# A new entropy model for RNA: part I. A critique of the standard Jacobson-Stockmayer model applied to multiple cross links

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

**[1]Department of Biotechnology, Bioinformation Engineering Laboratory, 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

The Jacobson-Stockmayer (JS) model is used in a number of standard programs for calculating the conformational entropy of RNA (and proteins). However, it is shown in this study that, in certain limiting cases, the current form of this model can lead to highly unphysical conclusions. The origin of this behavior can be traced to misunderstandings that occurred during the development of the model as applied to folded, single-stranded RNA. Here we show that an alternative model known as the cross linking entropy (CLE) model can overcome these issues. The principal object that causes entropy loss on a global scale in the CLE model is the *stem*, the primary measure of structural order in such coarse-grained calculations. The principal objects in the JS-model are various types of *loops*, and, with the exception of the hairpin loop, they are topologically local in character. To extract experimentally measurable variables, a simplified version of the CLE model is developed that resembles many features of the contact order model used in RNA and protein folding. These modifications are then applied to single molecule force-extension experiments (molecular tweezers) to extract quantitative information. It is further shown that a crude derivative of the CLE model itself can be derived directly from the JS-model when the misunderstandings are examined and corrected.

## Introduction

An important intermediate step in the ultimate goal of predicting the 3D structure of a RNA molecule from its sequence is to find information about the arrangement of the RNA molecule's base pairs. This arrangement of base pairs (bp) indicates the general topology of the RNA molecule, which is usually defined as its secondary structure, or, when present, its pseudoknot structure. It is essential that the topology information is correct. Base pairing information is, in part, dependent on achieving the correct prediction of the loop regions (Figure 1), where a loop consists of a region of a single-stranded RNA sequence that is closed by at least one stem (a set of base pairs, the hatched lines in Figure 1) and consists of nucleic acid bases that are relatively free and non-interacting due to weak non-Watson-Crick (non-WC) interactions inside the loop region (the blue segments in Figure 1). For loops, the Jacobson-Stockmayer (JS) equation (JS-model: see Appendix) is used in all of the current genre of thermodynamics based RNA secondary structure prediction approaches,[1-4] some RNA pseudoknot prediction approaches and in some of the protein topology prediction programsthat evaluate the entropy of loops (see Appendix for a definition of the term *secondary structure* as used for proteins and RNA).[5-9]

Correspondence: Wayne Dawson, Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, AIST Tokyo Waterfront BIO-IT Research Building, 2-4-7 Aomi, Koto-Ku, Tokyo, 135-0064, Japan. Tel. +81.3.3599.8630 - Fax: +81.3.3599.8085.
E-mail: dawson@bi.a.u-tokyo.ac.jp
wayne-dawson@aist.go.jp

Key words: entropy, RNA folding, protein folding, thermodynamics, bioinformatics.

Contributions: WD, wrote the manuscript and did the primary research; KY, advice, guidance and support during part of this research project; 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). We particularly thank Dr. Roderick Macrae for his very generous review, advice and comments on early versions of this manuscript and Dr. Robert J. Schneider (Berea College) for bringing Latin to life for wkd. We also thank Dr. Shingo Nakamura (Catalent Pharma Solutions), Profs Kentaro Shimizu, Shugo Nakamura, Tohru Terada, and Kazuya Sumikoshi (UofT), the students in Structural Biology Lab at CIT, Dr. Yasuhiro Futamura (IMCJ), Greg Rose, LMD, YZ, CS, YT, and countless others we may have missed, for their advice and encouragement.

Received for publication: 30 May 2011.
Revision received: 16 December 2011.
Accepted for publication: 6 February 2012.


*Journal of Nucleic Acids Investigation 2012; 3:e3*
*doi:10.4081/jnai.2012.e3*

The JS-model was originally developed as part of an effort to predict the fraction of ring polymer structures that would form in a condensation polymerization process.[1] Condensation polymerization involves processes like esterification of a monomer (mer) containing one alcohol group and one carboxylic acid group [*e.g.*, hydroxyundecanoic acid: $HOCH_2(CH_2)_9COOH$]. Whereas mers can only join at one point, the chemical reaction species cannot discriminate between a linear polymer chain and species resulting from a single polymer chain interacting with itself. As a result, part of the fraction would consist of ring structures (where the polymer had terminated the polymerization by closing up on itself) and the remaining fraction would consist of a linear polymer of some length that depended on concentration of monomer and other factors. The distribution in the size of the rings was found to correlate with the entropy, where the entropy was defined as a function of the length of the loop. The loops were closed at a single point because the chemical reactivity of the mers is restricted to two specific locations on the monomer with only one possible reaction mechanism.

The JS-model was later extended to any polymer that formed a loop structure. It was soon applied to early studies of double-stranded DNA (dsDNA) and dsRNA where multiple mismatches between sequences lead to the formation of interior loops that were closed by the mutual base pairing interaction between the independent DNA/RNA chains.[10-15] This approach continues to be successfully applied to this day.[16-18] Although the forces that cause loop formation in the dsDNA or dsRNA

structures arise from mere stacking interactions, collectively, they are often sufficient to maintain stable loop conformations. It was found that these loops could be roughly calibrated with the size and number of loops in the dsDNA or dsRNA structures and, therefore, could be evaluated as local free energy corrections.

Owing to the success of this approach, it was assumed that this could be applied in the same way in single-stranded DNA (ssDNA) and ssRNA structures that were folded in a similar way (locally dsRNA but resulting from a ssRNA sequence that folds back on itself with at least one hairpin loop).[19-21] For simple stem loops, this appeared to work. Along similar lines of reasoning, it has also been used to evaluate the free energy of turns in proteins.[8,9]

However, although the current (*in silico*) strategy of capping one end of a double-stranded DNA (or RNA) structure with a hairpin produces apparently superficial similarities between a dsDNA (or dsRNA) structure with mismatches in the base pairing, there is, in fact, no reason to assume that double strand and single strand interactions are universally the same with the exception of a few terminating hairpin loops capping the double strand architecture in the single strand cases. For one thing, although dsDNA can form a double helix that is four billion base pairs long in the human genome, a similar length for folded ssDNA or ssRNA has not been observed. Furthermore, the difference in energy between the predicted optimal structure and the observed structure can sometimes exceed 20 kcal/mol,[22] suggesting that the observed native state is metastable and quite far from thermodynamic equilibrium. Experimentally, growing single crystal samples of biomolecules in metastable conformations is usually a very difficult task. Some of these predicted structures do not involve particularly long sequences and, therefore, should be able to reach the optimal conformation within a few seconds. Yet it takes months to grow crystals and, most of the time, the crystal that grows is the one that is the most thermodynamically stable. Hence, valid or not, there is reason to question the conclusions that result from employing the JS-model in these folded ssRNA (or ssDNA) calculation approaches.

In previous work, we have introduced a new entropy model for calculating biopolymers, which we named the cross linking entropy (CLE) model.[23-28] In the first part of this Series, we examine the use of the standard entropy model that is based on the Jacobson-Stockmayer equation.[1] The objective is to look at the subtle role of this long range entropy effect that arises in polymers and to show why it is a misconception to equate the entropy evaluation of interactions between two independent strands in dsRNA structures with apparently similar single-stranded RNA secondary structure and likewise for other polymers such as dsDNA and protein topologies. Further, the objective is to show that the CLE model can be used to overcome these issues and provide a robust thermodynamics that can be applied to various experimental problems including manipulation of single RNA molecules using optical tweezers.[29] We find that the CLE model yields a consistent picture with quantitative results that are less sensitive to small differences in base pairing parameters or the particular sequence. Finally, we show that a type of CLE model could be anticipated from the JS-model once these misconceptions are removed. The focus is mainly on RNA secondary structure where the JS-model has been used with considerable success in the development of prediction algorithms that employ thermodynamic evaluation methods.[2-4] To understand the presentation in this work, the reader needs to be familiar with the basic concept of the Gaussian polymer chain (GPC),[30,31] the CLE model,[23-28] and the basic concepts of the thermodynamics of RNA secondary structure calculations.[32-38]

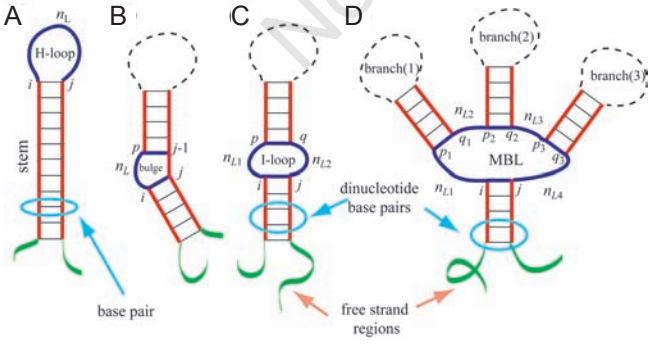


**Figure 1. Examples of secondary structure and the corresponding notations. (A) A simple hairpin loop (H-loop), (B) a bulge, (C) a interior loop (I-loop), (D) a multibranch loop (MBL). The parameters $n_L$, $n_{Lj}$ and $n_{Li}$ refer to the length of the free strand (blue) in a given loop. The stems are indicated by the red bars and black cross hatches. Base pairs and dinucleotide base pairs in the stem are marked in the Figure with the light blue circles. A distinction is made in this figure between free strand located in a loop region (blue) and free strand that has no loop associated with it (green).**

## The standard model: the loop penalty model (LP-model)

To this day, in RNA/DNA structure prediction, the entropy-loss due to folding is evaluated by a topologically *local* function derived from the JS-equation (1) (see also Appendix)

$$\Delta S(n) = -A_{JS} - \gamma k_B \ln(n_L) \quad (1)$$

where $A_{JS}$ is a fitted constant, $k_B$ is the Boltzmann constant (1.9872 cal/molK) and $\gamma$(=1.75) is a weight parameter that approximates the statistical characteristics of a self-avoiding random-walk where the walker must avoid points that have already been crossed in previous steps.[39] For hairpin loops (H-loops: Figure 1A, blue region), $A_{JS}$ is approximately 10 cal/molK and $n_L = j - i - 1$ is the enclosed, free single strand sequence length between bases $i$ and $j$ with $i<j$ (Figure 1A). The value of $A_{JS}$ in hairpin loops varies with the implementation of the Turner energy rules in various RNA secondary structure prediction programs.[35,37] In the GCG package, $A_{JS}$ is 10.6 cal/molK in the e98 set, and 9.74 cal/molK in the e99 set.[32] For the Vienna package (1.4 parameter set)[4,40] and mfold 3.0,[35] $A_{JS}$ is 12.9 cal/molK. The latter value is based largely on Serra *et al.*[41] In essence, $A_{JS}$ is about a factor of 5 larger than the Boltzmann constant.

For bulges $n_L = j - i - 1$ (Figure 1B, blue region), and, for interior loops (I-loops) $n_L = n_{L1} + n_{L2}$, where $n_{L1} = p - i - 1$ and $n_{L2} = j - q - 1$ (Figure 1C, blue region). Bulges and I-loops also have a fitted value for $A_{JS}$ that is of similar magnitude to the H-loop value. For multibranch loops (MBLs), an approximation is used

$$\Delta S = -C_0 - C_1 \sum_i n_{Li} - C_2 n_{br} \quad (2)$$

where $C_0$, $C_1$, and $C_2$ are all fitted parameters, $n_{Li}$ is the length of the free-strand segments of the MBL (Figure 1D, blue region), and $n_{br}$ is the number of branches. Branches consist of the stems that extend off from the MBL (Figure 1D).

Models conforming to Equations (1) and (2) assign penalties as a function of the total length of the free-strand segments enclosed by a given loop and by the type of loop that is formed. These free strand segments corre-

spond to the enclosed blue regions of the different types of loops shown in Figure 1A-D. The structures are topologically local because their entropy only depends on this free strand length in the immediate vicinity. Because the loop structures are considered isolated, they can be assigned penalties without taking into account any of the long range structure within the sequence the loop closes. Exceptions such as kissing loops are in principle admissible examples of long range tertiary structure; however, in general, even these are treated as though the interaction can be decomposed into two local penalties: one for each of the independent loops. We therefore call this approach the loop penalty model (LP-model).

In the LP-model, the penalties are combined with another topologically local base pairing free energy.[42,43] Evaluation of base pairing free energies usually consists of examining a lookup table of entropy and enthalpy terms for dinucleotide pairs (Figure 1) and evaluating the free energy (FE) for a given temperature,[32,35,37]

$$\Delta G_{bp}(T)=\Delta H_{bp}-T\Delta S_{bp} \quad (3)$$

where $T$ is the temperature, and $\Delta G_{bp}(T)$, $\Delta H_{bp}$ and $\Delta S_{bp}$ correspond to the FE, enthalpy and entropy of the base pairs, respectively. Supposing, for example, one were evaluating the sequence GGAGUAAUGUCC, then the underlined region would correspond to an H-loop (Figure 1A) and the bp sequence

$$\begin{matrix}5'\text{–}GGA\text{–}3'\\3'\text{–}CCU\text{–}5'\end{matrix}$$

would correspond to the stem region. The loop would be assigned the FE of a H-loop of length $n_L$=6; typically with some corrections for the closing bp

(which is $\begin{matrix}5'\text{–}AG\text{–}3'\\3'\text{–}UG\text{–}5'\end{matrix}$ in this example)

and sometimes other sequence related issues. Other structures such as bulges (Figure 1B), I-loops (Figure 1C) or multibranch loops (Figure 1D), would be handled similarly according to Equations (1) or (2), respectively. The stem region of folded single-stranded RNA (ssRNA) has the same physical geometry as an equivalent sequence of dsRNA. Therefore, it is logical to assume that they are identical. It follows that the topologically local base pairing terms would be assigned free energies corresponding to the particular dinucleotide pairs

( $\begin{matrix}5'\text{–}GG\text{–}3'\\3'\text{–}CC\text{–}5'\end{matrix}$ and $\begin{matrix}5'\text{–}GA\text{–}3'\\3'\text{–}CU\text{–}5'\end{matrix}$ )

and, when present, the FE evaluation would also include a single base or the first pair of non-WC bases that terminate the stem.

In the LP-model, both the loop terms and the base pairing terms are assigned according to the local environment regardless of where they are in the structure, only the local sequence relationships are assumed to require corrections. The FE is calculated by adding together these local dinucleotide bp contributions and the different local loop contributions that are assigned to the structure.

Protein structure models that utilize the JS-equation use reasoning identical to the LP-model used to calculate RNA secondary structure and sometimes neglect $A_{JS}$.[8,9] Typically, the JS-model is only applied to H-loops, so the model is not as general as the RNA structure model.

## Inspecting the standard (loop penalty) model

In this Section, we look at limiting cases of the JS-model where the calculations can render some surprising conclusions. The examples are meant to illustrate some of the pitfalls that can occur and are by no means the only cases where such unexpected behavior can occur. We first consider RNA in a little more detail, and generalize the point to proteins where the JS-model is also used.

### The loop penalty-model with RNA

For RNA, the base stacking FEs and loop penalty FEs are constant for a given temperature. One need only calculate these values once for a given temperature and then build a lookup table to calculate them for a particular structure.

Suppose we construct the following sequence; $A_{100}\underline{CCCC}U_{100}$. Suppose further that we measure this sequence at exactly the melting temperature ($T_m$) where $\Delta G_{bp}^{AU}(T_m)\equiv 0$ [kcal/mol], Equation (3). For this AU pairing, the corresponding enthalpy and entropy of base-pair stacking are $\Delta H_{bp}^{AU}=-6.82$ [kcal/mol] and $\Delta S_{bp}^{AU}=-0.019$ [kcal/molK], respectively.[35,44] Then $T_m=-\Delta H_{bp}^{AU}/\Delta S_{bp}^{AU}=359$K (86°C). There is also a small correction for the terminal AU: $\Delta H_{term}^{AU}=3.72$ [kcal/mol] and $\Delta S_{term}^{AU}=0.0105$ [kcal/molK], $\Delta G_{term}^{AU}(T_m)\equiv 0$. However, since this contribution cancels at $T_m$ we can neglect this interaction.

