Energy Parameters and Novel Algorithms for an Extended Nearest Neighbor Energy Model of RNA

Ivan Dotu[1], Vinodh Mechery[2], Peter Clote[1]*

[1] Biology Department, Boston College, Chestnut Hill, Massachusetts, United States of America, [2] Hofstra North Shore-LIJ School of Medicine, Hempstead, New York, United States of America

Abstract

We describe the first algorithm and software, RNAenn, to compute the partition function and minimum free energy secondary structure for RNA with respect to an *extended nearest neighbor* energy model. Our next-nearest-neighbor triplet energy model appears to lead to somewhat more *cooperative* folding than does the nearest neighbor energy model, as judged by melting curves computed with RNAenn and with two popular software implementations for the nearest-neighbor energy model. A web server is available at http://bioinformatics.bc.edu/clotelab/RNAenn/.

Citation: Dotu I, Mechery V, Clote P (2014) Energy Parameters and Novel Algorithms for an Extended Nearest Neighbor Energy Model of RNA. PLoS ONE 9(2): e85412. doi:10.1371/journal.pone.0085412

Editor: Danny Barash, Ben-Gurion University, Israel

Received October 20, 2013; **Accepted** December 4, 2013; **Published** February 21, 2014

Funding: Funded by NSF (http://www.nsf.gov) grant DMS-1016618, entitled "Energy parameters and novel algorithms for an extended nearest neighbor energy model of RNA", as well as NSF DBI-1262439. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: clote@bc.edu

Introduction

Thermodynamics-based *ab initio* RNA secondary structure algorithms are used to detect microRNAs [1], targets of microRNAs [2], non-coding RNA genes [3], temperature-dependent riboregulators [4], selenoproteins [5], ribosomal frameshift locations [6], RNA-protein binding sites [7], etc. The importance and ubiquity of RNA thermodynamics-based algorithms cannot be overemphasized – there are even applications in RNA design for novel cancer therapies and in synthetic biology. Indeed, in [8] Vashishta et al. used the RNA minimum free energy (MFE) structure prediction algorithm mfold [9] to design seven anti-pCD ribozymes, four of which were cloned, stably transfected in the highly metastatic human breast cancer cell line, MDA-MB-231, and shown to have a therapeutic potential by knocking down the expression of pCD. (Procathepsin D (pCD) is correlated with highly invasive malignancies, such as breast cancer. Ribozymes, first discovered by the Nobel laureats, T. Cech and S. Altman, are RNA enzymes that can cleave a molecule or catalyze a reaction.)

Following pioneering work of the Tinoco Lab and Freier et al. [10], a number of increasingly sophisticated *nearest neighbor* models have been defined: INN [11,12], INN-HB, also called *Turner99* [13], *Turner2004* [14,15], as well as models that incorporate knowledge-based parameters [16,17]. These free energy parameters of the *nearest neighbor* (NN) model form the foundation for essentially all current thermodynamics-based RNA algorithms: minimum free energy (MFE) secondary structure [9,18], Boltzmann partition function [19], maximum expected accuracy secondary structure [20], MFE secondary structure with pseudoknots [21], sampling suboptimal structures [22], RNA sequence-structure alignments [23], etc.

Benchmarking studies have shown that, on average, the minimum free energy structure includes 73% of base pairs in X-ray structures when domains of fewer than 700 nucleotides (nt) are folded [24]; i.e. prediction *sensitivity* of the MFE structure is 73%, although accuracy drops as sequence length increases. There is increasing evidence that by improving the free energy parameters, structure prediction accuracy can be improved. Andronescu et al. [16] used combinatorial optimization to determine optimal weights α,β for which energy parameters are determined by α-weighted contribution from Turner's free energies together with β-weighted contribution from knowledge-based potentials, the latter obtained from the negative logarithm of frequencies in existent structure databases. Free energy parameters in the Turner model are determined by a least-squares fit of UV absorption data based on the assumption that change in heat capacity, ΔC_P, is zero. This assumption is erroneous, as pointed out by Mikulecky and Feig [25], who observed that the hammerhead ribozyme does not fold in 2-state transition, but rather has 3 states: cold denatured, folded and hot denatured. In [17] M. Bon improved MFE structure prediction by defining new parameters for the nearest neighbor model that account for linear dependence of change ΔC_P of heat capacity on sequence length and by incorporating knowledge-based potentials from a hand-curated selection of Sprinzl's transfer RNA database [26].

Subsection 1.1: Motivation from protein helix-coil transition

Consider a coarse-grain classification of amino acids, where a polypeptide chain is given by an n-mer, or length n sequence $a_1,\ldots,a_n$ of amino acids, where each residue a_i is either in an H (α-helix) or C (coil) conformation. Assume that the energy of an α-helical residue is $E(H)=\epsilon_0<0$, while that of a coil residue is $E(C)=0$. A protein with many residues in an α-helical conformation at room temperature, such as hemoglobin, will unfold into a random coil at a higher temperature, where all previous H residues have been transformed into C residues. In particular, if

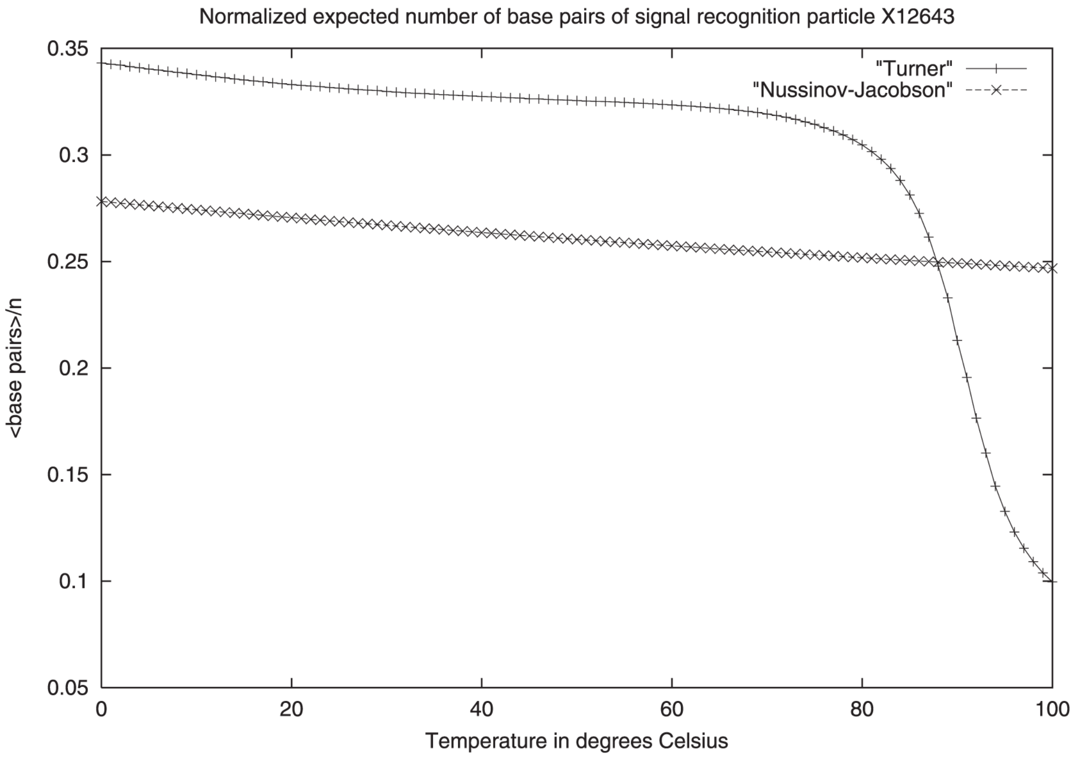


Figure 1. Graph of the expected number of base pairs as a function of temperature for signal recognition particle with Rfam [56] accession number X12643. Temperature in degrees Celsius is given on the x-axis, while the expected number of base pairs $\langle \text{base pairs}/n \rangle$, normalized by sequence length n, is given on the y-axis. Note the linear dependence on temperature for the non-cooperative Nussinov energy model, in contrast to the sigmoidal dependence on temperature for the cooperative Turner energy model. Data and figure taken from our paper [57].
doi:10.1371/journal.pone.0085412.g001

$a_1, \ldots, a_n$ is an α-helix, then at low temperature, all residues are H, while at high temperature all residues are C. The partition function Z of $a_1, \ldots, a_n$ is defined by $Z = \sum_s \exp(-E(s)/RT)$, where the sum is taken over all 2^n many sequences s of H's and C's. Using the (temperature-dependent) partition function, we can compute the expected number $\langle H \rangle$ of α-helical residues for the n-mer $a_1, \ldots, a_n$ at absolute temperature T, defined by

$$\langle H \rangle = \frac{\partial \ln Z}{\partial \ln s} \qquad (1)$$

where $s = \exp(-\epsilon_0/RT)$ – see [27]. Subsequently, it is possible to plot the *expected helical fraction* $\frac{\langle H \rangle}{n}$ as a function of temperature. Non-cooperative energy models show an approximately linear relation, where the expected helical fraction slowly decreases as temperature increases. In contrast, the plot of expected helical fraction versus temperature for *cooperative* energy models displays a *sigmoidal* shape, where there is an abrupt helix-coil transition from high to low values for the helical fraction that occurs at a *critical temperature* T_M.

Polymer theory provides several mathematical models to explain the temperature-dependent *helix-coil transition* for proteins. The simplest polymer model for the helix-coil transition of an α-helix is the *non-cooperative* model, where the probability that the each residue is H is independent of the conformation of every other residue. The cooperative, nearest-neighbor model for helix-coil transition, introduced by Zimm and Bragg [28], includes nucleation free energy $\delta > 0$ that is applied for each α-helical segment of contiguous H residues. Finally, the Ising model was introduced by E. Ising in 1925 [29] to explain ferromagnetism, but has subsequently been used to model protein temperature-dependent helix-coil transitions – see, for instance [30]. Progressing from the independent model to the Zimm-Bragg model to the Ising model, each model is increasingly cooperative, thus providing a better fit to the experimental data. See Dill and Bromberg [27] for a more detailed discussion.

