Journal of Computational Biology: http://mc.manuscriptcentral.com/liebert/jcb

Computing the probability of RNA hairpin and multiloop formation

Journal:	*Journal of Computational Biology*
Manuscript ID:	Draft
Manuscript Type:	Original Paper
Keyword:	algorithms, computational molecular biology, MACHINE LEARNING, secondary structure, RNA
Abstract:	We describe four novel algorithms, RNAhairpin, RNAmloopNum, RNAmloopOrder, RNAmloopHP, which compute the Boltzmann partition function for global structural constraints – respectively for the number of hairpins, the number of multiloops, maximum order (or depth) of multiloops, and the simultaneous number of hairpins and of multiloops. Given an RNA sequence of length n and a user-specified integer 0 ¶ K ¶ n, RNAhairpin [resp. RNAmloopNum resp. RNAmloopOrder] computes the partition functions Z(k) for each 0 ¶ k ¶ K in time O(K2n3) and space O(Kn2), while RNAmloopHP computes the partition functions Z(m, h) for 0 ¶ mm ¶ M multiloops and 0 ¶ h ¶ H hairpins, with run time O(M2H2n3) and space O(MHn2). In addition, programs RNAhairpin [resp. RNAmloopHP] sample from the low energy ensemble of structures having h hairpins [resp. m multiloops and h hairpins], for given h,m. Moreover, by using the fast Fourier transform (FFT), RNAhairpin and RNAmloopNum have been improved to run in time O(n4) and space O(n2), although this improvement is not possible for RNAmloopOrder. We present two applications of the novel algorithms. First, we show that for many Rfam families of RNA, structures sampled from RNAmloopHP are more accurate than the minimum free energy structure; for instance, sensitivity improves by almost 24% for transfer RNA, while for certain ribozyme families, there is an improvement of around 5%.

For Peer Review Only/Not for Distribution

	Second, we show that the probabilities p(k) = Z(k)/Z of forming k hairpins [resp. multiloops] provide discriminating novel features for a support vector machine or relevance vector machine binary classifier for Rfam families of RNA. Our data suggests that multiloop order does not provide any significant discriminatory power over that of hairpin and multiloop number, and since these probabilities can be efficiently computed using the FFT, hairpin and multiloop formation probabilities could be added to other features in existent noncoding RNA gene finders. Our programs, written in C/C++, are publicly available at http://bioinformatics.bc.edu/clotelab/RNAparametric.

SCHOLARONE™
Manuscripts

For Peer Review Only. Not for Distribution

Computing the probability of RNA hairpin and multiloop formation

Y. Ding[1], W.A. Lorenz[3], I. Dotu[3], E. Senter[3], P. Clote[3]*

[1]Department of Biology, University of Pennsylvania, 219 Lynch Labs, 433 S. University Ave., Philadelphia, PA 19104.
[2]Department of Mathematics and Computer Science, Denison University, Granville, OH 43023-0810.
[3]Dept. of Biology, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467.

Abstract

We describe four novel algorithms, `RNAhairpin`, `RNAmloopNum`, `RNAmloopOrder`, `RNAmloopHP`, which compute the Boltzmann partition function for global *structural constraints* – respectively for the number of hairpins, the number of multiloops, maximum order (or depth) of multiloops, and the *simultaneous* number of hairpins and of multiloops. Given an RNA sequence of length n and a user-specified integer $0 \leq K \leq n$, `RNAhairpin` [resp. `RNAmloopNum` resp. `RNAmloopOrder`] computes the partition functions $Z(k)$ for each $0 \leq k \leq K$ in time $O(K^2n^3)$ and space $O(Kn^2)$, while `RNAmloopHP` computes the partition functions $Z(m,h)$ for $0 \leq mm \leq M$ multiloops and $0 \leq h \leq H$ hairpins, with run time $O(M^2H^2n^3)$ and space $O(MHn^2)$. In addition, programs `RNAhairpin` [resp. `RNAmloopHP`] sample from the low energy ensemble of structures having h hairpins [resp. m multiloops and h hairpins], for given h, m. Moreover, by using the fast Fourier transform (FFT), `RNAhairpin` and `RNAmloopNum` have been improved to run in time $O(n^4)$ and space $O(n^2)$, although this improvement is not possible for `RNAmloopOrder`.

We present two applications of the novel algorithms. First, we show that for many Rfam families of RNA, structures sampled from `RNAmloopHP` are more accurate than the minimum free energy structure; for instance, sensitivity improves by almost 24% for transfer RNA, while for certain ribozyme families, there is an improvement of around 5%. Second, we show that the probabilities $p(k) = Z(k)/Z$ of forming k hairpins [resp. multiloops] provide discriminating novel features for a support vector machine or relevance vector machine binary classifier for Rfam families of RNA. Our data suggests that multiloop order does not provide any significant discriminatory power over that of hairpin and multiloop number, and since these probabilities can be efficiently computed using the FFT, hairpin and multiloop formation probabilities could be added to other features in existent noncoding RNA gene finders. Our programs, written in C/C++, are publicly available at `http://bioinformatics.bc.edu/clotelab/RNAparametric`.

1 Introduction

It has recently emerged that RNA plays surprising and previously unsuspected roles in many biological processes, including retranslation of the genetic code (selenocysteine insertion (Böck et al., 1991), ribosomal frameshift (Bekaert et al., 2003)), gene regulation by allostery (riboswitches) (Mandal et al., 2003) and by the RISC complex (microRNAs) (Lim et al., 2003), regulation of heat shock protein expression by temperature sensitive conformational switches (Chowdhury et al., 2003; Tucker and Breaker, 2005), pointwise editing of messenger RNA (guide RNA) (von Haeseler

*to whom correspondence should be addressed

For Peer Review Only/Not for Distribution

et al., 1992), chemical modification of specific nucleotides in the ribosome (small nucleolar RNAs) (Omer et al., 2000), regulation of alternative splicing (Cheah et al., 2007), regulation of chromatin remodeling (small interfering RNAs) (Cam et al., 2009) etc. RNA can encode genomic information (e.g. HIV and hepatitis C) and with no assistance from proteins can catalyze reactions such as peptidytransferase (at ribosomal P-site) (Weinger et al., 2004) and cleavage of RNA phosphodiester bonds at specific sites (group I introns) (Vicens and Cech, 2006). Since RNA plays various unsuspected regulatory and catalytic roles, and since it is known from the ENCODE consortium report (Project Consortium, 2007) that the human genome is "pervasively transcribed", most of whose RNA transcripts have completely unknown structure and function, it is clear that the development of noncoding RNA gene finders remains an important open problem, despite significant advances with tools such as `RNAz` (Gruber et al., 2007), `FOLDALIGN` (Havgaard et al., 2005), etc. The current paper provides novel computable features that could prove useful in enriching features sets for noncoding RNA gene finders.

In this paper, we present four novel thermodynamics-based algorithms to compute parametric structural aspects of the Boltzmann ensemble of low energy structures for a given RNA sequence. Specifically, given an RNA sequence $\mathbf{s} = a_1, \ldots, a_n$ and optionally an upper bound K, `RNAhairpin` computes, for each value of parameter k, for $0 \leq k \leq K \leq n$, the Boltzmann partition function $Z^h(k)$ and Boltzmann probability $p_h(k) = Z^h(k)/Z$ of all structures of $\mathbf{s}$ having exactly k hairpins. Here $Z^h(k)$ designates the sum of Boltzmann factors $\exp(-E(S)/RT)$, taken over all secondary structures S of $\mathbf{s}$ that have exactly k hairpins; the partition function Z denotes the sum of all Boltzmann factors, where the sum is taken over all secondary structures of $\mathbf{s}$. Analogously, `RNAmloopNum` computes, for each value of $0 \leq k \leq K \leq n$, the Boltzmann partition function $Z^m(k)$ and probability $p_m(k) = Z^m(k)/Z$ of all structures, that have exactly k multiloops. The program `RNAmloopOrder` computes the Boltzmann partition function $Z^d(k)$ and probability $p_d(k) = Z^d(k)/Z$ of all structures, having multiloops of order k but none of larger order, where *multiloop order* is is the maximum depth of multiloop nesting. (See Section 6 for formal definition.) Finally, `RNAmloopHP` simultaneously computes the Boltzmann partition function $Z(m, h)$ and probability $p(m, h) = Z(m, h)/Z$ of all structures, having m multiloops and h hairpins. Since our preliminary work showed that structures sampled from `RNAhairpin` could improve structure prediction for certain Rfam families, the program `RNAmloopHP` also supports sampling.

Other groups have shown an interest in global structural features of RNA families. Here we cite four specific examples. First, Hofacker et al. (Hofacker et al., 1998) determined the asymptotic number of hairpins, multiloops, and other structural features for random RNA, using the homopolymer model first introduced by Stein and Waterman (Stein and Waterman, 1978). Second, Giegerich et al. (Steffen et al., 2006) developed the program `RNAshapes`, which computes the minimum free energy structure for various *shapes*; for instance, the cloverleaf shape of tRNA is `[[][][]]`. Third, the RNA STRAND database (Andronescu et al., 2008) consists of 4666 RNA secondary structures collected from other databases, including the Nucleic Acid Database (HM et al., 2003), the Protein Data Bank (Berman et al., 2002), Sprinzl's tRNA database (Sprinzl et al., 1998), Gutell's database (Gutell et al., 2005), etc. RNA STRAND provides frequency analysis for sequence length, number of stems, hairpin loops, bulges, internal loops, multiloops, pseudoknots, etc., which can be generated for a class of RNAs selected by the user from a set of predefined RNA classes, such as 16S ribosomal RNA, 23S ribosomal RNA, 5S ribosomal RNA, 7SK RNA, ciliate telomerase RNA, cis-regulatory element, group I intron, etc. Fourth, Kazan et al. (Kazan et al., 2010) presented a machine learning algorithm `RNAcontext`, which used sequence profiles (sequence LOGOS) as well as local secondary structure profiles (structure LOGOS) to predict RNA nucleotides that bind to a particular riboprotein. Here, a structural profile computes the frequency, for each k in the putative binding region, that nucleotide position k is located in a hairpin, bulge/internal loop, multiloop, or

For Peer Review Only/Not for Distribution

base pair (frequencies are obtained by counting instances from Sfold samples).

