A new entropy model for RNA: part V, Incorporating the Flory-Huggins model in structure prediction and folding

Authors
Wayne Dawson[1,*], and Gota Kawai[2]

Institutions
[1] Bioinformation Engineering Laboratory, Department of Biotechnology, Graduate School of Agriculture and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

[2] Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino-shi, Chiba 275-0016 Japan.

*Current affiliation and address
Dept of Comp Bio, Fac Frontier Sci, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken 277-8561, Japan
CBRC, AIST, Tokyo Waterfront Bio-IT Research Bld, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

Corresponding Author
Correspondence should be addressed to:
Wayne Dawson: wayne-dawson@aist.go.jp

Keywords
RNA folding; virial coefficients; Flory-Huggins model; polymer-solvent effect; excluded volume effect; RNA structure; bioinformatics

Authors' contributions

Wayne Dawson: Wrote the manuscript and did the primary research and development of *vsfold*.
Gota Kawai: Advice, guidance and support in the software development of *vsfold*, contributed to writing the manuscript.

Abstract

The effect of solvent-biopolymer interactions is hardly negligible. Whereas the ideal (non-interacting) polymer consisting of N monomers in an ideal solvent is expected to have the terminal ends of its chain with a root-mean-squared (rms) end-to-end separation distance (r_{rms}) proportional to the square root of N, real interactions of a polymer both with itself and with the solvent often tend to strongly perturb r_{rms}. In poor solvent, the biopolymer can collapse into a small globule much smaller than the ideal r_{rms} due to excluding solvent. In good solvent, the biopolymer can swell to a size much larger than the ideal r_{rms} due to favoring solvent. These effects require corrections to an ideal polymer equation. We have been developing the cross linking entropy (CLE) model in this series. The model attempts find the maximum entropy of a folded polymer by taking into account the correlation caused by bonding and other interactions of the structure. In RNA, this mostly occurs in the stems. Here we adapt CLE model to handle polymer swelling and collapse for RNA molecules both in good and in poor solvent. This work is intended to introduce this type of study and to allow its systematic application in problems of RNA folding and structure prediction. The current study suggests that there may be some tendency for RNA to behave as a polymer in poor solvent and that this collapse may happen in sequences longer than 50 nt.

1 Introduction

In Parts I through IV, various aspects of an entropy model for predicting RNA structure have been presented. The model was termed the cross linking entropy (CLE) model and the unique aspect of the model is that the entropy loss has a global component that should be evaluated at *each* base pair (or cross link)[1] where, for RNA structure, this cross link concept is attributed mainly to the base pairs in the stems. Traditional approaches such as mfold [1]and RNAfold [2] assume that the base pairs in the stems can be calculated with a topologically independent set of base pairing rules, often referred to as the Turner energy rules [3-5]. Loops are then accounted for by assigning penalties for different types of loops based on the Jacobson-Stockmayer (JS) equation [6] (a topic of part I of this series). With the exception of the hairpin loop, these penalties are also topologically local to a given loop region [7]. In contrast, in the CLE model, it is assumed that the stem is where the conformational order is found and, therefore, every base pair (cross link) pays both a *local* and a *global* entropic cost in the free energy (FE). Other unique aspects of this entropy model are the capacity to evaluate entropy loss in both biopolymer folding and stretching and incorporating the parameters and conclusions of renormalization theory [8,9], including the concept of flexibility by scaling the number of degrees of freedom of the polymer with the Kuhn length ξ [8,10] and the option to employ non-Gaussian correlation functions within RNA folding and structure prediction calculations.

Up to this point, we have focused on developing the concepts of the CLE model around the ideal polymer in our presentation. The RNA was assumed to behave like a Gaussian polymer chain (GPC); an ideal polymer where the monomers (mers) are non-interacting and the variance in the distance between the ends of a polymer chain of length N in the denatured state has a root-mean-squared (rms) end-to-end separation distance (r_{rms}) such that $r_{rms} \propto N^{1/2}$ [11]. Some non-ideal corrections have already been introduced. For example, corrections for the fact that real polymers occupy space

[1] Perhaps a better choice than "cross linking entropy" might have been "bonding" or "contact", or maybe "correlation" entropy. The reason for choosing the word "cross link" was because this entropy is not just about bonding, it is about the freezing out of degrees of freedom and its consequences in all manner of polymers. "Cross link" seemed like a non-descript word that yielded the image of bonding without necessarily forcing the concept of "bonds". Maybe, that has only served to confuse matters. In retrospect, "correlation entropy" seems the closest to the actual

and, therefore, the polymer should be described in terms of a self-avoiding random walk [12,13] (the parameter γ). Adjustments for differences in flexibility of the polymer were introduced by employing the Kuhn length (ξ) in evaluating the entropy. Finally, because it is not entirely clear what correlation function precisely represents a true polymer (particularly in large RNA structural domains), the option to change the parameter δ [8,14] from the Gaussian function $\delta = 2$ to any value $0 < \delta < 10$ was introduced [15,16] to permit the use of an exponential $\delta = 1$ correlation function (for example).

In all coarse-grained thermodynamic models, a fundamental assumption in predicting RNA structure is that the the free energy (FE) of a given folded single-stranded RNA (ssRNA) can be obtained by evaluating the difference between two thermodynamically measurable states: the denatured state, which is defined in terms of the rms end-to-end distance (r_{rms}), and the native state, which is defined in terms of the base pair separation distance in the final structure (r_b). This is true whether assumed implicitly as in the JS equation or explicitly as in the CLE model.

The native state r_b has a fixed distance between the bps, however, what can we say about r_{rms}? Real biopolymers are hardly ideal and are rarely in an ideal solvent. The specific details about the type of solvent, temperature and ionic strength become important to an accurate description of r_{rms} [17,18]. The environmental conditions are incorporated into the ideal polymer equations by including higher order corrections to the basic equation for an ideal polymer. These corrections make up what is called the virial equation. A parameter that emerges from evaluating the second and third coefficients in the virial equation is a weight (ν) that yields $r_{rms} \propto N^{\nu}$ [9,10,19], where the ideal polymer has the weight $\nu = 1/2$ and this condition represents the case where the higher order terms in the virial equation just exactly cancel each other. In good solvent, such as denaturing agents, the polymer will tend to swell in an attempt to maximize interactions with the solvent. In such cases, the biopolymer can swell to a size $r_{rms} \propto N^{3/5}$, much larger than the GPC. On the other hand, in poor solvent, the solvent has limited carrying capacity and the polymer tends to shrink up into a small globule in

---

meaning and effects of this entropy.

an attempt to minimize the solvent accessible surface area in favor of self-interaction. In such cases, the biopolymer can collapse to a size $r_{rms} \propto N^{1/3}$, much smaller than the size of a GPC [10]. As a result, the global entropic behavior of the polymer is also strongly affected by the solvent conditions because r_{rms} is not a fixed value and depends on ν.

What is known of RNA? Biopolymers characteristically have a varying range of amphiphilic behavior that complicates predicting their overall interactions. Hydrophobic parts of the RNA molecule tend to repel water away, yet at the same time, charged functional groups like amines and phosphates tend to attract charged ions and the water that carries them. The bases of nucleic acids have aromatic rings that contribute both Van der Waals interactions (a significant source of stacking enthalpy [20]) as well as some hydrophobic effects that characterize the solubility of aromatic rings [21]. On the other hand, the hydrophilic sugars on nucleic acids contain negatively charged phosphates that are affected by the ionic strength [22-24], where an atmosphere of positively charged monovalent and divalent cations are thought to form a cloud around the nucleic acids, balancing and screening the large negative charge [24-26]. A fair number of charged Mg^{2+} ions are known to surround the x-ray structures like the Tetrahymena ribozyme (*T. thermophilus*) [27,28] and tRNA [29]. The Mg^{2+}-water complex ($[Mg(H_2O)_6]^{2+}$), although poor at forming specific binding sites, is excellent at stabilizing RNA [30,31]. Thus, although these effects are non-specific, the role of the counter ions induces characteristic effects that resemble "solvent". Is it a good solvent or a poor solvent? Studies of DNA with varying mixtures of alcohol produced a B-DNA to A-DNA transition at high alcohol content [32] adding further mystery to the issue of what stacking is. Though stacking is not well understood [33], it is generally agreed that stacking is largely enthalpic [5].

Given these complexities, measuring the virial coefficients that influence this parameter ν is not a simple task. The general objective is to find Flory's theta (Θ) temperature where the ideal polymer behavior is found [19,34]. If one can find this Θ temperature for some set of conditions such as ionic strength, solvent, buffer concentration, etc., then all of these amphiphilic interactions will cancel out.

In general, it remains a significant challenge to obtain a Θ temperature [34,35], let alone the virial coefficients, the Kuhn length or the precise parameters in the

polymer equations themselves. For example, denatured proteins have been studied extensively by Tanford [35,36] with considerable attention given to finding a good denaturing solvent. Yet the existence of residual structure in proteins even in denaturing solvents [37] is a matter of discussion to this day [38]. Nevertheless, a considerable amount has been written on what a denatured protein is, and what it isn't. Far less is known of RNA. Felsenfeld's group did some very careful studies to find the Θ temperature of poly(A) [39], poly(U) [40], and in single-stranded RNA (ssRNA) sequences with the purine bases removed [41] (apurinic ssRNA). The Θ temperature for poly(A) in 1 M NaCl is about 26^{o}C and poly(U) and apurinic ssRNA at 2 M NaCl is about 18^{o}C. Hence, some aspects of high salt concentration would appear to resemble features like a denaturant. On the other hand, mild concentrations of denaturing solvent (urea) have been used to induce folding of tertiary structure in trapped intermediates of RNase P [42,43] and the Tetrahymena ribozyme [27]. Although the Turner energy rules are often measured in 1M salt [5], the cellular environments of most RNA structures are quite different with lower salt and there are a plethora of interactions with other biomolecules *in vivo*. Hence, for any realistically plausible scenario, it is likely that $\nu \neq 1/2$.

In previous work, the denatured state with $\nu = 1/2$ was treated as given. Here, we turn to the Flory-Huggins (FH) model as a way to begin to account for the true solvent environment. Then FH model is named after Flory [44] and Huggins [45] who independently discovered the phenomena of polymer swelling. Here we show how to apply the Flory-Huggins model [19,44,45] to the CLE model to calculate RNA structure; including the Kuhn length within the framework of the RNA structure prediction. We can only lay the groundwork for polymer swelling and collapse, examine a few test cases where it may apply, and leave it to the experimentalists to ferret out the rich complexities of RNA under different solvent conditions, ionic strengths and denaturing solvents. This is the first generalized application of the FH model [44,45] to the context RNA structure prediction and folding, to the best of our knowledge.

The CLE model has been the theme of the previous four parts of this series (I-IV) and has been demonstrated in the literature [15,16,46-48]. A brief introduction to the material is given in Section 2; however, for a better understanding of the details and the

full range of the theory, the reader is encouraged to review the previous parts of this series first. Although the focus of this work is oriented to RNA, it is general and equally applicable to proteins (with additional considerations [46]).

2. The CLE model and its application to RNA structure prediction

The first two parts briefly review the essential concepts: the Kuhn length (Sec 2.1) and the equations in the CLE model (Sec 2.2). Further details can be found in Parts I-IV and Refs [15,16,46-48]. The second part briefly explains how this is applied to the prediction (Sec 2.3) and folding (Sec 2.4) of secondary structure and pseudoknot predictions. Finally, we comment on how these models generally account for the heterogeneous character of RNA (Sec 2.5).

2.1. The Kuhn length

The renormalization parameter with the most visible impact on RNA structure is the Kuhn length (ξ), measured here in units of mers. The Kuhn length [10] sets the scale of resolution or coarse-grained length scale of the polymer, where this length scale provides considerable direct information about the rigidity of the polymer. A Kuhn length of exactly one would mean that each base on the RNA sequence can flex over a full 4π solid angle. This is rarely feasible for simple monomers, and, for RNA, $\xi > 1$ nt. Single-stranded RNA (ssRNA) in the free strand (fs) regions exhibits a range for ξ between 3 nt and 5 nt. Hence, the first manifestation of this phenomenon that can be inferred from ξ is the tendency for RNA to exhibit a minimum loop size of 3 nt.

As argued in Part II of this series, the Kuhn length of the stem regions is likely to be on the same order as the stem length with a small variable region at the interface between the free strand and the terminal end of the stem. Typical RNA structure exhibits stem lengths ranging from 3 to 15 bps with an average around 7 or 8 bps. The stems of tRNA would have a Kuhn length of about 4 or 5 nt, but ribosomal RNA may have some regions that are as large as 12 nt. Therefore, it is important to have some sense of the Kuhn length before applying it to the RNA structure prediction.

The resulting coarse-grained resolution scale suggests that monomers are grouped

in a collective unit we have called an “effective mer”, or perhaps an *epimer* because the characteristics of the Kuhn length resemble epiphenomena and the Greek root “epi” has the sense of “in addition”. Moreover, the Kuhn length is not rigidly fixed to particular mers in the polymer chain but manifests itself locally as a result of coupling between the mers in the polymer chain.

Because the Kuhn length involves the collective coupling of mers, it differs from the concept of flexibility that is reported in terms of B-factors. The B-factor assumes independent, harmonic motions of the atoms in a sample measured using x-ray spectroscopy [49], not the large collective motions. Nevertheless, very stiff regions of a crystal will have a small B-factor (large ξ) and highly flexible regions will have a larger B-factor (small ξ). Therefore, qualitatively, the B-factor and ξ are inversely related.

For a sequence of N mers whose mer-to-mer separation distance is b, let the maximum stretched-out contour length of the polymer chain be

$$L = Nb. \tag{1}$$

Let the polymer also be expressed as a collection of effective mers (or *epimer*) of number ($\tilde{N} = N / \xi$) whose separation distance is $\tilde{b} = \xi b$. It follows that

$$L = Nb = \tilde{N}\tilde{b}. \tag{2}$$

Since N can be defined in terms of any positive integer and we have simply chosen a sequence from mer 1 to mer N, it is also true that we can choose an arbitrary mer i and mer j ($i < j$) and define $N_{ij} = j - i + 1$. This follows because we can cut the polymer at mer i and at mer j forming a new polymer of length N_{ij} with the same relationship of $r_{rms} \propto N^{1/2}$ where $N = N_{ij}$. The root mean square (rms) separation distance between mer i and mer j (ij-rmsd) is a function of the number of effective mers in the sequence

$$r_{rms,ij} = \langle r^2 \rangle_{ij}^{1/2} = \xi^{1-\nu} N_{ij}^{\nu} b \text{, or } \; r_{rms,ij} = \kappa \left| j - i + 1 \right|^{\nu} b \text{,} \tag{3}$$

where $\kappa = \xi^{1-\nu}$ and ν is the parameter that we wish to explore in this work. Roughly speaking, the range of ν should be around $1/3 \le \nu \le 3/5$; however, there is nothing *mathematically* wrong with $0 < \nu < 1$, where $\nu = 0$ means the volume is just $(\xi b)^3$ and $\nu = 1$ means a straight linear chain.