From this simple pattern, we can write the main contributions to this model as

$$\Delta G(T)=n_{bp}\left(\Delta H_{bp}^{AU}-T\Delta S_{bp}^{AU}\right)+T\left(A_{JS}+\gamma k_B\ln(n_L)\right) \quad (4)$$

where $n_{bp}$ is the number of base pairs and we assume a single value FE ( $\Delta G_{bp}^{AU}(T)$ ) for each bp. The first group of terms on the right-hand side corresponds to Equation (3) with a contiguous set of AU dinucleotide bps

$$\Delta G_{bp}^{AU}(T)=\Delta H_{bp}^{AU}-T\Delta S_{bp}^{AU} \quad (5)$$

and the last group of terms on the right hand side of Equation (4) comprises Equation (1) and is the FE contribution to the loop penalty for a H-loop of length $n_L$

$$\Delta G_L(n_L,T)=-T\Delta S_L(n_L,T)=T(A_{JS}+\gamma k_B\ln(n_L)).$$

Hence,

$$\Delta G(T)=n_{bp}\Delta G_{bp}^{AU}(T)+\Delta G_L(n_L,T). \quad (6)$$

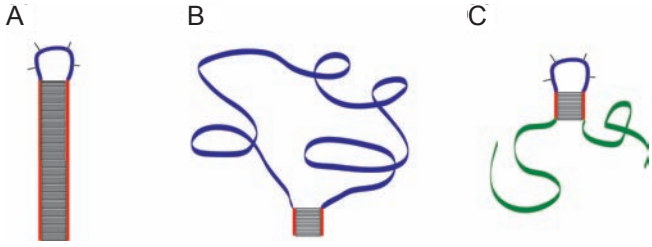


**Figure 2. Examples of configurations for an RNA molecule forming a hairpin loop. At ambient temperatures, the structures shown would correspond to the following sequences: (A) $A_{100}C_4U_{100}$, (B) $A_{15}C_{174}U_{15}$ and (C) $C_{85}A_{15}C_4U_{15}C_{85}$. At the melting temperature ($T_m$), none of these structures would be stable and these conformations would represent structures of low statistical weight in thermodynamic equilibrium with the coil structure (consisting of many random configurations of much higher statistical weight). During melting of structure (A), structures (B) and (C) would represent *partially unfolded* structures. In principle, the structures could be constrained to any of these configurations at $T_m$ and the response measured, if so desired. Free strand in the loop regions is indicated by the blue, and free strand outside the loop (C), is indicated by the green.**

Equation (5) contains both large entropy and enthalpy terms, but they exactly cancel each other at $T_m$ such that $\Delta G^{AU}_{bp}(T_m) = \Delta H^{AU}_{bp} - T_m \Delta S^{AU}_{bp} \equiv 0$. This permits us to isolate the entropy loss (due to folding) of $A_{100}C_4U_{100}$ from the contributions due to base stacking.

At face value, Figure 2A through 2C all represent possible but short lived and highly improbable conformations of $A_{100}C_4U_{100}$ at $T_m$. Their improbability is a function of the entropy loss (due to folding). Nevertheless, though naturally improbable, in principle, one could apply external forces (clamps) that constrain these structures and would expect to observe a measurable response to these constraints. Since entropy measures the direction the system should go, the expected entropy loss should be such that $-T_m\Delta S_{(\text{Figure 2A})} > -T_m\Delta S_{(\text{Figure 2B})}$ because there is more configurational order and restriction in Figure 2A, where *the reader should notice the very long tail* of the dsRNA helix that is not present in Figures 2B or 2C.

For AU bps, the LP-model data must be recalibrated to $T_m$ by a weight $T_m/T_{37}=359/310 \approx 1.16$, where $T_{37}$ is the temperature at 37°C. Since there are several versions of the hairpin loop penalties from different sources that differ substantially including a recent parameter set,[45] we simply adopt the values quoted in Table 6 of Matthews *et al.*[35] without the additional corrections. In particular, we neglect the oligo(C) corrections for the following reasons: i) There is no obvious way to account for oligo(C) when comparing the same sequence in different conformations. For example, does this correction apply to loops -ACCCCU- or -AACCCCUU- and if not, what rules do we use to model this correction? ii) These corrections are only used in some implementations of the LP-model. iii) The authors of the original study expressed some reservations about this observation. Nevertheless, we acknowledge that poly(C) loop structures are rarely observed and, therefore, perhaps this is because of unfavorable loop contributions.

Adopting a different implementation will not ultimately change the conclusion. From Matthews *et al.*,[35] for $n_L>9$ nt,

$$\Delta G_L(n_L,T_m) \approx (T_m/T_{37})\{6.4+1.75k_BT_{37}\ln(n_L/9)\} \quad (7)$$

Then, using the lookup table for $n_L \leq 9$ (at $T_{37}$), for $n_L=4$ nt (at 86°C),

$$\Delta G_L(n_L=4,T_m) \approx (1.16)(5.6) \approx 6.5 \text{ kcal/mol.}$$

Likewise, using Equation (7) for $n_L=174$ nt (at 86°C),

$$\Delta G_L(n_L=174,T_m) \approx (1.16)(9.6)=11.1 \text{ kcal/mol.}$$

Yet

$$\Delta G_L(174,T_m)[=-T_m\Delta S_{(\text{Figure 2B})}] > \Delta G_L(4,T_m)[-T_m\Delta S_{(\text{Figure 2A})}]. \quad (8)$$

This result is unphysical.

Whereas it is true that all configurations are *possible*, this does not mean that Figure 2A is more probable than Figure 2B. The JS-model is predicting that the structure in Figure 2A has the same entropy loss as the structure in Figure 2C. The entire long tail in Figure 2A contributes no observable entropy loss at $T_m$ except for the LP contribution associate with the loop of length $n_L=4$ nt; even if the RNA sequence were somehow constrained to the configuration in Figure 2A by external forces.

It is pertinent here to comment that the original experiments, on which the LP-model is based, were done measuring stems of essentially identical lengths.[41,46-48] Most of these measurements were carried out on sequences of the form $\text{GGBXN}_n\text{YBCC}$, where $^{5'\text{-GGB-}3'}_{3'\text{-CCB-}5'}$ forms the stem with $\text{B}\bar{\text{B}}=\{\text{AU,UA,GC,CG,GU,UG}\}$, X and Y consist of a selected pair of non-WC bases, and $N_n$ indicates any number $3 \leq n < 8$ of unspecified bases. This approach essentially corresponds to measuring $T_m$ for the structures in Figure 2B and 2C (*i.e.*, stems of the same length). Hence, the JS-model is consistent with the experimental observations in these conditions. However, Figure 2A and 2B do not have the same stem lengths. At $T_m$, the LP-model cannot differentiate between a structure with a *stem* length of Figure 2A and that of Figure 2B (or Figure 2C) in the model conditions currently specified, because the LP-model only assumes a local base-pair free energy. The LP-model base stacking is a highly local phenomenon that can only occur when the bases are in close proximity. Granted, after separating, some stacking can occur on a single chain (particularly poly(A)[49]); however, these corrections are not accounted for in the LP-model. Moreover, corrections for any such details are independent of the stacking free energy $\Delta G^{AU}_{bp}$ and, at most, influence the Kuhn length; a measure of the stiffness of a polymer and typically short in single-stranded RNA: about 5 nt for poly(A) and 3 nt for poly(U).[27,28] The Kuhn length is not an explicit property used in evaluating the loop penalty model either. Because these uncertainties are largely *local* in character, they are not likely in most cases to overcome the *global* issues discussed here.

Therefore, applying the above specified conditions, we are permitted (by the definitions) to examine the behavior of the loop penalty model in isolation from base stacking rules without claims to anything other than what is *explicitly* in these equations.

More important is the direction. The difference in the free energy is

$$\Delta G_L(4,T_m) - \Delta G_L(174,T_m) = 6.5 - 11.1 = -4.6 \text{ Kcal/mol} \quad (9)$$

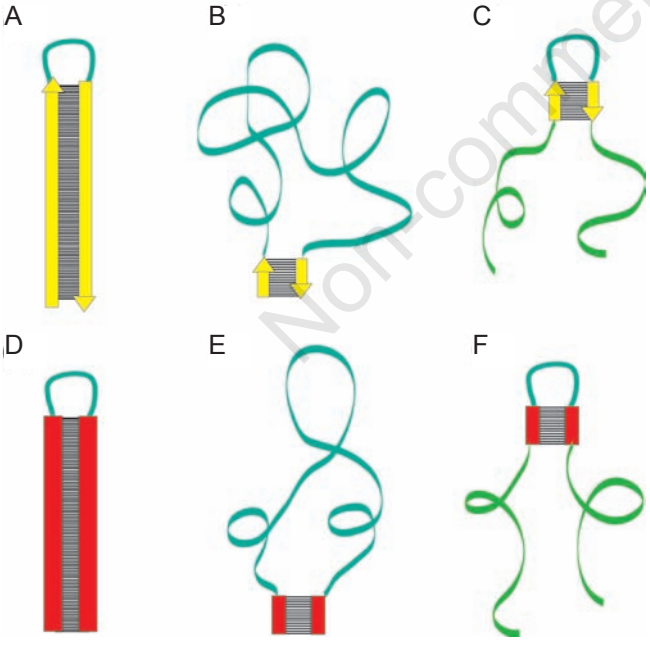


**Figure 3. Examples of Figure 2 as applied to proteins. Here, the topology of two types of protein secondary structure is shown: (A-C) various beta-strand topologies and (C-F) various alpha-helix topologies. In principle, just as in Figure 2, at $T_m$, the structures could be constrained to any of these configurations if we so desired. Without constraints, they would represent structures of low statistical weight in thermodynamic equilibrium. Free strand in the loop regions is indicated by the blue-green, and free strand outside the loop regions in (C) and (F), is indicated by the green.**

and because this is negative, the direction is toward Figure 2A and away from Figure 2B even though the latter should have less structural order. In thermodynamics, all pathways are equally possible. Though the thermodynamic probability is small that the RNA should fold first from the 5' and 3' ends (as in Figure 2B) and work backwards until an entire stem is formed in the shape of Figure 2A, this pathway is *possible*. If we work basepair-by-basepair to $n_L$=4 or in chunks $n_L$=174→164 …→4, the difference in the free energy [as in Equation (9)] is always negative. In other words, although both are configurations of low statistical weight, the LP-model predicts that we should encounter Figure 2A for $A_{100}C_4U_{100}$ far more frequently than Figure 2B and the system will move away from Figure 2B *toward* Figure 2A. Furthermore, the only criteria for deciding the entropy loss is the size of the H-loop, so both Figures 2A and 2C have the same entropy loss.

### The loop penalty-model for proteins

Protein structure predictions using the LP-model use a similar expression as Equation (1).[8,9]

In proteins, the same fundamental principles for loops closed by parallel β-strands (Figures 3A-C) or α-helices (Figure 3D-F) also apply.[50] For proteins, the equation describing the binding of the analogous double-stranded RNA structure would be applied to β-keratin for long parallel wound β-strands and to α-keratin for long wound α-helices.[51] The attractive interactions of α- and β-keratin will be more complex in the temperature dependence due to a complex interplay of hydrophobic and electrostatic interactions. Nevertheless, such sequences tend to be repetitive enough that an expression like Equation (5) could be constructed.

For a folded structure as in Figure 3A, long β-sheets are also seen in such structures as beta-barrels and these can certainly reach lengths of 20 amino acids. We can therefore propose to borrow a fragment of such a structure as a tangible example for this discussion. For α-helices, helix-helix interactions in loops like that in cytochrome C are in principle constructible and could join as shown in Figure 3D.[52-55] Hence, we can propose that there exists an amino acid sequence that satisfies the formation of a turn in which the β-stands bind into a β-sheet (Figure 3A).[52,53,56,57] Similarly, we can construct a turn that permits two alpha-helices to associate in a way similar to the beta strands (Figure 3D), for example, in the structure of cytochrome C.

Like in the case of RNA, for proteins, the same rules must apply and the entropy-loss must be evaluated in the same way. We can propose some temperature $T_m$ where the attractive interactions between the β-strand (Figure 3A) or α-helices (Figure 3D) become too weak to hold the structure together and the neighboring β-strands (or α-helices) can dissociate. It would likely require a rather highly contrived sequence to generate a structure with a single value $T_m$. Nevertheless, given one indulged in the effort to find such structures, surely at least one such protein sequence exists that generates a structure that conforms to Figure 3A (for β-strand) and similarly to Figure 3D (for α-helices). In turn, the same type of conditions used for the RNA (*i.e.*, that $\Delta G_{bp}^{AU}(T_m) = \Delta H_{bp}^{AU} - T_m \Delta S_{bp}^{AU} \equiv 0$) can be applied to dissociate the two chains in a well designed protein structure. Further, because the structures are in thermodynamic equilibrium, Figure 3B and 3C for β-strands also exist and, likewise, Figure 3E and 3F also exist for α-helices. At $T_m$, Figure 3A and 3D should be the most improbable of the examples.

For β-strands, we are, in principle, permitted to construct some device to constrain the β-strands to Figure 3A or Figure 3C and measure the response. If the LP-model is true, the energy required to constrain the structure in a form resembling Figure 3A is equal to that required to constrain the structure in a form resembling Figure 3C. Likewise, for α-helices, the constraining energy should be the same for Figure 3D and 3F. Because this is measured at $T_m$, the entropy for forming the secondary structure (the β-strands or α-helices) is zero.

## Deducing a perpetual motion machine from the model (*Deducens machinam in perpetuum moventem ex exemplari*)

The purpose of this section is to show that the unphysical outcomes that were found in limiting cases are not always inconsequential. Such behavior can directly affect the quality of prediction even for standard problems where the LP-model usually behaves properly.

In the previous Section, we found that neglecting the global contribution to the base-pairing entropy results in an unphysical response. Since the phenomenon does not employ any stacking whatsoever, this property should also work for non-WC sequences such as poly(A). Although extended structures are rare in nature, poly(A) can bind through interaction of Hoogsteen pairing along the back side of the nucleic acid base and, at low enough temperatures, should even be observable from its optical properties.[58,59] This would permit us to remove all temperature considerations from the calculations and run this experiment independent of temperature from above freezing to where the sample degrades.

This unphysical tendency (*were it true*) could be used to violate the second law of thermodynamics. We biotinylate the 5' and 3'-ends of the sequence $A_{204}$ (poly(A)) and force the 5' and 3'-ends toward each other using a direct manipulation approach like optical tweezers on the system.[29] Using a squeezing strategy to manipulate the 5' and 3' ends, Equation (9) predicts that

$$\Delta G_L((n_L=4),T) - \Delta G_L((n_L>>4),T) < 0$$

Therefore, the structure can (at the very least) spontaneously collapse into the highly ordered stem-loop structure in Figure 2A if we do work greater than $\Delta G_L(4,T_{37})$=5.6 kcal/mol (at 37°C) on the system, where $n_L$=4 is the minimum allowed loop size and $\Delta G_L(4,T_{37})$ is the approximate FE penalty (at 37°C). It is spontaneous because the difference in the FE is negative: $\Delta G_L(4,T_{37})$ - $\Delta G_L(174,T_{37})$=5.6 -9.6=-4.0 kcal/mol, where the LP-model predicts that the structure with the long tail in Figure 2A has less configurational entropy loss than Figure 2B. Whether we work with standard Watson-Crick pairs at melting temperatures or non-WC pairing where we can ignore melting temperatures, the LP-model sees Figure 2A as a case where we need only supply an external force sufficient to produce Figure 2C, and entropy will drive the structure to Figure 1A. Since there is no particular binding from the base pairs for poly(A), now we simply release our applied force, and the structure expands to its equilibrium structure doing $\Delta G_{L(174)}(T_{37}) \approx 9.6$ kcal/mol of work. This is because i) we biotinylated the 5' and 3' ends so the structure will do work against that and ii) the *difference* between the entropy-loss of the structure in Figure 2B [- $A_{JS}$ -$k_B\gamma$ln(174)] and the denatured structure (0, set by definition) is then $\Delta G_{L(174)}$. These were the definitions. This means we gained 4.0 kcal/mol of excess work that we can extract to do additional useful work.