Additionally, a number of groups have developed algorithms to compute the minimum free energy structure and partition function by integrating base pairing constraints. These constraints may be *hard*, in the sense that certain nucleotides are required to pair with certain other nucleotides, while other nucleotides may be required to be unpaired. Alternatively, constraints may be *soft*, in the sense that certain nucleotides are more likely to be paired or unpaired. Since chemical and enzymatic probing data (SHAPE, in-line probing, PARS) is not binary 0/1, such soft constraints permit a better mathematical integration of such footprinting data in structure prediction. For instance, the methods of Deigan et al. (Deigan et al., 2009) and Zarringhalam et al. (Zarringhalam et al., 2012) obtain accuracies of $96 - 100\%$ for RNA structure prediction of moderate size. See the papers of Mathews et al. (Mathews et al., 2004), Deigan et al. (Deigan et al., 2009), Zarringhalam et al. (Zarringhalam et al., 2012), and Washietl et al. (Washietl et al., 2012).

Our contribution in this paper is to extend such constrained structure prediction to more global features, such as requiring secondary structures to have a certain number of hairpins, a certain number of multiloops and multiloops of a certain maximum order. In addition to computing the number of structures having k hairpins and the partition function $Z^h(k)$ for each $0 \leq k \leq K \leq n$, the program RNAhairpin can additionally sample a user-specified number of low energy structures having a user-specified number of hairpins. Similarly, the program RNAmloopHP samples low energy structures simultaneously having m multiloops and h hairpins, for user-specified values of m, h. In future work, we hope to extend RNAmloopNum and RNAmloopOrder to sample low energy structures having a user-specified number or order of multiloops, and to extend all algorithms to compute parametric minimum free energy structures – for instance, in the case of RNAmloopHP, to compute the minimum free energy structure over all structures having m multiloops and h hairpins.

The following is the plan of the paper. Section 2 introduces standard definitions and notation to be used. Since our algorithms derive from McCaskill's algorithm (McCaskill, 1990) to compute the partition function $Z = \sum_S \exp(-E(S)/RT)$, for the benefit of the reader, we present that algorithm in Section 3. Sections 4, 5, and 6 respectively describe the algorithms to compute the partition functions $Z^h(k)$, $Z^m(k)$, $Z^d(k)$ for formation of k hairpins k multiloops and (maximum) order k multiloops, for all k. Section 8 describes two applications of the new algorithms, and Section 9 presents a discussion and conclusion of the paper. In the Appendix, we describe how the fast Fourier transform is used to speed up the computations of RNAhairpin and RNAmloopNum.

2 Basic definitions

In this section, we introduce some notation and definitions used in the description of the algorithms RNAhairpin, RNAmloopNum and RNAmloopOrder. Let $a = a_1, \ldots, a_n$ be an arbitrary RNA sequence, and let $a[i, j]$ denote the subsequence $a_i, \ldots, a_j$. Given an RNA sequence $a = a_1, \ldots, a_n$, a secondary structure is a set of ordered pairs corresponding to base pair positions, which satisfies the following requirements.

1. *Watson-Crick or GU wobble pairs:* If (i, j) belongs to S, then pair (a_i, a_j) must be one of the following canonical base pairs: (A, U), (U, A), (G, C), (C, G), (G, U), (U, G).

2. *Threshold requirement:* If (i, j) belongs to S, then $j - i > \theta$.

3. *Nonexistence of pseudoknots:* If (i, j) and (k, ℓ) belong to S, then it is not the case that $i < k < j < \ell$.

For Peer Review Only/Not for Distribution

4. *No base triples:* If (i, j) and (i, k) belong to S, then $j = k$; if (i, j) and (k, j) belong to S, then $i = k$.

Following standard convention, the threshold θ, or minimum number of unpaired bases in a hairpin loop, is taken to be 3.

Secondary structures are often portrayed in *dot bracket notation*, consisting of a balanced parenthesis expression with dots. Positions i, j occupied by left and right parenthesis correspond to the base pair (i, j), while positions occupied by a dot correspond to an unpaired position i. The dot bracket notation for the minimum free energy (MFE) structure for the selenocysteine insertion element `fdhA` is

```
CGCCACCCUGCGAACCCAAUAAUAAAAUAUACAAGGGAGCAAGGUGGCG
(((((((.(((...(((.................))).))).)))))))
```

with free energy -20.53 kcal/mol. A *pseudoknot* (not considered in our software `RNAhairpin`, `RNAmloopNum`), and `RNAmloopOrder` consists of two *unnested* base pairs, (i, j), (k, ℓ), where $i < k < j < \ell$.

In defining multiloops below, we will have recourse to the notion of *component*, defined as follows. For $1 \leq i \leq \ell \leq r \leq j \leq n$, the base pair (ℓ, r) is an *exterior* base pair in the interval $[i, j]$, if there is no base pair (ℓ', r') with $i \leq \ell' < \ell < r < r' \leq j$. When the interval $i = 1$ and $j = n$, then we drop mention of the interval $[i, j]$ and simply speak of *exterior* base pair. If S is a secondary structure on RNA sequence $a_1, \ldots, a_n$ and $1 \leq i \leq j \leq n$, then the number of exterior base pairs in the interval $[i, j]$ is said to be the number of components of S in $[i, j]$.

Free energy parameters

Following the pioneering work of I.Tinoco, Jr., Freier et al. (Freier et al., 1986) measured the free energy and enthalpy of numerous RNA hybridization duplexes, such as $5'$-`GAACGUUC`-$3'$ with its reverse complement. By least-squares fitting, *base stacking* free energies were determined. By similar methods, the Turner Lab (Matthews et al., 1999; Xia et al., 1999) has extended and refined base stacking free energies, loop free energies for hairpins, bulges, internal loops, multiloops, and *dangles*, which latter are stacked single-stranded nucleotides adjacent to a *canonical* $5'$ or $3'$ base pair. In this paper, we use the energy parameters from the Turner 1999 model (Matthews et al., 1999; Xia et al., 1999) as implemented in Vienna RNA Package 1.8.5 (Hofacker, 2003), except that we do not consider dangles. In future work, we plan to extend to the algorithms to the Turner 2004 energy model with dangles (Turner and Mathews, 2009).

3 McCaskill's partition function

Since our work extends McCaskill's algorithm (McCaskill, 1990), for the paper to be self-contained, we give a brief presentation of McCaskill's algorithm. This presentation follows the very lucid account given by Bompfunewerer et al. in (Bompfunewerer et al., 2008).

Given RNA nucleotide sequence $a_1, \ldots, a_n$, we will use the standard notation $\mathcal{H}$ to denote the free energy of a hairpin, $\mathcal{I}$ to denote the free energy of an internal loop (combining the cases of stacked base pair, bulge and proper internal loop), while the free energy for a multiloop containing N_b base pairs and N_u unpaired bases is given by the affine approximation $a + bN_b + cN_u$.

For RNA sequence $a_1, \ldots, a_n$, for all $1 \leq i \leq j \leq n$, 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 $a[i, j]$, $E(S)$

For Peer Review Only/Not for Distribution

is the free energy of secondary structure S, R is the universal gas constant with value $R = 1.987$ cal/mol^{-1} K^{-1}, and T is absolute temperature.

Definition 1 (McCaskill's partition function)

- $Z(i,j)$: *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).
- $ZM(i,j)$: *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* 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* $i < r \leq j$.

With this, we have the unconstrained partition function

$$Z(i,j) \quad = \quad Z(i,j-1) + \sum_{r=i}^{j-\theta-1} Z(i,r-1)\cdot ZB(r,j). \tag{1}$$

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

$$\begin{aligned} ZB(i,j) \quad = \quad & e^{-\mathcal{H}(i,j)/RT} + \sum_{i\leq k\leq \ell \leq j} e^{-\mathcal{I}(i,j;k,\ell)/RT}\cdot ZB(k,\ell) + \\ & e^{-(a+b)/RT}\cdot \left(\sum_{r=i+1}^{j-\theta-2} ZM(i+1,r-1)\cdot ZM1(r,j-1)\right). \end{aligned} \tag{2}$$

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

$$Z^{M1}(i,j) \quad = \quad \sum_{r=i+\theta+1}^{j} Z^{B}(i,r)\cdot e^{-c(j-r)/RT}. \tag{3}$$

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

$$\begin{aligned} ZM(i,j) \quad = \quad & \sum_{r=i}^{j-\theta-1} ZM1(r,j)\cdot e^{-(b+c(r-i))/RT} + \\ & \sum_{r=i+\theta+2}^{j-\theta-1} ZM(i,r-1)\cdot ZM1(r,j)\cdot e^{-b/RT} \end{aligned} \tag{4}$$

See Figure 1 for a pictorial representation of the recursions of McCaskill's (original) algorithm (McCaskill, 1990); note that the recursions are are not quite the same as those given in (Hofacker et al., 1994). We now turn to our parametric versions of the partition function.
================ FIGURE 1 GOES ABOUT HERE ===================

For Peer Review Only/Not for Distribution

4 Hairpins

We begin by defining some abbreviations for the partition function for hairpins

$$ZH(i,j) = \begin{cases} 0 & \text{if } j-i \le \theta \\ e^{-\mathcal{H}(i,j)/RT} & \text{else} \end{cases}$$

and internal loops having h hairpins

$$ZI^h(i,j) = \sum_{i \le k \le \ell \le j} e^{-\mathcal{I}(i,j;k,l)/RT} \cdot ZB^h(k,\ell)$$

where the sum is over k, ℓ such that $(k-i)+(j-\ell) > 0$. This combines the treatment of both left and right bulges with proper internal loops.

