A new entropy model for RNA: part IV, The Minimum Free Energy and the thermodynamically most-probable folding pathway

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: dawson@bi.a.u-tokyo.ac.jp

Keywords

RNA folding; Entropy; Dynamic programming algorithm; RNA structure; functional RNA; 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 *vsfold5* and *vs_subopt*.

Abstract

Here we discuss four important questions (1) how can we be sure that the thermodynamically most-probable folding-pathway yields the minimum free energy for secondary structure using the dynamic programming algorithm (DPA) approach, (2) what are its limitations, (3) how can we extend the DPA to find the minimum free energy with pseudoknots, and finally (4) what limitations can we expect to find in a DPA approach for pseudoknots. It is our supposition that some structures cannot be fit uniquely by the DPA, but may exist in real biology situations when disordered regions in the biomolecule are necessary. These regions would be identifiable by using suboptimal structure analysis. This grants us some qualitative tools to identify truly random RNA sequences, because such are likely to have greater degeneracy in their thermodynamically most-probable folding-pathway.

1. Introduction

RNA structure prediction methods usually rely upon a method known as the dynamic programming algorithm (DPA) [1-3] to find the optimal solution for the minimum free energy and related suboptimal structures [4,5]. Familiar examples are mfold [6,7] and RNAfold [8,9]. In this series, we have been exploring the prediction behavior of an entropy model that attempts to address the collective elastic response of multiple contact points in a folded RNA molecule, called the cross linking entropy (CLE) model [10,11]. All implementations of the CLE model also utilize a form of the DPA to compute either the optimal structure (*vsfold5*) or suboptimal structures (*vs_subopt*). Up to this point in this series, the DPA was treated like a black box that magically determined optimal structures for any RNA sequence. Whereas the DPA appears to be a successful strategy for prediction of secondary structure and even pseudoknots, it is important to understand why it succeeds and what are its limitations.

The DPA is generally applied to problems or processes where each new solution is built up progressively [1-3,12,13]. It requires some way to describe the process recursively, such that previous optimal solutions can be reused without solving them more than once [2,12]. Applications of the DPA typically involve time dependent processes (both deterministic and stochastic processes are possible [1]) or stepwise procedures that require selecting an optimal solution at each step and using that solution within all subsequent steps of the process. For example, crystal growth involves the gradual accumulation of well-arranged atoms or molecules onto a seed crystal. Ignoring the delicate matters of thermodynamic equilibrium, in effect, as the crystal grows in size, there is no further change to the atoms or molecules that are already added to the structure. Adjusting inventory in a stable market could be managed with

a stochastic DPA approach by expressing the outcomes as probabilities [1]. Likewise, one could picture optimizing an automobile assembly line. There may be more than one good solution; nevertheless, rarely would one find it more efficient to install and assemble the engine as the last stage on the assembly line.

Historically, the DPA has been applied to RNA secondary structure and pseudoknots without questioning how RNA folds. Fortuitously, experimental evidence suggests that RNA tends to fold from smaller hairpin loops [14-16] and gradually accumulates more complex structures such as internal loops (I-loops), multibranched loops (MBLs) and even pseudoknots (PKs), Fig 1. Moreover, this accumulation is local enough that new structures do not completely disrupt the previously accumulated structure, but simply add to it from the best prior structure. This is similar to the image of crystal growth. Hence, the DPA is a good choice for finding the optimal structure via the minimum free energy because the best free energy (FE) for the smaller structures are readily used in computing the next larger structure and remain unchanged in the process. This in turn means that if a proper order of calculation can be discerned, then each solution need only be evaluated once and the result can be solved in polynomial time instead of exponential time [2].

In the CLE model, shorter sequence lengths tend to fold more rapidly than longer regions (Section 4), a characteristic of RNA folding first reported at least as far back as Refs [14,15]. The distance dependent rate provides a natural order like the crystal growth example and suggests that the DPA can essentially minimize the folding by following that order (Section 3). This order in the folding is what we define as the thermodynamically most-probable folding-pathway (TMPFP). This is consistent with experimental information like the hierarchical folding model [17,18], the related kinetic

models [19-24] and the contact order model [16,25-30], which uses some of the same principles as the CLE model [11]. However, we have not really examined the limitations of the DPA in solving folding problems, particularly in the context of pseudoknots.

In the case of RNA secondary structure, the CLE can define distinguishable folding times for different structures. However, in the case of pseudoknots, the likelihood of encountering equal folding times is higher and we can imagine some case where two different structures compete for an identical site and have identical entropies and binding characteristics. The DPA aims at a single solution. Moreover, in a model that seriously addresses folding, the possibility of rearrangements that alters prior solutions cannot be ruled out. Therefore, it is important to understand the limitations of the DPA in the context of RNA folding problems.

The DPA, though a methodology, is not the only issue here. The biology also must select for sequences that avoid the above-mentioned pitfalls or make very good use of them. We therefore would like to understand what might distinguish a real biological sequence with some direction in the folding from a random sequence that could contain multiple conformations of equal likelihood [29].

Here, we dig into the minimizing strategy we have relied upon to examine how reliable the method is in pseudoknot prediction. In particular, we are interested in where the DPA is likely to fail in these problems, under what conditions the DPA will fail, and whether suboptimal structure evaluation can improve the result. Finally, we are also interested in whether the cases where the DPA fails are realistically possible to occur in real biology problems.

The presentation here is largely heuristic in form. It reflects our experience as we

developed the RNA structure prediction program *vsfold5* and the subsequent suboptimal structure prediction program *vs_subopt* to calculate pseudoknots and some of the myriad of questions that such problems generate. As a result, it is not our intent to rigorously address abstruse issues, but rather to explore practical and pragmatic issues that come with designing such an algorithm.

To understand this discussion, the reader should be familiar with the concept of the cross linking entropy (CLE) discussed in Parts I through III and explained in the literature in Refs [10,11], and particularly in relation to pseudoknots [10]. In particular, it is important to understand the definition of an "effective mer", the Kuhn length (or, in other parlance, the persistence length), and the general equations used to describe this entropy. Effective mers are used here because the global entropy and therefore the folding is a function that is largely dependent on the Kuhn length where the individual behavior of the monomers is grouped into these effective mers. The reader also needs to be familiar with the programming issues of applying optimization algorithms in the context of RNA or protein folding [10,31]. (A basic introduction to the concepts behind the CLE model can be found in Section 2 and its implementation using the DPA in Section 3.)

2. Review of the global entropy in the CLE model

Most of this material should be familiar for readers acquainted with Parts I to III of this series. At the end of this section, we provide some transformation notation that may be helpful for understanding the notation in latter parts of this work in addressing the general behavior of the DPA. This section will only address the global CLE in terms of Gaussian equations and will be written with minimal explanation. For details, the reader is strongly encouraged to review in parts I through III of this series or at least Ref [11] (particularly App A). These are all public access journals.

Let N be the number of monomers (mers) and b the distance between consecutive mers on the RNA polymer chain. Let i and j represent the indices of a pair of mers subject to $1 \le i < j \le N$.

2.1. Kuhn length

Let the Kuhn length (ξ), which expresses the stiffness of the RNA, be defined in units of mers and, for $\xi > 1$, let this quantity describe a group of mers that respond physically as though they comprised a single unit. Such a group of mers will be defined as an *effective mer*. The distance between effective mers is $b' = \xi b$. Hence, when $\xi = 1$, $b' = b$ and the monomer-to-monomer (mer-to-mer) separation distance along the RNA chain would be of unit length. In RNA, ξ is always longer than the mer-to-mer separation distance. Therefore, an RNA sequence consists of N / ξ effective mers (or perhaps *epimer*, where the Greek root *epi* is used in the sense of "addition to" in words like *epiphenomenon* and in some ways emerges from the underlying polymer chemistry like epiphenomena). Then the *epimer*-to-*epimer* separation distance is $b' = \xi b$.

2.2. End-to-end separation distance

When RNA is denatured with an appropriate denaturing solvent (such as urea [32,33]), the RNA tend to expand to a volume where the 5' and 3' ends of the chain span a distance that is proportional to its root mean square (rms) end-to-end separation distance (r_{rms}). In principle, r_{rms} is measurable and is a function of the length of the sequence,

$$r_{rms} = (N/\xi)^{\nu}\,\xi b, \tag{1}$$

where ν is a dimensionless parameter expressing the excluded volume [34,35]. The excluded volume depends on the type of solvent and buffer and it depends on N_{ij}. The value ν roughly ranges between $1/3 < \nu < 3/5$ (a subject of Part V in this series).

Rather than denaturing, which is a complex process in biomolecules that is not well understood even for proteins [36-40], suppose one could simply turn off the amphiphilic interactions of the RNA mers. Then the RNA would begin to exhibit the character of an ideal polymer where $\nu = 1/2$. In such an instance, Eqn (1) becomes

$$r_{rms}^2 = \xi N b^2. \tag{2}$$

Since the rms end-to-end separation distance is a simple function of N, if the sequence is truncated to a length N' ($N' < N$), then it should follow that $(r')_{rms}^2 = \xi N' b^2$. It follows that for mers i and j ($i < j$), this end-to-end separation distance can be

extrapolated to the concept of a rms separation distance between mers i and j (ij-rmsd)

$$r_{rms,ij}^2 = \xi N_{ij} b^2 = \xi (j - i + 1) b^2 . \tag{3}$$

2.3. The global entropy in single-stranded RNA base pairing

It turns out that this distance is also the variance when the arrangement of a polymer is approximated as a random flight model [35,41]. The likelihood of finding mers i and j separated by a distance r_{ij} at any given moment can be expressed as the radial part (r_{ij}) of a Gaussian function

$$p(r_{ij}, \beta_{ij}) \Delta r = 4\pi \left(\frac{\beta_{ij}^2}{\pi} \right)^{3/2} r_{ij}^2 \exp\left\{ -\beta_{ij}^2 r_{ij}^2 \right\} \Delta r \tag{4}$$

where

$$\beta_{ij}^2 = \frac{3}{2\xi N_{ij} b^2} = \frac{3}{2 r_{rms,ij}^2} . \tag{5}$$

Since Eqn (4) contains no explicit temperature dependence, the entropy of the interaction between mers i and j is

$$\begin{aligned} S(r_{ij}, \beta_{ij}, \xi > 1) &= \frac{k_B}{\xi} \ln\left(p(r_{ij}, \beta_{ij}) \Delta r \right) \\ &= \frac{k_B}{\xi} \left\{ \ln\left(4\pi \left(\beta_{ij}^2 / \pi \right)^{3/2} \right) + \ln\left(r_{ij}^2 \right) - \beta_{ij}^2 r_{ij}^2 \right\} \end{aligned} . \tag{6}$$

where k_B is the Boltzmann constant and ξ scales the entropy contribution due to stem formation by a corresponding reduction in degrees of freedom because the length scale is based on *effective mers* rather than mers.

The response of polymers is typically measured with a force-extension instrument such as the optical tweezers [42] and this is reported as $f_{ext}(r_{ij})$, where "ext" refers to the external force required to extend (or compress) the polymer. 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 this response. The mutual response of the mers is $f_{int}(r_{ij})$. This notation is discussed in more detail in Section 5 of Part I in this series. Readers accustomed to the traditional form should read $f_{int}(r) = (-f_{ext}(r))$.

Using the relationship $f_{int}(r_{ij}) = T(\partial S(r_{ij})/\partial r_{ij})_T$, Eqn (6) becomes

$$f_{int}(r_{ij}, \beta_{ij}) = 2k_B T\left(\frac{1}{r_{ij}} - \beta_{ij}^2 r_{ij}\right). \tag{7}$$

where Eqn (7) has a minimum ($R_{ij,c}$) at $R_{ij,c} = 1/\beta_{ij} = (2/3)^{1/2} r_{rms,ij}$. Hence, $r_{ij} < R_{ij,c} \Rightarrow f_{int}(r_{ij}, \beta_{ij}) > 0$ and $r_{ij} > R_{ij,c} \Rightarrow f_{int}(r_{ij}, \beta_{ij}) < 0$. This, in turn means that any pair of mers i and j ($j > i+1$) has this tendency. (Note that satisfying $f_{int}(r_{ij}, \beta_{ij}) = 0$ for all ij simultaneously is not feasible and some frustration will always be present in this system.)

In RNA folding, one measures the structure in the denatured state where $r_{ij} = r_{rms,ij}$ and the native state $r_{ij} = \lambda b$ (where the base pairs have a fixed separation distance in the native state). The mers are simple amorphous object in this model and the separation distance represents the distance to the centers of these objects, not the chemical H-bonding distances of GC, AU, etc. A good value is $\lambda = 2$, because the distance between the chains is about twice that of the mer-to-mer distance b. The entropy change is therefore

$$\begin{aligned}\Delta S_{bp}(N_{ij},\xi) &= S(\lambda b,\xi) - S\left(r_{rms,ij},\xi\right) \\ &= -\frac{k_{\mathrm{B}}}{\xi}\left\{\ln\left(\Psi_{1/2,\xi}N_{ij}\right) - \frac{3}{2}\left(1 - \frac{1}{\Psi_{1/2,\xi}N_{ij}}\right)\right\}.\end{aligned} \tag{8}$$

where $\Psi_{1/2,\xi} = \xi/\lambda^2$. A more general expression for Eqn (8) can be generated based on the material in Section 2 of Part II in Eqns (7) and (8)

$$\begin{aligned}\Delta S_{bp}(N_{ij},\xi) &= S(\lambda b,\xi) - S\left(r_{rms,ij},\xi\right) \\ &= -\frac{k_{\mathrm{B}}}{\xi}\left\{\delta\gamma\ln\left(\frac{r_{rms,ij}}{\lambda b}\right) - \frac{\zeta(\gamma,\delta)}{\xi^{\delta(1-\nu)}N_{ij}^{\delta\nu}}\left[\left(\frac{r_{rms,ij}}{b}\right)^{\delta} - \lambda^{\delta}\right]\right\}\end{aligned} \tag{9}$$

where $r_{rms,ij} = (N_{ij}/\xi)^{\nu}(\xi b)$. Upon substituting $r_{rms,ij}$, becomes

$$\Delta S_{bp}(N_{ij},\xi) = -\frac{k_{\mathrm{B}}}{\xi}\left\{\nu\delta\gamma\ln\left(\Psi_{\nu\xi}N_{ij}\right) - \zeta(\gamma,\delta)\left(1 - \frac{1}{(\Psi_{\nu\xi}N_{ij})^{\delta\nu}}\right)\right\} \tag{10}$$

where $\Psi_{\nu\xi} = \xi^{-1}(\xi/\lambda)^{1/\nu}$, δ is a finite positive constant [43,44] and γ (>0) is the a weight that corrects for the fact that real polymer chains cannot have more than one mer occupying the same space at the same time [45]. This is best called a "self-avoidance" parameter because it differs from the excluded volume associated with the pamameter ν (Part V). A common value used in RNA calculations is $\gamma = 1.75$ [14,45] compared to Gaussian statistics ($\gamma \equiv 1$). The parameter δ is a measure of the character of the correlation in the statistical model of the polymer. The standard value is $\delta = 2$ to reflect the Gaussian polymer chain character. However, the correlation could conceivably be exponential ($\delta = 1$) or even exponential square root ($\delta = 1/2$), at least in principle. The exact form is not well known, but is generally thought to be Gaussian for many problems. Finally,

$$\zeta(\gamma,\delta) = \left[\Gamma(\gamma + 3/\delta)/\Gamma(\gamma + 1/\delta)\right]^{\delta/2} \tag{11}$$

where $\Gamma(x)$ is the Gamma function. Here, it is assumed that $\delta = 2$. Hence, $\zeta(\gamma,2) = (\gamma + 1/2)$. When $\gamma \equiv 1$, Eqn (10) reduces to Eqn (8).

