JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 2, 2007

Pp. 190–215
DOI: 10.1089/cmb.2006.0012

Computing the Partition Function and Sampling for Saturated Secondary Structures of RNA, with Respect to the Turner Energy Model

J. WALDISPÜHL[1,3] and P. CLOTE[1,2]

ABSTRACT

An RNA secondary structure is *saturated* if no base pairs can be added without violating the definition of secondary structure. Here we describe a new algorithm, `RNAsat`, which for a given RNA sequence a, an integral temperature $0 \leq T \leq 100$ in degrees Celsius, and for all integers k, computes the Boltzmann partition function $Z_k^T(a) = \sum_{S \in \mathcal{SAT}_k(a)} exp(-E(S)/RT)$, where the sum is over all saturated secondary structures of a which have exactly k base pairs, R is the universal gas constant and $E(S)$ denotes the free energy with respect to the Turner *nearest neighbor* energy model. By dynamic programming, we compute Z_k^T simultaneously for all values of k in time $O(n^5)$ and space $O(n^3)$. Additionally, `RNAsat` computes the partition function $Q_k^T(a) = \sum_{S \in \mathcal{S}_k(a)} exp(-E(S)/RT)$, where the sum is over *all* secondary structures of a which have k base pairs; the latter computation is performed simultaneously for all values of k in $O(n^4)$ time and $O(n^3)$ space. Lastly, using the partition function Z_k^T [resp. Q_k^T] with stochastic backtracking, `RNAsat` rigorously samples the collection of saturated secondary structures [resp. secondary structures] having k base pairs; for Q_k^T this provides a parametrized form of `Sfold` sampling (Ding and Lawrence, 2003). Using `RNAsat`, *(i)* we compute the ensemble free energy for saturated secondary structures having k base pairs, *(ii)* show cooperativity of the Turner model, *(iii)* demonstrate a temperature-dependent phase transition, *(iv)* illustrate the predictive advantage of `RNAsat` for precursor microRNA `cel-mir-72` of *C. elegans* and for the pseudoknot PKB 00152 of Pseudobase (van Batenburg et al., 2001), *(v)* illustrate the RNA shapes (Giegerich et al., 2004) of sampled secondary structures [resp. saturated structures] having exactly k base pairs. A web server for `RNAsat` is under construction at *bioinformatics.bc.edu/clotelab/RNAsat/*.

Key words: secondary structure, RNA, Boltzmann partition function, kinetic trap.

1. INTRODUCTION

IN RECENT YEARS, it has emerged that RNA plays a surprising and previously unsuspected role in many biological processes, including *retranslation* of the genetic code [selenocysteine insertion (Böck et al., 1991; Heider et al., 1992), ribosomal frameshift (Moon et al., 2004)], post-transcriptional regulation

Departments of [1]Biology and [2]Computer Science, Boston College, Chestnut Hill, Massachusetts.
[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts.

via small interfering RNA and microRNA (Lim et al., 2003; Tuschl, 2003), conformational switches (Voss et al., 2006), metabolite-sensing *riboswitches* which interact with small ligands and up- or down-regulate certain genes (Barrick et al., 2004), small nucleolar RNAs which guide the methylation of specific ribosomal nucleotides (Schattner et al., 2004), etc.

The function of structural RNA certainly depends on RNA tertiary structure, which in Banerjee et al. (1993) has been shown to be largely determined by RNA *secondary structure*. The *Turner nearest neighbor model* (Matthews et al.,1999; Xia et al., 1999) consists of experimentally determined enthalpy and entropy values for stacked base pairs (both Watson-Crick and GU wobble pairs), as well as for hairpin loops, bulges and interior loops (arbitrary multi-loop free energies are approximated by an affine measure). Zuker's algorithm (Zuker and Stiegler, 1981) is a dynamic programming algorithm, which computes the minimum free energy (mfe) secondary structure of an input RNA nucleotide sequence, with respect to the nearest neighbor model. Though there are small discrepancies in certain energy parameters (e.g., treatment of coaxial stacking and dangles[1]), Zuker's `mfold` (Zuker, 2003), the Vienna RNA Package `RNAfold` (Hofacker et al., 1994) and Mathews and Turner's `RNAstructure` (Mathews et al., 2000) are all implementations of Zuker's algorithm for the Turner energy model (Matthews et al., 1999; Xia et al., 1999) and run in $O(n^3)$ time and $O(n^2)$ space; i.e., time is cubic and space is quadratic in the length of the input RNA nucleotide sequence.

Most current work on RNA secondary structure concerns the thermodynamic equilibrium minimum free energy structure (`mfold`, `RNAfold`, `RNAstructure`), or the low energy ensemble of structures [`Sfold` (Ding et al., 2004), `RNAsubopt` (Wuchty et al., 1999)], or multiple sequence/structure alignment [`Foldalign` (Havgaard et al., 2005), `Dynalign` (Mathews and Turner, 2002)], or the general shape of RNA secondary structures [(`RNAshapes`) (Giegerich et al., 2004; Steffen et al., 2006; Voss et al., 2006)], or applications and extensions of such software. One important application area concerns the development of noncoding RNA genefinders, such as `RNAz` (Washietl et al., 2005a,b).

Considering molecular complexes, Dimitrov and Zuker (2004) described how to efficiently compute a multi-species partition function for interactions between two species of nucleic acid sequences (e.g., between DNA and RNA). This led to the hybridization web server `DINAMelt` and software `UNAFOLD` (Markham and Zuker, 2005a; Bernhart et al., 2006; Mückstein et al., 2006). Computation of the partition function for molecular complexes allows one to compute concentrations of various hybridization products, clearly relevant in probe design for gene expression chips.

In contrast to much of the previously mentioned work, the current paper considers secondary structure energy landscape from a different point of view. We divide up the exponentially large collection of *saturated* [resp. *all*] secondary structures for a given RNA sequence into disjoint families F_k, each indexed by k. Choosing F_k appropriately, we compute partition functions at temperature T for the following: *(i)* Z_k^T, for *saturated* secondary structures having k base pairs, *(ii)* Q_k^T, for *all* secondary structures having k base pairs, *(iii)* $\widehat{Z}_k^T$, for *saturated* secondary structures having k base pairs fewer than the maximum possible (i.e., Nussinov-Jacobson optimal), *(iv)* $\widehat{Q}_k^T$, for all secondary structures having k base pairs fewer than the maximum possible (i.e., Nussinov-Jacobson optimal). For each of these cases, the partition function is computed with respect to the Turner energy model, and rigorous sampling is provided.

Our guiding motivation is to use the partition function and derived thermodynamic parameters (e.g., ensemble free energy, heat capacity) to better understand the nature and distribution of kinetic traps in the folding landscape of RNA. To this end, given a secondary structure S for an RNA sequence $a_1, \ldots, a_n$, we say that S is *locally optimal* with respect to the Turner energy model, if $E(S) \leq E(S')$, for all secondary structures S', obtained from S by either the removal of addition of one base pair.[2] Unable currently to compute the partition function for secondary structures which are locally optimal with respect to the Turner energy model, we instead describe our software `RNAsat`, which computes the partition function for saturated secondary structures with respect to the Turner energy model. Computational experiments reported here

[1] Dangles, either $5'$ or $3'$, are unpaired bases immediately adjacent to a paired base; dangles stack below or on top of existent stacked base pairs, hence contribute a stabilizing energy. At $37^\circ\mathrm{C}$, a $3'$ dangle can have an energy of up to approximately $-2\,\mathrm{kcal/mol}$. When computing minimum free energy structures and partition functions, dangles constitute an important factor.

[2] Here $E(S)$ is the energy of S using the Turner energy model (Matthews et al., 1999; Xia et al., 1999). In the Nussinov-Jacobson energy model, the notion of locally optimal and saturated coincide, and the corresponding problem was solved in (Clote, 2005a,b).

with precursor microRNA `cel-mir-72` of *C. elegans* and with pseudoknot PKB 00152 from Pseudobase (van Batenburg et al., 2001) suggest that saturated structures and locally optimal structures are quite distinct concepts, and that there are far fewer locally optimal structures than saturated structures. Nevertheless, the algorithms described are non-trivial, and provide a tool to better understand the folding landscape of RNA. The goal of the remaining paper is to explain the algorithms behind `RNAsat`; in future work, we will investigate additional applications. We begin by some definitions.