Although we have constructed a rather artificial test of the LP-model, which was not intended in the original conception, this in not simply a matter of esoteric and abstruse curiosity. Because this tendency is built into the thermodynamics of the LP-model, it is easy to find examples of

predictions using the LP-model that have features of Figure 2A: *i.e.*, the straight stem structures are significantly over predicted.

Figures 4A-D show examples of 5S ribosomal RNA (5S rRNA) with a similar sequence homology obtained from a BLAST search and aligned using ClustalW2 (Figure 5). Three of the four structures predicted by the LP-model (using the Vienna Package 1.4 implementation) show a reasonable portion of the Y-shaped feature of 5S rRNA; however, the structure in Figure 4D shows a tendency toward Figure 2A. Results from *vsfold5* (an implementation of the CLE model) are shown on the right hand side. This is quite general. For example, it is not so infrequent to encounter predictions like Figure 2A in fitting tRNA structures using the LP-model (*e.g.*, Figure 7 of Dawson *et al.*).[25]

Figure 4E shows a fit of the HIV-1 virus in the first 833 nt of the sequence. The LP-model finds only a few of the well-established HIV-1 structures when asked to fit this very long sequence and the fit closes the structure in straight stems near the 5' and 3' ends quashing the Tar region and other known features (again resembling features of Figure 2A). *Vsfold5* finds many of the well known structural features of HIV-1,[60,61] even though the fit is *artificially constrained* to only one Kuhn length for the entire sequence. A second fit of the most poorly matched regions is shown at the bottom of Figure 4E and captures most of the critical features of the DIS and PBS regions of the sequence. The experimentally obtained reactivity data using $Me_2SO_4$ (A(N1),C(N3), red circles) is also indicated in this second fit.

It is certainly recognized that the LP-model has been instrumental in finding important structures in HIV-1 and other important RNA molecules. The point of these examples in Figure 4 is to show that the tendency to predict structures like Figure 2A is not all that uncommon with the LP-model, even for short sequences (Figure 4D), and the effect works very much against the LP-model as the sequence length gets longer (Figure 4E). The predictions in the LP-model are also more unstable, because the sequences in Figure 4A-D all have a fair degree of sequence homology (at least 91%, see alignments in Figure 5), yet the structures can vary all the way from the expected 5S rRNA structure (Figure 4A-C) to a structure closer to Figure 2A (Figure 4D).

Using *vsfold5*, a trial fit yields a Y shape using $\xi$ =10 nt for the structures in Figure 4A-C or $\xi$ =12 nt in Figure 4D, where the option *-xi_min 3* is used to increase the resolution of the stem regions. In all the results, the right-hand branch was exceptionally difficult to fit correctly. The fit was achieved by increasing the scanning distance option to *-cc_dist 9* so that the large (highly symmetric) loop region was treated as a partially interacting stem. Invoking the $Mg^{2+}$ option was also successful for the structures in Figure 4A and 4B. These results show that the large loop in the right branch is stabilized by weak binding between the chains and that stem-like structure extends within the large loop region. This can be seen in the actual 3D structure of 5S rRNA (*e.g.*, the protein data bank structure 1C2X). Although not a regular helix, the RNA chains run parallel. The chain-chain interaction is likely supported by the fair number of purine bases in this loop region, particularly since the $Mg^{2+}$ option was successful in some cases. Therefore, whereas the LP-model tends to do better on this right hand branch (when it does work), the real merit of the CLE model is that it can provide information about the stiffness and the long-range ordering within this branch and it provides this information consistently. If the CLE model costs some effort in thinking, it also can reward the user with some insights about the RNA under study.

For *vsfold5*, most of the failures in Figure 4 are a consequence of the implementation: inadequate stem analysis methods and the rigid evaluation using a single Kuhn length. In particular, when the least accurate region of HIV-1 is refit with a smaller Kuhn length in Figure 4E, the expected structure is recovered.[61] In some instances, the CLE model is

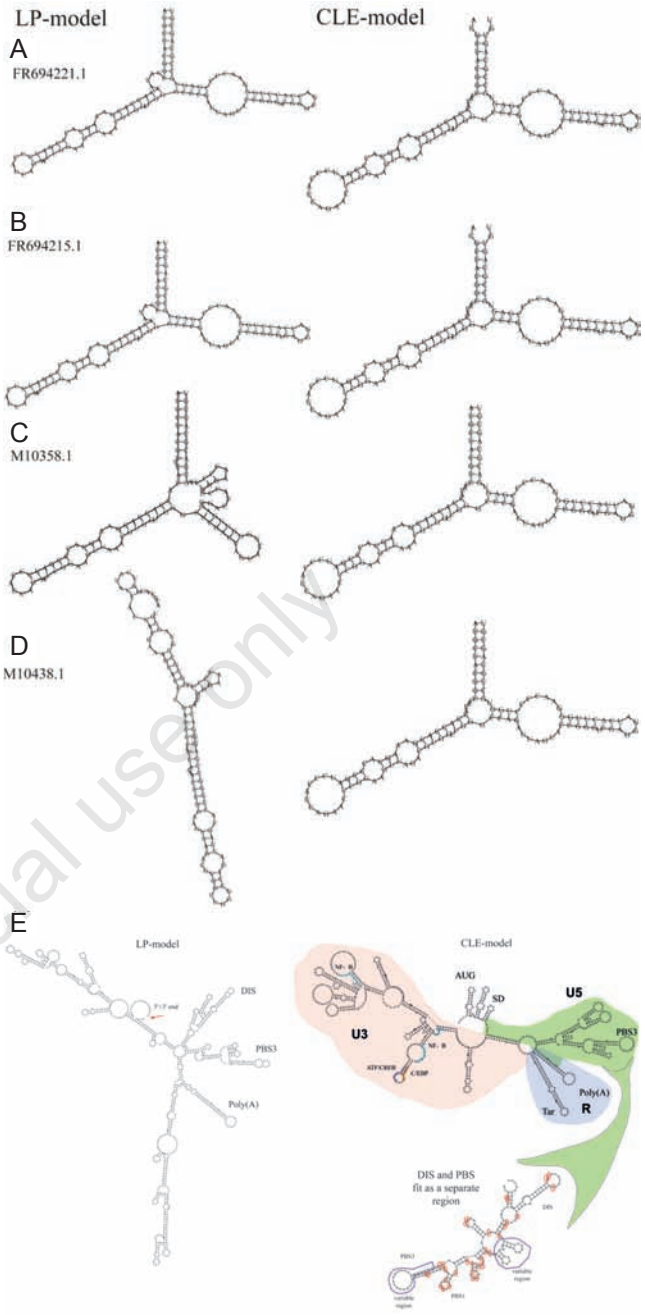


**Figure 4. A comparison of structures predicted using the LP-model (left) and the CLE (right). The CLE results are calculated using the *vsfold5* implementation. From (A) to (D), four homologous structures of 5S rRNA with at least 91% homology are shown fitted side by side. Neither the LP- nor the CLE-implementation predicts these structures perfectly; however, *vsfold5* predicts the Y shape for all four structures, and the LP-model predicts at least one structure corresponding to Figure 2A. E) An example of a much longer subsequence of HIV-1 (positions 1 to 833). With this longer sequence, the LP-model is more prone to predict a structure corresponding largely to Figure 2A. For this whole sequence, one Kuhn length was used with *vsfold5*. Many of the known features of HIV-1 can be found in this prediction.[60,61] Refitting the PBS/DIS region with a smaller Kuhn length (bottom) yields most of the correct structure of this region. This indicates that the PBS/DIS region is more flexible than the Tar/Poly(A) region. The red markings indicate experimentally observed probing with $Me_2SO_4$ and the corresponding points of reactivity (A(N1), C(N3)), as reported in Reference 61. Additional markings show the general structure of the whole 5' long terminal repeat region and interaction regions with enhancers and core factors: from http://www.hiv.lanl.gov/content/index .**

also hampered by inadequate or nonexistent experimental information.

In the Sections that follow, we will show further merits of the CLE model in that it offers a more complete thermodynamic analysis method that is applicable to molecular tweezers experiments. In Part III of this Series, we will show some aspects of the folding landscape in the CLE model, its ability to analyze two state molecular switching devices known as riboswitches, and how to estimate the maximum domain size of RNA quantitatively.

## Comparing the loop penalty- and cross linking entropy- models: the Carnot engine

Here, we apply the CLE model to general problems of heat engines. We show its behavior in a Carnot cycle and test the model in the context of recent molecular tweezers experiments.

### The Pressure *vs* Volume-diagram for the ideal gas

To put this in a context that is probably more familiar to readers, we first review the equations for the ideal gas.

The ideal gas equation of state has the form

$$p_{int}V=nRT \tag{10}$$

where $p_{int}$ is the *internal* pressure of a gas inside a vessel, $V$ is the corresponding volume, $n$ is the number of moles, $R$ (1.9872 cal/molK) is the universal gas constant, and $T$ is the temperature. Because the units often used in these problems involve kcal/mol, the values of $k_B$ and $R$ turn out to be the same. In other common units, $k_B$ =1.3806488×10$^{-23}$ [J/K] and the gas constant $R$=8.314462 [J/molK], from NIST (http://physics.nist.gov/cuu/Constants/index.html).

When the gas expands against an external pressure, it does positive work if the resulting volume is greater than the initial. Presumably, heat has entered the system to help the gas expand against the external pressure. This is referred to as work done by the system (the gas inside the vessel) and the differential change in the work done is $dW=-p_{int}dV$.[62] In experimental setups, it is often more convenient to measure the external pressure ($p_{ext}$); sometimes referred to as the work done *on* the system. In such cases, the direction of sign is opposite to the direction that the gas (in the system) expands and therefore the work done *by* the system should be expressed $dW=-p_{int}dV$. Although there is merit to expressing the problem from the perspective of how the experiment is done, here we want the frame of reference to be the system itself, not the means of probing it. Readers who are more accustomed to the alternative way of writing these equations should read $(p_{int})=(-p_{ext})$ for pressure and, $W=-\int p_{ext}dV=\int p_{int}dV$ for work.

The heat flow is $dq=dU+dW$, where $dU$ is the change in internal energy. The heat flow is defined as positive when heat flows into the system (*i.e.*, the gas inside the vessel expands) and negative when it flows out (the gas inside contracts). For a reversible thermodynamic process, the heat flow can be expressed in terms of the entropy, $TdS=dU+dW$. In general, the properties of an ideal gas should satisfy the conditions of reversible processes. Therefore, the heat flow for the ideal gas becomes

$$TdS=dU+p_{int}dV. \tag{11}$$

The work done by an ideal gas (*i.e.*, by the system) during isothermal expansion or contraction ($dT$=0) is

$$\Delta W=\int p_{int}dV=\int_{V_1}^{V_2}\frac{nRT}{V}dV=nRT\ln(V_2/V_1). \tag{12}$$

During an adiabatic expansion or contraction process of the ideal gas ($TdS$=0), we have

$$0=dU+dW=nc_v dT+\frac{nRTdV}{V} \tag{13}$$

where $c_v$ is the specific heat at constant volume. Integrating and rearranging,

$$c_v\ln(T_2/T_1)=-R\ln(V_2/V_1) \tag{14}$$

which yields the familiar expression

$$T_2(V_2)^{\eta-1}=T_1(V_1)^{\eta-1} \tag{15}$$

where $\eta=c_p/c_v$. For the ideal gas, Equation (10) can be derived from the Helmholtz free energy

$$p_{int}=-\left(\frac{\partial A}{\partial V}\right)_T \quad \text{or} \quad p_{ext}=\left(\frac{\partial A}{\partial V}\right)_T \tag{16}$$

### Thermodynamic equations for an ideal polymer

The equations for a polymer carry a similar form with the transformation $p_{int}\rightarrow f_{int}$ and $V\rightarrow r$, where r is the distance that separates the ends of the polymer chain and $f_{int}$ is the force (intrinsic to the system) acting at the ends. Traditionally, this expression is written in terms of the means of measurement ($f_{ext}(r_{ext})$). Presently, our interest is in understanding the polymer itself, so we explicitly write $f_{int}(r)$. Readers accustomed to the traditional form should read $f_{int}(r)=(-f_{int}(r))$. Unlike a gas where the volume of the vessel can be fairly accurately known, an experiment involving molecular tweezers has multiple interfaces between the device and the actual system being measured (*e.g.*, the beads and the lever arms).[29,63-65] Nevertheless, we currently want the frame of reference to be the system itself, not the means of probing it. Therefore, $r$ will be understood as $r_{int}$, even though the measured parameter is actually $r_{ext}$ and corrections are inevitably required to interpret $r_{int}$ based upon measurements obtained from $r_{ext}$.

Like the ideal gas, the equation of state for the GPC can be obtainable from the Helmholtz free energy,

$$f_{int}=-\left(\frac{\partial A}{\partial r}\right)_T \tag{17}$$

```
FR694221.1
ATGCTACGGTCATACCACCACGAAAGCACCCGATCCCATCAGAACTCGGAAGTTAAACGT 60
FR694215.1
ATGCTACGGTCATACCACCACGAAAGCACCCGATCCCATCAGAACTCGGAAGTTAAACGT 60
M10358.1
ATGCTACGTTCATACCACCACGAAAGCACCCGATCCCATCAGAACTCGGAAGTTAAACGT 60
M10438.1
ATGCTACGGTCATACCACCACGAAAGCACCCGATCCCATCAGAACTCGGAAGTTAGACGT 60
************************************************** ****

FR694221.1
GGTTGGGCTCGATTAGTACTGGGATGGGTGACCACCTGGGAACCCCGAGTGCCGTAGTGT 120
FR694215.1
GGTTGGGCTCGATTAGTACTGGGTTGAGGGATCACCTGGGAACCCCGAGTGCCGTAGTGT 120
M10358.1
GGTTGGGCTCGACTAGTACTGGGTTGAGGGATTACCTGGGAACCCCGAGTGACGTAGTGT 120
M10438.1
GGTTGGGCCAGATTAGTACTGGGTTGAGGGATCACTTGGGAACCCCTGGTGCTGTAGTGT 120
********  ** ********** ** * **  ** ***********  ***  *******
```

**Figure 5. Sequence alignment for 5S rRNA structures shown in Figure 4A-D. Sequences were initially obtained through a Blast search, aligned using ClustalW for DNA sequences and pruned of sequences that were either redundant or incomplete.**

and for the ideal polymer, the equation of state is

$$f_{\text{int}}(r,T) = 2k_{\text{B}}T\left(\frac{\gamma}{r} - \frac{(\gamma+1/2)}{\xi N b^2}r\right) = 2k_{\text{B}}T\left(\frac{\gamma}{r} - \alpha_N r\right) \quad (18)$$

where $\gamma$ is defined in Equation (1), $b$ is the monomer-to-monomer separation distance, $N$ is the number of monomers (mers) in the polymer sequence, $\alpha_N$ groups the prefactor terms as a function of $N$ and $\xi$ is a parameter known as the Kuhn length and is a measure of how coarse-grained the system is relative to the mer size.