The total entropy-loss is the sum of the local correction (which accounts for the coarse-grained character of the effective mers) and the global contribution caused by stem formation [11]

$$\Delta S_{cle} = \Delta S_{\xi\gamma\delta} + \sum_{bp(ij)} \Delta S_{bp}(N_{ij},\xi), \tag{12}$$

where $\Delta S_{bp}(N_{ij},\xi)$ is the global contribution given in Eqn (8) and the derivation of the

local entropy term ($\Delta S_{\xi\gamma\delta}$) is shown in Sections 3 and 4 of Part II. For a fixed Kuhn length, $\Delta S_{\xi\gamma\delta}$ is a constant for a given sequence length. In this work, this term can be treated as a constant. In effect, the CLE model integrates the contributions from the base pairs. The elements of Eqn (12) have been derived from first principles in numerous independent ways [11,46,47].

2.4. RNA folding and the TMPFP

Returning to Eqn (7), r_{rms} ($R_{ij,c}$) represents a separation distance where the maximum number of configurations are possible. Because both squishing and stretching from r_{rms} decreases the number of possible configurations of the RNA polymer, the entropy in Eqn (6) decreases. Because $R_{ij,c}$ depends on N_{ij}, folding rates are proportional to $\exp(\Delta S_{bp}(N_{ij})/k_{\mathrm{B}})$ and $\Delta S_{bp}(N_{ij})$ is a negative function for reasonable values of N_{ij}, a small N_{ij} yields a faster rate than a large N_{ij}. This is the essence of the thermodynamically most probable folding pathway (TMPFP). However, the unit of measure is the *epimer* (Sec 2.1) not the monomer. This requires introducing some specialized notation.

In Part I, it was shown that the effective mers in a stem are approximated by selecting the midpoint of the stem

$$\bar{i} = \sum_{k=1}^{\xi} i_k / \xi\,, \quad \bar{j} = \sum_{k=1}^{\xi} j_k / \xi \quad \text{and} \quad \bar{N}_{\bar{i}\bar{j}} = \bar{j} - \bar{i} + 1 \tag{13}$$

where there are ξ mers in an effective mer and it is assumed that ξ is an integer.

This is based on observations (Section 5 of Part II) that suggest that the stem length and ξ are roughly equal. For a uniform Kuhn length (ξ) of integer value, the indexing can be further simplified with the following notation

$$\tilde{i} = \overline{i} / \xi, \quad \tilde{j} = \overline{j} / \xi \text{ and } \tilde{N}_{\tilde{ij}} = (\overline{j} - \overline{i} + 1) / \xi = \overline{N}_{\overline{ij}} / \xi = \overline{N}_{\tilde{ij}} / \xi \tag{14}$$

which provides a convenient transformation between the indices of *effective mers* and the corresponding $\tilde{i}\tilde{j}$-rmsd,

$$r^2_{rms,\tilde{ij}} = \xi \overline{N}_{\tilde{ij}} b^2 = \tilde{N}_{\tilde{ij}} (\xi b)^2 . \tag{15}$$

Working from the concepts developed in Parts I and II of this series, it follows that Eqn (10) can be written in terms of Eqn (14) using effective mers (which are measured from the midpoint of a stem of length ξ)

$$\Delta S^{(g)}_{stem}(\overline{N}_{\tilde{ij}}, \xi) = -k_{\mathrm{B}} \left\{ \nu\delta\gamma \ln\left(\Psi_{\nu\xi} \overline{N}_{\tilde{ij}} \right) - \zeta(\gamma, \delta) \left(1 - \frac{1}{(\Psi_{\nu\xi} \overline{N}_{\tilde{ij}})^{\delta\nu}} \right) \right\}. \tag{16}$$

Where the $1/\xi$ in Eqn (10) is absent in Eqn (16) because the units in Eqn (16) are *epimers*, rendering further scaling is unnecessary.

Some care needs to be taken in reading the bar and the tilde notation. Eqns (14) and (15) are used in this work because it is easy to see the transformation. In general, more sophisticated notation methods should be employed. One can construct indices

and transformations for cases where the Kuhn length is neither a single-valued constant nor an integer. However, generalizations needlessly complicate the discussion in the latter part of this study without providing any greater insights.

3. RNA structure and the dynamic programming algorithm

Introductions to the dynamic programming algorithm (DPA) can easily be found in several textbooks [2,3,12,13]. A very clear and simple explanation of the DPA in application to sequence alignment and RNA secondary structure can be found in Refs [5,48] respectively.

The DPA is used in problems that can be solved recursively from a bottom-up (or possibly top-down) strategy. When a DPA method can be used, the DPA solves the problem in such an order that any prior optimal solutions are simply added to the best solution in the current evaluation step. One very simple example would be the way to calculate a Fibonacci number. This has a recursion relation $F_n = F_{n-1} + F_{n-2}$ with $F_0 = 0$ and $F_1 = 1$. If we start from F_0 and F_1, using the prior solution at each new evaluation of n, we are applying the basic mechanics of the DPA procedure [3]. Another important part of the DPA is that there is generally a decision that is made at each step of the recursion, and the best solution is used in subsequent operations. Therefore, an important aspect of this approach is that the prior solutions should be additive to the current solution and that past solutions do not change as a result of new information in subsequent steps.

The purpose of this section is to help the reader understand how the recursion is handled in current applications of the DPA for RNA secondary structure calculations and how the recursion differs when applied to the CLE model. In particular, the heuristics used to evaluate PKs needs to be understood. To understand the details of specific implementations of the current models like mfold [49] or RNAfold [9], the reader should consulted the respective literature. The details on how the implementation of *vsfold5* works on PKs are explained in Supplement 2 of Ref [10].

3.1. Using the DPA in current models for RNA structure prediction

For RNA secondary structure, the DPA was first introduced by Nusinov [50] in maximal matching of base pairs (bps) and later advanced by in work by Zuker and coworkers [7] by adding thermodynamic weights first introduced to RNA in work by Kalenbach [51] and Tinoco [52] and subsequently by Salser [53,54] and later Turner [55,56].

Applying the DPA to RNA structure prediction, it is logical to index the FE in terms of the indices of a matrix with $i < j$, because base pairing consists of the binding of mer i and mer j into a bp (i, j). We represent the region between i and j (inclusive) by the notation $[i \cdots j]$. If we can assume that the solution of $[i \cdots j]$ does not have any influence on the solution of $[p \cdots q]$, where $p > i$ and $q < j$, then the recursion relation for the DPA approach can be written

$$\Delta G_{ij} = \min\left\{\Delta G_{ij}^{\mathrm{fs}}, \Delta G_{ij}^{\mathrm{bp}}, \Delta G_{ij}^{t}, \Delta G_{ij}^{\mathrm{H}}, \Delta G_{ij,pq}^{\mathrm{I}}, \Delta G_{ij,\{pq\}}^{M(k)}\right\} \tag{17}$$

where $\Delta G_{ij}^{\mathrm{fs}}$ refers to leaving some part of the solution as a free strand (fs, green regions of Fig 1), $\Delta G_{ij}^{\mathrm{bp}}$ is the FE for forming a base pair (cyan circles in Fig 1), ΔG_{ij}^{t} is the FE at the terminal end of the stem (the 5' and 3' most position or tail of the stem, see labeling in Fig 1a), $\Delta G_{ij}^{\mathrm{H}}$ is the FE of a hairpin loop (H-loop, blue region of Fig 1a), $\Delta G_{ij,pq}^{\mathrm{I}}$ is the FE of a bulge or an internal loop (I-loop, blue regions of Figs 1b and 1c respectively), and $\Delta G_{ij,\{pq\}}^{M(k)}$ is the FE of a multibranch loop (MBL, blue region of Fig 1d). These individual terms will be explained subsequently.

The first term ($\Delta G_{ij}^{\text{fs}}$) indicates that i consists of an unpaired base sticking out from $i+1$ ($\Delta G_{i+1,j}$), or j has an unpaired base jutting out from $j-1$ ($\Delta G_{i,j-1}$), or both cases are true ($\Delta G_{i+1,j-1}$). Hence, the best free strand (fs) solution on $[i \cdots j]$ is evaluated as follows,

$$\Delta G_{ij}^{\text{fs}} = \min\left\{\Delta G_{i+1,j}, \Delta G_{i,j-1}, \Delta G_{i+1,j-1}\right\}. \tag{18}$$

The next term involves the Turner energy rules for base pairing formation. There are two cases here: (1) it can be a closing base pair for some loop (an H-loop, I-loop, or an MBL) or (2) it can be any other position in the given stem. The closing point will be discussed later in this section. Dinucleotide base pairs within the stem are a straight calculation

$$\Delta G_{ij}^{\text{bp}} = \Delta\Delta G_{ij}^{\text{bp}} + \Delta G_{i+1\,j-1} \tag{19}$$

where $\Delta G_{i+1\,j-1}$ is whatever contents were previously evaluated at $(i+1, j-1)$, which could be either a closing bp or another stem bp, and $\Delta\Delta G_{ij}^{\text{bp}}$ is the FE for dinucleotide bp formation based on the Turner energy rules [56]. The base pairs in the Turner energy rules consist of the base pair at (i, j) within the context of the nearest neighboring base pair at $(i+1, j-1)$. For the head of a stem with a closing point at $(i+1, j-1)$, the first evaluation of a bp contribution happens at (i, j).

The tail end of the stem can also be closed with a corresponding FE correction for

the nearest neighboring base at the terminal end of the stem,

$$\Delta G_{ij}^{t} = \min \left\{ \begin{array}{l} \Delta\Delta G_{i}^{\mathrm{bp},t} + \Delta G_{i+1,j}^{\mathrm{bp}}, \\ \Delta\Delta G_{j}^{\mathrm{bp},t} + \Delta G_{i,j-1}^{\mathrm{bp}}, \\ \Delta\Delta G_{i}^{\mathrm{bp},t} + \Delta\Delta G_{j}^{\mathrm{bp},t} + \Delta G_{i+1,j-1}^{\mathrm{bp}} \end{array} \right\} \tag{20}$$

where $\Delta\Delta G_{i}^{\mathrm{bp},t}$ is the FE for a terminal base jutting out from the 5'-most side of the stem and $\Delta\Delta G_{j}^{\mathrm{bp},t}$ from the 3'-most side of the stem.

The H-loop is calculated using the Jacobson-Stockmayer (JS) equation along with a closing bp

$$\Delta G_{ij}^{\mathrm{H}} = \Delta\Delta G_{ij}^{\mathrm{C}} + \Delta\Delta G_{n}^{\mathrm{H}} \tag{21}$$

where $\Delta\Delta G_{ij}^{\mathrm{C}}$ is the FE for forming a closing a bp (case 1 mentioned above), $n = j - i - 1$ and $\Delta\Delta G_{n}^{\mathrm{H}}$ is the JS equation and has the form

$$\Delta\Delta G_{n}^{\mathrm{H}} = T(A_{JS} + \gamma k_{\mathrm{B}} \ln(n)) . \tag{22}$$

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 self-avoiding random walk correction [14,45] (Section 2) and T is the temperature.

The point $[i \cdots j]$ may also close a bulge or an I-loop (Fig 1b and 1c,

respectively)

$$\Delta G_{ij}^{\mathrm{I}} = \Delta\Delta G_{ij}^{\mathrm{C}} + \Delta\Delta G_{n}^{\mathrm{I}}(n_1, n_2) + \Delta\Delta G_{\mathrm{I}}^{\mathrm{asym}}(n_1, n_2) + \Delta G_{pq}^{\mathrm{bp},t} \tag{23}$$

where $n_1 = p - i$, $n_2 = j - q$, $n = n_1 + n_2$, $\Delta\Delta G_{\mathrm{I}}^{\mathrm{asym}}(n_1, n_2)$ adds corrections for asymmetry in the loops ($n_1 \neq n_2$) [57] as generally implemented [9,56], $\Delta\Delta G_{n}^{\mathrm{I}}(n_1, n_2)$ is an internal loop penalty that depends primarily on the total enclosed length n but also depends on the asymmetry of the loop, when $n_1 \neq n_2$. For a bulge, either $n_1 > 1$ or $n_2 > 1$, but not both. The value of $\Delta\Delta G_{n}^{\mathrm{I}}(n_1, n_2)$ is essentially $\Delta\Delta G_{n}^{\mathrm{H}}$ with $n = n_1 + n_2$, but A_{JS} differs somewhat. Moreover, for both the H-loops ($n \leq 8$), bulges and I-loops ($n \leq 4$), the actual penalties are generally obtained from experimental measurements rather than Eqn (22).

Finally, $\Delta G_{ij,\{pq\}}^{M(k)}$ bifurcates the secondary structure into two independent sectors: one between $[i \cdots k]$ ($\Delta G_{i,k}$) and the other between $[k+1 \cdots j]$ ($\Delta G_{k+1,j}$). This is further broken down into whether the structure closes with a multibranch loop (MBL), or is just two independent regions

$$\Delta G_{ij,\{pq\}}^{M(k)} = \min \left\{ \begin{array}{l} \Delta\Delta G_{ij}^{\mathrm{C}} + \Delta\Delta G_{n,m}^{\mathrm{M}} + \min\limits_{i<k<j} \left\{ \Delta G_{i,k} + \Delta G_{k+1,j} \right\} \\ \min\limits_{i<k<j} \left\{ \Delta G_{i,k} + \Delta G_{k+1,j} \right\} \end{array} \right\} \tag{24}$$

where

$$\Delta\Delta G^{M}_{n,m} = T(C_0 + C_1\sum_{k=0}^{m} n_k + C_2 m), \tag{25}$$

C_0, C_1, and C_2 are all fitted parameters, m is the number of branches, $\{pq\}$ specifies the particular set of branches in terms of the tail of their respective stems and $n_k = p_{k+1} - q_k - 1$ is the length of the free-strand segments of the MBL in Fig 1d (blue region with $k = 0,...,m$, $q_0 = i$ and $p_{m+1} = j$). Branches consist of the stems that extend off from the MBL (Fig 1d). Note, $\Delta\Delta G^{\mathrm{C}}_{ij}$ is generally different depending on whether the closing bp is an H-loop, an I-loop or an MBL. However, the form is the same.

3.2. Using the DPA in the CLE models for RNA structure prediction

The form of the expression in Eqn (17) is the similar for the CLE model.

$$\Delta G^{cle}_{ij} = \min\begin{cases} \min\left\{\Delta G^{\mathrm{fs},cle}_{ij}, \Delta G^{\mathrm{bp},cle}_{ij}, \Delta G^{t,cle}_{ij}, \Delta G^{\mathrm{H},cle}_{ij}, \Delta G^{\mathrm{I},cle}_{ij,pq}, \Delta G^{M(k),cle}_{ij,\{pq\}}\right\} \\ \min\left\{\Delta G^{\mathrm{PK}}_{ij}\right\} \end{cases} \tag{26}$$

where $\Delta G^{\mathrm{fs},cle}_{ij}$ and $\Delta G^{t,cle}_{ij}$ are essentially treated and evaluated is the same way as Eqns (18) and (20) respectively. The term $\Delta G^{\mathrm{PK}}_{ij}$ refers to pseudoknots (PKs). The other terms have similar meaning as in Eqn (17) and their differences will be explained subsequently.