Given an RNA sequence $a = a_1, \ldots, a_n$, a secondary structure is a well-balanced parenthesis expression with dots, where the nucleotides a_i, a_j forming either Watson-Crick or GU wobble pairs correspond to well-balanced left and right parentheses. Formally, a secondary structure S on RNA sequence $a_1, \ldots, a_n$ is defined to be 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 $i < k < j < \ell$.
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 $i = k$.

In this paper, following convention, the threshold θ, or minimum number of unpaired bases in a hairpin loop, is taken to be 3. For any additional background on RNA and dynamic programming computation of secondary structures, the reader can consult the text (Clote and Backofen, 2000) and the recent review (Eddy, 2004).

Zuker (1986) defined a *saturated* structure to be a secondary structure in which no base pairs can be added without violating the definition of secondary structure (e.g., without introducing a pseudoknot). More recently, Evers and Giegerich (2001) defined a saturated structure to be a secondary structure, such that no additional base pairs can be stacked and for which there are no isolated base pairs. For this notion, Evers and Giegerich (2001) described an algorithm which computes the total number of saturated structures. It is not difficult to construct examples of structures which are saturated in the sense of Zuker, but not in the sense of Evers-Giegerich, and vice-versa—i.e., these notions are distinct; (for explicit examples, see Clote, 2006)

A related notion is that of *locally optimal* secondary structure. To define this notion, we first define *base pair distance*, a metric between two secondary structures for the same RNA sequence. If S, T are secondary structures for an RNA sequence $a_1, \ldots, a_n$, then $d(S, T)$ is the number of elements in the symmetric difference of S, T; i.e. the number of base pairs in S, but not T, or vice-versa. Structure T is a *neighbor* of S if the base pair distance between S, T is 1.

Given a fixed energy model, such as the base pair model of Nussinov and Jacobson (1980) or the nearest neighbor model of Turner (Matthews et al., 1999; Xia et al., 1999), a *locally optimal* secondary structure is a structure S, such that no neighbor T of S has lower energy. We developed an algorithm and web server to compute the number of locally optimal secondary structures with respect to the Nussinov-Jacobson energy model; for this energy model, local optimality coincides with Zuker's notion of saturated structure (Clote, 2005a,b).

Locally optimal structures correspond to potential *kinetic traps* in the folding process. Indeed, if S is a locally optimal secondary structure, which differs from the minimum free energy (mfe) structure S_0, then an energetically unfavorable neighbor S' of S must appear in the folding path from S to mfe structure S_0. The barrier tree of Flamm et al. (2002) illustrates this situation.

In this paper, we lift the algorithm of Clote, (2005a) to the Turner energy model; specifically we proceed as follows. Given an RNA nucleotide sequence $a = a_1, \ldots, a_n$, for each value of k, we compute the Boltzmann partition function $Z_k^T(a) = \sum_{S \in \mathcal{SAT}_k(a)} exp(-E(S)/RT)$, where $\mathcal{SAT}_k(a)$ designates the collection of all saturated secondary structures S of a having exactly k base pairs, R is the universal gas constant with value 1.98717 cal/mol per degree Kelvin, $0 \leq T \leq 100$ is temperature in degrees Celsius, and where $E(S)$ denotes the free energy, using the Turner *nearest neighbor* energy model (Matthews et al., 1999; Xia et al., 1999). By dynamic programming, we compute Z_k^T simultaneously for fixed T, for all values of k in time $O(n^5)$ and space $O(n^3)$. Although the notion of saturated structure (in the sense of Zuker) is distinct from that of locally optimal structure within the Turner nearest neighbor energy model, the work of this paper is a step towards the goal of computing the partition function for all locally optimal secondary structures.

Our software, RNAsat, additionally computes the full partition function

$$Q^T(a) = \sum_{S \in \mathcal{S}(a)} exp(-E(S)/RT)$$

and its parametrized form $Q_k^T(a) = \sum_{S \in \mathcal{S}_k(a)} exp(-E(S)/RT)$, where $\mathcal{S}(a)$ denotes the collection of *all* secondary structures of RNA sequence a, and $\mathcal{S}_k(a)$ designates the collection of *all* secondary structures S of a which have exactly k base pairs. The $O(n^3)$ time and $O(n^2)$ space algorithm to compute the partition function $Q^T(a)$ was first developed by McCaskill (1990). McCaskill's algorithm has additionally been implemented by Hofacker et al. (2004) in the Vienna RNA Package, by Mathews (2004), and by Markham and Zuker (2005b). However, unlike these authors, we additionally compute in time $O(n^4)$ and space $O(n^3)$ the partition function $Q_k^T(a)$ for all secondary structures having exactly k base pairs, for all values of $1 \leq k \leq n$. We omit details, since the method is similar, though simpler, to the method we present to compute the partition function of saturated structures having k base pairs.

RNAsat can sample from the collection of saturated secondary structures having k base pairs, using stochastic backtracking with $Z_k^T(a)$. Additionally, RNAsat can sample from the collection of *all* secondary structures having k base pairs, using stochastic backtracking with $Q_k^T(a)$. Our sampling procedure is mathematically rigorous, and is analogous to the method of Ding and Lawrence (2003) with web server Sfold (Ding et al., 2004), except that we sample with respect to $Z_k^T(a)$ and $Q_k^T(a)$ for any temperature $0 \leq T \leq 100$, whereas Ding and Lawrence use McCaskill's partition function $Q(a)$ to sample all secondary structures at 37°C. The algorithms of RNAsat to compute $Z_k^T(a)$ and sample saturated secondary structures having k base pairs are quite distinct from any existing algorithms. As well, despite the close relation of $Q_k^T(a)$ to $Q^T(a)$, to the best of our knowledge, no other software supports the computation of $Q_k^T(a)$ or samples, for given k, from the collection of secondary structures having k base pairs.

Using partition functions, we develop the following applications. *(i)* We compute ensemble free energies for saturated secondary structures having k base pairs, thus settling the question of thermodynamic stability of saturated structures. *(ii)* We demonstrate that the Turner nearest neighbor energy model (Xia et al., 1999) leads to *cooperative folding*, in the sense of Dill and Bromberg (2002). *(iii)* We compute and graph the *relative density of states* for saturated secondary structures having k base pairs as a function of temperature; these curves resemble temperature-dependent phase transitions. *(iv)* We illustrate the predictive advantage of RNAsat for precursor microRNA cel-mir-72 of *C. elegans* and for pseudoknot PKB 00152 of Pseudobase (van Batenburg et al., 2001), *(v)* We compute frequencies of RNA shapes (Giegerich et al., 2004) of sampled secondary structures [resp. saturated structures] having exactly k base pairs.

The plan of the paper is as follows. Section 2 presents key definitions used in our dynamic programming algorithm for saturated secondary structures. Additionally, this section describes several applications of RNAsat. Section 3 presents the pseudocode and an explanation of how to compute the Boltzmann partition function $Z_k^T(a)$ for all saturated secondary structures having exactly k base pairs. Section 4 gives pseudocode for the algorithm to sample, for each k, from the ensemble of saturated secondary structures having k base pairs. The sampling is rigorously performed according to the Boltzmann probability; i.e. lower energy structures in the ensemble $SAT_k(a)$ of saturated secondary structures of a having k base pairs are more heavily weighted than higher energy structures. Since sampling from the ensemble of saturated secondary structures is different from the procedure of Ding et al. (2004), we include a proof of correctness of our algorithm. Although our software RNAsat as well as the pseudocode for sampling given in Figures 12–14 below is for the Turner energy model Xia et al. (1999), the proof of correctness is given for the simpler Nussinov-Jacobson energy model, in the interests of readability.[3]

Finally, we note that in this paper, when the temperature T is clear from context, we may write Z_k, Z [resp. Q_k, Z] in place of Z_k^T, Z^t [resp. Q_k^T, Q^T], for example.

[3]The reader familiar with the Turner energy model should have no difficulty in constructing a similar argument for Proposition 1.

