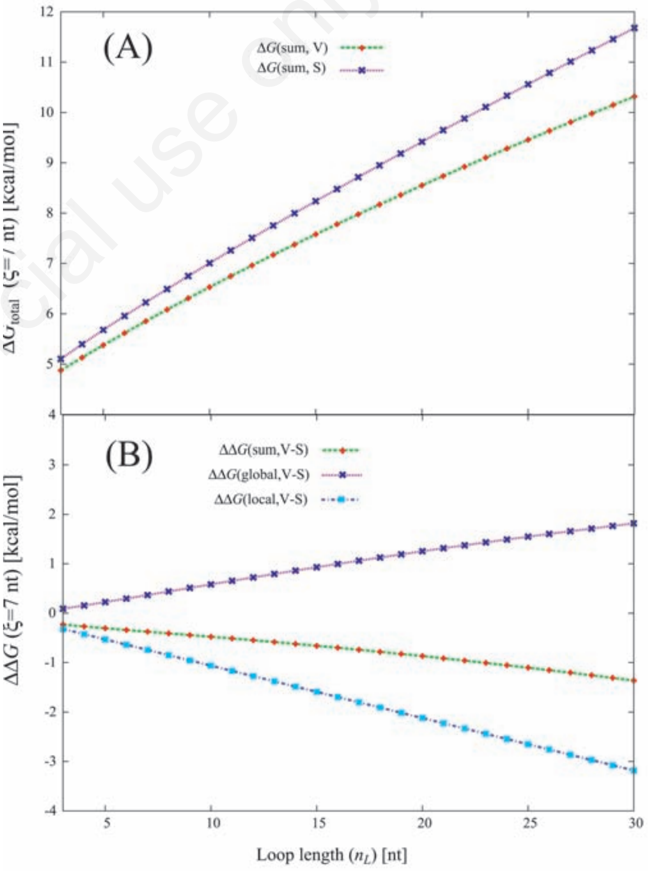

Figure 6. The error caused by using a static Kuhn length ξ=7 nt compared to the variable Kuhn length for the same stem-loop used in Figure 5, where the same parameters as Figure 5 are used for the variable Kuhn length. (A) The CLE contribution to the free energy (entropy loss) for the static Kuhn length (*G*(sum,S)) and the variable Kuhn length (*G*(sum,V)). (B) The differences in the CLE contribution to the free energy, where *G*(sum,V-S) = G(sum,V) – G(sum,S). These differences are further divided into the global *G*(global,V-S) and the local *G*(local,V-S) contributions. The negative slope in *G*(sum,V-S) indicates that the FE obtained by using the static Kuhn length tends to overestimate the entropy loss.

and facilitate corrections in a general object oriented programming approach, we suggest working from Equation (19).

As discussed in the Section deriving A_{JS}, the average Kuhn length in the free strand regions is approximately 3 nt, based on fitting A_{JS} to ξ_L and experimental data from References 37-39. Therefore, we use ξ_{bl}=3 nt as the baseline correction. The baseline FE is therefore $\Delta g(\xi=3)$. Now, to calculate other Kuhn lengths in the stem regions, we first rewrite Equation (18) in terms of Equation (19) as a free energy per nt

$$\Delta g_{nt}(\xi) = Cf(\xi)/\xi\,,\ C = (\gamma + 1/2)k_{\mathrm{B}}T/D\,, \tag{47}$$

and the local entropy contribution to the FE for a subsequence of length n is $n\Delta g_{nt}(\xi)$. If the Kuhn length differs from the baseline value (ξ_{bl}), the new Kuhn length must be added with the baseline value subtracted out. In a stem region where the length of the stem is L_{stem}, the Kuhn length for the dsRNA sequence (ξ_{ds}) becomes $\xi_{ds} \approx L_{stem}$ (Section on stem destabilization). Hence the corrections to the local entropy become

$$\begin{aligned} \Delta\Delta g_{bp}(\xi_{bl},\xi_{ds}) &= 2\left(\Delta g_{nt}(\xi_{ds}) - \Delta g_{nt}(\xi_{bl})\right) \\ &= 2C\left(f(\xi_{ds})/\xi_{ds} - f(\xi_{bl})/\xi_{bl}\right) \end{aligned} \tag{48}$$