2.2. The CLE model

Now that the ij-rmsd is defined, the next step is to define the entropy in terms of any separation distance between mers i and j (r_{ij}), which will be called the ij separation distance (ij-distance).

The general expression for the probability that one should find the separation distance between mers i and j equal to r_{ij} (now a variable) within a tolerance of Δr between r_{ij} and $r_{ij} + \Delta r$ is

$$p(r_{ij})\Delta r = A_{\delta\gamma} C_{ij}^{\gamma\delta} \left(\frac{r_{ij}}{b} \right)^{\delta\gamma} \exp\left\{ -\vartheta_{ij} \left(\frac{r_{ij}}{b} \right)^{\delta} \right\} \left(\frac{\Delta r}{b} \right) \tag{4}$$

where both δ and γ are finite positive constants. The parameter γ is a correction for the self-avoiding random walk [12,13] and the parameter δ changes the nature of the correlation from Gaussian ($\delta = 2$) to some other value such as exponential ($\delta = 1$), etc. Other remaining terms are

$$A_{\delta\gamma} = \frac{\delta \pi^{\gamma + 1/\delta}}{\Gamma(\gamma + 1/\delta)}, \tag{4a}$$

$$\vartheta_{ij} = \left(\frac{\Gamma(\gamma + 3/\delta)}{\Gamma(\gamma + 1/\delta)} \frac{b^2}{\langle r^2 \rangle_{ij}} \right)^{\delta/2} = \zeta(\gamma, \delta) \left(\frac{b^2}{r_{rms,ij}^2} \right)^{\delta/2}, \tag{4b}$$

$$\zeta(\gamma,\delta) = \left[\Gamma(\gamma+3/\delta)/\Gamma(\gamma+1/\delta)\right]^{\delta/2}, \tag{4c}$$

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

$$C_{ij}^{\gamma\delta} = \frac{\delta\vartheta_{ij}^{\gamma+1/\delta}}{A_{\delta\gamma}\Gamma(\gamma+1/\delta)} \tag{4d}$$

For $\xi > 1$, the entropy is

$$S(r_{ij}) = \frac{k_B}{\xi}\ln\left(p(r_{ij})\Delta r\right) = \frac{k_B}{\xi}\left\{\ln(A_{\delta\gamma}C_{ij}^{\gamma\delta}) + \delta\gamma\ln\left(\frac{r_{ij}}{b}\right) - \vartheta_{ij}\left(\frac{r_{ij}}{b}\right)^{\delta}\right\}. \tag{5}$$

where we have introduced renormalization scaling by ξ to account for the unit of measure being an *epimer* rather than a single mer.

For base pair formation, the following compact expression emerges after some manipulation,

$$\begin{aligned}\Delta S_{bp}(N_{ij}) &= S(\lambda b) - S\left(r_{rms,ij}\right)\\ &= -\frac{k_B}{\xi}\left\{\nu\delta\gamma\ln\left(\Psi_{\nu\xi}N_{ij}\right) - \zeta(\gamma,\delta)\left(1-1/(\Psi_{\nu\xi}N_{ij})^{\delta\nu}\right)\right\}\end{aligned} \tag{6}$$

where $\Psi_{\nu\xi} = (\xi/\lambda)^{1/\nu}/\xi$ and λ 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*). The distance between the mers in a base pair is not the hydrogen bond distance between the bases, but essentially the distance between the centers of the mers. The result is that $\lambda \approx 2$ because the distance between the mers of a bp is about twice the distance between mers on the RNA polymer chain.

The total entropy loss is the sum of the local correction (due to renormalization of the number of mers to *epimers*) and the global contribution caused by stem formation

$$\Delta S_{cle} = \Delta S_{\xi\gamma\delta} + \sum_{bp(ij)} \Delta S_{bp}(N_{ij}), \tag{7}$$

where $\Delta S_{bp}(N_{ij})$ is the global contribution given in Eqn (6) and $\Delta S_{\xi\gamma\delta}$ is the local entropy (derived in Part II of this series),

$$\Delta S_{\xi\gamma\delta} = -\left(\frac{N}{\xi}\right)\frac{k_B}{D}\int_{+1}^{\xi}\left\{\frac{(1-\nu)(\delta\gamma+1)\ln(x)}{1-x} + \varpi\zeta(\gamma,\delta)x^{\delta(2\nu-1)}\frac{1-x^{\delta(1-\nu)}}{1-x}\right\}dx \tag{8}$$

where, $\varpi = \left(\gamma + 1/\delta\right)/\zeta(\gamma,\delta)$ is a stretching weight and, in general, we assume the dimensionality is $D = 3$. Eqn (7) is summed over *all* bps of a given structure using Eqn (6). The current implementations of the CLE model are *vsfold5* and *vs_subopt*, both of which assume that the Kuhn length is a constant throughout the structure. Future implementations are aiming at using a variable Kuhn length.

In the absence of information, the standard values $\delta = 2$, $\gamma = 1.75$ and $\nu = 1/2$ are recommended because these are the implicit parameters used in other models for RNA structure prediction. However, mathematically, it is not wrong to select anything where $\delta > 0$, $\gamma > 0$ and $0 < \nu < 1$, though not necessarily physically meaningful. Since $\gamma = 1$ for random walk without corrections for self-avoiding effects, $\gamma > 1$ seems wise. In non-bonding polymers, renormalization theory suggests $\delta = 2.3$ [50] and recent measurements of DNA folding may suggest $3 < \delta < 4$ is possible [51]. One might also think that correlation will become more non-local in RNA and proteins, suggesting $1/2 \le \delta \le 2$ with self-avoiding parameter $1 < \gamma < 2.3$. At present, these remain unresolved issues. The parameter ν is the subject of the current study.

2.3. Application to structure prediction methods

The dynamic programming algorithm (DPA) and the details of how this method is applied to RNA structure prediction are explained in Part IV. A good primer on the subject can be found in Ref [52]. Here, we simply outline the general approach. A base pair between mers i and j will be indicated as (i, j).

Traditional DPA calculations:

Traditional models include programs like mfold [1] and RNAfold [2]. In traditional models, the prediction consists of choosing the best solution for the free energy at each position i and j in a triangle matrix

$$\Delta G_{ij} = \min\left\{\Delta G_{ij}^{\mathrm{bp}}, \Delta G_{ij}^{\mathrm{fs}}, \Delta G_{ij}^{\mathrm{H}}, \Delta G_{ij,pq}^{\mathrm{I}}, \Delta G_{ij,\{pq\}}^{\mathrm{M}}\right\} \tag{9}$$

where ij refers to the mers at i and j in the sequence, $\Delta G_{ij}^{\mathrm{bp}}$ is the base pairing FE at ij, $\Delta G_{ij}^{\mathrm{fs}}$ is the case where there is only the free strand (fs) interaction between ij, $\Delta G_{ij}^{\mathrm{H}}$ indicates a hairpin loop (H-loop) that closes at ij, $\Delta G_{ij,pq}^{\mathrm{I}}$ indicates an interior loop (I-loop) that closes at ij with an adjoining point at pq $(i < p < q < j)$, and $\Delta G_{ij,\{pq\}}^{\mathrm{M}}$ indicates a multibranch loop (MBL) closing at ij and forming branches at $\{pq\}$ where $p_k q_k$ satisfies $(i < p_1 < q_1 < p_2 < q_2 .... < j)$.

The base pair FE consists of a lookup table of dinucleotide bps with the corresponding FE known as the Turner energy rules [53] $\{\Delta\Delta G_{bp}^{Turner}\}$, where the $\{\cdots\}$ indicates a set of data and the subscript (bp) indicates a particular dinucleotide base pair such as $\begin{smallmatrix}\text{5'-AG-3'}\\ \text{3'-UC-5'}\end{smallmatrix}$ or $\begin{smallmatrix}\text{5'-GU-3'}\\ \text{3'-CG-5'}\end{smallmatrix}$. For a given bp at (i, j), $\Delta\Delta G_{ij}^{Turner}$ corresponds to the bases comprising (i, j) and $(i+1, j-1)$. This quantity is added to the previous bp at $(i+1, j-1)$

$$\Delta G_{ij}^{\mathrm{bp}} = \Delta\Delta G_{ij}^{Turner} + \Delta G_{i+1,j-1} \tag{10}$$

where the subscript ij only depend on the particular residues comprising (i, j) and $(i+1, j-1)$.

Both H-loops and I-loops use the Jacobson-Stockmayer (JS) equation to compute the entropy loss for formation of the loop

$$\Delta G_{JS}(n) = T(A_{JS} + \gamma k_{\mathrm{B}} \ln(n)). \tag{11}$$

where $n = j - i$ for a H-loop and $n = j - q - (p - i)$ for an I-loop. The JS equation is examined in detail in Parts I and II of this series. Briefly, A_{JS} is a constant expressing the average local entropy (Part II), γ is the same as defined in Eqns (4) and (5) and T is the temperature.

For H-loops,

$$\Delta G_{ij}^{\mathrm{H}} = \Delta\Delta G_{ij}^{C} + \Delta G_{JS}(n) \tag{12}$$

where $\Delta\Delta G_{ij}^{C}$ refers to the Turner closing base pair FE.

Similarly, for I-loops,

$$\Delta G_{ij,pq}^{\mathrm{I}} = \Delta\Delta G_{ij}^{C} + \Delta\Delta G_{pq}^{C} + \Delta G_{JS}(j - q - (p - i)). \tag{13}$$

For MBLs, an approximation for the penalty is used

$$\Delta G_{n,m}^{\mathrm{M}} = T(C_0 + C_1 \sum_{k=0}^{m} n_k + C_2 m), \tag{14}$$

where C_0, C_1, and C_2 are all fitted parameters, m is the number of branches that extend off of the MBL and $n_k = p_{k+1} - q_k - 1$ is the length of the free-strand segments of the MBL with $k = 0,...,m$, $q_0 = i$ and $p_{m+1} = j$. The loop term becomes

$$\Delta G_{ij,\{pq\}}^{\mathrm{M}} = \Delta\Delta G_{ij}^{C} + \Delta G_{n,m}^{\mathrm{M}} + \sum_{\{pq\}} \Delta G_{pq} \tag{15}$$

where ΔG_{pq} corresponds to a given branch in the MBL.

This recursive approach of the dynamic programming algorithm yields an optimal

solution because every ij in the matrix is evaluated and the answer yielded at $i=1$ and $j=N$ [54]. Hence, assuming the model is correct and the recursive method to evaluate the model is also correct, the optimal solution should be the best solution. A considerable amount of detail has been skipped in this summary and the specifics depend on the particular implementation. Nevertheless, the general concept is as described.

The CLE model in DPA calculations:

In the CLE model, the DPA looks similar to the traditional method

$$\Delta G_{ij} = \min\left\{\Delta G_{ij}^{\mathrm{bp}}, \Delta G_{ij}^{\mathrm{fs}}, \Delta G_{ij}^{\mathrm{H}}, \Delta G_{ij,pq}^{\mathrm{I}}, \Delta G_{ij,\{pq\}}^{\mathrm{M}}, \Delta G_{ij}^{\mathrm{PK}}\right\} \tag{16}$$

where $\Delta G_{ij}^{\mathrm{PK}}$ is the FE for a pseudoknot (PK) and the remaining terms resemble Eqn (9). Although apparently similar, the CLE is a very different approach.

First, the base pair FE is modified

$$\Delta G_{ij}^{\mathrm{bp}} = \Delta\Delta G_{ij}^{Turner} + \Delta G_{i+1,j-1} - T\Delta S_{bp}(N_{ij}) \tag{17}$$

where Eqn (6) is added to Eqn (10) and $N_{ij} = j - i + 1$.

Second, the various types of loops are different. For the H-loop,

$$\Delta G_{ij}^{H} = \Delta\Delta G_{ij}^{C} - T\Delta S(N_{ij}). \tag{18}$$

For the I-loop,

$$\Delta G_{ij,pq}^{I} = \Delta\Delta G_{ij}^{C} - T\Delta S(N_{ij}) + \Delta G_{pq}^{bp}. \tag{19}$$

Likewise, for the MBL,

$$\Delta G^{M}_{ij,\{pq\}} = \Delta\Delta G^{C}_{ij} - T\Delta S(N_{ij}) + \sum_{\{pq\}} \Delta G_{pq} \,. \tag{20}$$

Hence, for the secondary structure, there is one simple entropy calculation that is added to any bp or closing bp and that is all. However, because the CLE model processes the FE in terms of *stems*, these stems have to be scanned for changes in their structure, and there are a multitude of issues involved with defining a stem that go far beyond this brief introduction. There are also coaxial stacking issues in the MBLs and other structural matters with stems and loops that cannot be discussed here. Suffice it to say that over a length scale of at least ξ (within all sub-structures), additional processing is often required to take into account the stiffness and stability of the structure.

Finally, there is the term $\Delta G^{\mathrm{PK}}_{ij}$. Pseudoknots require handles to indicate what type of PK structure is involved, how these PK parts are connected to the remaining secondary structure that might extend out from the basic PK structure, and other types of information. However, whereas the PK requires a considerable amount of information processing, the FE is essentially calculated using Eqns (17) to (20). There are, of course, corrections to account for the closer proximity of the chains, an option to consider Mg^{2+} binding and information about the 3D structure must be inferred and calculated. Nevertheless, the global entropy $\Delta S(N_{ij})$ is simply entropy, and doesn't need further ado.

Hence, perhaps what is different about the CLE model is that it has only one type of global entropy evaluation for all contexts, it is straightforward to evaluate and, from Eqn (5), mainly depends on the relative distance between mers i and j and the magnitude of ξ.

As observed in Part II, the first term in Eqn (6) resembles the variable term in the Jacobson-Stockmayer (JS) equation, Eqn (11). In Eqn (6), the prefactor is $\nu\delta\gamma$. In standard implementations of RNA secondary structure prediction, it is implicitly assumed that $\nu = 1/2$, $\delta = 2$ and $\gamma = 1.75$ ($\nu\delta\gamma = 1.75$). However, recently, values ranging around $\gamma \approx 2.1$ have been reported in coarse-grained lattice calculations of bubbles (I-loops) in double-stranded DNA (dsDNA) [55,56] and in various types of

RNA loops [57]. This may reflect the case where $\nu = 3/5$, $\delta = 2$ and $\gamma = 1.75$ ($\nu\delta\gamma = 2.1$). Given this interpretation is correct, the bases in the chain (in the region around the dsDNA bubble) actually have expanded into the solvent, supporting the views proposed by Makarov and coworkers [51] and Weise and coworkers [58] that $\nu \approx 0.6$.