2. OVERVIEW AND APPLICATIONS

In 1980, Nussinov and Jacobson introduced a simple energy model, where each base pair in a secondary structure S contributes an energy of -1.[4] In this model, known as the *Nussinov-Jacobson energy model*, the energy $E(S)$ of secondary structure S equals $-|S|$, where $|S|$ denotes the number of base pairs in S. In Nussinov and Jacobson (1980), a dynamic programming algorithm[5] was given to compute the optimal secondary structure for a given RNA sequence; i.e., the optimal secondary structure with respect to the Nussinov-Jacobson energy model is that structure having the maximum number of base pairs (for implementation details, see Clote and Backofen, 2000).

Much earlier, a realistic *nearest neighbor* energy model was developed by Tonoco et al. (1973) with experimentally measured stacking free energies; i.e., energy contributions come from stacked base pairs rather than base pairs. The well-known Zuker algorithm (Zuker and Stiegler, 1981) computes the minimum free energy (mfe) over all secondary structures by adding contributions of negative (stabilizing) energy terms for stacked base pairs and positive (destabilizing) energy terms for hairpin loops, bulges, internal loops and multiloops. Initially implemented with energy parameters from Tonoco et al. (1973), successive refinements of Zuker's algorithm have incorporated more accurate energy parameters from Matthews et al. (1999) with refinements from Xia et al. (1999) for stacked base pairs and various loops (hairpin, bulge, internal loop, multi-branch loop) and $5'$ and $3'$ dangles, i.e., unpaired nucleotides which stack onto a base pair. The energy term contributed by a base pair depends on the base pair (if any) upon which it is stacks; for instance, Turner's current rules (Xia et al., 1999) at 37 degrees Celsius assign stacking free energy of $-2.24\,\mathrm{kcal/mol}$ to $\begin{smallmatrix}5'\text{-AC-}3'\\3'\text{-UG-}5'\end{smallmatrix}$ of $-3.26\,\mathrm{kcal/mol}$ to $\begin{smallmatrix}5'\text{-CC-}3'\\3'\text{-GG-}5'\end{smallmatrix}$ and of $-2.08\,\mathrm{kcal/mol}$ to $\begin{smallmatrix}5'\text{-AG-}3'\\3'\text{-UC-}5'\end{smallmatrix}$. Our software `RNAsat` uses free-energy values for stacked base pairs, dangles, hairpins, bulges, internal loops and multi-loops from `mfold` 3.0 when $T = 37^\circ\mathrm{C}$ and energy values from `mfold 2.3` otherwise. In particular, following the approach of Zuker's `mfold`, when loop size N exceeds 30, the free energy, due to loss of entropy, is $c_T \cdot \ln(N/30)$, where c_T is a temperature-dependent constant; e.g., $c_{37} = 1.079$.

Following Clote, (2005a), given an RNA sequence $a = a_1, \ldots, a_n$, a secondary structure S for a is defined to be k-*saturated*, if S is locally optimal, and S contains k base pairs fewer than the maximum for a; i.e. k fewer base pairs than that of the Nussinov-Jacobson optimal structure. If m denotes the number of base pairs in the Nussinov-Jacobson optimal structure for RNA sequence a, then clearly any k-saturated secondary structure for a has $m - k$ many base pairs, and any saturated secondary structure S on a having k base pairs must be $(m - k)$-saturated.

In Clote, (2005a,b), we developed a dynamic programming algorithm called `RNALOSS`, an acronym for RNA locally optimal secondary structures, computed with respect to the Nussinov-Jacobson energy model. This software computes, for a given RNA nucleotide sequence and each integer k, the number of k-saturated secondary structures. Since the Boltzmann partition function is essentially a weighted count, the algorithm of Clote, (2005a) can be used to compute, for given RNA sequence a and all integers k, the Boltzmann partition function with respect to the Nussinov-Jacobson energy function for all saturated secondary structures on a having k base pairs.

In this paper, we lift the algorithm `RNALOSS` (Clote, 2005a,b) to an $O(n^5)$ time and $O(n^3)$ space algorithm `RNAsat` to compute, for a given RNA sequence a and all integers k, the Boltzmann partition function $Z_k^T(a)$ with respect to the Turner energy model, for any integral temperature T in degrees Celsius, $0 \leq T \leq 100$.

2.1. Visible nucleotides and visible positions

To fix ideas, consider the toy sequence `GGGGCCCCC`. The maximum number of base pairs is 3, and there is one 0-saturated structure (with 3 base pairs) given by `(((...)))`, twelve 1-saturated structures given by

[4] Of course, this model has trivial variants, such as assigning -3 for GC, -2 for AU and -1 for GU pairs. In particular, Proposition 4.3 uses the notation $bp(i, j)$ for the energy of base pair (i, j) in the RNA sequence $a_1, \ldots, a_n$.

[5] This algorithm, known as the Nussinov-Jacobson algorithm, is also called the maximum circular matching algorithm (Hofacker et al., 2004).

```
GGGGCCCCC   GGGGCCCCC   GGGGCCCCC   GGGGCCCCC
..(...))    .(.(...))   .((....))   ((...)..)
(..(...))   .((...).)   (.(....))   (.(...)).
((....).)   ((....)).   ((...).).   ((...))...
```

and three 2-saturated secondary structures, given by

```
GGGGCCCCC
(...)....
(....)...
...(....)
```

Note that the Nussinov-Jacobson algorithm predicts an optimal secondary structure of `(((...)))` with three base pairs, whereas both `mfold` and `RNAfold` predict a minimum free energy structure `((....))`. of $-1.40\,\mathrm{kcal/mol}$ with one stacked base pair, one exterior base pair, and dangles.

With over 12 trillion saturated structures for the 54 nucleotide hammerhead ribozyme with EMBL accession number AF170517 [Rfam family RF00008 (Griffiths-Jones et al., 2003)], it is impossible in general to enumerate all saturated structures. Instead, we apply dynamic programming using the notions of *visible nucleotides* and *visible positions* defined in Clote, (2005a). Given RNA sequence $a = a_1, \ldots, a_n$, a position $1 \leq i \leq n$ is visible in secondary structure S if for all base pairs $(x, y) \in S$, it is not the case that $x \leq i \leq y$. A nucleotide $X \in \{A, C, G, U\}$ is visible if $X = a_i$, for some visible position i. Define $VisNuc(S) \subseteq \{A, C, G, U\}$ to be the set of visible nucleotides occurring in some position $i < n - \theta$, i.e.

$$VisNuc(S) = \{a_z : \text{ for all } (x, y) \in S\, [(z < x \text{ or } z > y) \text{ and } z < n - \theta]\}\,.$$

To account for visible positions between $n - \theta, \ldots, n$, we define $VisPos(S) = b$, where $0 \leq b \leq \theta + 1$, and b is the greatest value in $0, \ldots, \theta + 1$, such that for all $0 \leq x < b$, $n - x$ is external to every base pair of S. We say that structure S is s, b-visible if $VisNuc(S) = s$ and $VisPos(S, 1, n) = b$. For example, if a is the sequence `AUUCCGGCA` and if the minimum number of unpaired bases in a hairpin loop is $\theta = 1$,[6] then the four secondary structures

```
AUUCCGGCA   AUUCCGGCA   AUUCCGGCA   AUUCCGGCA
(.)(.)...   ...(.)...   (.)..(.).   ..(.(.).)
```

are respectively $\{G\}$, 2-visible, $\{A, G, U\}$, 2-visible, $\{C\}$, 1-visible and $\{A, U\}$, 0-visible. For given RNA sequence $a = a_1, \ldots, a_n$, for fixed temperature T, for all integers k, for all $2^4 = 16$ possible sets s of visible nucleotides and for all possible values $0 \leq b \leq \theta + 1 = 4$ of visible positions, our main algorithm, given in Figures 10 and 11, computes the partition function

$$Z_k^T(a, i, j, s, b) = \sum_{S \in \mathcal{SAT}_k(a,s,b)} exp(-E(S)/RT)$$

where $\mathcal{SAT}_k(a, s, b)$ denotes the set of saturated secondary structures on a which are s, b-visible and have exactly k base pairs. The definitions of $VisNuc$ and $VisPos$ are precisely what is needed to compute $Z_k^T(a, i, j, s, b)$, assuming that all values of $Z_k^T(a, i', j', s', b')$ have been computed and stored in a table for $|j' - i'| < |j - i|$. Using dynamic programming, for all integers k, we compute as well the partition function

$$Q_k^T(a, i, j) = \sum_{S \in \mathcal{S}_k(a)} exp(-E(S)/RT)$$

where $\mathcal{S}_k(a)$ denotes the set of all secondary structures S having exactly k base pairs (i.e., in Q_k^T, secondary structure S need not be saturated).

