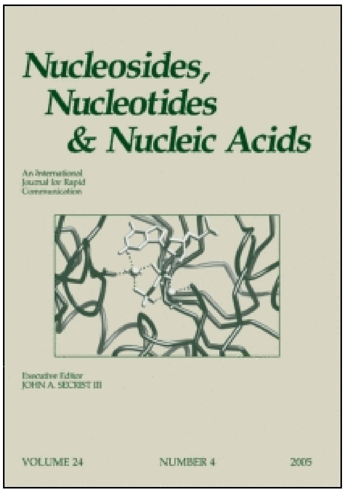

Nucleosides, Nucleotides and Nucleic Acids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

A method for finding optimal RNA secondary structures using a new entropy model (vsfold)

Wayne Dawson[a]; Kazuya Fujiwara[a]; Gota Kawai[a]; Yasuhiro Futamura[b]; Kenji Yamamoto[b]
[a] Chiba Institute of Technology, Tsudanuma, Narashino-shi, Chiba, Japan [b] International Medical Center of Japan, Toyama, Shinjuku-ku, Tokyo, Japan

To cite this Article Dawson, Wayne , Fujiwara, Kazuya , Kawai, Gota , Futamura, Yasuhiro and Yamamoto, Kenji(2006) 'A method for finding optimal RNA secondary structures using a new entropy model (vsfold)', Nucleosides, Nucleotides and Nucleic Acids, 25: 2, 171 — 189

To link to this Article: DOI: 10.1080/15257770500446915

URL: http://dx.doi.org/10.1080/15257770500446915

Nucleosides, Nucleotides, and Nucleic Acids, 25:171–189, 2006
Copyright © Taylor & Francis Group, LLC
ISSN: 1525-7770 print / 1532-2335 online
DOI: 10.1080/15257770500446915

A METHOD FOR FINDING OPTIMAL RNA SECONDARY STRUCTURES USING A NEW ENTROPY MODEL (VSFOLD)

Wayne Dawson, Kazuya Fujiwara, and Gota Kawai □ *Chiba Institute of Technology, Tsudanuma, Narashino-shi, Chiba, Japan*

Yasuhiro Futamura and Kenji Yamamoto □ *International Medical Center of Japan, Toyama, Shinjuku-ku, Tokyo, Japan*

□ *We are developing a program to calculate optimal RNA secondary structures. The model uses di-nucleotide pairing energies as with most traditional approaches. However, for long-range entropy interactions, the approach uses an entropy-loss model based on the accumulated sum of the entropy of bonding between each base-pair weighted inversely by the correlation of the RNA sequence (the Kuhn length). Stiff RNA forms very different structures from flexible RNA. The results demonstrate that the long-range folding is largely governed by this entropy and the Kuhn length.*

Keywords RNA secondary structure; Free energy; Cross linking entropy; Kuhn length; Folding; Renormalization theory

INTRODUCTION

The RNA world is rapidly expanding with a disproportionate number of non-coding RNA sequences.[1,2] It is increasingly important to develop and improve our prediction abilities to keep up with this growth. In considering how to improve the predictive capacity of RNA secondary structure prediction using thermodynamics, little has been done to examine the assumptions behind the penalties used to estimate the entropy-loss due to folding of RNA. Here we describe some recent encouraging results of this approach.

There are many web servers that provide a variety of structure prediction capabilities.[3] The largest and probably most successful are the knowledge based methods.[4–9] However, all such methods are based on

Received 25 August 2005; accepted 18 October 2005.

This work was supported by, in part, a grant-in-aid for high technology research from the ministry of Education, Science, Sports and Culture, Japan.

Address correspondence to Wayne Dawson, Department of Life and Environmental Sciences, Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino-shi, Chiba 275-0016, Japan. E-mail: wayne.dawson@it-chiba.ac.jp

obtaining knowledge of known secondary structures. The second approach involves some approach connected with thermodynamics.[3,10–15] This latter approach does not require known structures to construct a secondary structure, only a good set of fundamental parameters and a solid theory to support it.

Here we report a new way to estimate the entropy-loss using a model we call the "cross-linking entropy" (CLE) model.[14] The CLE is a function of the *accumulated sum* from the independent formation of base-pairs and inversely weighted by the local correlation in the RNA sequence. This new approach is relatively tolerant of the uncertainties in the base-stacking parameters, is generalizable to pseudoknot problems and permits an expanded set of non-Watson-Crick (non-WC) base pairs.[4] It is also generally consistent with the hierarchal folding hypothesis (see Tinoco and Bustanante,[16] and references therein).

THEORY

A detailed description of the theory will be discussed elsewhere. The foundations of the theory used here are discussed in considerable detail in Dawson et al.[14] Brief derivations for the equations in this section can be found in the Appendices.

RNA Secondary Structure Definitions

First we briefly describe RNA secondary structure. Given an RNA sequence of N nucleotides (bases) with the position of each base indexed from 1 (at the $5^{'}$ end) to N (at the $3^{'}$ end), a base pair (bp) formed between two such nucleotides i and $j\,(1 \leq i < j \leq N)$ is expressed here by the ordered pairs (i, j). In a single sequence, the formation of a bp also creates a loop and a group of contiguous bps forms a stem, where a "stem" should consist of at least two bps (Figure 1). Loops consist of hairpin loops (H-loops: Figure 1a), bulges (Figure 1b), interior loops (I-loops: Figure 1c), and multibranch loops (MBL: Figure 1d). A collection of these stems and loops forms a domain. A more detailed description can be found in Dawson et al.[14]

The $5'$ and $3'$ most end of a stem is called the "tail" and denoted by the ordered pair (p_t, q_t) and the other end of the stem is referred to as the "head" and denoted by the ordered pair (p_h, q_h): see Figure 1e. A necessary condition for any ordered pair of nucleotides in a stem is that $p_h > p_t$, $q_t > q_h$ and $(q_t - q_h) \equiv (p_h - p_t) \geq 1$. (A generalization of this concept is shown in Figure 1f.)