2.4. Application of the CLE model to RNA folding

Because Eqn (5) suggests that the global entropy mainly depends on the distance between mers i and j and the Kuhn length (ξ), and because Eqn (5) is a general expression that can be used with any distance r_{ij}, it is possible to calculate the global entropy for any configuration. The main point of this discussion is to give a good idea of how versatile the approach really is, and then it should be clear that the approach we use yields information about RNA folding.

RNA folding in general:

First, it is important to clarify some notation that applies to these problems.

A typical way to physically measure the folding of RNA in real time is by way of force-extension experiments [59,60]. The force-extension response of a polymer is typically measured with an experimental apparatus such as the optical tweezers [61,62] and the force is reported as $f_{\text{ext}}(r)$, where "ext" refers to the external force required to extend (or compress) the polymer. The distance r is also measured as r_{ext} because a measure of the internal ij-distance (r_{ij}) is difficult to obtain in a force extension measurement and r_{ij} is inferred. The ij-distance can be inferred directly from bulk measurements of the radius of gyration in small angle x-ray scattering [63]. However, no information about the force can be obtained. Here, the main interest is the response of mers i and j when r_{ij} deviates from its ideal ij-rmsd value ($r_{rms,ij}$), not the response of the experimental device used to measure the polymer. The mutual response of the mers is $f_{\text{int}}(r_{ij})$ where r_{ij} is understood to express the relative distance between mers i and j in terms of the mer frame of reference ($r_{\text{int},ij}$). Further detail can be found in Section 5 of Part I in this series. Readers accustomed to the traditional

form [62] should read $f_{\text{int}}(r_{\text{int}}) = (-f_{\text{ext}}(r_{\text{ext}}))$.

In Part I of this series, the following thermodynamic relationships were used

$$TdS = dU + f_{\text{int}}(r)dr \quad \text{[heat flow equation]} \quad (21a)$$

$$A = U - TS \quad \text{[Helmholtz equation]} \quad (21b)$$

$$dA = -SdT - f_{\text{int}}(r)dr \quad (21c)$$

$$f_{\text{int}}(r) = -\left(\frac{\partial A}{\partial r}\right)_T = T\left(\frac{\partial S(r)}{\partial r}\right)_T \quad \text{[force]} \quad (21d)$$

$$S = -\left(\frac{\partial A}{\partial T}\right)_r \quad \text{[entropy]} \quad (21e)$$

$$H = U + rf_{\text{int}}(r) \quad \text{[enthalpy]} \quad (21f)$$

$$dH = TdS + rdf_{\text{int}}(r) \quad (21g)$$

where U is the internal energy.

Eqn (21d) requires comment. In principle, one would write

$$f_{\text{int}}(r) = -\left(\frac{\partial A}{\partial r}\right)_T = -\left(\frac{\partial U}{\partial r}\right)_{T,V} + T\left(\frac{\partial S}{\partial r}\right)_{T,V} \quad (22)$$

because $(\partial U / \partial V)_{r,T}$ may be significant, particularly since the focus of this work is on the non-ideal behavior of the RNA polymer. Early experiments on the stretching of rubber [19,64,65] suggested that the internal energy of the polymer was negligible (less than 10% for stretching up to 3 times the initial length of the rubber [64]). From these studies, it was deduced that the mere deformation process should come without significant change in the internal energy, at least when considered apart from the mixing with solvent [19]. Hence, just as $(\partial U / \partial V)_T = 0$ for the ideal gas, $(\partial U / \partial r)_T = 0$. Nevertheless, issues like Mg^{2+} binding in force extension experiments [66] may suggest a certain degree of non-ideal character that may require including other information.

Using Eqn (21d) on Eqn (5), the force acting between two mers i and j on a polymer chain becomes

$$f_{\text{int}}(r_{ij}) = \frac{\delta k_{\text{B}} T}{\xi}\left(\frac{\gamma}{r_{ij}} - \frac{\vartheta_{ij} r_{ij}^{\delta-1}}{b^{\delta}}\right) \tag{23}$$

where Eqn (23) has a minimum ($R_{ij,c}$) at

$$R_{ij,c} = \left(\frac{\gamma}{\vartheta_{ij}}\right)^{1/\delta} b \tag{24}$$

Hence, $r_{ij} < R_{ij,c} \Rightarrow f_{\text{int}}(r_{ij}) > 0$ and $r_{ij} > R_{ij,c} \Rightarrow f_{\text{int}}(r_{ij}) < 0$. This resembles the familiar spring equation using Hooke's Law: $f(x) = -k(x - x_o)$ with $0 \le x < \infty$.

Eqns (17) and (24) need some major refinement for large extensions of r_{ij} when $r_{ij} \to (j - i + 1)b$; the maximum extension. The interested reader can consult Part I of this series to find an example of such a correction in the hybrid worm like chain model (Part I, Section 5). For our purposes, the interest is directed to folding where these corrections can be neglected.

These equations can be applied generally to the structure of RNA, including the 3D structure. At any given stage of the folding process, the structure can be "frozen" and its free energy calculated using these equations for the chain entropy at least. A Heaviside function could be used to introduce the base pairing potentials with their limits on range. Hence, this is a highly versatile modeling approach that can be developed much further.

RNA folding used in this work:

The subject of the previous 4 parts of this series largely fell on the question of whether or not the CLE model could obtain a reasonable value for the global entropy. Whereas not everything is perfect, a considerable amount of testing already has been done that often shows comparable (or better) results than other approaches. Since the CLE modeling approach itself is *entirely* different, it is all the more significant.

Hence, with the CLE model, it follows that the entropic response of any partially folded structural configuration can be evaluated at the coarse-grained level for mers i and j. Since suboptimal structures are also states of the RNA structure during folding, this means that suboptimal structures should also fall into this category. RNA folding in this work will be limited to looking at the suboptimal structures of a folded sequence.

2.5. Accounting for the heterogeneous nature of RNA

At this point, it is natural to ask "how do these models account for the fact that RNA is a heteropolymer?". All coarse-grained entropy models (all models that use the JS equation [67-69], all lattice models [57,70-72] and, to a large extent, this CLE model) assume that the RNA at the coarse-grained level can be treated as though it were a homopolymer. There are several reasons why all these approaches can essentially sidestep this issue, at least to some extent.

First, for base stacking, empirical parameters from the Turner energy rules are used. For some limited range of temperatures and buffer environments, these can be expressed as $\Delta\Delta G_{bp}^{Turner} = \Delta H_{bp}^{Turner} - T\Delta S_{bp}^{Turner}$, where ΔH_{bp}^{Turner} is the enthalpy and ΔS_{bp}^{Turner} is the entropy of dinucleotide base pair formation respectively. Inspecting $\{\Delta S_{bp}^{Turner}\}$, one observes a range of values for RNA [73] between -19 and -37 kcal/molK (for AU and GC type bps, respectively) and for DNA [74] between -20 and -27 kcal/molK. The difference between GC and AU pairing is far more pronounced in RNA. This entropy term mostly arises from the freezing out of a large part of the free motion of the heterogeneous monomers due to coupling between the base pairs in the stack [20] and interactions with the solvent environment [33]. Since $\{\Delta S_{bp}^{Turner}\}$ represents empirical parameters for particular dinucleotide bps, part of the issue of heterogeneity of RNA polymers is solved by using accurate empirical parameters in the base pairs.

Second, in traditional RNA structure prediction methods [7,68], these interactions are often further augmented with corrections to the penalties for certain types of I-loops and important hairpin loops. Likewise, lattice models can take into account the physical space occupied by the bases in the loop region [72]. The CLE model makes

similar corrections based in part from information gathered on the corrections in the traditional approaches, though more needs to be done to consider physical space as in the lattice models. For pseudoknots, the CLE model implementations (*vsfold5* and *vs_subopt*) attempts to infer a fair number of structural issues based upon real RNA structures. Hence, at least feeble attempts at considering structure and heterogeneity are embedded in most of these approaches including the CLE implementations.

Third, another major correction for heterogeneity manifests itself in the flexibility of the structure itself. Only the CLE model and the lattice model Kinfold [70,75] consider the Kuhn length in any calculations. Not only is the flexibility largely frozen out within the bases (and accounted for as a heterogeneous contribution through $\{\Delta S_{bp}^{Turner}\}$), this flexibility can couple over a distance of several mers in a chain in ssRNA and in double-stranded RNA (dsRNA), it can extend over the length of a stem. Part of this flexibility results from the type of bases (e.g., a stem of GC bps versus one of AU bps), other aspects emerge from the general stacking in the stem, and still other sources are regions where Mg^{2+} can localize in pockets [29]. In the CLE model, Eqn (8) describes local corrections that account for long-range coupling and heterogeneous interactions between the mers (independent of $\{\Delta S_{bp}^{Turner}\}$) in the coarse-grained level of interactions. At present, the Kuhn length is a user adjustable parameter because there is no experimental data on how to objectively define the Kuhn length on the local scale of stems. Nevertheless, plans are in place to develop *vsfold5* and *vs_subopt* such that some of this can be automated.

Fourth, it is apparent from the work of Felsenfeld that RNA approaches an ideal polymer in behavior at room temperatures in high salt [39-41], suggesting that some aspects of the Flory Θ temperature can be mimicked even for heteropolymers like RNA. This is far more uncertain for proteins [35,38]. Yet even the amino acid heterogeneity of proteins can be address to some extent as the mere interaction of hydrophobic and hydrophilic residues in lattice models [43,76,77]. Perhaps the diversity of interactions in proteins drowns out the impact of specific interactions, rendering such approaches feasible. The bases in RNA and DNA are similar to each other when compared with the chemical diversity of proteins, making this a far less contentious issue. This permits the use of more generic functions in these problems.

Certainly, a lot more needs to be done to gain further insights on the influence of heterogeneous interactions, particularly Mg^{2+} interactions with RNA. Nevertheless, we think the heteropolymer issue of RNA is not so serious given the use of empirical base pairing free energies. Moreover, in the case of the CLE model, the empirical bp parameters are further augmented with a user definable Kuhn length.

3. A generalized solvent-polymer interaction model

The Flory-Huggins (FH) model for the polymer-solvent interactions originally addressed polymer swelling in which the ij-rmsd (Sec 2.1) was seen to increase quite dramatically in good solvent compared to the GPC model [19]. This effect occurs because there are strong attractive interactions between the polymer and the solvent and therefore, the solvent quickly coordinates with the polymer [8,10] occupying space normally taken up by the ideal polymer. The effect is so pronounced, that the rms end-to-end separation distance in Eqn (5) is seen to increase exponentially. In other instances, the same polymer can shrink or collapse to a much smaller size.

In this section, we derive a general expression for modeling collapse and swelling due to solvent interactions within the RNA structure prediction strategy of the CLE model as reviewed in Section 2. Swelling and collapse involve isotropic changes in *volume*, not simply the ij-distance (Section 2.4). Therefore, Section 3.1 introduces Flory's concept of the elastic free energy, Section 3.2 introduces corrections to the ideal polymer using the virial equation, and Section 3.3 shows how to adapt this treatment to handle various Kuhn lengths and discusses how it is implemented in RNA structure prediction and folding.

3.1. The elastic free energy

The first task is to obtain the elastic effect of changing the volume. A convenient approach to this problem is to start from Flory's development of the isotropic linear expansion parameter (α) to define the elastic free energy. Let the ideal polymer state ($\nu = 1/2$) be defines as follows

$$r_{rms}^{\mathrm{GPC}} = (\xi N)^{1/2} b \tag{25}$$

and let α be defined as the isotropic linear expansion parameter

$$r_{rms} = r_{rms}^{\mathrm{GPC}} \alpha , \tag{26}$$

where $\alpha \equiv 1$ corresponds to the ideal polymer state (r_{rms}^{GPC}). In good solvent, $\alpha > 1$.

In poor solvent, $0 < \alpha < 1$. From Eqn (26), the swelling effect yields

$$r_{rms} = \xi(N/\xi)^{3/5} b = r_{rms}^{\mathrm{GPC}}(N/\xi)^{1/10} \quad \text{or} \quad \alpha = (N/\xi)^{1/10}$$

As reasoned in Section 2, Eqn (3), in terms of the ij-rmsd, if one were to cut the terminal ends of the polymer at i and at j $(i < j)$ leaving a sequence of length $N_{ij} = j - i + 1$, one should observe behavior consistent with Eqn (3) in Eqn (26), substituting $N = N_{ij}$. The objective is to find r_{rms} for each reference ij in the polymer. Defining $N = N_{ij}$, $r = r_{ij}$ and $\alpha = \alpha_{ij}$ (with r and α variables) and using Eqn (26) to express the effect of solvent

$$r_{ij}^{'} = \alpha_{ij} r_{ij}, \tag{27}$$

and Eqn (4b) must also be modified

$$\vartheta_{ij}^{'} = \frac{\vartheta_{ij}}{\alpha_{ij}^{\delta}} \tag{28}$$

where substitution of $r_{ij}^{'}$ and $\vartheta_{ij}^{'}$ into Eqn (4) shows that Eqn (4) is invariant under these transformations. To simplify the notation in the rest of this section, we will continue to use $N = N_{ij}$, $r = r_{ij}$, $\alpha = \alpha_{ij}$ and $\vartheta = \vartheta_{ij}$.

Let N define a particular RNA model consisting of $N = N_{ij}$ mers. One can then picture a sample consisting of Q such RNA molecules, measured using X-ray spectroscopy. Alternatively, one can obtain Q samples of data from one single RNA molecule, where Q is assumed to contain a sufficiently large number of measurements.

The value of $r = r_{ij}$ in Eqn (4) represents a measurable thermodynamic state variable. For a given r, let r_k define a microstate k of r where r_k has an

ij-separation distance ranging between r_k and $r_k + \Delta r$. For the state k, there will be Q_k identical RNA molecules with this range of ij-separation distances. For a large ensemble, $Q_k \sim Qp(r_k)\Delta r$. Let $\Delta r \approx b$ and

$$\omega_k = p(r_k)(\Delta r / b) \tag{29}$$

where $\Delta r / b$ is used to set the *resolution scale* to b.

Actually, the *resolution scale* is an important dimension to remember in these models because empirical parameters like the Turner energy rules already account for the entropy internal to the RNA bases via $\{\Delta S_{bp}^{Turner}\}$ (Section 2.5). In as much as the bp entropy (internal to the mer) is decoupled from the chain entropy considered here, corrections for the Turner energy rules only emerge for $\Delta r < b$. Likewise, the coarse-grained corrections of Eqn (8) emerge for $\Delta r > b$ (Sec 2.1 and 2.2). Therefore, it is not just a convenient trick to whisk away unaesthetic terms, it demonstrates understanding as long as the *resolution scale* used is firmly enbedded in the mind's eye.