Ding and Lawrence (2003) describe how to *sample* secondary structures in a mathematically rigorous fashion, by using stochastic backtracking with the partition function as computed by McCaskill's algorithm

[6] In this paper, as well as in `mfold`, `RNAfold` and `RNAstructure`, θ is taken to be 3. We have temporarily taken $\theta = 1$ for perspicacity in this toy example.

(McCaskill, 1990). Though details are substantially more complicated, we have implemented a sampling algorithm for the set of saturated secondary structures having exactly k base pairs (see Figs. 12–14 below). Though not described here, using our computation of $Q_k^T(a)$, our software `RNAsat` includes a rigorous sampling as well for the collection of all secondary structures having k base pairs; i.e. we provide a *parametrized* version of `Sfold` (Ding et al., 2004). In this fashion, `RNAsat` can additionally sample all [resp. saturated] secondary structures irrespective of number of base pairs.

2.2. Applications

For an input RNA sequence, fixed temperature T and fixed value of k, we can generate (say) 1000 samples of saturated secondary structures, each having exactly k base pairs and sampled using the Boltzmann partition function Z_k^T. Subsequent frequency analysis of the samples, when T and k are varied, allow us to investigate aspects of the ensemble of low energy saturated structures. In this section, we present *(i)* ensemble free energies for saturated secondary structures having k base pairs, *(ii)* a demonstration of *cooperative folding* in the Turner energy model, *(iii)* a temperature-dependent phase transition for saturated structures, *(iv)* additional experiments including precursor microRNA `cel-mir-72` from *C. elegans* and pseudoknot PKB 00152 from Pseudobase (van Batenburg et al., 2001), *(v)* a frequency analysis of RNA shapes at various temperatures for all [resp. saturated] secondary structures of a 149 nt. SAM riboswitch having k base pairs fewer than that of the Nussinov-Jacobson optimal structure (for the definition of RNA shape, see Giegerich et al., 2004).

Ensemble free energy. Figure 1 depicts the ensemble free energy $G_k = -RT \ln Z_k^T$, where the universal gas constant R is Avogadro's number times the Boltzmann constant and T is temperature in degrees Kelvin. Note the thermodynamic stability of saturated secondary structures, leading support to the suggestion that saturated structures form a potential kinetic trap in the folding process.

Cooperative folding. Zimm and Bragg (1959) introduced a nearest neighbor model for helix-coil transition of a protein, i.e. where the probability that a residue is in the helix (resp. coil) state is conditional on the state of its nearest neighbor. As indicated by the sigmoidal curves of the number of residues in coil state as a function of temperature, the Zimm-Bragg model illustrates *cooperativity* in the helix to coil transition of a polypeptide (for a good treatment of cooperativity and the Zimm-Bragg and Ising models, Dill and Bromberg, 2002). Figure 2 shows that the Turner model gives rise to a sigmoidal curve, typical of cooperative folding, while the Nussinov-Jacobson model does not. Assuming that the average number of base pairs in the ensemble of secondary structures is inversely proportional to the UV absorbance[7] in a spectrophotometer, notice the similarity between the Turner model curve in Figure 2 with the experimentally determined RNA melting curve of Jaeger et al. (1990).

Apparent phase transition. Define the *relative density of states* ρ_k^T for saturated secondary structures of a given RNA sequence having exactly k base pairs at temperature T to be the relative Boltzmann probability, i.e. $\rho_k^T = Z^T$, for each integer k. Since secondary structure S for a is k-saturated exactly when $|S| = m - k$, where m denotes the maximum number of base pairs in a secondary structure on a, this immediately allows us to compute the relative density of states $\widehat{\rho}_k^T = \widehat{Z}_k^T / \widehat{Z}^T$ for k-saturated secondary structures. Figure 3 displays overlaid graphs, for each fixed k, of the relative density of states for k-saturated structures as a function of temperature. The 0-saturated density is close to 1 at 0°C, gradually declines with rising temperature. With rising temperature, the 1-saturated density gradually increases to a peak, then diminuishes, etc.

It should be mentioned that Figures 2 and 3 of the current paper are similar in form, although distinct to Figure 2 of Dimitrov and Zuker (2004) and Figure 1 of Markham and Zuker (2005b). Figure 2 of the current

[7] Absorbance of UV light at 260 nm is due to especially pyrimidines when light impinges vertically to the plane of the base. Base pairing reduces likelihood of vertical impingement, hence the more base pairing that occurs in a secondary structure, the less UV absorbance will be detected (for more on UV absorbance, see Zheng et al., 2001).

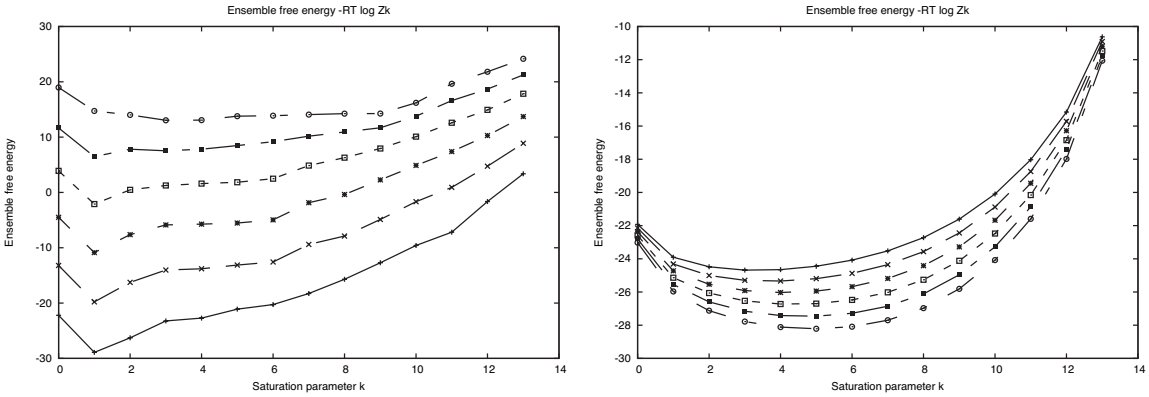

FIG . 1. Graph of the ensemble free energy $G_k = -RT \ln \widehat{Z}_k^T$ for k-saturated secondary structures for hammerhead type III ribozyme with Rfam [19] accession number AF170517 for both the Turner energy model (left panel) and the Nussinov-Jacobson energy model (right panel). As explained in the text, $\widehat{Z}_k^T$ is the sum over all k-saturated structures S of $exp(-E(S) = RT)$. The x-axis of each graph indicates the value of k, the saturation parameter – recall that a secondary structure S is k-saturated if S is saturated and has k fewer base pairs than that of the Nussinov-Jacobson optimal structure. Ensemble free energy G_k is given on the y-axis. It follows that energy values for $k = 0$ correspond to the ensemble free energy of 0-saturated structures, those having the maximum number of base pairs. Each graph displays six curves, corresponding respectively to temperature in degrees Celsius of 0, 20, 40, 60, 80, 100. The curve for $T = 0^\circ$ lies on the bottom, with curves for higher temperatures lying above those for lower temperatures. Note that the lowest ensemble free energy for the Turner model (left panel) occurs when $k = 1$. For the Turner energy model, ensemble free energy is in units of kcal/mol. Energy units are (roughly) in kcal/mol for for the Nussinov-Jacobson (NJ) energy model, for the following reason. For each of the 57 seed hammerhead type III ribozymes in Rfam, we computed the minimum free energy using Vienna RNA `Package RNAfold`. Additionally, we computed the maximum number of base pairs using our implementation of the Nussinov-Jacobson algorithm. From this, we obtain an average minimum free energy per base pair of –0:923385 for Rfam family RF00008 of type III hammerhead ribozymes. Since this value is close to the value –1 assigned per base pair in the Nussinov Jacobson model, it follows that the energy units in the left and right graphs of this figure are comparable.