In the Nussinov energy model [31] for RNA secondary structure, the free energy of a secondary structure S is defined to be -1 times the number $|S|$ of base pairs of S; i.e. in the Nussinov model, each base pair contributes an energy of -1, and there is no energy term for entropic considerations. The Turner energy model [13,32] for RNA secondary structure contains negative free energies for base stacks, which depend on the nucleotides involved, such as the base stacking free energy of -2.24 kcal/mol at 37^0C for $\begin{array}{c} 5'-\mathrm{AC}-3' \\ 3'-\mathrm{UG}-5' \end{array}$ and of -3.26 kcal/mol for $\begin{array}{c} 5'-\mathrm{CC}-3' \\ 3'-\mathrm{GG}-5'. \end{array}$. Additionally the Turner energy model contains free energies for various loops (hairpin, bulge, internal loop, multiloop) that include entropic considerations. Clearly, the Turner energy model for RNA secondary structure is analogous to the cooperative, nearest-neighbor model for helix-coil transitions introduced by Zimm and Bragg. Figure 1 contrasts the temperature-dependent cooperativity of the Turner energy model with the temperature-independent non-cooperativity of the Nussinov energy model. The motivation for this paper is to solve the equation: $\frac{?}{\text{Turner}} = \frac{\text{Ising}}{\text{Zimm}-\text{Bragg}}$. Though we do not determine

Table 1. Values of sensitivity and positive predictive value (ppv) for RNAfold and RNAenn with respect to various RNA families.

	RNAfold -d ∅		RNAenn (Turner99)		RNAenn (Turner04)	
RNA family	sens	ppv	sens	ppv	sens	ppv
16s	0.3940	0.3326	0.3779	0.3153	0.3099	0.2674
23sd	0.5974	0.5311	0.5527	0.4813	0.4409	0.4003
23s	0.4685	0.3972	0.4453	0.3738	0.3516	0.3061
5s	0.7575	0.6606	0.7319	0.6366	0.5713	0.5093
ec	0.5869	0.5314	0.6184	0.5519	0.5338	0.4982
grp1	0.6625	0.5832	0.5047	0.4650	0.4837	0.4589
grplii	0.6616	0.6409	0.6084	0.5976	0.4555	0.4334
rnap1	0.4051	0.3813	0.3514	0.3221	0.3154	0.3000
rnap2	0.4241	0.4046	0.4935	0.4648	0.3285	0.3187
short	0.4048	0.3400	0.3690	0.3298	0.3155	0.2760
srp	0.7228	0.5632	0.6286	0.4897	0.5677	0.4544
telomerase	0.4285	0.3074	0.3417	0.2404	0.3605	0.2662
tmRNA	0.2248	0.1958	0.1911	0.1622	0.1526	0.1326
trna2	0.4960	0.4697	0.5344	0.5005	0.3962	0.3828
avg	0.5213	0.4575	0.4866	0.4279	0.4020	0.3678

Sensitivity is the ratio of number of correctly predicted base pairs divided by the number of base pairs in the native structure; positive predictive value is the ratio of the number of correctly predicted base pairs divided by the number of base pairs in the predicted structure. Since RNAenn currently does not include energy contributions for dangles (single stranded, stacked nucleotides), RNAfold was used without dangles (version 1.8.5 with -d ∅ flag). To our knowledge, there has not been a careful benchmarking of structure prediction accuracy between the Turner 1999 energy model and the newer Turner 2004 energy model, though it is interesting to note that RNAenn has better structure prediction when using Turner 1999 for base stacking. Overall, it is clear that RNAfold outperforms RNAenn (Turner99), although a few cases, such as **ec** and **rnap2** RNAenn have better sensitivity. Nevertheless, we expect much better performance in the future when our triplet and base stacking energy terms have been refined by using knowledge-base potentials. The database of RNA structures in this benchmarking set comes from a data collection of D.H. Mathews (personal communication), which derives from published databases [26,54], etc. See [55] for a citation of original data sources.

doi:10.1371/journal.pone.0085412.t001

the analogue of the Ising model for RNA secondary structure formation, we do introduce an *extended nearest-neighbor* model, also called *triplet* or *next-nearest-neighbor* model, which displays somewhat more cooperativity, as displayed in the sharpness of the transition from folded to unfolded state in a figure shown later in the paper.

Subsection 1.2: Triplet model

As previously mentioned, the nearest-neighbor energy model [13,32] assigns free energies for base stacks of the form $\begin{array}{c}5'-\mathrm{AB}-3'\\3'-\mathrm{DC}-5'\end{array}$ for the formation of a stacked base pair between $5'-AB-3'$ with $5'-CD-3'$. In contrast, the extended-nearest-neighbor triplet model assigns free energies for triplexes of the form $\begin{array}{c}5'-\mathrm{ABC}-3'\\3'-\mathrm{FED}-5'\end{array}$ where a stacked triple (two contiguous base stacks) between $5'-ABC-3'$ and $5'-DEF-3'$. In this case, we expect that the triplet free energy of $\begin{array}{c}5'-\mathrm{ABC}-3'\\3'-\mathrm{FED}-5'\end{array}$ can be approximated by the average of the base stacking free energies for $\begin{array}{c}5'-\mathrm{AB}-3'\\3'-\mathrm{FE}-5'\end{array}$ and $\begin{array}{c}5'-\mathrm{BC}-3'\\3'-\mathrm{ED}-5'\end{array}$; however, we expect the triplet energies to more accurately model the formation of secondary RNA structure.

The extended-nearest-neighbor triplet model for hybridized DNA duplexes and DNA-RNA hybrids was considered in experimental work of D.M. Gray, who in Table 1 of [12] determined the theoretical number of independent hybridized sequences that must be considered in UV absorbance experiments, in order to obtain triplet stacking free energies by least-squares fitting of data. In [33] Gray et al. experimentally determined *in vivo* inhibition parameters for next-nearest-neighbor triplets in the case of antisense DNA – RNA hybridization to inhibit protein expression. In [34], Najafabadi et al. applied a neural network to predict the thermodynamic parameters for the next-nearest-neighbor triplet model, using existent UV absorbance data from the thermodynamic database for nucleic acids, NTDB version 2.0 [35].

Though at present there are no experimentally determined free energies for triplet stacking, Binder et al. [36] did show a strong correlation between microarray fluorescence intensities and DNA-RNA base stacking free energies of Sugimoto et al. [37]. More precisely, Binder et al. showed that linear combinations of triple-averaged probe sensitivities provide nearest-neighbor *sensitivity* terms, that rank in similar order as the base stacking free energy parameters for DNA-RNA in solution [37]. It is our hope that future improvements in RNAseq, microarray or other technologies will ultimately furnish experimentally determined triplet and even k-tuple stacking free energies. New triplet free energies could immediately be incorporated into our algorithms, and it is tedious, but clear how one can modify our algorithms to handle k-tuple free energies.

Subsection 1.3: Plan of the paper

In this paper, we describe the first algorithms to compute the partition function and minimum free energy structure for single-stranded RNA, with respect to the full *next-nearest-neighbor triplet* energy model for RNA. In the *Introduction*, we gave the motivation for this work, coming from the Zimm-Bragg and Ising models in biopolymer theory. The plan for the remainder of the paper is as follows. In the *Results* section, Section 2.1 gives the notation and definitions needed for the sequel, while Section 2.2 presents the extended nearest neighbor model and method used to obtain energy parameters. In the *Discussion*, we give secondary structure benchmarking results for the nearest neighbor (NN) and extended nearest neighbor (ENN) energy models. Additionally, the cooperativity of folding is compared with both energy models. In the *Methods* section, Section 3.1 [resp. Section 3.2] presents recursions for the partition function [resp. minimum free energy structure] computation. In addition to the software RNAnn and RNAenn developed for this paper, we use Vienna RNA Package RNAfold [18], RNAstructure [38], and mfold [9]. As illustration for the cooperativity of folding, we compare melting curves for two small nucleolar RNAs (snoRNA), with respect to the NN and ENN energy models; additional melting curves are available on the web server http://bioinformatics.bc.edu/clotelab/RNAenn/. These results suggest that the the extended nearest-neighbor energy model may lead to more cooperative folding than does the nearest-neighbor model, which was our motivation to study the ENN energy model.

The goal of this paper is to describe the non-trivial RNAenn algorithms, which are implemented in C/C++. Our work points toward a future potential improvement in RNA secondary structure prediction, either by incorporating triplet knowledge-based potentials or experimentally inferred extended nearest-neighbor free energy parameters.

Results: Extended nearest neighbor model algorithms

Assume that $a_1,\ldots,a_n$ is a given RNA sequence. In this section, we describe pseudocode for the partition function and minimum free energy computation for an extended nearest neighbor model. Although our software, RNAenn, does depend on the exact values of the extended nearest-neighbor energy parameters, the description of the algorithms does not.

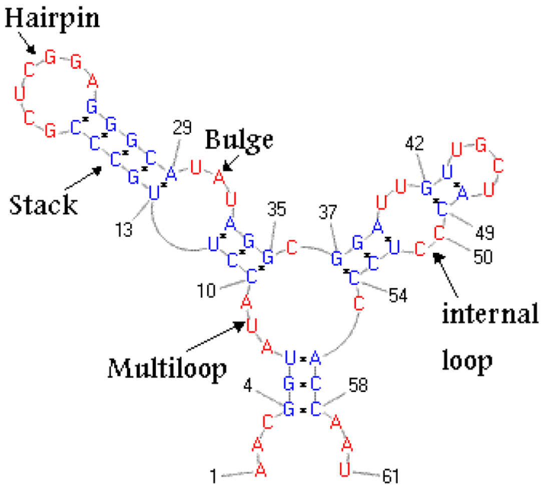


Figure 2. Depiction of elements of RNA secondary structure for which experimentally determined free energy parameters are available. In this 61 nt RNA, the hairpin loop closed by base pair between nucleotides at position 43 and 48 is known as a *tetraloop*, or hairpin loop of size 4. Similarly, the hairpin loop of size 7 is closed by a base pair between nucleotides at positions 17 and 25. Free energy parameters for bulges and internal loops (two-sided bulges, not shown in the figure) are available, while an affine approximation is used for the free energy of a multiloop or junction.
doi:10.1371/journal.pone.0085412.g002

Subsection 2.1: Notation and definitions

Let $a=a_1,\ldots,a_n$ be an arbitrary RNA sequence, and let $a[i,j]$ denote the subsequence $a_i,\ldots,a_j$. A *secondary structure* S for a given RNA sequence $a=a_1,\ldots,a_n$ is a set of *base pairs* (i,j), $1\le i<j\le n$, such that (1) a_i,a_j forms a Watson Crick AU, UA, GC, CG or wobble GU, UG pair; (2) each base is paired to at most one other base, i.e. $(i,j),(i,k)\in S$ implies that $j=k$, and $(i,j),(k,j)\in S$ implies that $i=k$; (3) there are no pseudoknots in S, where a pseudoknot consists of base pairs $(i,j),(k,\ell)$ where $i<k<j<\ell$; (4) each hairpin loop has at least θ unpaired bases; i.e. $(i,j)\in S$ implies that $j-i\ge\theta+1$.