Now, the system is perturbed such that the parameters in Eqn (4) reflect the character of a distorted chain: $r_k' = \alpha r_k$ (Eqn (27)) and $\vartheta' = \vartheta / \alpha^{\delta}$ (Eqn (28)), which leaves the overall character of Eqn (4) unchanged. We approximate the number of molecules having this state as $Q_k' \sim Qp(r_k / \alpha)$ and use these relationships in the Maxwell-Boltzmann expression

$$\Omega = Q! \prod_k \frac{\omega_k^{Q_k'}}{Q_k'!} \tag{30}$$

and, taking the logarithm of both sides and applying Stirling's approximation $\ln(Q!) \approx Q\ln(Q) - Q$, the expression becomes

$$\ln \Omega = \sum_k Q_k' \ln\left(\frac{\omega_k Q}{Q_k'}\right). \tag{31}$$

Using the definition for $p(r_k)$ and $p(r_k/\alpha)$, the logarithmic term in Eqn (31) becomes

$$\ln(\omega_k Q/Q_k^{'}) = \ln\left(\frac{p(r_k)}{p(r_k/\alpha)}\right) = \vartheta\left(\frac{r_k}{b}\right)^{\delta}\left(\frac{1}{\alpha^{\delta}}-1\right)+(\delta\gamma+1)\ln(\alpha). \tag{32}$$

Eqn (32) is summed over all microstates of r_k. Although the matter of *resolution size* was belabored on, it is customary to turn to integration at this point. The elastic contribution to the FE as a function of α becomes

$$A_{el}(\alpha) = -Qk_BTC_{ij}^{\delta\gamma}\int_0^{\infty}\exp\left\{-\vartheta\left(\frac{r}{\alpha b}\right)^{\delta}\right\}\ln\left(\frac{p(r)}{p(r/\alpha)}\right)4\pi\left(\frac{r}{\alpha b}\right)^{\delta\gamma}dr \tag{33}$$

and the result simplifies to

$$A_{el}(\alpha) = (\delta\gamma+1)Qk_BT\left\{\frac{1}{\delta}(\alpha^{\delta}-1)-\ln\alpha\right\} \tag{34}$$

where inspection of Eqn (34) shows that it is positive for both $0<\alpha<1$ and $\alpha>1$. Substituting $\delta\equiv 2$ and $\gamma\equiv 1$ for the GPC solution yields $A_{el}(\alpha) = 3Qk_BT[(\alpha^2-1)/2-\ln\alpha]$, which is exactly Flory's solution for the elastic contribution to the FE when the chain is distorted when the system is a GPC [19].

Now, differentiating this expression yields the optimal value for α

$$\frac{\partial A_{el}(\alpha)}{\partial\alpha} = Qk_BT(\delta\gamma+1)\left(\alpha^{\delta-1}-\frac{1}{\alpha}\right) \tag{35}$$

and solving Eqn (35) for $\partial F_{el}/\partial\alpha = 0$, yields $\alpha = 1$. This should be expected because, so far, *nothing* has been done to perturb the system from its ideal state. The source of

this contribution will come from higher order terms associated with volume.

Eqn (34) can also be expressed in terms of V. Let $V_o = r_o^3$ and let x_o^2, y_o^2 and z_o^2 be the rms deviation in Cartesian coordinates for the 3D Gaussian distribution function. Then using Eqn (27),

$$V = (r_o\alpha)^3 = V_o\alpha^3. \tag{36}$$

and, working from V directly for the case of the GPC ($\delta = 2$ and $\gamma = 1$) and ignoring the Q, which is arbitrary because it depends on the sample size,

$$A_{el}(V) = k_{\mathrm{B}}T\left\{3\left(\frac{V}{V_o}\right)^{2/3} - \ln\left(\frac{V}{V_o}\right)\right\} \tag{37}$$

and evaluating the derivative, one finds

$$\left(\frac{\partial A_{el}(V)}{\partial V}\right)_T = k_{\mathrm{B}}T\left\{2\frac{1}{V_o^{2/3}V^{1/3}} - \frac{1}{V}\right\} = k_{\mathrm{B}}Tg(V). \tag{38}$$

This shows the ideal polymer resists both swelling and collapse away from an ideal size $V_{\mathrm{c}} = (1/2)^{3/2}V_o$, just like Eqns (23) and (24) for R_c.

Eqn (34) expresses the FE difference caused by a change in volume from the ideal state (V_o) to a swelled or collapsed state (V)

$$\Delta A_{el} = A_{el}(V) - A_{el}(V_o) = k_{\mathrm{B}}T\left\{3\left(\frac{V}{V_o}\right)^{2/3} - \ln\left(\frac{V}{V_o}\right)\right\} - 3k_{\mathrm{B}}T \tag{39}$$

and, substituting $V = V_o\alpha^3$, one obtains Eqn (34).

3.2. A model for the volume contributions in polymers

To find the internal energy contributions of a polymer, a model for the virial equation is needed. The virial equation for a real gas or liquid is expressed in the following general form [78]

$$P = -\left(\frac{\partial A}{\partial V}\right)_T = k_B T \left\{ B_1 \left(\frac{Q}{V}\right) + 2B_2 \left(\frac{Q}{V}\right)^2 + 3B_3 \left(\frac{Q}{V}\right)^3 \ldots \right\} \tag{40}$$

where V is the volume, Q is the number of atoms (or molecules) of gas, P is the pressure, and one can see that the first term of this expression corresponds to the ideal gas (*i.e.*, $B_1 = 1$ and *c.f.*, $PV = Qk_B T$). The terms following involve two-body (B_2) and three-body (B_3) interactions [8]. The B_2 term is largely attributed to repulsive interactions that arise when two gas molecules come within too close a proximity to each other and the B_3 term (and higher order terms) often arises due to weak attractive interactions between the gas molecules.

For an ideal polymer,

$$P = -\left(\frac{\partial A}{\partial V}\right)_T = -\left(\frac{\partial U}{\partial V}\right)_T + T\left(\frac{\partial S}{\partial V}\right)_T \tag{41}$$

which is analogous to Eqn (22). Just like the ideal gas, $(\partial U / \partial V)_T = 0$ and this leads to

$$P = T\left(\frac{\partial S}{\partial V}\right)_T \tag{42}$$

and, for convenience, we will assume that Eqn (42) is correct.

Polymer swelling (or collapse) involves a process associated with volume (V). The *volume* is connected with the ij-rmsd through Eqn (26); $V = (r_{rms})^3 = (\alpha r_{rms}^{\mathrm{GPC}})^3$. To construct a solution for the equation of state for a polymer, it is easiest to start from

the free energy

$$A(V) = A_{el}(V) + A_{vol}(V) + \text{ terms independent of } V \tag{43}$$

where $A_{el}(V)$ is Eqn (37) and $A_{vol}(V)$ expresses the *volume interactions* [10],

$$A_{vol}(V) = Qk_{\mathrm{B}}T\left\{B_2\frac{Q}{V} + B_3\left(\frac{Q}{V}\right)^2\cdots\right\}, \tag{44}$$

where Q/V reflects the concentration of the RNA polymer, it is assumed that the total volume is occupied by Q such molecules, and B_2 and B_3 are the second and third virial coefficient (as in Eqn (40)). In principle, for the true ideal polymer, one finds $B_2 = 0$ and $B_3 = 0$ (Eqns (34) and (37)). However, real polymers are not anything remotely Gaussian in character, and this means that what happens at Flory's Θ temperature, is that all the higher order terms in Eqn (44) almost exactly cancel each other. Nevertheless, we assume the ideal polymer exists (or at least that the corresponding B_2 and B_3 are small enough). Substituting Eqn (37) and Eqn (44) into Eqn (43), one obtains

$$\begin{aligned} A(V, B_2, B_3) &= k_{\mathrm{B}}T\left\{3\left(\frac{V}{V_o}\right)^{2/3} - \ln\left(\frac{V}{V_o}\right)\right\} \\ &+ k_{\mathrm{B}}T\left\{QB_2\frac{Q}{V} + QB_3\left(\frac{Q}{V}\right)^2\right\} + \text{ terms independent of } V \end{aligned} \tag{45}$$

Since the purpose of this work is to obtain an expression for A in terms of N (the number of mers) and the present units of Q/V are molecules per equivalent volume, further transformations are needed. Let $V_m = r_{rms}^3$, $Q = V/V_m$, and N be the sequence length of the RNA. Then, using Eqn (36), Q/V is redefined such that

$$N\frac{Q}{V}=\frac{N}{V_m}=\frac{N}{r_{rms}^3}=\frac{N}{(r_o\alpha)^3}. \tag{46}$$

where the concentration now measures the number of mers contained in the volume of a single molecule of RNA and corrective weights such as $4\pi/3$ (i.e., a "spherically symmetric" RNA molecule) are assumed to be absorbed into the virial coefficients. It is justified to set $Q=1$ in Eqn (46) because RNA folding and structure prediction is directed to what can be effectively termed a single molecule measurement. Since the equations presumably already reflect the statistical properties of any number of such molecules, the result for one RNA molecule is the same as for many.

There is a further complication. Eqn (37) has an implicit N in the volume term already. Moreover, we are looking at the volume generated by a sequence of length $N=N_{ij}$ where the equations say absolutely nothing whatsoever about $N_{i+1,j-1}$. Though it may be natural to infer some relationship, the equations could care less. As a result, we are only counting ij-rmsd. Therefore, just like we don't multiply Eqn (23) by N_{ij}, so we also do not multiply Eqn (37) by such an N. On the other hand, Eqn (44) is simply imported from Eqn (40). In essence, what has been assumed is that the monomers of a *single RNA molecule* can be represented as a "gas". This depends on the number of "gas particles" in the volume of the polymer; i.e., the sequence length N. Rewriting Eqn (45) in terms of V_m and N, one obtains

$$\begin{aligned}A(V_m,B_2,B_3)&=k_{\mathrm{B}}T\left\{3\left(\frac{V_m}{V_o}\right)^{2/3}-\ln\left(\frac{V_m}{V_o}\right)\right\}\\&+k_{\mathrm{B}}T\left\{B_2\frac{N^2}{V_m}+B_3\frac{N^3}{V_m{}^2}\cdots\right\}+\text{ terms independent of }V_m\end{aligned} \tag{47}$$

Evaluating Eqn (47) in terms of Eqn (41), an expression resembling Eqn (40) is obtained

$$P = -\left(\frac{\partial A}{\partial V_m}\right)_T = T\left(\frac{\partial S}{\partial V_m}\right)_T = k_B T\left\{ NB_1 g(V_m) + B_2\left(\frac{N}{V_m}\right)^2 + 2B_3\left(\frac{N}{V_m}\right)^3 \cdots \right\} \tag{48}$$

where $g(V)$ is defined in Eqn (38) and $B_1 = 1/N$. Substituting $V_m = V_o$ into Eqn (38) and assuming $B_2 = 0$ and $B_3 = 0$, one obtains $P = k_B T / V_o$, which is the osmotic pressure of an ideal polymer with $B_1 = 1/N$. Since N is proportional to the molecular weight, Eqn (48) is consistent with the equations for osmotic pressure [79].

Using the definition of a perfect polymer $B_2 = 0$, $B_3 = 0$ and employing Eqn (46),

$$\begin{aligned} &\Delta A(\alpha) = A(\alpha, B_2, B_3) - A(1,0,0) = \Delta A_{el} + \Delta A_{vol} \\ &= 3k_B T\left\{\frac{1}{2}(\alpha^2 - 1) - \ln\alpha\right\} + k_B T\left\{ B_2\frac{N^2}{(r_o\alpha)^3} + B_3\frac{N^3}{(r_o\alpha)^6}\right\}. \end{aligned} \tag{49}$$

The general expression for Eqn (49) is

$$\Delta A(\alpha) = (\delta\gamma + 1)k_B T\left\{\frac{1}{\delta}(\alpha^\delta - 1) - \ln\alpha\right\} + k_B T\left\{ B_2\frac{N^2}{(r_o\alpha)^3} + B_3\frac{N^3}{(r_o\alpha)^6}\right\}. \tag{50}$$

Now, taking the derivative of Eqn (49)

$$\left(\frac{\partial A}{\partial \alpha}\right)_T = 3k_B T\left\{\alpha - \frac{1}{\alpha}\right\} - 3k_B T B_2\frac{N^2}{r_o^3\alpha^4} - 6k_B T B_3\frac{N^3}{r_o^6\alpha^7} \tag{51}$$

and solving for the stationary points, one obtain

$$\alpha^5 - \alpha^3 = \frac{B_2 N^{1/2}}{b^3} + \frac{2B_3}{b^6\alpha^3} \tag{52}$$

where in good solvent, one can typically ignore the constant contribution of B_3. The

relationship between B_2 and B_3 can, in principle, be found either by calculation or by experiment.

If B_2 and B_3 are zero, $\alpha = 1$. This is known as the athermal condition where the polymer behaves as though it were a GPC (or somehow we find $T \approx \Theta$).

In the limiting case of large N and in good solvent conditions, one can see that the dominant term is B_2. Solving Eqn (52) for $B_2 >> B_3$

$$\alpha_{max} \sim \left(\frac{B_2 N^{1/2}}{b^3} \right)^{1/5} = \frac{B_2^{1/5} N^{1/10}}{b^{3/5}} \tag{53}$$

and

$$r_{max} = \alpha_{max} r_o = \left(\frac{B_2}{b^3} \right)^{1/5} N^{3/5} b \,. \tag{54}$$

For a polymer in poor solvent, one finds a situation where B_2 is negative and both B_2 and B_3 are significant. Rearranging Eqn (52) yields a polynomial equation that can be solved using numerical methods

$$\alpha^8 - \alpha^6 - \frac{B_2 N^{1/2} \alpha^3}{b^3} - \frac{2B_3}{b^6} = 0 \tag{55}$$

In the limiting case of large N, Eqn (55) reduces to

$$\alpha_{min} \sim \left(\frac{2B_3}{-B_2 N^{1/2} b^3} \right)^{1/3} = \frac{1}{b N^{1/6}} \left(\frac{2B_3}{-B_2} \right)^{1/3} \tag{56}$$

and

$$r_{min} = \alpha_{min} r_o = \left(\frac{2B_3}{-B_2} \right)^{1/3} N^{1/3} \tag{57}$$

where B_2 has units of volume and B_3 volume squared.

We have obtained the excluded volume $r \propto N^{3/5}$ dependence and the corresponding globular case where $r \propto N^{1/3}$ [10].