For $h \ge 0$, define the base cases $Z^h(i,i) = 1$, $ZB^h(i,i) = ZM^h(i,i) = ZM1^h(i,i) = 0$. The unconstrained partition function for secondary structures restricted to the interval $[i,j]$ that contain h hairpins is given by

$$Z^h(i,j) = \begin{cases} 1 & \text{if } h = 0 \\ Z^h(i,j-1) + \sum_{r=i}^{j-\theta-1} \sum_{h_0+h_1=h} Z^{h_0}(i,r-1) ZB^{h_1}(r,j) & \text{if } h > 0. \end{cases}$$

The partition function for secondary structures restricted to the interval $[i,j]$ that contain h hairpins and are closed by the base pair (i,j) is given by $ZB^h(i,j) = 0$, if $h = 0$; $ZB^h(i,j) = ZH(i,j) + ZI^h(i,j)$, if $h = 1$; and for $h \ge 2$ by

$$ZB^h(i,j) = ZI^h(i,j) + \sum_{r=i+\theta+3}^{j-\theta-2} \sum_{k=1}^{h-1} ZM^k(i+1,r-1) \cdot ZM1^{h-k}(r,j-1) \cdot e^{-(a+b)/RT}.$$

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

$$ZM1^h(i,j) \quad = \quad \sum_{r=i+\theta+1}^{j} ZB^h(i,r) \cdot e^{-c(j-r)/RT}. \tag{5}$$

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

$$\begin{aligned} ZM^h(i,j) \quad = \quad & \sum_{r=i}^{j-\theta-1} ZM1^h(r,j) \cdot e^{-(b+c(r-i))/RT} + \\ & \sum_{r=i+\theta+2}^{j-\theta-1} \sum_{k=1}^{h-1} ZM^k(i,r-1) \cdot ZM1^{h-k}(r,j) \cdot e^{-b/RT} \end{aligned} \tag{6}$$

5 Number of multiloops

As before, define the abbreviations for the partition function for hairpins

$$ZH(i,j) = \begin{cases} 0 & \text{if } j-i \le \theta \\ e^{-\mathcal{H}(i,j)/RT} & \text{else} \end{cases}$$

For Peer Review Only/Not for Distribution

and internal loops having k multiloops

$$ZI^m(i,j) = \sum_{i \leq k \leq \ell \leq j} e^{-\mathcal{I}(i,j;k,l)/RT} \cdot ZB^m(k,\ell)$$

where the sum is over k, ℓ such that $(k-i)+(j-\ell) > 0$. As in the previous section, this combines the treatment of both left and right bulges with proper internal loops.

Define $Z^0(i,i) = 1$, and for $m > 0$, define $Z^m(i,i) = 0$. For the remaining base cases, define $ZB^m(i,i) = ZM^m(i,i) = ZM1^m(i,i) = 0$. The unconstrained partition function for secondary structures restricted to the interval $[i,j]$ that contain m multiloops is given by

$$Z^m(i,j) = Z^m(i,j-1) + \sum_{r=i}^{j-\theta-1} \sum_{k=0}^{m} Z^k(i,r-1) \cdot ZB^{m-k}(r,j)$$

The partition function for secondary structures restricted to the interval $[i,j]$ that contain m multiloops and are closed by the base pair (i,j) is given by $ZB^m(i,j) = ZI^m(i,j)$, if $m = 0$, and in the case that $m > 0$ and i,j can form a base pair by

$$\begin{aligned} ZB^m(i,j) &= ZI^m(i,j) + \sum_{r=i+\theta+3}^{j-\theta-2} \sum_{k=0}^{m-1} ZM^k(i+1,r-1) \\ &\quad \cdot ZM1^{m-k-1}(r,j-1) \cdot e^{-(a+b)/RT}. \end{aligned}$$

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

$$ZM1^m(i,j) = \sum_{r=i+\theta+1}^{j} ZB^m(i,r) \cdot e^{-c(j-r)/RT}. \qquad (7)$$

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

$$\begin{aligned} ZM^m(i,j) &= \sum_{r=i}^{j-\theta-1} ZM1^m(r,j) \cdot e^{-(b+c(r-i))/RT} + \\ &\quad \sum_{r=i+\theta+2}^{j-\theta-1} \sum_{k=1}^{m-1} ZM^k(i,r-1) \cdot ZM1^{m-k-1}(r,j) \cdot e^{-b/RT} \end{aligned} \qquad (8)$$

================ FIGURE 2 GOES ABOUT HERE ===================

6 Multiloop order

The *order* (or *depth*) of a secondary structure is the maximum *depth* of nesting of its multiloops. Formally, multiloop order can be defined via a finite analogue of the Cantor-Bendixson topological derivative (Clote, 1984; Kechris, 1995). The *derivative* $D(\mathcal{S})$ of secondary structure $\mathcal{S}$ is equal to the set of base pairs $(i,j) \in \mathcal{S}$, within which there is an internal branching; i.e.

$$\begin{aligned} D(\mathcal{S}) &= \{(i,j) \in \mathcal{S} : \text{there exist distinct } (x,y),(u,v) \in \mathcal{S} \\ &\quad \text{such that } i < x < y < u < v < j\}. \end{aligned}$$

For Peer Review Only/Not for Distribution

The *order* of a secondary structure $\mathcal{S}$ is now defined to be $n-1$, where n is the least integer such that $D(\mathcal{S}) = \emptyset$. For readers familiar with the notion of RNA *shape* (Giegerich et al., 2004), it follows that the order of a helix is zero, with shape `[]`, while the order of a tRNA cloverleaf secondary structure is one, with shape `[[][][]]`. Examples of order 2 secondary structures, with shape `[[[][][]][]]`, are furnished by certain RNase P RNA molecules, such as STRAND database (Andronescu et al., 2008) sequence ASE00001 from *Acidianus ambivalens*, and by some transfer-messenger RNA, such as STRAND database sequence TMR00040, from *Azos.oryz.*.

For typographic reasons, we denote the multiloop partition function by Z^d, rather than Z^o. As before, define the partition function for hairpins

$$ZH(i,j) = \begin{cases} 0 & \text{if } j-i \leq \theta \\ e^{-\mathcal{H}(i,j)/RT} & \text{else} \end{cases}$$

and internal loops having multiloop order or depth d

$$ZI^d(i,j) = \sum_{i \leq k \leq \ell \leq j} e^{-\mathcal{I}(i,j;k,\ell)/RT} \cdot ZB^d(k,\ell)$$

where the sum is over k, ℓ such that $(k-i)+(j-\ell) > 0$. Define $Z^0(i,i) = 1$ and for $d > 0$, define $Z^d(i,i) = 0$. For $d \geq 0$, define $ZB^d(i,i) = ZM^d(i,i) = ZM1^d(i,i) = 0$. The unconstrained partition function for secondary structures of multiloop order d, when restricted to the interval $[i,j]$, is given by

$$Z^d(i,j) = Z^d(i,j-1) + \sum_{r=i}^{j-\theta-1} \sum_{0 \leq k,\ell \leq d, \max(k,\ell)=d} Z^k(i,r-1) \cdot ZB^\ell(r,j)$$

The partition function for secondary structures of multiloop order d when restricted to the interval $[i,j]$ and are closed by the base pair (i,j) is given as follows. For $d=0$, let

$$ZB^d(i,j) = ZH(i,j) + ZI^d(i,j)$$

while for $d > 0$, define

$$\begin{aligned} ZB^d(i,j) &= ZI^d(i,j) + \sum_{r=i+\theta+3}^{j-\theta-1} \sum_{0 \leq k,\ell \leq d, \max(k,\ell)=d} \\ & \quad ZM^k(i+1,r-1) \cdot ZM1^\ell(r,j-1) \cdot e^{-(a+b)/RT}. \end{aligned}$$

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

$$ZM1^d(i,j) = \sum_{r=i+\theta+1}^{j} ZB^d(i,r) \cdot e^{-c(j-r)/RT}. \tag{9}$$

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

$$\begin{aligned} ZM^d(i,j) &= \sum_{r=i}^{j-\theta-1} ZM1^d(r,j) \cdot e^{-(b+c(r-i))/RT} + \\ & \quad \sum_{r=i+\theta+2}^{j-\theta-1} \sum_{0 \leq k,\ell \leq d, \max(k,\ell)=d} ZM^k(i,r-1) \cdot \\ & \quad ZM1^\ell(r,j) \cdot e^{-b/RT}. \end{aligned} \tag{10}$$

For Peer Review Only/Not for Distribution

7 Simultaneous multiloop number and hairpin number

Given the algorithms described in Sections 4 and 5, it is straightforward to design the algorithm `RNAmloopHP`, which computes the partitiion function $Z(m,h)$ simultaneously for m multiloops and h hairpins. Sampling low energy structures is done by a straightforward variation of the sampling method introduced by Ding and Lawrence (Ding and Lawrence, 2003). For purposes of brevity, further details of the partition function and sampling will not be discussed, though the interested reader can study our publicly available source code.