In software such as mfold [9], Unafold [39], RNAfold [40], and RNAstructure [38], the parameter θ, denoting the minimum number of unpaired bases in a hairpin loop, is taken to be equal to 3, due to steric constraints of RNA molecules.

The nearest-neighbor and extended nearest-neighbor triplet models are additive energy models that entail free energy values for *loops*, as explained in [41]. A *hairpin* in a secondary structure S is defined by the base pair (i,j), where $i+1,\ldots,j-1$ are unpaired. A *left bulge* in S is defined by the two base pairs $(i,j),(k,j-1)\in S$, where $i+1<k$ and $i+1,\ldots,k-1$ are unpaired. A *right bulge* in S is defined by the two base pairs $(i,j),(i+1,k)\in S$, where $k<j-1$ and $k+1,\ldots,j-1$ are unpaired. An *internal loop* in S is defined by the two base pairs $(i,j),(k,\ell)\in S$, where $i+1<k$ and $\ell<j-1$ and $i+1,\ldots,k-1$ and $\ell+1,\ldots,j-1$ are unpaired. Finally, a k-way junction, or multiloop with $k-1$ components, is defined by the closing base pair (i,j) and $k-1$ inner base pairs $(x_1,y_1),\ldots,(x_{k-1},y_{k-1})$, where $i<x_1<y_1<x_2<y_2<\cdots<x_{k-1}<y_{k-1}<j$, and the nucleotides in intervals $[i+1,x_1-1]$, $[y_1+1,x_2-1],\ldots,[y_{k-1}+1,j-1]$ are all unpaired. See Figure 2 for an illustration.

Given the RNA nucleotide sequence $a_1,\ldots,a_n$, we use the notation $\mathbb{H}$ to denote the free energy of a hairpin, $E(\mathbb{S})$ to denote the free energy of a stacked base pair, $E(\mathbb{I})$ to denote the free energy of an internal loop, $E(\mathbb{B})$ to denote the free energy of a bulge, while the free energy $E(\mathbb{M})$ for a multiloop containing N_b base pairs and N_u unpaired bases is given by the affine approximation $a+bN_b+cN_u$. The free energy $E(\mathbb{M}1)$ of a multiloop having exactly one component is then given by $a+b+cN_u$.

For RNA sequence $a_1,\ldots,a_n$, for all $1\le i\le j\le n$, the 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 $a[i,j]$, $E(S)$ is the free energy of secondary structure S, R is the universal gas constant with value $R=0.001987$ kcal/mol^{-1} K^{-1}, and T is absolute temperature. In the Zuker [9,18,38] and McCaskill [19] algorithms, $E(S)$ is the Turner nearest neighbor energy model; in contrast, when discussing the extended nearest-neighbor energy model, we use $E(S)$ to denote the triplet energy model.

Given an RNA sequence $a_1,\ldots,a_n$, in order to compute the partition function $Z_{1,n}$ [resp. minimum free energy $E_{1,n}$] for $a_1,\ldots,a_n$, we need inductively to determine the partition function $Z_{i,j}$ [resp. minimum free energy $E_{i,j}$] for all smaller subsequences $a_i,\ldots,a_j$. In so doing, we need to know which structures involve a triple stack $(i,j),(i+1,j-1),(i+2,j-2)$, which structures involve only a stacked pair $(i,j),(i+1,j-1)$, and which structures involve a base pair (i,j) which closes a loop region. This is accomplished by terms $ZBB_{i,j}$ [resp. $EBB_{i,j}$]. Moreover, the Turner energy model stipulates that a base pair (i,j), which closes a left bulge of size 1, as in $(i,j),(i+2,j-1)$, or a right bulge of size 1, as in $(i,j),(i+1,j-2)$, is considered to stack on the subsequent base pair. This consideration requires the introduction of special terms $ZBBL_{i,j}$,

$ZBBR_{i,j}$ [resp. $EBBL_{i,j}$, $EBBR_{i,j}$]. With that, we have the following definition.

Definition 1 (Energies and partition function for triplet loop model)

- $E_{i,i+1;j-1,j}$ *denotes the base stacking free energy from the NN model, while* $E_{i,i+1,i+2;j-2,j-1,j}$ *denotes the triplet stacking free energy from the ENN model.*
- $Z_{i,j}$: *partition function over all secondary structures of* $a[i,j]$.
- $ZB_{i,j}$: *partition function over all secondary structures of* $a[i,j]$, *which contain the base pair* (i,j).
- $ZBB_{i,j}$: *partition function over all secondary structures of* $a[i,j]$, *which contain the base pairs* $(i,j),(i+1,j-1)$.
- $ZBBL_{i,j}$: *partition function over all secondary structures of* $a[i,j]$, *which contain the base pairs* $(i,j),(i+2,j-1)$.
- $ZBBR_{i,j}$: *partition function over all secondary structures of* $a[i,j]$, *which contain the base pairs* $(i,j),(i+1,j-2)$.
- $ZM_{i,j}$: *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.*
- $ZM1_{i,j}$: *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* $i<r\leq j$.
- $E_{i,j}$: *minimum free energy over all secondary structures of* $a[i,j]$.
- $EB_{i,j}$: *minimum free energy over all secondary structures of* $a[i,j]$, *which contain the base pair* (i,j).
- $EBB_{i,j}$: *minimum free energy over all secondary structures of* $a[i,j]$, *which contain the base pairs* $(i,j),(i+1,j-1)$.
- $EBBL_{i,j}$: *minimum free energy over all secondary structures of* $a[i,j]$, *which contain the base pairs* $(i,j),(i+2,j-1)$.
- $EBBR_{i,j}$: *minimum free energy over all secondary structures of* $a[i,j]$, *which contain the base pairs* $(i,j),(i+1,j-2)$.
- $EM_{i,j}$: *minimum free energy 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.*
- $EM1_{i,j}$: *minimum free energy 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* $i<r\leq j$.

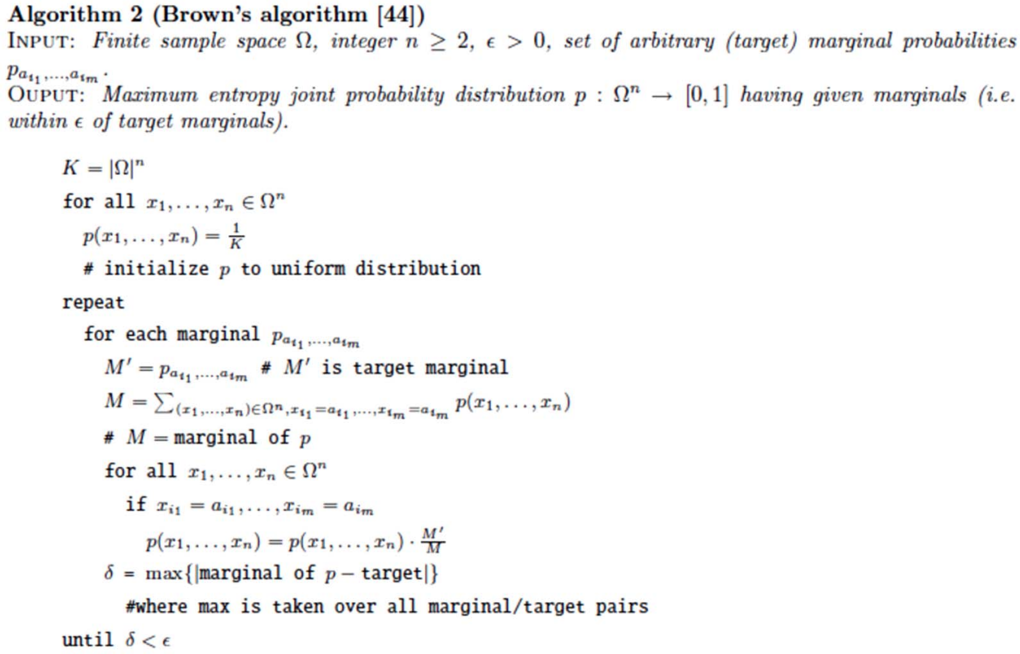
Details for the recursions necessary to compute the ENN minimum free energy secondary structure and ENN partition function are given in the *Methods* section.

Subsection 2.2: Extended nearest-neighbor energy model ENN-13

Here we describe details for the extended nearest-neighbor energy model parameters, which we denote by ENN-13, since our code RNAenn was completed in 2013.

Though some related experimental work has been done, especially by D.M. Gray and co-workers [11,12,33], there are no available experimentally determined parameters for triplet stacking. Rather than using the triplet stacking free energy parameters INN-48 [34], which are incomplete since GU-wobble pairs were not included, we instead infer RNA triplet stacking free energies by a novel use of Brown's algorithm [42], which computes the *maximum entropy* joint probability distribution that is consistent with given user-specified marginal probabilities. Though Brown's algorithm has been used by C. Burge to predict intron-exon splice sites in the human gene finder, *GenScan* [43], this appears to be the first use of Brown's algorithm to infer free energy parameters.

Brown's algorithm for maximum entropy joint distribution. In [42], D.T. Brown described an efficient algorithm to compute the *maximum entropy* joint probability distribution given certain marginal probabilities, where we recall that the entropy of a joint probability distribution $p:\Omega^n \rightarrow [0,1]$ is defined by $H(p)=-\sum_{x_1,\ldots,x_n\in\Omega} p(x_1,\ldots,x_n)\cdot \ln p(x_1,\ldots,x_n)$. Although the algorithm was correct, there was an error in Brown's



Algorithm 2 (Brown's algorithm [44])

INPUT: *Finite sample space Ω, integer $n \geq 2$, $\epsilon > 0$, set of arbitrary (target) marginal probabilities* $p_{a_{i_1},\ldots,a_{i_m}}$.