For completeness, we write the general form for Eqn (52), which can be found by using the same procedures as used to derive Eqn (50)

$$\alpha^{\delta+3} - \alpha^3 = \frac{3B_2 N^{1/2}}{(\delta\gamma+1)b^3} + \frac{6B_3}{(\delta\gamma+1)b^6\alpha^3} \tag{58}$$

where one can quickly see that for $\delta \equiv 2$ and $\gamma \equiv 1$, Eqn (58) reduces to Eqn (52). Moreover, applying the renormalization group solution of $\delta \approx 2.5$ into Eqn (58) and a small positive constant for B_2, one quickly observes the renormalization group estimate for 2ν: $2\nu = 13/11 \approx 1.18$. Hence, this independent approach of the FH model appears to have generated some of the critical exponent values that are surprisingly consistent with renormalization group theory [8].

3.3. Application to RNA structure prediction and folding

The general approach of RNA structure prediction and folding was explained in Section 2.4. The approach introduced here is effectively a refinement of the approach already developed in the CLE model. To adapt the FH model to RNA structure prediction and folding, the first task is to solve Eqn (52) or (58) for α_{ij}. All equations referencing bp ij will now explicitly contain ij indices.

In Section 3.2, the Kuhn length was neglected. RNA has a Kuhn length that influences the stiffness of the RNA and reflects how nature has reduced the number of degrees of freedom on the biopolymer to produce the structure one sees in journals and textbooks. Having a Kuhn length $\xi > 1$ means that we must divide the sequence into N_{ij}/ξ effective mers (or *epimers*) each of which has a length ξb, Eqn (2).

To do RNA structure prediction with the CLE model, one must estimate the ij-rmsd for each base pair (i, j) in the polymer chain rather than just the extreme ends

($i \equiv 1$ and $j \equiv N$). The α_{ij} parameter depends on N_{ij} and ξ. At the same time, α_{ij} only depends on the difference $n = j - i + 1$, hence, it is only necessary to compute α_n and then use that value for all $n = j - i + 1$. In Eqn (27), α_{ij} is a function of N_{ij} and, for each base pair (i, j),

$$r_{ij} = \alpha_{ij} r_{ij_o} = \left(N_{ij} / \xi \right)^{\nu_{ij}} \xi b \tag{59}$$

$$\alpha_{ij} = \left(N_{ij} / \xi \right)^{\nu_{ij} - 1/2} \tag{60}$$

Making appropriate substitutions, Eqn (56) becomes

$$\alpha_{ij}^5 - \alpha_{ij}^3 = \frac{B_2 (N_{ij} / \xi)^{1/2}}{(\xi b)^3} + 2 \frac{B_3}{(\xi b)^6 \alpha_{ij}^3} \tag{61}$$

or, for $\delta \neq 2$ and $\gamma \neq 1$, (61) takes on the form in Eqn (58)

$$\alpha_{ij}^{\delta+3} - \alpha_{ij}^3 = \frac{3 B_2 (N_{ij} / \xi)^{1/2}}{(\delta\gamma + 1)(\xi b)^3} + \frac{6 B_3}{(\delta\gamma + 1)(\xi b)^6 \alpha_{ij}^3} \tag{62}$$

Only positive roots of Eqns (61) and (62) can be used and the dominant term is B_2 when N_{ij} is large. If $\alpha_{ij} > 1$ and N_{ij} is large, then $\alpha_{ij}^{\delta+3} > \alpha_{ij}^3$ and B_2 must be positive. If $0 < \alpha_{ij} < 1$ and N_{ij} is large, then $\alpha_{ij}^{\delta+3} < \alpha_{ij}^3$ and B_2 would tend to be negative, B_3 positive and $\nu_{ij} < 1/2$ (for any large value of N_{ij}). Hence, specifying a negative value for B_2 will insure that the response of the polymer will resemble a globular system for N_{ij} larger than some length that depends on the particular coefficients in the two-body and three-body interaction terms. Values of B_3 must be positive to generate a positive real root in this case. If B_2 is positive, then, irrespective of B_3 (usually positive), $\nu_{ij} > 1/2$ and B_3 can be neglected in most cases. Finally,

when $B_2 = 0$ and $B_3 = 0$, the equation reduces to the GPC for large N_{ij}.

The exponent ν_{ij} can be expressed as a function of α_{ij} through the following relationship

$$\nu_{ij} = \frac{1}{2}\left(1 + \frac{2\ln(\alpha_{ij})}{\ln(N_{ij}/\xi)}\right). \tag{63}$$

Eqn (63) can now be substituted into Eqn (6) for each ν_{ij} yielding

$$\Delta S(N_{ij}, \nu_{ij}, \xi) = \frac{k_B T}{\xi}\left\{\nu_{ij}\delta\gamma \ln\left(\Psi(\nu_{ij}, \xi)N_{ij}\right) - \zeta(\delta, \gamma)\left(1 - \left(\frac{1}{\Psi(\nu_{ij}, \xi)N_{ij}}\right)^{\nu_{ij}\delta}\right)\right\} \tag{64}$$

where λ is defined in Eqn (6), ζ in Eqn (4c) and

$$\Psi(\nu_{ij}, \xi) = \frac{1}{\xi}\left(\frac{\xi}{\lambda}\right)^{1/\nu_{ij}} b \tag{65}$$

and the remaining formalism of Eqns (7) and (8) are also used. In Eqn (63), for $0 < \alpha_{ij} < 1$, $0 < \nu_{ij} < 1/2$, and for $\alpha_{ij} \geq 1$, $\nu_{ij} \geq 1/2$. In general, physical values for ν_{ij} range from roughly $0. < \nu_{ij} < 0.6$.

In the limit of large N_{ij} and $B_2 < 0$, $\alpha_{ij} = \left\{2B_3 / \left[-B_2 (N_{ij}/\xi)^{1/2} (\xi b)^3\right]\right\}^{1/3}$ and

$$r_{rms,ij} = r_{rms,ij}^{GPC}\alpha_{ij} = \left\{2B_3 N_{ij} / \left(-B_2 \xi\right)\right\}^{1/3} \tag{57'}$$

For the infinite chain with this kind of compaction, it is very important that the result of $r_{\min}$ satisfy Eqn (57); i.e., Eqn (57') with $\xi = 1$. For spherically symmetric beads on a chain with a volume $(\xi b)^3$, the density (ρ) of the material in the limit of large N is

$\rho = N/(r_{rms})^3 \propto N/\left\{2B_3 N/\left(-B_2\xi\right)\right\}^{3\nu}$ [80,81]. For $\xi > 1$, over some initial range of N, ν can be smaller than $\nu = 1/3$. This is because the volume of an RNA base is smaller than the volume of the effective mer: c.f., $b^3 < (\xi b)^3$. Likewise, for the ideal polymer, $[N^{1/2}b]^3 < [(\xi N)^{1/2}b]^3$. Therefore, a considerable amount of space is unused for $\xi >> 1$. Nevertheless, this sets a further restriction on acceptable sizes for ν that should be remembered.

Table 1 summarizes the general characteristics of ν_{ij} for a given set of polymer-solvent conditions and the corresponding virial coefficients. Table 2 shows the solutions for virial coefficients B_2 and B_3 as a function of ξ for a coexistence region with critical length at $N_c \sim 50$ nt (where the transition occurs) and a critical width of $R_c \sim \pm 10$ nt (the main span over which the drop occurs) and standard conditions $\delta \equiv 2$, and $\gamma \equiv 1.75$, Fig 1 and 2 plot the results of Table 2. Particularly for the larger values of ξ in Table 2, there is a rough tendency for B_2 and B_3 to follow $B_3 \sim (B_2)^2/2$ and $B_2(\xi) = \prod_{k=1}^{\xi-\xi_0} (\xi_0 + k)/(\xi_0 + k - 1)$, where ξ_0 is a reference Kuhn length and ξ and ξ_0 should be treated as integers in the product.

Figure 3 shows a calculation of the coexistence region using Eqn (62) with $\xi = 4$ nt and using the parameters in Table 2. From Eqn (62) with $B_2 < 0$ and $B_3 > 0$, the value of a ranges from 0 to 1. The coexistence region lies between $N_{c1} = 40$ and $N_{c2} = 60$ nt ($R_c = \pm 10$) in Fig 3, and the mid-point is at $N_c = 50$ nt. This is assigned as the critical point (N_c). The width on each side of N_c is 10 nt. Translating α into ν using Eqn (63), the values are shown in the lower part of Fig 3 as a function of sequence length N for the same conditions ($\xi = 4$ nt).

Solutions tend to follow the general tendencies predicted in Table 1, at least in the limits. For example, the solution in Fig 3 for large N gradually increases from 0.17 to 0.283 (at $N > 20000$). However, not all combinations of B_2 and B_3 are reasonable values when $B_2 < 0$. For one thing, some solutions of Eqn (63) can be negative around the coexistence region and gradually turn positive only as N

increases away from N_c. Both B_2 and B_3 tend to show geometric growth with increasing ξ. It is important therefore not only to seek solutions with positive roots for α_{ij}, but also roots that yield a positive value for ν_{ij} in Eqn (63).

Eqns (62) through (65) are new to the RNA structure prediction; particularly when generalized for γ, δ and ξ. Further, using the CLE formalism where the focus is on the stem interactions and base pairing (i, j) with a weight $N_{ij} = j - i + 1$, this perspective is new. However, what is far more important is to model the environment of a biopolymer correctly. We have shown in the previous four parts and in previous work that this formalism is certainly competitive with existing formalisms that focus on loops. Presumably, if a model is good, it gets better with better information.

3.4. Experimental values of B_2 in relation to RNA

Up to this point, everything has been done from the standpoint of theory. Here we look at what we can glean out from the experimental side of the picture.

Experimentally, Eqn (52) is usually measured using some form of the following expression [19]

$$\alpha^5 - \alpha^3 = C\left(1 - \frac{T}{\Theta}\right) \tag{52'}$$

where, in practice, C is a positive empirical constant. Since the dominant term tends to be B_2, the right hand side is roughly proportional to B_2 and this means that B_2 changes sign as it passes through the Θ temperature.

In applying these equations to experiments, the "volume" (V or V_m) in these equations describes the volume of the polymer, but a real experiment measures B_2 and B_3 via osmotic pressure and this is measured in terms of the concentration of solute c_{RNA} mixed together within a significant volume of solvent (V_{expt}). The equations conveniently ignored these "details". Like magic, V_m conveniently arrived on the scene. A real V_m requires someone to measure it in an experiment via the radius of

gyration to infer the value of r_{rms}. How might we connect these equations to any experiments?

The virial coefficients need to be weighted by the volume of the polymer and solvent mixture (V_{expt}) and the quantity of solute (Q_{RNA}), where $c_{\text{RNA}} \leftrightarrow Q_{\text{RNA}} / V_{\text{expt}}$. Since concentration is usually measured in units of M or mg/ml, for example, B_2 needs to be weighted by a conversion factor w_c. A conversion (w_c) between nm^3/mer to M^{-1} is 0.602 $\text{merM}^{-1}\text{nm}^{-3}$. In addition, c_{RNA} contains the number of moles of the molecule, not the number of mer in the molecule. It is then necessary to rescale V_m to reflect V_{expt}. This is done by weighting B_2 by $Q_{\text{RNA}} V_m / V_{\text{expt}}$, where V_m is weighted by Q_{RNA} in order to account for the number of RNA molecules in the volume V_{expt}. Therefore, a corresponding experimentally measured value for the n^{th} virial coefficient (B_n') would require formulating the equations in this section to the following

$$B_n' = B_n \left(w_c \frac{Q_{\text{RNA}}}{V_{\text{expt}}} \frac{V_m}{N} \right)^n$$

where B_n' is calculated using c_{RNA}. This means that one needs to know the value of V_m at the Θ temperature. It also assumes that the Kuhn length doesn't change, which is unlikely to be the case. Such a myriad of issues cannot be addressed here and will have to be ignored. The point is, these are very difficult experiments to do and the interpretation is complex to say the least.

There is very little experimental information on B_2 values. The only measured B_2' data available is from Felsenfeld and coworkers [39,40]. For poly(A) in 1M NaCl, the low temperature (10^{o}C) and high temperature (50^{o}C) data shown a positive B_2', but at room temperature, B_2' was negative. For poly(U) in 2M NaCl, B_2' increased monotonically from a negative value at 10^{o}C to a positive value up to 50^{o}C with the Θ temperature located around 18^{o}C. It is clear from Felsenfeld's data that more salt actually causes B_2 to go negative sooner and perhaps more strongly. Therefore,

lower salt conditions inside the cell may yield a tendency toward a positive B_2.

How to combine these results with real RNA is not obvious. In general, B_2 would be more positive at higher temperatures because the water is able to solvate the RNA better. Interestingly, adding divalent cations to RNA tends to cause the Tetrahymena ribozyme to collapse into a globular structure [27,82]. Felsenfeld's data also suggests that increasing the ionic strength drives B_2 more in the direction of a poor solvent at typical experimental temperatures. This would suggest that a negative B_2 is one way to model the long range effects of divalent cations.

In the ambient temperature range for poly(A), Felsenfeld and coworkers [39] reported B_2 values ranging -0.001 to -0.020 for sequences lengths of order $N \approx 1400$ nt for a sample with a concentration of 170 mg of solute in 32 ml of solvent. Estimating $\xi = 8$ nt and $b = 5.9$ Å, $r_{rms} = (\xi N)^{1/2} b \approx 62.4$ nm, $V_m = r_{rms}^3 = 2.4 \times 10^5 \,\mathrm{nm}^3$, $V_{\mathrm{expt}} = 37 \text{ ml} \rightarrow 3.7 \times 10^{22} \,\mathrm{nm}^3$, and

$$(170 \text{ mg poly(A)})(10^{-3}\text{g/mg})\left[\text{mol}/(338\text{g} \times 1400 \text{ mers})\right]$$
$$\rightarrow 3.6 \times 10^{-7} \text{ mol poly(A)}$$
$$\rightarrow Q_{\mathrm{RNA}} = 2.2 \times 10^{17} \text{ poly(A)}$$

Then, using the value for B_2 in Table 2,

$$B_2' = B_2 \left[w_c \frac{Q_{\mathrm{RNA}}}{V_{\mathrm{expt}}} \frac{V_m}{N} \right] = (-12)\left[6.1 \times 10^{-4} \mathrm{M}^{-1} \right]$$
$$\rightarrow -7 \times 10^{-3} \mathrm{M}^{-1}$$

which is in the ball park of what Felsenfeld and coworkers observed, though it is not clear exactly how strong the stacking effect and corresponding stiffness of the poly(A) was in the sample. This is little more than a shot in the dark and the existing data is surely inadequate for an accurate interpretation of anything pertaining to RNA in general. Nevertheless, if we are willing to extrapolate on the observations of Felsenfeld et al., then the tendency for RNA to collapse as a result of the addition of Mg^{2+} is largely a peculiar and counterintuitive consequence of solvent collapse. Perhaps the RNA begins to prefer itself more than the solvent in the presences of the

divalent cations.