8 Applications

In this section, we mention two main applications of the new algorithms, though first we mention that `RNAhairpin` presents a novel method to generate suboptimal secondary structures.

In the literature, there are a number of approaches to compute *suboptimal* secondary structures. Historically, the first method was due to Zuker (Zuker, 1989), as implemented in `mfold mfold` (Zuker, 1989) and `Unafold` (Markham and Zuker, 2008), who for certain base pairs (i,j) computed the minimum free energy structure containing (i,j) that was sufficiently distinct from previously generated suboptimal structures. Next, the program `RNAsubopt` of Wuchty et al. (Wuchty et al., 1999) generated all secondary structures within a user-specified energy above the minimum free energy. In contrast the program `Sfold` (Ding and Lawrence, 2003) samples from the low energy Boltzmann ensemble of structures; indeed, our implementation of sampling in `RNAhairpin` is a modification of the method of Ding and Lawrence (Ding and Lawrence, 2003). (Note that the `Sfold` algorithm is implemented in the Vienna RNA Package program `RNAsubopt` with flag `-p`; as well the program `RNAstructure` (Mathews et al., 2004) also supports sampling.) The program `RNAshapes` of Steffen et al. (Steffen et al., 2006) computes the minimum free energy structure from each shape class. The program `RNAbor` of Freyhult et al. (Freyhult et al., 2007) computes, for each k, the minimum free energy structure $MFE(k)$ having base pair distance k from a user-specified reference structure, while the program `RNA2Dfold` of Lorenz et al. (Lorenz et al., 2009) computes, for each k, ℓ, the minimum free energy structure $MFE(k, \ell)$ having base pair distance k [resp. ℓ] from a first [resp. second] user-specified reference structure. The program `RNAlocopt` of Lorenz and Clote (Lorenz and Clote, 2011) samples low locally optimal secondary structures, where a locally optimal structure has the property that its free energy cannot be lowered by the addition or removal of a single base pair. The program `RNAbormea` of Lou and Clote (Clote et al., 2012) determines for each k, the maximum expected accuracy structure among all structures having base pair distance k from a user-specified reference structure. To this list of previous methods, `RNAhairpin` generates suboptimal secondary structures in a manner that seems orthogonal to previous methods.

8.1 Improved structure prediction for certain RNA families

Certain RNAs have a characteristic structure that is important for their function. For instance, the cloverleaf structure of transfer RNA generally has three hairpins, which then form an L-shaped tertiary structure by additional pseudoknots. Transfer RNAs usually contain a small number of chemically modified nucleotides, making their structure at times difficult to predict using minimum free energy structure methods. In such cases, `RNAhairpin` [resp. and expecially `RNAmloopHP`] can improve structure prediction by sampling low energy structures that are required to have a specific number of hairpins [resp. number m of multiloops and h of hairpins].

Table 1 presents a comparison of `RNAhairpin` and `RNAfold` statistics for sequences taken from the seed alignments of several families from Rfam 11.0 (Burge et al., 2013) (August 2012, 2208

For Peer Review Only/Not for Distribution

families). For each sequence, we sampled only one low energy structure having H hairpins. For a given sequence and structure computed either by `RNAhairpin` or `RNAfold`, the sensitivity, or true positive rate, is computed, defined as the ratio of number of correctly predicted base pairs in the Rfam structure over the number of base pairs in the Rfam structure. The average and standard deviation of sensitivity is given, for each Rfam family of the table, for both `RNAhairpin` and `RNAstructure`. For these computations, version 1.8.5 of `RNAfold` was used without dangles, so that both `RNAhairpin` and `RNAfold` employed the same energy model. In future work, we plan to lift `RNAhairpin` to the Turner 2004 energy model and implement dangles, which then would support the same energy model as version 2.0 and higher of `RNAfold` (Lorenz et al., 2011).

In the case of tRNA, there is more than 20% improvement in sensitivity of `RNAhairpin` over `RNAfold`; `RNAhairpin` has greater sensitivity than `RNAfold` for other instances, such as in the case of the hammerhead ribozyme (around 4% improvement). On the other hand, `RNAfold` has greater sensitivity than `RNAhairpin` for several classes, including HIV primer binding site RF00375 (over 5% improvement), and purine riboswitch aptamers RF00167 (around 4.5% improvement). Clearly `RNAhairpin` is not a better structure prediction tool than `RNAfold`; however, for particular classes of functional RNA, which require certain hairpin structures for function, `RNAhairpin` may provide a useful tool. See Section 9 for more discussion.

The program `RNAmloopHP`, which samples low energy structures having m multiloops and h hairpins, improves the structure prediction accuracy of `RNAhairpin` (e.g. an improvement of over 4% for RF000167 purine riboswitches), and also outperforms `RNAfold` for a larger number of cases on the previously described Rfam families. For instance, there is an improvement of almost $approx$24% in RF00005 (tRNA), over 4% in RF00008 (type III hammerhead ribozyme), 5% in RF00504 (glycine riboswitch), etc. On the other hand, `RNAmloopHP` has significantly lower sensitivity than `RNAfold` in the following two cases, where the difference is over 5% for RF00375 (HIV primer binding site), and 8% for RF00635 (HAR1A). The consensus structures for these Rfam families have large loop regions, which may in fact be base-paired, which could explain the poorer performance of `RNAmloopHP`. (Recall that Rfam consensus structures are determined by covariation found in multiple alignments, thus loop regions in consensus structures could indeed by base-paired and involve additional hairpins and/or multiloops.) In any case, we do not propose the use of `RNAmloopHP` in place of minimum free energy structure software, such as `RNAfold`; instead, if a biologist has knowledge or intuition about the existence of a certain number of multiloops and hairpins, then `RNAmloopHP` may prove to be a useful tool.

8.2 Support vector machine results

In this section, we describe receiver operating characteristic (ROC) curves, computed by 5-fold cross-validation, where in each case, the positive instances were taken to be sequences from the seed alignment of a given Rfam family, and negative instances were taken to be random sequences having the same number of dinucleotides, as computed by our implementation of the Altschul-Erikson algorithm (Altschul and Erikson, 1985). (Similar results were obtained, when we took negative instances to be sequences the seed alignments of all other Rfam families – data not shown.)

For each positive instance, we generated 10 random negative instances. Using `libSVM` (Chang and Lin, 2001), we performed a *stratified training* by selecting one-fifth of the positive instances together with an equal number of negative instances (one of the 10 negative instances was selected for each positive instance) for training. The remaining four-fifths of the positive sequences, together with all corresponding negative instances, constituted the test set (note that in testing, there were 10 negative instances per positive sequence). A radial basis kernel was chosen with cost $C = 1$, and parameter γ taken to be the reciprocal of the number of features.

For Peer Review Only/Not for Distribution

We considered three features sets: HP, HP/M, and HP/M/O, where HP features were the 21 probabilities of hairpin formation $p_h(0), \ldots, p_h(20)$, where M features were the 6 probabilities of multiloop formation $p_m(0), \ldots, p_m(5)$, and where O features were the 6 probabilities of multiloop order (or depth) $p_d(0), \ldots, p_d(5)$. Thus the SVM binary classifier HP (hairpins) has 21 features, though in most cases all but a small number of the features are 0; the classifier HP/M (hairpin and multiloop number) has 27=21+6 features; the classifier HP/M/O (hairpin, multiloop number, multiloop order) has 33=21+6+6 features. The R packages `e1071` (Meyer et al., 2012) and `pROC` (Robin et al., 2011) were used with `libSVM` (Chang and Lin, 2001).

Table 3 summarizes the area under curve (AUC) values for ROC curves for the three different SVM classifiers HP, HP/M, HP/M/O, while Figure 3 depicts the corresponding ROC curves. Note that in all cases, inclusion of multiloop order probabilities as features does not add any discriminatory power, and even in certain cases reduces the AUC. This is fortunate, since `RNAmloopOrder` cannot be sped up by using the fast Fourier transform, unlike `RNAhairpin` and `RNAmloopNum`. The results of this table and figure indicate that, although hairpin and multiloop formation probabilities may not be sufficient to be used solely as the feature set of a noncoding RNA gene finder, we believe that, when added, these features could lead to improvements in performance of existent RNA gene finders. Moreover, to the best of our knowledge, current noncoding RNA gene finders do not take into account global propensity to form hairpins or multiloops.

Table 4 presents the ratio of ROC area under curve values for support vector machines (SVM) over that for relevance vector machines (RVM). A value greater [resp. less] than unity in the table indicates that SVM outperforms [resp. underperforms] RVM using the same features. Figure 4 shows an unexpected situation for the 5-fold (stratified) cross-validation experiments of the Rfam family RF00027, Using the feature set consisting of only 21 hairpin formation probabilities $p_h(0), \ldots, p_h(20)$, the ratio of SVM/RVM AUC is 1.4234, indicating that SVM far outperforms RVM for this family using these features, while for the full feature set of hairpin formation probabilities $p_h(0), \ldots, p_h(20)$, multiloop number probabilities $p_h(0), \ldots, p_h(6)$, and multiloop maximum order (depth) probabilities $p_h(0), \ldots, p_h(6)$, the ratio of SVM/RVM AUC is 0.8986, indicating that RVM outperforms SVM.

================= FIGURE 3 GOES ABOUT HERE ====================
================= FIGURE 4 GOES ABOUT HERE ====================

9 Discussion and conclusion

We terminate the paper with a discussion of strengths and shortcomings of each application shown: improved structure prediction and support vector machine classification.