The base pair formation rules are significantly changed in the CLE model. First, the closing bp FE is changed to

$$\Delta\Delta G_{ij}^{\mathrm{C},cle} = \Delta\Delta G_{ij}^{\mathrm{C}} + \Delta\Delta G_{bp}^{cle}(j-i+1,\xi) \tag{27}$$

where

$$\begin{aligned}&\Delta\Delta G_{bp}^{cle}(n_{ij} = j-i+1,\xi)\\&= \frac{k_{\mathrm{B}}T}{\xi}\left\{\nu\delta\gamma\ln\left(\Psi_{\nu\xi}n_{ij}\right) - \zeta(\gamma,\delta)\left(1-1/(\Psi_{\nu\xi}n_{ij})^{\delta\nu}\right)\right\}\end{aligned} \tag{28}$$

corresponds to Eqns (10) and (11) with $\Psi_{\nu\xi} = \xi^{-1}(\xi/\lambda)^{1/\nu}$ and $\lambda = 2$. The parameters γ, δ and ν are usually set to 1.75, 2 and 1/2, respectively (Section 2) with $\Psi_{\nu\xi} = \xi/\lambda^2$. The explicit parameter $\gamma = 1.75$, and the implicit parameters $\delta = 2$ and $\nu = 1/2$ are all the same in Eqn (22); Part II, Section 6. The user can alter these parameters in the *vsfold5* and *vs_subopt* implementations.

The term $\Delta G_{ij}^{\mathrm{bp},cle}$ requires a correction for the global entropy and has the form

$$\Delta G_{ij}^{\mathrm{bp},cle} = \Delta\Delta G_{ij}^{\mathrm{bp}} + \Delta\Delta G_{bp}^{cle}(j-i+1,\xi) + \Delta G_{i+1\,j-1} \tag{29}$$

where $\Delta\Delta G_{ij}^{\mathrm{bp}}$ is weighted with the global entropy, $\Delta\Delta G_{bp}^{cle}(j-i+1,\xi)$. Finally, there is a corrective term for when the stem length (L_{stem}) is shorter than the Kuhn length (ξ), as discussed in Part II, section 5 of this series. As the stem length increases, the FE for the stem must be updated. Hence, whereas the value of $\Delta G_{ij}^{\mathrm{bp},cle}$ can be calculated in the standard way of a DPA, some information about stem length and some *backtracking* must be included in the procedures to evaluate Eqn (29).

The H-loop also applies Eqn (28)

$$\Delta G_{ij}^{\mathrm{H},cle} = \Delta\Delta G_{ij}^{\mathrm{C},cle} + \Delta\Delta G_{bp}^{cle}(j-i+1,\xi) \tag{30}$$

where the form of Eqn (21) is retained.

The case of an I-loop (or a bulge) is quite different

$$\Delta G_{ij}^{\mathrm{I},cle} = \Delta\Delta G_{ij}^{\mathrm{C},cle} + \Delta\Delta G_{bp}^{cle}(j-i+1,\xi) + \Delta\Delta G_{\mathrm{I}}^{\mathrm{asym}}(p-i,j-q) + \Delta G_{pq}^{t,cle} \tag{31}$$

where $\Delta\Delta G_{\mathrm{I}}^{\mathrm{asym}}(p-i,j-q)$ only retains the asymmetry aspects ($j-q \neq p-i$) of an I-loop [57] and a few other considerations associated with interior loops. There is no JS weight except for a small weight at $\Delta\Delta G_{bp}^{cle}((j-q)/2-(p-i)/2+1,\xi)$, where the midpoint of the I-loop is $(j+i-q-p)/2+1$.

Likewise, $\Delta G_{ij,\{pq\}}^{M(k),cle}$ bifurcates the FE between two independent sectors: one between i,k and the other between $k+1,j$

$$\Delta G_{ij,\{pq\}}^{M(k),cle} = \min\left\{\begin{array}{l}\Delta\Delta G_{ij}^{\mathrm{C},cle} + \Delta\Delta G_{bp}^{cle}(j-i+1,\xi) + \min\limits_{i<k<j}\left\{\Delta G_{i,k}^{cle} + \Delta G_{k+1,j}^{cle}\right\} \\ \min\limits_{i<k<j}\left\{\Delta G_{i,k}^{cle} + \Delta G_{k+1,j}^{cle}\right\}\end{array}\right\} \tag{32}$$

where in the CLE model, there is are none of the penalties. This is because the CLE model focuses on stems. Some consideration of coaxial stacking [58-60] is included and other information could be included in the analysis such as flexibility of the

branches.

The major difference between Eqn (17) and (26) is the additional processing of pseudoknots in a separate buffer and then either grafting the result onto ΔG_{ij}^{cle} or choosing the secondary structure buffer. The details of pseudoknots and their treatment are explained in in Supplement 2 of Ref [10]. Briefly, there are two basic motif for a pseudoknot, the core PK (or H-type pseudoknot), shown in Fig 1e, and an extended PK, shown in Fig 1f.

The extended PK involves joining existing secondary with a PK (Fig 1f). Examples of this are kissing loops. These are evaluated in place in the interval between i and j and for a particular extended PK, the closing point of the structure is (i, j), just like a stem. A PK, which is located at (i, j), has handles that explain how the PK should be processed from position (i, j). The interested reader should consult Suppl 2 of Ref [10] for details.

Because of the 5' to 3' folding direction in the implementation of the CLE model (*vsfold5*), the core PK (H-type) takes advantage of a lead sequence L_{ls} ahead at $j + L_{ls}$ and saves the result until $\Delta G_{i,j+L_{ls}}^{\mathrm{PK}}$ is evaluated. Hence, if the best solution at $[i \cdots j]$ is a PK and $\Delta G_{i,j}^{\mathrm{PK}}$ represents an H-type PK, then it was actually previously evaluated at $\Delta G_{i,j-L_{ls}}^{\mathrm{PK}}$ (with $i < j - L_{ls}$). This is to help prevent the possibility of secondary structure "crowding out" a good PK solution.

Even when ΔG_{ij}^{cle} favors $\min\{\Delta G_{i,j}^{\mathrm{PK}}\}$ between i and j, a tag is added at (i, j) to allow further stem building. This prevents the situation where a PK is selected at (i, j) that blocks a stem from forming with a more favorable FE between

(i', j') ($i' < i$ and $j' > j$) and (i'', j'') ($i'' \geq i$ and $j'' \leq j$) when the DPA evaluates $[i'....j']$. This sometimes happens, though not very often for real RNA in our experience.

The language in the previous paragraph is, unfortunately, a little vague because *vsfold5* works with *effective stems*; stems that can have small bulges or I-loops breaking the contiguous dinucleotide bp pattern yet do not break the definition of a stem. Suppose that there is a simple stem between $(i - k_t, j + k_t)$ and $(i + k_h, j - k_h)$ with $k_t > 0$ (the tail) and $k_h \geq 0$ (the head), Fig 1a. Then the bp (i, j) is clearly contained in this stem. Now imagine that we add some "defects" to this perfect contiguous stem with a couple of small I-loops but keep (i, j) in part of one of these stems. This is the image of an *effective stem*; a stem that may have defects, but functions as a single unit in all other respects. The details are explained in the *vsfold5* manual and do not seem important for this discussion. The important point is that v*sfold5 backtracks* and *updates* the local stem structure (or structures). Stem bps are not merely calculated once and, in all subsequent operations, simply added to the accumulating solution as in Eqn (17).

In this respect, *vsfold5* is not a simple DPA, at least as originally proposed by Bellman [1] and as applied to computer-based algorithms in textbooks like Cormen et al. [2]. On the other hand, thinking in terms of effective mers (or maybe *epimers*) with notation like Eqn (14), the image is still that of a DPA, although there is a "fuzzy region" where regular updating is required. Excluding PKs, the solution is still optimal; however, there is a band in the matrix elements where the solution may be transitory. The overall coarse-grained solution requires more details than a cavalier evaluation of *epimers*, yet neither is it appropriate to think only on a *monomer* scale of

resolution as in the traditional implementations for RNA secondary structure.

By natural extension, the entropy model for the PKs operates on the same principle. However, because of the greater proximity of stems and the possibility for RNA structure to fold up in complex ways, the algorithms for scanning PKs must infer a considerable amount of structural information to build physically feasible PKs. Post structural editing is also required. The details are published in the Supplement of Ref [10]. The "editing", in particular, suggests some similarities with kinetic models.

3.3. Using kinetic models for RNA structure prediction

There are a vast variety of alternative approaches to the DPA that try to model RNA folding using some kind of sampling or kinetic algorithm. Briefly, the earliest approaches involved Monte Carlo methods [61,62] that have evolved into the genetic algorithm [63,64]. Other approaches are based on the kinetics of RNA folding [21,22,65,66]. Still other methods deal much closer to 3D structure and folding with various levels of coarse-grained polymer chain approximation [16,67,68].

Kinetic models work from the assumption that the biopolymer doesn't necessarily search its entire conformation space for the minimum free energy. Rather, it selects the best local solution that is thermally stable and cannot easily come undone. Stable in such a model is defined by the likelihood that the structure can unfold and be captured by a structure with a lower free energy. There are merits to this approach because it is not necessary to evaluate all the solutions in the search space for the best one, only the best one at the moment. These approaches are sometimes successful at finding native state structure.

Moreover, an exhaustive search for PKs is believed to require an exponential

number of computations to test every possible configuration (NP-hard) [69,70]. Therefore, some type of heuristic is required to address the pseudoknot problem. Rivas and Eddy [71] proposed a nearly complete DPA that processes at $O(N^6)$. More recent approaches offer faster prediction of [72] order $O(N^4)$. Therefore, the alternative of using a kinetic approach that might find the structure faster is a reasonable proposition.

A fundamental assumption in the DPA is that the recursion elements in prior solutions do not require any revision with new information. Kinetic models can have some advantage here if there is restructuring after connection. For example, if there is post formation swapping of neighboring base pairs. Although we have not observed a significant amount of such swapping because most such swaps rarely offers significant gain, the possibility still remains and sometimes happens.

The algorithm in *vsfold5* utilizes mapping to parse through the existing built up structure and, in some cases, to edit prior structure to fit a good pseudoknot candidate into a particular configuration. The heuristic assumes that once a good candidate is found, then editing can occur on the previously determined structure. Even in secondary structure calculations using *vsfold5*, to a limited extent, the stem FE is revised as the stem lengthens from its initial stub. When the Kuhn length is long, this can involve a large number of consecutive base pairs before the FE stabilizes to an incremental value as in the case of traditional DPAs. In this respect, there are some elements of a hybrid kinetic model implicitly built into the heuristics of the *vsfold5* approach.

4. Path independence for a secondary structure model

The most important factor in insuring that we can find a minimum free energy (mFE) is that the model itself is path independent. A consequence of path independence is that, for a polymer with a unique minimum free energy (mFE), all folding pathways will eventually reach the ground state and this ground state yields the same free energy (FE) regardless of what path was taken to arrive at the native state.

RNA secondary structure represents a special case where every base-pair (i.e., cross-link) combination (i, j) and (i', j') in the set satisfies one of the following conditions (with $i < j$, $i' < j'$ and $i' \neq i \neq j' \neq j$): $i < i' < j' < j$, $i', j' < i$, $j < i', j'$ or $i' < i < j < j'$.

Lemma 1:

Given the set of structures $\{S\}$ *contain only secondary structures, the minimum free energy (mFE) is non-degenerate and distinguishable in* $\{S\}$*, the folding model is path independent and no rearrangements occur after formation of the secondary structure: then the structure with the mFE can be found using the dynamic programming algorithm.*

Proof: The mFE is assumed to exist. If we can construct at least one pathway and all pathways to the ground state are independent, then we automatically find the mFE if we can find just one such path. Since, by definition, this must also include the thermodynamically most-probable folding pathway (TMPFP), all we have to do is search for this one pathway to find the mFE. Therefore, the DPA, which finds the

optimal solutions for every sequence fragment, can also find the mFE.

In the remainder of this section, we explore the characteristics of TMPFP within the framework of the CLE model.

In thermodynamics, all paths are, in principle, possible in a folding model that claims reversibility; however, not all such paths have equal thermodynamic probability. In Fig 2, we show the full range of reversible pathways for a sequence of RNA with Kuhn length ξ ($=5$ nt) as it transitions between the denatured state to the native state through a collection of simplified intermediates. Here we assume the sequence contains residues that yield no strong binding interactions except for the following specific binding sites (Fig 2a): 1 with $\overline{1}$, 2 with $\overline{2}$, and 3 with $\overline{3}$, where the bar indicates the strand's complement. In Fig 2b, D is the denatured state, N is the full native state, I_k represents the transition intermediates along the path between D and N, and $\overset{k}{\longleftrightarrow}$ refers to the path taken.

Using the notation introduced at the end of Section 2 (and first used in Ref [47]), let $\tilde{i}$ and $\tilde{j}$ represent effective mers where $\tilde{i} < \tilde{j}$ and the effective mers are all numbered sequentially from 1 to $\tilde{N}$ with $\tilde{N}$ the total number of effective mers. Let the distance between these effective mers $\tilde{i}$ and $\tilde{j}$ be referred to here as $r(\tilde{i}\tilde{j})$. Let $r_i(\tilde{i}\tilde{j})$ be the initial distance separating effective mers $\tilde{i}$ and $\tilde{j}$ and let that distance represent the denatured state (D) where $r_i(\tilde{i}\tilde{j}) = \overline{r}(\tilde{i}\tilde{j}) = r_{rms,\tilde{i}\tilde{j}}$ (the $\tilde{i}\tilde{j}$-rmsd). From Eqn (15),

$$r_{rms,\tilde{i}\tilde{j}}^2 = \xi \overline{N}_{\tilde{i}\tilde{j}} b^2 = \tilde{N}_{\tilde{i}\tilde{j}} (\xi b)^2 \tag{33}$$

where ν in Eqn (1) is assumed to have the value $\nu = 0.5$. Let $r_f(\tilde{\tilde{ij}})$ be the final distance separating $\tilde{\tilde{ij}}$ and let that distance represent the native state (N: no *italics*) where $r_f(\tilde{\tilde{ij}}) = \lambda b$. Working from Eqn (9), the leading term in the entropy equation has the form

$$\begin{aligned}\Delta S(\tilde{\tilde{ij}}) &= S\left(r_f(\tilde{\tilde{ij}})\right) - S\left(r_i(\tilde{\tilde{ij}})\right) \\ &= k_{\mathrm{B}}\left\{\delta\gamma \ln\left(r_f(\tilde{\tilde{ij}}) / r_i(\tilde{\tilde{ij}})\right) - \zeta(\delta,\gamma)\left[\left(\frac{r_f(\tilde{\tilde{ij}})}{\overline{r}(\tilde{\tilde{ij}})}\right)^{\delta} - \left(\frac{r_i(\tilde{\tilde{ij}})}{\overline{r}(\tilde{\tilde{ij}})}\right)^{\delta}\right]\right\}\end{aligned} \tag{34}$$

where $[\overline{r}(\tilde{\tilde{ij}})]^{\delta} = [\xi^{1-\nu}\overline{N}_{\tilde{ij}}^{\nu} b]^{\delta}$ in Eqn (9). Using $r_i(\tilde{\tilde{ij}}) = \overline{r}(\tilde{\tilde{ij}}) = r_{rms,\tilde{ij}}$ and $r_i(\tilde{\tilde{ij}}) >> r_f(\tilde{\tilde{ij}})$ for large $\overline{N}_{\tilde{ij}}$ (Eqn (33)), it follows that

$$\Delta S(\tilde{\tilde{ij}}) \sim \delta\gamma k_{\mathrm{B}} \ln\left(r_f(\tilde{\tilde{ij}}) / r_i(\tilde{\tilde{ij}})\right) + k_{\mathrm{B}}\zeta(\delta,\gamma)$$

where we assume the usual values $\gamma = 1.75$ [14,45], $\delta = 2$ and, from Eqn (11), $\zeta(\gamma,\delta) = \gamma + 1/2$.