4. An approximation model: the effective Flory-Huggins model

In Sections 3 we developed the theory for Flory-Huggins model. That model can be extended by including the additional virial coefficients B_n and these virial coefficients can even be measured experimentally, in principle. In this way, the study in Section 3, supplemented with the contributions from the CLE model, permits us to move out of the world of "ideal polymers" to a more accurate and experimentally sound model of the polymer chemistry. However, it is also clear from the latter part of Section 3, that obtaining these parameters is not easy. Indeed, some solutions are inadmissible because they yield unphysical results. Since the virial parameters depend upon environmental conditions, it requires considerable study and experimental measurement under a variety of conditions to develop a full set for such a model. That set is largely lacking. At the current stage of development, we also have not really been able to test this model beyond a few example problems such as those we provide in Section 5. The precise position, size and solvent dependence of the coexistence region are largely unknown. Therefore, although formally and aesthetically pure, at a practical level, calculations with genuine virial coefficients are not particularly convenient at this time. We therefore introduced an option "−pflory" in *vsfold5* (and *vs_subopt*) that attempts to imitate the behavior of ν_{ij}, without the complications of seeking appropriate values for B_2 and B_3.

An example of the behavior of ν_{ij} for some hypothetical RNA conditions is shown schematically in Fig 4. In Fig 4, the RNA is depicted either existing in a swelled state ($\nu > 0.5$) or existing as a default state like the GPC ($\nu = 0.5$) at short lengths (due to the exposure of RNA to solvent) and becoming more globular at very long lengths where the polymer has more chances to squeeze out solvent and increase the number of contacts with itself. The region labeled N_c indicates the critical length where this transition occurs. The parameters R_c indication the coexistence region where ν_{ij} transitions between ν_1 ($N_{ij} < N_c$) and ν_2 ($N_{ij} > N_c$).

In the FH model with the real parameters, we modeled the transition in the coexistence region by a sudden drop at the midpoint. Here, we opted to control this

behavior by allowing parameters R_c to define the sharpness of that drop and permitting the equation to be a linear transition. The precise character of ν_{ij} in the coexistence region is not well understood for most biomolecules in any environment. Therefore, since the user can have full control over the parameters R_c and N_c, rather than enforce some arbitrary and literal evaluation scheme, we have left it up to the user to decide how best to construct this region.

The model is intended to approximated the attraction effect when $B_2 < 0$ and $B_3 > 0$. However, swelling can be also modeled this way by setting $\nu_1 = \nu_2$ and using an arbitrary R_c and N_c. The user can alter these values with the options "−pfNc N_c" and "−pfRc R_c" (where N_c and R_c reflect appropriate input values).

The advantage to approaching the problem this way is that one can obtain and control the properties of this polymer swelling effect without having to solve for the second and third virial coefficients (B_2 and B_3). This model does not replace the FH model as much as circumvent the difficulties of solving non-integral sixth order polynomial equations. The virial coefficients are very strongly dependent on the Kuhn length (ξ), and therefore, finding the exact parameters that produce a particular shape for N_c (with R_c) depends strongly on ξ, δ, and γ.

Since short RNA sequences tend to behave similar to a GPC due to significant solvent exposure and long RNA sequences tends to form compact structures with little solvent occupying the space inside, at present, the default parameters used in *vsfold5* are $N_c \sim 50$ nt, $R_c \sim 10$, $\nu_1 = 0.5$ and $\nu_2 = 0.3$. Whereas these parameters appear to be consistent with the general tendencies of RNA polymers, the user has the option to change the parameters if better information is available.

5. Application of the model

To this point, only the theoretical aspects of this model have been explained. Whereas the model has considerable potential, and was provided with the earliest web versions of *vsfold* (http://www.rna.it-chiba.ac.jp), few opportunities have emerged to test the model rigorously.

In Part III, the folding landscape of structures for tRNA(Phe) using straight GPCs parameters was examined. The resulting set of structures is shown in Fig 5. In Fig 5(top), we carry out the same calculation including the Flory-Huggins (FH) parameters for $\xi = 5$ nt listed in Table 2. The bottom listing is the same as found in Part III. The funnel shape shows some visible improvement when the option for the FH model is used. Nearly all the structure examined from the bottom up show an order that might reasonably be expected in a denaturing/refolding experiment. The bottom most structure contains the partially folded tRNA intermediate on 5' domain and a weak hairpin on the 3' domain. The 3' domain vanishes in the next frame and a gradually compacting pseudoknot intermediate takes its place. From there, the structure takes up the familiar final course.

The SAM ribozyme in Fig 6 of Part III was also redone using the Flory option. For that sequence, with the exception of differences in the energies, there was no significant difference between the suboptimal structures predicted with the FH model and those predicted with the parameters (Table 2) for $\xi = 9$ nt and all other conditions identical. Although adding the Flory option is likely to introduce more complexity and bring about observable structure changes, it still does not necessarily mean that anything will actually happen. Nevertheless, the energy was lower, which is likely to have provided some additional favorability to the P1 stem.

As a third example, we show how the Flory contribution affects the behavior of tmRNA. The tmRNA consists of four pseudoknots PK1-4 (Fig 6(a)) [83,84], where tRNA(Ala) is compared on the top right side of the same panel. Individually, *vsfold5* (and *vs_subopt*) can fit all four of these pseudoknots (PK). Hence, treated on a domain level as discussed in Part IV, each PK module is solvable using *vsfold5* for some Kuhn length. However, three of the four PKs (PK2,3, and 4) have at least one long stem that is much longer than those on PK1 (which is the most critical and has the short stems). In

Part II (Section 7), we observed that using improper Kuhn lengths leads to errors. When the variations in the Kuhn length are large, there may not be a tractable solution for a single monolithic Kuhn length. In Ref [15], several of these PK modules were shown to fit with separate Kuhn lengths. Fig 6(b) shows a fit of the tmRNA sequence of E. coli (X16382, accession 2482406) with standard parameters and a Kuhn length of $\xi = 4$ nt (Fig 6(b)), and Fig 6(c) shows the same fit with $\xi = 4$ nt and the Flory-Huggins parameters $B_2 = -0.9$ [nm^3] and $B_3 = 1.0$ [nm^6] ($N_c = 55$ and $R_c = \pm 20$). In Fig 6(c), using the FH parameters, we see many of the characteristic features of longer RNA sequences: the compacting of the general structure and the higher stability of larger domains. We also see that PK1 is visibly present in the figure. This is little more than a crude fit, but *vsfold* is showing that it can find some of the relevant structure. Using default parameters, the effective Flory-Huggins model (Fig 6(d)) discussed in Section 3 is able to detect some of the structure, but does not show the same extent of important compact features seen in Fig 6(c) using the Flory-Huggins model. Perhaps more important is the visible compacting of the long-range structure with the introduction of the Flory-Huggins model.

To improve the fit, it will be necessary to include a variable Kuhn length. It is also likely that interactions between different PK modules influence the overall structure of this complex molecule, for which only primitive 3D structure analysis tools are built into the *vsfold5* algorithm. Nevertheless, even in this crude calculation, this may explain some of the qualitative structural features suggesting a tendency for large domains of RNA to be somewhat more compact.

The precise location of N_c is not currently known because the role of a real polymer with excluded volume (the B_2 term) and understanding the tendency toward globular behavior rather than swelling (resulting from the combined effect of B_2 and B_3) does not appear to have been a subject of much interest in experimental studies of nucleic acids. However, based on the observed behavior of tRNA in our studies, it is reasonable to think that it is around $40 < N_c < 60$. Since the virial coefficients are both specific for each N_c and, for each N_c, strongly dependent on the Kuhn length, we have focused our current efforts on the one example of $N_c = 50$ for simplicity.

The results reported here are not intended as a rigorous study of the full potential of the model. Rather, they represent routine tests we made after adding the functionality. It is all the more encouraging therefore, that they also turn out to produce some reasonable results.

In this study, we have aimed at collapse of the RNA as $0 < \nu < 1/2$. However, studies by by Wiese and coworkers [58,85,86] and by Makarov and coworkers [51] suggest that $\nu = 0.6$. There is nothing restricting the possibility of using this software with $1/2 < \nu < 1$ in principle. We favor the use of $0 < \nu < 1/2$ because when Mg^{2+} is added to the Tetrahymena ribozyme [28], the ribozyme is observed to collapse and form a state resembling the 'molten globule' state observed in some proteins [87]. In some ways, the *absence* of Mg^{2+} amounts to a "denaturing solvent" . Several other systems have been observed to compact with addition of Mg^{2+}: RNase P [88], 5S ribosomal RNA [89] and even tRNA [90-93]. Wiese and coworkers were studying the general properties of RNA and use the JS equation which is not as general as Eqns (5) through (8). At least in the context of trying to model this collapse, we favor a picture closer to that in Figure 3. Nevertheless, given the paucity of experimental data offering any clue whatever, we don't insist on this position. The flexibility of *vsfold5* and *vs_subopt* also permits a user to choose a denaturing condition, if so desired.

Conclusions

In this work, we expand the scope of the cross linking entropy model to include the option to predict RNA structure and folding under the non-ideal solvent conditions that are largely expected to exist for RNA in most *in vivo* and *in vitro* conditions in which it is measured. The model finds a generalized equation for evaluating the second and third virial coefficients using the Flory-Huggins model of excluded volume. Since there is very little experimental information about the virial coefficients or their influence on RNA, this work is simply meant to present these concepts.

Many questions remain about the proper way to model the conditions of RNA, particularly in vivo. The Kuhn length is strongly affected by the salt content with high salt tending to reduce its magnitude [17,18], yet almost all the measurements of RNA thermodynamic parameters are done in high salt. Do non-specific electrostatic interactions with proteins and nucleic acids in the cell help to stabilize RNA structures? Judging the available information on B_2, it is more prone to become negative in high salt, suggesting that the Mg^{2+} causes the RNA to behave as though in poor solvent. Maybe stability comes at the expense of the RNA becoming more insoluble. The tendency to accelerate the rate of folding of RNA to the native state by addition of urea [27,42] suggests that cellular environment may play a significant role. Ironically, in such cases, the "denaturing solvent" actually "lubricates" the folding process. Extrapolating these observations to the cell, this suggests that it is important to understand the interaction of RNA in the presence of proteins and in an *in vivo* environment. For example, histones show rather specific binding with the DNA in the *globular* regions of the protein [94]. Is it structured or is it not? Some capsid proteins can mix very thoroughly with the RNA in some viruses [95]. What sort of polymer solution is a system like that? It would be of considerable value to know how the *in vivo* cellular interactions influence RNA structures.

In conclusion, what this work has most revealed is that there is a lot more to be learned particularly with respect to RNA in its true environmental conditions. Both theory and experiment would be greatly facilitated by empirical knowledge of the virial coefficients in these complex polymer systems.

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 thank the students at CIT: Michiko Fujii who provided the structure figure for tmRNA and we thank Amiu Shino, Misaki Imai and Kenta Kondo for their encouragement. We also thank Dr. Shingo Nakamura (Catalent Pharma Solutions), Profs Kentaro Shimizu, Shugo Nakamura, Tohru Terada, and Kazuya Sumikoshi (UofT) and Dr. Yucong Zhu for their ongoing encouragement.

Software

Binary versions of *vsfold5* and *vs_subopt* (where both support the Flory-Huggins model calculations) are available upon request to the corresponding author and upon written consent to the license agreement. Available formats are 64 bit Linux (x86_64), or 32 bit Linux, Visual C++ Window XP/Window 7, and Mac OSX 10.4 (32 bit).

Figures

Figure 1

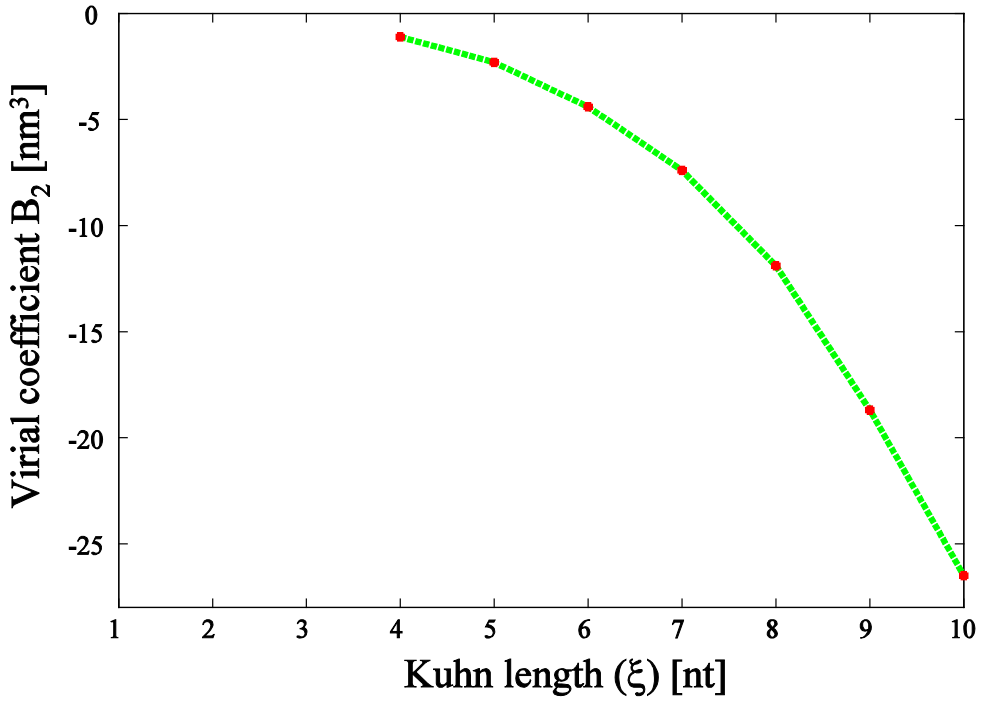


Figure 2

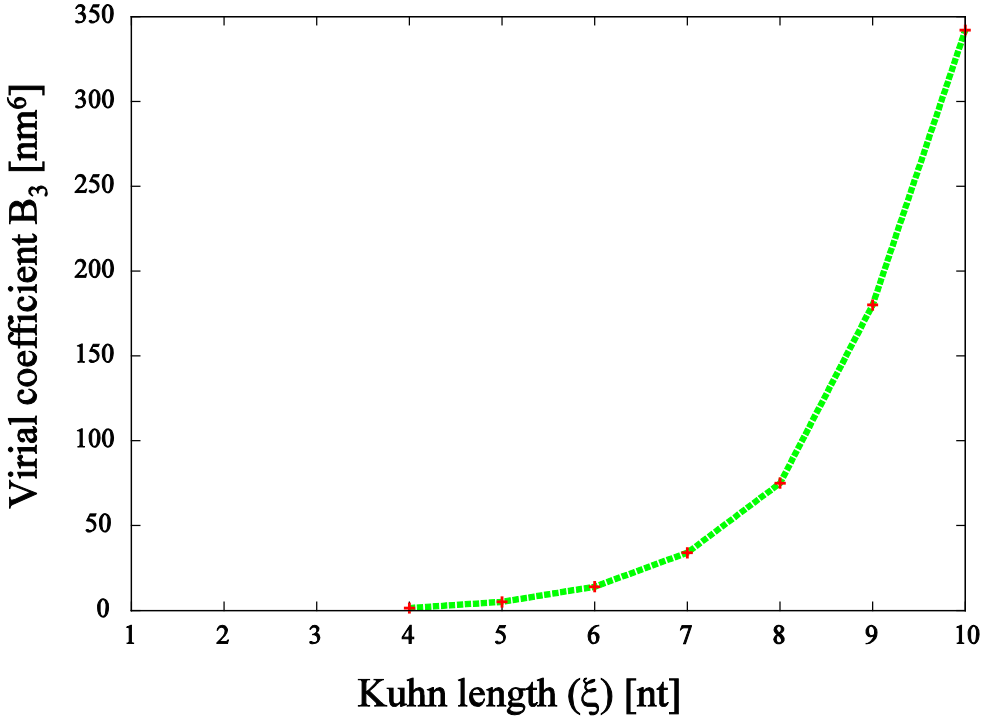

Virial coefficient B_3 [nm^6]
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
Kuhn length (ξ) [nt]