Parametric structure prediction

For benchmarking purposes in Table 1, we sampled only one low energy structure having H many hairpins, where in most cases H was taken to be the number of hairpins in the Rfam consensus structure of the first member of the Rfam family. This explains how it could happen that the `RNAhairpin` sensitivity for a certain sequence could at times be different than the `RNAhairpin` sensitivity for the same sequence, even when the sampled structure and the minimum free energy structure have the same number of hairpins. Of course, in general, our code `RNAhairpin` will be used to sample a large number (1,000 or 10,000) of structures per sequence.

Since the base pairs that appear in Rfam consensus structures are inferred by covariation observed in a multiple alignment, many valid base pairs do not appear in the consensus structure. For this reason, we did not compute *positive predictive rate*, defined as the ratio of the number of

For Peer Review Only/Not for Distribution

correctly predicted base pairs in the Rfam consensus structure, divided by the number of base pairs in the predicted structure. This is the reason that Table 1 only reports average sensitivity values.

As previously mentioned, we computed the average sensitivity of the minimum free energy (MFE) structure obtained from Vienna RNA Package `RNAfold` (Hofacker, 2003), version 1.8.5 without dangles, in order to ensure that both `RNAhairpin` and `RNAfold` employ the same energy model. In future work, we plan to extend `RNAhairpin` to the Turner 2004 energy model with dangles (Turner and Mathews, 2009). By the same token, it is not conceptually difficult to modify the program `RNAmloopNum`, in order to sample low energy structures having a specified number of multiloops. Such sampled structures could yield better structure predictions for certain types of RNA. Finally, it would be possible to combine the algorithms `RNAhairpin` and `RNAmloopNum` in order sample structures having *both* a specified number of hairpins and a possibly distinct number of multiloops. Nevertheless, such an algorithm would run in time $O(H^2M^2n^3)$ and space $O(HMn^2)$, where H [resp. M] is an upper bound on the number of hairpins [resp. multiloops]. For relatively small values of H, M, such an algorithm would indeed be feasible, and could prove useful for certain classes of RNA, whose function is known to depend on certain structural motifs.

Table 1 presents examples of Rfam families, where the average sensitivity of `RNAhairpin` exceeds that of `RNAfold`. Improvements were obtained for RNA families, where a certain number of hairpins are known to be functionally important, as in the cloverleaf tRNA, typically having three hairpins. In this case, the sensitivity of `RNAhairpin` exceeds that of `RNAfold` by approximately 20%. For certain ribozymes, such as type III hammerhead ribozyme (RF00008) and glycine ribozyme (RF00504), the improvement was over 4% resp. 3%. Not shown in the table are RNA families, where the sensitivity of `RNAfold` exceeded that of `RNAhairpin` – for instance, for 5S rRNA (RF00001), `RNAhairpin` average sensitivity was 0.621306 compared with `RNAfold` sensitivity of 0.633189; for purine riboswitches (RF00167), `RNAhairpin` obtained had 0.811327 compared with `RNAfold` sensitivity of 0.856764. We believe that `RNAhairpin` showed better sensitivity than `RNAfold` in the case of tRNA, because of two reasons: (1) tRNA has a well-known cloverleaf structure usually involving 3 hairpins, and (2) there may be a large sequence and energy difference among especially bacterial tRNAs. Item (2) could cause minimum free energy structures to be quite distinct from the usual cloverleaf – manual investigation confirms this hypothesis in some randomly chosen instances, while item (2) ensures that `RNAhairpin` will sample structures having 3 hairpins, hence more likely to adopt the functional cloverleaf structure. It could be that similar reasons explain the small improvement of `RNAhairpin` over `RNAfold` for some of the other examples, including certain ribozymes. However, along this line of reasoning, we expected `RNAhairpin` to outperform `RNAfold` for purine riboswitch aptamers, which have a very well-defined multiloop with two hairpins; Table 1 shows this is not the case. As described in the caption of Table 2, by sampling low energy structures that simultaneously have a specified number m of multiloops and number h of hairpins, we substantially improve the prediction accuracy of `RNAfold`. However, such improvements tend to occur when the Rfam families show a prounounced common fold, as in the case of tRNA and certain ribozymes, and when there are no large loop (undefined) regions in the Rfam consensus structures. In any case, we believe that minimum free energy structure prediction algorithms, such as `RNAfold`, `UNAFold`, `mfold`, `RNAstructure`, remain the best universal thermodynamics-based tool for structure prediction.

Features for SVM classifiers

The development of noncoding RNA gene finders is important for the analysis and classification of the pervasively transcribed RNA from the human genome, most of which has no previously known structure or function. In this paper, we have described four novel thermodynamics-based

For Peer Review Only/Not for Distribution

algorithms, `RNAhairpin`, `RNAmloopNum`, `RNAmloopOrder`, and `RNAmloopHP` which compute global, parametric features of the ensemble of low energy secondary structures for a given RNA sequence. For the first three algorithms, we have shown that there is a significant *global* signal, as witnessed by ROC area under curve, that suggests that probability and multiloop formation probabilities present useful features that could be added to existent noncoding RNA gene finders – note that this remark only concerns gene finders for specific noncoding RNA families, not general ncRNA gene finders.

One of our goals in developing these parametric algorithms was to provide additional discriminatory features that can be added to other features within the context of a support vector machine, in order to improve the accuracy of noncoding RNA gene finders. Indeed, by adding novel features, it is known that one can improve the accuracy of SVM classifiers. For instance, the state-of-the-art precursor microRNA (pre-miRNA) SVM developed by Ng and Mishra (Ng and Mishra, 2007) uses features MFEI2, MFEI1, %G+C, dP, dG, dQ, dD, dF,,zG, zQ, zD, zP, zF, etc. (see (Ng and Mishra, 2007) for explanation), which outperforms the simpler triplet kernel pre-miRNA SVM developed by Xue et al. (Xue et al., 2005).

Acknowledgments

We would like to thank Hosna Jabbari for improving the Figure 1. The initial part of the research in this paper was performed while authors Y. Ding and W.A. Lorenz were part of our lab at Boston College. This research was funded by the National Science Foundation under grant DBI-1262439. 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 National Science Foundation.

References

S.F. Altschul and B.W. Erikson. Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol*, 2(6):526–538, 1985.

M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC. Bioinformatics*, 9:340, 2008.

M. Bekaert, L. Bidou, A. Denise, G. Duchateau-Nguyen, J.P. Forest, C. Froidevaux, I. Hatin, J.P. Rousset, and M. Termier. Towards a computational model for −1 eukaryotic frameshifting sites. *Bioinformatics*, 19:327–335, 2003.

H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki. The Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.*, 58(Pt):899–907, June 2002.

A. Böck, K. Forschhammer, J. Heider, and C. Baron. Selenoprotein synthesis: An expansion of the genetic code. *Trends Biochem. Sci.*, 16:463–467, 1991.

A. F. Bompfunewerer, R. Backofen, S. H. Bernhart, J. Hertel, I. L. Hofacker, P. F. Stadler, and S. Will. Variations on RNA folding and alignment: lessons from Benasque. *J. Math. Biol.*, 56(1-2):129–144, January 2008.

S. W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner, and A. Bateman. Rfam 11.0: 10 years of RNA families. *Nucleic. Acids. Res.*, 41(Database):D226–D232, January 2013.

H. P. Cam, E. S. Chen, and S. I. Grewal. Transcriptional scaffolds for heterochromatin assembly. *Cell*, 136(4): 610–614, February 2009.

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

For Peer Review Only/Not for Distribution

M. T. Cheah, A. Wachter, N. Sudarsan, and R. R. Breaker. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, 447(7143):497–500, May 2007.

S. Chowdhury, C. Ragaz, E. Kreuger, and F. Narberhaus. Temperature-controlled structural alterations of an RNA thermometer. *J. Biol. Chem.*, 278(48):47915–47921, November 2003.

P. Clote. Théorème de cantor-bendixson. In *Séminaire Général de Logique*, volume 27, pages 73–88. Publications Mathématiques de l'Université Paris VII, 1984. ISSN 0073-8301.

P. Clote, F. Lou, and W. A. Lorenz. Maximum expected accuracy structural neighbors of an RNA secondary structure. *BMC. Bioinformatics*, 13:S6, 2012.

K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, 106(1):97–102, January 2009.

Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic. Acids. Res.*, 31:7280–7301, 2003.

S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. U.S.A.*, 83(24):9373–9377, December 1986.

E. Freyhult, V. Moulton, and P. Clote. Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, 23(16):2054–2062, August 2007.

R. Giegerich, B. Voss, and M. Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Res.*, 32(16):4843–4851, 2004.

A. R. Gruber, R. Neubock, I. L. Hofacker, and S. Washietl. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic. Acids. Res.*, 35(Web):W335–W338, July 2007.

R. Gutell, J. Lee, and J. Cannone. The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology*, 12:301–310, 2005.

J. H. Havgaard, R. B. Lyngso, and J. Gorodkin. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic. Acids. Res.*, 33(Web):W650–W653, July 2005.

Berman HM, Westbrook J, Feng Z, Iype L, Schneider B, and Zardecki C. The nucleic acid database. *Methods Biochem Anal.*, 44:199–216, 2003.

I.L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31:3429–3431, 2003.

I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsch. Chem.*, 125:167–188, 1994.

Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 88:207–237, 1998. URL `citeseer.nj.nec.com/1454.html`.

H. Kazan, D. Ray, E. T. Chan, T. R. Hughes, and Q. Morris. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS. Comput. Biol.*, 6:e1000832, 2010.

