

Abundance of pseudoknots: theory and observation

Daniel P. Aalberts* and Evan M. Miller

Physics Department, Williams College, Williamstown, MA 01267, USA

(Dated: May 31, 2007)

The pseudoknot fold is often seen in auto-catalytic RNA and in viruses. The relative probabilities of different pseudoknot folds predicted by a statistical-mechanical theory is consonant with a database of known folds. We extend that result to estimate the abundance of pseudoknot folds in RNA sequences, finding approximately 1 pseudoknot per 40,000 nucleotides. This theoretical probability density compares favorably to what we infer from structure databases and has implications for genome organization, RNA folding algorithms, and the RNA World.

The pseudoknot fold is an uncommon RNA secondary structure of unusual importance. Defined by its non-nested topology (Figure 1b), it forms the catalytic core in self-splicing introns and in many ribozymes, it enables reading-frame shifts in viruses, and it participates in translation control [1]. Pseudoknots have a catalytic propensity. This insight emerges from directed evolution experiments: simple stem-loop RNA structures are sufficient to bind to small molecules; however, when directed to act as ligase enzymes, pseudoknot structures emerge [2, 3]. In addition, pseudoknots are essential in most of the natural ribozymes, see Table I.

How rare are pseudoknot structures? In this paper, with a thermodynamic model, we predict the abundance of pseudoknots in RNA, compute the density of pseudoknots observed in nature, and find the comparison agreeable.

Pseudoknots are relatively unlikely in sequence space because of the base-pairing constraints of the stems; however, they have more base pairing and thus a lower energy than typical nested (Figure 1a) structures [12]. Although pseudoknots may in principle be of arbitrary complexity, the simple ABAB-type is by far the most common, accounting for 97% of the pseudoknots in PseudoBase [12, 13], most with no nucleotides separating the stems (see Figure 1b). This fact supports our use of only this dominant group for our theoretical estimates.

The probability of an ABAB pseudoknotted native state p_ψ can be computed [12] from physical principles. The base-pairing constraints of two stems of length s_1 and s_2 (defined in Figure 1b) must be satisfied to permit the ABAB pattern. Complementarity occurs with prob-

ability $1/4^{s_1+s_2}$. The one-quarter chance of each base is both the most unbiased assumption and the actual composition of bases in PseudoBase [13] sequences.

If the ABAB pattern is present, the sequence may adopt either an ABAB pseudoknot fold or a nested fold, or remain unfolded. The equilibrium between these options is expressed in terms of Boltzmann weights. Thus,

$$p_\psi = \frac{1}{4^{s_1+s_2}} \frac{e^{-\beta G_{\text{ABAB}}}}{e^{-\beta G_{\text{ABAB}}} + e^{-\beta G_{\text{nested}}} + e^{-\beta G_{\text{unfolded}}}}, \quad (1)$$

with $\beta^{-1} = RT$ and $G_{\text{unfolded}} \equiv 0$. We now must specify the free energies G_{ABAB} and G_{nested} .

The ABAB pseudoknot fold includes base-paired stems and polymer loops,

$$G_{\text{ABAB}} = G_{\text{stems}}(s_1, s_2) - TS_{\text{loops}}(s_1, s_2, L_1, L_3), \quad (2)$$

where L_1 and L_3 are the numbers of nucleotides (nt) in the loops (defined in Figure 1b). Let's consider each contribution separately.

| Ribozyme | Pseudoknot | Essential |
|------------------------------|------------|-----------|
| RNase P [4] | yes | yes |
| Hepatitis Delta Virus [5] | yes | yes |
| Neurospora VS [6] | yes | yes |
| Group I intron [7] | yes | yes |
| Group II intron [8] | yes | yes |
| Hammerhead [9] | alt. fold | yes |
| Hairpin [10] | no | no |
| glmS [11] | yes | enhances |
| Diels-Alder [1] (artificial) | yes | ? |

TABLE I: A listing of ribozymes, whether they are pseudoknotted, and whether the pseudoknot is essential for function.