The Kuhn length indicates the scale of the system, where almost always $\xi > 1$, the effective mer size becomes $\xi b$, and the number of effective mers becomes $N/\xi$. This scaling ($\xi$) means not only that the monomers become effective mers, it means that the number of bps (cross links) become effective cross links. How to implement this entropy model will be the subject of Parts II and III. Here, it need only be understood that we are dealing with effective cross links.

When $\gamma$=1, Equation (18) yields the historic Gaussian polymer chain (GPC).[30,31]

The minimum of Equation (18) is at $r=R_s=(\gamma/\alpha_N)^{1/2}$ and it represents the point where the sign changes for $f_{\text{int}}(r) : f_{\text{int}}(r)>0$ for $r<(\gamma/\alpha_N)^{1/2}$ and $f_{\text{int}}(r)<0$ for $r>(\gamma/\alpha_N)^{1/2}$. The root mean square (rms) value of $r$ is obtained by evaluating the second moment of a given probability distribution function of the polymer:

$$\int r^2 p(r)\Omega(r)dr = r_{\text{rms}}^2 = \xi N b^2$$

where $\Omega(r)$ is the weight function for a given probability density function.

Following the definitions and procedures for constructing thermodynamics potentials as outlined in Sears *et al.*,[62] Equation (11) transforms to

$$TdS = dU + f_{\text{int}}dr \quad (19)$$

where $dU=c_r dT$ is the internal energy of the polymer and $c_r$ is the specific heat at constant end-to-end separation distance of the polymer; analogous to the specific heat at constant volume for an ideal gas ($c_v$). Early experiments on the stretching of rubber[31,66,67] revealed that the internal energy of the polymer was negligible (less than 10% for stretching up to 3 times the initial length of the rubber).[66] The experiments did not test polymer compression; nevertheless, for good reason, the elastic properties of typical polymers like rubber are largely attributed to entropic effects arising from $f_{\text{int}}$.[31,66]

Using the relation $f_{\text{int}}(r,T)=T(\partial S(r,T)/\partial r)_T$, where $(\ldots)_T$ indicates evaluation at constant temperature, the corresponding expression for the work done in Equation (12) becomes

$$\Delta W = \int_{r_1}^{r_2} f_{\text{int}}(r)dr = T\left[S(r_2) - S(r_1)\right], \quad (20)$$

where $r_1$ and $r_2$ refer to the measured end-to-end separation distance in different states of the system. Similar equations to Equations (13) and (14) are obtained for adadiabatic expansion or contraction of the polymer

$$0 = c_r dT + f_{\text{int}}dr = c_r dT + T\left(\frac{\partial S}{\partial r}\right)_T dr, \quad (21)$$

and this in turn leads to

$$c_r \ln\frac{T_2}{T_1} + \left(S(r_2) - S(r_1)\right) = 0 \quad (22)$$

where $T_1$ and $r_1$ refer to one state of the system and $T_2$ and $r_2$ refer to another state.

## The Carnot cycle for a polymer with a single cross link

From Dawson *et al.*,[24,25] the cross linking entropy (CLE) equation is

$$S(r) = S_o + k_{\text{B}}\left\{\ln\left(C_{\gamma\xi N}\right) + 2\gamma\ln\left(\frac{r}{b}\right) - \zeta\frac{1}{\xi N}\left(\frac{r}{b}\right)^2\right\} \quad (23)$$

where $S_o$ is a constant, $\zeta=\gamma+1/2$ and $C_{\gamma\xi N}=2(\zeta/\xi N)^{\zeta}/\Gamma(\zeta)$. (The various parameters in Equation (23) are explained in further detail within Dawson *et al.*).[23-25] Evaluation of Equation (18) using Equation (20) with $f_{\text{int}}(r, T)=T(\partial S(r,T)/\partial r)_T$ yields Equation (23) to within a constant. The maximum entropy occurs at $S(R_s)$, where all other values are smaller.

The work done in Equation (20) during an isothermal process becomes

$$\Delta W = \int_{r_1}^{r_2} f_{\text{int}}(r)dr = k_{\text{B}}T\left\{2\gamma\ln\frac{r_2}{r_1} - \frac{\zeta}{\xi N}\left[\left(\frac{r_2}{b}\right)^2 - \left(\frac{r_1}{b}\right)^2\right]\right\}, \quad (24)$$

and the change in temperature and separation distance ($r=r_{\text{int}}$) during an adiabatic process becomes

$$c_r \ln\frac{T_2}{T_1} = k_{\text{B}}\left\{2\gamma\ln\frac{r_1}{r_2} - \frac{\zeta}{\xi N}\left[\left(\frac{r_1}{b}\right)^2 - \left(\frac{r_2}{b}\right)^2\right]\right\} \quad (25)$$

The right hand side of Equation (25) can be modified to the following form

$$c_r \ln\frac{T_2}{T_1} = 2\gamma k_{\text{B}}\ln\left(\frac{r_1}{r_2}\omega(r_1,r_2)\right) \quad (26)$$

where

$$\omega(r_1,r_2) = \exp\left(-\frac{\zeta}{2\gamma\xi N}\left[\left(\frac{r_1}{b}\right)^2 - \left(\frac{r_2}{b}\right)^2\right]\right) \quad (27)$$

and we obtain an expression resembling Equation (15)

$$T_2 \approx T_1\left(\frac{r_1}{r_2}\omega(r_1,r_2)\right)^{\eta}, \text{ with } \eta = 2\gamma k_{\text{B}}/c_r \quad (28)$$

In general, the parameters $r_1$ and $r_2$ in Equations (24) and (25) correspond to states in thermodynamic equilibrium. Assuming $r_1>r_2$ for the given polymer, this would usually be a denatured state ($r_1^2=\xi N b^2$) and a bound state $r_2>r_b$.

To gain a graphical picture of the force, extension, temperature and entropy (heat flow), the force-extension is shown in Figure 6A for two temperatures in a Carnot cycle, a cycle comprising isothermal and adiabatic processes.[62] Figure 6B shows the corresponding entropy for the force-extension in Figure 6A. Figure 6C is an entropy-temperature diagram of the processes shown in Figures 6A and 6B. Suppose we start from state a in Figure 6A, and run through the following cyclic process

- a-b: (reversible) isothermal contraction $r_a>r_b$ at constant temperature on the $T_{ab}$ isotherm. In the subsequent exposition, the variable $b$ (*italics*) denotes the mer-to-mer separation distance (a constant) and should not to be conflated with the thermodynamic state that is labeled b (*i.e.*, without italics).
- b-c: (reversible) adiabatic contraction $r_b>r_c$ with temperature change $T_{cd}>T_{ab}$.
- c-d: (reversible) isothermal expansion $r_d>r_c$ at constant temperature on the $T_{cd}$ isotherm.
- d-a: (reversible) adiabatic expansion $r_a>r_d$ with temperature change $T_{cd}>T_{ab}$.

For simplicity, we assume that any internal binding forces are all located at the two ends of the polymer chain. However, there is no loss in generality in assuming a sub-section of the polymer chain from posi-

tion $i$ and $j$, where $j>i$, $N_{ij}=j-i+1$ and $r=r_{ij}$ ($=r_{ij,\text{int}}$) describes the relative separation distance between mers $i$ and $j$.

Tracing the Carnot cycle in Figure 6, from Equation (24), the path going from state a→b on isotherm $T_{ab}$ (Figure 6B, red curve) the work done is

$$\Delta W_{a\to b} = k_B T_{ab}\left\{2\gamma \ln\frac{r_b}{r_a} - \frac{\zeta}{\xi N}\left[\left(\frac{r_b}{b}\right)^2 - \left(\frac{r_a}{b}\right)^2\right]\right\} \quad (29a)$$

where, because $r_a>r_b$ (and $f_{int}dr<0$ in this region), heat flows out of the system during this process (Figure 6C, red line). At this stage, any internal local binding forces will do negative work on the polymer chain pulling the ends together.

From Equation (25), the adiabatic transition from b→c results in a temperature change $T_{ab}\to T_{cd}$ with $T_{cd}>T_{ab}$ (Figure 6B, green curve)

$$c_r \ln\frac{T_{cd}}{T_{ab}} = k_B\left\{2\gamma \ln\frac{r_b}{r_c} - \frac{\zeta}{\xi N}\left[\left(\frac{r_b}{b}\right)^2 - \left(\frac{r_c}{b}\right)^2\right]\right\}, \quad (29b)$$

In the process, the system heats up (Figure 6C, green line). For the process a→b→c, internal binding interactions would be needed to overcome the repulsive response of the polymer chain. More work is done because $r_b>r_c$ and $f_{int}\neq 0$. For small $\Delta T=T_{cd}-T_{ab}$, this is approximately

$$\Delta W_{b\to c} = c_r \Delta T = k_B T_{ab}\left\{2\gamma \ln\left(\frac{r_b}{r_c}\right) - \frac{\zeta}{\xi N b^2}\left[r_b^2 - r_c^2\right]\right\}.$$

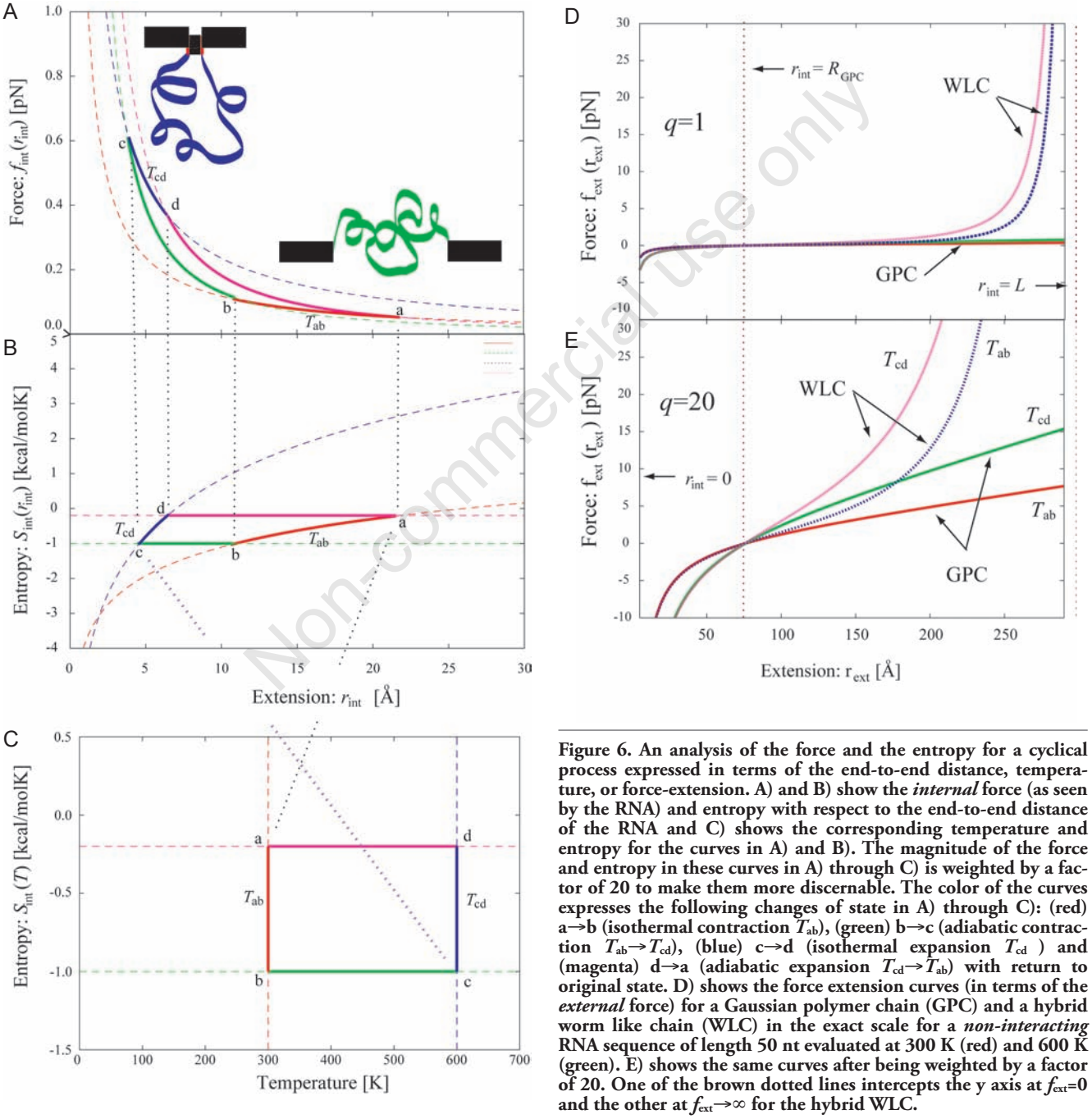


**Figure 6. An analysis of the force and the entropy for a cyclical process expressed in terms of the end-to-end distance, temperature, or force-extension. A) and B) show the *internal* force (as seen by the RNA) and entropy with respect to the end-to-end distance of the RNA and C) shows the corresponding temperature and entropy for the curves in A) and B). The magnitude of the force and entropy in these curves in A) through C) is weighted by a factor of 20 to make them more discernable. The color of the curves expresses the following changes of state in A) through C): (red) a→b (isothermal contraction $T_{ab}$), (green) b→c (adiabatic contraction $T_{ab}\to T_{cd}$), (blue) c→d (isothermal expansion $T_{cd}$ ) and (magenta) d→a (adiabatic expansion $T_{cd}\to T_{ab}$) with return to original state. D) shows the force extension curves (in terms of the *external* force) for a Gaussian polymer chain (GPC) and a hybrid worm like chain (WLC) in the exact scale for a *non-interacting* RNA sequence of length 50 nt evaluated at 300 K (red) and 600 K (green). E) shows the same curves after being weighted by a factor of 20. One of the brown dotted lines intercepts the y axis at $f_{ext}=0$ and the other at $f_{ext}\to\infty$ for the hybrid WLC.**

Going from state c→d on isotherm $T_{cd}$ (Figure 6B, blue curve), the work done is

$$\Delta W_{c\to d} = k_B T_{cd}\left\{2\gamma \ln\frac{r_d}{r_c} - \frac{\xi}{\xi N}\left[\left(\frac{r_d}{b}\right)^2 - \left(\frac{r_c}{b}\right)^2\right]\right\}, \quad (29c)$$

where, because $r_d > r_c$, heat flows into of the system during this process (Figure 6C, blue line) and solvent or external forces must do work on the system to pull the ends of the polymer chain apart.

Finally, going from state d→a in the adiabatic transition from $T_{cd} \to T_{ab}$ (Figure 6B, magenta curve) is

$$c_r \ln\frac{T_{ab}}{T_{cd}} = k_B\left\{2\gamma \ln\frac{r_d}{r_a} - \frac{\xi}{\xi N}\left[\left(\frac{r_d}{b}\right)^2 - \left(\frac{r_a}{b}\right)^2\right]\right\} \quad (29d)$$

and the polymer will cool down with $T_{cd} > T_{ab}$ (Figure 6C, magenta line). For a biopolymer, this condition would be akin to the denatured state, where the external forces of the solvent cause the biomolecule to lose its native structure. An equal and opposite amount of work as Equation (29b) is done in this step with $r_d < r_a$, and for small $\Delta T (= T_{cd} - T_{ab})$

$$\Delta W_{d\to a} = c_r(-\Delta T) = k_B T_{cd}\left\{2\gamma \ln\left(\frac{r_d}{r_a}\right) - \frac{\xi}{\xi N b^2}\left[r_d^2 - r_a^2\right]\right\}.$$

The state changes depicted in Figure 6 are greatly exaggerated and the equations are scaled by a factor of 20 to make the separation between the curves more *visible* to the eye. For a weakly interacting polymer as assumed with a GPC, $r_b \approx r_c$ and $r_a \approx r_d$. As mentioned earlier, the elastic effect originates mostly from the entropy change;[31,66,67] hence, for a GPC, $c_r$ and $\Delta T = T_2 - T_1$ are rather small quantities. If this is a true Carnot engine, then the process is completely reversible: the system returns to its original configuration. Hence, polymers exhibit a rather poor efficiency of $1 - T_{ab} / T_{cd} = \Delta T / T_{cd}$ (even assuming all the frictionless, lossless, etc. conditions of a reversible process apply). This property is certainly a good thing when it involves tires, because tires don't heat up easily when the rubber meets the road.