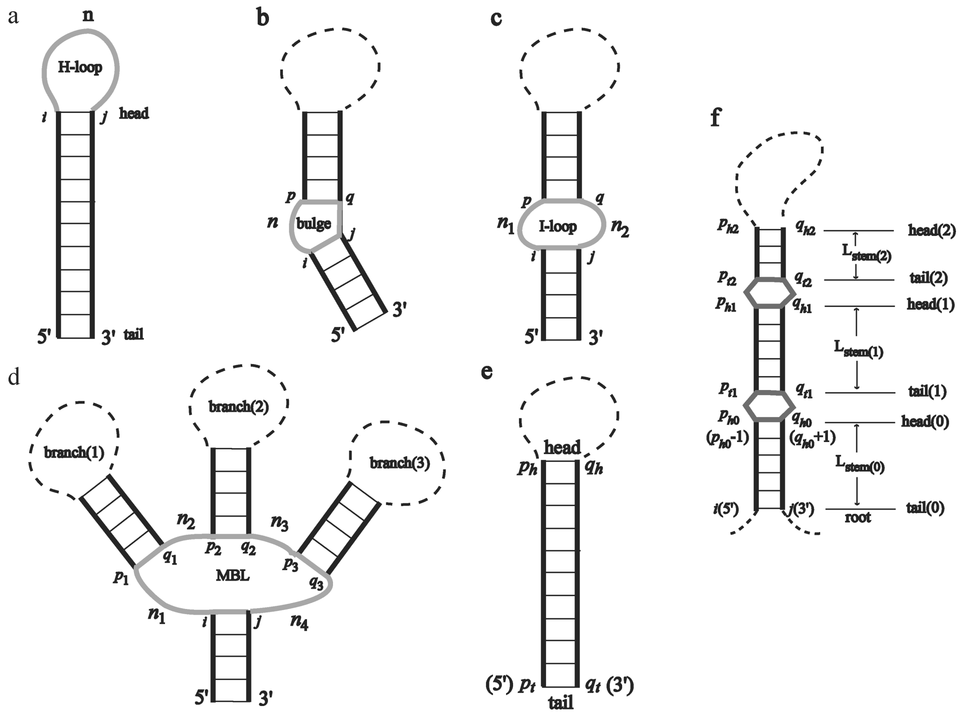

FIGURE 1 Schematics of secondary structure: (a) hairpin loop (H-loop), (b) bulge, (c) interior loop (I-loop), (d) multiloop or multibranch loop (MBL), and (e) the definition of head and tail on a stem. The thick black lines with thin black cross-hatching indicate the base pairs (stems), the thick gray lines indicate regions of various types of loops (see labels), and the dotted lines indicate regions where additional unspecified secondary structure may be present. The indices are discussed in the text. (f) The general features of an effective stem and the hierarchy of stems it comprises with respect to a given origin (i,j). The I-loops are a dark gray in Figure 1f because the stems are contiguous and therefore the real physical distinction between "interior loop" and "stem" is strongly blurred.

The Cross-Linking Entropy-Loss Calculation Method

The model we use here is called cross-linking entropy (CLE).[14] The CLE model assigns entropy-loss due to base pair formation. In addition, the CLE approach requires understanding of the concept of a persistence length (ξ), or, more precisely, a Kuhn length. "Persistence" is a measure of the local correlation between adjacent monomers of RNA (A,C,G and U) on a single chain of RNA. Locally, these monomers (or mers) tend to correlate over short ranges. As a result, one should not think of RNA folding as that of individual mers, but rather as groups of mers that fold together.

To illustrate this concept, we show (in Figures 2a and b) a set of beads (blue) forming the chain of mers. Overlaid on top of these beads are rods (orange) that approximate the distance wherein the beads tend to be roughly straight. In Figure 2a, the length scale of the group is about 4 mers. In

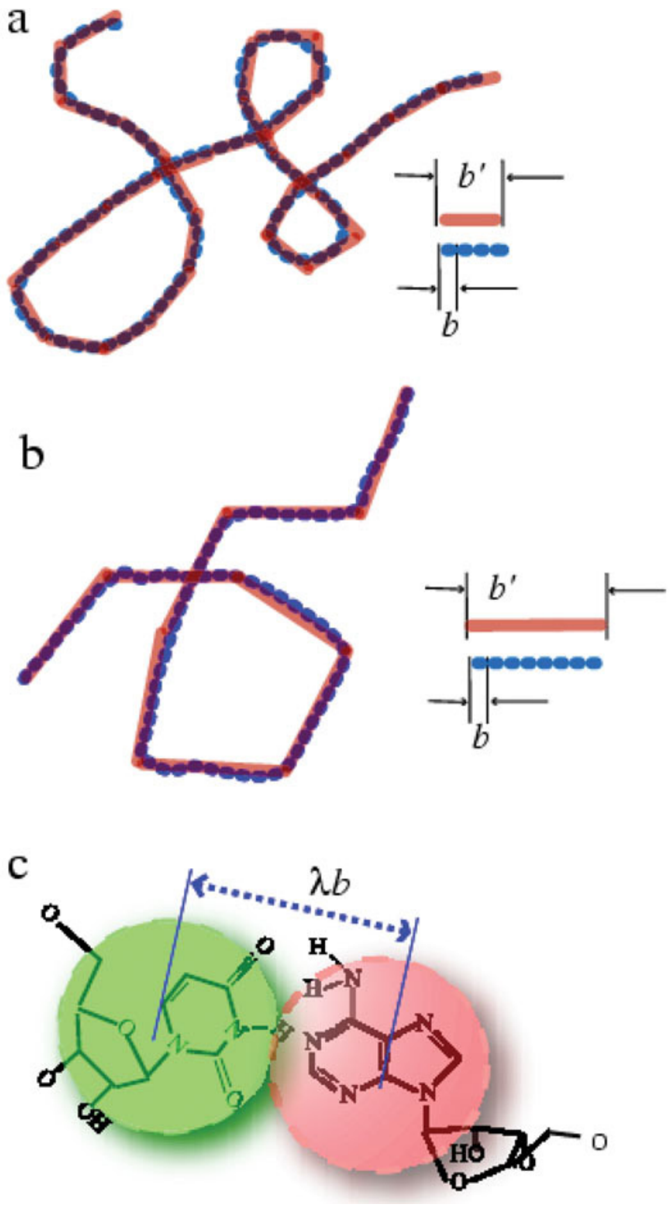

FIGURE 2 (a) An example of a random walk where each step is expressed by a blue bead shaped dot whose bead-to-bead separation distance is b, and an "effective step" (the Kuhn length of each step) is indicated by the transparent orange bars of length b'. In transforming this random walk to a structure for a randomly oriented polymer chain, the blue beads correspond to the monomers and the orange bars to the "effective mers" or the Kuhn length of the mers. (b) Same as Figure 2a, only here the Kuhn length b' is longer. Note that the structure is more spread out and does not fold up as tightly as it did in Figure 2a. (c) A schematic of how the base pair stacking separation distance (λb) is determined between nucleotides in an RNA sequence.

Figure 2b, that length scale has extended to a group of 8 mers. The chain of mers in Figure 2a forms a much tighter structure than the structure in Figure 2b because the flexible rods can turn on their joints to form a more compact structure. The orange rods are what we call "effective mers" because, in general, the actual characteristics of the polymer on a global scale are measured by the Kuhn length (ξ), not the individual mers. Again, this is discussed in far more detail in Dawson et al.[14]

When the CLE model is greatly simplified, it can be shown to be consistent with the constants that are used in the traditional entropy models. Most of this is discussed in detail in Dawson et al.[14] The only new parameter is therefore the Kuhn length (ξ).