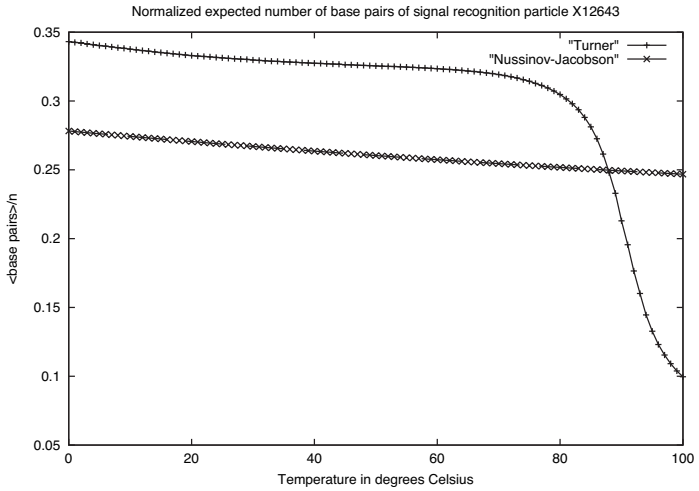

FIG . 2. Graph of the expected number of base pairs as a function of temperature for signal recognition particle with Rfam (Griffiths-Jones et al., 2003) accession number X12643. Temperature in degrees Celsius is given on the x-axis, while the expected number of base pairs $\langle \text{base pairs}/n \rangle$, normalized by sequence length n, is given on the y-axis. Expected number of base pairs is computed by $\langle \text{base pairs} \rangle = \sum_k k \cdot Q_k^T / Q^T$, where Q_k^T is the partition function at temperature T for *all* secondary structures having exactly k base pairs, and $Q^T = \sum_k Q_k^T$ is the partition function at temperature T for *all* secondary structures. Although Q^T is the value obtained by McCaskill's (1990) algorithm, and can be obtained using `RNAfold -p`, the values Q_k^T can only be obtained with our software.

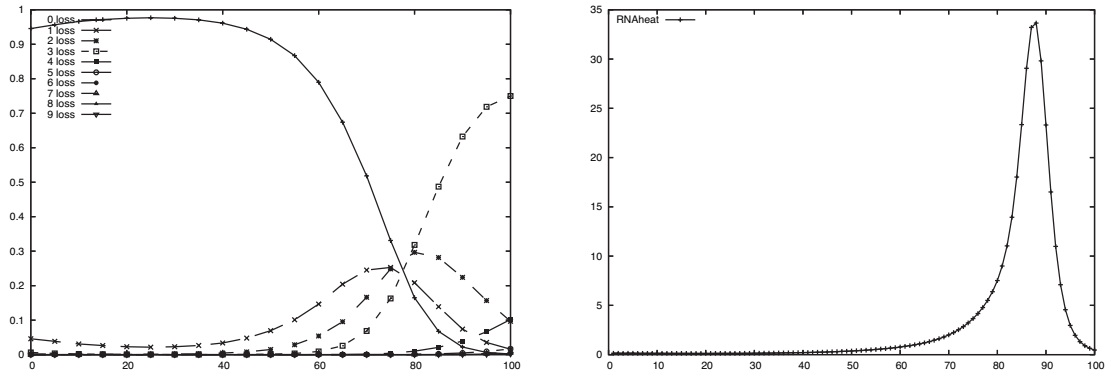

FIG . 3. (Left panel) Graph of relative density of states for k-saturated secondary structures of precursor microRNA `ath-MIR156c` of *Arabidopsis thaliana*, for values of $0 \leq k \leq 10$, as a function of temperature in degrees Celsius. Notice that for $T = 0$, over 90% of the saturated structures are 0-saturated (i.e., have the maximum number of base pairs). As temperature rises, fewer saturated structures are 0-saturated, as indicated by the gradual decline of the Ã˜rst curve. At the same time, there is a gradual increase, as temperature increases, of the proportion of 1-saturated structures (i.e., have one fewer base pair than the maximum). **(Right panel)** Curve of specific heat for `ath-MIR156c`, produced by `RNAheat` from Vienna RNA Package. Temperature is given on the x-axis and specific heat on the y-axis.

paper depicts the expected number of base pairs of a single RNA molecule as a function of temperature,[8] while one of the curves in Figure 2 of Dimitrov and Zuker (2004) depicts the fraction of random coil as a function of temperature. Figure 3 of the current paper depicts the relative density of states of k-saturated secondary structures, for small values of k, as a function of temperature. Though unrelated, Figure 2 of Dimitrov and Zuker (2004) depicts the fraction of various species (duplex, hairpin, random coil) as a function of temperature—both graphs depict rising and falling concentrations as a function of temperature.

Experiments with microRNA `cel-mir-72` *and pseudoknot* `PKB 00152`. In this section, we describe several additional experiments. Figures 4 and 5 illustrate the difference between the minimum free energy (mfe) secondary structure for 96 nt. precursor microRNA `cel-mir-72`, from *C. elegans*, with accession number MI0000043 from miRBase linked to Rfam (Griffiths-Jones et al., 2003). The left panel of Figure 4 depicts the mfe structure, as computed with RNAfold from Vienna RNA Package 1.5. The right panel of this figure displays the consensus structure,[9] as determined by 1000 saturated structures sampled using `RNAsat`. The consensus structure, which by definition consist of the base pairs occurring in strictly more than half the sampled structures, clearly resembles the familiar approximate stem-loop structure which precursor microRNAs are believed to adopt. Figure 5 displays dotplot presentations of Boltzmann pair frequencies; in the upper triangular portion, square are depicted in the position (i, j) with area proportional to frequency that base pair (i, j) occurs in the sampled ensemble. In the lower triangular portion, the minimum free energy structure is represented, in that a black square appears in position (j, i) if base pair (i, j) occurs in the mfe structure, as computed using version 1.5 of `RNAfold`. The left panel of Figure 5 depicts the base pair frequencies from 1000 sampled secondary structures, using the implementation of McCaskill's (1990) algorithm in `RNAsat`; since `RNAsat` computes Q_k^T for each k, by sampling, we first compute the Boltzmann probability that a secondary structure has k base pairs, then we sample accordingly. The middle panel depicts the dotplot of base pair frequencies from 1000 samples of approximately locally optimal secondary structures, produced by a greedy algorithm suggested by one of the referees. The greedy algorithm first samples a low energy structure using McCaskill's algorithm, and subsequently adds that base pair which can be added which would lead to the greatest decrease in energy. Clearly, the dotplot in the left and middle panels are similar. In contrast, the right panel depicts base pair frequencies from 1000 sampled saturated secondary structures. It follows that approximately locally optimal structures, as determined by the greedy algorithm, resemble the

[8] This expected number could alternatively be obtained by summing over all base pairs the Boltzmann probability of base-pairing, as computed by McCaskill's algorithm – remark due to M. Zuker (personal communication).

[9] The consensus structure consists of those base pairs which appear in strictly greater than half the sampled structures.