## Cyclic process for a polymer with multiple cross links

The expressions in Equations (23) to (29a-29d) are only valid for one specific cross link. Here we show how to extend the model to multiple cross links.

Starting from Equation (18), we know that when there are no external forces acting on the ends of a polymer, $f_{int}(R_s)=0$, where $R_s$ is discussed in Appendix. Yet this is not simply limited to the experimentally observable ends-to-end distance. For any mers $i$ and $j$ ($i \neq j$ and $i<j$), there exists some function $f_{ij,int}(r_{ij})$ in which $r_{ij}$ identifies the interaction between mers $i$ and $j$

$$R_{s,ij} \propto N_{ij}^{1/2} = (j-i+1)^{1/2} \text{ and } f_{ij,int}(R_{s,ij}) = 0.$$

To see that this must be so, consider that if we were to cut this same polymer at mers $i$ and $j$ to make a new polymer of length $N_{ij}$, then we would observer that $r_{rms,ij}^2 \propto N_{ij}$ (from the radius of gyration) and therefore $R_{s,ij}^2 \propto N_{ij}$. The terminal ends of the polymer simply correspond to $i$=1 and $j=N$.[30]

As a result, the total force acting on the polymer is

$$f_{net} = \sum_k^q f_{k,int}(r_k) \quad (30)$$

where $k \in \{i_1 j_1, i_2 j_2 .... i_n j_n, ..., i_q j_q\}$ $(0 < i_n < j_n \le N$, and $k_n \Rightarrow i_n j_n)$

identifies specific correlations between particular mers $i_n$ and $j_n$ *in* the chain, which typically involve base pairing in RNA ($i_n j_n$), $f_{k,int}$ specifies the internal force for a particular $N_k (N_{k=n} \Rightarrow j_n - i_n + 1)$ and $q$ identifies the number of such correlations.

From here, it should be a small step to understand that since $f_{k,int}(r_k,T) = T(\partial S_k(r_k,T)/\partial r_k)_T$ and the entropy of an ideal polymer has no explicit temperature dependence, the global contribution to the entropy of the polymer can be estimated from the net force in Equation (30),

$$\Delta S = \frac{1}{T}\sum_k^q \int_{r_{1,k}}^{r_{2,k}} f_{k,int}(r_k)\,dr_k = \sum_k^q \left(S_k(r_{2,k}) - S_k(r_{1,k})\right) \quad (31)$$

where we emphasize again that Equations (18) and (23) depend on $N_k$ and $\xi$ and therefore, so does $f_{k,int}$ and $S_k$. Equation (31) is the basis of the CLE model.

Up to this point, the methods are, in principle, general. However, from here, our interests are to find the general features of the CLE model that can be measured by common experimental techniques such as melting, denaturing solvents, or molecular tweezers. These generalizations require that we make some important assumptions. First, we assume that enough is known about the RNA under study and the subject of interest is a single domain of RNA structure; where the closing base pair $(i, j)$ of a *single domain* is such that there are no other base pairs $(i', j')$ for which $j' < i$ or $j < i'$. Second, if there is more than one such domain in the structure, the methods are used only within the respective domains separately and it is understood that additional methods are required to express the collective response. Finally, we assume that these generalizations are applied to relatively simple domains where most of the experimental techniques have been applied in the past. In short, this development is intended more for the purpose of understanding rather than for the purpose of prediction.

The quantity $f_{net}$ is actually a statistical quantity and a scalar that weights the contributions from $f_{ij,int}$ due to $r_{ij,int}$. Further, as previously pointed out, the stable states of a system in thermodynamic equilibrium are reasonably well defined: the denatured state ($r_1^2 = r_{rms}^2 = \xi N b^2$) and the bound state ($r_2 = r_b$). Both $f_{int}$ and $r_{int}$ are one dimensional statistical quantities, not vectors. As a result, for a particular configuration of an RNA polymer, these correlations between mers $i$ and $j$ can easily be averaged over a domain of the structure

$$\bar{r}_b^2 = \frac{1}{q}\sum_k^q r_{b,k}^2 \quad (32a)$$

$$\bar{r}_{rms}^2 = \frac{1}{q}\sum_k^q r_{rms,k}^2 \quad (32b)$$

$$\bar{N} = \frac{1}{q}\sum_k^q N_k \quad (32c)$$

$$\bar{i} = \frac{1}{q}\sum_k^q i_k \text{ and } \bar{j} = \frac{1}{q}\sum_k^q j_k \quad (32d)$$

where the bar over the quantity refers to an average. The bound state ($r_b$) is the same for all base pairs $(i, j)$. Equation (32b) is also found from Equation (32c): $\bar{r}_{rms,k}^2 = \xi \bar{N} b^2$. Likewise, since $N_{ij} = j - i + 1$, Equation (32c) can be derived from Equation (32d)

$$\bar{N} = \frac{1}{q}\sum_k^q (j_k - i_k + 1)$$
$$= \frac{1}{q}\sum_k^q j_k - \frac{1}{q}\sum_k^q i_k + 1 = \bar{j} - \bar{i} + 1$$

Hence, a quadratic evaluation of $\bar{r}_{rms,k}$ assures us of a consistent single value relationship between $\bar{N}$, $\bar{r}_{rms}$, and $\bar{R}_s$. Likewise, Equations

(32a-b) can be generalized to

$$\bar{r}^2 = \frac{1}{q}\sum_k^q r_k^2 \tag{32e}$$

Now applying Equation (32a-e) to Equation (30), it follows that the statistical internal force interactions will be

$$f_{\text{net}} = \sum_k^q f_{k,\text{int}}(r_k) \approx q\bar{f}_{\text{int}}(\bar{r}) .$$

Hence,

$$\bar{f}_{\text{net}}(\bar{r}) \approx q\bar{f}_{\text{int}}(\bar{r}) \tag{33}$$

Extrapolating from Equation (31)

$$\Delta\bar{S} \approx \frac{q}{T}\int_{\bar{r}_1}^{\bar{r}_2} \bar{f}_{\text{int}}(r)dr = q\left(\bar{S}(\bar{r}_2) - \bar{S}(\bar{r}_1)\right) \tag{34}$$

Equation (34) is largely the basis for the contact order model.[68-71] A main feature of the contact order model is that protein (or RNA) folding rates ($k_{fold}$) depend on the largest domain in the native structure.[52,69-76] The folding time of the largest domain ($\tau_{fold}$) is related to Equation (34) through $k_{fold} = 1/\tau_{fold} \propto \exp(\Delta\bar{S}/k_B)$.

Hence, one can understand the observations of the contact order model as effectively measuring the parameter $\bar{N}$ for the largest domain.[25]

Transforming Equation (25) to multiple cross links in the CLE model via Equation (31) yields

$$\left(\sum_k^q c_{rk}\right)\ln\frac{T_2}{T_1} = k_B\sum_k^q\left\{2\gamma\ln\frac{r_{1k}}{r_{2k}} - \frac{\xi}{\xi N_k}\left[\left(\frac{r_{1k}}{b}\right)^2 - \left(\frac{r_{2k}}{b}\right)^2\right]\right\} \tag{35}$$

where $c_{rk}$ is the heat capacity at constant length for the k$^{th}$ cross link, $r_{1k}$ and $r_{2k}$ are the initial and final distances for the k$^{th}$ cross link (respectively). Each state has a definite measurable temperature. Moreover, Equation (25) is separable into $q$ distinct equations

$$c_{rk}\ln\frac{T_2}{T_1} = k_B\left\{2\gamma\ln\frac{r_{1k}}{r_{2k}} - \frac{\xi}{\xi N_k}\left[\left(\frac{r_{1k}}{b}\right)^2 - \left(\frac{r_{2k}}{b}\right)^2\right]\right\} \tag{36}$$

If one state $r_{1k}$ or $r_{2k}$ is known, the other can be uniquely obtained (at least in principle). Therefore, we have all the information that is necessary to evaluate this expression for some given $q$ cross links and a well specified system.

However, it is usually not practical to measure $r_{1k}$ or $r_{2k}$ or $c_{rk}$ in a system in this way because we cannot probe the system with sufficient sensitivity. Instead, following Equation (32a-e), it is more realistic to measure the average value of $r_{1k}$, $r_{2k}$ and $c_{rk}$ with respect to a given domain,

$$\bar{r}_1^2 = \frac{1}{q}\sum_k^q r_{1k}^2 \ , \ \bar{r}_2^2 = \frac{1}{q}\sum_k^q r_{2k}^2 \tag{37a}$$

$$\langle c_r\rangle = \frac{1}{q}\sum_k^q c_{rk}, \text{ or } C_r = q\langle c_r\rangle \tag{37b}$$

where $C_r$ represents the collective weight of the heat capacity. In this way, $\bar{r}_1$ and $\bar{r}_2$ can be understood as the *observed* rms distance between effective mers $\bar{i}$ and $\bar{j}$ located at the midpoint of a domain of RNA structure. Likewise, $\bar{N}$ finds an effective midpoint in the domain: $\bar{r}_{\text{rms}}^2$. Again, this is what the contact order effectively measures.[22,25,69,76]

It follows from Equation (34) that the work done on this system is approximately

$$\begin{aligned}\Delta W &= qk_BT\left\{2\gamma\ln\frac{\bar{r}_2}{\bar{r}_1} - \frac{\xi}{\xi\bar{N}}\left[\left(\frac{\bar{r}_2}{b}\right)^2 - \left(\frac{\bar{r}_1}{b}\right)^2\right]\right\} \\ &= q\int_{\bar{r}_1}^{\bar{r}_2} f_{\text{int}}(r,\xi,\bar{N})dr\end{aligned} \tag{38}$$

The approximate total work done by this RNA to transition from a denatured state to a folded state involves the following substitution: $\bar{r}_1 = \bar{r}_{\text{rms}}$ (the denatured state for a domain) and $\bar{r}_2 = r_b$ (the bound or native state). Then, according to the contact order model, the domain with the largest $\bar{r}_{\text{rms}}$ will take the longest time to fold (of course, with some predictable provisos about various experimental conditions, etc.).

For an adiabatic processes

$$C_r\ln\frac{T_2}{T_1} = qk_B\left\{2\gamma\ln\frac{\bar{r}_1}{\bar{r}_2} - \frac{\xi}{\xi\bar{N}}\left[\left(\frac{\bar{r}_1}{b}\right)^2 - \left(\frac{\bar{r}_2}{b}\right)^2\right]\right\} \tag{39}$$

This also means that Equation (39) can be written in a form similar to Equation (28),

$$T_2 \approx T_1\left(\frac{\bar{r}_1}{\bar{r}_2}\omega(\bar{r}_1,\bar{r}_2)\right)^\eta \text{ , with } \eta = \frac{2\gamma k_B}{\langle c_r\rangle} \tag{40}$$

Therefore, the CLE model is entirely consistent with the basic thermodynamics that are expected of an idealized system. When the system as a whole is averaged over a domain, it takes on a similar form to Equations (24) and (25), but equations are now weighted by a factor of $q$.

For special conditions, Equation (32a-e) permits some approximations. If the main interest is in binding, the following *rough approximation* is sufficient

$$\bar{f}_{\text{net,int}}(\bar{r}) \approx q\bar{f}_{\text{int}}(\bar{r}) \approx 2qk_BT/\bar{r} \ , \ \bar{f}_{\text{net,int}}(\bar{r}) \approx q\bar{f}_{\text{int}}(\bar{r}) \approx 2qk_BT/\bar{r} \tag{41}$$

Given this approximation is acceptable, using Equation (40) the force can be solved for directly in adiabatic transitions,

$$f_{\text{net,int}}(\bar{r}_2) = q\bar{f}_{\text{int}}(\bar{r}_1)\left(\frac{\bar{r}_1}{\bar{r}_2}\right)^{1+\eta} \text{ with } \eta = \frac{2\gamma k_B}{\langle c_r\rangle} \tag{42}$$

where we have used the definition of $f_{\text{int}}(r)$ in Equation (33) where one effective cross link is weighted by $q$ contact points. Again, the magnitude of the observed external force, measuring the response by the system, will appear to be $q$ times stronger than if that force were acting as $q$=1. For the usual forces involved in folding and refolding a polymer, the approximations in Equations (41) and (42) are sufficient because the contributions from the stretching part of the force are not so large.

In general, if more precise values were needed, for Gamma-function derivative based-equations like the GPC, it is possible to rearrange $f_{\text{int}}(r)$ such that it is expressed in terms of $r$,

$$r = -\frac{f_{\text{int}}}{4\alpha_N k_BT} + \left(\left(\frac{f_{\text{int}}}{4\alpha_N k_BT}\right)^2 + \frac{\gamma}{\alpha_N}\right)^{\frac{1}{2}} .$$

Inspection shows that $r\to 0$ for large positive values of $f_{\text{int}}$ and $r\to\infty$ for large negative values of $f_{\text{int}}$.

To end this Section, it is pertinent to point out that the merit in the CLE model is that we do not have to restrict the definition of $\bar{r}$ to large objects like domains. Values of $\bar{r}_{\bar{i}\bar{j}}$ corresponding to mers $\bar{i}$ and $\bar{j}$ can be defined in terms of effective mers of Kuhn length $\xi$ that interact at the midpoint of a given stem. The domain is then subdivided into these

separate stems, and the entropy evaluated as in Equation (31) for different stems. In such cases, in previous work,[23-25,28] we have identified such objects with a tilde notation:

$$\tilde{i}_k = \frac{1}{q_k}\sum_n^{q_k} i_n \,,\; \tilde{i}_k = \frac{1}{q_k}\sum_n^{q_k} i_n \,,\; \tilde{N}_k = \tilde{j}_k - \tilde{i}_k + 1 \tag{43a}$$

and

$$\tilde{r}^2_{\tilde{i}\tilde{j},k} = \frac{1}{q_k}\sum_n^{q_k} r_n^2 \tag{43b}$$

where $k$ and $q_k$ reference a particular group of base pairs that collectively form a stem, and $\tilde{i}_k$, $\tilde{j}_k$ and $\tilde{r}_{\tilde{i}\tilde{j},k}$ specify the index and distance relationship between the effective mers that form the stem. From the standpoint of computation, such a strategy permits a fairly quantitative estimate of the contribution of this global entropy component to the overall free energy.

## Application to experiments with multiple cross links

At this point, it is instructive to return to Figure 6, where the equations were weighted by a factor of 20. The justification for this was actually based upon Equations (33) and (34). Up to this point, we have emphasized $f_{\text{int}}$ because we were interested in the RNA polymer itself. However, it is practical to modify the approach to be more amenable to experiment. In this regard, the equations are all inverted in Figure 6A-C with respect to the laboratory frame of reference. From here, explicit references to $r_{\text{ext}}$ and $f_{\text{ext}}$ will be made. If no explicit reference to external coordinates is used, then it is assumed that $r$=$r_{\text{int}}$.

Although we have focused on Gaussian functions (actually Gamma functions) in a considerable part our study, we hardly endorse this function as one reflecting the full behavior of a polymer. In particular, whereas studies using solvents in RNA folding focus on the *squishing aspects* of biopolymers, the current studies with molecular tweezers focus on the *stretching aspects* of biopolymers. In the stretching domain, the weakness of the Gamma functions is that the force response is simply linear and the polymer can be stretched to infinite length; a feature that is obviously unphysical.