In the CLE model, the entropy-loss due to bp formation becomes Dawson et al.[14] and Appendix C)

$$\Delta S(N_{ij}) = -\frac{k_B}{\xi}\{\gamma \ln(\Psi_{1/2} N_{ij}) - (\gamma + 1/2)[1 - 1/(\Psi_{1/2} N_{ij})]\} \tag{1}$$

where N_{ij} represents the number of monomers between base i and base j (with $j > i$ and $1 \leq i < j \leq N$). Also, k_B is the Boltzmann constant (1.98 cal/mol) and $\gamma (= 1.75)$ is a weight that approximates the statistical characteristics of a self-avoiding random-walk where the walker must avoid points that have already been crossed in previous steps. Finally, $\Psi_{1/2} = \xi/\lambda^2$, where λ describes the distance between the bases (see Figure 2c and Appendix C).

The total entropy loss is evaluated by summing up the contributions from each base pair corresponding to the indices i and j

$$\Delta S_{cle} = \Delta S_{\gamma\xi} + \sum_{hp(ij)} \Delta S_{bp}(N_{ij}) \tag{2}$$

where we emphasize that Eq. (2) is summed over *all* bps of a given structure using Eq. (1) and the local entropy is

$$\Delta S_{\gamma\xi} = -\left(\frac{N}{\xi}\right)\frac{(\gamma + 1/2)k_B}{\omega}\int_{+1}^{\xi}\left\{\frac{\ln(x)}{1-x} + 1\right\}dx \tag{3}$$

where ω indicates the dimensionality (usually $\omega = 3$). Equation (3) accounts for the fact that we are evaluating rods of length ξ rather than monomers (of length $\xi = 1$ nt); see Appendix E.

There should also be interactions between the unstacked bases in the loop structures because the bases are still in greater correlation than they would be if they were completely free; loops like stems are also structural elements. The loop formation entropy for MBLs, I-loops and bulges is calculated in a similar way as above. From Appendix D, an approximate

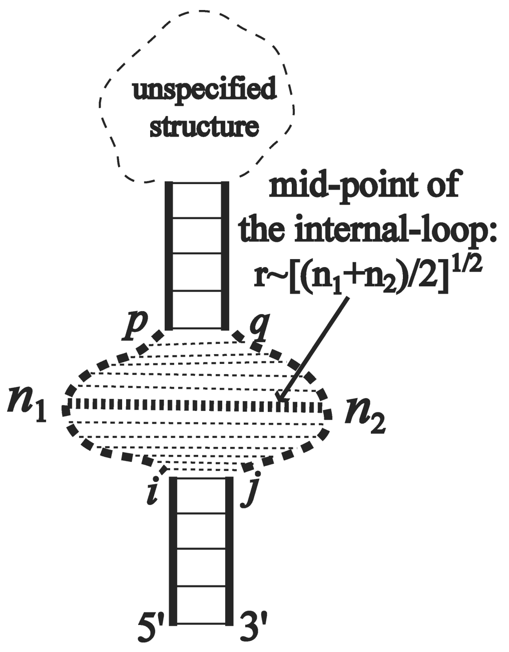

FIGURE 3 A qualitative description of the correlation interactions between the free strands in an interior loop. The dotted lines indicate weak correlation interactions between chains 1 and 2; the maximum rms separation-distance between the chains is found at the center of the two chains (the black dashed line). Although the chain regions are far freer than the stems, they are still more localized than would be the case if the entire sequence (from 5′ to 3′) were completely unstructured.

formula is

$$\begin{aligned}&-T\Delta S(N_{ij}, n, s, \lambda)\\&\quad\approx \frac{w[n+(s-1)\lambda]k_BT}{2\xi}\left[\gamma \ln\left(\frac{2N_{ij}}{n+(s+1)\lambda}\right)\right.\\&\quad\left.-(\gamma+1/2)\left(1-\frac{n+(s+1)\lambda}{2N_{ij}}\right)\right]\end{aligned} \tag{4}$$

where w is a weight, N_{ij} is the enclosed length from position (i, j), and s is the number of stems. For an I-loop, $s = 1$. For a MBL of k-branches, $s = k$. Hence, from Figures 1c and 3 (the I-loop), $n = n_1 + n_2$ where n_1 and n_2 are the lengths of the two single-strand segments ($n_1 = p - i - 1$, and $n_2 = j - q - 1$). For an MBL with three branches (Figure 1d), $s = 3$ and $n = n_1 + \cdots + n_4 = \sum_{k=1}^{s+1} n_k$.

METHODS

Implementation

Calculations are done on a Mac OSX 10.3 and LINUX Red Hat version 9.3 and Fedora core operating systems. The program is written in C and compiles on most unix operating systems including Sun, SGI, and DEC/HP machines.

RNA Secondary Structure Calculations

In this work, the parameters used in the vsfold evaluations are all identical to the standard thermodynamic models[17] except that vsfold introduces the Kuhn length (ξ). Calculations are done to the same level of detail as these traditional programs; however, vsfold rejects isolated base pairs and absurdly short stems as a rule. Vsfold currently cannot evaluate suboptimal structure,[18] coaxial stacking,[19] or pseudoknots (see Dowell and Eddy,[7] and references therein), although these are planned in future revisions. Other options such as adaptation of Flory's polymer-solvent interaction model or including Mg interactions can be selected but are not discussed here.

RESULTS AND DISCUSSION

We applied the CLE approach to some familiar RNA structures: Figures 4 and 5. Calculations using the CLE model were carried out using vsfold version 4.11 (http://www.rna.it-chiba.ac.jp/vsfold4) using the mfold 3 Turner energy parameters (the most recent).[4,17]

Examples for tRNA

All of the secondary structure predictions for the tRNAs shown in Figure 4 are essentially perfect with all the correct base pairs and the t-shape structure.

The structure of tRNA(phe) (Figure 4a) has been studied using numerous approaches. All studies on tRNA(phe) indicate that the t-shaped secondary structure is quite stable in a broad range of solvent environments. The unmodified transcript has also been studied and shows no evidence of losing the t-shape in low Mg conditions under a wide range of monovalent salt conditions.[20] Although the line widths broaden and some peaks are lost in the NMR spectrum, there is sufficient evidence that the structure remains with the same approximate base pairing[20,21] and aminoacylation still occurs,[22] although at considerably reduced levels. The t-shape is present in high salt conditions (1M), indicating that we should expect this structure with the Turner base-stacking parameters.

Using the CLE-model, this structure is so stable that it fits over a wide range of Kuhn lengths ($3 \leq \xi \leq 5$ nt). This amounts to a rather large difference in the global entropy contribution (about 1/3, or 5 kcal/mol [20 kJ/mol]) that is independent of the base-stacking parameters.