where it is assumed that ξ_{bl}=3 nt and $\xi_{ds} \approx L_{stem} \geq 3$ nt. Therefore, to the baseline free strand (ξ_{bl}=3 nt), we simply add Equation (48) for any cases where we find ξ_{ds}>3 bps. Then the total local entropy becomes

$$\Delta G_{local} = N\Delta g_{nt}(3) + \sum_{\{s\}} \xi_s \Delta\Delta g_{bp}(3,\xi_s) \tag{49}$$

where $\{s\}$ is the set of stems that are formed in the RNA structure.

The remaining correction is the loop regions in contiguous stems, a group of stems joined by interior loops that are too short to render the group effectively independent and are therefore effectively a single stem.[4] We think corrections should be made for these because they are probably stiffer than the standard free strand, but their presence in a stem indicates that they are not simply the same as the rest of the stem. Functional RNA likely takes advantage of this to help melt the stem slightly to achieve a more desirable flexibility. At this point, we only suggest that, as in Equation (48), we can apply the same strategy

$$\Delta\Delta g_{nt}(3,\xi_l) = C\left(f(\xi_l)/\xi_l - f(3)/3\right) \tag{50}$$

where $\xi_l \geq 3$ could be evaluated as a function of the interior-loop separation-distance of the two stems and the total Kuhn length of the stem itself. These aspects will be examined in a future study and should also be tested experimentally.

Future plans are to incorporate a variable Kuhn length model into the vsfold program.

Now that we have constructed the calculation methodology, as a final step, it is worth testing the model in a real problem to see if, in fact, the model yields results that make sense.

To help motivate this, we use the sequence in Table 1 of part I of this series where the FE of a modified version of the HIV-1 tar sequence was evaluated using both mfold 2.3 and 3.0 base pairing parameters. It was clear in that study that the CLE model is less sensitive to the base pairing parameters. However, at that stage, we could not dig into the details of the part played by the local entropy.

For computing the structures in the previous two sections, $\xi \approx 3$ nt in the free strand region and $\xi \approx L_{stem}$ for the stem regions. Whereas this is certainly computationally convenient, the sharp boundaries between the stem and free strand Kuhn lengths seem better suited to a continuous function in which the Kuhn length of the hairpin loop extends partway into the stem region (presumably gradually increasing). Edge effects are known for stems,[69,79,84-86] perhaps extending 2 to 4 bps into the stem. Figure 7 shows this penetration into the stem region in terms of Equation (49), where the loop region is corrected for as though the loop region penetrated 2 bp, 4 bp, 6 bp and 8 bp into the stem.

A permeation of the free strand flexibility 8 bp into the stem region seems excessive. However, the edge effects from the 5'-3' ends of the stem-loop could add 4 bps and the hairpin side contribute the other 4 bps. (We did not correct for the global entropy at the edges, which should sharpen the boundaries.) If penetration into the stem is at least 2 bps from each end, this would explain why tRNA tends to have a Kuhn length of 4 nt throughout and why the Kuhn length does not expand if an extended structure can form. For tRNA, the free strand region (at both ends of the stem) penetrates too deeply for the RNA to stiffen into a straight stick shape.

In Figure 7, the FE tends to flatten out over a wide range of the Kuhn lengths for a penetration distance of 8 bps, ξ_{ds} could range between 13 to 20 bps in the center region of the stem. This suggests an average Kuhn length of about 12 nt, as suggested in the previous Section as a good choice for a fixed ξ. A variable Kuhn length emerges as a consequence of a synergy between base pairing FE effects (changing the global CLE) and the freezing out costs of at least ξ degrees of freedom in each chain (changing the local CLE). Introducing a penetration distance smooths out the competition: when a certain range of penetration of the loop ξ is included, the minimum FE is achieved. Therefore, an ideal model would require optimization of global and local entropy contributions to the FE in the context of base pairing and would require a continuous function for the Kuhn length. The precise nature of these function needs to be examined experimentally. This is the first attempt that we know of where variation in the Kuhn length is attributed to changes in the stem length.