The major contribution to the folding kinetics begins with the diffusion of the polymer chain. Suppose we have two sites $\tilde{\tilde{ij}}$ and $\tilde{\tilde{i'j'}}$ whose binding FE is identical (where the binding FE for RNA consists of the Turner energy rules [53,56,73] when applied to a full stem: $\Delta G^{\mathrm{bp}}_{stem}$). In such a case, the rate of folding will depend mainly on the global entropy over the majority of the course of diffusion to the bound state

$$k(\tilde{\tilde{ij}}, i \to f) = k_o \exp\left\{-\frac{\Delta G^{\mathrm{bp}}_{stem} - T\Delta S(\tilde{\tilde{ij}})}{k_{\mathrm{B}}T}\right\}. \qquad (35)$$
$$\sim C \exp\left\{\gamma\delta \ln\left(r_f(\tilde{\tilde{ij}}) / r_i(\tilde{\tilde{ij}})\right)\right\}$$

where T is the temperature, $k_o \approx k_{\mathrm{B}}T/h$, h is the Planck constant and C is a constant whose details are the same for all bonds in this simple example:

$$C = k_o \exp\left(-\frac{\Delta G^{\mathrm{bp}}_{stem}}{k_{\mathrm{B}}T} + \zeta(\delta,\gamma)\right).$$

.

Given $\bar{r}(\tilde{\tilde{ij}}) \sim r_i(\tilde{\tilde{ij}}) >> r_i(\tilde{i}'\tilde{j}')$, $r_f(\tilde{\tilde{ij}}) = r_f(\tilde{i}'\tilde{j}') = \lambda b$ (i.e., the binding interactions of $\tilde{\tilde{ij}}$ and $\tilde{i}'\tilde{j}'$ are identical), then the relative rates will be

$$\frac{k(\tilde{\tilde{ij}}, i \to f)}{k(\tilde{i}'\tilde{j}', i \to f)} \sim \frac{\exp\left\{\delta\gamma \ln\left(r_f(\tilde{\tilde{ij}}) / r_i(\tilde{\tilde{ij}})\right)\right\}}{\exp\left\{\delta\gamma \ln\left(r_f(\tilde{i}'\tilde{j}') / r_i(\tilde{i}'\tilde{j}')\right)\right\}} = \left(\frac{r_i(\tilde{i}'\tilde{j}')}{r_i(\tilde{\tilde{ij}})}\right)^{\delta\gamma}. \qquad (36)$$

In other words, the relative folding time depends mainly on the relative magnitude of the $\tilde{\tilde{ij}}$-rmsd. Since $r_i^2 = \tilde{N}(\xi b)^2 = (\tilde{j} - \tilde{i} + 1)(\xi b)^2$,

$$\left(\frac{r_i(\tilde{i}'\tilde{j}')}{r_i(\tilde{\tilde{ij}})}\right)^{\delta\gamma} = \left(\frac{\tilde{j}' - \tilde{i}' + 1}{\tilde{j} - \tilde{i} + 1}\right)^{\nu\delta\gamma} << 1. \qquad (36a)$$

Hence, the rate of folding $\tilde{\tilde{ij}}$ will be much slower than $\tilde{i}'\tilde{j}'$.

In fact, $\Delta G_{stem}^{\text{bp}}$ should be a function that is only turned on when the folding structure is within some critical proximity

$$\Delta G_{stem}^{\text{bp}} = \begin{cases} \Delta G, & r(\tilde{\tilde{ij}}) \le r_{\text{range}} \\ 0, & r(\tilde{\tilde{ij}}) > r_{\text{range}} \end{cases} \tag{36b}$$

where $r_{\text{range}} \approx O(\lambda b)$ is the critical region where binding can occur. Even if $\Delta G \to -\infty$, the chain still requires time to come within proximity of the actual site. Therefore, Eqn (36a) actually holds for any scenario at least in terms of folding times and the influence of the stem binding interactions can only be invoked when the stem is well within r_{range} of meeting this condition.

In the full CLE equation, Eqn (12) becomes

$$\Delta S_{cle} = \Delta S_{\xi\delta\gamma} + \sum_{\{\tilde{\tilde{ij}}\}} \Delta S_{stem}^{(g)}(\bar{N}_{\tilde{ij}}, \xi). \tag{37}$$

where $\Delta S_{\xi\delta\gamma}$ is the local entropy in Eqn (12) and is treated as a constant in this study (see Part II, Sections 3 and 4 for details). The second term ($\Delta S_{stem}^{(g)}(\bar{N}_{\tilde{ij}}, \xi)$) is Eqn (16) and expresses the global entropy with effective mer separation distance $\tilde{N}_{\tilde{ij}}$ and a physical chain separation distance $\bar{N}_{\tilde{ij}} = \xi \tilde{N}_{\tilde{ij}}$ (Eqn (15)). The global entropy is summed over all cross links $\{\tilde{\tilde{ij}}\}$. In Ref [11], it was shown that Eqn (37) satisfies Gaussian statistics. Different folding paths only change the order in which the

individual entropies are added, but do not change the total sum.

In the notation of *epimers* (Section 2), for every *stem* $(\tilde{i},\tilde{j})$ and $(\tilde{i}',\tilde{j}')$ (where $\tilde{i}<\tilde{j}$, $\tilde{i}'<\tilde{j}'$ and $\tilde{i}'\neq\tilde{i}\neq\tilde{j}\neq\tilde{j}'$), RNA secondary structure satisfies one of the following conditions: $\tilde{i}<\tilde{i}'<\tilde{j}'<\tilde{j}$, $\tilde{i}',\tilde{j}'<\tilde{i}$, $\tilde{j}<\tilde{i}',\tilde{j}'$ or $\tilde{i}'<\tilde{i}<\tilde{j}<\tilde{j}'$. In all such instances, $\Delta S_{\tilde{i}\tilde{j}}^{(g)}(\bar{N}_{\tilde{i}\tilde{j}},\xi)$ can be folded in any order and generate a unique structure because it is either folded in a separate region ($\tilde{i}',\tilde{j}'<\tilde{i}$ or $\tilde{j}<\tilde{i}',\tilde{j}'$) or as a sub-region ($\tilde{i}<\tilde{i}'<\tilde{j}'<\tilde{j}$ or $\tilde{i}'<\tilde{i}<\tilde{j}<\tilde{j}'$) and we assumed there exists a unique mFE for the final state.

In Eqn (15), the $\tilde{i}\tilde{j}$-rmsd is $r_{\tilde{i}\tilde{j}}^{\,2}=(\tilde{j}-\tilde{i}+1)(\xi b)^2$. Viewed as a continuous function $r\propto\sqrt{\tilde{N}}$ which forms a parabola around the $\tilde{N}$ axis. Now let $r_{rms,1}$ define the $\tilde{i}\tilde{j}$-rmsd for stem 1, where $\tilde{i}$ and $\tilde{j}$ point to the midpoint of the stem. Similarly, let $r_{rms,2}$ and $r_{rms,3}$ correspond to the $\tilde{i}\tilde{j}$-rmsd for stem 2 and 3, respectively. This $\tilde{i}\tilde{j}$-rmsd is diagrammed in Fig 3a. The final structure, where stems 1, 2 and 3 are fully formed, is diagrammed in Fig 3b. The order of folding of the structure along the TMPFP, corresponding to the structures shown in Fig 2, is diagrammed in Fig 3c along with the corresponding structures. Based on Eqn (36a) and (36b), the folding will be fastest for stem 1 and slowest for stem 3.

To precisely model the intermediate structures in Fig 2 (and particularly the interior loop regions), the entropy for the specific configurations of $r_i(\tilde{i},\tilde{j})$ and $r_f(\tilde{i},\tilde{j})$ (for all relevant members) would need to be evaluated in Eqn (37); however, for $r_3>>r_2>>r_1$, the zeroth order approximation is the separate entropy of each stem: $\Delta S_1^{(g)}(\bar{N}_1,\xi)$, $\Delta S_2^{(g)}(\bar{N}_2,\xi)$ and $\Delta S_3^{(g)}(\bar{N}_3,\xi)$. Therefore, given $\Delta G_{stem}^{\mathrm{bp}}$ is equal for

all these stems and the folding is confined to the states given in Fig 2 for all parts of the folding process, then the folding is most likely to follow the course $S_0 \to S_1 \to S_2 \to S_3$ (Fig 3c).

Since Eqn (37) does not depend on the order of the summation, it follows that the CLE describes a path independent process. Therefore, for RNA secondary structure, the CLE equation will yield entropy that is path independent and reversible for every intermediate structure all the way to the native state. Hence, the DPA is guaranteed to yield an optimal solution. Moreover, because the diffusion times are a function of the sequence length separating two effective mers $(\tilde{i}, \tilde{j})$, the DPA follows the TMPFP

$$D \overset{1}{\leftrightarrow} I_1 \overset{12}{\leftrightarrow} I_{12} \overset{123}{\leftrightarrow} N \tag{38}$$

where the least probable path is

$$D \overset{3}{\leftrightarrow} I_3 \overset{32}{\leftrightarrow} I_{32} \overset{321}{\leftrightarrow} N. \tag{39}$$

This order could only be reversed on the TMPFP by finding a case where $\Delta G^{bp}_{stem}(3\overline{3}) << \Delta G^{bp}_{stem}(2\overline{2}) << \Delta G^{bp}_{stem}(1\overline{1})$. Is this likely to happen? If stem 3 is composed of GC, stem 2 of AU and stem 1 of GU, then this is a possibility. However, it seems unlikely for several reasons.

First, according to Eqn (36b), the weight of the Turner rules only become significant within a reasonable proximity of r_{range}.

Second, even in bacteria from hyperthermal vents, there is, at most, a 12% bias

toward higher GC content because of the higher temperatures [74]. Hence, there is still a lot of AU in the genome of hyperthermal bacteria. In general, one can understand the equal distributions of ACGU from Shannon entropy [75] where the information is maximized by maximizing the randomness of the sequences. A sequence with N_{A} base of A, N_{C} of C, N_{G} of G and N_{U} of U (with $N = N_{\mathrm{A}} + N_{\mathrm{C}} + N_{\mathrm{G}} + N_{\mathrm{U}}$) has $N!/(N_{\mathrm{A}}!N_{\mathrm{C}}!N_{\mathrm{G}}!N_{\mathrm{U}}!)$ different arrangements. A sequence rich in GC would tend to reduce this maximum (by reducing the distinguishability of different arrangements). This, in turn, would reduce the number of options for mutations. Hence, it reduces the "information" entropy (i.e., freedom of choice) as defined in Shannon entropy. We also saw in Part III in the discussion around Table 1, that the RNA is probably selecting sequences that are closer to an even mix of ACGU based on the size of the domains that are observed experimentally. This means that large fluctuations in $\Delta G_{stem}^{\mathrm{bp}}$ are not all that likely.

Third, this would eventually reach a limit because the energy difference between different base-pair (bp) combinations is not so large (maximum 1 kcal/mol) and the global entropy grows in a non-linear fashion. Taken together, whereas such a scenario is possible, it is not so likely for typical RNA sequences where equal fractions of ACGU are found. Moreover, even if a case could be found, the FE does not depend on the order of the operations in Eqn (37), only on the final FE. Likewise, for secondary structure, the order is not important in calculation. Therefore, the DPA can find that optimal solution.

Since, in thermodynamic equilibrium, the result only depends on the initial and final conformations, and the model is completely reversible, the result can be calculated as though the process always proceeds along the TMPFP. The dependence of the

folding on r_3 in Fig 3 is consistent with the contact order model [11,16,25], which predicts that the longest separation distance between the monomers determines the rate [11,25-27].

Not only are we assured that the DPA essentially calculates along the TMPFP in a biological process, even if folding along the TMPFP actually deviates significantly from the DPA calculation recursion, we are guaranteed an optimal solution as long as there exists a minimum FE. The DPA can find the minimum FE even when challenged with a very peculiar order of the stem-stem binding. *Vsfold* takes advantage of this thermodynamic property to manage the sequential folding efficiently.

5. Limitations on the dynamic programming algorithm (DPA) for secondary structure

We have shown that if we restrict ourselves to secondary structure (ss) and if the free energy is non-degenerate (there exists a unique mFE), then the TMPFP can guide us to finding the correct structure and we can use the DPA to find this mFE. However, what happens if the free energy is degenerate and, in particular, if the ground state has more than one structure with the same mFE?

In Fig 4, two structures are shown whose free energies are degenerate: structure $A\bar{A}_1$ and $A\bar{A}_2$, where $\bar{A}_1$ and $\bar{A}_2$ refer to complements of A. Fig 4a shows the pairing of these stems in Rivas-Eddy Feynman type diagrams [71,76]. Figure 4b shows the overall folding pathways and resulting structures. In short, Fig 4 indicates that

$$\Delta\Delta G(A\bar{A}_1, A\bar{A}_2) = \Delta G(A\bar{A}_1) - \Delta G(A\bar{A}_2) = 0 .$$

Since only one structure is allowed, one of the stems cannot form. Since the DPA searches for a unique mFE, the DPA cannot distinguish which structure $A\bar{A}_1$ and $A\bar{A}_2$ is the mFE if $\Delta\Delta G = 0$.

Because there are a multitude of possible structures that can be generated for a sequence of length *N*, as many as 1.8^N for secondary structure alone [77], there is reason to think that some of these patterns could have the same free energy and that the minimum free energy could be a multitude of structures. This is why it is quite reasonable to search for other structures and consequently, to find suboptimal structures

and possibly alternative optimal structures. Although $\Delta G(\mathrm{A}\bar{\mathrm{A}}_1) = \Delta G(\mathrm{A}\bar{\mathrm{A}}_2)$, the indexing of the matrices for $\mathrm{A}\bar{\mathrm{A}}_1$ and $\mathrm{A}\bar{\mathrm{A}}_2$ in the DPA are different. If suboptimal structures are evaluated, then it is possible to scan over the solved set of structures and find $\mathrm{A}\bar{\mathrm{A}}_1$ or $\mathrm{A}\bar{\mathrm{A}}_2$. Therefore, it is possible to find structures with degenerate FEs with a systematic backtracking strategy.

The type of structure that is observed in this sequence is more likely to depend on the way the RNA is folded. Let $N(\mathrm{A}\bar{\mathrm{A}}_1)$ be the number of mers between $\mathrm{A}\bar{\mathrm{A}}_1$ and similarly $N(\mathrm{A}\bar{\mathrm{A}}_2)$. If one simply denatures the RNA and refolds it, both $\mathrm{A}\bar{\mathrm{A}}_1$ or $\mathrm{A}\bar{\mathrm{A}}_2$ have equal opportunity to compete. Based on the dependence of $\Delta S^{(g)}_{stem}(\bar{N}(\mathrm{A}\bar{\mathrm{A}}_1))$ and $\Delta S^{(g)}_{stem}(\bar{N}(\mathrm{A}\bar{\mathrm{A}}_2))$ in Eqn (16), if $N(\mathrm{A}\bar{\mathrm{A}}_1) < N(\mathrm{A}\bar{\mathrm{A}}_2)$, then Eqn (36a) suggests that a higher fraction of structures containing $\mathrm{A}\bar{\mathrm{A}}_1$ will be observed because the $\tilde{\tilde{ij}}$-rmsd for $N(\mathrm{A}\bar{\mathrm{A}}_2)$ is larger and therefore, the folding time will take longer (statistically). If $N(\mathrm{A}\bar{\mathrm{A}}_1) = N(\mathrm{A}\bar{\mathrm{A}}_2)$, then equal fractions of both structures are likely.