In many respects, the Kuhn length in folded RNA structures appears to be primarily a function of the stem length. Since tRNA has stems of length

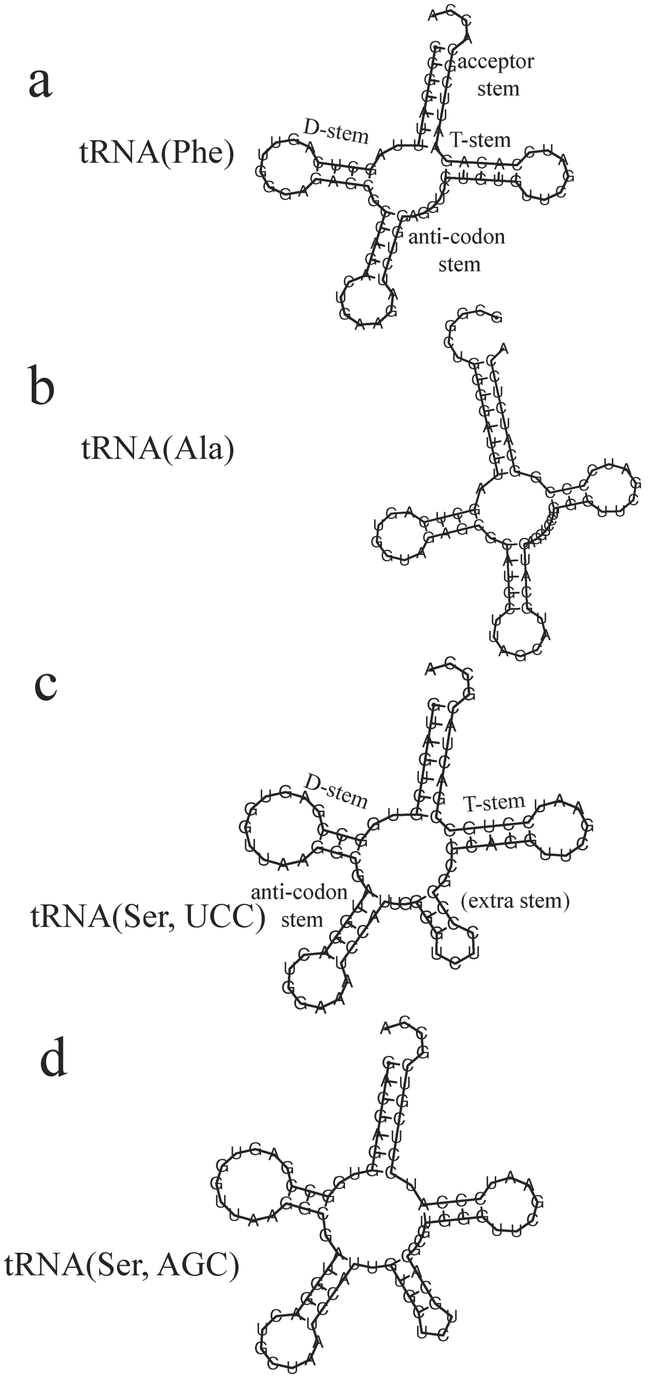

FIGURE 4 Results of tRNA secondary structures using the CLE-model. (a) Optimal secondary structure prediction of tRNA(phe) for *E. coli*. (b) Optimal secondary structure prediction of tRNA(ala). (c) Optimal secondary structure prediction of tRNA(Ser) corresponding to codon UCC. (d) Optimal secondary structure prediction of tRNA(Ser), codon AGC.

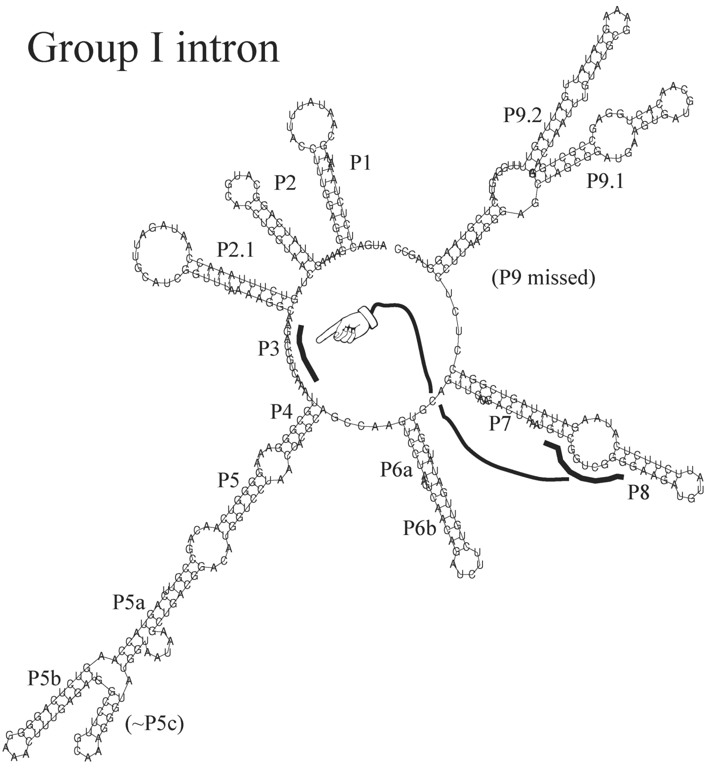

FIGURE 5 Results of the optimal secondary structure of the *Tetrahymena thermophila* group I intron (the L-21 Scal ribozyme) using the CLE model. The arrow in the CLE predicted structure indicates a later step in the hierarchal folding process where the pseudoknot stem (P3) is formed.

approximately 4 bps, it should come as no surprise that ξ would yield a reasonable value in these calculations. Although the CLE-model does not regularly perform with such astounding stability, it is consistent with the often asserted view in protein folding that the FE landscape should be funnel shaped[23] and is a convenient demonstration of the surprising robustness of this model.

Currently, our own algorithm does not support coaxial stacking. Therefore, we are presently forced to benchmark our results without this option. We expect that coaxial stacking is important in stabilizing secondary structure.[19] However, according to the hierarchical folding hypothesis,[16] tertiary structure effects (which applies to coaxial stacking) should only serve to further stabilize an already established RNA secondary structure. The structure in Figure 4a is already made without coaxial stacking and is now poised for further stabilization by the addition of coaxial stacking. This is consistent with the general observation that tertiary structure should form after the major secondary structure has formed.[16] Experimental evidence indicates that, in the absence of Mg^{2+} even at ionic strengths up to 2M Na^{+}, the flexibility of the MBL region and the angle between the acceptor stem and the anticodon stem changes drastically (about 70°) without any loss of secondary structure in tRNA(phe).[20] The coaxial angle between the T-loop and the acceptor stem was not measured directly; however, the strong Mg^{2+} dependence on this angle and the localization of $[Mg(H_2O)_6]^{2+}$ in tRNA(phe) both suggest that the tertiary structure may be affected. Further citations within Friederich et al.[20] suggest other examples where the tRNA class of structures may lose tertiary structure while retaining an unchanged secondary structure under a variety of conditions.