OUPUT: *Maximum entropy joint probability distribution $p:\Omega^n \rightarrow [0,1]$ having given marginals (i.e. within ϵ of target marginals).*

```
K = |Ω|^n
for all $x_1,\ldots,x_n \in \Omega^n$
  $p(x_1,\ldots,x_n) = \frac{1}{K}$
  # initialize $p$ to uniform distribution
repeat
  for each marginal $p_{a_{i_1},\ldots,a_{i_m}}$
    $M' = p_{a_{i_1},\ldots,a_{i_m}}$ # $M'$ is target marginal
    $M = \sum_{(x_1,\ldots,x_n)\in\Omega^n, x_{i_1}=a_{i_1},\ldots,x_{i_m}=a_{i_m}} p(x_1,\ldots,x_n)$
    # $M$ = marginal of $p$
    for all $x_1,\ldots,x_n \in \Omega^n$
      if $x_{i_1}=a_{i_1},\ldots,x_{i_m}=a_{i_m}$
        $p(x_1,\ldots,x_n) = p(x_1,\ldots,x_n)\cdot\frac{M'}{M}$
    $\delta$ = max{|marginal of $p$ − target|}
      #where max is taken over all marginal/target pairs
until $\delta < \epsilon$
```

Figure 3. Pseudocode for Brown's algorithm.
doi:10.1371/journal.pone.0085412.g003

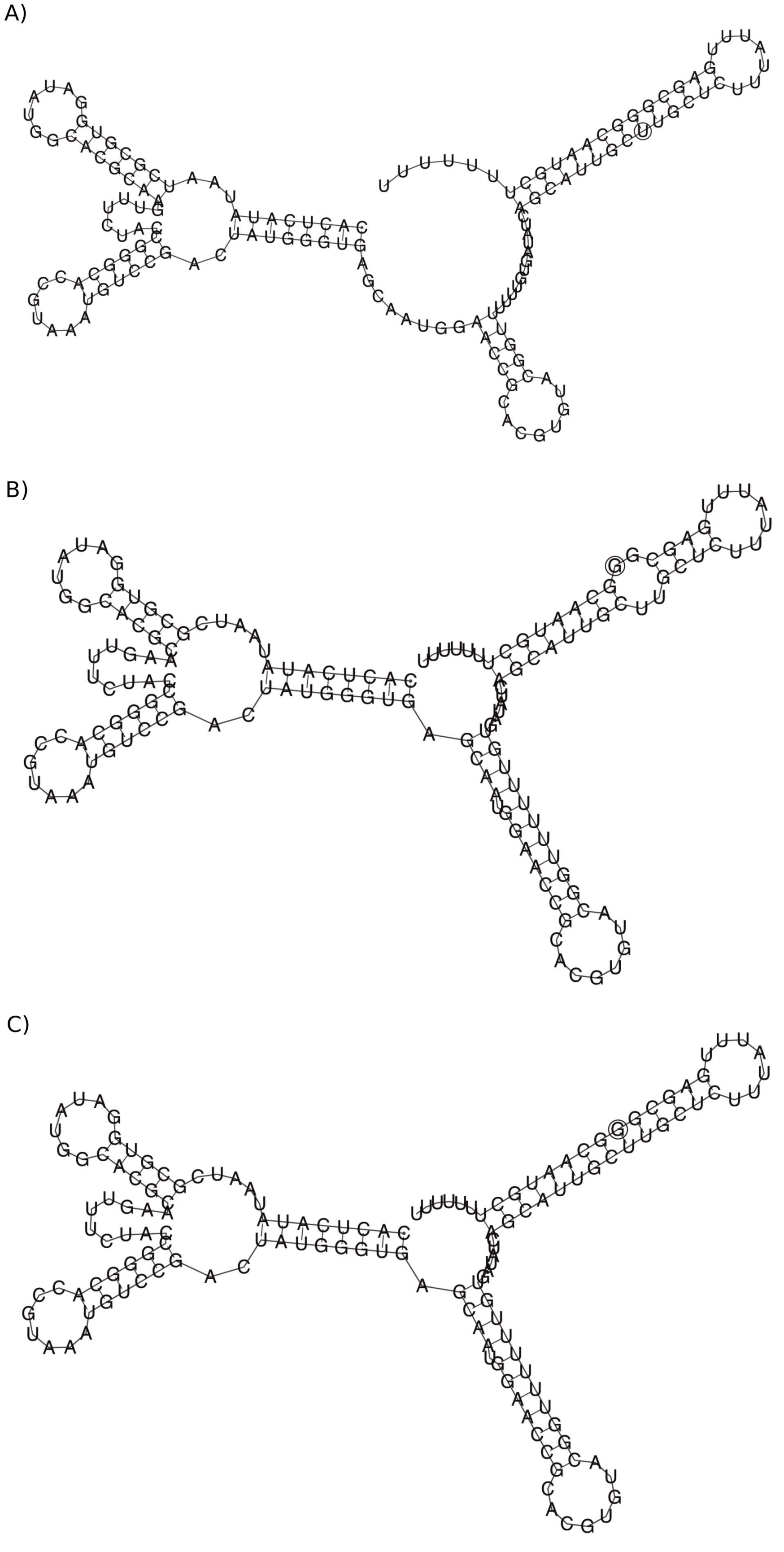

A)
B)
C)

Figure 4. Secondary structure of the XPT guanine riboswitch from *Bacillus subtilis*, with experimentally determined 148 nt sequence CACUCAUAUA AUCGCGUGGA UAUGGCACGC AAGUUUCUAC CGGGCACCGU AAAUGUCCGA CUAUGGGUGA GCAAUGGAAC CGCACGUGUA CGGUUUUUUG UGAUAUCAGC AUUGCUUGCU CUUUAUUUGA GCGGGCAAUG CUUUUUUU taken from Wakeman et al. [58]. *(Left)* Gene off structure, determined by in-line probing – see [59] for X-ray structure of aptamer, which is consistent with the secondary structure. *(Center)* Minimum free energy (MFE) structure, determined by RNAnn, our implementation of the nearest-neighbor energy model. This structure is identical to the MFE structures computed by Vienna RNA Package RNAfold [18], RNAstructure [38], and mfold [9]. *(Right)* Minimum free energy (MFE) structure, determined by RNAenn, our implementation of the extended nearest-neighbor energy model. The only difference with the nearest-neighbor MFE structure lies in two missing GU base pairs (116,134), (117,133) indicated by a circle.
doi:10.1371/journal.pone.0085412.g004

proof of correctness, which was subsequently repaired by Ireland and Kullback [44], who additionally showed that the maximum entropy distribution is *not* the maximum likelihood distribution.

Suppose that $p(x_1,...,x_n)$ is a given joint probability distribution on Ω^n, where Ω is the alphabet $\{A,C,G,U\}$ of RNA nucleotides. Recall that a marginal probability distribution $p_{a_{i_1},\ldots,a_{i_m}} : \Omega^{n-m} \rightarrow [0,1]$ is defined by the projection

$$\sum_{(x_1,\ldots,x_n)\in\Omega^n, x_{i_1}=a_{i_1},\ldots,x_{i_m}=a_{i_m}} p(x_1,...,x_n)$$

Given an integer $n \geq 2$, a value $\epsilon > 0$, and a set of arbitrary marginal probabilities $p_{a_{i_1},\ldots,a_{i_m}}$ the idea is to initialize p to the uniform distribution, then repeatedly update p so that it has the correct currently considered marginal.

Algorithm 1 (Brown's algorithm [42]) Input: *Finite sample space* Ω*, integer* $n \geq 2$, $\epsilon > 0$, *set of arbitrary (target) marginal probabilities* $p_{a_{i_1},\ldots,a_{i_m}}$. Ouput: *Maximum entropy joint probability distribution* $p : \Omega^n \rightarrow [0,1]$ *having given marginals (i.e. within* ϵ *of target marginals).*

Idea:

```
    initialize $p$ to the uniform distribution
    repeat
      for each target marginal $M'$
        compute current marginal $M$ of $p$
        $p = p \cdot \frac{M'}{M}$
    until $p$ has all the desired marginals
```

See Figure 3 for more detailed pseudocode of Brown's algorithm.

Conversion between free energies and probabilities. To compute triplet stacking free energies, base stacking free energies from Turner 1999 [or alternatively Turner 2004] energy model are converted to marginal probabilities in the following manner. Given a triplet stack

$$5' - \mathrm{X}_1\mathrm{X}_2\mathrm{X}_3 - 3'$$

$$3' - \mathrm{Y}_1\mathrm{Y}_2\mathrm{Y}_3 - 5'$$

where the outermost base pair occurs at (i,j), let α denote the outermost base pair (i,j) with nucleotides X_1, Y_1, let β denote the middle base pair $(i+1,j-1)$ with nucleotides X_2, Y_2, and let γ denote the innermost base pair $(i+2,j-2)$ with nucleotides X_3, Y_3. It is a well-known principle, first proved by Jaynes [45] and subsequently exploited in protein threading algorithms [46,47], that a representative database of biomolecular sequences and structures has the property that motif occurrences are Boltzmann distributed – i.e. motif frequencies are of the form $\frac{\exp(-E(\mathrm{motif})/RT)}{Q}$, where the partition function Q is the sum of Boltzmann factors of all motifs. For this reason, we define the *left*, *middle* and *right marginal* probabilities of stacked base pairs by:

$$leftMargProb(\beta,\gamma) = \frac{\sum_\delta \mathrm{stack}(\delta,\beta) + \mathrm{stack}(\beta,\gamma)}{Q}$$

$$midMargProb(\alpha,\gamma) = \frac{\sum_\delta \mathrm{stack}(\alpha,\delta) + \mathrm{stack}(\delta,\gamma)}{Q}$$

$$rightMargProb(\alpha,\beta) = \frac{\sum_\delta \mathrm{stack}(\alpha,\beta) + \mathrm{stack}(\beta,\delta)}{Q}.$$

where δ ranges over the six base pairs GC, CG, AU, UA, GU, UG, stack(α,β) denotes base stacking free energies from the Turner 1999 model [or alternatively Turner 2004 model], and the partition function

$$Q = \sum_{\alpha,\beta,\gamma} \mathrm{stack}(\alpha,\beta) + \mathrm{stack}(\beta,\gamma).$$

In words, the left/middle/right marginal probability is defined by the quotient of the sum over all six base pairs in the left/middle/right position, while fixing the remaining two base pairs, divided by the partition function. We then apply Brown's algorithm to compute the joint probability distribution $P(\alpha,\beta,\delta)$ for all base pairs α,β,δ, and thus obtain triplet stacking free energies

$$E(\alpha,\beta,\gamma) = -RT\ln(Q \cdot P(\alpha,\beta,\gamma)). \qquad (2)$$

An additional potential advantage of the extended nearest neighbor energy model is that the MFE structure is perhaps less likely to have isolated base pairs, than when using base stacking free energies. In particular, Bompfunewerer et al. [48] described an $O(n^3)$ algorithm to compute the MFE structure and partition function over all *canonical* secondary structures; i.e. those having no *isolated base pairs*, where an isolated base pair (i,j) has no adjacent base pair $(i+1,j-1)$ or $(i-1,j+1)$. Bompfunewerer et al. stated that preliminary studies indicated that canonical MFE structure prediction is both faster and more accurate. In [49] we provided theoretical reasons for the computational speed-up, by using complex analysis to prove that the asymptotic number of canonical secondary structures is $2.1614 \cdot n^{-3/2} \cdot 1.96798^n$, compared to the much larger number $1.104366 \cdot n^{-3/2} \cdot 2.618034^n$ of all secondary structures, a result obtained in [50] by a different method.