Figure 7 also affirms the previous assertion that the renormalization process tends to leave the overall FE unchanged. Of course, the fixed Kuhn length shows an increase with longer Kuhn length; however, when the problem is solved more realistically with a variable Kuhn length, there is a wide range of values ξ where the change is small. Moreover, even for

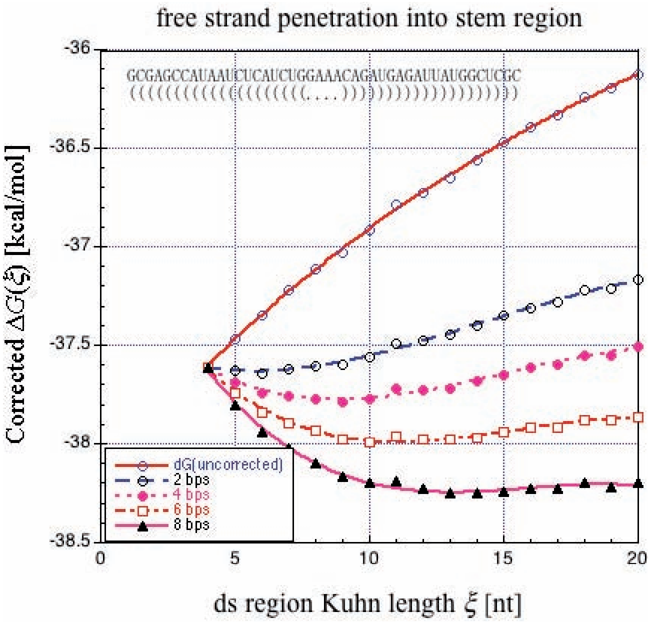

Figure 7. Calculation of a modified HIV-1 tar sequence (shown at the top of the graph) using a variable Kuhn length in the double strand helix part and the effects of the free strand Kuhn length penetrating into the stem region. Here a linear slope is used to express the change in Kuhn length in the transition region. Penetration is expressed in terms of bps and the free strand penetrates both ends of the stem equally. There is a clear flattening out of the free energy over a large range of Kuhn lengths when even 2 bp of penetration is used.

the fixed Kuhn length, the change is almost within thermal FE differences even though the Kuhn length varies from 4 to 20 nt.

This study of a variable Kuhn length also reveals another interesting aspect of RNA and protein folding. On the length scale of a Kuhn length, melting temperatures are clearly cooperative, because the melting of RNA has the typical sigmoid signature characteristic of this.[13,87-89] However, the thermodynamics evinced from this study suggest that on the macroscale, the overall character is more akin to order-disorder, if we take the fundamental unit of the polymer to be the effective mer. In other words, order-disorder on the scale of effective mers, cooperativity on the scale of monomers. Since the stem is largely the *de facto* mer in these systems, the effective mer is the unit of meaning in RNA biology.

Finally, the loop penalty (LP) model is the standard approach for calculating secondary structure in RNA, the most common version are mfold and the Vienna package. The LP-model has successfully predicted secondary structure in many cases.[78,90,91] The model relies on a Jacobson-Stockmayer model to evaluate the loop entropy at the closing point of a loop, and simply adds the base pair parameters for the dinucleotide base pairs in the structure.[77,92] This strategy appears to be successful in many cases and does not require all of the extensive effort that was needed to develop the CLE model.

However, the LP model succeeds because of two errors that just so happened to cancel each other.

First, as shown in Part I, the standard model underestimates the global entropic correction in long contiguous stems. This unphysical behavior encourages a straightening effect as a consequence. Because the training set, used to tune these prediction approaches, tends to include very stiff structures (long Kuhn lengths), it produces deceptive success when the target structures have inherently long Kuhn lengths. Such structures were easier to isolate experimentally and were the first found and determined. The standard entropy model is fairly successful at finding long straight stem structures, and if these also happen to be the training set of structures, it is no surprise that at least some of the predicted results would agree with the experimental data.

Second, all the experimentally obtained base-pairing thermodynamic parameters for RNA were built from oligonucleotide sequences of the same length and forming relatively short double-stranded sequences (typically 8 bps), where it is assumed that these parameters are valid for all stem lengths. It may indeed be possible to extrapolate their value for all dsRNA sequence lengths; however, it is not known.