All the experimental evidence on tRNA(phe) suggests that we should expect the framework of the secondary structure to be present independent of coaxial stacking or any other tertiary structure contributions. Thus, we are in a far better position to make estimates of the intrinsic tertiary structure contributions to the FE using vsfold.

In the case of tRNA(ala) (Figure 4b), the structure is not quite as thoroughly studied as tRNA(phe). Nevertheless, it takes on the t-shape in the X-ray crystallography and all the stems are present in the NMR measurements. Structures like tRNA are "passive" biological entities and there is little reason we should expect them to take on structures other than the t-shape without introducing extreme conditions. Moreover, typical metastable states do not usually crystallize. The same structure is obtained over a range of Kuhn lengths (also $3 \leq \xi \leq 5$ nt). Base modifications here should only improve the stability.

In the case of Figures 4c and d, rat-tRNA(Ser, UCC) and rat-tRNA(Ser, AGC) respectively, both sequences are similar but not identical. An additional branch is observed in the prediction; however, the indicated D-loop, T-loop, and anticodon stems are all consistent with known consensus structure of tRNA.[24] Base modifications for these two structures can be found at http://www.uni-bayreuth.de/departments/biochemie/trna under the accession numbers RS9160 and RS9161. Modifications appear on the D-stem, T-stem, and the anticodon stem. However, these particular modifications are generally thought to further stabilize the tRNA.[25]

Using the CLE method with unmodified sequences is important in of itself. First, the bulwark of tRNA modifications are posttranscriptional.[25] Hence, when the CLE method fails to predict the correct tRNA structure from the unmodified sequence, this can help quickly identify targets where the post-transcriptional modifications may play a "keystone" role. Second, the most common of modifications to tRNA are observed to *stabilize* the final structure.[25] With the CLE information, we now can accurately evaluate the extent to which base modification "temper" or "buttress" the nascent tRNA structure. Whether a base modification serves as the "principal load-bearing support beam" for the structure or simply buttresses that structure, target structures can be quickly identified and studied.[25] Finally, the secondary structure is arguably the most important means whereby the post transcription machinery identifies the base-modification sites and maintains the fidelity of this process. Knowing what structures this machinery actually recognizes also identifies the important structural targets.

A relatively robust prediction method should be tolerant to these common and recognizable variations in sequence. The FE landscape should be funnel shaped and one should expect that minor variations in the sequence will have only a minor effect on the structure at most; particularly for functional RNA where we already know these variations exist and do not change the structure significantly.

Example of the Group I Intron

With the group I intron (Figure 5) using a fixed Kuhn length throughout the entire structure, we obtain some problems with the $3'$ end. The size and structure of the domains with the exception of P9 are all nearly close to what is expected of this structure.[26,27] When the sequence is then divided between P5 and P6 and ξ in the P6-P9 region is reduced to from $\xi = 10$ to $\xi = 9$ nt, the correct structure is obtained for P9 suggesting that the latter half is more flexible. This later point is significant because ξ is a measure of the flexibility of the RNA. One can see that most of the stems in the group I intron are at least 10 bp long. Hence a long Kuhn length is quite reasonable. Current secondary structure prediction programs offer no advice about the flexibility of the RNA they search. Moreover, the fact that the optimal structure can be found so easily with the CLE-model with an inaccurate monolithic Kuhn length is a sound indication that the CLE-model is a better way to approach this problem.

Persistence Length/Kuhn Length Dependence

To illustrate the importance of Kuhn length in calculations of secondary structure using vsfold4, we compare the structure of the P4-P5 stem of the group I intron with two different ranges of Kuhn lengths (Figure 6). The P5 region is known to fold up very rapidly during refolding experiments.[16,28] The remaining structure begins to form after Mg addition. Divalent cations would have the tendency to "harden" the nascent stems. The P4-P5 stem structure serves largely as a scaffold for the rest of the group I intron's structure. With $\xi = 10$ nt, the P4-P5 segment is very stiff.

The predictions using the CLE model are clearly not perfect, and one of the clearest problems is that ξ is assumed to be constant throughout the sequence. In retrospect, variable flexibility should not have come as a big surprise. Nevertheless, the predictive capacity of vsfold 4.11 was surprisingly good, at least *in these examples.* Using an MFOLD suboptimal structure reevaluation program, we showed that similar suboptimal structures of rRNA and the group I intron tended to group or cluster together in Dawson

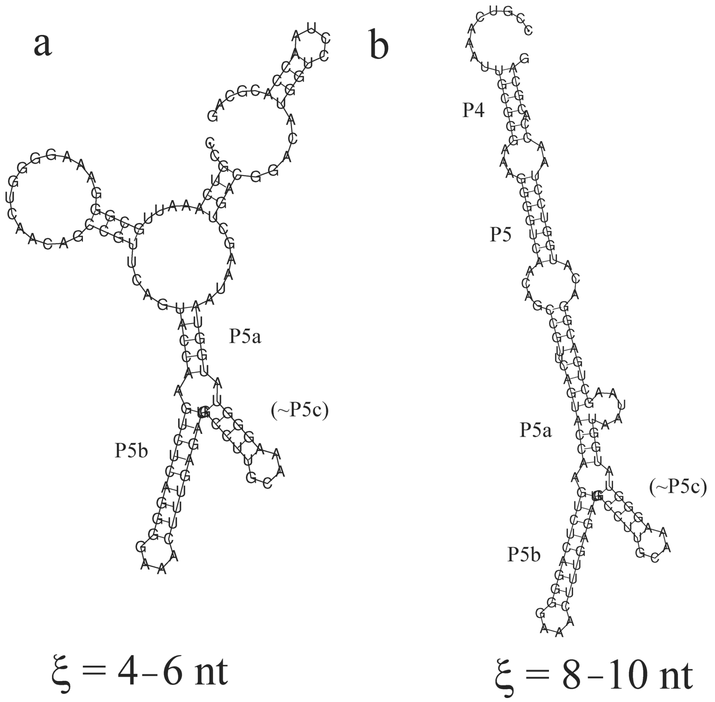

FIGURE 6 Comparison of the structure of the P4-P5 stem of the group I intron (shown in Figure 5) for different Kuhn lengths (ξ):(a) when $\xi = 4$ nt, and (b) when $\xi = 8$ nt.

et al.,[14] where these same structures were scattered hither thither and yon throughout the connect files in the original mfold 2.3 calculation. “Grouping” is characteristic of what one should expect of a funnel-shaped FE-landscape of secondary structures. Developing a partition function model and allowing for suboptimal structures should only improve these results.

The main weakness in this model is that we have little or no information on how to independently decide what the Kuhn length should be. Currently, the user must make a decision about the character of the RNA; *i.e.*, the user must decide if the RNA is stiff or flexible. A guideline would be from the function of the RNA. Stiff RNA is typically involved in scaffolding, whereas flexible RNA would be involved in function or recognition. The context dependence of the Kuhn length is beyond the scope of the current work. Nevertheless, adding the concept of a Kuhn length (persistence length) to RNA structure prediction is quite instructive in understanding the character of RNA in general.

