Computing the Partition Function for Kinetically Trapped RNA Secondary Structures

William A. Lorenz¹, Peter Clote²

¹ Department of Mathematics and Computer Science, Denison University, Granville, Ohio, United States of America, ² Biology Department, Boston College, Chestnut Hill, Massachusetts, United States of America

Abstract

An RNA secondary structure is locally optimal if there is no lower energy structure that can be obtained by the addition or removal of a single base pair, where energy is defined according to the widely accepted Turner nearest neighbor model. Locally optimal structures form kinetic traps, since any evolution away from a locally optimal structure must involve energetically unfavorable folding steps. Here, we present a novel, efficient algorithm to compute the partition function over all locally optimal secondary structures of a given RNA sequence. Our software, RNALOCOPT runs in $O(n^2)$ time and $O(n^2)$ space. Additionally, RNALOCOPT samples a user-specified number of structures from the Boltzmann subensemble of all locally optimal structures. We apply RNALOCOPT to show that (1) the number of locally optimal structures is far fewer than the total number of structures – indeed, the number of locally optimal structures approximately equal to the square root of the number of all structures, (2) the structural diversity of this subensemble may be either similar to or quite different from the structural diversity of the entire Boltzmann ensemble, a situation that depends on the type of input RNA, (3) the (modified) maximum expected accuracy structure, computed by taking into account base pairing frequencies of locally optimal structures, is a more accurate prediction of the native structure than other current thermodynamics-based methods. The software RNALOCOPT constitutes a technical breakthrough in our study of the folding landscape for RNA secondary structures. For the first time, locally optimal structures (kinetic traps in the Turner energy model) can be rapidly generated for long RNA sequences, previously impossible with methods that involved exhaustive enumeration. Use of locally optimal structure leads to state-of-the-art secondary structure prediction, as benchmarked against methods involving the computation of minimum free energy and of maximum expected accuracy. Web server and source code available at http://bioinformatics.bc.edu/clotelab/RNALocopt/.

Introduction

Kinetics of RNA secondary structure formation plays an important role in many biological functions, as shown by co-transcriptional folding [1] of large RNA molecules, the host-killing (hok) and suppression of killing system (sok) system [2] to control plasmid copy number in E. coli, the kinetically driven trans-splicing of a S' codon in Leptomonas collosoma [3], and the kinetic control in the formation of the Tetrahymena ribozyme [4].

RNA secondary structure kinetics depends on the distribution of locally optimal secondary structures, where a structure is said to be locally optimal if it is not the case that by adding or removing a single base pair, one can obtain a structure having lower free energy. In the context of the Nussinov energy model [5], where the energy of a base pair is $-1$, locally optimal structures are exactly the saturated secondary structures, as first defined by M. Zuker [6]. (A secondary structure is saturated if one cannot add any base pairs without violating the definition of a secondary structure; i.e. without either creating a base triple or pseudoknot.) In the paper [7] we developed an algorithm to compute the partition function for all saturated secondary structures of a given RNA sequence. Exploiting the idea behind this algorithm, in the papers [8,9], we subsequently proved that the asymptotic number of saturated secondary structures is $1.07427n^{-3/2}2.35467n$, which (surprisingly) is not substantially less than the asymptotic number $1.104366n^{-3/2}2.618034n$ of all secondary structures, a result earlier proved by Stein and Waterman [10]. In Waldispühl and Clote [11], we extended our previous algorithm [7] to compute the partition function of all saturated secondary structures, with respect to the widely used Turner energy model [12]. In the Turner energy model, a secondary structure is decomposed into loops, as described in Zuker [13], and the free energy is computed by summing the energy contributions of all loops. A $K$-loop consists of $K-1$ base pairs (excluding the closing base pair) and $u$ unpaired bases. The energies of 1-loops (hairpins), 2-loops (stacks if $u=0$, bulges or interior loops if $u>0$), 3-loops and 4-loops (also known as 3-way and 4-way multiloop junctions) are obtained by least squares fit of enthalpy and free energy change at $37^\circ C$, determined by optical melting (UV absorption) of small model systems [14,15]. Even though free energies for the most common multiloops (3-way
and 4-way junctions) have been experimentally determined [15], for computational efficiency it is usual to define the free energy of arbitrary multiloops \( k > 2 \) by the affine approximation \( a + b(k - 1) + c \), where \( a, b \) and \( c \) are constants.

Computational studies of RNA kinetics are currently performed either by repeated Monte-Carlo simulations, as in software Kinfold of Flamm, Fontana, Hofacker and Schuster [16], Kinefold of Xayaphoummine, Bucher, and Isambert [17], and RNAKinetics of Danilova, Pervouchine, Favorov, and Mironov [18], or by direct solution of the master equation from chemical kinetics

\[
\frac{dP_j(t)}{dt} = \sum \left( P_j(t)k_{ij} - P_i(t)k_{ji} \right).
\]

Here, \( P_j(t) \) is the probability that the RNA molecule is in secondary structure \( S_j \) at time \( t \), and \( k_{ij} \) is the transitional probability of moving from structure \( S_i \) to neighboring structure \( S_j \), which differs from \( S_i \) by the addition or removal of a single base pair, and where \( \frac{1}{k_{ij}} = \exp(-(-E_j - E_i)/RT) \). By constructing probabilistic roadmaps for RNA secondary structure formation, a technique derived from robotic motion planning, Tang, Kirkpatrick, Thomas, Song and Amato [19] and Tang, Thomas, Tai, Giedroc and Amato [20] are able to apply both Monte Carlo methods and the master equation over a smaller set of structures.

Flamm, Fontana, Hofacker and Schuster [16] describe RNA folding at an elementary step resolution, by using a Monte Carlo algorithm to study the kinetics of folding. Their Kinfold program is an implementation of Gillespie’s Monte Carlo algorithm [21,22] for stochastic folding, where elementary steps consist of either adding, removing or shifting a single base pair. In that paper, Flamm et al. describe the barrier tree, whose leaves are those locally optimal secondary structures having free energy that lies below a user-defined threshold. The barrier tree is constructed by using the program RNAsubopt [23] to exhaustively generate all secondary structures, whose free energy lies below a user-defined threshold, then aggregating structures into basins containing a locally optimal structure. As more structures are aggregated, using the imagery of flooding a landscape, two basins may be gradually joined together by folding paths, all of whose intermediate structures lie in one of the two basins, for which there exists a saddle structure of highest free energy along the path. Flamm, Hofacker, Stadler and Wolfinger [24] present additional applications of the Barriers program, while Wolfinger et al. [25] describe a coarse-grained approach by applying the master equation of chemical kinetics to macrostates consisting of basins of structures aggregated near locally optimal structures. For additional results on saddle points and energy barriers, see Stadler and Flamm [26], Flamm, Hofacker, Stadler, and Stadler [27], as well as the recent paper by Hofacker, Flamm, Heine, Wolfinger, Scheuermann et al [28], who introduce the notion of barmap which “links macrostates of temporally adjacent landscapes and defines the transfer of population densities from one ‘snapshot’ to the next”.

Other groups have studied various aspects of kinetically driven RNA folding. Shapiro, Bengali, Kasprzak and Wu [29] compute likely folding intermediates in the earlier described hok/sok system. Danilova, Pervouchine, Favorov, and Mironov [18] describe the web server, RNAKinetics, which models the secondary structure kinetics of an elongating RNA molecule. Xayaphoummine, Bucher, and Isambert [17] and Isambert [30] introduce the Kinefold web server, which stochastically folds a user-given RNA sequence into a low energy structure that may include pseudoknots. Quite recently, Dotu, Lorenz, Van Hentenryck and Clotte [31] describe an efficient program RNAyABPath to compute near-optimal folding pathways between two secondary structures of a given RNA sequence. For an overview of RNA folding kinetics, see the review articles by Chen [32] and Al-Hashimi and Walter [1].

In this paper, we describe a novel, efficient algorithm, RNAlocopt, to compute the partition function over all secondary structures that are locally optimal in the Turner energy model. Locally optimal structures form ‘kinetic traps’, hence create ‘basins of attraction’ in the energy landscape. The structure of this paper is as follows. In the introduction sections, we provide background definitions for the Turner energy model and loop decomposition. To allow the paper to be self-contained, we additionally describe McCaskill’s classical algorithm for the partition function [33].

In the Results section, we present three types of analysis using the software RNAlocopt. First, by performing computational experiments on RNA sequences of increasing length, we show that the number of locally optimal structures is asymptotically the square root of the number of all structures, as depicted in Figure 5. Secondly, we compare the structural diversity, as measured by four different metrics, of the set of locally optimal structures with that of the Boltzmann ensemble of all secondary structures. Structural diversity appears to depend on the type of RNA; for instance, in the case of precursor microRNAs and 5S-rRNA, the structural diversity of the collection of locally optimal secondary structures is markedly lower than that over the Boltzmann ensemble, while structural diversity for TTP riboswitch aptamers appears to be about the same. Thirdly, we demonstrate how to combine McCaskill base pairing probabilities with those from sampled locally optimal structures in order to compute a modified maximum expected accuracy structure [34,35], which appears to be closer to the native structure than structures produced by other thermodynamics-based algorithms. The Discussion section provides additional comments on the energy model of RNAlocopt and benchmarking issues, and as well describes intended future applications and possible extensions of the software. In particular, in forthcoming work, we will introduce a new method using RNAlocopt to quickly and accurately determine the mean folding time for a given RNA sequence, a synthetic biology application for de novo RNA design.