Figure 3

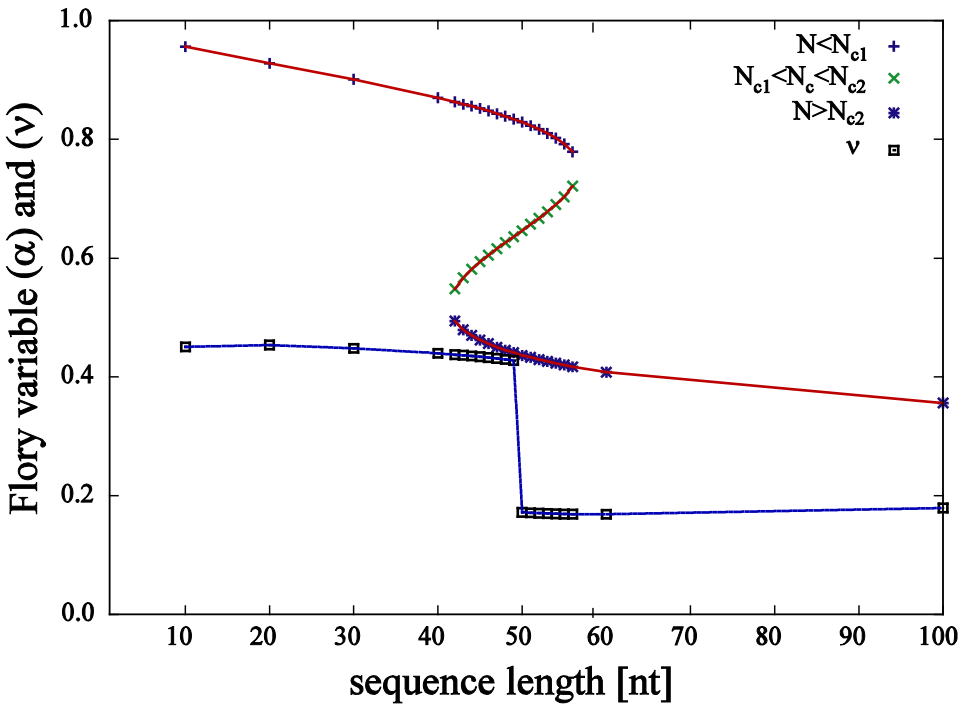

Flory variable (α) and (ν)
sequence length [nt]
N<Nc1
Nc1<Nc<Nc2
N>Nc2
ν
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50
60
70
80
90
100

Figure 4

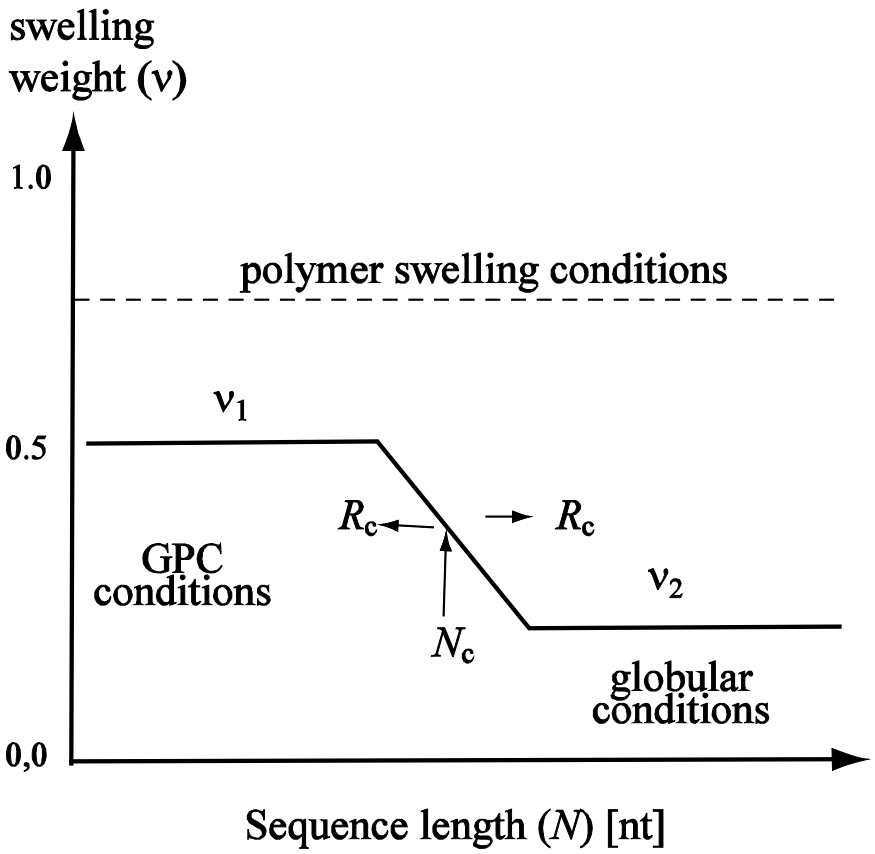
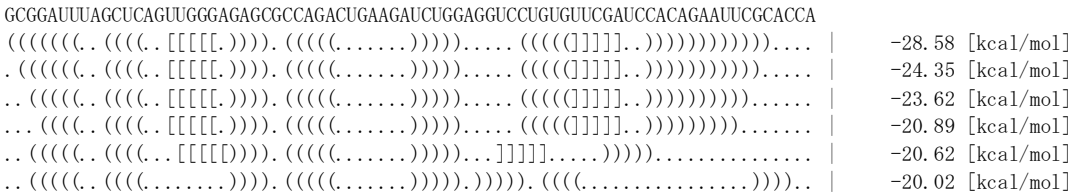

swelling
weight (ν)
1.0
polymer swelling conditions
ν1
0.5
Rc
Rc
GPC
conditions
ν2
Nc
globular
conditions
0,0
Sequence length (N) [nt]

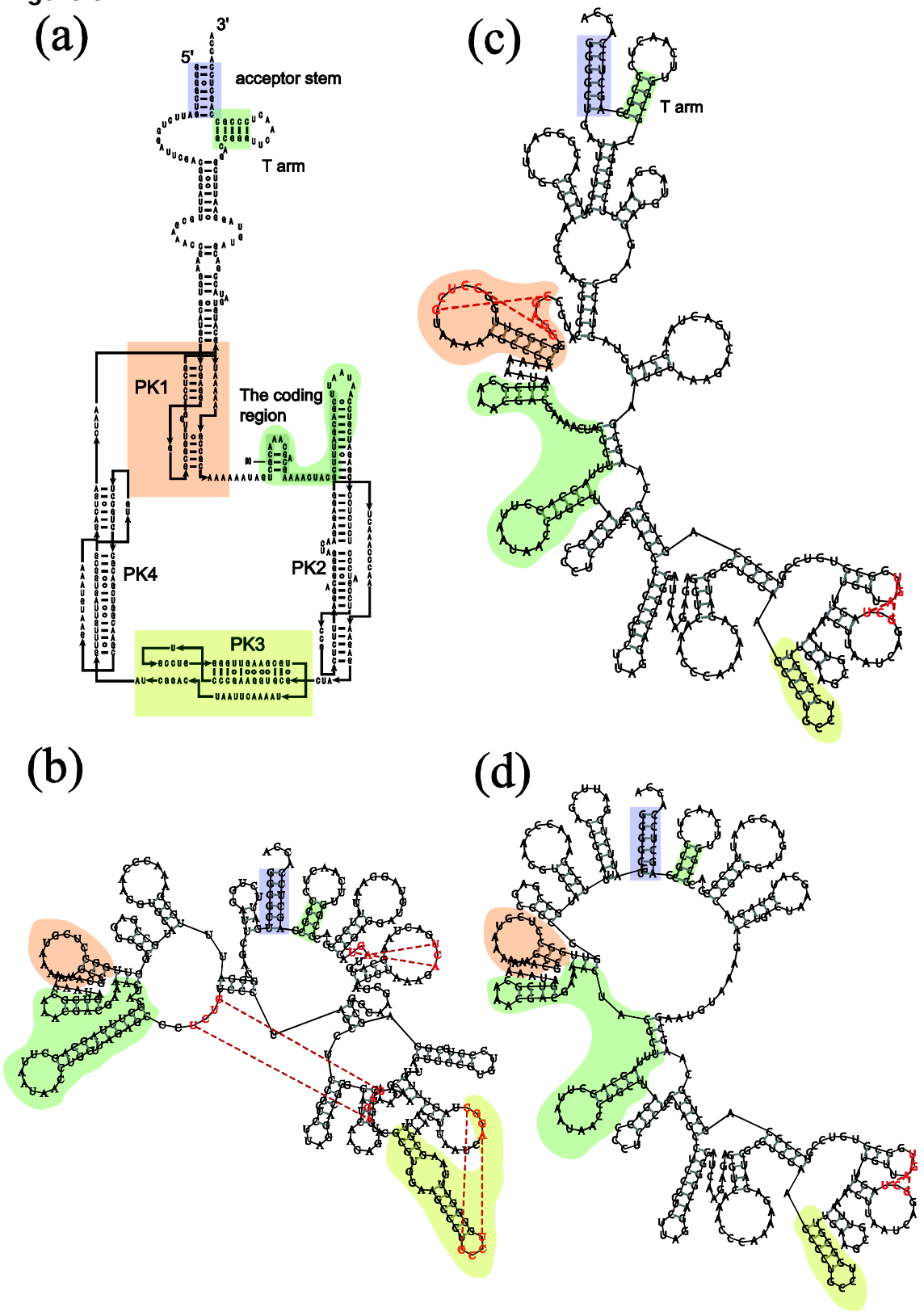
Figure 5

With Flory-Huggins corrections



Without Flory-Huggins corrections

Figure 6


(a)
3'
5'
acceptor stem
T arm
PK1
The coding region
PK4
PK2
PK3
(c)
T arm
(b)
(d)

Tables

Table 1

ν	solvent conditions	characteristics of the polymer	Virial coefficients
0.33	poor	the globular state $r \propto N^{1/3}$	$B_2 < 0$, $B_3 > 0$
0.50	Athermal	Obeys Gaussian statistics $r \propto N^{1/2}$	$B_2 = 0$, $B_3 = 0$
0.60	good	polymer swells $r \propto N^{3/5}$	$B_2 > 0$

Table 2

ξ [nt]	B_2 [nm^3]	B_3 [nm^6]
4.0	-1.1	1.4
5.0	-2.3	5.0
6.0	-4.4	14.0
7.0	-7.4	34.0
8.0	-11.9	75.0
9.0	-18.7	180.0
10.0	-26.5	342.0

Figure Captions

Figure 1

Figure 1. Second virial coefficient (B_2) plotted as a function of Kuhn length (ξ) with parameter conditions $\delta \equiv 2$ and $\gamma \equiv 1.75$. Coupled with B_3 (the third virial coefficient), these parameterizations yield a coexistence region at $N_c \sim 50$ nt and $R_c \sim \pm 10$ nt.

Figure 2

Figure 2. Third virial coefficient (B_3) plotted as a function of Kuhn length (ξ) with parameter conditions $\delta \equiv 2$ and $\gamma \equiv 1.75$. Coupled with B_2 (the second virial coefficient), these parameterizations yield a coexistence region at $N_c \sim 50$ nt and $R_c \sim \pm 10$ nt.

Figure 3

Figure 3. The solution for α with $\xi = 4$ nt in which the coexistence region is observed. Here the coexistence region lies between $N_{c1} \sim 40$ nt and $N_{c2} \sim 60$ nt with a mid-point around $N_c \sim 50$ nt ($R_c = \pm 10$). In this region, the structure exists in two possible states. The effective Flory-Huggins model is meant to approximate the average result of this graph. The corresponding value for ν is also plotted and shows that it lies between 0.0 and 0.5.

Figure 4

Figure 4. A approximate model of the transition of ν as a function of N. Here, N_c is the critical length where the transition between the solvent expanded structure and the globular state occurs. R_c expresses the range of the *coexistence region*: a property of Van der Waals like models such as this. Values for ν at $N < N_c$ and $N > N_c$ are ν_1 and ν_2 respectively. Solved using the standard Flory-Huggins model, these

weights are determined by the virial coefficients B_2 and B_3.

Figure 5

Comparison of the folding landscape of tRNA(Phe) shown with the Flory-Huggins correction and without. Top, structure calculated with Flory-Huggins correction. Bottom, structure calculated without (Same as in Part III).

Figure 6

Example of using the Flory-Huggins (FH) model with tmRNA. (a) The structure of tmRNA for *E. coli*. (b) The predicted structure of this tmRNA sequence using *vsfold5* without the FH model. (c) The predicted structure using the FH model. (d) The predicted structure using the effective FH model. The parameters are the current default values that are used on the web site without any changes. More important to understand is how the FH model tends to make the structures far more compact than expected compared to the standard CLE model.