A. S. Kechris. *Classical Descriptive Set Theory.* Springer-Verlag, 1995. ISBN 978-0-387-94374-9.

L.P. Lim, M.E. Glasner, S. Yekta, C.B. Burge, and D.P. Bartel. Vertebrate microRNA genes. *Science*, 299(5612): 1540, 2003.

R. Lorenz, C. Flamm, and I.L. Hofacker. 2D projections of RNA folding landscapes. In I. Grosse, S. Neumann, S. Posch, F. Schreiber, and P.F. Stadler, editors, *German Conference on Bioinformatics 2009*, volume 157 of *Lecture Notes in Informatics*, pages 11–20, 2009.

R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. Viennarna Package 2.0. *Algorithms. Mol. Biol.*, 6:26, 2011.

For Peer Review Only/Not for Distribution

W.A. Lorenz and P. Clote. Computing the partition function for kinetically trapped RNA secondary structures. *Public Library of Science One (PLoS ONE)*, 6(1):316178, 2011. doi:10.1371/journal.pone.0016178.

M. Mandal, B. Boese, J.E. Barrick, W.C. Winkler, and R.R. Breaker. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell*, 113(5):577–586, 2003.

N. R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, 453:3–31, 2008.

D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, 101(19):7287–7292, May 2004.

D.H. Matthews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2012. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6-1.

K. L. Ng and S. K. Mishra. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23(11):1321–1330, June 2007.

A.D. Omer, T.M. Lowe, A.G. Russell, H. Ebhardt, S.R. Eddy, and P.P. Dennis. Homologues of small nucleolar RNAs in Archaea. *Science*, 288:517–522, 2000.

ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frdrique Lisacek, Jean-Charles Sanchez, and Markus Mller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

E. Senter, S. Sheikh, I. Dotu, Y. Ponty, and P. Clote. Using the fast fourier transform to accelerate the computational search for RNA conformational switches. *PLoS. One.*, 7(12):e50506, 2012.

M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26:148–153, 1998.

P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

P. R. Stein and M. S. Waterman. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Mathematics*, 26:261–272, 1978.

B. J. Tucker and R. R. Breaker. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, 15(3): 342–348, June 2005.

D. H. Turner and D. H. Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, 0(O):O, October 2009.

Q. Vicens and T.R. Cech. Atomic level architecture of group I introns revealed. *Trends Biochem. Sci.*, 31(1):41–51, 2006.

A. von Haeseler, B. Blum, L. Simpson, N. Sturm, and M. S. Waterman. Computer methods for locating kinetoplastid cryptogenes. *Nucleic Acids Res.*, 20(11):2717–2724, 1992.

S. Washietl, I. L. Hofacker, P. F. Stadler, and M. Kellis. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic. Acids. Res.*, 40(10):4261–4272, May 2012.

For Peer Review Only/Not for Distribution

J. S. Weinger, K. M. Parnell, S. Dorner, R. Green, and S. A. Strobel. Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nat. Struct. Mol. Biol.*, 11:1101–1106, 2004.

S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–164, 1999.

T. Xia, Jr. J. SantaLucia, M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, and D.H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–35, 1999.

C. Xue, F. Li, T. He, G. P. Liu, Y. Li, and X. Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC. Bioinformatics*, 6:310, 2005.

K. Zarringhalam, M. M. Meyer, I. Dotu, J. H. Chuang, and P. Clote. Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS. One.*, 7(10):e45160, 2012.

M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(7):48–52, 1989.

For Peer Review Only/Not for Distribution

RNA family	H	RNAhairpin $\mu \pm \sigma$	RNAfold $\mu \pm \sigma$	avg len	num seq
RF00001	2	0.6213 ± 0.2667	0.6332 ± 0.2721	116.6	712
RF00004	5	0.7548 ± 0.1840	0.7104 ± 0.2058	190.5	208
RF00005	3	0.7345 ± 0.2313	0.5370 ± 0.1992	73.4	960
RF00008	2	0.9565 ± 0.1284	0.9154 ± 0.1894	55.4	84
RF00031	1	0.7679 ± 0.1748	0.7657 ± 0.1945	64.5	61
RF00045	4	0.4420 ± 0.2983	0.4205 ± 0.3274	202.6	66
RF00094	2	0.3080 ± 0.2131	0.3604 ± 0.2091	91.1	33
RF00167	2	0.8113 ± 0.2301	0.8568 ± 0.2290	100.8	133
RF00375	2	0.8278 ± 0.3060	0.8814 ± 0.2044	99.0	130
RF00504	2	0.5940 ± 0.2711	0.5603 ± 0.2895	100.9	44
RF00635	4	0.3024 ± 0.1127	0.3707 ± 0.1204	117.9	13
RF01055	4	0.5821 ± 0.2725	0.5787 ± 0.2641	142.0	160

Table 1: Comparison between RNAhairpin and RNAfold of the average sensitivity (ratio of number of correctly predicted base pairs in Rfam structure over number of base pairs in Rfam structure) for various Rfam families. RNAhairpin was used to sample a single secondary structure having H many hairpins, and the average sensitivity of RNAhairpin and RNAfold was computed over all sequences in the seed alignment of the following Rfam families: RF00001 (5S rRNA), RF00004 (splicesomal U2 RNA), RF00005 (tRNA), RF00008 (type III hammerhead ribozyme), RF00031 (selenocysteine insertion sequence I), RF00045 (snoRNA), RF00094 (HDV ribozyme), RF00167 (purine riboswitch), RF00375 (HIV primer binding site), RF00504 (glycine riboswitch), RF00635 (HAR1A), RF01055 (moco RNA motif).

For Peer Review Only/Not for Distribution

RNA family	M	H	RNAhairpin $\mu \pm \sigma$	RNAfold $\mu \pm \sigma$	avg len	num seq
RF00001	1	2	0.6308 ± 0.2571	0.6332 ± 0.2721	116.6	712
RF00004	0	5	0.6980 ± 0.1780	0.7104 ± 0.2058	190.5	208
RF00005	1	3	0.7740 ± 0.1946	0.5370 ± 0.1992	73.4	960
RF00008	1	2	0.9582 ± 0.1005	0.9154 ± 0.1894	55.4	84
RF00031	0	1	0.7679 ± 0.1748	0.7657 ± 0.1945	64.5	61
RF00045	1	4	0.4456 ± 0.2977	0.4205 ± 0.3274	202.6	66
RF00094	0	2	0.3464 ± 0.1951	0.3604 ± 0.2091	91.1	33
RF00167	1	2	0.8511 ± 0.1726	0.8568 ± 0.2290	100.8	133
RF00375	1	2	0.8283 ± 0.3063	0.8814 ± 0.2044	99.0	130
RF00504	1	2	0.6101 ± 0.264	0.5603 ± 0.2895	100.9	44
RF00635	1	3	0.2930 ± 0.1059	0.3707 ± 0.1204	117.9	13
RF01055	1	4	0.60170 ± 0.277	0.5787 ± 0.2641	142.0	160

Table 2: Comparison between RNAmloopHP and RNAfold of the average sensitivity for the same Rfam families, as in Table 1. By now sampling a single secondary structure having simultaneously M many multiloops and H many hairpins, the average sensitivity improved over that of RNAhairpin in essentially all cases. Moreover, RNAmloopHP provides more accurate structure prediction (sensitivity) than RNAfold for a number of Rfam families. There is an improvement of almost *approx*24% in RF00005 (tRNA), over 4% in RF00008 (type III hammerhead ribozyme), 2.5% in RF00045 (snoRNA), 5% in RF00504 (glycine riboswitch), over 2% in RF01055 (moco RNA motif). On the other hand, RNAmloopHP has significantly lower sensitivity than RNAfold in the following two cases, where the difference is over 5% for RF00375 (HIV primer binding site), and 8% for RF00635 (HAR1A). Insignificant differences, such as 0.6308 for RNAmloopHP versus 0.6332 in RF00001 (5S rRNA) are likely to be due to the stochastic nature of sampling low energy structures, rather than computing the MFE structure having a specified number of multiloops and hairpins.

For Peer Review Only/Not for Distribution

Family name and description	H	HM	HMO	num seq	avg len	avg GC %
RF00004 U2 spliceosomal RNA	0.9217	0.9282	0.9328	208	204.26	48.0%
RF00005 tRNA	0.6367	0.9038	0.9017	959	73.4	47.0%
RF00008 hammerhead III	0.9191	0.9705	0.9562	84	55.4	48.4%
RF00027 let 7 microRNA precursor	0.8338	0.8766	0.8617	67	79.6	43.7%
RF00031 SECIS 1	0.7917	0.8361	0.7941	61	64.5	49.0%
RF00045 SNORA73	0.6306	0.6515	0.6609	66	202.6	53.1%
RF00167 purine riboswitch	0.6508	0.8608	0.8529	136	100.8	38.1%

Table 3: Area under curve (AUC) for receiver operating characteristic (ROC) curves for seven Rfam families, each family tested under 5-fold cross-validation with support vector machines (SVM) using a radial basis kernel with cost $C = 1$ and γ equal to the inverse of the number of features. In the case of H (hairpin number), there were 21 hairpin formation probabilities $p(0), \ldots, p(20)$ taken as features, (though in most cases all but a very small number of these probabilities were zero); in the case of HM (hairpin and multiloop number), there were 27=21+6 hairpin and multiloop formation probabilities taken as features, and in the case of HMO (hairpin and multiloop number with maximum multiloop order), there were 27=21+6 hairpin and multiloop formation probabilities taken as features along with 6 multiloop maximum order probabilities, hence altogether 33=21+6+6 features. The R packages `e1071` (Meyer et al., 2012) and `pROC` (Robin et al., 2011) were used with `libSVM` (Chang and Lin, 2001).