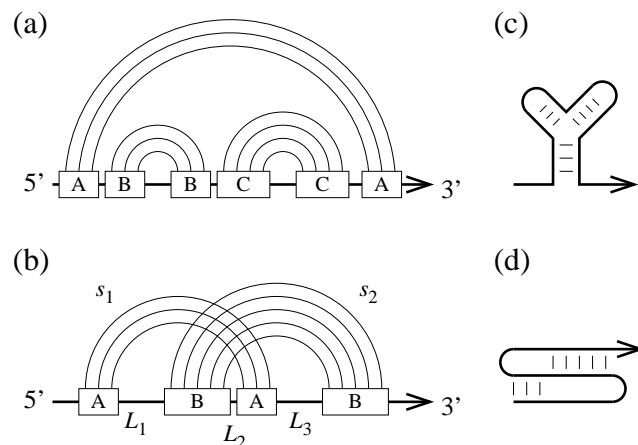


FIG. 1: (a) Nested stem-loop hairpin or branched hairpin folds are the most common RNA structures, depicted here with base pair connections. Stems are labeled with letters. (b) An ABAB-pseudoknot is a non-nested structure with crossing base-pairing lines. Stem lengths s_1 and s_2 and loop lengths L_1 , L_2 , and L_3 are indicated. $L_2 = 0$ in most ABAB pseudoknots because stacking between the stems stabilizes the structure. (c) and (d) are two-dimensional representations of (a) and (b).

The average free energy of paired stems G_{stems} is linear in the number of base pairs in the stem. Likewise, the average free energy of the optimal nested fold G_{nested} is found to be linear in the number of nucleotides in the sequence N .

$$\begin{aligned} G_{\text{stems}} &= -2.14(N - 2.5) \text{ kcal/mol}, \\ G_{\text{nested}} &= -0.287(N - 17) \text{ kcal/mol}. \end{aligned} \quad (3)$$

Free energies were obtained using version 3.2 of MFOLD [14] which fixes the temperature at 37°C. The G_{nested} values are averages of the free energy of the optimal fold of 100 random RNA sequences at each length in the range $20 \leq N \leq 100$. G_{stem} is obtained using MFOLD with random complementary stems of length $6 \leq N \leq 100$; each stem is bookended with mismatching pairs. Again 100 sequences are averaged for each N . These linear fits have utility beyond the present context; our group is currently using them to estimate unfolding barriers to pass through nanopores, bind small molecules, microarrays, and so on.

The entropy S_{loops} is estimated using the standard Gaussian approximation from polymer physics. The probability that a chain of N links of length a has an end-to-end separation distance between D and $D + d$ is

$$p_G(D, N) = 4\pi D^2 d \left(\frac{3}{2\pi N a^2}\right)^{3/2} \exp\left\{-\frac{3D^2}{2Na^2}\right\}. \quad (4)$$

If a loop has L nucleotides it has $N = L + 1$ links of length $a = 6.2\text{\AA}$. Using Eq. (4), the entropic contribution for a loop spanning a distance D then is $S_{\text{loop}} = R \ln[p_G(D, L + 1)]$. As in Ref. [12], the distances D are calculated from the geometry of the A-form RNA helix; these D prove to differ for loops crossing the major and minor grooves. The shell thickness d is the one free parameter of the theory [12], and $d = 0.002\text{\AA}$ optimizes and provides good agreement with the observed distribution of stem lengths in PseudoBase as seen in Figure 2.

With p_ψ fully parameterized, we can now estimate the density of pseudoknots in the space of all possi-

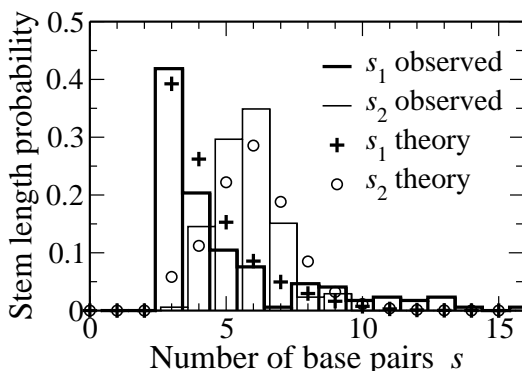


FIG. 2: Stem lengths s_1 and s_2 are observed to follow different distributions. The Aalberts and Hodas theory [12] for p_ψ explains the asymmetry. The shell thickness d in Eq. (4) is fixed to optimize agreement with the observed distribution.

ble RNA sequences, our primary interest in this paper. The probability divided by the length of the pseudoknot and summed over all configurations yields the probability density:

$$\rho = \sum_{s_1, s_2, L_1, L_3} \frac{p_\psi(s_1, s_2, L_1, L_3)}{2s_1 + 2s_2 + L_1 + L_3}. \quad (5)$$

Our numerical prediction for ρ uses $d = 0.002\text{\AA}$ in Eq. (5). We predict the pseudoknot probability density $\rho \approx 1$ pseudoknot per 40,000 nt.