In Reference 25, we introduced the worm like chain (WLC) model in the place of the unrealistic linear stretching part of GPC, and showed that this was a permissible solution in the family of polymer generating functions. This permits us to propose a hybrid WLC (hWLC)

$$f_{\text{int}}(r) = 2k_{\text{B}}T\left\{\frac{\gamma}{r} - \frac{\gamma_w}{\xi Lb}\left[r + \frac{L}{4(1-r/L)^2} - \frac{L}{4}\right]\right\} \tag{44a}$$

where $L$=$Nb$, $\gamma_w$ is a constant that we currently define such that Equation (44a) matches the same intercept as the GPC model: $f_{\text{int}}$ ($R_{\text{s,GPC}}$)=0. For $\xi$ =5 nt and $\gamma$=1, we found $\gamma_w$=0.85 yielded a relatively close match. At present, $\gamma_w$ can be chosen somewhat freely and the value $R_{\text{s}}$ solved for numerically. However, barring experimental evidence to the contrary, the value should have a plausible relationship such as $R_{\text{s}}$=($\gamma/\alpha_N$)$^{1/2}$ and, in particular, a convenient one is $R^2_{\text{s,GPC}} = 2\xi Nb^2/3$ (Appendix).

Weighting the hWLC equation by $q$ is also permitted, and, transforming to the experimental frame of reference where Equation (44a) is most likely to be applied, we write

$$f_{\text{ext}}(\bar{r}) = -qf_{\text{int}}(\bar{r}) = -2qk_{\text{B}}T\left\{\frac{\gamma}{\bar{r}} - \frac{\gamma_w}{\xi\bar{L}b}\left[\bar{r} + \frac{\bar{L}}{4(1-\bar{r}/\bar{L})^2} - \frac{\bar{L}}{4}\right]\right\} \tag{44b}$$

where $\bar{L} = \bar{N}b$ and all other definitions follow. Note that $r_{\text{ext}}$ will usually require some form of transformation to make it compatible with $\bar{r}$, for example $\bar{r} = r_{\text{ext}} - r_{\text{ext,o}}$, where $r_{\text{ext,o}}$ is some reference point in the experimental setup.

Figures 6D and 6E show $f_{\text{ext}}$ for the GPC and the hWLC using two different weights $q$=1 (Figure 6D) and $q$=20 (Figure 6E) for a sequence of length 50 nt and Kuhn length $\xi$ =5 nt. Like Figure 6A through C, greatly exaggerated temperature differences are used: $T_{\text{ab}}$=300K and $T_{\text{cd}}$=600K. The weight $q$=1 in Figure 6D corresponds to a standard ideal polymer where there are only weak interactions between monomers in the polymer chain. An experiment using molecular tweezers on short sequences of 50 nt would yield almost no response as a function of distance ($r_{\text{ext}}$) until nearly the full extension was achieved. On the other hand, the weight $q$=20 shows a far more measureable response.

Up to this point, the CLE model has been applied to problems were the forces were internal to the molecule. This required that equations should be averaged according to Equations (32a-e) and (37a-b) to approximate the response of the RNA. In experiments using molecular tweezers, the forces are externally applied to particular points on the RNA chain.[29,65,77,78] In this respect, unlike the averaging strategies we have used up to now, response measurements from molecular tweezers should be analyzed with respect to the position at which the *external* force is applied. Therefore, it is important to express interactions in terms of where $r_{\text{ext}}$ is actually pulling on the polymer chain with a force $f_{\text{ext}}$.

Figure 7 shows the results of an experiment that was reported in Collin *et al.*[77] In the experiment, $r_{\text{ext}}$ refers to pulling at the 5' and 3' ends of the RNA chain. The experimental data is indicated by the red + and green ×, where the red data points refer to the stretching of the RNA and the green data points reference the return trip after release of the external force. The small 44 nt sequence has additional RNA sequence attached at the 5' and 3' ends of the 44 nt stem-loop structure. This extra sequence is used to hybridize with (bound to or mount to) the corresponding complimentary strands of ssDNA; one ssDNA affixed to a surface, and the other ssDNA pulled with some optical tweezers (the general setup can be found in the Supplement of Collin *et al.*).[77] In these experiments, the 44 nt sequence forms a stem of 20 bps, and this helix rips (unzips) when the pulling force reaches a critical value causing the helix to melt (the red data points jump to the curve on the right). When this force is relaxed, the structure also relaxes; however, not on the same path as during the pulling phase (the green data points jump to the curve on the left).

Calculation of the FE (Table 1) for this sequence at the measured temperature of 25°C was found to be -37 kcal/mol using *vsfold5* (*mfold* 3.0 parameter set). We also tried the older *mfold* 2.3 parameter set and found the FE was significantly less stable (-31 kcal/mol, Table 2). In general, it is our observation that the newer parameter set is better. Using the Vienna package 1.4 implementation of the LP-model, all the FE estimates in Table 1 are significantly more favorable than the experimentally obtained FE (-37 kcal/mol). Moreover, the deviation would suggest that the older parameter set is a better fit for the LP-model. In both the CLE- and LP-models, the difference between the old and new parameter set was about 5 to 6 kcal/mol (same sign). This suggests that this difference is some feature inherent in the dinucleotide pairing potentials of the old parameter set when applied to this particular stem.

In Figure 7A, the estimated force extension for the hWLC (based on Equation 44b) is superimposed on the experimental data for a sequence length of 50 nt and 20 contacts, a Kuhn length of 5 nt and a mer-to-mer separation distance of 5.9 Å (to mesh the WLC and GPC together, we have employed the observation that the persistence length is roughly 1/2 the Kuhn length).[30] A more precise estimate for $\xi$ in the ssRNA will be discussed in Part II of this Series, however, the difference in FE between using $\xi$ =5 nt and using $\xi$ =20 nt is less than 1 kcal/mol in this case and therefore not significant in these rough approximations. In Figure 7A, Equation (44b) is shown plotted together with a pure WLC,[79] an expres-

sion that does not contain the logarithmic *squishing* contribution in Equation (44b). The native state is approximated by the vertical line at 3570 Å and the maximum pulling length for the sequence is, in principle, 295 Å upstream (about 3870 Å). Strong resistance from the RNA chain is apparent from about 2/3 of the maximum possible length.

Figure 7B shows Equation (44b) for $q$=1 and $q$=20. The response shown for Equation (44b) cannot not be made to fit the experimental data for $q$=1. Figure 7C shows the GPC fit for $q$=20 along with the hWLC. As expected, the GPC can fit data for $r_{int} \leq R_S$; however, it cannot fit the data at large extensions.

In the experimental set up, the additional leader ssRNA sequence (that is added on each side of the 44 nt sequence under study) is mounted onto the corresponding complimentary ssDNA chain forming a hybrid ssRNA/ssDNA chain (hdsRD) that resembles dsDNA. One ssDNA chain is clamped down and the other chain is pulled by the optical tweezers.[77] The hdsRD is also a polymer that behaves like a single strand when pulled at its ends. In Figures 7A-C, the red data points on the left hand side correspond to the response from the hdsRD chain when partially stretched out. Along the left hand curve, when the force acting on the hdsRD chains is large enough, the folded ssRNA suddenly rips and the hdsRD chain shifts to the right by approximately 200 Å. At that point, pulling on the hdsRD chain again resumes and becomes the primary force response. Upon release, the hdsRD begins to relax until the ssRNA can refold, at which point it suddenly returns with a *zipping* (refolding) action. The relative positions of the hdsRD are shown in Figure 7 using the dotted brown construction lines overlaying the experimental data. The second curve overlaying the experimental data (after the ripping occurs) is right shifted 200 Å from the initial position: 2/3 the maximum extension of the ssRNA sequence. The two curves essentially lie on top of the data indicating that the hdsRD resumes further stretching with only minor contributions from the 2/3 stretched out ssRNA.

The hdsRD chains are fit to an assumed sequence length of approximately 1000 bps. A reasonable fit of the data yields $\xi$ =5 nt and $q$=2. In unstressed equilibrium experiments, the typical Kuhn length of dsDNA or dsRNA is quite long; about 50 to 200 bps.[80-83] Pulling experiments on ssDNA and ssRNA tend to be very flat until near maximum extension,[84-86] characteristic of a fairly long Kuhn length. However, in Figure 7, the rise of this hdsRD polymer is more characteristic of ssRNA or ssDNA. Perhaps there is a considerable amount of ssRNA in the leader sequence or perhaps pulling on the hdsRD chains is causing micro-fractures or uniform stretching along the double strand pulling axis. It seems possible that there may be torsion effects that are visible in the hdsRD response in such rapid pulling experiments.

The experiment is done under non-equilibrium conditions and Collin *et al.*[77] interprets the results using the Jarzynski equality.[87] It is known that these experiments will show hysteresis when they are done in non-equilibrium conditions.[66,67] The merit of the CLE model is that it can be *used* in these non-equilibrium conditions to analyze the experimental data.

We observe three aspects where the CLE model is consistent with the experimental data. First, the calculated weight of the response in the model ($f_{ext}(r)$) is a factor of 20 heavier than a simple worm like chain, $q$=1. A chain with 1/20 the response would not fit the experimental data (Figure 7B). This suggests that the response is due to the weight of 20 cross links, consistent with the CLE model predictions in Equations (33) and (34). Second, with a good parameter set, *vsfold5* could predict the structure and FE accurately to within experimental error and even the poorer parameter set could be used. None of the RNA parameter sets match the experiment with the LP-model (Table 1). A DNA parameter set was used in Collin *et al.*[77] to calculate the FE at 25°C; DNA typically

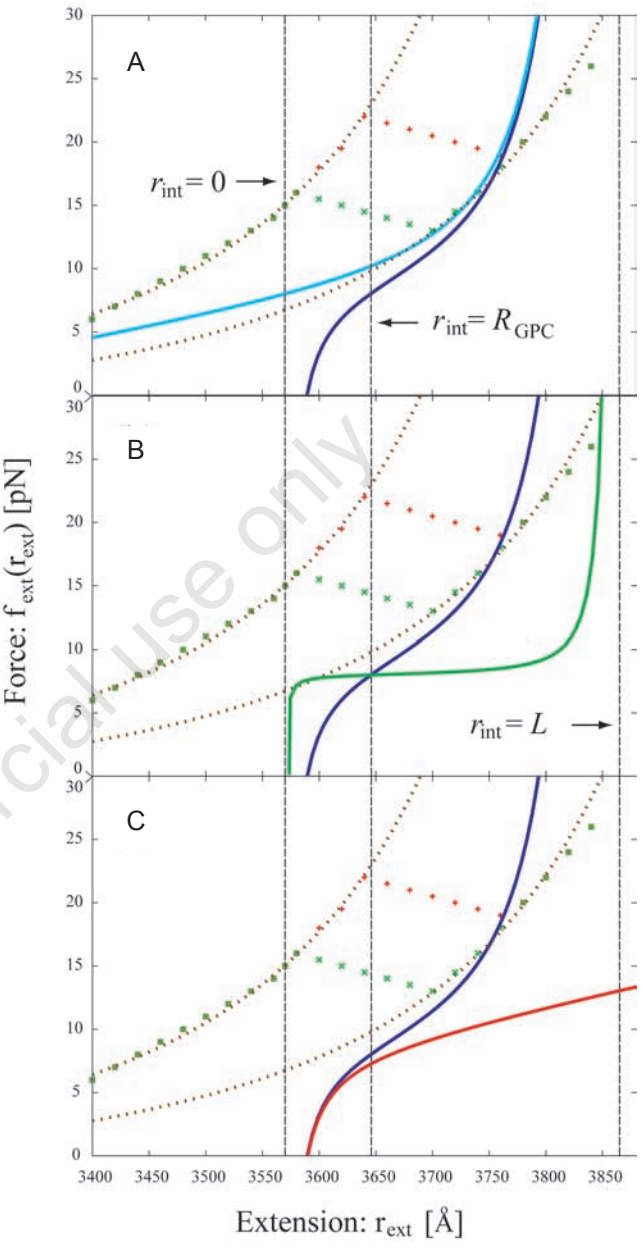


**Figure 7. An analysis of the experimental data reported in Collin *et al.*,[77] using the CLE model. Here, a hybrid worm like chain (WLC) equation is used in A) through C), blue curve. In A), a pure WLC (cyan) and the hybrid-WLC are shown together, both weighted by a factor of 20 ($q$=20). In B), one hybrid-WLC has a weight $q$=1 and the other $q$=20 . In C), the hybrid-WLC is shown along with the Gaussian polymer chain, both weighted by $q$=20. The brown dotted curves are used to fit the experimental data before and after the ssRNA rips (unzips) or zips. The red '+' indicates the ssRNA during the pulling phase of the experiment, and the green '×' indicates the path after release. The brown hatched curve on the left is for dsDNA chain stretching for the applied force-extension before the RNA hairpin breaks and the corresponding curve on the right is the same stretching after the RNA hairpin breaks and the whole complex shifts 200 Å, where all other parameters are the same in both curves.**

shows stacking energies that are about 30% smaller than RNA. Third, the whole experiment including the hdsRD connectors could be analyzed quantitatively and interpreted using the CLE model, not just the factors influencing the jumps in the experimental data due to the sudden state change of the ssRNA sample. Therefore, the interpretation is consistent with the experimental observations and the CLE model permits a straightforward picture of the events that unfolded (or *refolded*).

In another experiment reported in Liphardt *et al.*,[88] various fragments of the P5abc structure in the group I intron were studied. In Tables 2-3, results from *vsfold5* and the LP-model are shown row by row using the old and new RNA parameters for the three structures. An alignment of the structures is shown in Table 3.[88] The predicted structures by both models are identical. Energy evaluation using the LP-model (using the Vienna Package 1.4 implementation) shows large variation in prediction depending on the parameter set used. In this case, *vsfold5* was not particularly sensitive to the bp parameter set. In general, *vsfold5* is less sensitive to small changes in parameterization. Again, the CLE model could analyze the experimental data successfully.

In this Section, we have evaluated experimental results using the CLE model and observed that the forces in the pulling experiments can be approximated by a single contact weight that is proportional to the number of cross links. The CLE model is versatile enough to be used to quantitatively describe the dynamics of RNA and hybrid ssDNA/RNA linkers that pulled on the RNA, it is able to make the same correct predictions of the structures as the LP-model, it was able to correctly predict the FE at 25°C with a good bp parameter set, and it is at least as stable, if not more stable, in its FE calculations. Moreover, the CLE model is informative about the flexibility of these structures because it measures quantities like the Kuhn length, whereas the LP-model has no such concept. Finally, because the theory is adaptable, we are in a position to measure quantities like $C_r = q\langle c_r \rangle$ and $C_f = q\langle c_f \rangle$ though we have not done so in the current discussion. Therefore, there is considerable merit to using the CLE model in analyzing RNA and other biopolymers.