Apart from the triplet stacking energy, ENN-13 contains free energies for base stacks (used only at stem ends), hairpins, bulges, internal loops and multiloops from the Turner NN model – here, the user may choose between the Turner 1999 parameters and the Turner 2004 parameters, the former taken from Vienna RNA Package 1.8.5 and the latter taken from the Nearest Neighbor

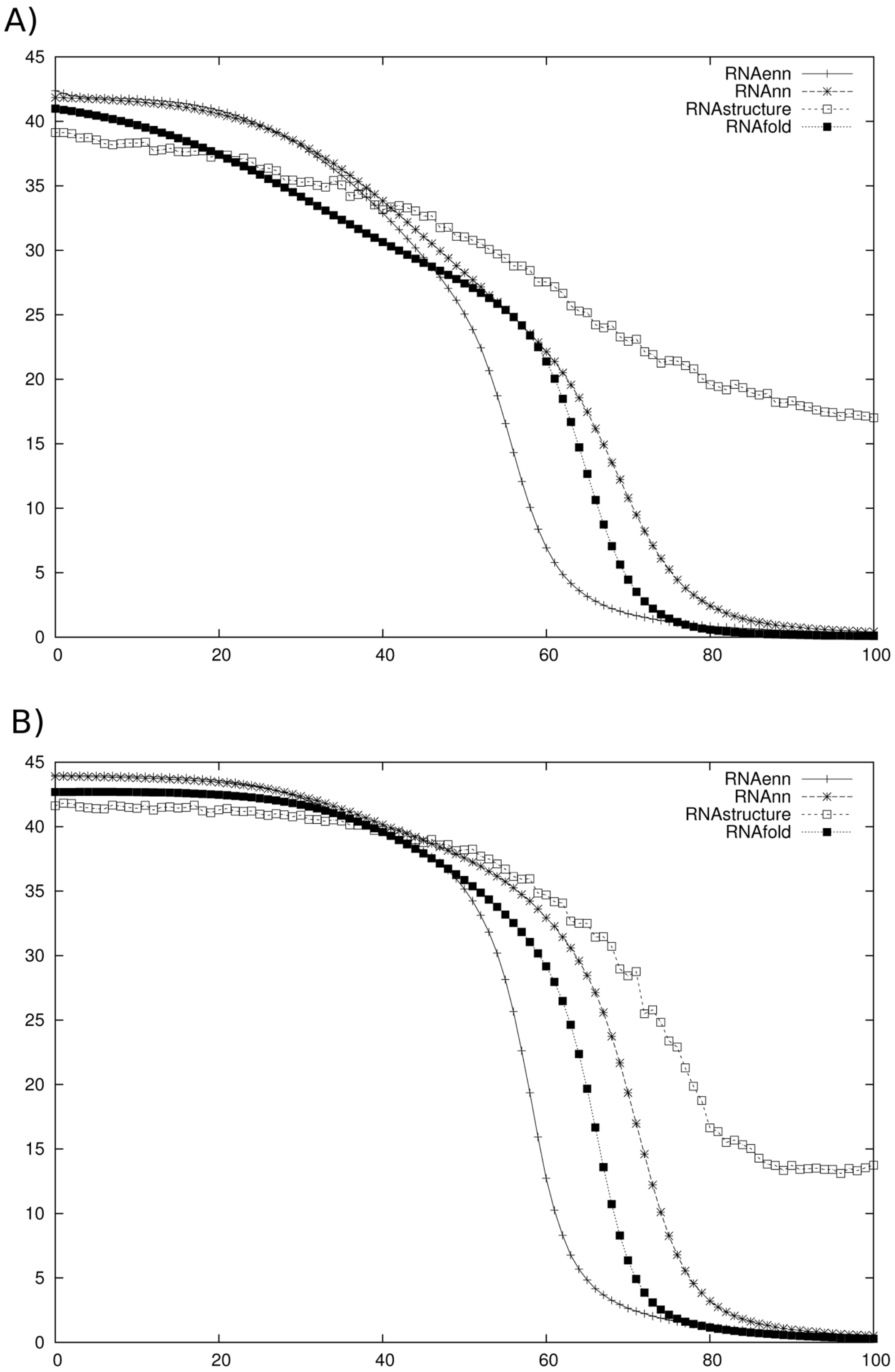


Figure 5. Melting curves for two small nucleolar RNAs (snoRNA) from family RF00158 from Rfam version 9.0 [56]. For each RNA sequence, over a range of temperatures, temperature-dependent base pair probabilities were computed using four different software packages: RNAenn, RNAnn, version 1.8.5 of RNAfold [40] and RNAstructure [38]. The software RNAenn (RNA extended nearest-neighbor) is our implementation of the algorithms described in this paper, while the software RNAnn (RNA nearest-neighbor) is our implementation of the following algorithms: Zuker's minimum free energy structure algorithm [51], McCaskill's partition function algorithm [19], and the Ding-Lawrence sampling algorithm [22].

Each algorithm was run without dangle or coaxial free energies. At each temperature T, for each algorithm, the expected number $\langle BP \rangle$ of base pairs was computed as $\langle BP \rangle = \sum_{1 \le i < j \le n} p_{i,j}$; for each algorithm, the collection of such $(T, \langle BP \rangle)$ points generates a melting profile obtained by that algorithm. *(Left)* Melting curves for the 72 nt small nucleolar RNA (snoRNA) from *Ornithorhynchus anatinus* (platypus) with GenBank accession code AAPN01359272.1/4977–5048 and sequence given by AGCACAAAUG AUGAGCCUAA AGGGACUUAA UACUGAAACC UGAUGUAACU AAAUAAUAUA UGCUGAUCGU GC *(Right)* Melting curves for the 69 nt small nucleolar RNA (snoRNA) from *Otolemur garnetti* (small-eared galago) with GenBank accession code AQR01179445.1/1047–1115 and sequence given by GGCACAAAUG AUGAAUGACA AGGGACUUAA UACUGAAACC UGAUGUUACA UUACAAUGUG CUGAUGUGC.
doi:10.1371/journal.pone.0085412.g005

Data Base (NNDB) http://rna.urmc.rochester.edu/NNDB/ [15]. For readers interested in the exact nature of the NN energy parameters, we recommend the excellent overview by Zuker et al. [58].

Discussion

There are some deviations between the MFE structure computed for the ENN model, compared with the nearest-neighbor (NN) model. In particular, Figure 4 shows the secondary structure for the XPT riboswitch from *Bacillus subtilis*, obtained by experimental in-line probing (left panel), minimum free energy structure computation for the NN model (middle panel) and minimum free energy structure computation for the ENN model (right panel). The MFE structure for the NN model was identical, using four different software packages: mfold [9], RNAfold [40], RNAstructure [38] and our own program RNAnn for the nearest-neighbor model. Our software RNAenn for the ENN model differs from the NN minimum free energy structure, only by missing two GU-wobble base pairs at positions (116,134), (117,133). Adjacent wobble pairs are energetically weak, so we do not view this as a failure of our software, but rather the need for additional scrutiny of the ENN energy parameters. Specifically, in the future, we intend to include a dependence on the heat capacity ΔC_P as proposed by M. Bon [17], and knowledge-based potentials [16,17]. By such energy re-parametrization, we expect to improve the sensitivity values reported in Table 1.

Our next-nearest-neighbor *triplet* energy model appears to lead to somewhat more *cooperative* folding than does the nearest neighbor energy model, as indicated by sharper sigmoidal transition in the melting curves obtained by RNAenn, compared to melting curves obtained by RNAfold and RNAstructure – see Figure 5. Here, melting curves were computed in the following manner. For each RNA sequence, over a range of temperatures, temperature-dependent base pair probabilities were computed. At each temperature T, for each algorithm, the expected number $\langle BP \rangle$ of base pairs was computed by $\langle BP \rangle = \sum_{1 \le i < j \le n} p_{i,j}$. For

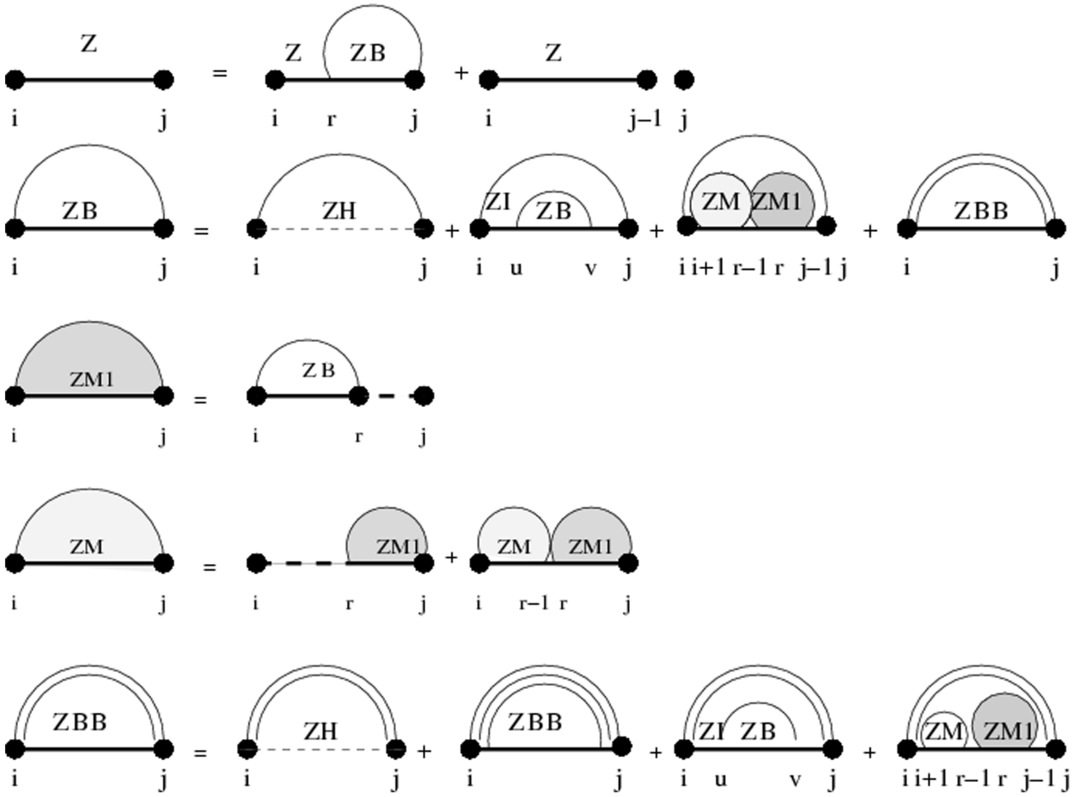


Figure 6. Feynman diagram to pictorially describe recursions described in this proposal for partition function with respect to extended nearest neighbor model. For simplicity, this diagram depicts ZBB, but not $ZBBL$,$ZBBR$, which correspond to a special treatment for particular left/right bulges of size 1, that are treated as stacked base pairs.
doi:10.1371/journal.pone.0085412.g006

each algorithm, the collection of all points with (x,y) coordinates given by $(T,\langle BP\rangle)$ generates a melting profile.