Future Plans

Future plans are to adapt the model to pseudoknot applications (see Dowell and Eddy[7] and references therein), introduce a variable Kuhn length, expand the polymer-solvent capabilities of the model, introduce suboptimal structures, introduce partition function calculations, and work on speeding up the algorithm, which is currently very inefficient. The effects of correlation extending beyond the range of the Kuhn length in the folding process have been ignored and we appear to have evaded any serious repercussions; however, this is an area that needs to be looked into further. Finally, the universality condition of this model needs to be examined rigorously.

REFERENCES

1. Huttenhofer, A.; Schattner, P.; Polacek, N. Non-coding RNAs: Hope or hype? Trends Genet. **2005**, 21, 289–297 (review).
2. Storz, G. An expanding universe of noncoding RNAs. Science **2002**, 296, 1263 (review).
3. Fox, J.A.; Butland, S.L.; McMillan, S.; Campbell, G.; Ouellette, B.F.F. The Bionformatics links directory: A compilation of molecular biology web servers. Nucleic Acids Res. **2005**, 33, W3–W24
4. Mathews, D.H.; Turner, D.H. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. J. Mol. Biol. **2002**, 317, 191–203.
5. Tahi, F.; Gouy, M.; Régnier, M. Automatic RNA secondary structure prediction with comparative approach. Comp. and Chem. **2002**, 26 521–530.
6. Knudsen, B.; Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. **2003**, 31, 3423–3428.
7. Dowell, R.D.; Eddy, S.R. Evaluation of several lightweight stochastic conext-free grammars for RNA secondary structure prediction. BMC Bioinformatics **2004**, 5, 71.
8. Hu, Y.-J. GPRM: A genetic programming approach to finding common RNA secondary structure elements. Nucleic Acids Res. **2003**, 31, 3446–3449.
9. Havgaard, J.H.; Lyngso, R.B.; Gorodkin, J. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. Nucleic Acids Res. **2005**, 33, W650–W653.
10. Hofacker, I.L. Vienna RNA secondary structure server. Nucleic Acids Res. **2003**, 31, 3429–3431.
11. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids. Res. **2003**, 31, 3406–3415.
12. Ding, Y.; Chan, C.-Y.; Lawrence, C.E. Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res. **2004**, 32, W135–W141.
13. Xayaphoummine, A.; Bucher, T.; Isambert, H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. Nucleic Acids Res. **2005**, 33 W605–W610.
14. Dawson, W.; Suzuki, K.; Yamamoto, K. A physical origin for functional domain structure in nucleic acids as evidenced by cross-linking entropy. J. Theor. Biol. 213 **2001**, 359–386 and 387–412.
15. Reeder, J.; Giegerich, R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics **2004**, 5, 104.
16. Tinoco, I.; Bustamante, C. How RNA folds. J. Mol. Biol. **1999**, 293, 271–281.
17. Mathews, D.H.; Sabina, J.; Zuker, M.; Turner, D.H. Expanded sequence dependence of the thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **1999**, 288, 911–940.
18. Zuker, M. Calculating nucleic acid secondary structure. Curr. Opin. Struct. Biol. **2000**, 10, 303–310.
19. Walter, A.E.; Turner, D.H. Sequence dependence of stability for coaxial stacking of RNA helixes with Watson-Crick base paired interfaces. Biochemistry **1994**, 33, 12715–12719.
20. Friederich, M.W.; Vacano, E.; Hagerman, P.J. Global flexibility of tertiary structure in RNA: Yeast tRNA(phe) as a model system. Proc. Natl. Acad. Sci. USA **1998**, 95, 3572–3577.
21. Hall, K.B.; Sampson, J.R.; Uhlenbeck, O.C.; Redfield, A.G. Structure of an unmodified tRNA molecule. Biochemistry **1989**, 28, 5794–5801.
22. Sampson, J.R.; Uhlenbeck. O.C. Biochemical and physical characterization of an unmodified yeast phenylalanine transfer RNA transcribed in vitro. Proc. Natl. Acad. Sci. USA **1994**, 85, 1033–1037.

23. Wales, D.J. *Energy Landscapes: With Applications to Clusters, Biomolecules and Glasses*, Cambridge University Press, Cambridge, 2003.
24. Westhof, E.; Dumas, P.; Moras, D. Loop stereochemistry and dynamics in transfer RNA. J. Biomolec. Struct. Dyn. **1983**, 1, 337–355.
25. Grosjean, H.; and Benne, R. *Modification and Editing of RNA*; ASM Press, Washington DC, 1998.
26. Cech, T.R. Conserved sequences and structures of group I introns: Building an active site for RNA catalysis: A review. Gene **1988**, 73, 259–271.
27. Zarrinkar, P.P.; Williamson, J.R. Kinetic intermediates in RNA folding. Science **1994**, 265, 918–924.
28. Pan, T.; Sosnick, T.R. Intermediates and kinetic traps in the folding of a large ribozyme revealed by circular dichroism and UV absorbance spectroscopies and catalytic activity. Nat. Struct. Biol. **1997**, 4, 931–938.
29. Grosberg, A.Y.; Khokhlov, A.R. *Statistical Physics of Macromolecules*; AIP Press, New York, 1994.
30. Flory, P.J. *Statistical Mechanics of Chain Molecules*; Wiley, New York, 1969.
31. McKenzie, D.S. Polymers and scaling. Phys. Rep. **1976**, 27C, 35–88.
32. Ma, S.-K. Introduction to renormalization group. Rev. Mod. Phys. **1973**, 45, 589–614.
33. Rudnick, J.; Gaspari, G. *Elements of the Random Walk*; Cambridge University Press, Cambridge, 2004.
34. Flory, P.J. *Principles of Polymer Chemistry*; Cornell University Press, New York, 1953.

APPENDICES

A rigorous foundation for the cross linking entropy model is presented from two independent standpoints in Dawson et al.,[14] both from straightforward physical considerations and from a purely mathematical standpoint. The reader is encouraged to consult the references cited in this section for a deeper understanding.

A. The End-to-End Separation Distance

Inasmuch as any polymer in the denatured state can be approximated by a random walk model (Figures 2a and b) like a Gaussian polymer chain (GPC), it has been shown that for any pair of mers i and j, the root-mean-square (rms) end-to-end separation distance between them ($\langle r^2 \rangle_{ij,\xi}^{1/2}$: rms-distance) is a function of the difference in the number of mers that separate them[29]

$$\langle r^2 \rangle_{ij\xi}^{1/2} = \xi^{1-\nu} N_{ij}^{\nu} b \tag{A1}$$

where N_{ij} is the number of residues separating i and j ($N_{ij} = j - i + 1$) and ν measures the excluded volume of a polymer that depends on the specific polymer-solvent system involved and ranges between $1/3 < \nu < 3/5$. In all RNA structure prediction problems, the implicitly assumed value is $\nu \equiv 1/2$ or $\langle r^2 \rangle_{ij,\xi} = \xi N_{ij} b^2$. We are quite explicit; however, we will adopt this value within the current monograph.