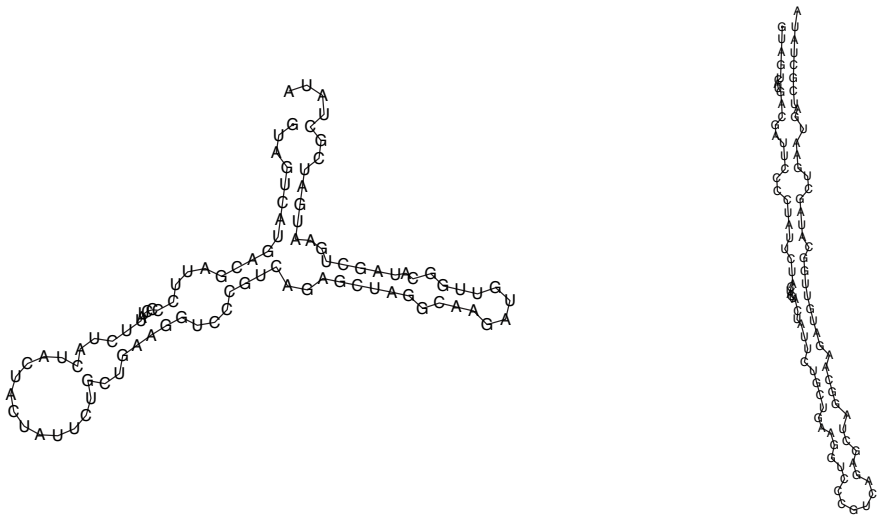

FIG. 4. Precursor microRNA `cel-mir-72` of length 96 from *Caenorhabditis elegans* with sequence GUAGUCAUGACGAUUCCCCUAUUCUACUACUACUAUUCUGCUGAAGGUCCCGUCAGAGCUAGGCAAGAU GUUGGCAUAGCUGAAUGAUCGCUAUA. [Data from Sanger Center miRBase linked by Rfam (Griffiths-Jones et al., 2003)]. **(Left)** Minimum free energy structure, as computed with RNAfold from Vienna RNA Package 1.5. **(Right)** Consensus structure from 1000 sampled saturated structures, using `RNAsat`. Here, the consensus structure consists of those base pairs occurring in strictly greater than half the sampled structures.

structures sampled by the method of Ding and Lawrence (2003). These appear to be distinct from sampled saturated structures.

In the set of experiments displayed in Figures 6–8 and Table 3, we considered the 26 nt. pseudoknot PK2 of the upstream pseudoknot domain (UPD) of the 3′ UTR of RNA beta, with EMBL accession number X03854 and start position 3088. This data is taken from the *Pseudobase* database (van Batenburg et al., 2001) with accession number PKB 00152. The RNA sequence and secondary structure, with annotated pseudoknot, is as follows:

```
UGGUGCCCAUCAACCAUAUGAUGGGA
.(((.[[[[[[[)))...]]]]]]].
```

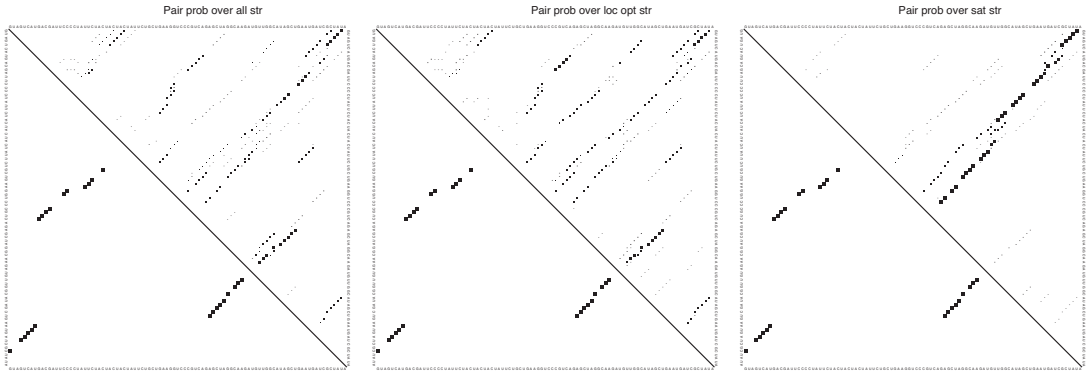

FIG. 5. Precursor microRNA `cel-mir-72` of length 96 from *Caenorhabditis elegans*. **(Left)** Dot plot of base pair frequencies from 1000 samples over all secondary structures. This is an implementation of McCaskill's (1990) algorithm within RNAsat) **(Middle)** Dot plot of base pair frequencies from 1000 samples of approximately locally optimal secondary structures, produced by a greedy algorithm described in the text. **(Right)** Dot plot of base pair frequencies from 1000 samples of saturated secondary structures produced by RNAsat. It is of interest that the Boltzmann centroid output by the Sfold (Ding et al., 2004) server is empty; i.e., the centroid has no base pairs.

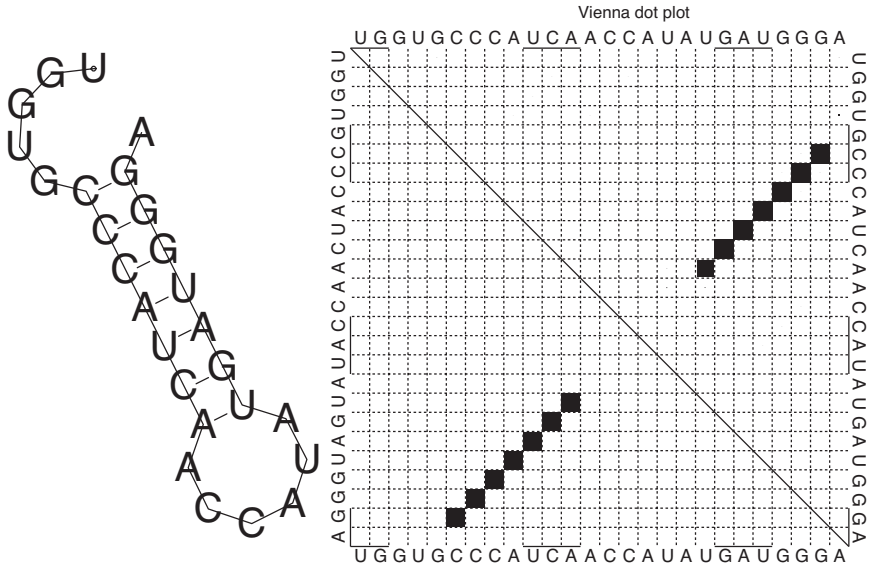

FIG . 6. Pseudoknot PK2 of the upstream pseudoknot domain (UPD) of the 3′ UTR of RNA beta, with Pseudobase (van Batenburg et al., 2001) number PKB 00152, EMBL accession number X03854 and start position 3088. The RNA sequence is UGGUGCCCAUCAACCAUAUGAUGGGA and the secondary structure with annotated pseudoknot is .(((.[[[[[[[)))...]]]]]]].. **(Left panel)** Minimum free energy secondary structure predicted by version 1.5 of Vienna RNA Package `RNAfold`. **(Right panel)** Boltzmann pair probabilities, as calculated using McCaskill's (1990) algorithm using `RNAfold -p`.

Figure 6 displays the minimum free energy (mfe) secondary structure and corresponding dot plot of Boltzmann pair probabilities, as computed using version 1.5 of Vienna RNA Package `RNAfold`. `RNAfold` correctly detects one of the helices, computed to have energy of $-10.5\,\mathrm{kcal/mol}$, as illustrated as follows:

```
UGGUGCCCAUCAACCAUAUGAUGGGA
.....(((((((.....)))))))..
```

The left panel of Figure 7 displays the dot plot of pair frequencies, computed by sampling 1000 structures using `RNAsat`. For each k, `RNAsat` computes the Boltzmann probability $Q_k^T / \sum_k Q_k^T$ of all secondary structures having k base pairs, and then samples appropriately. The same dot plot is produced by `Sfold` (Ding and Lawrence, 2003) (data not shown). Note that this dot plot is identical to the right panel of Figure 6. The right panel of Figure 7 displays the dot plot of pair frequencies computed by sampling 1000 locally optimal structures. All locally optimal structures were computed by brute force, by first enumerating *all* secondary structures by using Vienna RNA Package `RNAsubopt`, and subsequently checking whether the structure is locally optimal with respect to the Turner energy model; i.e., no neighbor has lower energy. This produces a set $LocOpt$ of all 780 locally optimal structures for PKB 00152. Note by contrast that there are 111,014 saturated structures of PKB 00152. We then sample 1000 locally optimal structures from $LocOpt$, each with probability $Pr[S] = exp(-E(S)/RT)/\sum_S exp(-E(S)/RT)$, where the sum is taken over $LocOpt$. Among these 1000 sampled structures, with few exceptions, the mfe structure occurs. In contrast, Table 3 presents data on the frequency of occurrence of saturated structures, when 1000 saturated structures are sampled from the collection of 111,014 saturated structures of PKB 00152.

The left panel of Figure 7 displays the dot plot of 1000 saturated structures sampled by `RNAsat`. Notice how alternative base pairs appear, corresponding to the pseudoknot, a feature not detected using mfe calculation, McCaskill's algorithm, or sampling over all secondary structures. The right panel of Figure 7 presents the histogram of energies of all sampled saturated structures.