If instead, the RNA is transcribed in the usual way (5' to 3') in vivo with RNA polymerase [78] , the fraction should be initially biased toward $\mathrm{A}\bar{\mathrm{A}}_1$. (A similar in vivo folding process of N to C occurs for proteins [79] and therefore, protein folding would also show this behavior for a similar protein sequence.) Likewise, synthetic RNA is synthesized from 3' to 5' [80] and, given a sufficiently rapid rate of transcription, the initial fraction should be biased toward $\mathrm{A}\bar{\mathrm{A}}_2$. It is important to remember that

thermodynamic equilibrium will eventually eliminate this initial bias, approaching fractions similar to the refolding results over time. Nevertheless, the role of the thermodynamically most-probably folding pathway (TMPFP) is seen to play out in different folding scenarios. Aiming the DPA to follow the folding pathway is more likely to bring about agreement with the physical process.

A degenerate minimum FE is likely to be a problem for truly random sequences [29]. Nevertheless, we think it arguable that most functional biomolecules are unlikely to have degenerate mFEs because function depends on specific signaling and recognition. Such examples as we mention here in Fig 4 clearly defeat these purposes and it is likely that natural selection would quickly eliminate such poor candidates to increase the activity. Exceptions would include structures like riboswitches which we saw in Part III tend to have two state systems that are often just slightly different in FE. Even these structures are selected for equilibrium distributions that clearly favor one of the structures [81]. Disordered biomolecules [82-85] are some other candidates that may have multiple states, at least until encountering the target ligand. Nevertheless, it is also likely that such degenerate structures can result in genetic disorders and disease related mutations.

Regardless of the particular TMPFP, since the FE is equal, the choice made by DPA depends on the selection mechanism. Nevertheless, both structures are possible and both should be reported. Since a DPA approach can only report one answer, this is a limitation. This is why we have worked to expand *vsfold* to handle suboptimal structure calculations with *vs_subopt* (Part III). We saw that this was of some value in studying riboswitches. Currently, since *vsfold* computes the structure in the 5' to 3' direction, *vsfold* will certainly choose the first of the candidates in Fig 4 because the

sequential folding (5’ to 3’) is clearly organized to choose the first candidate with the shortest loop in Fig 4. The added functionality of calculating suboptimal structures helps to overcome these issues.

6. Path independence for a pseudoknot model

The adaptations to the DPA to build and evaluate the local stem structure only require some simple backtracking procedures to adjust the CLE for each bp that is added. In the case of pseudoknots, there are long range interactions due to accounting for 3D structural considerations and there are occasional cases where further editing occurs around the region where the PK is constructed. Whereas the 3D structural issues do not change prior solutions, the editing steps do. Therefore, the PK heuristic employs a hybrid of some kinetic features to the general DPA architecture. Here we show that the minimum FE structure for a pseudoknot (PK) is also a global minimum free energy (mFE) structure and can be found through folding the structure 5' to 3' only if the structure has a unique mFE and if internal rearrangements after folding can be neglected. Minimal rearrangements such as exchanging chains in the PK stem or adding stem-stem packing interactions are not guaranteed to yield a mFE solution.

We return to regular monomers i and j in this discussion. Let i and j represent a starting and ending position of a segment of an RNA sequence such that $1 \le i < j \le N$ where N is the sequence length. Let the notation $[i \cdots j]$ indicate the sequence between i and j including i and j. The arrangements of base pairs in RNA pseudoknot structures, which involve base pairing of the form $i < i' < j < j'$ or $i' < i < j' < j$, distinguishes pseudoknots from standard definition of RNA secondary structure that are described in Sections 4 and 5. Examples of PKs are shown in Fig 1e and 1f.

Lemma 2:

If a pseudoknot (PK) structure has a unique minimum FE, the folding to form the PK is path independent and there are no internal rearrangements upon formation of a PK, then the thermodynamically most-probable folding pathway (TMPFP) also yields the minimum free energy for a pseudoknot.

Proof: The easiest way to see this is so is by the example shown in Fig 5. Fig 5a shows the folding pathways to form a core pseudoknot (H-type). Because we assume no internal rearrangements, the folding of the core PK is path independent regardless of whether the system might fold in a biological process of 5' to 3', a synthetic process of 3' to 5', or "refolding" in which all parts of the sequence can fold simultaneously. Fig 5b shows the folding pathways to form an extended PK. The folding structure is far more complicated. Particularly notable are the dashed lines indicating some unique pathways available in the independent process of refolding. In refolding, diverse parts of the sequence can fold simultaneously with a high likelihood of the two stem-loops at the bottom of Fig 5b forming first. Nevertheless, every path is accessible, there are no rearrangements, and the FE of the PK is assumed to be unique and a minimum. Since any pathway is permitted, the TMPFP is also allowed. It follows that, for this (comparatively) simple system, the DPA is able to find the minimum FE by following the sequence of calculations from 5' to 3' most characteristic of the actual folding environment and therefore the TMPFP.

A schematic of RNA folding along the TMPFP is shown in Fig 6 where an initial folding structure is shown in Fig 6a and the final structure in Fig 6f. Structures jutting

out from the main chain are called level 0 structures, which are indicated by the hatched black lines surrounding different structures in Fig 6. As the structure grows (progressing along the arrows in Fig 6), the levels increase with the previous level 0 structures shifting to subdomains at level 1 (c.f., Figs 6a and 6b, the purple hatched lines) and level 2 (c.f., Figs 6d and 6e, the green hatched lines). Fig 6a shows two independent stem-loops jutting out of the main chain at level 0. These become subdomains of an MBL in Fig 6b and the MBL becomes the level 0 structure and the branches make up subdomains. In Fig 6c, a new stem-loop begins forming 3' of the MBL in a separate domain of level 0 structure. This stem-loop also becomes a subdomain via an I-loop in Fig 6d. Fig 6e shows yet another stem-loop forming and the MBL and second stem-loop complex forming an extended PK. The level 1 structures stem-loops are promoted to level 2 structures, the MBL and stem-loop complex to level 1. Fig 6f shows the new stem-loop forms a core PK and the extended PK is observed to go through some internal rearrangement. These structures show a modular behavior and hierarchical folding largely as proposed by Westhof and coworkers [17,86-88] The core pseudoknot (H-type pseudoknot) and extended pseudoknot modules of Fig 6f that have a 5'-input and a 3'-output and exist as independent (thermodynamically stable) entities (domains) [10], depicted in Figs 1e and 1f respectively.

The first question is whether forming the PK can corrupt an alternative secondary structure forming in the same local region $[i \cdots j]$ of an RNA sequence. Though it occasionally occurs, it is not so common for biological RNA to share a common closing point (i, j) in a stem of secondary structure with the 5'-input/3'-output of a pseudoknot (a PK module).

To overcome this potential problem, in *vsfold*, two buffers are used. One buffer contains the best secondary structures and the optimal level 1 (and greater) structures (the current structure and the previously solved modules). For extended PKs, the other buffer contains attempts to fit new structure against the first buffer with pseudoknots (if any are found) on the same interval $[i \cdots j]$. For core PKs, the PK search buffer scans ahead 3' of j with a leader free strand 2ξ nt in length. Therefore, a core PK found on $[i \cdots j]$ was determined prior to evaluation at $[i \cdots j]$. This forward scanning is important because sometimes the secondary structure can "crowd out" a good PK. When a PK module is found at some 5'-input/3'-output point (i, j) with a better free energy (FE) than the existing structure on the interval, the location is tagged and the pointer (or link, as described in Supplement 2 of Ref [10]) of whatever existing optimal secondary structure domains present on the interval $[i \cdots j]$ (without the PK) is saved. Although the point (i, j) is tagged as a PK because it is the mFE on $[i \cdots j]$, information about the secondary structure on that interval is not destroyed. For example, when moving to $[i-1 \cdots j+1]$, a stem that is replaced by a PK tag at (i, j) can be recovered when a contiguous part of that stem is preset at $(i-1, j+1)$. If the stem turns out to be more stable than the PK on the new interval, it will dominate the level 0 domain structures and, although optimal at (i, j), the PK will generally be ignored in further calculations (running 5' to 3'). The tag is preserved. In the CLE model, because it also considers Kuhn length, the strength of a stem can grow as its effective length increases (e.g., see the discussion in Part II, Section 5). Over a "fuzzy" interval of ξ (i to $i+\xi$ and $j-\xi$ to j), the FE can change.

In effect, the DPA is actually optimizing the best solution of two buffers. Given a unique mFE exists on $[i \cdots j]$, choosing the best solution (status quo vs new

pseudoknot) guarantees that an optimal solution is found in the modules comprising level 1 and greater structures if they remain otherwise unchanged.[1] Moreover, because the level 0 structures themselves build up gradually along the 5' to 3' folding pathway, and eventually become themselves level 1 and greater subdomains, the internal structure of the subdomains are also optimal as long as there are no significant internal rearrangements after folding and becoming structure at level 1 or higher.

We assume that it is given that we have a solution for the mFE of the secondary structure on any interval $[i \cdots j]$. We are assured that given a secondary structure with a mFE exists on the interval $[i \cdots j]$, we can find it with the TMPFP and DPA. If that secondary structure is also a (level 0) isolated secondary structure, then the structure is also the optimal secondary structure in $[i \cdots j]$.

Given structure (a) in Fig 7a, with mFE $\Delta G(\text{a})$, is the best secondary structure on $[i \cdots j]$, and let the structure in Fig 7b be the mFE for some core pseudoknot on the same $[i \cdots j]$ and let $\Delta G(\text{b}) < \Delta G(\text{a})$. Since the core PK (structure (b)) is a module on $[i \cdots j]$ and $\Delta G(\text{b})$ represents the mFE of this segment, according to Lemma 2, (b) is also the best structure on $[i \cdots j]$ and (b) is a registered (tagged) structure that exists on this interval.

If $[i \cdots j]$ is expanded to $[i' \cdots j']$, where

[1] In constructing the *vsfold* algorithm, we have of necessity assumed that only reasonably sparse distributions of PKs are found. In principle, such a strategy will distinguish the best structure given a mFE exists for each interval considered. However, practically speaking, building an algorithm to withstand the full onslaught of complexities of some test sequence that has wildly varying binding energy properties (if such exists) is surely not for the faint of heart. *Vsfold* is fairly robust, but has not been tested to an extreme level. The intended purpose of *Vsfold* is to solve observed biologically relevant RNA and no effort has been made to find artificial sequences that could achieve extreme levels of testing.

$$[i' \cdots j'] \rightarrow \begin{cases} i' < i,\ j' = j \\ i' < i,\ j' > j \\ i' = i,\ j' > j \end{cases} \tag{40}$$

then the interval is changed and we must know the set of structures that occupy this new interval. Hence, if further sequence is fit and $\Delta G(\mathrm{c}) < \Delta G(\mathrm{b})$ on $[i' \cdots j']$ (Fig 7c) according to Eqn (40), then structure (c) is the mFE. Likewise, if further we can say that $\Delta G(\mathrm{d}) < \Delta G(\mathrm{c})$, then structure (d) is the mFE on $[i' \cdots j']$, Fig 7d.

If these modules are then incorporated into multibranch loops (MBLs) or internal loops (I-loops) and, excluding special boundary conditions, the modules remain independent and form as before without change over the boundary. Therefore, if, within an I-loop or an MBL, the pseudoknot (b) on $[i \cdots j]$ or (d) on $[i' \cdots j']$ is found to yield the best connecting element, then the secondary structure plus this pseudoknot represent the mFE for the sequence. On the other hand, if (i, j) is an intermediate point on a stem and the forming stem is the best FE, then, whereas the PK tag remains, the solution set becomes the stem. With *vsfold*, for level 1 structures and greater, the modules are independent and therefore assumed to form a stable sub-domain that is not changed. With suboptimal structure calculations, selection is directed to finding suboptimal modules of the specified level.

Hence, giving that there exists some unique mFE on an interval $[i \cdots j]$ (and given that the number of viable PKs is not too excessive), we are assured that we can find that solution because we can explore two different buffers in a given interval (one containing the best level 0 domain structures and the other containing new pseudoknot information) to determine whether PKs, stems or a combination thereof form that best

structure.

Nevertheless, it is important to point out that there is some reasonable possibility of an internal sub-domain rearrangement (as in the transition between Figs 6e and 6f). As the structure compacts, there would be room for additional sub-domain interactions to form. There is no guarantee that the structure will not change. Nevertheless, though exceptions may exist, it is more likely that a precursor structure resembling the final one will naturally fold up, such that the degree of rearrangement should not be enormous. There are some provisions for some internal structural interactions worked into the *vsfold* algorithm when folding PKs. In particularly, considerable attention was given to chain swapping as in Fig 6e and 6f in the extended PK (and similarly for core PKs). Large scale restructuring would be very difficult to model this way.

7. Limitations on the dynamic programming algorithm (DPA) for pseudoknots

The problem of a degenerate ground state for secondary structures was considered in Section 5. Here, the case of pseudoknots (PKs) is considered.

In Fig 8a, a group of equally possible secondary structure loops and a pseudoknot linkage are shown. The weights are such that

$$\Delta G(\mathrm{A}\overline{\mathrm{A}}_1) = \Delta G(\mathrm{A}\overline{\mathrm{A}}_2) = \Delta G(\mathrm{A}\overline{\mathrm{A}}_3) = \Delta G(\mathrm{A}\overline{\mathrm{A}}_4) \quad \text{and} \quad \Delta G(\mathrm{A}\overline{\mathrm{A}}_1) = \Delta G(\overline{\mathrm{A}}_1\overline{\mathrm{A}}_3),$$

in short, *somehow,* their free energies are all equal with a similar thermodynamically most-probable folding pathway (TMPFP). Given this is so, which is indeed rather unlikely, the native state would contain two possibilities: N_{ss} and N_{pk} (Fig 8b the gray box to the right). The pathways are equally distributed, and any and all of these structures searched. In such an example, it is likely that *vsfold* would choose the pseudoknot simply for the fact that this structure could be created first along the TMPFP. There is no way for a DPA to decide which structure (N_{pk} or N_{ss}) is the correct one in such a case.

Therefore we make the following observations.

(1) *If a sequence contains a unique minimum FE (mFE) and there are no internal rearrangements, then the DPA strategy is sufficient to find the optimal structure.* This was shown in Sections 4 through 6. The DPA can be designed with the particular TMPFP model (e.g., 5' to 3' folding or refolding from the denatured state) because the variance in the average base pair FE is less than 1 kcal/mol and

therefore a significant determining factor on the rate mostly results from the global entropy (as shown in Section 2, Eqn (36a)).

(2) *If the mFE is degenerate and there are no internal rearrangements, then more than one solution is equally possible and the solution is not unique; however, evaluating suboptimal structures can overcome this weakness.* All that can be expressed in such a case is the distribution of structures at (or near) the mFE. In such a case, most likely the first structure encountered will be selected by the DPA. To find these alternatives, suboptimal structure approaches are required. Again, the DPA can be designed with the particular TMPFP model because there are no internal rearrangements and because the folding pathway will tend to follow the kinetics suggested by Eqn (36a). Since the DPA solves all structures and there are no internal rearrangements of the FE, these alternative mFE structures can be found by backtracking.

(3) *If the mFE is unique but there are internal rearrangements, then an optimal solution can only be conditionally guaranteed.* For example, if the PK is optimized by chain swapping after an editing operation, then the PK module is optimal and stable. Then, as in case (1), the DPA can be organized to follow a similar recursion order as the TMPFP. However, if the PK later disassembles, the FE is no longer optimal or requires corrections to make it optimal. Such procedures deviate, at least, from the general concepts of the DPA strategy.