An essential property of the rms-distance is that when one shifts by a distance k to $(i + k,\ j + k)$, the same behavior is observed: $\langle r^2 \rangle_{ij,\xi} = \langle r^2 \rangle_{i+k,\,j+k,\xi}$ where $-i < k < N - j$.[29]

B. The Probability Density Function in the CLE-Model

There are a variety of ways in which to represent the configuration of a polymer. When represented mer-by-mer, as in the blue dots in Figures 2a and b, one is tempted to model the structure using matrices as in the Flory isolated pair model.[30] However, as we move away from the convenience of the mer-by-mer representation to effective mers, our modeling skills are seriously challenged. We reflect that the determinant of the product of matrices resembles a multinomial distribution, which itself resembles a normal distribution. Thus emerges the essence of the Gaussian polymer chain: a probability distribution function (pdf). In its most general form, the probability that one should find the end-to-end distance between mers i and j equal to $r_{ij}$with a tolerance of $\Delta r(r_{ij} - \Delta r/2 < r_{ij} < r_{ij} + \Delta r/2)$ is expressed by the following probability

$$p(r_{ij})\Delta r = A_{\delta\gamma} C^{\gamma\delta}_{ij,\xi}(r_{ij}/b)^{\delta\gamma} \exp[-\vartheta_{ij,\xi}(r_{ij}/b)^{\delta}](\Delta r/b) \tag{B1}$$

where b is the mer-to-mer separation distance, δ is a finite positive constant, and $\gamma(> 0)$ (see McKenzie [31] and references therein). Of the other parameters, $A_{\delta\gamma}$ is the spherically symmetric contribution to the volume weight element $(\Delta V_{ij} = A_{\delta\gamma} r^{\delta\gamma}_{ij} \Delta r)$,

$$A_{\delta\gamma} = \frac{\delta\pi^{\gamma+1/\delta}}{\Gamma(\gamma + {}^{1}/_{\delta})} \tag{B2}$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the Gamma-function; $C^{\gamma\delta}_{ij,\xi}$ is a normalization constant for the pdf

$$C^{\gamma\delta}_{ij,\xi} = \frac{\delta\vartheta^{\gamma+1/\delta}_{ij,\xi}}{A_{\delta\gamma}\Gamma(\gamma + {}^{1}/_{\delta})} \tag{B3}$$

and $\vartheta_{ij,\xi}$ is an exponential weight,

$$\vartheta_{ij,\xi} = \left(\frac{\Gamma(\gamma + {}^{3}/_{\delta})}{\Gamma(\gamma + {}^{1}/_{\delta})}\frac{b^2}{\langle r^2\rangle_{ij,\xi}}\right)^{\delta/2} \tag{B4}$$

with $\langle r^2\rangle_{ij,\xi}$ defined in Eq. (A1).

Since $r \gg \Delta r$ and the resolution is on the scale of mers, one can set $\Delta r \sim b$. In the case of the GPC (where $\delta \equiv 2, \gamma \equiv 1$ and $\nu \equiv 1/2$): $A_{\delta\gamma} = 4\pi$, $\langle r^2\rangle_{ij,\xi} = \xi N_{ij} b^2$, $\vartheta_{ij,\xi} = 3/(2\xi N_{ij})$, and $C_{ij,\xi} = [3/(2\pi\xi N_{ij})]^{3/2}$.[30] Eq. (B1) has the property of being self-similar and scalable.[31]

C. Base Pair Stacking in the CLE-Model

The main issue is that effective mers are measured on a length scale of ξ nt yet the standard base pair stacking parameters are measured on a length scale of $\xi = 1$ bp. To reconcile this problem, we invoke renormalization theory[29,31–33] where it is well know that self-similar and scalable systems exhibit the property that their entropy scales roughly as $S(\xi = 1) \rightarrow S(\xi)/\xi^d$ where $d \sim 1$ is the dimensionality of the system and $\xi > 1$ is the length scale of the correlation.[32] In the case of RNA, the entropy-loss is reduced in inverse proportion to the correlation between bases in the sequence and equal to the Kuhn length (ξ).

Since there is no implicit temperature dependence in Eq. (B1), the global contribution to the entropy for mers i and j becomes

$$\begin{aligned} S(r_{ij}) &= \frac{k_B}{\xi} \ln[p(r_{ij})\Delta r^2] \\ &\approx \frac{k_B}{\xi}\left[\ln\left(A_{\delta\gamma} C_{ij,\xi}^{\gamma\delta}\right) + \delta\gamma \ln(r_{ij}/b) - \vartheta_{ij,\xi}(r_{ij}/b)^{\delta}\right] \end{aligned} \tag{C1}$$

where k_B is the Boltzmann constant and ξ scales the entropy contribution of stem formation by a corresponding reduction in degrees of freedom. In effect, we are averaging the entropy contributions from these mers over the range of a Kuhn length.

In Eq. (C1), we have implicitly assumed that ξ is a constant because the mers are similar in chemical behavior and composition. Since RNA is likely to exhibit variable flexibility, we should expect that ξ is *not* a constant. However, if the model itself is soundly robust, these shortcomings should be either surmountable or rectifiable in future developments where a variable ξ is included. For small N_{ij}, there is local structure in a real polymer that simply cannot be approximated properly as "spherically symmetric balls" as one can see in Figure 2c. We currently must add these local corrections from known experimental measurements.

The global change in entropy is measured by considering two stable states of the system: the denatured structure, where $r_{ij} \rightarrow R_{ij} = \langle r^2 \rangle_{ij,\xi}^{1/2} = \xi^{1-v} N_{ij}^{v} b$, and the native state, where $r_{ij} \rightarrow \lambda_{ij} b$. For RNA, $\lambda_{ij} = \lambda$ (Figure 2c). Since the polymer model involves very crude approximations of shape, the value of λb is not equivalent to the chemical bond length of a nucleic acid. Rather it reflects the distance between two similar spheres (shown schematically in Figure 2c). The transition between the initial denatured state ($r_{ij} = R_{ij}$) and the final native state ($r_{ij} = \lambda b$) defines the entropy-loss. Using $R_{ij} = \langle r^2 \rangle_{ij,\xi}^{1/2}$, the entropy-loss due to bp formation as the structure folds from the denatured state to the native state has the form

$$\begin{aligned}\Delta S_{bp}(N_{ij}) &= S(\lambda b) - S(R_{ij})\\ &= -\frac{k_B}{\xi}\{\delta\gamma \ln(R_{ij}/(\lambda b)) - \vartheta_{ij,\xi}[(R_{ij}/b)^{\delta} - \lambda^{\delta}]\}. \qquad \text{(C2)}\end{aligned}$$

Substituting Eqs. (A1) and (B4) into Eq. (C2) and simplifying, a relatively compact expression emerges

$$\Delta S_{hp}(N_{ij}) = -\frac{k_B}{\xi}\{\nu\delta\gamma \ln(\Psi_\nu N_{ij}) - \zeta(\gamma,\delta)[1 - 1/(\Psi_\nu N_{ij})^{\delta\nu}]\} \qquad \text{(C3)}$$

where $\Psi_\nu = \xi^{(1/\nu)-1}(1/\lambda)^{1/\nu}$ and $\zeta(\gamma,\delta) = [\Gamma(\gamma + {}^3/_\delta)\Gamma(\gamma + {}^1/_\delta)]^{\delta/2}$.