In the Methods section, we begin by describing the intuition behind the new \( O(n^3) \) time and \( O(n^2) \) space algorithm, whose details and recurrence relations are then provided. Though our software RNAlocopt additionally can sample a user-specified number of structures from the Boltzmann subensemble of locally optimal structures, we do not describe details of the construction, since it is analogous to the construction of Ding and Lawrence [36,37], albeit where the McCaskill partition function is replaced by the partition function for locally optimal structures.

**Background**

An RNA molecule is a biopolymer consisting of nucleotides, adenine (A), cytosine (C), guanine (G) and uracil (U), oriented in a natural left-to-right fashion given by the 5’ to 3’ direction. Given an RNA sequence \( a_1, \ldots, a_n \) of length \( n \), an RNA secondary structure \( S \) is defined to be a set of base pairs \( (i,j) \), where \( (i,j) \in S \), then \( (a_i,a_j) \in \{ A,U,A,GC,CG,GU,U/G \} \) (base pairs are canonical, i.e. either Watson-Crick or wobble pairs); (i) if \( (i,j) \in S \), then \( j > i + 1 \), where by convention \( 0 = 3 \) (minimum of 0 unpaired bases in a hairpin loop); (ii) if \( (i,j),(k,l) \in S \), then \( j = k \) and if \( (i,j),(k,l) \in S \), then \( l = k \) (non-existence of base triples); (iii) if
(i,j), (k,ℓ) ∈ S, and i < k, then ℓ < j (non-existence of pseudoknots). See Figure 1 for three equivalent representations of the secondary structure for RNA from human accelerated region HAR1F, a region of the human genome that seems to have been under evolutionary pressure in the divergence of humans from great apes [38]. While secondary structures satisfy a planarity condition, pseudoknots violate that condition, as shown in Figure 2. Although pseudoknots and non-canonical base pairs play important roles in RNA tertiary structure formation [39], the secondary structure forms rapidly and serves largely as a scaffold for the formation of tertiary contacts [40]. In this paper, we are interested in developing an efficient algorithm to explore the energy landscape of kinetically trapped RNA structures. Since Lyngsø and Pedersen [41] have proved that it is NP-complete to compute the minimum free energy structure for a given RNA sequence, when general pseudoknots are permitted, we will restrict our attention throughout the paper to secondary structure.

Nearest neighbor energy model

The Turner nearest neighbor energy model is an additive model, where the free energy of an RNA secondary structure is computed as the sum of distinct loop free energies in a unique decomposition of the structure. Figure 3 illustrates the different types of possible loops for an example RNA secondary structure. The structure contains 8 loops of 4 basic different types. Hairpins are formed when a base pair \((i,j)\) encloses an unpaired region of RNA; thus a hairpin contains the nucleotides \(a_i, \ldots, a_j\), where due to steric constraints, \(j - i > 0\), for \(θ = 3\), and positions \(i+1, \ldots, j-1\) are unpaired. Stacked base pairs are loops containing adjacent base pairs, \((i,j), (i+1,j-1)\), as shown in loops \(L_2\) and \(L_1\). Left bulges are loops containing the closing base pairs \((i,j), (i+1,j-k)\) for \(k > 1\), where \(j = k + 1, \ldots, j-1\) are unpaired; right bulges contain the closing base pairs \((i,j), (i+k,j-1)\) for \(k > 1\), where positions \(i+1, \ldots, i+k-1\) are unpaired. Loop \(L_3\) depicts a left bulge. Internal loops are loops bordered by 2 base pairs \((i,j), (k,ℓ)\), where \(i+1 < k\) and \(ℓ + 1 < j\). Loop \(L_5\) depicts an internal loop. A multiloop is a loop bordered by 3 or more base pairs. For instance, \(L_4\) is a multiloop closed by the base pair \((3,28)\), which here is a 3-way junction (i.e. bordered by three base pairs) and which has two components (i.e. stems bordered by base pairs \((5,15)\) and \((17,26)\)). The number \(k\) of base pairs that border a loop can be used to classify the loop; \(k = 1\) in hairpins, \(k = 2\) in stacked base pairs, bulges, and internal loops, and \(k > 2\) in multiloops. Finally, external loops, depicted in \(L_8\), are technically not loops, but rather are defined to be regions containing nucleotide positions \(x\) for which there is no base pair \((i,j)\) satisfying \(i \leq x \leq j\).

In the Turner energy model, there are free energies for each type of loop. For the example structure \(S\) depicted in Figure 3, if we denote the energy of loop \(L_i\) by \(E(L_i)\), it follows that the free energy of \(S\) is

\[
E(S) = \sum_{i=1}^{8} E(L_i).
\]

The Turner rules were fit to enthalpy and folding free energy change at 37°C, determined by optical melting of small model systems [12,51]. For instance, Turner's rules assign stacking free energy of \(-2.24\) kcal/mol to \(\delta = AC - \overline{3}\) and of \(-3.26\) kcal/mol to \(\delta = CC - 3'\). Stacked base pairs constitute negative (stabilizing) free energy contribution; hairpins, bulges, internal loops, and multiloops generally contribute positive (destabilizing) free energies, although certain \(1 \times 1\) and \(2 \times 2\) internal loops contribute stabilizing energies.

Figure 1. RNA from human accelerated region HAR1F, a region of the human genome that differs from highly conserved regions of our closest primate relatives and is active in the developing human brain between the 7th and 18th gestational weeks [38]. Secondary structure representation in conventional form (left), as a circular Feynman diagram (center) and as a linear Feynman diagram (right). Sequence and consensus secondary structure taken from Rfam [42]; graphics produced with jViz software [43].

doi:10.1371/journal.pone.0016178.g001
Figure 2. Long range pseudoknot PKB239 in the 5’ untranslated region (UTR) of human immunodeficiency virus HIV-1. Secondary structure with pseudoknots displayed in conventional form (left) and as a circular Feynman diagram (right). Sequence and structure of PKB239 taken from Pseudobase [44]; graphics produced with jViz software [43].

doi:10.1371/journal.pone.0016178.g002

Figure 3. The Turner energy model is an additive loop model, whereby the free energy of an RNA secondary structure is defined to be the sum of loop free energies in a unique decomposition of the structure into loops. In this figure, the free energy of the depicted structure is the sum of free energies of loops $L_1$ through $L_8$. The Turner rules include free energy parameters for different types of loops, illustrated here for hairpins ($L_1$, $L_4$), stacked base pairs ($L_2$, $L_7$), bulges ($L_3$), internal loops ($L_5$), multiloops ($L_4$) and external loops ($L_8$). The Turner parameters are derived from a series of UV absorption (optical melting) experiments described in a number of papers including the references [12,45–49]. For a complete list of all references, see http://rna.urmc.rochester.edu/NNDB/ref.html. Images created using the software VARNA [50].

doi:10.1371/journal.pone.0016178.g003
Important aspects of the Turner energy model are additivity and locality. Both of these properties are critical in the development of an efficient computation of the partition function; indeed, it is this local nature of the energy model that renders it possible to inductively determine all locally optimal structures.

The algorithm to compute the partition function of all locally optimal structures is a modification of McCaskill’s algorithm, which we will review now. McCaskill’s algorithm recursively computes the partition function for structures on subsequence \(a_1, \ldots, a_j\) by table look-up of the previously computed partition function values for proper subwords of \(a_1, \ldots, a_j\). Each recursion step involves the addition of either one base pair or one unpaired base to groups of structures whose partition function is already known. Our modification to McCaskill’s algorithm is to make sure at each step that the base pair or base added does not cause the occurrence of non-optimal structures. This will require additional information to be stored at each step, but does not change the basic structure of the McCaskill recursions.

### McCaskill’s partition function

In order to provide a self-contained treatment, we now review the construction of McCaskill’s algorithm [33] to construct the partition function for RNA secondary structures.

Given RNA nucleotide sequence \(a_1, \ldots, a_n\), we let \(E_{HP}(i,j)\) denote the free energy of a hairpin closed by base pair \((i,j)\), while \(E_H(i,j,f)\) denotes the free energy of an internal loop enclosed by the base pairs \((i,j)\) and \((i',j')\), where \(i < i' < j' < j\). (Internal loops comprise the cases of stacked base pairs, left/right bulges and proper internal loops.) The free energy for a multiloop containing \(N_b\) base pairs and \(N_u\) unpaired bases is given by the affine approximation \(a + bN_b + cN_u\).

Given an RNA sequence \(a_1, \ldots, a_n\), for \(1 \leq i \leq j \leq n\), the McCaskill partition function \(Z(i,j)\) is defined by \(\sum_S e^{-E(S)/RT}\), where the sum is taken over all secondary structures \(S\) of \(a_1, \ldots, a_j\), \(E(S)\) is the free energy of secondary structure \(S\), \(R\) is the universal gas constant, and \(T\) is absolute temperature. In the sequel we write \(a[i,j]\) to abbreviate \(a_i, \ldots, a_j\).

**Definition 1 (McCaskill’s partition function)**

- \(Z(i,j)\): partition function over all secondary structures of \(a[i,j]\).
- \(Z_B(i,j)\): partition function over all secondary structures of \(a[i,j]\), which contain the base pair \((i,j)\).
- \(Z_M(i,j)\): partition function over all secondary structures of \(a[i,j]\), subject to the constraint that \(a[i,j]\) is part of a multiloop and has at least one component.
- \(Z_M(i,j)\): partition function over all secondary structures of \(a[i,j]\), subject to the constraint that \(a[i,j]\) is part of a multiloop and has at exactly one component. Moreover, it is required that \(i\) base-pair in the interval \([i,j]\); i.e. \((i,r)\) is a base pair, for some \(i < r \leq j\).