The result was a chimera of relatively short range thermodynamic parameters for the stems weighted against a highly biased long (double strand) stem-search strategy. For a certain domain of problems often encountered in the field, this balance has worked to some extent.

However, in some cases, correctly predicted structures can require some careful cutting at the 5' and 3' ends (*manicuring*) to force them to come out with the structure that is actually observed experimentally. By neglecting the Kuhn length altogether and underestimating the global entropy costs, such approaches tend to generate poor predictions for sequences significantly longer than 100 nt (though obviously, we can expect some exceptions to any generalization). If one tries to insert additional randomly generated sequences or even repetitive sequences of the same structure plus some spacer, the original desired structure is often destroyed (though exceptions exist). Minor changes in sequence can produce major changes in the structure even when it is known that it should not (see for example, the comparison of predictions for tRNA sequences in Reference 1, Figure 7 and part I of this series). There is no discernible order in the energy of suboptimal structures and the location of the experimentally observed structure can be mixed in with the suboptimal structures like a needle in a haystack.[61]

In our first attempt to apply the CLE model in Reference 61, suboptimal structures from mfold 2.3 calculations (using the GCG package with e99 parameters) were recalculated using the CLE-entropy with the loop penalties used in mfold 2.3 subtracted out. It was essentially a filtering or re-ranking program. Because of the tendency for the LP-mode to over-predict long straight stem structures (as shown in Part I of this series), this made it easy for the filtering program to sort out the best available predictions from the list of suboptimal structures. However, because our first approach lacked any computational methodology to enforce a straightening effect of its own as a function of the Kuhn length (as developed in the Section on stem destabilizing FE), when only the global entropy was used to independently predict structures rather than just filter prediction results, the local effects of the Kuhn length soon became apparent. It was not obvious that structure predictions require straightening and structural constraints or filtering of the degrees of freedom to capture the characteristic features of polymer behavior. In particular, it was not obvious because of misleadingly successful predictions obtained by the commonly used LP-model (examined in Part I of this series).

Conclusions

We have derived and generalized the local entropy contributions in the CLE model in this work for studies of RNA secondary structure and RNA pseudoknot predictions. We have shown that the CLE model can be reduced to the standard Jacobson-Stockmayer model using first principles and simplifying approximations. Hence, the CLE generalizes the classic Jacobson-Stockmayer equation. We have also shown how to extend these concepts to variable Kuhn lengths, and how to handle heterogeneous monomers with different mer-to-mer separation distances. We have also shown how the Kuhn length is attributed to stem length and how this flexibility can be modeled into biopolymers using the CLE model. Whereas the use of a heterogeneous Kuhn length is far more desirable both aesthetically and from the viewpoint of accuracy of information, we also show here that, even when only an average Kuhn length is used, the CLE model is often robust enough to compensate for these deficiencies.

Future work is aimed at applying these concepts to double-stranded RNA, DNA and protein systems. In addition, greater focus will be aimed at modeling with a variable Kuhn length to help improve the prediction abilities of the vsfold program.

Postscript

At the time of printing, further developments have been introduced to describe the interface between the stem and free strand boundaries and improvements have been made in the predictions in Table 2 with this new information. These results are expected to be published sometime in 2014.

Software

A binary version of vsfold5 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/7, and Mac OSX4-8. Requests can be emailed to vsfold@gmail.com or dawson@bi.a.u-tokyo.ac.jp. A web version of the program can be found at http://www.rna.it-chiba.ac.jp/vsfold5 .

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;2:001-23.
2. Grosberg AY, Khokhlov AR. Statistical physics of macromolecules. New York: AIP Press; 1994.