Methods

The top level recursion in the computation of the partition function [resp. minimum free energy structure] is identical to that of McCaskill's algorithm [19] [resp. Zuker's algorithm [51]]. The technical difficulty lies in a kind of "2-look-ahead" strategy, to determine if a base pair (i,j) not only stacks onto the adjacent base pair $(i+1,j-1)$, but the latter also stacks onto the base pair $(i+2,j-2)$. This leads to technical issues, including a special treatment for bulges of size 1, since these are considered to stack on the following base pair.

Subsection 3.1: Partition function algorithm

This section presents the recursions to compute the partition function in the extended nearest-neighbor energy model. Figure 6 depicts the recursions as a Feynman diagram. (For simplicity, the Feynman diagram in Figure 6 depicts ZBB, but not $ZBBL,ZBBR$, which correspond to a special treatment for particular left/right bulges of size 1, that are treated similarly to stacked base pairs.) The unconstrained partition function $Z_{i,j}$ for $a_i,\dots,a_j$ is defined below. Note that the recursions for $Z_{i,j}(\mathbb{I})$ entail a maximum internal loop of size 30, which follows the Vienna convention to reduce run time of the algorithm to $O(n^3)$; however, our implementation actually uses the more complicated treatment of Lyngsø et al. [52], which ensures a cubic run time while not arbitrarily bounding the maximum size of internal loops. A similar remark applies to the treatment of internal loops and bulges of size 1 in Section 3.2.

$$Z_{i,j}=\begin{cases}1 & \text{if } j-i\leq\theta\\ Z_{i,j-1}+ZB_{i,j}+\sum\limits_{k=i+1}^{j-\theta-1} Z_{i,k-1}\cdot ZB_{k,j} & \text{else}\end{cases}$$

We now in turn describe the partition functions $ZM1_{i,j}$ for a multiloop having a single component, $ZM_{i,j}$ for a multiloop having one or more components, $ZB_{i,j}$ where (i,j) pair together, and for $ZBB_{i,j}$ where (i,j) and $(i+1,j-1)$ pair together.

$$ZM1_{i,j}=\begin{cases}0 & \text{if } j-i\leq\theta\\ \sum\limits_{k=i+\theta+1}^{j} \exp\left(-\frac{c\cdot(j-k)}{RT}\right)\cdot ZB_{i,k} & \text{else}\end{cases}$$

$$ZM_{i,j}=$$

$$\begin{cases}0 & \text{if } i\leq j \text{ and } j-i\leq\theta\\ \sum\limits_{k=i}^{j-\theta-1} \exp(-\frac{b+c\cdot(k-i)}{RT})ZM1_{k,j} & \\ \quad & \text{else}\\ +\sum\limits_{k=i}^{j-\theta-2} \exp\left(-\frac{b}{RT}\right)\cdot ZM_{i,k}\cdot ZM1_{k+1,j} & \end{cases}$$

$$ZB_{i,j}=\begin{cases}0 & \text{if } j-i\leq\theta\\ Z_{i,j}(\mathbb{S})+Z_{i,j}(\mathbb{H})+Z_{i,j}(\mathbb{B})+Z_{i,j}(\mathbb{I})+Z_{i,j}(\mathbb{M}) & \text{else}\end{cases}$$

where

$$Z_{i,j}(\mathbb{S})=ZBB_{i,j}$$

$$Z_{i,j}(\mathbb{H})=\exp\left(-\frac{\mathbb{H}(j-i-1,T)}{RT}\right)$$

$$Z_{i,j}(\mathbb{LB})=\exp\left(-\frac{\mathbb{B}(1,T)}{RT}\right)\cdot ZBBL_{i,j}+ \sum_{k=i+3}^{j-\theta-2}\exp\left(-\frac{\mathbb{B}(k-i-1,T)}{RT}\right)\cdot ZB_{k,j-1}$$

$$Z_{i,j}(\mathbb{RB})=\exp\left(-\frac{\mathbb{B}(1,T)}{RT}\right)\cdot ZBBR_{i,j}+ \sum_{k=i+\theta+2}^{j-3}\exp\left(-\frac{\mathbb{B}(j-k-1,T)}{RT}\right)\cdot ZB_{i+1,k}$$

$$Z_{i,j}(\mathbb{I})=\sum_{\ell-i-2\leq 2}^{j-\theta-3}\ \sum_{j-r-1+\ell-i-2\leq 30}^{j-2}\exp\left(-\frac{\mathbb{I}((\ell-i-1)+(j-r-1))}{RT}\right)\cdot ZB_{\ell,r}$$

$$Z_{i,j}(\mathbb{M})=\exp\left(-\frac{a+2b+TMM(i,j,i+1,j-1)}{RT}\right)\cdot \sum_{k=i+\theta+3}^{j-\theta-2} ZM_{i+1,k-1}\cdot ZM1_{k,j-1}$$

$$ZBB_{i,j}=$$

$$\begin{cases}0 & \text{if } j-i\leq\theta+2\\ Q_{i,j}(\mathbb{S})+Q_{i,j}(\mathbb{H})+Q_{i,j}(\mathbb{B})+Q_{i,j}(\mathbb{I})+Q_{i,j}(\mathbb{M}) & \text{else}\end{cases}$$

where

$$Q_{i,j}(\mathbb{S})=\exp\left(-\frac{\mathbb{E}(i,i+1,i+2;j-2,j-1,j)}{RT}\right)\cdot ZBB_{i+1,j-1}$$

$$Q_{i,j}(\mathbb{H})=\exp\left(-\frac{\mathbb{E}(i,i+1;j-1,j)+\mathbb{H}(j-i-3,T)}{RT}\right)$$

$$Q_{i,j}(\mathbb{LB})=\exp\left(-\frac{\mathbb{E}(i,i+1,i+3;j-2,j-1,j)+\mathbb{B}(1,T)}{RT}\right)\cdot ZBBL_{i+1,j-1}+ \sum_{k=i+4}^{j-\theta-3}\exp\left(-\frac{\mathbb{E}(i,i+1;j-1,j)+\mathbb{B}(k-i-2,T)}{RT}\right)\cdot ZB_{k,j-2}$$

$$Q_{i,j}(\mathbb{RB}) = \exp\left(- \frac{\mathbb{E}(i,i+1,i+2;j-3,j-1,j) + \mathbb{B}(1,T)}{RT} \right) \cdot ZBBR_{i+1,j-1} + \sum_{k=i+\theta+3}^{j-3} \exp\left(- \frac{\mathbb{E}(i,i+1;j-1,j) + \mathbb{B}(j-k-2,T)}{RT} \right) \cdot ZB_{i+2,k}$$

$$Q_{i,j}(\mathbb{I}) = \sum_{\ell=i+3}^{j-\theta-4} \sum_{r=\ell+\theta+1}^{j-3} \exp\left(- \frac{\mathbb{E}(i,i+1;j-1,j) + \mathbb{I}((\ell-i-2)+(j-r-2))}{RT} \right) \cdot ZB_{\ell,r}$$

$$Q_{i,j}(\mathbb{M}) = \exp\left(- \frac{\mathbb{E}(i,i+1;j-1,j) + a + 2b}{RT} \right) \cdot \sum_{k=i+\theta+4}^{j-\theta-3} ZM_{i+2,k-1} \cdot ZM1_{k,j-2}$$

$$ZBBL_{i,j} = \begin{cases} 0 & \text{if } j-i \leq \theta+3 \\ QL_{i,j}(\mathbb{S}) + QL_{i,j}(\mathbb{H}) + QL_{i,j}(\mathbb{B}) + QL_{i,j}(\mathbb{I}) + QL_{i,j}(\mathbb{M}) & \text{else} \end{cases}$$

where

$$QL_{i,j}(\mathbb{S}) = \exp\left(- \frac{\mathbb{E}(i,i+2,i+3;j-2,j-1,j)}{RT} \right) \cdot ZBB_{i+2,j-1}$$

$$QL_{i,j}(\mathbb{H}) = \exp\left(- \frac{\mathbb{E}(i,i+2;j-1,j) + \mathbb{H}(j-i-4,T)}{RT} \right)$$

$$QL_{i,j}(\mathbb{LB}) = \exp\left(- \frac{\mathbb{E}(i,i+2,i+4;j-2,j-1,j) + \mathbb{B}(1,T)}{RT} \right) \cdot ZBBL_{i+2,j-1} + \sum_{k=i+5}^{j-\theta-3} \exp\left(- \frac{\mathbb{E}(i,i+2;j-1,j) + \mathbb{B}(k-i-3,T)}{RT} \right) \cdot ZB_{k,j-2}$$

$$QL_{i,j}(\mathbb{RB}) = \exp\left(- \frac{\mathbb{E}(i,i+2,i+3;j-3,j-1,j) + \mathbb{B}(1,T)}{RT} \right) \cdot ZBBR_{i+2,j-1} + \sum_{k=i+\theta+4}^{j-4} \exp\left(- \frac{\mathbb{E}(i,i+2;j-1,j) + \mathbb{B}(j-k-2,T)}{RT} \right) \cdot ZB_{i+3,k}$$

$$QL_{i,j}(\mathbb{I}) = \sum_{\ell=i+4}^{j-\theta-4} \sum_{r=\ell+\theta+1}^{j-3} \exp\left(- \frac{\mathbb{E}(i,i+2;j-1,j) + \mathbb{I}((\ell-i-3)+(j-r-2))}{RT} \right) \cdot ZB_{\ell,r}$$

$$QL_{i,j}(\mathbb{M}) = \exp\left(- \frac{\mathbb{E}(i,i+2;j-1,j) + a + 2b}{RT} \right) \cdot \sum_{k=i+\theta+5}^{j-\theta-3} ZM_{i+3,k-1} \cdot ZM1_{k,j-2}$$

$$ZBBR_{i,j} = \begin{cases} 0 & \text{if } j-i \leq \theta+3 \\ QR_{i,j}(\mathbb{S}) + QR_{i,j}(\mathbb{H}) + QR_{i,j}(\mathbb{B}) + QR_{i,j}(\mathbb{I}) + QR_{i,j}(\mathbb{M}) & \text{else} \end{cases}$$

where

$$QR_{i,j}(\mathbb{S}) = \exp\left(- \frac{\mathbb{E}(i,i+1,i+2;j-3,j-2,j)}{RT} \right) \cdot ZBB_{i+1,j-2}$$