Following McCaskill [33], the unconstrained partition function is defined by

\[
Z(i,j) = Z(i,j - 1) + Z_B(i,j) + \sum_{r = i + 1}^{j - \theta - 1} Z(i, r - 1) Z_B(r,j).
\]

The constrained partition function closed by base pair \((i,j)\) is given by

\[
Z_B(i,j) = \exp(-E_{HP}(i,j)/RT) + \sum_{i \leq l \leq r \leq j} \exp(-E_H(l,r,i,j)/RT) Z_B(l,r) + \exp(-(a+b)/RT) \left( \sum_{r = i + 1}^{j - \theta - 2} Z_M(l+1,r-1) Z_M(r,j-1) \right).
\]

The multiloop partition function with a single component and where position \(i\) is required to base-pair in the interval \([i,j]\) is given by

\[
Z_M(i,j) = \sum_{r = i + \theta + 1}^{j - \theta - 1} Z_B(i,r) \exp(-c(j-r)/RT). \]

Finally, the multiloop partition function with one or more components, having no requirement that position \(i\) base-pair in the interval \([i,j]\) is given by

\[
Z_M(i,j) = \sum_{r = i + \theta + 2}^{j - \theta - 1} Z_M(i,r-1) Z_M(r,j) \exp(-b/RT).\]

See Figure 4 for a pictorial representation of the recursions of McCaskill’s (original) algorithm [52]; note that the recursions are equivalent to, but not quite the same as, those given in [53].

### Results

**Number of locally optimal structures**

In this section, we compare the values of the partition function, \(Z_{LO}\), of all locally optimal structures, and the total number, \(N_{LO}\), of locally optimal structures, with those for all structures. The number of locally optimal structures, \(N_{LO}\), is determined by removing all energy factors in the previous equations for the Boltzmann partition function. This is equivalent to setting the temperature to \(+\infty\), since all energetic factors are of the form \(e^{-E/RT}\).

In Figure 5, for lengths between 20 and 200 nt, 100 RNA were randomly generated for each length in the simplest possible manner, with 1/4 probability of A, C, G, and U at each location. For each such RNA, the number of locally optimal structures as well as the number of all secondary structures is determined. These are averaged over the 100 randomly generated RNA sequences of that length, and plotted in the graph shown in Figure 5. We find there is exponential growth in the average, or expected, number of locally optimal structures, as a function of sequence length. Moreover, the slope of the curve in Figure 5 for the total number of structures is approximately twice that of the number of locally optimal structures, hence implying that the number of structures is approximately the square of that for locally optimal structures. Indeed, by fitting the data with a least-squares approximation, we find that the number \(N_{LO}(n) \approx 10^0.254579n^{-1.95771}\) with
$R^2 = 0.999815$, while the number $\text{Num}_{LO}(n)$ of locally optimal structures for length $n$ random RNA satisfies $\text{Num}_{LO}(n) \approx 10.130366^n - 1.50238$ with $R^2 = 0.999407$. (The coefficient of determination, $R^2$, is the square of Pearson correlation coefficient of the least squares [linear] fit of the logarithm of the average number of structures.)

In Figure 6, we compare the partition function, $Z_{LO}$, of all locally optimal structures, with the partition function, $Z$, of all structures, by plotting the ratio, $Z_{LO}/Z$, by the same method, averaging over 100 RNA at each length. This ratio, depicted with error bars, represents the percentage of structures, as weighted by their Boltzmann factor, that are locally optimal. By numerical fitting the data from this curve, it appears that the ratio is approximately $1.0053 \exp(-0.0123n)$ with coefficient of determination $R^2 = 0.9876$ (see [54] for explanation of how to compute the coefficient of determination).

Another interesting computational experiment we performed was to determine the sum of the Boltzmann factors for a non-redundant subset of 1000 sampled locally optimal structures, produced by RNA LOCOPT, compared with the sum of the Boltzmann factors for a non-redundant subset of secondary structures, sampled by the Ding-Lawrence algorithm [36], as implemented in RNA SUBOPT -p. Table 1 presents these results for RNA generated in the previously described manner from an order 0 Markov chain, for lengths from 20 to 200 in steps of 20. For each length, we averaged statistics over 10 runs, where for each run, we computed the percent coverage of the partition function; i.e. sum of the Boltzmann factors of a non-redundant subset from 1000 samples generated by RNA SUBOPT [resp. RNA LOCOPT], divided by the partition function $Z$ [resp. partition function $Z_{LO}$ of locally optimal structures]. The number of locally optimal structures is far fewer than that of all structures (see Figure 5), hence, there is proportionately more redundancy among sampled locally optimal structures than that over all structures. As well, the percentage coverage of the partition function for sampled locally optimal structures is higher than that for the Boltzmann ensemble.

**Structural diversity of ensemble of locally optimal structures**

In our paper on RNA saturated structures [55], we suggested that (a) there are far fewer locally optimal structures than there are of saturated structures, and (b) base pairing probabilities over locally optimal structures are similar to the base pair probabilities over all structures. In the previous section, we have shown that (a) holds; indeed, Figure 5 shows that the number of locally optimal structures is approximately the square root of the number of all structures, while the papers [7–9,56] show that the number of saturated structures lies closer to that of all structures. While statement (b) holds in some cases, such as for purine riboswitch aptamers, in other cases, such as for precursor microRNAs and 5S-rRNA, it does not hold.

To numerically quantify how closely the ensemble of locally optimal structures resembles the Boltzmann ensemble of all structures, we consider four measures: the *pseudo-entropy* for base pairing probabilities, the *average entropy* for the base pairing probabilities, and two forms of structural diversity, the first due to Morgan and Higgs [57] and the second described in the Vienna RNA Package [58].

For a fixed RNA sequence $a_1, \ldots, a_n$ with base pairing probabilities $p_{ij}$, the *pseudo-entropy* is defined by

$$H_0 = -\sum_{i,j} p_{ij} \ln p_{ij}.$$

Since the collection of base pairing probabilities $p_{ij}$ does not form a probability distribution (although it does for fixed $i$, as exploited in the next definition), we cannot speak of its entropy, but rather use the term *pseudo-entropy*. The average (Shannon) entropy is defined by

$$\langle H \rangle = \frac{\sum_{i=1}^n H(i)}{n} = \frac{\sum_{i=1}^n \sum_{j=1}^n -p_{ij} \ln p_{ij}}{n}.$$

Both pseudo-entropy and the average entropy are measures of how well-defined are the base pairs. Indeed, if position $i$ base-pairs...
S, T, and P(S) is the Boltzmann probability $P(S) = \exp(-E(S)/RT)$ for structure $S$ (and similarly for $T$). If there is no structural diversity whatsoever, so that $p_{ij} = 1$ for all base pairs $(i,j)$ in the minimum free energy structure $S_0$, then clearly the Morgan-Higgs diversity $\langle D_{\text{MH}} \rangle$ will take on the least possible value, $n - |S_0| \approx n/2$, while the Vienna diversity will equal 0.

Variants of the above measures are given as well for the ensemble of locally optimal secondary structures, where we use base pairing frequencies $p_{ij}$ over a sampled collection of 1000 locally optimal structures for a given RNA sequence. Table 2 summarizes these four measures for 14 families of seed alignments from the Rfam 10.0 database [42]. For essentially all of these measures, we see that the structural diversity of the ensemble of locally optimal structures appears to be less than that for all structures. Notable exceptions are the riboswitch aptamers from Rfam.

By using the new algorithm RNAlocopt, we have shown that the collection of locally optimal structures constitutes an ensemble that is smaller (see Figure 5) and structurally less diverse in general than that of all structures. This provides additional evidence for the hypothesis advanced in [24,25,27] that locally optimal structures form basins of attraction in the folding landscape of RNA secondary structures. For this reason, RNAlocopt may prove valuable in the study of kinetics of RNA folding.

Basepair probabilities lead to better RNA secondary structure prediction

In ground-breaking work, Knudsen and Hein [59], followed by Do, Mahabashyam, Brudno and Batzoglou [60] and by Kiryu, Kin and Asai [34], introduced the notion of maximum expected accuracy secondary structure, shown to be closer to the native structure, compared to the minimum free energy structure, when benchmarked against known structures. The underlying idea of this new approach is that there is a strong signal in the Boltzmann ensemble of low energy structures—a signal that is ignored when one computes the minimum free energy (MFE) structure, which is the maximum likelihood structure with respect to Boltzmann probability. Independently and at the same time, Ding, Chan and Lawrence [61] also realized the benefit of considering the Boltzmann ensemble rather than the MFE structure in their construction of the Boltzmann centroid of a cluster of sampled structures.

Following [34,35,59,60], we define the maximum expected accuracy (MEA) structure for a given RNA sequence to be that which is obtained by tracebackings, using the matrix $M$, defined as follows:

$$M_{ij} = \begin{cases} 
0 & \text{if } j - i \leq 4 \\
\max\left(M_{i-1,j} + \beta q_j, \max_{k=1}^{4} 2x_{ik} p_{ij} + M_{i,j-1} + M_{i+1,j-1}\right) & \text{else}
\end{cases}$$

where $q_j = 1 - \sum_{i=0}^{n} p_{ij}$, and $x, \beta$ are non-negative constants. In the previous studies [34,35], optimal values of $x, \beta$ were found to be $x = 1, \beta = 1$. In this paper, we have set $\beta = 1$ and performed benchmarking for a range of values $x$ in $\{2^{-4} -2, 2^{-3}, 2^{-2}, 2^{-1}, 1, 2, 2^2, 2^3, 2^4\}$. If most structures in the Boltzmann ensemble contain the base pair $(i,j)$, then $p_{ij}$ will be large, and it can happen that $(i,j)$ will belong to the MEA structure even though $(i,j)$ does not belong to the MFE structure. The values $M_{ij}$ can be computed by a simple modification of the Nussinov-Jacobson algorithm [5], and the maximum expected accuracy structure with score $M_{ij}$ can be subsequently computed by tracebackings. See the references for more
Figure 6. Plot of ratio, with error bars, of the restricted Boltzmann partition function $Z_{LO}$ and the total Boltzmann partition function, as a function of RNA length, for the same random RNA generated as described in the Figure 5. This ratio represents the percentage of structures, as weighted by their Boltzmann factor, that are locally optimal. By numerical fitting, we find that this ratio is approximately $1.0053 \exp(-0.0123n)$ with coefficient of determination (see [54]) $R^2 = 0.9876$.