3. Flory PJ. Statistical mechanics of chain molecules. New York: Wiley; 1969.
4. Dawson W, Fujiwara K, Kawai G. Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. PLoS One 2007;2:905.
5. Dawson W, Fujiwara K, Kawai G, et al. A method for finding optimal RNA secondary structures using a new entropy model (vsfold). Nucleotides Nucleosides Nucleic Acids 2006;25:171-89.
6. Dawson W, Kawai G, Yamamoto K. Modeling the long range entropy of biopolymers: a focus on protein structure prediction and folding. Rec Res Dev Experiment Theoretic Biol 2005;1:57-92.
7. Ma S-K. Introduction to renormalization group. Rev Mod Phys 1973;45:589-614.
8. McKenzie DS. Polymers and scaling. Phys Rep 1976;27C:35-88.
9. Gonzalez R, Zeng Y, Ivanov V, Zocchi G. Bubbles in DNA melting. J Phys Condens Matter 2009;21:034102.
10. Lando DY, Fridman AS. Role of small loops in DNA melting. Biopolymers 2001;58:374-89.
11. Ringrose L, Chabanis S, Angrand PO, et al. Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances. EMBO J 1999;18:6630-41.
12. Rouzina I, Bloomfield VA. Force-induced melting of the DNA double helix. 2. Effect of solution conditions. Biophys J 2001;80:894-900.
13. Rouzina I, Bloomfield VA. Force-induced melting of the DNA double helix 1. Thermodynamic analysis. Biophys J 2001;80:882-93.
14. Thomas TJ, Bloomfield VA. Chain flexibility and hydrodynamics of the B and Z forms of poly(dG-dC).poly(dG-dC). Nucl Acids Res 1989;11:1919-30.
15. Williams MC, Wenner JR, Rouzina I, Bloomfield VA. Effect of pH on the overstretching transition of double-stranded DNA: evidence of force-induced DNA melting. Biophysics J 2001;80:874-81.
16. Hearst JE, Schmid CW, Rinehart FP. Molecular weights of homogeneous samples of deoxyribonucleic acid determined from hydrodynamic theories for the wormlike chain. Macromolecules 1968; 1:491-4.
17. Record MT Jr., Mazur SJ, Melancon P, et al. Double helical DNA: conformations, physical properties, and interactions with ligands. Annu Rev Biochem 1981;50:997-1024.
18. Hagerman PJ. Investigation of the flexibility of DNA using transient electric birefringence. Biopolymers 1981;20:1503-35.
19. Harrington RE. DNA chain flexibility and the structure of chromatin nu-bodies. Nucl Acids Res 1977;4:3519-35.
20. Harrington RE. Opticohydrodynamic properties of high-molecular-weight DNA. III. the effects of NaCl concentration. Biopolymers 1978;17:919-36.
21. Moukhtar J, Faivre-Moskalenko C, Milani P, et al. Effect of genomic long-range correlations on DNA persistence length: from theory to single molecule experiments. J Phys Chem B 2010;114:5125-43.
22. Frontali C, Dore E, Ferrauto A, et al. An absolute method for the determination of the persistence length of native DNA from electron micrographs. Biopolymers 1979;18:1353-73.
23. Schellman JA. Flexibility of DNA. Biopolymers 1974;13:217-26.
24. Smith SB, Cui Y, Bustamante C. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. Science 1996;271:795-9.
25. Smith SB, Finzi L, Bustamante C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. Science 1992;258:1122-6.
26. Rief M, Pascual J, Saraste M, Gaub HE. Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. J Mol Biol 1999;286:553-61.