Ratio SVM/RVM	RF00004	RF00005	RF00008	RF00027	RF00031	RF00045	RF00167
HP	0.9874	1.0657	0.9874	1.4234	1.1965	0.9895	1.1894
HP/M	0.9863	0.9798	0.9977	1.0625	1.0954	0.9808	1.0153
HP/M/O	0.9818	0.9855	1.0025	0.8986	1.2324	1.0237	1.0031

Table 4: Ratio of ROC area under curve values for two types of machine learning methods: support vector machines (SVM) and relevance vector machines (RVM), using the same seven Rfam families that were considered in Table 3. In 11 out of 21 tests, AUC values for SVMs were greater than those for RVMs. In the case of RF00027, it is interesting to note that when using only hairpin features, SVM AUC was much higher than RVM AUC (SVM/RVM 1.4234), while for the same class, when using the larger feature set for hairpins, multiloop number and multiloop order, SVM AUC was lower than RVM AUC (SVM/RVM 0.8986). At present, the reason for this surprising result is unclear. The R packages `e1071` (Meyer et al., 2012) and `pROC` (Robin et al., 2011) were used for SVM and RVM computations; for SVM, the radial basis kernel (rbfkernel) was employed with default parameters, while for RVM, rvmbinary rbfdot kernel was used with default parameters and 1000 iterations.

For Peer Review Only/Not for Distribution

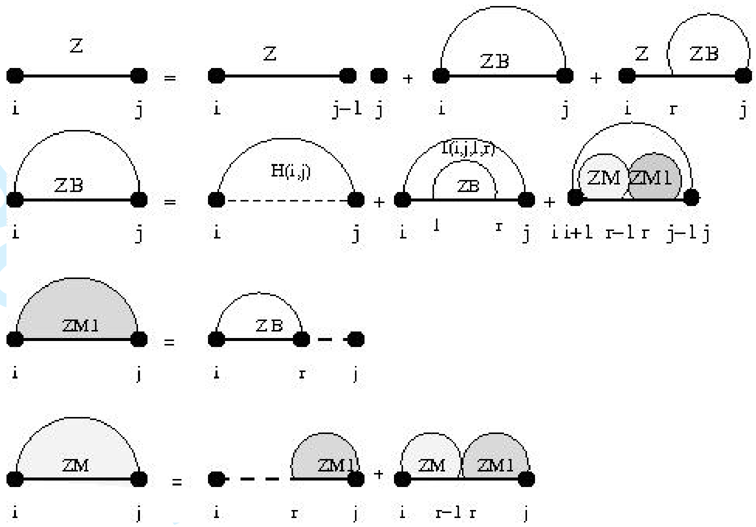

Figure 1: Feynman diagram of original recursions from McCaskill's algorithm (McCaskill, 1990) to compute the partition function. Dashed lines present intervals of unpaired bases, and shaded arcs represent structures in which i and j will not necessarily pair.

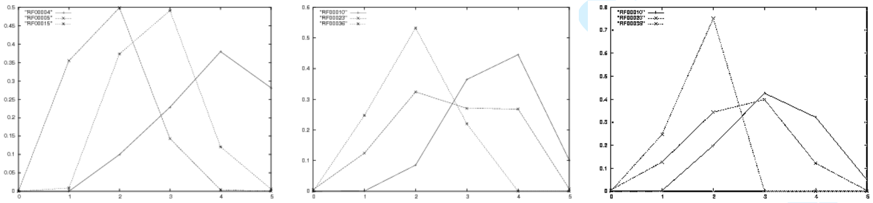

Figure 2: *(Left)* Hairpin profile of Rfam families: U2 spliceosomal RNA (RF00004), transfer RNA (tRNA, RF00005) and U4 spliceosomal RNA (RF00015). *(Center)* Multiloop number profile of Rfam families: RNaseP (RF00010), transfer messenger RNA (tmRNA, RF00023), and Rev response element of HIV env gene (RF00036). *(Right)* Multiloop order (or depth) profile of Rfam families: RNaseP (RF00010), transfer messenger RNA (tmRNA, RF00023), and Rev response element of HIV env gene (RF00036). Notice that we chose Rfam families consisting of long RNA sequences for multiloop number/order profiles, since multiloops are energetically unfavorable, hence are not generally present in small RNA.

For Peer Review Only/Not for Distribution

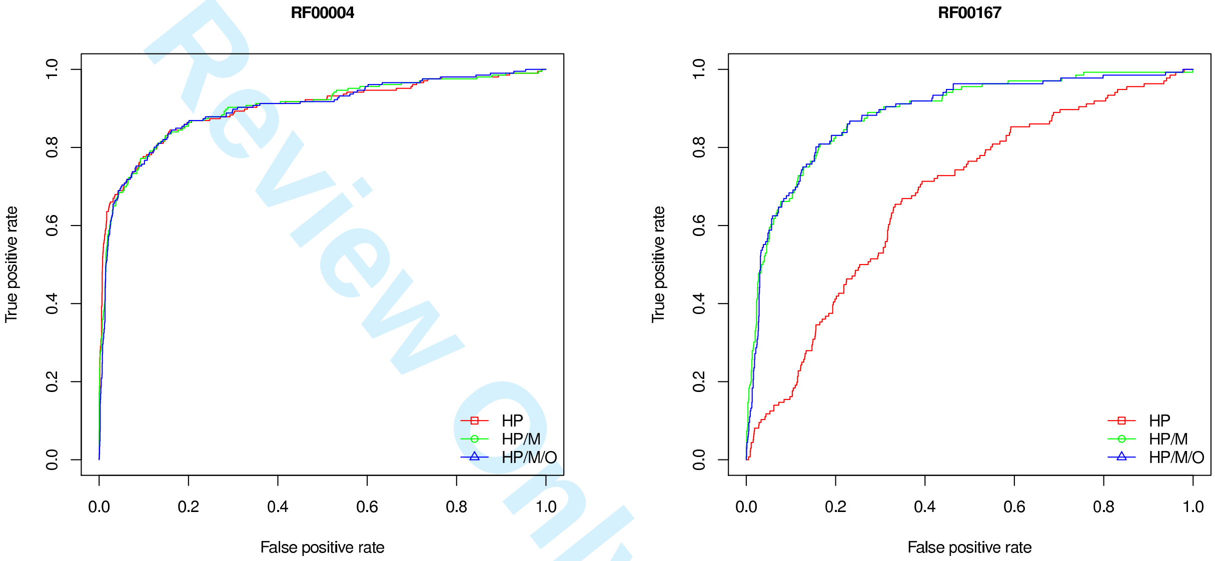

Figure 3: Receiver operating characteristic (ROC) curves for the performance of support vector machine binary classification using a feature set consisting of probabilities $p(0),\ldots,p(20)$ for the number hairpins (HP), probabilities $p(0),\ldots,p(5)$ for the number of multiloops (M), and probabilities $p(0),\ldots,p(5)$ for the maximum order of multiloops (O). In the case of HP (hairpin number), there were 21 features, though in most cases all but at most 6-8 features had the value 0; in the case of HP/M (hairpin and multiloop number), there were 27=21+6 features, and in the case of HP/M/O (hairpin and multiloop number with maximum multiloop order), there were 33=21+6+6 features. The R packages `e1071` (Meyer et al., 2012) and `pROC` (Robin et al., 2011) were used with `libSVM` (Chang and Lin, 2001). A radial basis kernel was used in each case with cost $C = 1$; parameter γ was taken to be the inverse of the number of features, i.e. for HP, $\gamma = 1/21 = 0.0476$, for HP/M, $\gamma = 1/27 = 0.0370$, for HP/M/O, $\gamma = 1/33 = 0.0303$. As shown in this figure, accounting for multiloop order did not improve classification ROC curves, and data presented in Table 3 shows that in some cases, ROC area under curve is lessened by taking into account maximum multiloop order. This is in fact fortunate, since the fast Fourier transform can be applied to reduce time and space requirements for `RNAhairpin` and `RNAmloopNum`, but not `RNAmloopOrder`. *(Left)* Rfam family RF00004 (U2 spliceosomal RNA). *(Right)* Rfam family RF00167 (purine riboswitch).