Details, where the previously cited authors list benchmarking statistics to determine optimal parameters $\alpha, \beta$ for which the MEA structure is closer to the native structure than is the MFE structure.

We compare variants of the MEA construction, obtained by using (i) base pairing probabilities $P_{ij}^M$ computed by McCaskill’s algorithm [33] using RNAFOLD, (ii) base pairing probabilities $P_{ij}^{LO}$ for locally optimal structures computed by relative frequency count from 10,000 sampled locally optimal structures, and (iii) base pairing probabilities $P_{ij}^{\text{min}}$, and unpaired probabilities $q_{ij}^{\text{min}}$, defined as the minimum of both probabilities; i.e. $P_{ij}^{\text{min}} = \min(P_{ij}^M, P_{ij}^{LO})$, and $q_{ij}^{\text{min}} = \min(q_{ij}^M, q_{ij}^{LO})$. Note that in case (iii) it is no longer the case that $\sum_{ij} P_{ij}^{\text{min}} + \sum_{ij} q_{ij}^{\text{min}} = 1$. Cases (i), (ii), and (iii) yield the base pairing distributions $P$ (McCaskill), $P_{LO}$ (locally optimal) and $P_{MIN}$ (minimum of McCaskill and locally optimal).

We can determine corresponding MEA structures, denoted by MEA and MEA$_{LO}$, according to the use of $P$ resp. $P_{LO}$. We see in Figures 7 and 8 that predictions based on these MEA structures are better than the MFE structure, as predicted by RNAFOLD. However the predictions based on local optima are consistently worse.

However, we can create a third matrix, denoted by $P_{MIN}$, where for each base pair $(i,j)$,

$$P_{MIN}(i,j) = \min(P((i,j)), P_{LO}(i,j)).$$

This will in essence emphasize those base pairs that occur prominently in both samples of local optima and samples of all structures. As shown in Figure 7, this consistently increases the sensitivity and positive predictive value.

Discussion

In this paper, we describe a novel and efficient algorithm to compute the partition function over all locally optimal secondary structures of a given RNA sequence. The software, RNAlocopt, runs in $O(n^2)$ time and $O(n^3)$ space, the same time and space complexity as that of McCaskill’s algorithm to compute the partition function over all secondary structures. Additionally, RNAlocopt samples a user-specified number of structures from the Boltzmann subensemble of all locally optimal structures. Our work completely solves a line of investigation begun originally by M. Zuker [6], who first defined the notion of saturated structure (for which no base pair can be added without violating the definition of secondary structure).

The energy model implemented in RNAlocopt is the Turner nearest neighbor energy model without dangles; in contrast, the energy model used in the software RNAfold and RNAsubopt is the Turner model with dangles. Our computation of sensitivity and positive predictive value (PPV) is exact; i.e. with no allowed slippage. In contrast, some authors, such as Lu and Mathews [35], benchmark sensitivity and positive predictive values by allowing a slippage of $\pm 1$; i.e. if base pair $(i,j)$ belongs to the native structure, then the predicted base pair $(x,y)$ is counted as correctly predicted if $(x,y)$ is one of the following: $(i-1,j), (i,j), (i+1,j), (i,j-1), (i,j+1)$. In [35], sensitivity and PPV values are reported with slippage for the maximum expected accuracy (MEA) method using the software RNAstructure [62], which includes energy terms for coaxial stacking.

There may be some discrepancies between reported sensitivity and PPV values from various groups. Such discrepancies will occur due to a combination of benchmarking with respect to
compared to all structures, we are led to the hypothesis that in
diversity between locally optimal and all structures. 

riboswitch aptamers should display a difference in structural
essentially missing from the Rfam data, a situation we will address
riboswitch aptamers, which do not undergo any significant
measured by four different parameters, of samples of locally
approximately equals the square root of the number of all
accuracy (MEA) methods.

is a consistent improvement of MEA MIN, as shown in this paper,
over both minimum free energy (MFE) and maximum expected

By applying RNALOCOPT to randomly generated RNA, we have
shown that there are far fewer locally optimal structures than that of all
structures, and MFE and maximum expected accuracy (MEA) methods.

By applying RNALOCOPT to randomly generated RNA, we have
shown that there are far fewer locally optimal structures than that of all
structures, and MFE and maximum expected accuracy (MEA) methods.

As we have shown, the structural diversity, as measured by four different parameters, of samples of locally
optimal structures can either be similar or quite distinct from
samples from the Boltzmann ensemble of all structures – a
dependence that is particular to the RNA family. While most
RNA families we investigated displayed smaller locally optimal
diversity than total structure diversity, notable exceptions were the
riboswitch aptamers from Rfam. One might think that this is due
to the fact that two distinct low energy conformations (gene-on and
gene-off) are present in both the local optimal and Boltzmann
ensemble. However, the Rfam database contains only the
riboswitch aptamers, which do not undergo any significant
conformation change. (Indeed, the riboswitch portion that under-
goes conformation change, called the expression platform, is
essentially missing from the Rfam data, a situation we will address
in a future publication.) Thus it remains unclear exactly why
riboswitch aptamers should display a difference in structural
diversity between locally optimal and all structures.

Since there are relatively few locally optimal structures, compared to all structures, we are led to the hypothesis that in
certain circumstances, a collection of sampled locally optimal
structures can more succinctly represent the folding landscape of a
given RNA sequence. In forthcoming work, we will describe an
application of this observation, by presenting a new method for de
nova RNA structure design, where kinetic properties are taken into
account.

Theoretical studies of RNA folding kinetics have primarily
focused on unit-step resolution, where a single base pair is added
or removed in each time step. For such studies, RNALOCOPT will
prove to be a valuable new tool. There is some possibility of
extending RNALOCOPT to allow the formation or removal of
entire helices in each time step, a direction we are currently
considering. The idea would be to redefine a locally optimal
structure to be one for which no addition or removal of any stem
region would lower the free energy. An extension of RNALOCOPT
in this direction would allow more rapid exploration of the
folding process.

Locally optimal structures \( S \) form kinetic traps, in the sense that
there does not exist a structure \( T \), obtained from \( S \) by the removal
or addition of a single base pair, which has lower free energy.
Since thermal noise can overcome the energy barrier between
certain conformations in the low energy ensemble, a better model
of kinetic trap might arguably be a that of a basin of attraction
located around locally optimal structure \( S \). Such a basin would be a
set \( S \) of low energy structures, such that: (i) there is a folding path
whose barrier energy is less than a fixed energy threshold \( \varepsilon \) that
cannot be overcome by thermal noise, and (ii) if \( T \) is reachable by
a folding pathway from \( S \) with barrier energy less than \( \varepsilon \), then
\( T \in S \). Though it is currently unclear what value of \( \varepsilon \) should
be taken, it may be possible to extend RNALOCOPT in this direction.
This is a possible avenue for future research. (A folding pathway
from \( S \) to \( T \) is a sequence \( S = S_0 S_1 \ldots S_n = T \) of secondary
structures, such that \( S_i S_{i+1} \) is obtained by adding or removing
a single base pair from \( S_{i+1} \), for each \( 0 \leq i < n \). The barrier energy of a
folding pathway is \( \max \{ E(S_i) - E(S_{i+1}) : 0 \leq i < n \} \). Computing
optimal folding pathways between any two secondary structures is
known to be NP-complete, though there are exponential time exact
algorithms [24,63] and efficient near optimal algorithms [31].)

Finally, we have shown the utility of locally optimal structures
by demonstrating that the variant of maximum expected
accuracy structure, MEA MIN, provides the most accuracy
structure prediction currently available via thermodynamic
methods. The improvement in sensitivity and PPV for this
method depends on the fact that we take into account the base
pairing frequency of pairs \((i,j)\) within the ensemble of locally
optimal structures as well as that of the Boltzmann ensemble of all
structures.