(4) *If the mFE is unique, but the heuristic cannot detect the structure, then the approach will fail.* For example, if two completely closed stems should spontaneously swap chains with each other, then there is nothing for the *vsfold* heuristic to grab hold of. Both the DPA and the TMPFP would fail. The TMPFP assumes that global folding

dominates and that the variance is small in the local interactions and, fundamentally, the DPA depends on static solutions. It may still be possible to overcome this with a more sophisticated approach such as the $O(N^6)$ DPA proposed by Rivas and Eddy [71] or and Lyngsø and Pedersen [70]. However, the concept of "folding" is lost. The *vsfold5* heuristic assumes that there is some extant structure already formed that leads and directs the formation process.

(5) *If the base pair FE varies drastically and arbitrarily over the entire free energy surface, then an optimal solution can only be conditionally guaranteed.* The DPA depends on distinguishable free energies and works best when the fluctuations between successive steps are not large [1]. If bp FE varies drastically (e.g., jumping erratically between -1 kcal/mol and -10^6 kcal/mol), then case (1) and (2) could be solved with a DPA (in principle), but cases (3) and (4) are far from guaranteed. For the TMPFP, it is important to apply Eqn (36b) to the bp FEs. Nevertheless, the success of current approaches largely depends on the relatively small variance of the bp FE around ± 1 kcal/mole. A highly variable landscape would be better treated with other approaches.

In general, most of the problems so far encountered in practical application of this heuristic to RNA structure prediction including PKs satisfy the criteria in case (1). The RNA structures that form PKs often already have nearly correct modules of secondary structure ready for forming a PK, as in Fig 6a-e. Case (2) may happen, but perhaps most of the degenerate pairing is limited to non-Watson-Crick interactions in biological RNA. It appears the chain swapping in Fig 6f (case (3)), although allowed, rarely improves the FE. This is probably largely because this happens at the expense of

losing a bp in the complementary module. In the end, the improvement is negligible. It appears that natural selection has eliminated aberrant sequences and that these calculations are largely doable with a DPA and the heuristic strategy employed in *vsfold5*.

In case (2), natural selection is likely to have eliminated structures highly degenerate in FE in functional RNA because folding becomes highly error prone and no native state can be distinguished. Alternatively, natural selection may choose these structures if the purpose is to form disordered regions. Therefore, just as there are regions of disorder in proteins [82-85], it is likely that there are regions of disorder in RNA and DNA. For example, H-loop (Fig 1a) with lengths greater than 8 nucleotides are often evaluated using the Jacobson-Stockmayer equation and therefore have a uniform FE with no well-defined order.

Most uncertain is case (3) where internal editing is required to build a full PK. The two buffer system where secondary structure and PKs can be compared and a scratch buffer, which tests these editing solutions before writing them into the PK buffer, appears to be sufficient for many problems when editing on the PK yields a stable PK that does not break up at a later step in the DPA calculation. However, selection rules are not simple, and sometimes the PK can break up. In the object oriented code of *vsfold5*, many of these issues are addressed by various memory tags; however, this remains a stability issue with the heuristic. Because of the instabilities in the structure that result from this type of issue, it remains a current issue of research in developing future versions of this algorithm.

Whereas case (4) is possible, up to now, this has not proved to be a problem. The most likely situation where this might occur is when there is chain swapping between

two fully formed stems. For biological sequences, such a case is more likely when a mis-folded structure refolds to the correct one. Normally, this should be a suboptimal structure. The other possibility is the parallel alignment of two or more stem structures where the coupling is between stems. This can be addressed to some extent even in the current implementation if coupling information is available.

Cases (3) and (4) are most likely to occur if one attempts to fit randomly generated RNA sequences [29]. Such structures can exist in equilibrium with neither a means to distinguish which structure is correct nor a way to know if the minimum FE is found.

Case (5) is mentioned because it is important to understand the limits of these methods. If the FE surface took on extreme and arbitrary values that had no correlation with the natural folding of a sequence, for example, arbitrarily large negative binding free energies in arbitrary locations (i, j) in the sequence mixed with moderate values, this problem would become far more difficult. However, it is likely that the reason why a model with unrestrained delta functions appears so foreign is because the free energy expressions are too simple. The DPA depends largely on the fact that there is not so much variation in the free energy. It is because of the general regularity of the monomers and their binding responses that we can depend on a TMPFP strategy even in principle. If the energies actually varied so arbitrarily, as in (5), then it is likely that the only way to know a solution is to do an exhaustive search.

Therefore, there are limits to a general fitting heuristic (like the *vsfold5* approach) but for natural functional RNA, such heuristics are probably sufficient in general. If very complex landscapes were the subject of these problems, perhaps only an exhaustive search would be adequate. It is exactly because structures fold in a regular way in biologically relevant RNA that we can turn to heuristics to help make the

search for RNA structure more successful.

In conclusion, in this study, the underlying optimization method known as the dynamic programming algorithm (DPA) has been examined in the context of RNA folding. Under conditions in which no rearrangement of chains occur and a unique minimum free energy mFE exists, if the recursion in the DPA follows the thermodynamically most-probable folding pathway, then an optimal solution can be reached (for a given folding scenario). This would be true even when the problem involves pseudoknots. On the other hand, if the mFE is degenerate or there are internal rearrangements of the chains that occur, then the mFE cannot be guaranteed.

At least for typical RNA biopolymers, it appears the DPA is largely successful. Most functional RNA appear to have little or no degeneracy and do not appear to rearrange. The reason it is successful is mainly because the natural selection has essentially tuned the folding behavior to reach the native state. It is because we have of these regularities that it possible to succeed at these types of calculations.

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 Dr. Shingo Nakamura, Profs Kiyoshi Asai, Kentaro Shimizu, Shugo Nakamura, and Kazuya Sumikoshi and Dr. Yucong Zhu for their encouragement.

Figures

Figure 1

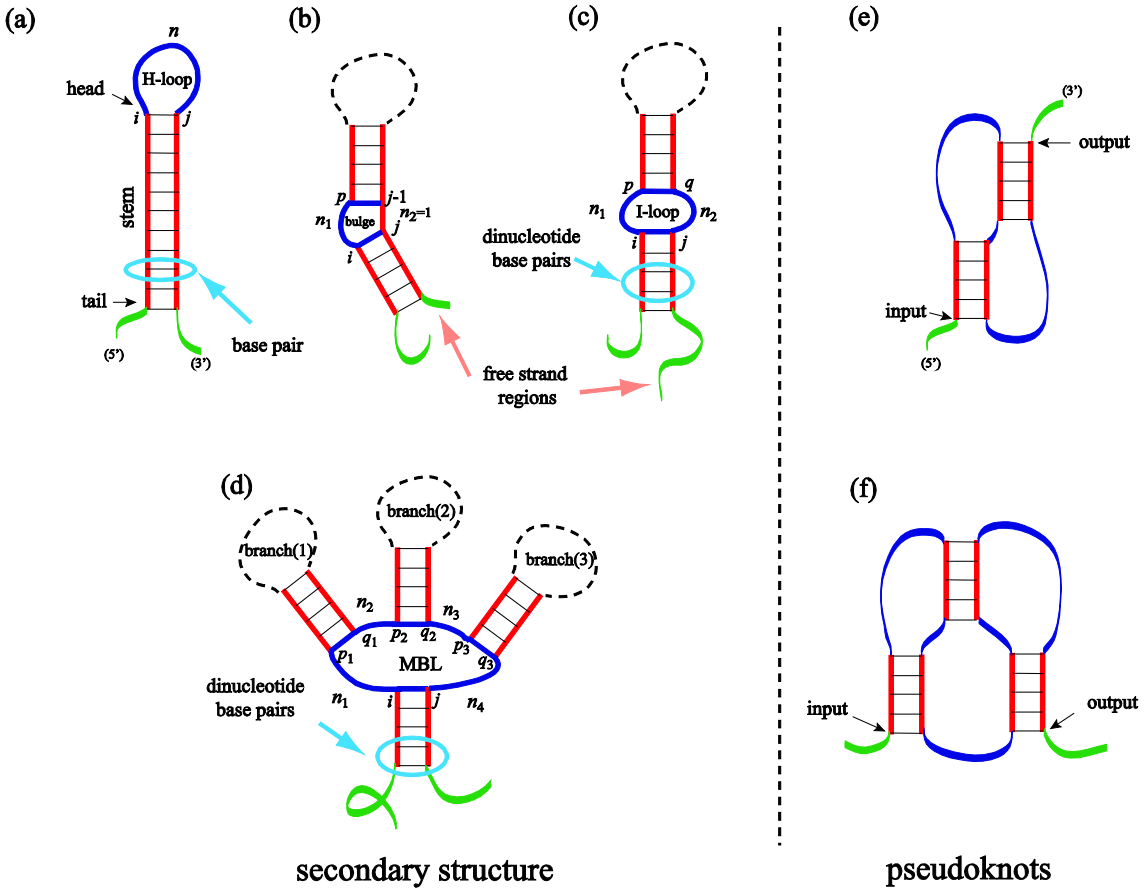

Figure 2

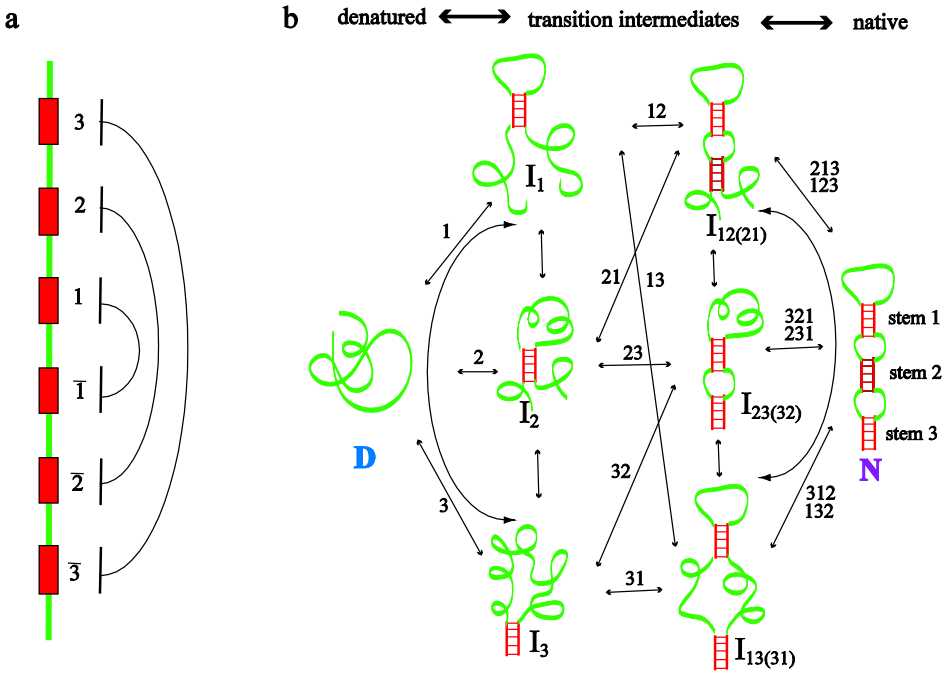
a
b
denatured
transition intermediates
native
3
2
1
1̄
2̄
3̄
D
I1
I2
I3
I12(21)
I23(32)
I13(31)
N
stem 1
stem 2
stem 3
1
2
3
12
21
13
23
32
31
213
123
321
231
312
132