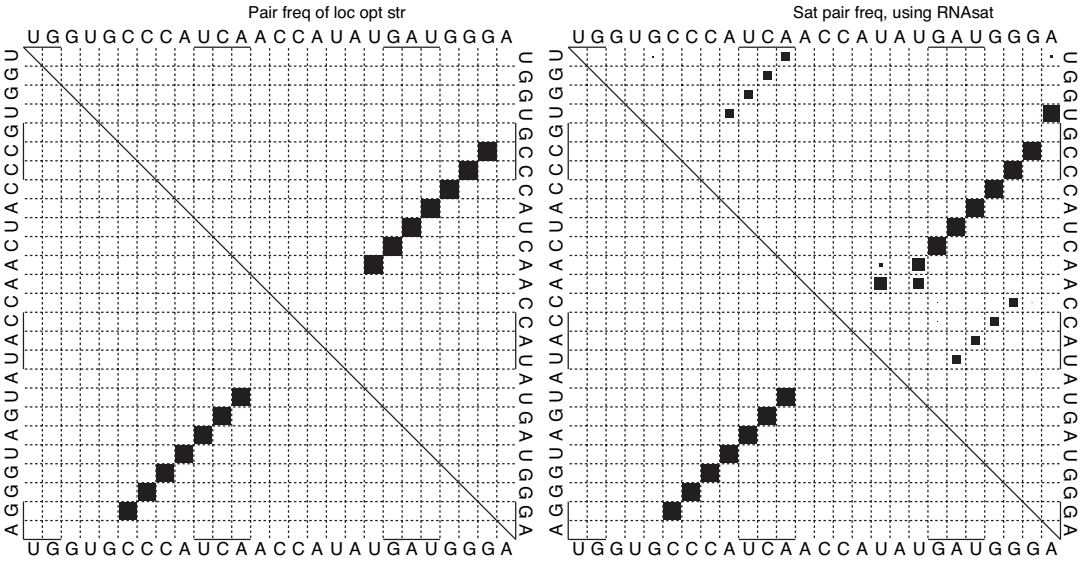

FIG . 7. Dotplots of pseudoknot PKB 00152 as described in the caption of Figure 6, where the lower triangular portion depicts the minimum free energy structure, while the upper triangular portion describes pair frequencies according to the experiment performed. **(Left panel)** Pair frequencies, computed by sampling 1000 locally optimal structures, where all locally optimal structures were computed brute force by exhaustive search (see text for details); subsequently structure S was sampled with probability $exp(-E(S)/RT)/\sum_S exp(-E(S)/RT)$, where the sum is taken over all locally optimal structures. Note that this dot plot is Identical to that of Figure 6, and is identical to the dot plot produced by `Sfold` (Ding et al., 2004). **(Right panel)** Pair frequencies, computed by sampling 1000 *saturated* structures using `RNAsat`. Notice how alternative base pairs appear, a feature not detected using mfe calculation, McCaskill's algorithm, or sampling over all secondary structures. This figure again illustrates the point that saturated structures are different in nature than locally optimal structures, which latter seem to be similar in nature to the Boltzmann low energy ensemble. One might speculate that this data suggests the existence of a folding funnel for RNA secondary structure.

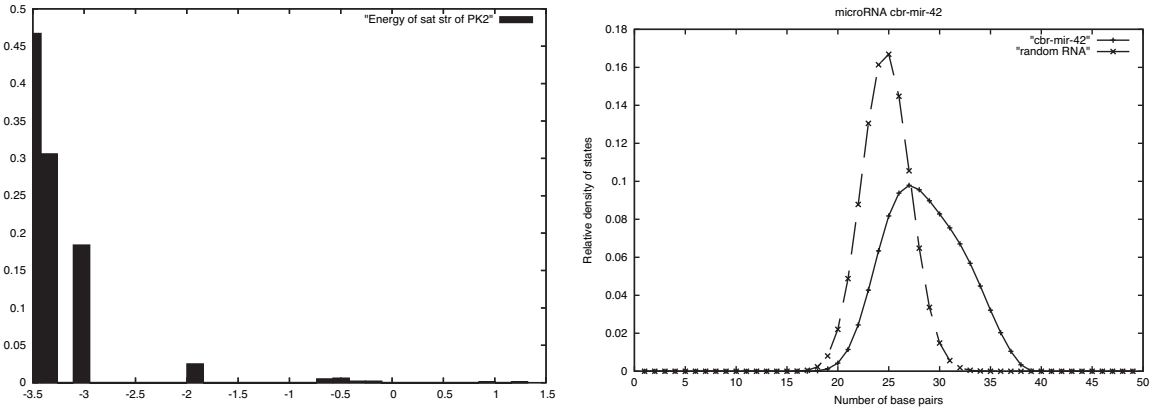

FIG . 8. **(Left panel)** Energy histogram for pseudoknot PKB 00152, the RNA sequence described in the caption of Figure 6. Histogram of energies of all saturated secondary structures. All saturated structures were determined brute force by exhaustive search. Note the energy variation for saturated structures. The histogram of energies of all locally optimal structures, as determined brute force by exhaustive search, is not shown because all locally optimal structures have the same energy as that of the mfe structure, i.e., -10.5 kcal/mol. **(Right panel)** Graph of the relative density of states $\rho(k) = Z_k/Z$ as a function of k, the number of base pairs, for precursor microRNA `cbr-mir-42` from *C. briggsae*. The lower curve with wider spread is that of `cbr-mir-42`, while the higher curve with more narrow spread is that of random RNA of the same length and dinucleotide frequency, as produced by our implementation of the Altschul-Erikson algorithm (Altschul and Erikson, 1985; Clote et al., 2005).

FIG . 9. Diagram for inductive extension of visibility predicates and overview of the three cases which arise in computing the partition function $Z_k^T(i,j,s,b)$ for saturated (s,b)-visible structures on the subsequence $a_i,\ldots a_j$ having k base pairs. The Feynman diagrams explaining the recursions are not different than those for McCaskill's (1990) algorithm, with the exception that the number of base pairs and visibility predicates *VisNuc, VisPos* depend in a precise manner on previously computed values which are stored in the dynamic programming table. For that reason, in this figure, we describe that dependence. The solid left rectangle in each of the four panels represents the set $s = VisNuc(S,i,j-1)$ of visible nucleotides in positions $i,\ldots,(j-1)-\theta-1$. The dotted left rectangle in the first panel represents the $s' = VisNuc(S,i,j)$ of visible nucleotides in positions $i,\ldots,j-\theta-1$. The solid right rectangle in each of the four panels represents the value $b = VisPos(S,i,j-1)$ equal to the number of visible positions $(j-1)-\theta,\ldots,j-1$ which occur to the right of the rightmost position basepaired in interval $[i,j-1]$. The dotted right rectangle in the first panel represents the value $b' = VisPos(S,i,j)$ equal to the number of visible positions $j-\theta,\ldots,j$ which occur to the right of the rightmost position basepaired in interval $[i,j]$. The black dot to the right of the rectangular regions represents the position j. By induction, we assume that $Z_\ell^T(i,j-1,u,v)$ has been computed for all values $\ell,u \subseteq \{A,C,G,U\}$ and $b \leq \theta$. We now describe how the contributions from (s',b')-visible structures on $a_i,\ldots,a_j-1$ can be extended to contributions from (s',b')-visible structures on $a_i,\ldots,a_j$. **(First panel)** Schematic view of how to update the visibility parameters inductively. **(Second panel)** In case $1,j$ basepairs with i, and the resulting structure is (s',b')-visible, where $s' = \theta, b' = 0$. Thus $Z_k^T(i,j,s',b')$ receives contributions from hairpin loops, bulges, internal loops and multi-loops closed by base pair (i,j) having k base pairs and visibility parameters (s,b). **(Third panel)** In case 2, j basepairs with intermediate r, for some $i < r \leq j-\theta-1$, and the resulting structure is (s',b')-visible, where $s' = Vis(S,i,r-1), b' = 0$. Thus $Z_k^T(i,j,s',b')$ receives contributions from hairpin loops, bulges, internal loops and multi-loops closed by base pair (i,j) having k base pairs and visibility parameters (s,b). **(Fourth panel)** In case 3, j does not basepair. In this case, the resulting structure is (s',b')-visible, where $s' = s \cup \{a_{j-\theta-1}\}$ if $VisPos(S) = \theta$, and otherwise $s' = s$. The new visible position $b' = \max(b,\theta)$. Thus $Z_k^T(i,j,s',b')$ receives contributions from hairpin loops, bulges, internal loops and multi-loops on $(i,j-1)$ having k base pairs and visibility parameters (s,b).