$$QR_{i,j}(\mathbb{H}) = \exp\left(- \frac{\mathbb{E}(i,i+1;j-2,j) + \mathbb{H}(j-i-4,T)}{RT} \right)$$

$$QR_{i,j}(\mathbb{LB}) = \exp\left(- \frac{\mathbb{E}(i,i+1,i+3;j-3,j-2,j) + \mathbb{B}(1,T)}{RT} \right) \cdot ZBBL_{i+1,j-2} + \sum_{k=i+4}^{j-\theta-4} \exp\left(- \frac{\mathbb{E}(i,i+1;j-2,j) + \mathbb{B}(k-i-2,T)}{RT} \right) \cdot ZB_{k,j-3}$$

$$QR_{i,j}(\mathbb{RB}) = \exp\left(- \frac{\mathbb{E}(i,i+1,i+2;j-4,j-2,j) + \mathbb{B}(1,T)}{RT} \right) \cdot ZBBR_{i+1,j-2} + \sum_{k=i+\theta+3}^{j-5} \exp\left(- \frac{\mathbb{E}(i,i+1;j-2,j) + \mathbb{B}(j-k-3,T)}{RT} \right) \cdot ZB_{i+2,k}$$

$$QR_{i,j}(\mathbb{I}) = \sum_{\ell=i+3}^{j-\theta-5} \sum_{r=\ell+\theta+1}^{j-4} \exp\left(- \frac{\mathbb{E}(i,i+1;j-2,j) + \mathbb{I}((\ell-i-2)+(j-r-3))}{RT} \right) \cdot ZB_{\ell,r}$$

$$QR_{i,j}(\mathbb{M})=\exp\left(-\frac{\mathbb{E}(i,i+1;j-2,j)+a+2b}{RT}\right)\cdot \sum_{k=i+\theta+4}^{j-\theta-4} ZM_{i+2,k-1}\cdot ZM1_{k,j-3}$$

Subsection 3.2: Minimum free energy algorithm

Assume that $a_1,\ldots,a_n$ is a given RNA sequence. Throughout this section, we let $E_{i,j}$ denote the minimum free energy of $a_i,\ldots,a_j$, which is computed and stored in arrays by a dynamic programming algorithm corresponding to the following recursions. Once $E_{1,n}$ is computed, then the minimum free energy structure can be computed by tracebacks. The following recursions are obtained from those in the previous section, by systematically replacing sum by minimum, product by sum and Boltzmann factor by energy.

$$E_{i,j}=\begin{cases}0 & \text{if } j-i\leq\theta\\ \min\left\{E_{i,j-1},EB_{i,j},\min\limits_{k=i+1}^{j-\theta-1}E_{i,k-1}+EB_{k,j}\right\} & \text{else}\end{cases}$$

$$1_{i,j}=\begin{cases}+\infty & \text{if } j-i\leq\theta\\ \min\limits_{k=i+\theta+1}^{j}(c\cdot(j-k))+EB_{i,k} & \text{else}\end{cases}$$

$$EM_{i,j}=$$
$$\begin{cases}+\infty & \text{if } i\leq j \text{ and } j-i\leq\theta\\ \min\left\{\min\limits_{k=i}^{j-\theta-1}EM1_{k,j}+b+c\cdot(k-i),\ \min\limits_{k=i}^{j-\theta-2}b+EM_{i,k}+EM1_{k+1,j}\right\} & \text{else}\end{cases}$$

$$EB_{i,j}=\begin{cases}+\infty & \text{if } j-i\leq\theta\\ \min\left\{E_{i,j}(\mathbb{S}),E_{i,j}(\mathbb{H}),E_{i,j}(\mathbb{B}),E_{i,j}(\mathbb{I}),E_{i,j}(\mathbb{M})\right\} & \text{else}\end{cases}$$

where

$$E_{i,j}(\mathbb{S})=EBB_{i,j}$$

$$E_{i,j}(\mathbb{H})=\mathbb{H}(j-i-1,T)$$

$$E_{i,j}(\mathbb{LB})=$$
$$\min\left\{\min\limits_{k=i+3}^{j-\theta-2}\mathbb{B}(k-i-1,T)+EB_{k,j-1},\mathbb{B}(1,T)+EBBL_{i,j}\right\}$$

$$E_{i,j}(\mathbb{RB})=$$
$$\min\left\{\min\limits_{k=i+\theta+2}^{j-3}\mathbb{B}(j-k-1,T)+EB_{i+1,k},\mathbb{B}(1,T)+EBBR_{i,j}\right\}$$

$$E_{i,j}(\mathbb{I})=\min\limits_{\ell=i+2}^{j-\theta-3}\min\limits_{r=\ell+\theta+1}^{j-2}\mathbb{I}((\ell-i-1)+(j-r-1))+EB_{\ell,r}$$

$$E_{i,j}(\mathbb{M})=a+2b+\min\limits_{k=i+3}^{j-\theta-2}\left(EM_{i+1,k-1}+EM1_{k,j-1}\right)$$

$$EBB_{i,j}=$$
$$\begin{cases}+\infty & \text{if } j-i\leq\theta+2\\ \min\left\{G_{i,j}(\mathbb{S})+G_{i,j}(\mathbb{H})+G_{i,j}(\mathbb{B})+G_{i,j}(\mathbb{I})+G_{i,j}(\mathbb{M})\right\} & \text{else}\end{cases}$$

where

$$G_{i,j}(\mathbb{S})=\mathbb{E}(i,i+1,i+2;j-2,j-1,j)+EBB_{i+1,j-1}$$

$$G_{i,j}(\mathbb{H})=\mathbb{E}(i,i+1;j-1,j)+\mathbb{H}(j-i-3,T)$$

$$G_{i,j}(\mathbb{LB})=\min\ \{\min\limits_{k=i+4}^{j-\theta-3}\mathbb{E}(i,i+1;j-1,j)+\mathbb{B}(k-i-2,T)+EB_{k,j-2},$$
$$\mathbb{E}(i,i+1,i+3;j-2,j-1,j)+\mathbb{B}(1,T)+EBBL_{i+1,j-1}\}$$

$$G_{i,j}(\mathbb{RB})=\min\{\min\limits_{k=i+\theta+3}^{j-4}\mathbb{E}(i,i+1;j-1,j)+$$
$$\mathbb{B}(j-k-2,T)+EB_{i+2,k},$$
$$\mathbb{E}(i,i+1,i+2;j-3,j-1,j)+\mathbb{B}(1,T)+EBBR_{i+1,j-1}\}$$

$$G_{i,j}(\mathbb{I})=\min\limits_{\ell=i+3}^{j-\theta-4}\min\limits_{r=\ell+\theta+1}^{j-3}\mathbb{E}(i,i+1;j-1,j)+$$
$$\mathbb{I}((\ell-i-2)+(j-r-2))+EB_{\ell,r}$$

$$G_{i,j}(\mathbb{M})=\mathbb{E}(i,i+1;j-1,j)+a+2b+$$
$$\min\limits_{k=i+4}^{j-\theta-3}\left(EM_{i+2,k-1}+EM1_{k,j-2}\right)$$

$$EBBL_{i,j}=$$
$$\begin{cases}+\infty & \text{if } j-i\leq\theta+3\\ \min\left\{GL_{i,j}(\mathbb{S})+GL_{i,j}(\mathbb{H})+GL_{i,j}(\mathbb{B})+GL_{i,j}(\mathbb{I})+GL_{i,j}(\mathbb{M})\right\} & \text{else}\end{cases}$$

where

$$GL_{i,j}(\mathbb{S})=\mathbb{E}(i,i+2,i+3;j-2,j-1,j)+EBB_{i+2,j-1}$$

$$GL_{i,j}(\mathbb{H})=\mathbb{E}(i,i+2;j-1,j)+\mathbb{H}(j-i-4,T)$$

$$\begin{aligned} GL_{i,j}(\mathbb{LB}) = \min\{ & \min_{k=i+5}^{j-\theta-3} \mathbb{E}(i,i+2;j-1,j) + \\ & \mathbb{B}(k-i-3,T) + EB_{k,j-2}, \\ & \mathbb{E}(i,i+2,i+4;j-2,j-1,j) + \mathbb{B}(1,T) + EBBL_{i+2,j-1}\} \end{aligned}$$

$$\begin{aligned} GL_{i,j}(\mathbb{RB}) = \min\{ & \min_{k=i+\theta+4}^{j-4} \mathbb{E}(i,i+2;j-1,j) + \\ & \mathbb{B}(j-k-2,T) + EB_{i+3,k}, \\ & \mathbb{E}(i,i+2,i+3;j-3,j-1,j) + \mathbb{B}(1,T) + EBBR_{i+2,j-1}\} \end{aligned}$$

$$\begin{aligned} GL_{i,j}(\mathbb{I}) = & \min_{\ell=i+4}^{j-\theta-4} \min_{r=\ell+\theta+1}^{j-3} \mathbb{E}(i,i+2;j-1,j) + \\ & \mathbb{I}((\ell-i-3)+(j-r-2)) + EB_{\ell,r} \end{aligned}$$

$$\begin{aligned} GL_{i,j}(\mathbb{M}) = & \mathbb{E}(i,i+2;j-1,j) + a + 2b + \\ & \min_{k=i+5}^{j-\theta-3} \left(EM_{i+3,k-1} + EM1_{k,j-2} \right) \end{aligned}$$

$$EBBR_{i,j} = \begin{cases} +\infty & \text{if } j-i \leq \theta+3 \\ \min\{GR_{i,j}(\mathbb{S}) + GR_{i,j}(\mathbb{H}) + GR_{i,j}(\mathbb{B}) + GR_{i,j}(\mathbb{I}) + GR_{i,j}(\mathbb{M})\} & \text{else} \end{cases}$$

where

$$GR_{i,j}(\mathbb{S}) = \mathbb{E}(i,i+1,i+2;j-3,j-2,j) + EBB_{i+1,j-2}$$

$$GR_{i,j}(\mathbb{H}) = \mathbb{E}(i,i+1;j-2,j) + \mathbb{H}(j-i-4,T)$$

$$\begin{aligned} GR_{i,j}(\mathbb{LB}) = \min\{ & \min_{k=i+4}^{j-\theta-4} \mathbb{E}(i,i+1;j-2,j) + \\ & \mathbb{B}(k-i-2,T) + EB_{k,j-3}, \\ & \mathbb{E}(i,i+1,i+3;j-3,j-1,j) + \mathbb{B}(1,T) + EBBL_{i+1,j-2}\} \end{aligned}$$

$$\begin{aligned} GR_{i,j}(\mathbb{RB}) = \min\{ & \min_{k=i+\theta+3}^{j-4} \mathbb{E}(i,i+1;j-2,j) + \\ & \mathbb{B}(j-k-3,T) + EB_{i+2,k}, \\ & \mathbb{E}(i,i+1,i+2;j-4,j-2,j) + \mathbb{B}(1,T) + EBBR_{i+1,j-2}\} \end{aligned}$$

$$\begin{aligned} GR_{i,j}(\mathbb{I}) = & \min_{\ell=i+3}^{j-\theta-5} \min_{r=\ell+\theta+1}^{j-4} \mathbb{E}(i,i+1;j-2,j) + \\ & \mathbb{I}((\ell-i-2)+(j-r-3)) + EB_{\ell,r} \end{aligned}$$

$$\begin{aligned} GR_{i,j}(\mathbb{M}) = & \mathbb{E}(i,i+1;j-2,j) + a + 2b + \\ & \min_{k=i+4}^{j-\theta-4} \left(EM_{i+2,k-1} + EM1_{k,j-3} \right) \end{aligned}$$

Conclusion

In this paper, we have introduced a new energy model ENN for RNA secondary structure prediction and implemented it in a tool called RNAenn along with new energy parameters for triplet stacking inferred using Brown's algorithm. RNAenn is implemented in C/C++, without any function calls or dependence on other programs, such as mfold [9], RNAfold [18], and RNAstructure [38]. Recursions from the partition function have been cross-checked by setting free energy parameters to zero, in which case the program returns the number of secondary structures, which can be determined by independent simpler methods.