## Where did the misunderstanding creep in?

It should be clear by now that the primary object that causes entropy loss in the global perspective of the CLE model is the *stem*, a major source of long range structural order in a coarse-grained model. On the other hand, the primary objects in the JS-model are loops and, with the exception of an isolated hairpin loop closed by a small stem, the accounting is topologically local. In dsDNA (and likewise dsRNA), a topologically local model is valid because the formation of base-pairs is local and the JS-model has been used successfully to model *defects* in the double helix (*i.e.*, mismatches in the DNA/RNA sequence)[11,12,14,15,89] and continues to be used to some extent to model *bubbles*.[16-18,90] The CLE model would also approach the problem in a similar way because the



**Table 1. Free energy calculations at 25°C of the structure reported in Collin *et al.*[77] using CLE model (*vsfold5*) and LP-model (*RNAfold*) for the *mfold* 3.0 parameter set (new) and the *mfold* 2.3 set (old). The experimentally measured value is listed in the right most column.**

| Model implementation | New parameters [kcal/mol] | Old parameters [kcal/mol] | Experimental values [kcal/mol] |
|---|---|---|---|
| CLE (*vsfold5*) | -37.22 | -31.38 | -37 |
| LP (*RNAfold*) | -46.87 | -41.70 | -37 |

**Table 2. Free energy calculations at 25°C or 28°C of the structures reported in Liphardt *et al.*[88] using CLE model (*vsfold5*) and LP-model (*RNAfold*) for the *mfold* 3.0 parameter set (new) and the *mfold* 2.3 set (old). The experimentally measured values are in the right most columns.**

| Sequence | Calc method | Parameter set New [kJ/mol] | Parameter set Old [kJ/mol] | Measured No Mg2+ [kJ/mol] | Measured With Mg2+ [kJ/mol] |
|---|---|---|---|---|---|
| P5ab | *vsfold5* | -139.9 | -146.5 | -144 +/- 20 | -157 +/- 20 |
| | LP-model | -175.4 | -145.6 | | |
| P5abc_dA | *vsfold5* | -170.1 | -168.8 | -144 +/- 20 | -169 +/- 27 |
| | LP-model | -194.6 | -173.6 | | |
| P5abc | *vsfold5* | -149.2 | -145.0 | -140 | ----- |
| | LP-model | -169.2 | -149.8 | | |

Temp used with parameter set: (New) 25°C and (Old) 28°C.

**Table 3. An alignment of the sequences listed in Table 2.**

| Sequence | Alignment |
|---|---|
| p5ab: | acagccguucaguaccaagucucaggggaaacuuugagaugg----------ggu-----gcugacggaca |
| p5abc_dA: | acagccguucaguaccaagucucaggggaaacuuugagauggccuugcaaaggguauggu-----gcugacggaca |
| p5abc: | acagccguucaguaccaagucucaggggaaacuuugagauggccuugcaaaggguaugguaauaagcugacggaca |

major interactions of concern are largely local in character. However, the entropy is not local when a single strand chain is folded into RNA/DNA secondary structure. The main focus of this non-local entropy is not the regions of disorder (the loops), but the regions of order (the *stems*).

The main misconception is the assumption that a topologically local entropy model for these loops is sufficient: that modeling dsRNA (or dsDNA) and folded ssRNA (or ssDNA) is essentially identical except for capping the double strand with a loop to close one end (or several ends) to make it apparently folded ssRNA (or ssDNA) and that long range entropy originates in the loops and is local to the loops. The base pairing interactions are only topologically local for independent polymer chains like dsRNA/dsDNA. Early versions of the RNA secondary structure prediction approach for ssRNA focused on solving the secondary structure of tRNA.[2,91] The main features of typical tRNA molecules are three hairpin loops of similar length and the acceptor stem that closes the structure. This means the only complex feature to challenge well selected tRNA sequences is the acceptor stem that closes the multi-branch loop. Hence, researchers assumed that failures in the predictions were caused by poor parameterizations, which was partly true because refined thermodynamic tables did obtain better results,[32,35,92] as was visible in the previous Section (Table 1). The large entropy change caused by base pairing often contributes far more overall entropy loss in terms of order of magnitude. So the focus was sound. Nevertheless, lurking in the shadows, the model tends to fall into problems when limiting cases are used to test it, as the third and fourth Sections showed. It was for this reason that the JS-model tends to create problems, particularly when very long sequences were tested (Figure 4E). As will be shown in the next two subsections (and further developed in Part III), it is in long sequences that the global CLE begins to contribute in visible ways in predictions that neglect the effect.

Protein models generally focused on the hairpin and only recently have considered interior loop issues.[8,9,93] Although these limitations appear to be less problematical in proteins (perhaps due to the types of folding topologies found in most protein structures), these issues also apply to proteins.

## How does the CLE model overcome the issues previously listed

The details on how to derive the entropy loss due to the formation of structure using the CLE model are discussed in detail in Dawson *et al.*[23,25,27] In words, the entropy loss is measured by considering the difference in the entropy for the unfolded polymer chain and the folded structure.

From Dawson *et al.*,[25-27] the simplest form of the free energy expression for a Gaussian polymer chain with a fixed Kuhn length ($\xi$) is

$$\Delta G = -T\Delta S = \frac{k_B T}{\xi}\sum_{\{ij\}}\left\{\gamma \ln(\Psi N_{ij}) - (\gamma + 1/2)\left(1 - 1/(\Psi N_{ij})\right)\right\} \qquad (45)$$

where $\{ij\}$ is the set of cross links i and j that comprise a specified structure (for example, in Figures 1 through 4), $\gamma$ is often set to 1.75 to correct for the fact that real polymer chains are self-avoiding, $\Psi=\xi/\lambda^2$ and $\lambda^2$ represents the ratio of the cross link distance between the mers (measured as coarse-grained beads on a chain) and the mer-to-mer separation distance ($b$). Here, $N_{ij}=|j-i|+1$ with $j>i$. It is assumed that $\xi > 1$ mer. Weighting Equation (45) by $1/\xi$ reflects a renormalization/scaling of the cross links to form effective cross links. This permits a simple implementation of the CLE. It should not be difficult to see that application of Equations (34) and (43a-b) to Equation (45) for a fixed $\xi$ (effective cross link size) yields essentially equal expressions.

Essentially, Figure 2 is describing a domain of structure (as first introduced in Dawson *et al.*).[27,28] More specifically, Figure 4E in the CLE results shows the Tar, Poly(A), SD, $\psi$ and AUG regions as separate domains from the PBS and DIS regions. The CLE model was originally developed to find these closed off domains on the basis of the weight contributed from the formation of base pairs. The CLE model tends to discard solutions like the LP-model found (where the major domain of structure is closed off not so far away from the 5' and 3' ends) because the weight of such an entropy loss is far too large. For the examples in Figure 2, because the region is long and contiguous, it was shown in Dawson *et al.*[25,27] that this entropy grows as $\max\{N\}1n(\max\{N\})$, where $\max\{N\}=\max\{j-i+1\}$ for a specified domain. This was seen to influence the domain size of the RNA structure. Regardless of the order in which we add the entropy loss due to cross link formation in Equation (45), the entropy will consistently increase in a non-linear fashion.

The model can also be generalized. We will show in Part II of this Series that the stem length tends to be proportional to the Kuhn length. In essence, the unit of measure is the Kuhn length in these coarse-grained calculations. Figure 8 shows all the possible pathways of stem formation for an example of a simple RNA molecule consisting of three stems, two interior loops and one closing hairpin loop. According to Equation (45), regardless of the order in which the stems form, and even if some of the stems come apart and recombine later in thermodynamic equilibrium, the source of entropy loss will depend on the stems present (the source of order in a coarse-grained model) and the distance $N_{ij}$ for each bp ($ij$). There is no plausible stratagem or expedient that could gain any advantage by changing the order or manipulating the structure with various levers, as we previously observed. Based upon the approximations of the CLE model in Equations (43a-b), the stems should be evaluated in terms of the midpoint of the stems that are formed rather than at the ends.

There is also the local entropy that will be discussed in detail in Part II of this Series. The local entropy results from local restriction on the motion of the polymer and has the range of a Kuhn length. The local entropy is independent of location in the structure (topologically local) and can be a large value that is a function of the Kuhn length. In the case of base pairing, the coupling between the chains due to stacking adds further entropy costs. We are not talking about the local entropy in Part I; we are talking about the global entropy contribution due to stem formation of folded single-stranded RNA (and extrapolating the general observations to DNA, proteins, etc.).

Therefore, with the CLE model, since the cost of stem formation is always increasing and unique for each new cross link that is formed, the test we devised in the previous Sections would not prove to be a productive Maxwell daemon even in principle. Somewhat ironically, we are confronted with the paradox that the entropy (*i.e.*, disorder) is actually a major determinant in the order of biopolymers. The flip side is that this entropy (disorder) grants us a lot of mechanical action that a biopolymer needs to do useful work as a genuine molecular machine, not just to pose as a pretty picture on the page of a journal.

### Correcting the Jacobson Stockmayer -model

To finish this monograph, it is perhaps instructive to consider how the JS-model could be used to derive an expression similar to the CLE model. A more rigorous derivation of the CLE model can be found in Dawson *et al.*[25,27,28]

From the Appendix, we can infer that JS assumes a volume $v_s$ is occupied by a *cross link* segment of the polymer, where the two ends of the

chain are closed up in a loop. Indeed, JS originally addressed a problem in which the loop that formed was a polymerized ring where all information about the closing monomers is lost in the symmetry of the ring. For RNA, that $v_s$ should be understood as the *stem region* and evaluated from the mid-point of that *stem*. This would cover the volume change for a single cross link ($i, j$), where the cross link is a *stem*. Let us suppose this volume is associated with the binding of the longest segment of the polymer chain (1, $N$). Then the region around position 1 and $N$ is occupying a volume $v_s$ as proposed by JS. However, let us now suppose that there is another *stem* at ($i', j'$)=($i', N'$) that also binds. Since 1<$i'$ and $N'$<$N$, this cross link is also occupying a volume $v_s$ independent of (1, $N$), where we assume the volume of two chains forming a cross link is the same for both cases (*i.e.*, the same Kuhn length). This means that the change in entropy will be the sum of the independently formed cross links

$$\Delta S = \Delta S(1,N) + \Delta S(i',N')$$
$$= -k_B\left\{\left(A_{JS} + \gamma \ln(N-1+1)\right) + \left(A_{JS} + \gamma \ln(N'-i'+1)\right)\right\}$$

Extrapolating to a set of cross links, with $N_{ij} = |j - i|+1$ and $j>i$, we find

$$\Delta G = -T\Delta S = k_B T \sum_{\{ij\}} \left(A_{ij} + \gamma \ln(N_{ij})\right) \quad (46)$$

where $A_{ij}=A_{JS}$ for a fixed Kuhn length and the constant $A_{JS}$ incorporates various scaling corrections for the Kuhn length which will be the subject of Part II. The form of the expression is similar to that proposed in the CLE model if the stretching term is neglected, as was done in the example in Equation (41). It can also be seen in Figures 6D-E and Figure 7C that for $r<R_S$, the dominant feature is the logarithmic contribution. Hence, we have shown that the JS-model could also be used to anticipate the CLE model and the reason why it is used in the way described in Equation (1) is because of some misunderstandings about the derivation of the single cross link in the original paper.[1]

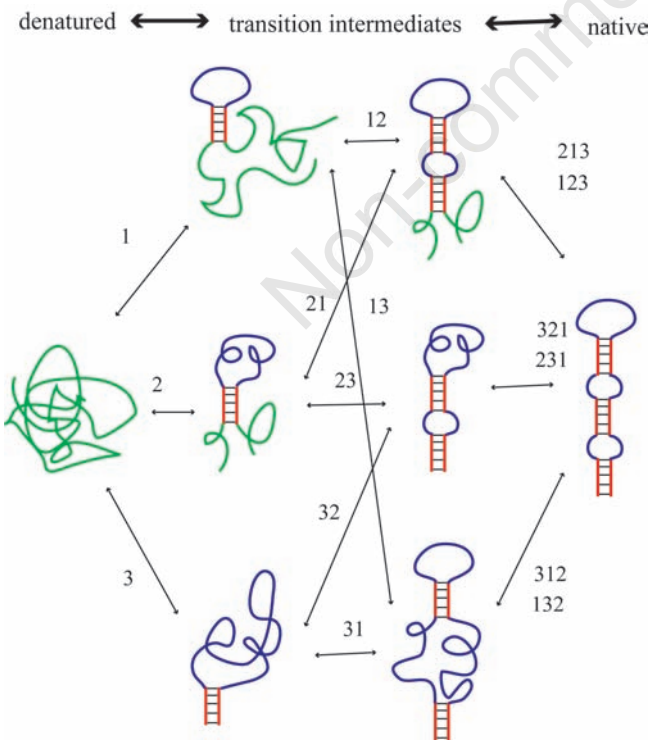


**Figure 8. An example of different folding pathways for a simple RNA structure with three stems, two interior loops, and one hairpin loop from the denatured state to the native state.**

Nevertheless, even these corrections are somewhat inadequate without some consideration about the flexibility of the RNA in the conditions involved. For example, in Figure 9, the two structures need to be compared with the same (or similar) stem flexibility. In Figure 9A, it seems relatively intuitive to assess that the single stem might be associated with a single Kuhn length and that the large H-loop region might have another, Figure 9B). The notation for the lattice constant and $b_{ds}$ is because it is quite common for authors to describe the contour length in terms of the rise in the stack for the dsDNA/dsRNA, not the effective bond distance between mers in the RNA/DNA (*e.g.*, Forties *et al.*).[16] Based on Figure 9B, we might reason that the long tail in Figure 9C has a similar Kuhn length as Figure 9B. This would mean that the long tail in Figure 9C is divided into six groups of stems (Figure 9D), each weighted by Equation (46). In this case, $A_{ij}=A_{JS}$ would be a reasonable approximation for both Figure 9B and D. However, it is also possible that the stem now becomes much stiffer with a new Kuhn length that is also the length of the new stem, Figure 9E. If this be the case, then we must correct $A_{JS}$ to reflect the fact that the Kuhn length has changed in the two Figures. To some extent, this should also be considered for the loop regions of Figures 9B, D, and 9E, which are different. Finally, the Kuhn length is finite and a typical maximum is perhaps 200 bps.[94] Therefore, a very long sequence of dsRNA (*e.g.*, 200 kbps) capped at one end by a loop is not likely to be found, even if the Kuhn length does reach the $\xi$ =200 nt maximum in the double strand region of the folded ssRNA because the multitude of *clamps* (*i.e.*, Figure 9D) that would be required is not strong enough to hold the domain closed.

Estimating $A_{JS}$ is important because, in both Tables 1 and 2, part of the reason the LP-model does poorly is not just because it neglects the global entropy, but because it does not properly estimate $A_{JS}$. For example, if only the global entropy is considered, Figure 9E has a smaller global weight than Figure 9D. However, it will be shown, in Part II of the series, that after the local entropy ( $A_{JS}$) is evaluated for the stem in Figure 9E, the total entropy (local *plus* global) of Figure 9D,E are nearly the same magnitude. Hence, it is still important to go further into the details of how a proper value for can be found and to understand this problem in terms of both global and local entropy issues.

In part II of this Series, we will show that a theoretical expression for $A_{JS}$ can be derived from the CLE model and that this (currently) empirical constant can be derived from first principles. We will also consider the issue of a variable Kuhn length and one that changes with the formation of different types of stems and loops. In Part III, we will return to the combined role of global and local issues.

## Conclusions

In this work, we have shown that the thermodynamic model that is commonly used to predict RNA structure and protein structure in some cases has a flaw that can lead to unphysical predictions. For the system that the model was originally developed for, the parameters were tuned to render a sensible result. It is only when the model is extrapolated to more complex cases that issues may arise. We have shown that the CLE model is an alternative that is more general and helps overcome these issues. Moreover, the CLE model easily expands into a fully adaptable

thermodynamic model allowing a complete description of the long range entropy for any configuration of a polymer. Such flexibility would prove useful in studying the dynamics of RNA, both in the secondary structure and in 3D structure calculations. The thermodynamic equations of the CLE model are applicable to pulling experiments with molecular tweezers with simple modifications and the predictions can be quantitative. Finally, a crude version of the CLE model can be derived from the Jacobson-Stockmayer model based on a clearer understanding of the model itself. Inasmuch as the dynamics of real polymers can be approximated by a coarse-grained model such as the CLE model, we have shown that the approach is consistent and can model the dynamics of biopolymers.