27. Destainville N, Manghi M, Palmeri J. Microscopic mechanism for experimentally observed anomalous elasticity of DNA in two dimensions. Biophys J 2009;96:4464-9.
28. Manghi M, Palmeri J, Destainville N. Coupling between denaturation and chain conformations in DNA: stretching, bending, torsion and finite size effects. J Phys Condens Matter 2009;21: 034104.
29. Palmeri J, Manghi M, Destainville N. Thermal denaturation of fluctuating finite DNA chains: the role of bending rigidity in bubble nucleation. Phys Rev E Stat Nonlin Soft Matter Phys 2008;77:011913.
30. Bar A, Kafri Y, Mukamel D. Dynamics of DNA melting. J Phys Condens Matter 2009;21:034110.
31. Rahi SJ, Hertzberg MP, Kardar M. Melting of persistent double-stranded polymers. Phys Rev E Stat Nonlin Soft Matter Phys 2008;78:051910.
32. Rapti Z, Smerzi A, Rasmussen KO, et al. Healing length and bubble formation in DNA. Phys Rev E Stat Nonlin Soft Matter Phys 2006;73:051902.
33. Ramprakash J, Lang B, Schwarz FP. of single strand DNA base stacking. Biopolymers 2008;89:969-79.
34. Murphy MC, Rasnik I, Cheng W, et al. Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy. Biophys J 2004;86:2530-7.
35. Seol Y, Skinner GM, Visscher K. Elastic properties of a single-stranded charged homopolymeric ribonucleotide. Phys Rev Lett 2004;93:118102.
36. Mills JB, Vacano E, Hagerman PJ. Flexibility of single-stranded DNA: use of gapped duplex helices to determine the persistence lengths of poly(dT) and poly(dA). J Mol Biol 1999;285:245-57.
37. Achter EK, Felsenfeld G. The conformation of single-strand polynucleotides in solution: sedimentation studies of apurinic acid. Biopolymers 1971;10:1625-34.
38. Eisenberg H, Felsenfeld G. Studies of the temperature-dependent conformation and phase separation of polyriboadenylic acid solutions at neutral pH. J Mol Biol 1967;30:17-37.
39. Inners LD, Felsenfeld G. Conformation of polyribouridylic acid in solution. J Mol Biol 1970;50:373-89.
40. 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.
41. 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.
42. 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.
43. Hagerman PJ. Flexibility of RNA. Ann Rev Biophys Biomolec Struct 1997;26:139-56.
44. Kebbekus P, Draper DE, Hagerman P. Persistence length of RNA. Biochemistry 1995;34:4354-7.
45. Abels JA, Moreno-Herrero F, van der Heijden T, et al. Single-molecule measurements of the persistence length of double-stranded RNA. Biophys J 2005;88:2737-44.
46. Gerland U, Bundschuh R, Hwa T. Force-induced denaturation of RNA. Biophys J 2001;81:1324-32.
47. Liphardt J, Dumont S, Smith SB, et al. Equilibrium Information from nonequilibrium measurements in an experimental test of Jarzynski's equality. Science 2002;296:1832-5.
48. Liphardt J, Onoa B, Smith SB, et al. Reversible unfolding of single RNA molecules by mechanical force. Science 2001;292:733-7.
49. 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.
50. Li PT, Collin D, Smith SB, et al. Probing the mechanical folding kinetics of TAR RNA by hopping, force-jump, and force-ramp methods. Biophys J 2006;90:250-60.