We turn now to comparing our theoretically predicted probability density with the observed abundances of naturally occurring pseudoknotted structures. We use the Protein Data Bank (PDB) [15] on 17 November 2006 as our source of high quality RNA fold data. Duplicate PDB sequences are eliminated, as are substrings [16]. Stems which contain at least three base pairs and which cross as in Figure 1b are identified as pseudoknotted structures, and each crossing stem is labeled. The number of pseudoknots N_ψ is obtained by dividing the number of stem labels by 4 in order to process more complicated pseudoknot patterns. Thus an ABAB pattern gives $N_\psi = 1$, an ABACDCDB pattern gives $N_\psi = 2$, and an ABACBC pattern gives $N_\psi = 1.5$.

Pseudoknot densities are calculated separately for each species in the database simply by dividing the number of pseudoknots by the total length of PDB sequences. There are, however, two reasons this method may overestimate the natural abundance of pseudoknots. First, pseudoknot folds are more constrained than nested structures and therefore more likely to crystallize and give a good X-ray images. Second, researchers are typically most interested in the structures which are functional, as pseudoknots often are (Table I). But by composing a lower density bound for each organism, dividing the number of pseudoknots by the total length of the genome, we can account for such selection biases. The range of observed pseudoknot densities is

$$N_\psi/L_{\text{PDB}} \gtrsim \rho \gtrsim N_\psi/L_{\text{genome}}, \quad (6)$$

where L_{PDB} is the number of nucleotides in the PDB, and L_{genome} is the number in the genome. In Figure 3 with species from all domains of life, the empirical bounds given by Eq. (6) are shown.

Stochastic RNA folding simulations like KINEFOLD can generate complex pseudoknot folds [17]. The theory behind KINEFOLD lacks important 3D structural details such as the major and minor groove asymmetry [12]. Recently KINEFOLD has been used to calculate, at 37°C, that the average fraction of pseudoknotting base pairs is a few percent [18]. This translates to a predicted density of the order of one pseudoknot per 10^2 nt. Figure 3 shows that actual PDB pseudoknots are at least one, and more likely two orders of magnitude more rare than KINEFOLD predicts.

Conventional deterministic RNA folding algorithms such as MFOLD or VIENNA RNA ignore the possibility of

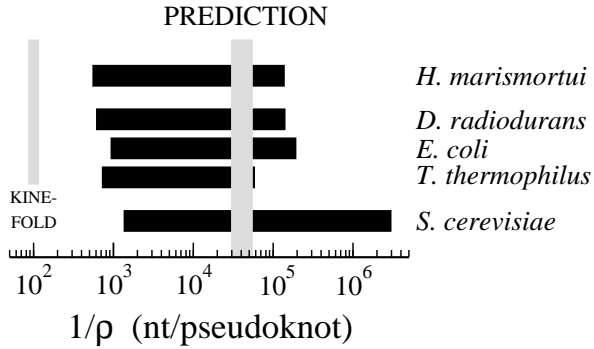


FIG. 3: The horizontal bar indicates the pseudoknot density range for each species $L_{\text{PDB}}/N_{\psi} < 1/\rho < L_{\text{genome}}/N_{\psi}$. Note that archaeobacteria, bacteria, and eukaryote domains of life are all represented. Our theoretical prediction, depicted as a vertical bar, is in good agreement with experiment. The KINFOLD simulation results of Ref. [18] predict a high density inconsistent with observation.

pseudoknotted structures. Our calculations and observations place a weak bound on the accuracy, given that assumption. There have been several recent attempts to develop deterministic RNA secondary structure algorithms that include pseudoknot folds, for example Ref. [19].

It is interesting that $1/\rho$ is roughly the length of pre-mRNA for a typical eukaryotic gene; this suggests that longer pre-mRNA would typically have pseudoknots which might disrupt splicing or translation, or might result in catalytic activity.