Why is the MEA MIN structure apparently closer to the
native structure, at least in the benchmarking study performed
in this paper? Since there is no clear answer to this question, we
can only formulate a guess. Recall that there are far fewer
locally optimal structures than there are of all secondary
structures, and that the ensemble of locally optimal structures
appears to be more consistent (i.e. less structurally diverse, at
least in most cases) than the ensemble of all structures. For these
two reasons, certain unlikely, pathological candidate base pairs
have diminished likelihood of contributing to the MEA LO
structure. However, certain important intermediate structures,
which do not appear in the ensemble of locally optimal
structures, could contribute to the accuracy of the MEA
structure. By taking the minimum of base pairing probabilities
over both ensembles, MEA MIN is closer to the native structure.
Though reasonable, we must stress that this explanation can only be speculative.

| SeqLen | RNAsubopt | RNALOCOPT | %Z subopt | ZLO |
|--------|-----------|-----------|-----------|-----|
| 20     | 42.9±2.7  | 93.0±0.6  | 0.9696±0.0199 | 0.9981±0.0202 |
| 40     | 150.9±11.0| 37.2±3.3  | 0.7886±0.0302 | 0.9852±0.0199 |
| 60     | 352.7±20.9| 93.2±8.3  | 0.5173±0.0319 | 0.9504±0.0199 |
| 80     | 540.4±28.2| 178.7±14.8| 0.2607±0.0291 | 0.8811±0.0211 |
| 100    | 719.3±28.7| 250.7±18.4| 0.1437±0.0225 | 0.8034±0.0262 |
| 120    | 813.7±25.9| 332.3±25.6| 0.0609±0.0144 | 0.7314±0.0290 |
| 140    | 831.4±27.7| 374.0±27.0| 0.0442±0.0115 | 0.6134±0.0336 |
| 160    | 936.7±22.6| 536.3±30.3| 0.0035±0.0014 | 0.4853±0.0335 |
| 180    | 954.7±22.6| 576.4±30.1| 0.0046±0.0032 | 0.3845±0.0348 |
| 200    | 987.0±20.1| 644.9±26.5| 0.0030±0.0016 | 0.3802±0.0322 |

For each value of \( n \), 1000 structures were sampled, by applying the Ding-
Lawrence sampling algorithm [36], as implemented in RNAsubopt with flag -p,
and by applying RNALOCOPT. For each run, the number of non-redundant
samples is computed, yielding the expected number \( n \) for RNAsubopt and
RNALOCOPT, where \( n \) is the error bound (standard deviation \( \varepsilon / \sqrt{50} \), since 50
sequences generated). For each run the percent coverage of the partition
function was computed; i.e. the sum of the Boltzmann factors of the non-
redundant collection from 1000 samples generated by RNAsubopt [resp.
RNALOCOPT], divided by the partition function \( Z \) [resp. partition function \( Z_{\text{LO}} \)
of locally optimal structures]. Since the number of locally optimal structures is far
fewer than that of all structures (see Figure 5), it is not surprising that there is
proportionately more redundancy among sampled locally optimal structures
than over all structures. As well, the percentage coverage of the partition
function for sampled locally optimal structures is higher than that for the
Boltzmann ensemble.

doi:10.1371/journal.pone.0016178.t001
Table 2. Structural diversity comparison between ensemble of locally optimal structures and Boltzmann ensemble of all structures.

| Rfam family | L/M | $H_0$ | $\mu_H$ | $\sigma_H$ | $\langle D_{obs} \rangle$ | $\langle D \rangle$ | num | corCoeff($H_0$, $\mu_H$) |
|-------------|-----|-------|---------|---------|------------------|-----------------|-----|----------------------|
| RF00001     | L   | 14.403| 0.247   | 0.231   | 23.276           | 15.160          | 710 | 0.6934               |
|             | M   | 19.122| 0.327   | 0.283   | 28.543           | 18.918          |     |                      |
| RF00003     | L   | 29.167| 0.357   | 0.285   | 42.046           | 29.487          | 100 | 0.70183              |
|             | M   | 36.703| 0.450   | 0.3318  | 51.372           | 36.147          |     |                      |
| RF00004     | L   | 23.770| 0.248   | 0.251   | 38.010           | 24.478          | 212 | 0.68387              |
|             | M   | 28.086| 0.294   | 0.295   | 42.623           | 27.779          |     |                      |
| RF00005     | L   | 11.637| 0.315   | 0.259   | 18.013           | 11.811          | 1052| 0.62225              |
|             | M   | 12.058| 0.325   | 0.272   | 18.322           | 11.979          |     |                      |
| RF00008     | L   | 3.088 | 0.108   | 0.127   | 5.014            | 3.318           | 84  | 0.53836              |
|             | M   | 3.435 | 0.116   | 0.161   | 6.051            | 3.594           |     |                      |
| RF00017     | L   | 30.883| 0.205   | 0.247   | 48.606           | 33.072          | 104 | 0.63080              |
|             | M   | 45.936| 0.307   | 0.296   | 66.852           | 46.554          |     |                      |
| RF00031     | L   | 6.600 | 0.201   | 0.215   | 10.697           | 6.903           | 61  | 0.80180              |
|             | M   | 9.259 | 0.278   | 0.250   | 14.199           | 9.137           |     |                      |
| RF00050     | L   | 29.743| 0.441   | 0.327   | 43.230           | 29.449          | 147 | 0.61214              |
|             | M   | 26.096| 0.382   | 0.348   | 36.694           | 25.447          |     |                      |
| RF00059     | L   | 17.626| 0.318   | 0.278   | 26.416           | 18.008          | 118 | 0.64729              |
|             | M   | 17.979| 0.320   | 0.287   | 26.645           | 18.123          |     |                      |
| RF00162     | L   | 12.448| 0.228   | 0.226   | 20.356           | 13.343          | 228 | 0.62482              |
|             | M   | 12.219| 0.222   | 0.255   | 19.610           | 12.756          |     |                      |
| RF00167     | L   | 13.227| 0.261   | 0.219   | 20.411           | 12.771          | 133 | 0.71049              |
|             | M   | 13.344| 0.262   | 0.237   | 20.299           | 12.537          |     |                      |
| RF00168     | L   | 29.169| 0.316   | 0.282   | 43.205           | 29.396          | 47  | 0.66707              |
|             | M   | 37.286| 0.405   | 0.341   | 52.967           | 36.244          |     |                      |
| RF00174     | L   | 34.448| 0.338   | 0.297   | 52.693           | 35.340          | 439 | 0.59984              |
|             | M   | 42.376| 0.417   | 0.361   | 58.045           | 40.617          |     |                      |
| RF00380     | L   | 18.675| 0.219   | 0.234   | 30.579           | 20.039          | 96  | 0.72220              |
|             | M   | 19.438| 0.228   | 0.25112 | 30.801           | 20.266          |     |                      |

Given the collection of base pairing probabilities $p_{ij}$ over all locally optimal structures [resp. over all structures] of a given RNA sequence $a_1, \ldots, a_n$, we define four measures of structural diversity. (1) The pseudo-entropy $H_0$ is defined by $H_0 = -\sum_{j=1}^n p_{0,ij} \ln p_{0,ij}$. (2) The average entropy $\langle H \rangle$ is defined by $\langle H \rangle = \sum_{j=1}^n \sum_{i=1}^n p_{ij} \ln p_{ij}$. (3) The Morgan-Higgs structural diversity $\langle D_{obs} \rangle$ is defined by $\langle D_{obs} \rangle = n - \sum_{j=1}^n \sum_{i=1}^n p_{ij} \ln p_{ij}$, where we define $p_{0,ij} = 0 - \sum_i p_{ij}$. (4) The Vienna structural diversity $\langle D \rangle$ is defined by $\langle D \rangle = \sum_{j=1}^n \sum_{i=1}^n P_{ij} (1-p_{ij})$. In the table above, we consider these measures with respect to locally optimal structures (L) and with respect to all (M) structures. (‘L’ stands for locally optimal, and ‘M’ for McCaskill.) The table depicts the number of structures for each Rfam family considered, as well as the correlation coefficient between structural diversity.

Methods

We begin by providing an intuitive overview of the construction, while subsequent sections provide full details and the recurrence relations for the RNALOCOPt algorithm.

Conditional local optimality

To implement our algorithm, at each step we wish to calculate the partition function of only the locally optimal structures. Since the Turner energy model is a loop-based model, it can largely be construed as a local model. Therefore we can locally check whether or not adding the base pair $(i,j)$ makes some structures suddenly no longer locally optimal simply by looking at nucleotides near $(i,j)$. To do this, we need to keep track of a bit more information during our recursion than is done in McCaskill’s algorithm.

In this section we show through a simple example the key idea behind the recursions. Consider the partial sequence-structure shown in the left side of Figure 9. The Boltzmann factor (portion of the Boltzmann partition function) of this structure would be shown in the left side of Figure 9. The Boltzmann factor (portion of the Boltzmann partition function) of this structure would be

$$Z_B = \sum_{i=1}^n \sum_{j=1}^n p_{ij} \ln p_{ij}.$$
any base pair would raise the energy, or keep it the same. But in our example structure, there is one base pair for which we do not have sufficient information to know the change in energy caused by removing it — namely, the outer base pair \((i,j)\). In the next recursive steps, this structure could be extended in several different possible ways, perhaps with a base pair \((x,y)\) shown in the right-hand side of Figure 9. At that point, we will know the energy of the two loops in which the base pair \((i,j)\) is contained. But until then, this energy is unknown.

Since we do not yet know how removing the base pair \((i,j)\) will affect the energy, the best we can do is to inductively assume that the structure is conditionally locally optimal, conditioned on the fact that \((i,j)\) must base pair. It will not be until we add the next base pair \((x,y)\) that we will know whether the base pair \((i,j)\) causes the structure to not be locally optimal, that is if removing the base pair \((i,j)\) decreases the energy.

Consider then the structure including the base pair \((x,y)\) on the left side of Figure 10. Remember that we could not determine the change in energy caused by removing base pair \((i,j)\) before. That change in energy is now given by the energy of the new loop, \(E(L_5)\) minus the energy of the old loops, \(E(L_4)+E(L_3)\), as indexed in Figure 10. However, to determine the energies \(E(L_4)\) and \(E(L_5)\), we need to know the location of the base pair \((a,b)\).