It is known from experimental work of Silverman and Cech on Tetrahymena group I intron P4–P6 domain [53] that RNA folds cooperatively. The melting curves in Figure 5 demonstrate that our ENN model leads to somewhat more *cooperative* folding than does the nearest neighbor energy model, in the same manner that the melting curves of Figure 1 demonstrate that the nearest neighbor energy model leads to more cooperative folding than the simple Nussinov energy model. For this reason, we feel that RNAenn supports a mathematical model that better reflects the experimental data concerning cooperativity of RNA folding.

From the benchmarking comparison in Table 1, it is clear that triplet stacking free energy parameters need further refinement to produce better agreement with RNA secondary structures, as determined by comparative sequence alignment or X-ray structure. This situation is not unlike the situation with nearest neighbor software mfold, RNAfold, which over the years underwent a series of refinements, with the introduction of additional energy parameters (energy parameters for particular triloops, tetraloops, bulges of size one, etc.). At the present time, software such as Unafold, RNAfold, RNAstructure remain state-of-the-art for RNA secondary structure prediction. However, in future work, we plan to optimize the triplet stacking energy parameters, by using knowledge-base potential as in the work [16,17] for the nearest neighbor model.

Acknowledgments

We would like to thank D.M. Gray for providing us references on his work on the next-nearest-neighbor triplet model [11,12,33], P.F. Stadler for discussions about the the possibility of using microarray data to infer RNA base stacking and triplet stacking free energies [4], and C. Burge for introducing us to Brown's algorithm. Thanks as well to D.H. Mathews for use of his RNA secondary structure database, to V. Ilyin for some assistance in the implementation of Brown's algorithm to obtain triplet stacking energies, and to Hosna Jabbari and Andy Lorenz for preparation of the Figure 2. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Author Contributions

Conceived and designed the experiments: PC. Performed the experiments: ID VM PC. Analyzed the data: ID VM. Contributed reagents/materials/analysis tools: ID VM PC. Wrote the paper: PC ID. Project concept, design of pseudocode, implementation of treatment for internal loops: PC. Implementation of partition function, minimum free energy computation and sampling algorithms: ID. Implementation of all energy models, in particular the NN and ENN energy paramters for multiple energy models, as well as auxiliary implementation: VM.

References

1. Lim L, Glasner M, Yekta S, Burge C, Bartel D (2003) Vertebrate microRNA genes. Science 299(5612): 1540.
2. Rajewsky N (2006) microrna target predictions in animals. Nat Genet 38: S8–S13.
3. Washietl S, Hofacker I, Stadler P (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A 102: 2454–2459.
4. Waldminghaus T, Kortmann J, Gesing S, Narberhaus F (2008) Generation of synthetic RNA-based thermosensors. Biol Chem 389: 1319–1326.
5. Kryukov GV, Kryukov VM, Gladyshev VN (1999) New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. J Biol Chem 274: 33888–33897.
6. Bekaert M, Bidou L, Denise A, Duchateau-Nguyen G, Forest J, et al. (2003) Towards a computational model for −1 eukaryotic frameshifting sites. Bioinformatics 19: 327–335.
7. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput Biol 6:e1000832.
8. Vashishta A, Ohri SS, Proctor M, Fusek M, Vetvicka V (2007) Ribozyme-targeting procathepsin D and its effect on invasion and growth of breast cancer cells: an implication in breast cancer therapy. Int J Oncol 30: 1223–1230.
9. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31(13): 3406–3415.
10. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, et al. (1986) Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci USA 83: 9373–9377.
11. Gray DM (1997) Derivation of nearest-neighbor properties from data on nucleic acid oligomers. I. Simple sets of independent sequences and the influence of absent nearest neighbors. Biopolymers 42: 783–793.
12. Gray DM (1997) Derivation of nearest-neighbor properties from data on nucleic acid oligomers. II. Thermodynamic parameters of DNA.RNA hybrids and DNA duplexes. Biopolymers 42: 795–810.
13. Xia T, SantaLucia J, Burkard M, Kierzek 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 37: 14719–35.
14. 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.
15. Turner DH, Mathews DH (2009) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res 0: O.
16. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP (2007) Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics 23: i19–i28.
17. Bon M (2009) Prédiction de structures secondaires d'ARN avec pseudo-noeuds. Ph.D. thesis, Ecole Polytechnique. Ph.D. dissertation in Physics.
18. Hofacker I (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31: 3429–3431.
19. McCaskill J (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29: 1105–1119.
20. Kiryu H, Kin T, Asai K (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. Bioinformatics 23: 434–441.
21. Reeder J, Giegerich R (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics 5: 104.
22. Ding Y, Chan CY, Lawrence CE (2004) Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res 32: 0.
23. Mathews D, Turner D (2002) Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317: 191–203.
24. Mathews D, Sabina J, Zuker M, Turner H (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J Mol Biol 288: 911–940.
25. Mikulecky PJ, Feig AL (2004) Heat capacity changes in RNA folding: application of perturbation theory to hammerhead ribozyme cold denaturation. Nucleic Acids Res 32: 3967–3976.
26. Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 26: 148–153.
27. Dill K, Bromberg S (2002) Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology. Garland Publishing Inc. 704 pages.
28. Zimm B, Bragg J (1959) Theory of the phase transition between helix and random coil in polypeptide chains. J Chem Phys 31: 526–531.
29. Ising E (1925) Beitrag zur theorie des ferromagnetisumus. Z Phys 31: 253–258.
30. Baumgärtner A, Binder K (1979) Dynamics of the helixcoil transition in biopolymers. J Chem Phys 70: 429–437.
31. 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.
32. Matthews D, Sabina J, Zuker M, Turner D (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288: 911–940.
33. Gray DM, Gray CW, Yoo BH, Lou TF (2010) Antisense DNA parameters derived from nextnearest- neighbor analysis of experimental data. BMC Bioinformatics 11: 252.
34. Najafabadi HS, Goodarzi H, Torabi N, Banihosseini SS (2006) Applying a neural network to predict the thermodynamic parameters for an expanded nearest-neighbor model. J theor Biol 238: 657–665.
35. Chiu WL, Sze CN, Ma NT, Chiu LF, Leung CW, et al. (2003) NTDB: Thermodynamic Database for Nucleic Acids, Version 2.0. Nucleic Acids Res 31: 483–485.
36. Binder H, Kirsten T, Hofacker I, Stadler P, Löffler M (2004) Interactions in oligonucleotide hybrid duplexes on microarrays. J Phys Chem B 108: 18015–18025.
37. Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, et al. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry 34: 11211–11216.
38. Reuter J, Mathews D (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11: 129.
39. Markham N, Zuker M (2005) DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res 33: W577–81.
40. Hofacker I (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31: 3429–3431.
41. Zuker M, Mathews DH, Turner DH (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In: Barciszewski J, Clark B, editors, RNA Biochemistry and Biotechnology, Kluwer Academic Publishers, NATO ASI Series. pp. 11–43.
42. Brown D (1959) A note on approximations to discrete probability distributions. Information and Control 2: 386–392.
43. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78–94.
44. Ireland C, Kullback S (1968) Contingency tables with given marginals. Biometrika 55(1): 179–188.
45. Jaynes ET (1957) Information theory and statistical mechanics. Physical Review 106: 620–630.
46. Sippl M (1990) Calculation of conformation ensembles from potentials of mean force. J Mol Biol 213: 859–883.
47. Kihara D, Lu H, Kolinski A, Skolnick J (2001) TOUCHSTONE: An ab initio protein structure prediction method that uses threading-bases tertiary restraints. Proc Natl Acad Sci USA 98(18): 10125–10130.
48. Bompfunewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, et al. (2008) Variations on RNA folding and alignment: lessons from Benasque. J Math Biol 56: 129–144.
49. Clote P, Kranakis E, Krizanc D, Salvy B (2009) Asymptotics of canonical and saturated RNA secondary structures. J Bioinform Comput Biol 7: 869–893.
50. Hofacker IL, Schuster P, Stadler PF (1998) Combinatorics of RNA secondary structures. Discr Appl Math 88: 207–237.
51. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9: 133–148.
52. Lyngsø RB, Zuker M, Pedersen CN (1999) Fast evaluation of internal loops in RNA secondary structure prediction. Bioinformatics 15: 440–445.
53. Silverman SK, Cech TR (1999) Energetics and cooperativity of tertiary hydrogen bonds in RNA structure. Biochemistry 38: 8691–8702.
54. Gutell R, Lee J, Cannone J (2005) The accuracy of ribosomal RNA comparative structure models. Current Opinion in Structural Biology 12: 301–310.
55. Mathews D (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. Bioinformatics 15: 2246–2253.
56. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, et al. (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37: D136–D140.
57. Waldispuhl 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: 190–215.
58. Wakeman CA, Winkler WC, Dann C (2007) Structural features of metabolite-sensing riboswitches. Trends Biochem Sci 32: 415–424.
59. Serganov A, Keiper S, Malinina L, Tereshko V, Skripkin E, et al. (2005) Structural basis for Diels-Alder ribozyme-catalyzed carbon-carbon bond formation. Nat Struct Mol Biol 12: 218–224.