Figure 3

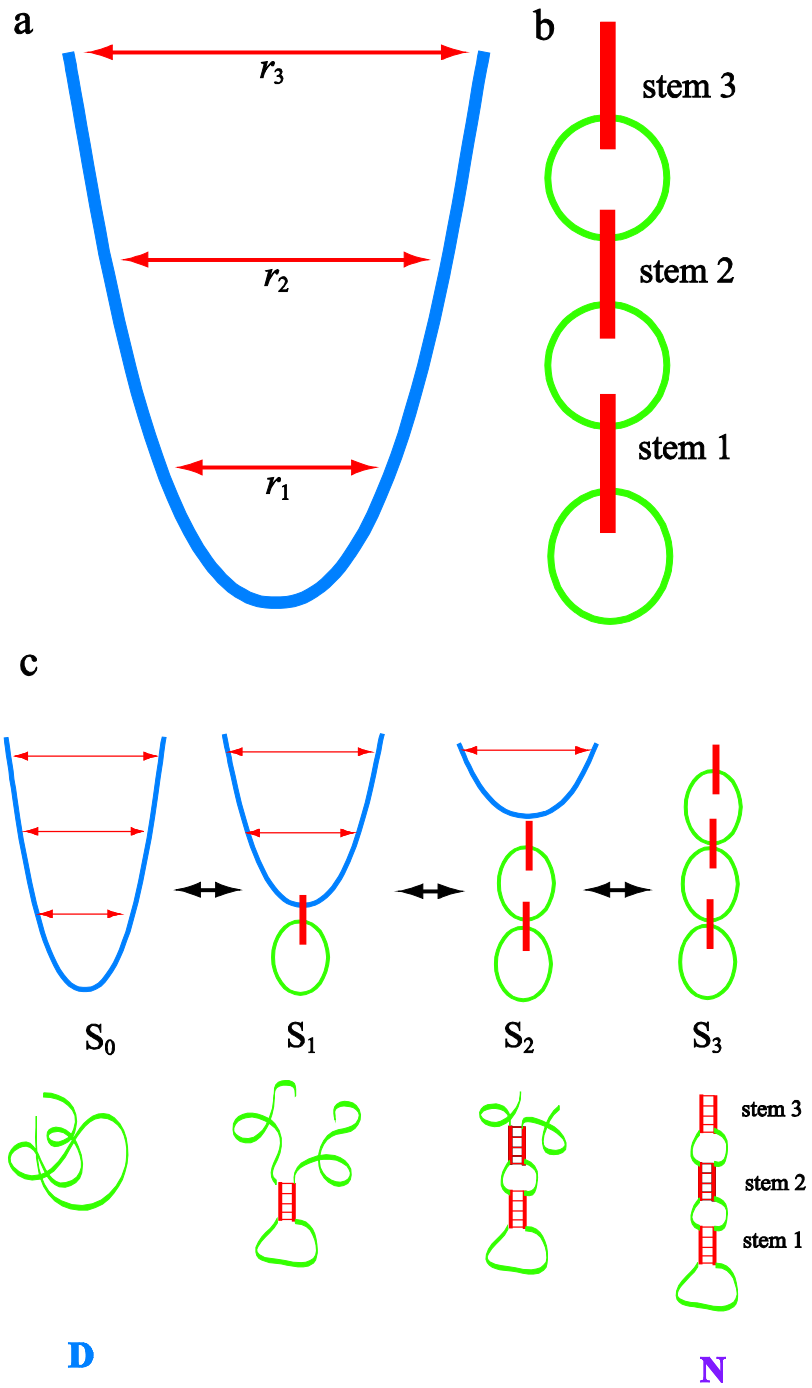

Figure 4

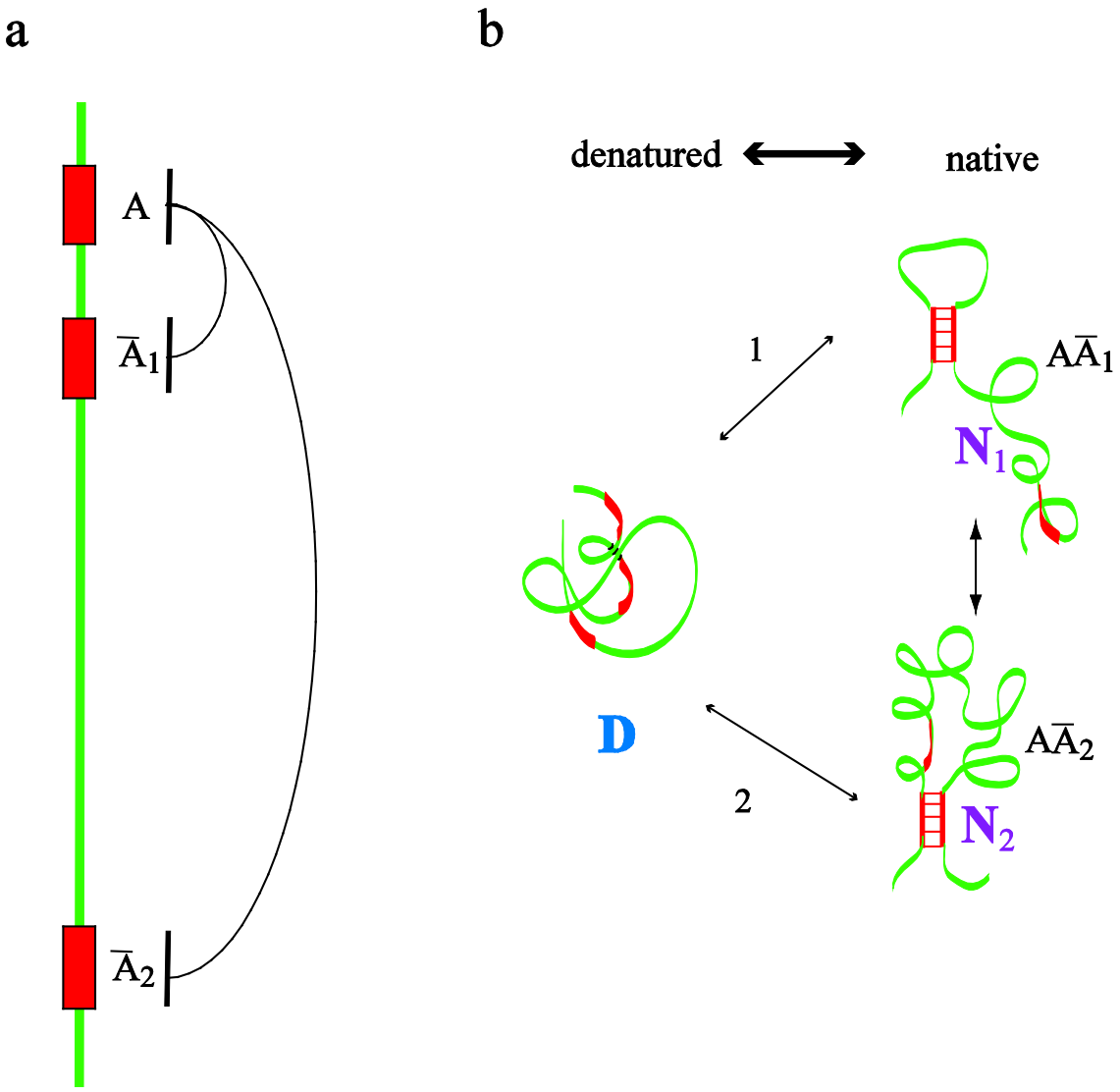

Figure 5

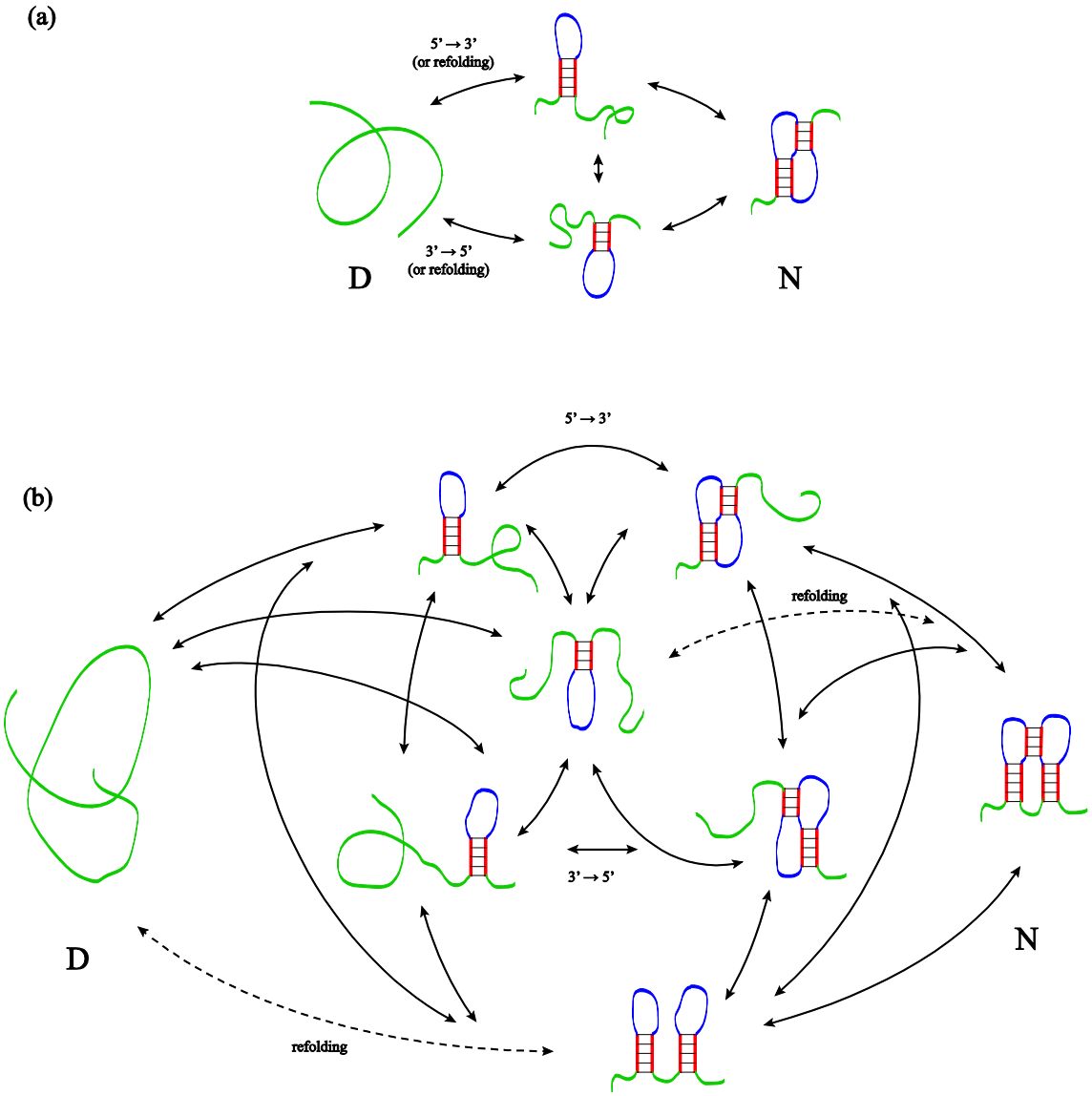
(a)
5' → 3'
(or refolding)
3' → 5'
(or refolding)
D
N
(b)
5' → 3'
refolding
3' → 5'
D
N
refolding

Figure 6

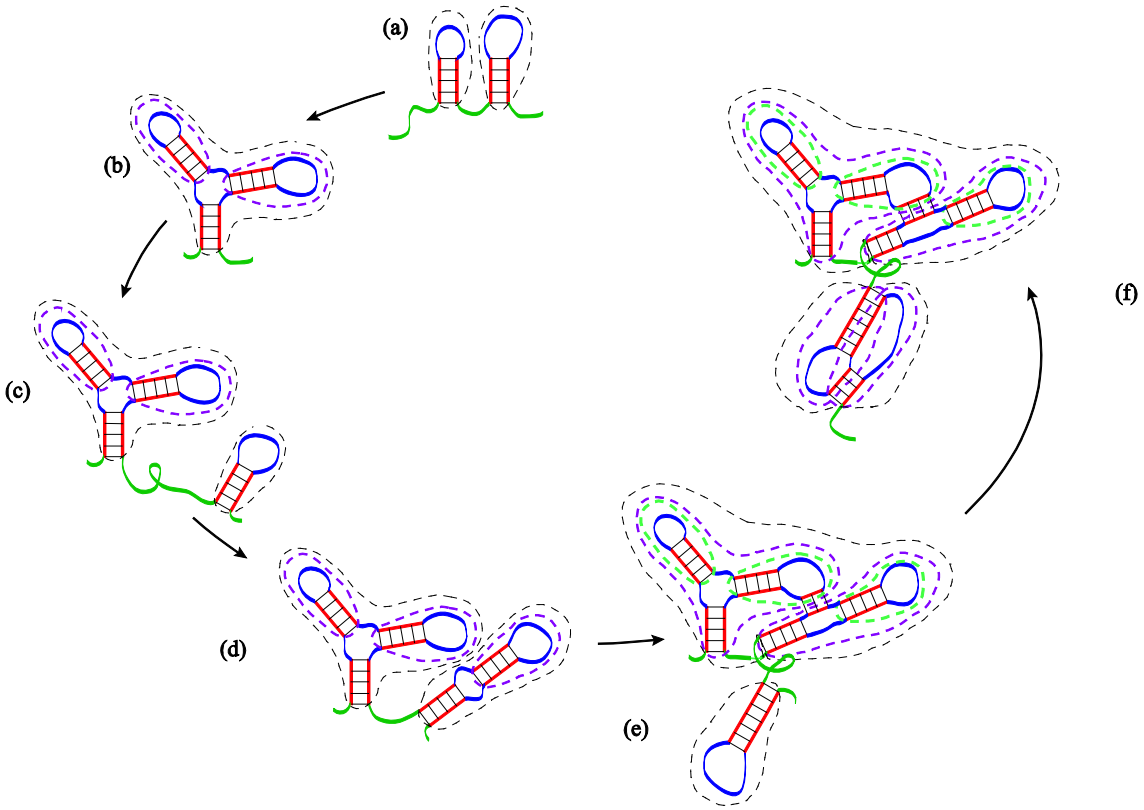

Figure 7

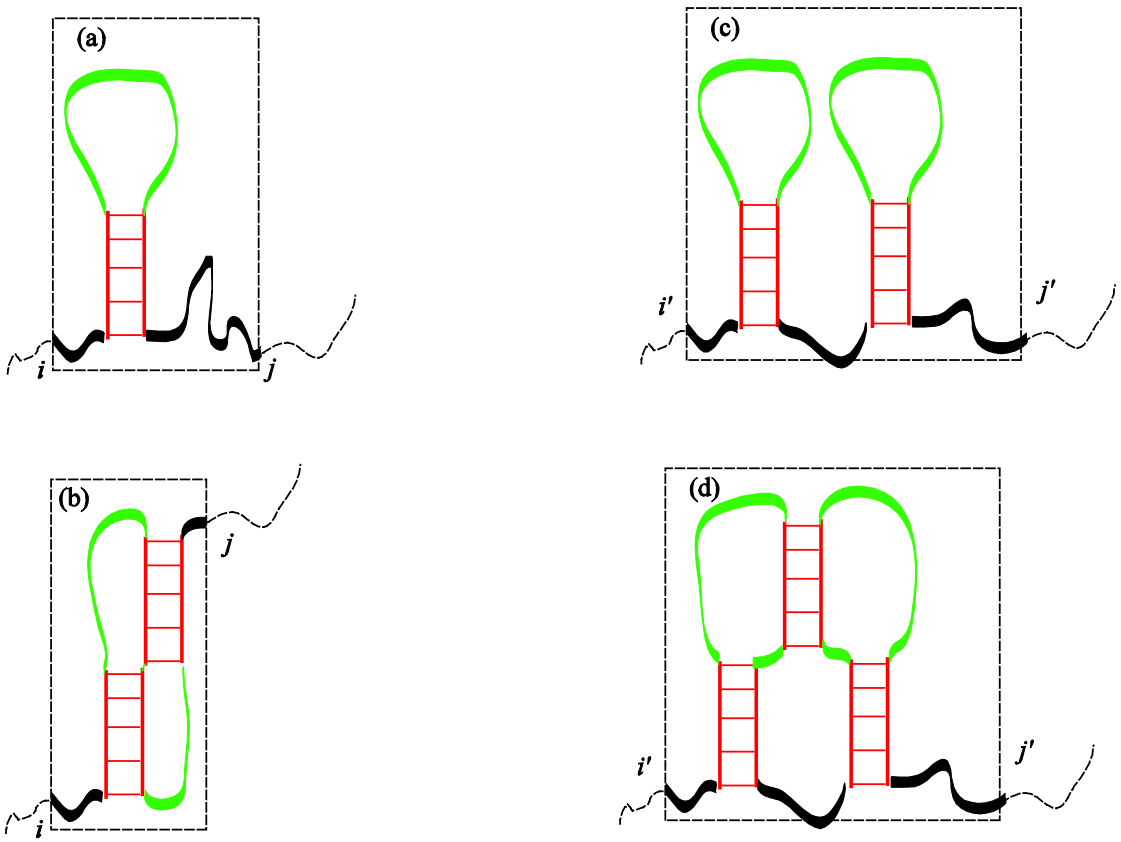

Figure 8

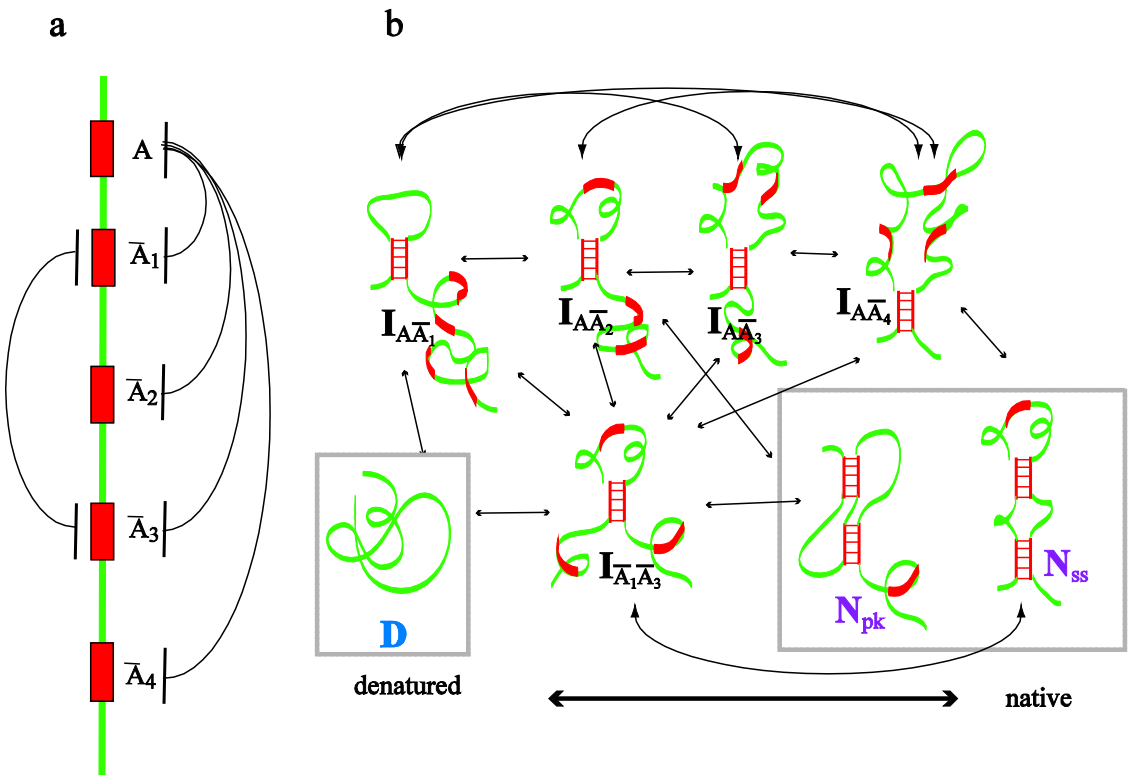
a
b
A
$\bar{A}_1$
$\bar{A}_2$
$\bar{A}_3$
$\bar{A}_4$
$I_{A\bar{A}_1}$
$I_{A\bar{A}_2}$
$I_{A\bar{A}_3}$
$I_{A\bar{A}_4}$
$I_{\bar{A}_1\bar{A}_3}$
D
denatured
N_{pk}
N_{ss}
native

Figure Captions

Figure 1

Figue 1. Examples of secondary structures and pseudoknots and the corresponding notations. The green chains represent free strand regions outside of the particular structural module. (a) A simple hairpin loop (H-loop, blue region) and stem (crosshatched regions), (b) a bulge, (c) an interior loop (I-loop), (d) a multibranch loop (MBL), (e) a core pseudoknot (H-type) and (f) an extended pseudoknot. The parameters n, n_i and n_j refer to the length of the free strand (blue) in a given loop. The stems are indicated by the red bar and black cross hatch. Base pairs and dinucleotide base pairs in the stem are marked in the Figure with the light blue circles. A distinction is made in this figure between free strand located in a loop region (blue) and free strand that has no loop associated with it (green).