Our approach for this internal loop example is to induct on the last two base pairs, not just the last base pair. So in our example, our example structure on the left-hand side of Figure 9 will contribute to the term \(Z_B((i,j),(a,b))\), which denotes the partition function of all locally optimal structures with the outermost two base pairs \((i,j)\) and \((a,b)\). Then, if removing the base pair \((i,j)\) doesn’t lower the energy, that is if \(E(L_4)-(E(L_4)+E(L_3))\geq 0\), the structure on the right-hand side of Figure 9 will contribute to the term \(Z_B((x,y),(i,j))\).

We must also check if any base pairs can be added. In our example, when adding the base pair \((x,y)\), we check if any base pairs can be added within the internal loop \(L_5\) defined by \((i,j)\) and \((x,y)\) (see Figures 9 and 10). Any other base pair additions would already have been considered earlier in the recursion, and the energy change of adding different base pairs is independent due to the loop energy model.

The previous discussion deals with internal loops. For external loops and multiloops, the motivation is similar, but the approach is more difficult, and the solution, which is more time-consuming and depends at least theoretically on the parameters of the Turner energy model, is less satisfying. As the recursion progresses, the conditionality of the optimality will be pushed outward, and in checking the final external loop, the conditionality will be
free energy (MFE) structure predictions, benchmarked by using RNAFOLD a consistent improvement other MEA methods, as well as over minimum average. Results are similar; data not shown.) The MEA MIN method yields were taken over all sequences in Rfam, without first computing a family alignment in Rfam 10.0. (We performed a similar analysis where averages are computed by Rfam family, then averaged over all families of seed alignments from all seed alignments of Rfam 10.0 database [42]. Values of $\alpha = 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3$ given on x-axis, while values of sensitivity and PPV are given on the y-axis. Sensitivity and PPV are computed by Rfam family, then averaged over all families of seed alignment in Rfam 10.0. (We performed a similar analysis where averages were taken over all sequences in Rfam, without first computing a family average. Results are similar; data not shown.) The MEA MIN method yields a consistent improvement other MEA methods, as well as over minimum free energy (MFE) structure predictions, benchmarked by using RNAfold from the Vienna RNA package [58]. The best sensitivity and the best PPV are given by method MEA MIN; the next best by MEA LO, and the last by method MEA. Two horizontal lines indicate the sensitivity (top line) and PPV (bottom line) for the minimum free energy structure, as computed by RNAfold from the Vienna RNA Package. doi:10.1371/journal.pone.0016178.g008

removed, giving the full partition function and completing the recursion.

Details of recursion for locally optimal structures
To do our recursion, we need to know the energies of various internal loops, hairpins, and the energies associated with a multiloop in the Turner energy model. These are available as temperature-dependent parameters. For simplicity, all calculations will be at 37°C.

We let $E_{IL}(i,j,i',j')$ denote the free energy of an internal loop enclosed by two base pairs, $(i,j)$ and $(i',j')$, where $i < i' < j' < j$. The energy of a hairpin enclosed by a base pair $(i,j)$ will be denoted by $E_{HP}(i,j)$. For a multiloop, such notation is not possible. The accepted energy of a multiloop is given by a multiloop penalty, $a$, a penalty for unpaired bases in a multiloops, $b$, and a penalty or bonus for a base pair within a multiloop, $c$, which can depend on the type of base pair being considered. The energy of the multiloop is then

$$E_{ML} = a + k \cdot b + l \cdot c$$

where $k$ is the number of unpaired bases in the multiloop, and $l$ is the number of bases in the multiloop. This is standard, as used in McCaskill’s algorithm, and is done in part for computational reasons. There is no affine energy term associated with external loops, but their treatment is somewhat analogous to that of multiloops (indeed, a multiloop can be formed by adding a closing base pair to an external loop).

Explanation of deltas
The method of calculating local optima is straightforward. We will calculate the partition function of locally optimal structures with the same basic McCaskill algorithm used to calculate the partition function over all secondary structures. However, some modifications must be made, for at each step in our recursion, we must make sure that no base pair can be added or removed that would lower the energy. Anything that does not satisfy this property is dropped from the partition function.

The way this is done is to realize all of the different ways a single base pair can be added and removed that can lower the total energy, and to build in a check for all of these cases as we build the partition function. Figure 11 shows all of the possibilities. A base pair can be added to or removed from a hairpin, (Types 3 and 4), an internal loop, which includes bulges and stacked base pairs (Types 1 and 2), or a multiloop (types 5 and 6).

In our recursion, we will have six different delta functions corresponding to these six different cases, where each delta function is 1 if adding, or removing, the relevant base pair does not lower the energy. Such deltas will act as checks whether the structures built so far are locally optimal.

For example, to check whether we can remove a base pair from between two internal loops, we have, from type 1 in Figure 11,

$$\delta^1(i,j,i',j') = \begin{cases} 0, & \text{if removing bp } (i',j') \text{ lowers the energy} \\ 1, & \text{otherwise} \end{cases}$$

This delta is calculated using the energies of a given segment. The energy of the internal loops before removing the base pair are

![Figure 8. Graph showing sensitivity (black, increasing curves) and positive predictive value (PPV, red, decreasing curves) as a function of $\alpha$ (explained in text and in Figure 7) for methods MEA, MEA LO, and MEA MIN. as benchmarked with consensus structures from all seed alignments of Rfam 10.0 database [42]. Values of $\alpha = 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3$ given on x-axis, while values of sensitivity and PPV are given on the y-axis. Sensitivity and PPV are computed by Rfam family, then averaged over all families of seed alignment in Rfam 10.0. (We performed a similar analysis where averages were taken over all sequences in Rfam, without first computing a family average. Results are similar; data not shown.) The MEA MIN method yields a consistent improvement other MEA methods, as well as over minimum free energy (MFE) structure predictions, benchmarked by using RNAfold from the Vienna RNA package [58]. The best sensitivity and the best PPV are given by method MEA MIN; the next best by MEA LO, and the last by method MEA. Two horizontal lines indicate the sensitivity (top line) and PPV (bottom line) for the minimum free energy structure, as computed by RNAfold from the Vienna RNA Package. doi:10.1371/journal.pone.0016178.g008](http://www.plosone.org)[Figure 8. Graph showing sensitivity (black, increasing curves) and positive predictive value (PPV, red, decreasing curves) as a function of $\alpha$ (explained in text and in Figure 7) for methods MEA, MEA LO, and MEA MIN. as benchmarked with consensus structures from all seed alignments of Rfam 10.0 database [42]. Values of $\alpha = 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3$ given on x-axis, while values of sensitivity and PPV are given on the y-axis. Sensitivity and PPV are computed by Rfam family, then averaged over all families of seed alignment in Rfam 10.0. (We performed a similar analysis where averages were taken over all sequences in Rfam, without first computing a family average. Results are similar; data not shown.) The MEA MIN method yields a consistent improvement other MEA methods, as well as over minimum free energy (MFE) structure predictions, benchmarked by using RNAfold from the Vienna RNA package [58]. The best sensitivity and the best PPV are given by method MEA MIN; the next best by MEA LO, and the last by method MEA. Two horizontal lines indicate the sensitivity (top line) and PPV (bottom line) for the minimum free energy structure, as computed by RNAfold from the Vienna RNA Package. doi:10.1371/journal.pone.0016178.g008)

![Figure 9. Example structure in recursion. In the left structure, we do not yet know the two loops bordered by the base pair $(i,j)$. Therefore we do not yet know whether by removing this base pair, the free energy will be lowered. In the right structure, one step further in the recursion, we now know which loops border the base pair $(i,j)$ – namely, loops $L_4$ and $L_5$. Images created using the software VARNA [50]. doi:10.1371/journal.pone.0016178.g009](http://www.plosone.org)[Figure 9. Example structure in recursion. In the left structure, we do not yet know the two loops bordered by the base pair $(i,j)$. Therefore we do not yet know whether by removing this base pair, the free energy will be lowered. In the right structure, one step further in the recursion, we now know which loops border the base pair $(i,j)$ – namely, loops $L_4$ and $L_5$. Images created using the software VARNA [50]. doi:10.1371/journal.pone.0016178.g009)
The energy change effected by removing the base pair \((i,j)\) is \(E(L_6) - (E(L_5) + E(L_4))\). To calculate this, we need to keep track of base pair \((a,b)\). Images created using the software VARNA [50].

doi:10.1371/journal.pone.0016178.g010

One note is that some base pairs are never favorable, and thus do not need to be calculated. The important case is adding a base pair to a multiloop, which would split the multiloop into two multiloops when the multiloop is closed. This type of base pair is shown in Figure 12. Provided that there are no energy terms for either dangles within a multiloop, or coaxial stacking, this base pair will never lower the energy. This is fortunate, since it is computationally more difficult to inductively include such base pairs.

### Tails, conditional optimality

Just as there are hairpins, internal loops, multiloops, and external loops in the Turner energy model, there are recursion terms for hairpins, internal loops, multiloops, and external loops. However, as we need to keep a little more context to keep track of whether we still have a set of local optima, there will be some extra information.

Note that all these structures will be conditionally locally optimal. We commonly can’t know if the most exterior base pair will be locally optimal, as that will depend on future base pairs, thus we need this conditional optimality in order to perform the recursion.

---

**Figure 10.** Example structure in recursion. The energy change effected by removing the base pair \((i,j)\) is \(E(L_6) - (E(L_5) + E(L_4))\). To calculate this, we need to keep track of base pair \((a,b)\). Images created using the software VARNA [50].

doi:10.1371/journal.pone.0016178.g010

**Figure 11.** The six ways that a single base pair can be added to or removed from a structure and possibly reduce the overall energy. Images created using the software VARNA [50].

doi:10.1371/journal.pone.0016178.g011
For example, for internal loops, we will denote

$$Z_{M}(i,j)$$

for the partition function of all locally optimal structures on the subinterval (i,j) with an internal loop with base pairs (i,j) and (i’,j’), where i < i’ < j’. This local optimality is conditional on (i,j) being a base pair, that is, we assume (i,j) is a base pair, and will check later if this is a problem. We cannot tell whether, in the future, removing the base pair (i,j) will lower the energy or not, as we don’t know the structure outside of (i,j).