Experiments are influenced by the understanding of the theory of the time; therefore, future work on these subjects should include at least some of the following: i) new experiments need to be done to refine the statistical model itself. We introduced the hybrid worm like chain model. However, some parameters could only be estimated. The fundamentals should be explored experimentally; ii) the CLE model appears to satisfy the thermodynamics, but, just like the ideal gas model, it is likely to have these properties only over a limited range. In particular, very little is currently addressed on the temperature dependence of these entropy equations. It is unlikely that they are pure linear functions of temperature. RNA hybridization should be studied in context dependent environments such as the case where helices are packed side-by-side as opposed to solvent exposed helices. We need to understand the range of applicability for the current equation of state.

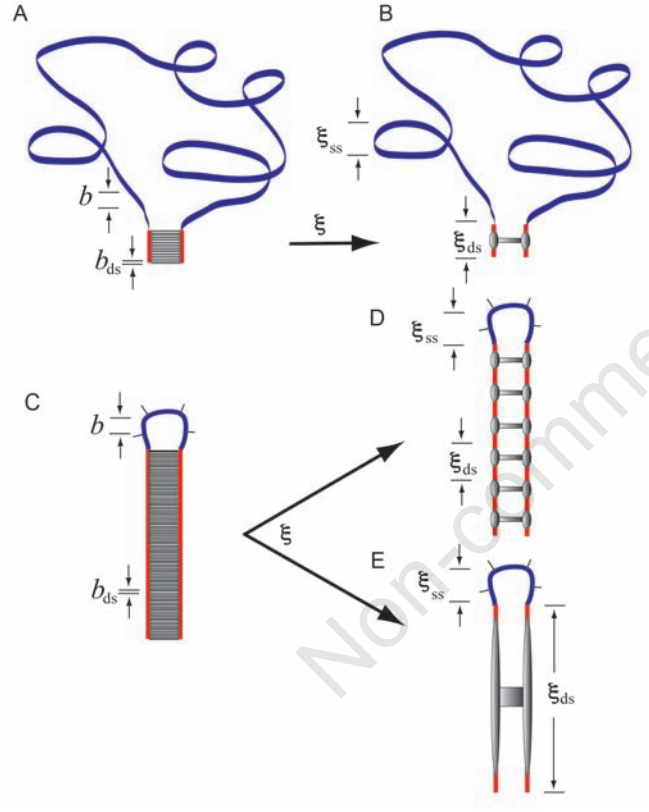


**Figure 9. A cartoon showing qualitatively how the Jacobson-Stockmayer equation needs to be transformed to make it able to overcome the issues expressed in the text and in Figure 2A,B. A) shows the base pair interactions in Figure 2B and B) shows the corresponding effective cross link for the structure. The evaluation of the JS-equation is associated with a stem of length 15 bps and, therefore, the Kuhn length $\xi_{ds}$ should also be understood as approximately 15 bps long (B). The Kuhn length and the (effective) cross links (located at the center of the stem) are indicated by the metallic beads and cross bar (B). One should also consider that the Kuhn length of the single strand region ($\xi_{ds}$) is probably different and smaller than $\xi_{ds}$. (C,D and E) Evaluation of the structure in Figure 2A (Figure 9C) could change in two different directions. In the first case (D), the Kuhn length ($\xi_{ds}$) is the same as in A) and B), and the other (E) is that the Kuhn length becomes similar in length to the stem itself.**

## References


1. Jacobson H, Stockmayer W. Intramolecular reaction in polycondensations. I. the theory of linear systems. J Chem Phys 1950;18:1600-06.
2. Tinoco I, Uhlenbeck OC, Levine MD. Estimation of secondary structure in ribonucleic acids. Nature 1971;230:362-7.
3. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 1981;9:133-48.
4. Hofacker IL, Fontana W, Stadler PF, e al. Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie 1994;125:167-88.
5. Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol 1999:285:2053-68.
6. Dirks RM, Pierce NA. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. J Comput Chem 2004;25:1295-304.
7. Ren JH, Rastegara B, Condon A, Hoos HH. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. RNA 2005;11:1494-04.
8. Pace CN, Grimsley GR, Thomson JA, Barnett BJ. Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. J Biol Chem 1998;263:11820-5.
9. Alm E, Baker D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc Natl Acad Sci USA 1999;96:11305-10.
10. Crothers DM, Zimm BH. Theory of the melting transition of synthetic polynucleotides: evaluation of the stacking free energy. J Mol Biol 1964;9:1-9.
11. Zimm BH.Theory of melting of the helical form in double chains of the DNA type. J Chem Phys 1960;33:1349-56.
12. Poland D, Scheraga HA. Phase transitions in one dimension and the helix-coil transition in polyamino acids. J Chem Phys 1966;45:1456-63.
13. Poland D, Scheraga HA. Occurrence of a phase transition in nucleic acid models. J Chem Phys 1966;45:1464-9.
14. Uhlenbeck OC, Martin FH, Doty P. Self-complementary oligoribonucleotides: effects of helix defects and guanylic acid-cytidylic acid base pairs. J Mol Biol 1971;57:217-29.
15. Fink TR, Crothers DM. Free energy of imperfect nucleic acid helices: I. The bulge defect. J Mol Biol 1972;66:1-12.
16. Forties RA, Bundschuh R, Poirier MG. The flexibility of locally melted DNA. Nucleic Acids Res 2009;37:4580-86.
17. Lando DY, Fridman AS. Role of small loops in DNA melting. Biopolymers 2001;58:374-89.

18. Rahi SJ, Hertzberg MP, Kardar M. Melting of persistent double-stranded polymers. Phys Rev E Stat Nonlin Soft Matter Phys 2008;78:51910.
19. Kallenbach NR. Theory of thermal transitions in low molecular weight RNA chains. J Mol Biol 1968;37:445-66.
20. Scheffler IE, Elson EL, Baldwin RL. Helix formation by d(TA) oligomers I. Hairpin and straight-chain helices. J Mol Biol 1968;36:291-304.
21. Scheffler IE, Elson EL, Baldwin RL. Helix formation by d(TA) oligomers II. Analysis of the helix-coil trasitions of linear and circular oligomers. J Mol Biol 1970;48:145-71.
22. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 2004;5:105.
23. Dawson W, Fujiwara K, Kawai G. Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. PLoS One 2007;2:905.
24. Dawson W, Fujiwara K, Kawai G, et al. A method for finding optimal RNA secondary structures using a new entropy model (vsfold). Nucleotides, Nucleosides, and Nucleic Acids 2006;25:171-89.
25. Dawson W, Kawai G. Modeling the chain entropy of biopolymers: unifying two different random walk models under one framework. J Comput Sci Syst Biol 2009;2:1-23.
26. Dawson W, Kawai G, Yamamoto K. Modeling the long range entropy of biopolymers: A focus on protein structure prediction and folding. Recent Research Developments in Experimental & Theoretical Biology 2005;1:57-92.
27. 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.
28. 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.
29. Moffitt JR, Chemla YR, Smith SB, Bustamante C. Recent advances in optical tweezers. Annu Rev Biochem 2008;77:205-28.
30. Flory PJ. Statistical mechanics of chain molecules. New York: Wiley; 1969.
31. Flory PJ. Principles of polymer chemistry. Ithaca: Cornell University Press; 1956.
32. 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.
33. Jaeger JA, Turner DH, Zuker M. Improved predictions of secondary structures for RNA. Proc Natl Acad Sci USA 1989;86:7706-10.
34. Jaeger JA, Turner DH, Zuker M. Predicting optimal and suboptimal secondary structure for RNA. Meth in Enzymology 1990;183:281-306.
35. 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.
36. SantaLucia J, Turner DH. Measuring the thermodynamics of RNA secondary structure formation. Biopolymers 1998;44:309-19.
37. Turner DH, Sugimoto N, Freier SM. RNA structure prediction. Ann Rev Biophys Biophys Chem 1988;17:167-92.
38. Zuker M, Mathews DH, Turner DH. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski J, Clark BFC, editors. NATO ASI Series; 1999.
39. Fisher ME. Effect of excluded volume on phase transitions in biopolymers. J Chem Phys 1966;45:1469-73.
40. Hofacker IL. Vienna RNA secondary structure server. Nucleic Acids Res 2003;31:3429-31.
41. Serra MJ, Barnes TW, Betschart K, et al. Improved parameters for the prediction of RNA hairpin stability. Biochemistry 1997;36:4844-51.
42. Borer PN, Bengler B, Tinoco I. Stability of ribonucleic acid double-stranded helices. J Mol Biol 1974;86:843-53.
43. Gray DM, Tinoco I Jr. A new approach to the study of sequence-dependent properties of polynucleotides. Biopolymers 1970;9:223-44.
44. 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.
45. 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;34:4912-24.
46. Giese MR, Betschart K, Dale T, et al. Stability of RNA hairpins closed by wobble base pairs. Biochemistry 1998;37:1094-100.
47. Serra MJ, Axenson TJ, Turner DH. A model for the stabilities of RNA hairpins based on a study of the sequence dependence of stability for hairpins of six nucleotides. Biochemistry 1994;33:14289-96.
48. Vecenie CJ, Morrow CV, Zyra A, Serra MJ. Sequence dependence of the stability of RNA hairpin molecules with six nucleotide loops. Biochemistry 2006;45:1400-7.
49. Privalov PL, Filimonov VV. Thermodynamic analysis of transfer RNA unfolding. J Mol Biol 1978;122:447-64.
50. Alm E, Baker D. Matching theory and experiment in protein folding. Curr Opin Struct Biol 1999;9:189-96.
51. Lehninger AL. Biochemistry. New York: Worth Publishers, INC; 1975.
52. Hofrichter J, Thompson PA, Munoz V, et al. Dynamics of alpha-helices, beta-hairpins and loops. Kuwajima K, Arai M, editors. Amsterdam: Elsevier Science B.V.; 1990
53. Roder H, Shastry MCR, Sauder J, Park SH. Kinetic and structural characterization of early events in protein folding; Kuwajima K, Arai M, editors. Amsterdam: Elsevier Science B.V.; 1999.
54. Shastry MC, Roder H. Evidence for barrier-limited protein folding kinetics on the microsecond time scale. Nat Struct Biol 1998;5:385-92.
55. Takahashi S, Akiyama S, Ishimori K, Morishima I. CD measurements on the rarly folding intermediate of cytochrome c using the fast flow mixer; Kuwajima K, Arai M, editors. Amsterdam: Elsevier Science B.V.; 1999.
56. Schenck HL, Gellman SH. Use of a designed triple-stranded antiparallel beta-sheet to probe beta-sheet cooperativity in aqueous solution. J Am Chem Soc 1998;120:4869-70.
57. Sharman GJ, Searle MS. Cooperative interaction between the three strands of a designed antiparallel beta-sheet. J Am Chem Soc 1998; 120:5291-300.
58. Soyfer VN, Potaman VN. Triple-helical nucleic acids. New York: Springer-Verlag, Inc; 1995. p 360.
59. Burkard ME, Turner DH, Tinoco I Jr. Structure of base pairs involving at least two hydrogen bonds. Gesteland RF, Cech TR, Atkins JF, editors. Cold Springs Harbor: Cold Springs Harbor; 1999.
60. Berkhout B, van Wamel JL. The leader of the HIV-1 RNA genome forms a compactly folded tertiary structure. RNA 2000;6:282-95.
61. Paillart JC, Dettenhofer M, Yu XF, First snapshots of the HIV-1 RNA structure in infected cells and in virions. J Biol Chem 2004;279:48397-403.
62. Sears FW, Salinger GL. Thermodynamics, kinetic theory, and statis-

tical thermodynamics. Menlo Park: Addison-Wesley; 1986. p 464.
63. Green L, Kim CH, Bustamante C, Tinoco I Jr. Characterization of the mechanical unfolding of RNA pseudoknots. J Mol Biol 2008;375:511-28.
64. Li PT, Tinoco I Jr. Mechanical unfolding of two DIS RNA kissing complexes from HIV-1. J Mol Biol 2009;386:1343-56.
65. Manosas M, Wen JD, Li PT, et al. Force unfolding kinetics of RNA using optical tweezers. II. Modeling experiments. Biophys J 2007;2:3010-21.
66. Anthony BL, Caston RH, Guth E. Equations of state for natural and synthetic rubber-like materials. I Unaccelerated natural soft rubber. J Phys Chem 1942;46:826-40.
67. Elliott DR, Lippmann SA. The thermodynamics of rubber at small extensions. J App Phys 1945;16:50-4.
68. Calloni G, Taddei N, Plaxco KW, et al. Comparison of the folding processes of distantly related proteins. Importance of hydrophobic content in folding. J Mol Biol 2003;330:577-91.
69. Ivankov DN, Garbuzynskiy SO, Alm E, et al. Contact order revisited: influence of protein size on the folding rate. Prot Sci 2003;12:2057-62.
70. Makarov DE, Keller CA, Plaxco KW, Metiu H. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proc Natl Acad Sci USA 2002;99:3535-9.
71. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985-94.
72. Fernandez A. Folding a protein by discretizing its backbone torsional dynamics. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 1999;59:5928-39.
73. Michel F, Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J Mol Biol 1990;216:585-610.
74. Morgan SR, Higgs PG. Evidence for kinetic effects in the folding of large RNA molecules. J Chem Phys 1996;105:7152-7.
75. Schultes EA, Spasic A, Mohanty U, Bartel DP. Compact and ordered collapse of randomly generated RNA sequences. Nat Struct Mol Biol 2005;12:1130-6.
76. Sosnick TR, Pan T. Reduced contact order and RNA folding rates. J Mol Biol 2004;342:1359-65.
77. 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.
78. Wen JD, Manosas M, Li PT, et al. Force unfolding kinetics of RNA using optical tweezers. I. Effects of experimental variables on measured results. Biophys J 2007;92:2996-3009.
79. Marko JF, Siggia ED. Stretching DNA. Macromolecules 1995;28:8759-70.
80. Harrington RE. Opticohydrodynamic properties of high-molecular-weight DNA. III. the effects of NaCl concentration. Biopolymers 1978;17:919-36.
81. 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.
82. Rouzina I, Bloomfield VA. Force-induced melting of the DNA double helix. 2. Effect of solution conditions. Biophys J 2001;80:894-900.
83. Rouzina I, Bloomfield VA. Force-induced melting of the DNA double helix 1. Thermodynamic analysis. Biophys J 2001;80:882-93.
84. 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.
85. 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.
86. Abels JA, Moreno-Herrero F, van der Heijden T, e al. Single-molecule measurements of the persistence length of double-stranded RNA. Biophys J 2005;88:2737-44.
87. Jarzynski C. Nonequilibrium equality for free energy differences. Phys Rev Lett 1997;78:2690-3.
88. Liphardt J, Onoa B, Smith SB, et al. Reversible unfolding of single RNA molecules by mechanical force. Science 2001;292:733-7.
89. Poland DC, Scheraga HA. Statistical mechanics of noncovalent bonds in polyamino acids. VIII. Covalent loops in proteins. Biopolymers 1965;3:379-99.
90. Gonzalez R, Zeng Y, Ivanov V, Zocchi G. Bubbles in DNA melting. J Phys Condens Matter 2009;21:034102.
91. Fresco JR, Alberts BM, Doty P. Some molecular details of the secondary structure of ribonucleic acid. Nature 1960;188:98-101.
92. Salser W. Globin mRNA sequences: analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp Quant Biol 1977;42:985-1103.
93. Cheng RR, Uzawa T, Plaxco KW, Makarov DE. Universality in the timescales of internal loop formation in unfolded proteins and single-stranded oligonucleotides. Biophys J 2010;99:3959-68.
94. Hagerman PJ. Flexibility of RNA. Ann Rev Biophys Biomol Struct 1997;26:139-56.