The likelihood of pseudoknots in finite fragments can be computed in our theory. The probability of zero pseudoknots in a chain fragment of length f nucleotides is:

$$p_0(f) = \prod_{n=1}^f \left(1 - \sum_{N=1}^{f-n+1} \frac{p_{\psi}(N)}{N} \right) \approx e^{-\rho(f-29)}, \quad (7)$$

where $N = 2s_1 + 2s_2 + L_1 + L_3$ is the total length of the pseudoknot. For large f ,

$$p(\text{pseudoknot}) \approx 1 - \exp\{-\rho(f-29)\}, \quad (8)$$

is the probability of having one or more pseudoknots in the fragment.

The probability of zero pseudoknots in K fragments of length f is:

$$[p_0(f)]^K. \quad (9)$$

Pseudoknots characteristically appear when this probability decays to $1/e$, when the total material $M = Kf$ is:

$$M = -f / \ln[p_0(f)]. \quad (10)$$

The dependence on fragment length is depicted in Figure 4. We find that $1/\rho = 40,000$ nt means that a single

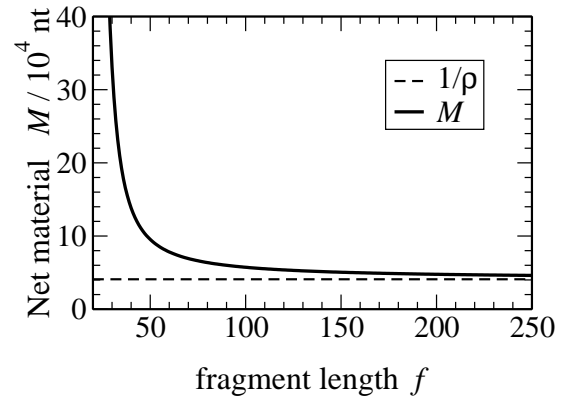


FIG. 4: Because of the finite size of pseudoknots, the total amount of RNA material M depends on the fragment size f . In the limit of long fragments, M approaches $1/\rho$. Results are presented for 37° conditions.

40,000 nt sequence, one hundred 430 nt, or six hundred 100 nt RNA sequences will, with approximately equal probabilities, contain pseudoknotted structures.

A few comments about the role of evolution are in order. It is interesting that nature's RNA, even after billions of years of evolution has basically equal numbers of the four nucleotides. Pseudoknots might be selected for if they provide a useful function, selected against if they interfere with the workings of the cell, or may spontaneously appear or disappear due to mutations. Nevertheless, the evidence suggests that our thermodynamic model produces predictions which fall within the bounds set by the available data. We observe that this estimate is correct to an order-of-magnitude predictor for pseudoknot abundance in all domains of life.

Our theory may have something to say about the barrier to the emergence of the first enzymes. In the RNA World that likely predated our present-day biology, catalytic activity and information storage are embodied in RNA rather than protein and DNA. One key question [20] is how much raw RNA material is required to permit self-replication?

Given the prevalence of pseudoknots in ribozymes, the density of pseudoknots might be regarded as a proxy for the density of ribozymes in random RNA. In the warmer conditions ($55^\circ\text{C} < T < 85^\circ\text{C}$) of 3.5 billion years ago [21], we predict $\rho(60^\circ\text{C}) \sim 1/10^6$ nt. The directed evolution literature [2, 3] suggests that convergence to efficient ribozymes from seed structures is relatively rapid. In this way, estimates for the probability of producing functional molecular structures in the RNA World and today can be derived from physical principles.

In summary, the probability $p_{\psi}(s_1, s_2, L_1, L_3)$ for pseudoknots with specified stem and loop lengths can be calculated with a statistical-mechanical theory [12]; the predictions of this one-parameter theory agree well with observed pseudoknot structures (see Figure 2) and explain

the observed asymmetry of stem lengths. The probability density of pseudoknots today is estimated with Eq. (5). Then, by analyzing PDB structures, we demonstrate our theoretical estimate is correct to an order of magnitude.

We have argued that pseudoknots are common structures in ribozymes, so one may expect that the pseudoknot density and ribozyme density are of the same order of magnitude. This permits us to make a factor-of-

ten estimate of the amount of RNA material necessary for the first ribozymes in the RNA World and address one of the key barriers in the emergence of life.