For Peer Review Only/Not for Distribution

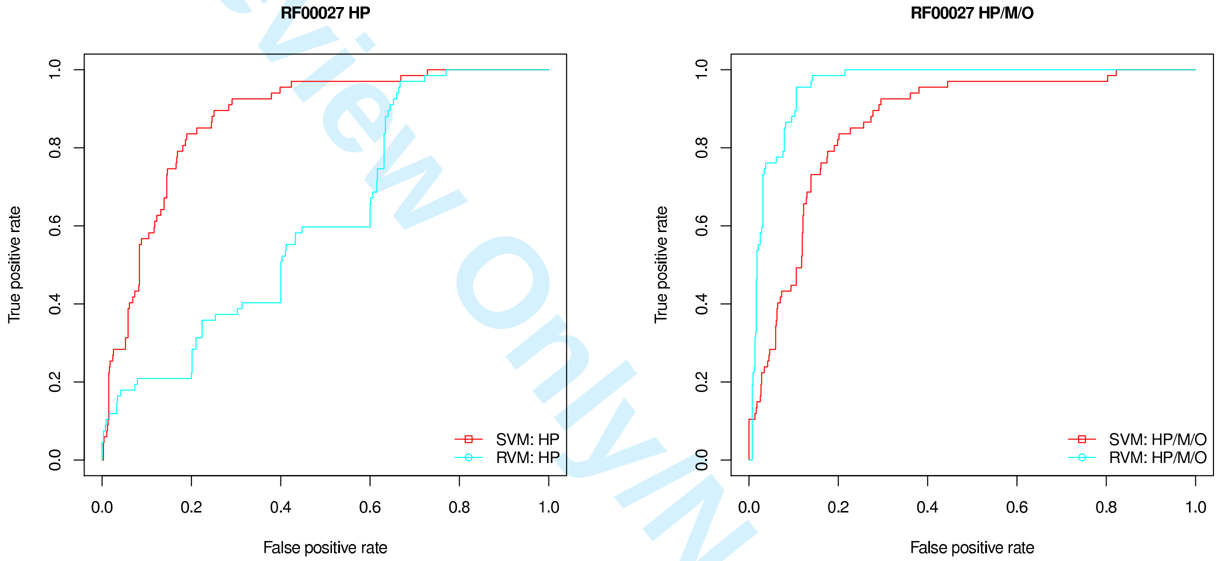

Figure 4: Receiver operating characteristic (ROC) curves for 5-fold cross-validation for sequences from the seed alignment of RF00027. The left panel shows an overlay of support vector machine (SVM) and relevance vector machine (RVM) for the feature set consisting of 21 hairpin formation probabilities $p_h(0), \ldots, p_h(20)$, while the right panel presents an overlay of SVM and RVM for the full feature set of hairpin formation probabilities $p_h(0), \ldots, p_h(20)$, multiloop number probabilities $p_h(0), \ldots, p_h(6)$, and multiloop maximum order (depth) probabilities $p_h(0), \ldots, p_h(6)$. As explained in the caption of Table 4, it seems unusual that SVM outperforms RVM using only hairpin probability features, while the reverse is true when using the full feature set.

For Peer Review Only/Not for Distribution

Appendix: Using FFT to compute RNAhairpin.

In (Freyhult et al., 2007), we developed the algorithm RNAbor, which computes the minimum free energy structure $MFE(k)$ and the partition function $Z(k)$ for each integer k, where $Z(k)$ is the sum of Boltzmann factors $\exp(-E(S)/RT)$ and the sum is taken over all structures having base pair distance k from a user-specified reference structure. Like the parametric algorithms RNAhairpin, RNAmloopNum and RNAmloopOrder in this paper, RNAbor runs in time $O(n^5)$ and space $O(n^3)$, when all values $Z(k)$ are needed for $0 \leq k \leq n$.

In (Senter et al., 2012), we described a more efficient means to compute the partition functions $Z(k)$ for $0 \leq k \leq n$, by using the FFT to determine probabilities $p(k) = Z(k)/Z$ by polynomial interpolation. Since the partition function Z can be separately computed by McCaskill's algorithm (McCaskill, 1990), the new method yields the values $Z(k)$ for $0 \leq k \leq n$ in time $O(n^4)$ and space $O(n^2)$.

Given the algorithmic similarities between the parametric algorithms of this paper and RNAbor, we can use the same method of polynomial interpolation using the FFT to compute probabilities $p_h(k) = Z^h(k)/Z$ and $p_m(k) = Z^m(k)/Z$ for hairpin resp. multiloop formation, for all $0 \leq k \leq n$. For technical reasons clear to the careful reader, In (Senter et al., 2012), we described a more efficient means to we can not use this new method to compute probabilities $p_d(k)$ for multiloop order (depth). Moreover, although RNAhairpin can sample low energy structures having exactly k hairpins, for desires values k, and although it is not difficult (though labor intensive) to modify both RNAhairpin and RNAmloopNum to compute in time $O(n^5)$ and space $O(n^3)$ the minimum free energy structures $MFE_h(k)$ resp. $MFE_m(k)$, taken over all structures having k hairpins resp. multiloops, the $O(n^4)$ time and $O(n^2)$ space FFT method can *not* compute these minimum free energy structures.

With these remarks, we succinctly describe the overall recursions for the FFT version of RNAhairpin; similar recursions apply to the FFT version of RNAmloopNum. Both FFT algorithms have been implemented and are publicly available at http://bioinformatics.bc.edu/RNAparametric/.

FFT version of RNAhairpin

Let $\mathbf{s} = s_1, \ldots, s_n$ be a given RNA sequence. For all $1 \leq i \leq j \leq n$, we define the *polynomial*

$$\mathcal{Z}_{i,j}(x) = \sum_{k=0}^{n-1} z_{i,j}(k) \cdot x^k$$

where $z_{i,j}$ is the hairpin partition function for interval $[i,j]$; i.e. $z_{i,j}$ is the sum of Boltzmann factors $\exp(-E(S)/RT)$, taken over all secondary structures S of $s_i, \ldots, s_j$. Since the coefficients of any polynomial of degree strictly less than n can be efficiently determined by the FFT using polynomial interpolation, provided that one first evaluates the polynomial at n many complex nth roots of unity $1, \exp(\frac{2\pi i}{n}), \ldots, \exp(\frac{2\pi i(n-1)}{n})$. For a complex number α, in order to evaluate $\mathcal{Z}(\alpha) = \mathcal{Z}_{1,n}(\alpha)$, we proceed by recursions that resemble somewhat the recursions given in Section 4. To compute $\mathbf{Z}_{1,n}(\alpha)$, we use dynamic programming to evaluate $\mathbf{Z}_{i,j}(\alpha)$, for all $1 \leq i \leq j \leq n$; moreover, in order to compute $\mathbf{Z}_{i,j}(\alpha)$, we need to compute $\mathbf{ZB}_{i,j}(\alpha)$, $\mathbf{ZM}_{i,j}(\alpha)$, and $\mathbf{ZM1}_{i,j}(\alpha)$.

Now let $\mathbf{B}$ denote the set of canonical base pairs GC, CG, AU, UA, GU, UG. To compute $\mathcal{Z}(x) = \mathbf{Z}_{1,n}(x)$, we use the recursions

$$\mathbf{Z}_{i,j}(x) = \mathbf{Z}_{i,j-1}(x) \cdot x + \sum_{\substack{s_k\, s_j \in \mathbb{B}, \\ i \leq k < j}} \Big(\mathbf{Z}_{i,k-1}(x) \cdot \mathbf{ZB}_{k,j}(x) \Big). \tag{11}$$

For Peer Review Only/Not for Distribution

The sum is taken over all possible base pairs (k, j) with $i \leq k < j$.

We compute $\mathbf{ZB}(x)$ using the recursion

$$\begin{aligned}
\mathbf{ZB}_{i,j}(x) &= e^{-EH(i,j)/RT} \cdot x \\
&\quad + \sum_{\substack{s_k s_l \in \mathbb{B}, \\ i<k<l<j}} \mathbf{ZB}_{k,l}(x) \cdot e^{-EI(i,j,k,l)/RT} \\
&\quad + \sum_{\substack{s_k \in \mathbb{B}, \\ i<k<j}} \left(\mathbf{ZM}_{i+1,k-1}(x) \cdot \mathbf{ZM1}_{k,j-1}(x) \cdot e^{-(a+b)/RT} \right),
\end{aligned} \tag{12}$$

where $EH(i, j)$ is the energy of the hairpin loop with closing base pair (i, j), $EI(i, j, k, l)$ is the energy of the stack, bulge or interior loop with the closing base pair (i, j) To reduce complexity of the algorithm, the interior and bulge loop size can be limited to a maximum size of L, by requiring that $l > j - L$ in the above recursion.

The recursion for computing $\mathbf{ZM1}(x)$, is

$$\mathbf{ZM1}_{i,j}(x) = \sum_{\substack{s_k \in \mathbb{B}, \\ i<k\leq j}} \left(\mathbf{ZB}_{i,k}(x) \cdot e^{-(b+c(j-k))/RT} \right)$$

The final recursion, for computing $\mathbf{ZM}(x)$, is

$$\mathbf{ZM}_{i,j}(x) = \sum_{\substack{s_k \in \mathbb{B}, \\ i\leq k<j}} \left(\mathbf{ZM1}_{k,j}(x) \cdot e^{-(c(k-i))/RT} + \mathbf{ZM}_{i,k-1}(x) \cdot \mathbf{ZM1}_{k,j}(x) \cdot e^{-(c(k-i))/RT} \right).$$

Note that $\mathbf{ZM1}_{i,j}(x)$ [resp. $\mathbf{ZM}_{i,j}(x)$] are defined under the assumption that $[i, j]$ is part of a multi-loop for which the multiloop energy penalty a is already applied. Moreover, for $\mathbf{ZM}_{i,j}(x)$, there is either exactly one stem-loop structure in $[i, j]$, corresponding to the $\mathbf{ZM1}_{i,j}(x)$ term, or more than one stem-loop in $[i, j]$, corresponding to the $\mathbf{ZM}_{i,k-1}(x) \cdot \mathbf{ZM1}_{k,j}(x)$ term. Justification of recursions (11), (12), and (13) follow by induction.

Also, for all $i < j$ such that $j - i < \theta$ we initialize the recursions as follows

$$\mathbf{Z}_{i,j}(x) = 1 \ , \ \mathbf{ZB}_{i,j}(x) = 0 \ , \ \mathbf{ZM}_{i,j}(x) = 0 \ , \ \mathbf{ZM1}_{i,j}(x) = 0$$

Finally, we mention that as in (Senter et al., 2012), it is necessary to interpolate the probabilities $p_h(k) = \frac{Z^h_{1,n}(k)}{Z}$ due to numerical stability issues that arise when trying to interpolate very large partition function values. This completes the sketch of the FFT version of `RNAhairpin`; analogous recursions lead to an FFT interpolation of partition functions $Z^m(k)$ for multiloop number. For more details, please consult (Senter et al., 2012).