We now obtain Eq. (1), when we *assume* that $v \equiv \frac{1}{2}$ (the athermal condition[34]) and the function adheres to Gaussian-like statistics ($\delta \equiv 2$); $\nu\delta \to 1$, $\Psi_{1/2} = \xi/\lambda^2$, and $\zeta(\gamma, 2) \to (\gamma + {}^1/_2)$.

D. Interior Loops and Multibranch Loops in the CLE-Model

In the same way as we arrived at Eqs. (1) and (2) in the text, we consider what happens when the transition is from the denatured state ($r_{rms} = \langle r^2\rangle_{ij,\xi}^{1/2} = [\xi N_{ij}]^{1/2} b$) to a more localized and ordered structure associated with the interior loop. From theory, it is well known that the separation distance (r_I) at the center of a circular loop (Figure 3) will be proportional to half the number of residues in the loop: $r_I = [\xi(n + 2\lambda)/2]^{1/2} b$, where from Figure 3, $n = n_1 + n_2$ with $n_1 = (p - i - 1)$ and $n_2 = (j - q - 1)$ are the lengths of the two single-strand sequences.[29] Therefore, the entropy-loss at the center of the loop is approximately

$$\begin{aligned}\Delta S_I(n) &= S(r_I) - S(r_{rms})\\ &= -\frac{k_B}{\xi}\left\{\gamma \ln\left(r_{rms}^2/r_I^2\right) - (\gamma + 1/2)[1 - (r_{rms}/r_I)^2]\right\}. \qquad \text{(D1)}\end{aligned}$$

Now, using the fact that $(r_{rms}/r_I)^2 \sim 2N_{ij}/(n + 2\lambda)$ and weighting Eq. (D1) by $wn/2$ (i.e., the average length of one side of the I-loop times an additional weight), we obtain

$$\begin{aligned}&-T\Delta S(N_{ij}, n, s = 1, \lambda) \approx -T\left(\frac{wn}{2}\right)\Delta S_I(n)\\ &= \frac{wnk_BT}{2\xi}\left[\gamma \ln\left(\frac{2N_{ij}}{n + (s+1)\lambda}\right) - \left(\gamma + {}^1/_2\right)\left(1 - \frac{n + (s+1)\lambda}{2N_{ij}}\right)\right] \qquad \text{(D2)}\end{aligned}$$

where N_{ij} is the enclosed length from position (i, j) (Figure 3), $n = n_1 + n_2$ and s is the number of stems that branch from (i, j) (for an I-loop, this must be $s = 1$). Eq. (D2) is added to Eq. (2). Currently, $w = {}^1/_2$. Note that the weight on this entropy is $n = (n_1 + n_2)$, not $n + (s + 1)\lambda$. This is because the $5'$ and $3'$ ends of the I-loop are already accounted for by the CLE calculation of the stems.

For an MBL with s-branches, we must sum all the free strand regions: $n = \sum_1^{s+1} n_k$ (see, for example, Figure 1d with 3 branches). Treating this similar to an I-loop, the general expression becomes

$$\begin{aligned}
-T\Delta S(N_{ij}, n, s, \lambda) &\approx -T\left(\frac{w[n + (s-1)\lambda]}{2}\right)\Delta S_I(n) \\
&= \frac{w[n + (s-1)\lambda]k_B T}{2\xi}\left[\gamma \ln\left(\frac{2N_{ij}}{n + (s+1)\lambda}\right)\right. \\
&\quad \left. - (\gamma + 1/2)\left(1 - \frac{n + (s+1)\lambda}{2N_{ij}}\right)\right] \qquad \text{(D3)}
\end{aligned}$$

which yields Eq. (4).

E. The Local CLE Contribution

The local entropy contribution reflects the fact that we must apply a renormalization correction in the FE to account for the freezing out of the degrees of freedom of the individual mers to form the “effective mers.” This yields additional negative contribution to the entropy of the entire polymer chain as ξ is increased.

Derivation of Eq. (3) is quite involved. Only an outline is presented. Altering ξ in Eq. (C1) amounts to changing the chemical potential (i.e., adding or deleting mers).

We first look at a small segment $r = b\sqrt{\xi}$ (the effective length of a single Kuhn segment) and solve Eq. (C1) for a change in ξ (from $\xi = 1$ to ξ) at fixed r (without averaging the entropy over ξ because we are evaluating this change from the length scale and viewpoint of the monomers themselves)

$$\begin{aligned}
(\Delta S_{1\to\xi})_{r,U} &= S_{r,U}(\xi) - S_{r,U}(\xi = 1) \\
&= -k_B\left(\gamma + {}^1/_2\right)\left\{\ln\left(\frac{1}{\xi}\right) - \xi\left(\frac{1}{\xi} - 1\right)\right\} \qquad \text{(E1)}
\end{aligned}$$

where $(S)_{r,U}$ means the entropy at fixed r and internal energy (U). Likewise, the change in Kuhn length (per effective mer) over this process is

$$\Delta n \sim \xi - 1 = \xi\left(1 - \frac{1}{\xi}\right) \qquad \text{(E2)}$$

Using the relation between chemical potential and entropy $[\mu = -T(\partial S/\partial n)_{r,U}]$ and free energy, we obtain

$$\begin{aligned}\Delta G_{1\to\xi} &= \left(\frac{N}{\xi}\right)\mu = -\left(\frac{NT}{\omega\xi}\right)\int_{+1}^{\xi}\left(\frac{\Delta S_{1\to\xi}}{\Delta n}\right)d\xi \\ &= \frac{(\gamma + {}^{1}/_{2})\,Nk_BT}{\omega\xi}\int_{+1}^{\xi}\left(\frac{\ln(x)}{1-x} + 1\right)dx \qquad \text{(E3)}\end{aligned}$$

where ω is the dimensionality (where the straightening out of the polymer chain and consequently the change in Kuhn length is averaged over the three axes). Solving $\Delta G_{1\to\xi} = -T\Delta S_{\gamma\xi}$ yields Eq. (3).