Figure 2

Figure 2. Description of the all the folding pathways of a special RNA that only permits unique stacking at the specific locations stem 1, 2, 3. All other positions are given as impossible by specific pairing rules. (a) The strand segment's interaction and labeling where $\bar{1}$ is the complement of 1, *etc.* (b) A diagram of all the folding pathways that are possible for this particular structure. Here, D means denatured, N native state, and I_x (x=1,2,3, *etc.*) is an intermediate state. The numbers indicate the stem indices and the numbers over the arrows indicate which transition is taking place.

Figure 3

Figure 3. (a) Description of the end to end bond distance in terms of its parabolic dependence of r_k ($k = 1,2,3$) in the GPC model. Here, r_1, r_2, and r_3 correspond to the respective distances formed between effective mers $(\tilde{i}, \tilde{j})$. (b) A simplified stem-loop representation of the native state in which the red bars correspond to the stems and the green circles correspond to the loops. (c) Using the simplified representation in (a) and (b), a depiction of the character of the TMPFP (S_o, S_1, S_2, and S_3) as the RNA folds from the denatured state S_o through a series of intermediates (S_1 and S_2) to the native state S_3. The CLE can handle details of the problem were we to choose to calculated the minute details of such intermediates and the exact differences in their entropy. However, this is not necessary because the change in the entropy only depends on the initial state and the final state in thermodynamic equilibrium.

Figure 4

Figure 4. An example of a toy model with degenerate free energies — two competing secondary structures of equal mFE. (a) The Rivas and Eddy Feynman diagram [71,76] and a map of the structure lined out on the sequence (similar to Fig 2a). (b) The folding pathways available to this sequence.

Figure 5

Figure 5. A diagram of the folding of pseudoknots with no internal rearrangement at any stage of the folding. (a) Folding of a core (H-type) pseudoknot. (b) Folding of an extended pseudoknot. The notation 5'→3' refers to the way RNA is expected to fold during transcription and 3'→5' indicates a hypothetical folding direction possibly occurring during production of synthetic RNA. The dotted line in (b) indicates the unique features of a refolding experiment where the entire sequence can fold simultaneously. Upon folding, all pathways are equally accessible to all directions of folding; however, at the immediate point of removing denaturing solvents, this pathway is unique to refolding or some process that delays folding at the 5' end of the sequence in biological systems. The labels N and D indicate native state and denatured state, respectively.

Figure 6

Figure 6. Schematic depiction of RNA folding in its typical 5'→3' direction and the corresponding hierarchy of levels that gradually build up as modules of RNA structures build up. The folding progression from a to f is explained in the text. As the structure grows, the previous level is promoted to a higher level (subdomain) which roughly corresponds to a basic motif or module where level 0 structures are indicated by the black dashed lines, level 1 structures by the purple dashed lines and level 2 structures by the green dashed lines. The pseudoknot structures in f are treated as a single unit with an input and output point along the main chain (Figs 1e and 1f). The secondary structure inside is then decomposed into its original modules as it folded.

Figure 7

Figure 7. Examples for showing the minimum free energy (mFE) on interval $[i \cdots j]$ and $[i' \cdots j']$. (a) Secondary structure is the mFE on $[i \cdots j]$. (b) For the same secondary structure, after a pseudoknot is added such that $\Delta G(\mathrm{b}) < \Delta G(\mathrm{a})$, the pseudoknot on $[i \cdots j]$ must be mFE. (c) Secondary structure is mFE on $[i' \cdots j']$. (d) For the same secondary structure, after the pseudoknot is added such that $\Delta G(\mathrm{d}) < \Delta G(\mathrm{c})$, the PK must be the mFE on $[i' \cdots j']$.

Figure 8

Figure 8. A case where a pseudoknot structure and secondary structure all *somehow* have exactly the same free energies. (a) Rivas & Eddy Feynman diagram [71,76] of the interaction of these stems. (b) the full set of pathways possible for this system.

References

1. Bellman R (2003) Dynamic Programming. Mineola (New York): Dover. 340 p.
2. Cormen TH (2001) Introduction to algorithms. Cambridge, Mass.: MIT Press. xxi, 1180 p.
3. Wikipedia Dynamic Programming. Wikipedia: Wikipedia.
4. Waterman MS, Smith TF (1986) Rapid Dynamic Programming Algorithms for RNA Secondary Structure. Advances in Applied Mathematics 7: 455-464.
5. Eddy SR (2004) How do RNA folding algorithms work? Nat Biotechnol 22: 1457-1458.
6. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 31: 3406-3415.
7. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research 9: 133-148.
8. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Research 31: 3429-3431.
9. 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.
10. Dawson W, Fujiwara K, Kawai G (2007) Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. PLoS One 2: 905.
11. 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.
12. Dasgupta S, Papadimitriou CH, Vazirani UV (2008) Algorithms. Boston: McGraw-Hill Higher Education. x, 320 p. p.
13. Sedgewick R (1983) Algorithms. Reading, Mass.: Addison-Wesley. viii, 551 p. p.
14. 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.
15. Gralla J, Crothers DM (1973) Free energy of imperfect nucleic acid helices. 2. Small hairpin loops. J Mol Biol 73: 497-511.
16. 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.
17. Michel F, Westhof E (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J Mol Biol 216: 585-610.
18. Tinoco I, Bustamante C (1999) How RNA folds. Journal of Molecular Biology 293:

271-281.

19. Fernandez A, Cendra H (1996) In vitro RNA folding: the principle of sequential minimization of entropy loss at work. Biophys Chem 58: 335-339.
20. Morgan SR, Higgs PG (1996) Evidence for kinetic effects in the folding of large RNA molecules. J Chem Phys 105: 7152-7157.
21. Geis M, Flamm C, Wolfinger MT, Tanzer A, Hofacker IL, et al. (2008) Folding kinetics of large RNAs. J Mol Biol 379: 160-173.
22. Dotu I, Lorenz WA, Van Hentenryck P, Clote P (2010) Computing folding pathways between RNA secondary structures. Nucleic Acids Res 38: 1711-1722.
23. 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.
24. 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.
25. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, et al. (2003) Contact order revisited: influence of protein size on the folding rate. Protein Science 12: 2057-2062.
26. Makarov DE, Keller CA, Plaxco KW, Metiu H (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proc Natl Acad Sci U S A 99: 3535-3539.
27. Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277: 985-994.
28. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 5: 105.
29. Schultes EA, Spasic A, Mohanty U, Bartel DP (2005) Compact and ordered collapse of randomly generated RNA sequences. Nat Struct Mol Biol 12: 1130-1136.
30. Sosnick TR, Pan T (2004) Reduced contact order and RNA folding rates. J Mol Biol 342: 1359-1365.
31. 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.
32. 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.
33. 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.

34. Grosberg AY, Khokhlov AR (1994) Statistical Physics of Macromolecules. New York: AIP Press.
35. Flory PJ (1953) Principles of Polymer Chemistry. Ithaca: Cornell University Press.
36. Shortle D (1995) Staphylococcal nuclease: a showcase of m-value effects. Adv Protein Chem 46: 217-247.
37. Baldwin RL, Zimm BH (2000) Are denatured proteins ever random coils? Proc Natl Acad Sci (USA) 97: 12391-12392.
38. Bowler BE (2012) Residual structure in unfolded proteins. Curr Opin Struct Biol 22: 4-13.
39. Tanford C (1970) Protein denaturation. C. Theoretical models for the mechanism of denaturation. Adv Protein Chem 24: 1-95.
40. Tanford C (1968) Protein denaturation. Adv Protein Chem 23: 121-282.
41. Flory PJ (1969) Statistical Mechanics of Chain Molecules. New York: Wiley.
42. Moffitt JR, Chemla YR, Smith SB, Bustamante C (2008) Recent advances in optical tweezers. Annu Rev Biochem 77: 205-228.
43. Fisher ME (1966) Shape of a Self-Avoiding Walk or Polymer Chain. Journal of Chemical Physics 44: 616-&.
44. McKenzie DS (1976) Polymers and scaling. Physics Reports 27C: 35-88.
45. Fisher ME (1966) Effect of Excluded Volume on Phase Transitions in Biopolymers. Journal of Chemical Physics 45: 1469-1473.
46. 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.
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 II. J Theor Biol 213: 387-412.
48. Eddy SR (2004) What is dynamic programming? Nat Biotechnol 22: 909-910.
49. Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244: 48-52.
50. Nussinov R, Jacobson AB (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A 77: 6309-6313.
51. Uhlenbeck OC, Martin FH, Doty P (1971) Self-complementary oligoribonucleotides: effects of helix defects and guanylic acid-cytidylic acid base pairs. J Mol Biol 57: 217-229.

52. Tinoco I, Uhlenbeck OC, Levine MD (1971) Estimation of secondary structure in ribonucleic acids. Nature 230: 362-367.
53. Salser W (1977) Globin mRNA sequences: analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp Quant Biol 42: 985-1103.
54. Studnicka GM, Rahn GM, Cummings IW, Salser WA (1978) Computer method for predicting the secondary structure of single-stranded RNA. Nucleic Acids Res 5: 3365-3387.
55. Freier SM, Petersheim M, Hickey DR, Turner DH (1984) Thermodynamic studies of RNA stability. J Biomol Struct Dyn 1: 1229-1242.
56. 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.
57. Papanicolaou C, Gouy M, Ninio J (1984) An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. Nucleic Acids Res 12: 31-44.
58. Laing LG, Draper DE (1994) Thermodynamics of RNA folding in a conserved ribosomal RNA domain. J Mol Biol 237: 560-576.
59. Walter AE, Turner DH (1994) Sequence dependence of stability for coaxial stacking of RNA helixes with Watson-Crick base paired interfaces. Biochemistry 33.
60. Mathews DH, Turner DH (2002) Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. Biochemistry 41: 869-880.
61. Chen J-H, Le S-Y, Maizel JV (1992) A procedure for RNA pseudoknot prediction. Computer Applications in the Biosciences 8: 243-248.
62. Martinez HM (1990) Detecting pseudoknots and other local base-pairing structures in RNA sequences. Methods in Enzymology 183: 306-318.
63. Gultyaev AP, van Batenburg FH, Pleij CW (1995) The computer simulation of RNA folding pathways using a genetic algorithm. J Mol Biol 250: 37-51.
64. Chen JH, Le SY, Maizel JV (2000) Prediction of common secondary structures of RNAs: a genetic algorithm approach. Nucleic Acids Res 28: 991-999.
65. Morgan S, Higgs PG (1998) Barrier heights between ground states in a model of RNA secondary structure. J Phys A: Math Gen 31: 3153-3170.
66. Higgs PG (2000) RNA secondary structure: physical and computational aspects. Quarterly Rev Biophys 33: 1999-1253.
67. Fernandez A, Salthu R, Cendra H (1999) Discretized torsional dynamics and the folding of an RNA chain. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 60: 2105-2119.

68. Xayaphoummine A, Viasnoff V, Harlepp S, Isambert H (2007) Encoding folding paths of RNA switches. Nucleic Acids Res 35: 614-622.
69. Akutsu T (2000) DP algorithms for RNA secondary structure prediction with pseudoknots. Discrete Appl Math 104: 45-62.
70. Lyngsø RB, Pedersen CNS (2000) RNA pseudoknot prediction in energy-based models. Journal of Computational Biology 7: 409-427.
71. Rivas E, Eddy SR (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of Molecular Biology 285: 2053-2068.
72. Reeder J, Giegerich R (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics 5: 104.
73. Turner DH, Sugimoto N, Freier SM (1988) RNA Structure Prediction. Ann Rev Biophys Biophys Chem 17: 167-192.
74. Meyer TE, Bansal AK (2005) Stabilization against hyperthermal denaturation through increased CG content can explain the discrepancy between whole genome and 16S rRNA analyses. Biochemistry 44: 11458-11465.
75. Shannon CE, Weaver W (1949) The mathematical theory of communication. Urbana,: University of Illinois Press. v (i.e. vii), 117 p. p.
76. Rivas E, Eddy SR (2000) The language of RNA: a formal grammar that includes pseudoknots. Bioinformatics 16: 34-340.
77. Mathews DH (2006) Revolutions in RNA secondary structure prediction. Journal of Molecular Biology 359: 526-532.
78. Pan T, Sosnick T (2006) RNA folding during transcription. Annu Rev Biophys Biomol Struct 35: 161-175.
79. Stryer L (1995) Biochemistry. New York: W.H. Freeman. xxxiv, 1064 p. p.
80. Wikipedia Oligonucleotide synthesis. Wikipedia.
81. Garst AD, Batey RT (2009) A switch in time: detailing the life of a riboswitch. Biochim Biophys Acta 1789: 584-591.
82. Dixon SE, Bhatti MM, Uversky VN, Dunker AK, Sullivan WJ, Jr. Regions of intrinsic disorder help identify a novel nuclear localization signal in Toxoplasma gondii histone acetyltransferase TgGCN5-B. Mol Biochem Parasitol 175: 192-195.
83. Xue B, Dunker AK, Uversky VN Retro-MoRFs: Identifying Protein Binding Sites by Normal and Reverse Alignment and Intrinsic Disorder Prediction. Int J Mol Sci 11: 3725-3747.
84. Xue B, Hsu WL, Lee JH, Lu H, Dunker AK, et al. SPA: Short peptide analyzer of intrinsic disorder status of short peptides. Genes Cells 15: 635-646.

85. Xue B, Williams RW, Oldfield CJ, Goh GK, Dunker AK, et al. Viral disorder or disordered viruses: do viral proteins possess unique features? Protein Pept Lett 17: 932-951.
86. Brion P, Westhof E (1997) Hierarchy and Dynamics of RNA folding. Annu Rev Biophys Biomol Struct 26: 113-137.
87. Jaeger L, Westhof E, Leontis NB (2001) TectoRNA: modular assembly units for the construction of RNA nano-objects. Nucleic Acids Res 29: 455-463.
88. Lescoute A, Westhof E (2006) The interaction networks of structured RNAs. Nucleic Acids Research 34: 6587-6604.