A few of the recursive elements will contain tails. For example, for multiloops, we will let

$$Z_{ML}(i,j,p,q)$$

denote the partition function, on the interval (i,j), for all unclosed locally optimal multiloops (with more than one base pair) that have ‘tails’, regions of unpaired nucleotides, of lengths p and q on the left and right side respectively. See Figure 13.

These tails are needed. In McCaskill’s algorithm, for a multiloop closed by base pair (i,j), there is a recursion of the form

$$Z_{ML}(i,j) = e^{a_{wobble}} Z_{ML}(i,j-1) + \text{remaining terms}$$

We cannot use such a recursion, as adding an unpaired base may result in a structure that is no longer a local optimum. While there may be better approaches, we avoid this problem by indexing locally optimal multiloops by their tail length. We can then glue such multiloops together with tails. See Figure 14.

We have seen that p and q can be always less than 10, this is sufficient to avoid all possible base pairs in multiloops that lower energy. Almost all such base pairs can be avoided by setting p and q to be always less than 4; this allows for considerable speed-up with little loss of accuracy.

Note that we need the assumption that a single base pair cannot split a multiloop into two multiloops and thereby lower the energy. (This is true under the present Turner energy model. See Figure 12.) Otherwise, such a gluing method could result in a base pair being possible that lowers the energy – that is, the structure would not be locally optimal.

**Recursion Relations**

Let $$Z'(i,j)$$ denote the partition function of all structures ending in the base pair (i,j) which will enter a multiloop. Note that we know from the Turner energy parameters that only an internal loop can enter a multiloop. It follows that $$Z'(i,j)$$ will be the sum of all possible internal loops ending in (i,j).

$$Z'(i,j) = e^{-(s_{wobble} + b)/kT} \sum_{m=1}^{10} Z_{ML}(i,j,m) \delta_{m M}(i,j)$$

where $$\delta_{m M}$$ is 1 if we have an AU or GU base pair, $$p_{wobble}$$ is the corresponding energy penalty, b is the penalty of adding a base pair in a multiloop, and $$\delta_{m M}$$ is 0 if removing the base pair (i,j) (and exposing the base pair (i’,j’) to the multiloop) lowers the energy. Thus (by induction) $$Z'(i,j)$$ is the partition function for all structures that are locally optimal with respect to all of their base pairs, including (i,j).

$$Z_{ML}(i,j,p,q)$$ is the partition function for locally optimal multiloops closed by base pair (i,j) and having with exactly one component, while $$Z_{M}(i,j,p,q)$$ is the partition function for locally optimal multiloops with exactly one component, and which contains tails of length p and q. We let p, q range from 0 to 10, with one extra position, called “>10+”, which is reserved for long tails. Thus p and q each have 12 possible values. (However, in practice, most values of $$Z_{ML}$$ are not stored, but calculated as needed. Only those with 1 or 2 long tails need to be stored.)

The partition function $$Z_{ML}(i,j,p,q)$$ corresponds to having a base pair at (i+p−q) entering a multiloop, with tails out to (i,j), i and j not base-paired. $$Z_{M}(i,j,m,>10)$$ means an M1 element with large right tail, greater than 10. This is used because if either tail is of length >10, there are no longer any base pairs that can be.

**Figure 13. Example of the formation of a multiloop with tails of length p and q.** Images created using the software VARNA [50].

**Figure 14. Example of gluing together two pieces of a multiloop.** Note that if each piece is locally optimal, then the composite, obtained by gluing the pieces together, is as well. Images created using the software VARNA [50].

**Figure 12. Image of base pair that could not possibly lower the energy by creating a multiloop, since it creates two bordering multiloops.** Images created using the software VARNA [50].

doi:10.1371/journal.pone.0016178.g012
added that can reduce the energy. This was shown by exhaustive search. This allows for traditional induction, since we don’t have to worry about adding a base pair causing a formally locally optimal structure to become not locally optimal. For \( p, q \in \{0,1,2,\ldots,10\} \),

\[
Z_M(i,j,p,q) = Z(i+p-j-q)e^{-(p+q)/kT}d_M(i,j+p,j-q)
\]

\[
Z_M(i,j,p,> 10) = Z_M(i-1,p,> 10) + Z_M(i-1,p,10)
\]

\[
Z_M(i,j,> 10,q) = Z_M(i+1,j,> 10,q) + Z_M(i+1,j,10,q)
\]

\[
Z_M(i,j,> 10,> 10) = Z_M(i-1,j,> 10,> 10) + Z_M(i-1,j,10,10)
\]

where \( c \) is the energy penalty of an unpaired base in a multiloop, and \( d_M(i,j+p,j-q) = 1 \) if the base pair \((i+p,j-q)\) is such that no base pair \((x,y)\) can be added \( 1 \leq x < i+p, \ j-q < y \leq j \), that lowers the energy. That is, the base pair \((i+m,j-n)\) is locally optimal with tails in the multiloop of length \( m \) and \( n \) on the left and right respectively.

Note, we need another variable, \( Z_{ML} \), for the partition function of external loops with exactly one element. The recursion relations are almost identical. The only change is there is no base pair penalty.

\( Z_M \) is the partition function of multiloops with at least 2 exiting base pairs. Tails are glued together as in Figure 14. Notation is similar to \( Z_M \), for the same reasons. Here, the recursion is quite nice.

Define the set \( S = \{1,2,3,\ldots,10,> 10\} \). For \( p, q \in S \)

\[
Z_M(i,j,p,q) = \sum_{k=1}^{i+j} \sum_{a=1}^{10} (Z_M(i,k,p,a) + Z_M(i,k,p,a))Z_M(k+1-a,j,q)
\]

where in the expression \((k+1-a)\), we replace \( > 10 \) with 10. (This corresponds to unambiguously gluing the largest possible fixed tail. Otherwise there are several ambiguous ways to glue two long tails together.)

Thus we add a single exiting base pair with tails during the recursion. Note, with 12 possible tail lengths, the memory usage here is \( 144N^2/2 \). As cases of isolated base pairs far into a multiloop lowering the energy are rare, we can reduce the number of tail lengths recorded.

A similar equation for the external loop can be determined. Here we can always assume that the left end of the external loop (usually denoted with the variable \( i \)) is 1, since we never need to close an external loop. Also, we need only worry about the right tail, for the same reason. Remember, an external loop can contain 0, 1, or more entering base pairs, corresponding respectively to the empty structure, structure with one component, and structures with more than one component. In this way it is slightly different than a multiloop.

\[
Z_{ML}(i,j,q) = \sum_{k=1}^{i} \sum_{a=1}^{10} (Z_{ML}(i,k,a)Z_M(k+1-a,j,q) + \delta_{q-j,> 10,q> 10})
\]

where \( \delta_{q-j,> 10,q> 10} \) is 1 if \( q = j \) or \( q > 10 \) and \( j > 10 \), and where again in the expression \((k+1-a)\), \( > 10 \) is replaced by 10. The term \( \delta_{q-j,> 10,q> 10} \) actually represents the empty structure. Note that \( Z_{EL}(1,0,0) \) will be set to 1 by the above equation, as will \( Z_{EL}(1,j,j) \). These can be thought of as representing the empty structure, or equivalently as initial conditions.

The variable \( Z_{MLC}(i,j) \) represents the partition function of all locally optimal closed multiloops ending in base pair \((i,j)\). It is given by all of the ways to end a multiloop.

\[
Z_{MLC}(i,j) = e^{-\delta_{wobble}}\sum_{p,q} Z_{MLC}(i,j,p,q)ML(i+1,j,1-p,q)
\]

where \( \delta_{wobble} \) is as before, \( a \) is the closing penalty of a multiloop, and \( \delta_{MLC}(i,j,p,j-q) = 1 \) if there is no base pair \((x,y)\), \( i \leq x \leq i+p-j-q \leq x < j \), that would lower the energy of the multiloop. That is, within the available tails that close the ML, there is no way to add a base pair connecting these tails and lowering the energy.

All that is left is the partition function \( Z_{UL}(i,j,i',j') \). This is the partition function of all structures that are locally optimal, conditional on \( ij \) base-pairing, that exit in an internal loop with outmost base pairs \((i',j')\) and \((i,j)\), \( i < i' < j' < j \). Following standard convention, we consider only internal loops of size at most 30; i.e. we can restrict to the case \( i' = i+j-j' \leq 30 \).

There are 3 cases: (i) the internal loop borders a hairpin at \((i',j')\), (ii) the internal loop borders a multiloop at \((i',j')\), (iii) the internal loop borders another internal loop with base pairs \((i',j')\), \((i',j')\). In all 3 cases, we need to do our inductive checks on optimality. For the last case, we must sum over all possible internal loops. (In practice, there is a prerecorded set of possible internal loops, increasing speed considerably.) The recursion is a sum over these three cases and is given by

\[
Z_{UL}(i,j,i',j') = e^{-E_{UL}(i,j,i',j')/kT}d_{UL}(i,j,i',j')
\]

\[
+ \sum_{i'' = 1}^{j'} Z_{UL}(i',j',i'',j'')d_{UL}(i,j,i',i'')d_{UL}(i'',j'',j')
\]

\[
Z_{UL}(i',j')d_{UL}(i,j,i',j') + Z_{MLC}(i',j')d_{MLC}(i,j,i',j')
\]

where \( d_{UL}(i,j,i',j') = 0 \) if removing the base pair \((i',j')\) lowers the energy, and \( d_{UL}(i,j,i',j') = 1 \) if no base pair \((x,y)\), \( i < x < i' < j' < j \) can be added that will (split the multiloop in two and) lower the energy. \( d_{HP}(i,j,i',j') \) and \( d_{MLC}(i,j,i',j') \) both check if removing \((i',j')\) lowers the energy.