Table Captions:

Table 1

Table 1. Relationship between solvent conditions, the state of the polymer as expressed by the Flory-Huggins parameter (ν) and the virial coefficients.

Table 2

Table 2. Relationship between Kuhn length and the virial coefficients B_2 and B_3 that yield a coexistence region at $N_c \sim 50$ nt and $R_c \sim \pm 10$ nt. with the fixed conditions $\delta \equiv 2$, and $\gamma \equiv 1.75$. These data are also plotted in Figs 1 and 2.

References

1. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 31: 3406-3415.
2. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Research 31: 3429-3431.
3. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, et al. (1986) Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci U S A 83: 9373-9377.
4. Lu ZJ, Turner DH, Mathews DH (2006) A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. Nucleic Acids Res 34: 4912-4924.
5. Turner DH, Sugimoto N, Freier SM (1988) RNA Structure Prediction. Ann Rev Biophys Biophys Chem 17: 167-192.
6. Jacobson H, Stockmayer W (1950) Intramolecular reaction in polycondensations. I. the theory of linear systems. Journal of Chemical Physics 18: 1600-1606.
7. Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244: 48-52.
8. McKenzie DS (1976) Polymers and scaling. Physics Reports 27C: 35-88.
9. de Gennes P-G (1979) Scaling Concepts in Polymer Physics. Ithaca: Cornell University Press. 324 p.
10. Grosberg AY, Khokhlov AR (1994) Statistical Physics of Macromolecules. New York: AIP Press.
11. Flory PJ (1969) Statistical Mechanics of Chain Molecules. New York: Wiley.
12. Fisher ME (1966) Effect of Excluded Volume on Phase Transitions in Biopolymers. Journal of Chemical Physics 45: 1469-1473.
13. Scheffler IE, Elson EL, Baldwin RL (1970) Helix formation by d(TA) oligomers II. Analysis of the helix-coil trasitions of linear and circular oligomers. J Mol Biol 48: 145-171.
14. Fisher ME (1966) Shape of a Self-Avoiding Walk or Polymer Chain. Journal of Chemical Physics 44: 616-&.
15. Dawson W, Fujiwara K, Kawai G (2007) Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. PLoS One 2: 905.
16. Dawson W, Kawai G (2009) Modeling the Chain Entropy of Biopolymers: Unifying Two Different Random Walk Models under One Framework. J Comput Sci Syst Biol 2: 001-023.
17. Hearst JE, Schmid CW, Rinehart FP (1968) Molecular weights of homogeneous samples of deoxyribonucleic acid determined from hydrodynamic theories for the wormlike chain. Macromolecules 1: 491-394.
18. Thomas TJ, Bloomfield VA (1983) Chain flexibility and hydrodynamics of the B and Z forms of poly(dG-dC).poly(dG-dC). Nucleic Acids Res 11: 1919-1930.
19. Flory PJ (1953) Principles of Polymer Chemistry. Ithaca: Cornell University Press.
20. DeVoe H, Tinoco I, Jr. (1962) The stability of helical polynucleotides: base contributions. J Mol Biol 4: 500-517.
21. Porschke D, Eggers F (1972) Thermodynamics and kinetics of base-stacking interactions. Eur J Biochem 26: 490-498.

22. Manning GS (1978) The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. Q Rev Biophys 11: 179-246.
23. Record MT, Jr., Anderson CF, Lohman TM (1978) Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity. Q Rev Biophys 11: 103-178.
24. Volker J, Klump HH, Manning GS, Breslauer KJ (2001) Counterion association with native and denatured nucleic acids: an experimental approach. J Mol Biol 310: 1011-1025.
25. Ray J, Manning GS (1992) Theory of delocalized ionic binding to polynucleotides: structural and excluded-volume effects. Biopolymers 32: 541-549.
26. Laing LG, Gluick TC, Draper DE (1994) Stabilization of RNA structure by Mg ions. Specific and non-specific effects. J Mol Biol 237: 577-587.
27. Heilman-Miller SL, Pan J, Thirumalai D, Woodson SA (2001) Role of counterion condensation in folding of the Tetrahymena ribozyme. II. Counterion-dependence of folding kinetics. J Mol Biol 309: 57-68.
28. Shcherbakova I, Mitra S, Laederach A, Brenowitz M (2008) Energy barriers, pathways, and dynamics during folding of large, multidomain RNAs. Curr Opin Chem Biol 12: 655-666.
29. Misra VK, Draper DE (1998) On the role of magnesium ions in RNA stability. Biopolymers 48: 113-135.
30. Misra VK, Draper DE (2000) Mg(2+) binding to tRNA revisited: the nonlinear Poisson-Boltzmann model. Journal of Molecular Biology 299: 813-825.
31. Misra VK, Draper DE (2001) A thermodynamic framework for Mg2+ binding to RNA. Proceedings of the National Academy of Science (USA) 98: 12456-12461.
32. Roy KB, Antony T, Sakena A, Bohidar HB (1999) Ethanol-induced condensation of calf thymus DNA studied by laser light scattering. Journal of Physical Chemistry B 103: 5117-5121.
33. Mikulecky PJ, Feig AL (2006) Heat capacity changes associated with nucleic acid folding. Biopolymers 82: 38-58.
34. Gupta A, Mohanty B, Bohidar HB (2005) Flory temperature and upper critical solution temperature of gelatin solutions. Biomacromolecules 6: 1623-1627.
35. Tanford C (1968) Protein denaturation. Adv Protein Chem 23: 121-282.
36. Tanford C (1970) Protein denaturation. C. Theoretical models for the mechanism of denaturation. Adv Protein Chem 24: 1-95.
37. Shortle D (1995) Staphylococcal nuclease: a showcase of m-value effects. Adv Protein Chem 46: 217-247.
38. Bowler BE (2012) Residual structure in unfolded proteins. Curr Opin Struct Biol 22: 4-13.
39. Eisenberg H, Felsenfeld G (1967) Studies of the temperature-dependent conformation and phase separation of polyriboadenylic acid solutions at neutral pH. J Mol Biol 30: 17-37.
40. Inners LD, Felsenfeld G (1970) Conformation of polyribouridylic acid in solution. J Mol Biol 50: 373-389.
41. Achter EK, Felsenfeld G (1971) The conformation of single-strand polynucleotides

in solution: sedimentation studies of apurinic acid. Biopolymers 10: 1625-1634.
42. Pan T, Sosnick TR (1997) Intermediates and kinetic traps in the folding of a large ribozyme revealed by circular dichroism and UV absorbance spectroscopies and catalytic activity. Nature Structural Biology 4: 931-938.
43. Sosnick TR, Barrick D (2011) The folding of single domain proteins--have we reached a consensus? Curr Opin Struct Biol 21: 12-24.
44. Flory PJ (1942) Thermodynamcis of High Polymer Solutions. Journal of Chemical Physics 10: 51-61.
45. Huggins ME (1942) Theory of Solutions of High Polymers. Journal of the American Chemical Society 64: 1712.
46. Dawson W, Kawai G, Yamamoto K (2005) Modeling the long range entropy of biopolymers: A focus on protein structure prediction and folding. Recent Research Developments in Experimental & Theoretical Biology 1: 57-92.
47. Dawson W, Suzuki K, Yamamoto K (2001) A physical origin for functional domain structure in nucleic acids as evidenced by cross linking entropy: part I. J Theor Biol 213: 359-386.
48. Dawson W, Suzuki K, Yamamoto K (2001) A physical origin for functional domain structure in nucleic acids as evidenced by cross linking entropy: part II. J Theor Biol 213: 387-412.
49. Poon BK, Chen X, Lu M, Vyas NK, Quiocho FA, et al. (2007) Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-A crystallographic resolution. Proc Natl Acad Sci U S A 104: 7869-7874.
50. McKenzie D, Moore M (1971) Shape of a self-avoiding walk or polymer chain. J Phys A4: L82-86.
51. 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 99: 3959-3968.
52. Eddy SR (2004) How do RNA folding algorithms work? Nat Biotechnol 22: 1457-1458.
53. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology 288: 911-940.
54. Cormen TH (2001) Introduction to algorithms. Cambridge, Mass.: MIT Press. xxi, 1180 p.
55. Bar A, Kafri Y, Mukamel D (2009) Dynamics of DNA melting. J Phys Condens Matter 21: 034110.
56. Manghi M, Palmeri J, Destainville N (2009) Coupling between denaturation and chain conformations in DNA: stretching, bending, torsion and finite size effects. J Phys Condens Matter 21: 034104.
57. Einert TR, Netz RR Theory for RNA folding, stretching, and melting including loops and salt. Biophys J 100: 2745-2753.
58. David F, Hagendorf C, Wiese KJ (2007) Random RNA under tension. Epl 78.
59. Li PT, Tinoco I, Jr. (2009) Mechanical unfolding of two DIS RNA kissing complexes from HIV-1. J Mol Biol 386: 1343-1356.
60. Li PT, Bustamante C, Tinoco I, Jr. (2007) Real-time control of the energy landscape by force directs the folding of RNA molecules. Proc Natl Acad Sci U S A 104: 7039-7044.

61. Moffitt JR, Chemla YR, Smith SB, Bustamante C (2008) Recent advances in optical tweezers. Annu Rev Biochem 77: 205-228.
62. Tinoco I, Jr. (2004) Force as a useful variable in reactions: unfolding RNA. Annu Rev Biophys Biomol Struct 33: 363-385.
63. Pollack L (2011) Time resolved SAXS and RNA folding. Biopolymers 95: 543-549.
64. Anthony BL, Caston RH, Guth E (1942) Equations of state for natural and synthetic rubber-like materials. I Unaccelerated natural soft rubber. J Phys Chem 46: 826-840.
65. Elliott DR, Lippmann SA (1945) The thermodynamics of rubber at small extensions. Journal of Applied Physics 16: 50-54.
66. Liphardt J, Onoa B, Smith SB, Tinoco I, Jr., Bustamante C (2001) Reversible unfolding of single RNA molecules by mechanical force. Science 292: 733-737.
67. Tinoco I, Uhlenbeck OC, Levine MD (1971) Estimation of secondary structure in ribonucleic acids. Nature 230: 362-367.
68. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M (1994) Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie 125: 167-188.
69. Jaeger JA, Turner DH, Zuker M (1989) Improved predictions of secondary structures for RNA. Proc Natl Acad Sci (USA) 86: 7706-7710.
70. Xayaphoummine A, Bucher T, Isambert H (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. Nucleic Acids Res 33: W605-610.
71. Cao S, Furtig B, Schwalbe H, Chen SJ (2010) Folding kinetics for the conformational switch between alternative RNA structures. J Phys Chem B 114: 13609-13615.
72. Cao S, Chen S-J (2006) Predicting RNA pseudoknot folding thermodynamics. Nucleic Acids Research 34: 2634-2652.
73. Xia T, SantaLucia J, Jr., Burkard ME, Kierzek R, Schroeder SJ, et al. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37: 14719-14735.
74. SantaLucia J, Jr., Hicks D (2004) The thermodynamics of DNA structural motifs. Annu Rev Biophys Biomol Struct 33: 415-440.
75. Xayaphoummine A, Bucher T, Thalmann F, Isambert H (2003) Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. Proceedings from the National Academy of Science (USA) 100: 15310-15314.
76. Govorun EN, Khokhlov AR, Semenov AN (2003) Stability of dense hydrophobic-polar copolymer globules: Regular, random and designed sequences. European Physical Journal E 12: 255-264.
77. Chan S-C, Dill KA (1997) Solvation: how to obtain macroscopic energies from partitioning and solvation experiments. Annual Review Biophysics and Biomolecular Structucture 26: 425-459.
78. Plischke M, Bergersen B (1994) Equilibrium Statistical Physics. Englewood Cliffs.
79. Levine IN (2003) Physical Chemistry. Singapore: Mc Graw-Hill. 966 p.
80. Muller M, Krzakala F, Mezard M (2002) The secondary structure of RNA under tension. Eur Phys J E Soft Matter 9: 67-77.

81. Müller M (2003) Statistical physics of RNA folding. Phys Rev E Stat Nonlin Soft Matter Phys 67: 021914.
82. Heilman-Miller SL, Thirumalai D, Woodson SA (2001) Role of counterion condensation in folding of the ¥textit{tetrahymena} ribozyme. I. Equilibrium stabilization by cations. J Mol Biol 306: 1157-1166.
83. Chauhan AK, Apirion D (1989) The gene for a small stable RNA (10Sa RNA) of Escherichia coli. Mol Microbiol 3: 1481-1485.
84. Nameki N, Chattopadhyay P, Himeno H, Muto A, Kawai G (1999) An NMR and mutational analysis of an RNA pseudoknot of Escherichia coli tmRNA involved in trans-translation. Nucleic Acids Res 27: 3667-3675.
85. Lassig M, Wiese KJ (2006) Freezing of random RNA. Phys Rev Lett 96.
86. David F, Wiese KJ (2007) Systematic field theory of the RNA glass transition. Phys Rev Lett 98: 128102.
87. Woodson SA (2000) Compact but disordered states of RNA. Nat Struct Biol 7: 349-352.
88. Sosnick TR, Pan T (2004) Reduced contact order and RNA folding rates. J Mol Biol 342: 1359-1365.
89. Osterberg R, Sjoberg B, Garrett RA (1976) Molecular model for 5-S RNA. A small-angle x-ray scattering study of native, denatured and aggregated 5-S RNA from Escherichia coli ribosomes. Eur J Biochem 68: 481-487.
90. Cole PE, Yang SK, Crothers DM (1972) Conformational changes of transfer ribonucleic acid. Equilibrium phase diagrams. Biochemistry 11: 4358-4368.
91. Friederich MW, Hagerman PJ (1997) The angle between the anticodon and aminoacyl acceptor stems of yeast tRNA(Phe) is strongly modulated by magnesium ions. Biochemistry 36: 6090-6099.
92. Coutts SM, Gangloff J, Dirheimer G (1974) Conformational transitions in tRNA Asp (brewer's yeast). Thermodynamic, kinetic, and enzymatic measurements on oligonucleotide fragments and the intact molecule. Biochemistry 13: 3938-3948.
93. Maglott EJ, Deo SS, Przykorska A, Glick GD (1998) Conformational transitions of an unmodified tRNA: implications for RNA folding. Biochemistry 37: 16349-16359.
94. Nacheva GA, Guschin DY, Preobrazhenskaya OV, Karpov VL, Ebralidse KK, et al. (1989) Change in the pattern of histone binding to DNA upon transcriptional activation. Cell 58: 27-36.
95. Devkota B, Petrov AS, Lemieux S, Boz MB, Tang L, et al. (2009) Structural and electrostatic characterization of pariacoto virus: implications for viral assembly. Biopolymers 91: 530-538.