51. Li PT, Bustamante C, Tinoco I Jr, et al. Real-time control of the energy landscape by force directs the folding of RNA molecules. Proc Natl Acad Sci USA 2007;104:7039-44.
52. Li PT, Tinoco I, Jr. Mechanical unfolding of two DIS RNA kissing complexes from HIV-1. J Mol Biol 2009;386:1343-56.
53. Vieregg J, Cheng W, Bustamante C, Tinoco I Jr. Measurement of the effect of monovalent cations on RNA hairpin stability. J Am Chem Soc 2007;129:14966-73.
54. Green L, Kim CH, Bustamante C, Tinoco I Jr. Characterization of the mechanical unfolding of RNA pseudoknots. J Mol Biol 2008; 375:511-28.
55. Caliskan G, Hyeon C, Perez-Salas U, et al. Persistence length changes dramatically as RNA folds. Phys Rev Lett 2005;95:268303.
56. Hyeon C, Thirumalai D. Forced-unfolding and force-quench refolding of RNA hairpins. Biophys J 2006;90:3410-27.
57. Hagerman KR, Hagerman PJ. Helix rigidity of DNA: the meroduplex as an experimental paradigm. J Mol Biol 1996;260:207-23.
58. Honig B, Ray A, Levinthal C. Conformational flexibility and protein folding: rigid structural fragments connected by flexible joints in subtilisin BPN. Proc Natl Acad Sci USA 1976;73:1974-8.
59. Forties RA, Bundschuh R, Poirier MG. The flexibility of locally melted DNA. Nucleic Acids Res 2009;37:4580-6.
60. 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.
61. 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.
62. Feller W. An introduction to probability theory and its applications (pt I). New York: Wiley; 1968.
63. Flory PJ. Principles of polymer chemistry. Ithaca: Cornell University Press; 1953.
64. Nash LK. Elements of statistical Thermodynamics. Reading: Addison-Wesley; 1974.
65. Misra VK, Draper DE. A thermodynamic framework for Mg^{2+} binding to RNA. Proceedings of the National Academy of Science (USA) 2001;98:12456-61.
66. Volker J, Klump HH, Manning GS, Breslauer KJ. Counterion association with native and denatured nucleic acids: an experimental approach. J Mol Biol 2001;310:1011-1025.
67. Gray HB Jr., Hearst JE. Flexibility of native DNA from the sedimentation behavior as a function of molecular weight and temperature. J Mol Biol 1968;35:111-29.
68. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 1999;288:911-40.
69. Xia T, SantaLucia J Jr., Burkard ME, et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 1998;37:14719-35.
70. SantaLucia J Jr, Hicks D. The thermodynamics of DNA structural motifs. Ann Rev Biophys Biomol Struct 2004;33:415-40.
71. DeVoe H, Tinoco I Jr. The stability of helical polynucleotides: base contributions. J Mol Biol 1962;4:500-17.
72. Chan S-C, Dill KA. Solvation: how to obtain macroscopic energies from partitioning and solvation experiments. Ann Rev Biophys Biomolr Struct 1997;26:425-59.
73. Landau LD, Lifshitz EM. Statistical physics. Landau EML ed. London: Pergaman Press; 1958.
74. Alexandrov BS, Wille LT, Rasmussen KO, et al. Bubble statistics and dynamics in double-stranded DNA. Phys Rev E Stat Nonlin Soft Matter Phys 2006;74:050901.
75. Landau LD, Lifshitz EM. Theory of elasticity. London; Pergamon Press; 1989.
76. Lu ZJ, Turner DH, Mathews DH. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. Nucleic Acids Res 2006;34:4912-4.
77. Jacobson H, Stockmayer W. Intramolecular reaction in polycondensations. I. the theory of linear systems. J Chem Phys 1950;18:1600-6.
78. Freier SM, Kierzek R, Jaeger JA, et al. Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci USA 1986;83:9373-7.
79. Turner DH, Sugimoto N, Freier SM. RNA structure prediction. Ann Rev Biophys Biophys Chem 1988;17:167-92.
80. Giese MR, Betschart K, Dale T, et al. Stability of RNA hairpins closed by wobble base pairs. Biochemistry 1998;1094-1100.
81. Molinaro M, Tinoco I Jr. Use of ultra stable UNCG tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications. Nucleic Acids Res 1995;23:3056-63.
82. Zheng M, Wu M, Tinoco I Jr. Formation of a GNRA tetraloop in P5abc can disrupt an interdomain interaction in the Tetrahymena group I ribozyme. Proc Natl Acad Sci USA 2001;98:3695-700.
83. Shu Z, Bevilacqua PC. Isolation and characterization of thermodynamically stable and unstable RNA hairpins from a triloop combinatorial library. Biochemistry 1999;38:15369-79.
84. Applequist J, Damle V. Thermodynamics of the helix-coil equilibrium in oligoadenylic acid from hypochromicity studies. Journal of the American Chemical Society 1965;87:1450-1458.
85. Burkard ME, Kierzek R, Turner DH. Thermodynamics of unpaired terminal nucleotides on short RNA helixes correlates with staking at helix termini in larger RNAs. J Mol Biol 1997;290:967-82.
86. Freier SM, Petersheim M, Hickey DR, Turner DH. Thermodynamic studies of RNA stability. J Biomol Struct Dyn 1984;1:1229-42.
87. Hinz HJ, Filimonov VV, Privalov PL. Calorimetric studies on melting of tRNA Phe (yeast). Eur J Biochem 1977;72:79-86.
88. Privalov PL, Filimonov VV. Thermodynamic analysis of transfer RNA unfolding. J Mol Biol 1978;122:447-64.
89. Gluick TC, Draper DE. Thermodynamics of folding a pseudoknotted mRNA fragment. J Mol Biol 1994;241:246-62.
90. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 2003;31:3406-15.
91. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research 1981;9:133-48.
92. Jacobson H, Stockmayer W. Intramolecular reaction in polycondensations. II. the theory of linear systems. J Chem Phys 1950;18:1607-12.