\( Z_{HP}(i,j) \) is the partition function of a locally optimal hairpin with outer base pair \((i,j)\), conditional on \( i,j \) being base paired. We have

\[
Z_{HP}(i,j) = \delta_{HP}(i,j)e^{-E_{HP}(i,j)/kT}
\]

where \( E_{HP}(i,j) \) is the Turner energy for the hairpin with external base pair \((i,j)\), and \( \delta_{HP}(i,j) = 1 \) if the hairpin is locally optimal, that is if no base pair \((i',j')\), \( i < i' < j' < j \), can be added that would lower the energy.

This gives consistent recursions. To calculate the total partition function, simply sum up all of the external loops with different tail lengths to yield
\[ Z = \sum_{q} Z_{G}(q,q) \]

Acknowledgments

Research for this paper was carried out while W.A. Lorenz was at Boston College. A preliminary report of this method was described in the poster paper Q54, “Calculating Local Optima in the Turner Energy Model for RNA Secondary Structure”, presented by W.A. Lorenz and P. Clote at the Sixteenth Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2008). We would like to thank Hosam Jabbari for an improvement to Figure 4, and to anonymous referees for some helpful suggestions.

Author Contributions

Main algorithm was developed and implemented by W.A. conceived and designed the experiments: WAL, PGC. Performed the experiments: WAL. Analyzed the data: WAL. Contributed reagents/materials/analysis tools: WAL. Wrote the paper: WAL.

References

1. Al-Hashimi HM, Walter NG (2008) RNA dynamics: it is about time.Curr Opin Struct Biol 18: 325–330.
2. Fanch T, Gulyaev AP, Gerdes K (1997) Programmed cell death by hok/sok of plasmid r1: Processing at the hok mRNA 3H-end triggers structural rearrangements that allow translation and antisense RNA binding. J Mol Biol 273: 38–51.
3. Lecuyer K, Crothers D (1993) The Leptomons collosoma spliced leader RNA can switch between two alternate structural forms. Biochemistry 32(20): 5301–5311.
4. Heilman-Miller SL, Woodson SA (2003) Effect of transcription on folding of the eubacterial ribosome. RNA 9: 722–733.
5. Nussinov R, Jacobson AB (1980) Fast algorithm for predicting the secondary structure of single stranded RNA. Proceedings of the National Academy of Sciences, USA 77: 6309–6313.
6. Zuker M (1986) RNA folding: The continued need for interaction between biologists and mathematicians. In: Lectures on Mathematics in the Life Sciences. Springer-Verlag, volume 17: 87–124.
7. Clote P (2005) An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. J Comput Biol 12: 83–101.
8. Clote P (2006) Combinatorics of saturated secondary structures of RNA. J Comput Biol 13: 1640–1657.
9. Clote P, Kranakis E, Krizanc D, Salvy B (2009) Asymptotics of canonical and saturated RNA secondary structures. J Bioinform Comput Biol 7: 869–895.
10. Stein PR, Waterman MS (1978) On some new sequences generalizing the Catalan and Motzkin numbers. Discrete Mathematics 26: 261–272.
11. Wallach J, Clote P (2007) Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the Turner energy model. J Comput Biol 14: 190–213.
12. Xia T, J. Santalucia J, Burkard M, Krierz R, Schroeder S, et al. (1999) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 38: 14719–33.
13. Zuker M, Sankoff D (1984) RNA secondary structures and their prediction. Bulletin of Mathematical Biology 46: 591–621.
14. Mathews D, Sabina J, Zuker M, Turner D (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J Mol Biol 288: 911–940.
15. 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.
16. Flamme C, Fontana W, Hofacker I, Schuster P (2000) RNA folding at elementary rearrangements that allow translation and antisense RNA binding. J Mol Biol 288: 911–940.
17. Flamm C, Hofacker I, Stadler B, Stadler P (2007) Saddles and barrier in landscapes of generalized search operators. In: Foundations of Genetic Algorithms, Springer, volume 4436 of Lecture Notes in Computer Science, pp 194–212.
18. Hofacker I, Flamme C, Heine C, Wolflinger MT, Schustermann G, et al. (2010) Barmnap: RNA folding on dynamic energy landscapes. RNA 0: O.
19. Shapiro BA, Bengali D, Kauppi KL, Wu JC (2001) RNA folding pathway functional intermediates: their prediction and analysis. J Mol Biol 312: 27–44.
20. Isambert H (2009) The jerky and knotty dynamics of RNA. Methods 49: 190–196.
21. Dotu I, Lorenza WA, VAN Hentenryck P, Clote P (2010) Computing folding pathways between RNA secondary structures. Nucleic Acids Res 38: 1711–1722.
22. Chen SJ (2008) RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu Rev Biophys 37: 197–214.
23. McCaskill JL (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29: 1105–1119.
24. Kiryu H, Kim T, Asai K (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. Bioinformatics 23: 434–441.
25. Lu ZJ, Gloor JW, Mathews DH (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA 15: 1805–1813.
26. Ding Y, Lawrence C (2003) A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res 31(24): 7280–7301.
27. Ding Y, Chan CV, Lawrence CE (2004) Stochastic web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res 32: 4603–4615.
28. Pollard KS, Salama SR, Lambert N, Lambert MA, Coppen S, et al. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443: 167–172.
29. Leontis N, Wenzhof E (2003) Tools for the automatic identification and classification of RNA base pairs. Nucleic Acids Res 31(13): 3450–3460.
30. Banerjee AR, Jaeger J, Turner D (1993) Thermal unfoldung of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. Biochemistry 32: 153–163.
31. Lynge RR, Pedersen CN (2000) RNA pseudoknot prediction in energy-based models. J Comput Biol 7: 409–427.
32. Gardner PP, Deja B, Tate GE, Jaworski EP, Kolbe DH, et al. (2009) Rfam: The RNA families database. Nucleic Acids Res 37: D136–D140.
33. Wiese KC, Glen E, Vasudevan A (2005) JvRNA: A Java tool for RNA secondary structure visualization. IEEE Trans Nanobioscience 4: 212–218.
34. Tauber M, Lexon A, Araia R, Mirescu D, Van Buren BF, et al. (2009) Pseudoknot**, an extension of RNAfold, for easy secondary structure assignment, formatting and visualization of pseudoknots. Nucleic Acids Res 37: D127–D135.
35. Turner DH, Sugimoto N, Freier SM (1983) RNA structure prediction. Annu Rev Biophys Biophys Chem 12: 167–192.
36. Jaeger JA, Turner DH, Zuker M (1989) Improved predictions of secondary structures for RNA. Proc Natl Acad Sci USA 86: 7706–7710.
37. He L, Krierz R, Santalucia J, Jr., Walter AE, Turner DH (1991) Nearest-neighbor parameters for G.U mismatches: [formula; see text] and [formula; see text] destabilizing in the contexts [formula; see text] and [formula; see text] but stabilizing in [formula; see text]. Biochemistry 30: 11124–11132.
38. Periz AE, Krierz R, Sugimoto N, Turner DH (1991) Thermodynamic study of internal loops in oligoribonucleotides: symmetry loops are more stable than asymmetric loops. Biochemistry 30: 6426–6436.
39. Walter AE, Turner DH, Kim J, Lytle MH, Muller P, et al. (1994) Coaxial stacking of helices enhances binding of oligohomocleotides and improves predictions of RNA folding. Proc Natl Acad Sci USA 91: 9214–9222.
40. Darty K, Denise A, Ponty Y (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 25: 1974–1975.
41. Mathews D, Sabina J, Zuker M, Turner D (1999) Experimental dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288: 911–940.
42. McCaskill JL (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29: 1105–1119.
43. Hofacker I, Fontana W, Schuster P, Mirkovic D, Zuker M, et al. (1994) Fast folding and comparison of RNA secondary structures. Monatsch Chem 125: 167–188.
44. Zar J (1999) Biostatistical Analysis. Prentice-Hall, Inc.
55. Waldispühl J, Clote P (2007) Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the Turner energy model. J Comput Biol 14: 190–213.
56. Clote P (2005) RNALOSS: a web server for RNA locally optimal secondary structures. Nucleic Acids Res 33: W600–W604.
57. Morgan S, Higgins P (1996) Barrier heights between ground states in a model of RNA secondary structure. J Phys A: Math Gen 31: 3153–3170.
58. Hofacker I (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31: 3429–3431.
59. Knudsen B, Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res 31: 3423–3428.
60. Do CB, Mahabhashyam MS, Brutnho M, Batzoglou S (2005) Probcons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15: 330–340.
61. Ding Y, Chan CY, Lawrence CE (2003) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA 11: 1157–1166.
62. Mathews D, Disney M, Childs J, Schroeder S, Zuker M, et al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci USA 101: 7287–7292.
63. Thachuk C, Manuch J, Rafiey A, Mathieson LA, Stacho L, et al. (2010) An algorithm for the energy barrier problem without pseudoknots and temporary arcs. Pac Symp Biocomput.