This work is supported by grants from the National Institutes of Health (GM068485) and the National Science Foundation (MCB-0641995). Thanks to Teng Jian Khoo for data preparation and discussions. Thanks to Nathan Hodas and Wendy Raymond for discussions.

-
- [*] Contact: aalberts@williams.edu, 1-413-597-3520
- [1] One recent review of pseudoknot functions is: Staple, D.W. and S.E. Butcher. 2005. Pseudoknots: RNA Structures with Diverse Functions. *PLoS Biol* 3(6):e213.
- [2] Ekland, E.H., J.W. Szostak, and D.P. Bartel. 1995. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* 269:364-370.
- [3] Baskerville, S. and D.P. Bartel. 2002. A ribozyme that ligates RNA to protein. *Proc. Natl. Acad. Sci. USA* 99:9154-9159.
- [4] Mann, H., Y. Ben-Asouli, A. Schein, S. Moussa, and N. Jarrous. 2003. Eukaryotic RNase P: Role of RNA and protein subunits of a primordial catalytic ribonucleoprotein in RNA-based catalysis. *Molecular Cell* 12:925-935.
- [5] Wadkins T.S., A.T. Perrotta, A.R. Ferré-D'Amaré, J.A. Doudna, and M.D. Been. 1999. A nested double pseudoknot is required for self-cleavage activity of both the genomic and antigenomic hepatitis delta virus. *RNA* 5:720-727.
- [6] Rastogi, T., T.L. Beattie, J.E. Olive, and R.A. Collins. 1996. A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. *EMBO Journal*, 15:2820-2825.
- [7] Cech, T.R. 1990. Self-splicing of Group I introns. *Annu. Rev. Biochem.* 59:543-568.
- [8] Harris-Kerr, C.L., M. Zhang, and C.L. Peebles. 1993. The phylogenetically predicted base-pairing interaction between α and α' is required for group II splicing *in vitro*. *Proc. Natl. Acad. Sci. USA* 90: 10658-10662.
- [9] Song, S.I., S.L. Silver, M.A. Aulik, L. Rasochova, B.R. Mohan, and W.A. Miller. 1999. Satellite cereal yellow dwarf virus-RPV (satRPV) RNA requires a double hammerhead for self-cleavage and an alternative structure for replication. *J. Mol. Biol.* 293:781-793.
- [10] Doherty, E. A. and J. A. Doudna. 2000. Ribozyme structures and mechanisms. *Annu. Rev. Biochem.* 69:597-615.
- [11] Wilkinsin, S.R., and M.D. Been. 2005. A pseudoknot in the 3' non-core region of the glmS ribozyme enhances self-cleavage activity. *RNA*, 11:1788-1794.
- [12] Aalberts, D.P. and N.O. Hodas. 2005. Asymmetry in RNA pseudoknots: observation and theory. *Nucleic Acids Res.* 33:2210-2214.
- [13] Batenburg, H.F.D. van, A.P. Gulyaev, C.W.A Pleijj, J. Ng, and J. Oliehoek. 2000. Pseudobase: a database with RNA pseudoknots. *Nucleic Acids Res.* 28:201-204.
- [14] Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406-3415.
- [15] Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235-242.
- [16] A complete list of the structures used in our analysis is available at <http://rna.williams.edu/abundance/>
- [17] Isambert, H. and E.D. Siggia. 2000. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. USA* 97:6515-6520.
- [18] Xayaphoummine, A., T. Bucher, F. Thalmann, and H. Isambert. 2003. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. USA* 100:15310-15315.
- [19] Rivas, E. and S.R. Eddy. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285:2053-2068.
- [20] Joyce, G.F. and Orgel, L.E. 1999. Prospects for Understanding the Origin of the RNA World. *The RNA World*, second edition, eds. Gesteland, R.F., Cech, T.R., and Atkins, J.F. Cold Spring Harbor Lab. Press, Cold Spring Harbor, NY, 49-77.
- [21] Knauth L.P. and D.R. Lowe D.R. 2003. High Archaean climatic temperature inferred from oxygen isotope geochemistry of cherts in the 3.5 Ga Swaziland Supergroup, South Africa. *Bulletin Geological Society of America* 115:566-580.