Shape frequencies of saturated and locally optimal structures. In Tables 1 and 2, we present shape frequencies for sampled structures of the SAM riboswitch with EMBL accession number AL939119.1/177986-178133. As defined by Giegerich and co-workers (Giegerich et al., 2004; Steffen et al., 2006; Voss et al., 2006), the *shape* of a given secondary structure is a homomorphic mapping from the secondary structure into a more compact structure, which succinctly describes basic features of the secondary structure (for the formal definition, see Giegerich et al., 2004). For instance, the π-shape of the cloverleaf secondary structure of tRNA is `[[][][]]`, while the less succinct π'-shape is `_[_[_]_[_]_[_]_]_`. By a simple linear time algorithm not described here, we can compute the π- and π'-shape of a given secondary structure; for reasons of space, we only present π-shapes in the data shown. Notice how the frequency of certain shapes for the SAM riboswitch depends on temperature (25°C, 37°C, 50°C), and on the number k of base pairs fewer than that of the Nussinov-Jacobson optimal structure for this riboswitch, which has 55 base pairs. Note the quite different shape frequencies of saturated structures versus general structures sampled by `RNAsat`, as displayed in Tables 1 and 2. Sample data is presented for different temperatures (25°C, 37°C and 50°C) and for different values of k, where k designates the number of base pairs fewer than that of the Nussinov-Jacobson optimal structure (in the case at hand, this number is 55). By "gen", we mean a general sampling, i.e., without any restriction on the value of k.

Collectively, the data for these examples suggest that there may in general be far fewer locally optimal structures than saturated structures, and that the ensemble of locally optimal structures resembles the Boltzmann ensemble of low energy structures, both quite different from the ensemble of saturated structures.

3. COMPUTING THE BOLTZMANN PARTITION FUNCTION Z_K^T

3.1. Recurrence relations for McCaskill's algorithm

McCaskill (1990) described a cubic time algorithm to compute the partition function $Q = \sum_S e^{-\beta E(S)}$, where the sum is over all secondary structures of a given RNA sequence $a_1,\ldots,a_n$. (Recall that we 'Z' to

TABLE 1. SHAPE FREQUENCIES FOR 100 SAMPLED STRUCTURES (WITH NO RESTRICTION TO BEING SATURATED) OF THE 149 NT

Shape	*Frequency*	*Temperature*	*k*
[[[ ][ ]][ ]]	0.580000	37	gen
[ ][ ][ ][ ]	0.180000	37	gen
[[ ][ ][ ]]	0.130000	37	gen
[ ][ ][ ]	0.090000	37	gen
[ ][ ][ ]	0.620000	37	10
[ ][ ][ ][ ]	0.260000	37	10
[ ][ ]	0.110000	37	10
[ ][ ][ ]	0.620000	25	10
[ ][ ][ ][ ]	0.260000	25	10
[ ][ ]	0.110000	25	10
[ ][ ][ ]	0.780000	50	10
[ ][ ][ ][ ]	0.170000	50	10
[ ][ ]	0.040000	50	10

SAM riboswitch with EMBL accession number AL939119.1/177986–178133. Shapes with low frequency are not displayed for reasons of space economy. Notice how the frequency of certain shapes depends on temperature ($25°\mathrm{C}, 37°\mathrm{C}, 50°\mathrm{C}$), and on the number k of base pairs fewer than that of the Nussinov-Jacobson optimal structure for this riboswitch, which has 55 base pairs. By "sat," we mean that there is no restriction on k; i.e., the sampling is general over all secondary structures, where the probability of sampling a k-saturated structure is $\widehat{Q}_k^T / \sum_k \widehat{Q}_k^T$. The shape frequencies for these structures should be compared with those for saturated structures, presented in Table 2.

denote the partition function for saturated structures, while 'Q' denotes the partition function without the restriction to saturated structures.) In particular, McCaskill defined

$$Q_{i,j} = \sum_S e^{-\beta E(S)} \tag{1}$$

where the sum in equation (1) is taken over all secondary structures S of $a[i,j] = a_i, \ldots, a_j$. Following Hofacker et al. (1994), let $Q_{i,j}^B$ denote the partition function restricted to those secondary structures on $a[i,j]$ in which (i,j) is a base pair. Similarly, $Q_{i,j}^M$ [resp. $Q_{i,j}^{M1}$] denotes the partition function for those secondary structures on $a[i,j]$ in which $a[i,j]$ is part of a multiloop having one or more [resp. exactly one] helix. With this notation, the recurrence relations for McCaskill's algorithm are given as follows.

$$\begin{aligned}
Q_{i,j} &= Q_{i,j-1} + \textstyle\sum_{i \le x < j} Q_{i,x}^B Q_{x+1,j} \\
Q_{i,j}^B &= e^{-\beta H(i,j)} + \textstyle\sum_{i<x<y<j} Q_{x,y}^B e^{-\beta I(i,j,x,y)} + \sum_{i<x<j} Q_{i+1,x}^M Q_{x+1,j-1}^{M1} e^{-\beta a} \\
Q_{i,j}^M &= \textstyle\sum_{i<x<j} e^{-\beta(x-i-1)c} Q_{x+1,j}^M + \sum_{i<x<j} Q_{i,x}^M Q_{x+1,j}^B e^{-\beta b} + Q_{i,j-1}^M e^{-\beta c} \\
Q_{i,j}^{M1} &= Q_{i,j-1}^{M1} e^{\beta c} + Q_{i,j}^B e^{\beta b} \\
Q_{i,i} &= 1, Q_{i,i}^B = 0, Q_{i,i}^M = 0, Q_{i,i}^{M1} = 0
\end{aligned}$$

Here, $H(i,j)$ is the free energy of a hairpin loop closed by base pair (i,j), and $I(i,j,x,y)$ is the free energy of an interior loop between base pair (i,j) and base pair (x,y). The free energy of a multiloop having B base pairs and U unpaired bases is approximated by the affine function $a + bB + cU$.

It is now straightforward to modify these recursion equations, in order to compute $Q_{i,j}(k)$, the partition function for all secondary structures on $a[i,j]$ having exactly k base pairs. We have the following.

$$\begin{aligned}
&Q_{i,j}(k) = Q_{i,j-1}(k) + \textstyle\sum_{1 \le \ell < k} Q_{i,x}^B(\ell) Q_{x+1,j}(k-\ell) \\
&Q_{i,j}^B(1) = e^{-\beta H(i,j)} \\
&Q_{i,j}^B(k+1) = \textstyle\sum_{i<x<y<j} Q_{x,y}^B(k) e^{-\beta I(i,j,x,y)} +
\end{aligned}$$

TABLE 2. SHAPE FREQUENCIES FOR 100 SAMPLED SATURATED STRUCTURES OF THE 149 NT

Shape	*Frequency*	*Temperature*	*k*
[[ ][ ]][[ ][ ]]	0.500000	37	gen
[[[ ][ ]][ ]]	0.320000	37	gen
[[ ][ ][ ]]	0.150000	37	gen
[[ ][ ][ ][ ]]	0.280000	37	10
[[[ ][ ]][ ][ ]]	0.160000	37	10
[[ ][ ][ ]]	0.120000	37	10
[[ ][[ ][ ][ ]]]	0.060000	37	10
[ ][[ ][ ]][ ][ ][ ]	0.050000	37	10
[ ][ ][ ][ ][ ]	0.050000	37	10
[[ ][ ][ ][ ]]	0.200000	25	10
[[[ ][ ]][ ][ ]]	0.180000	25	10
[[ ][ ][ ]]	0.140000	25	10
[[ ][ ][ ][ ]]	0.290000	50	10
[ ][ ][ ][ ][ ]	0.150000	50	10
[[ ][ ][ ]]	0.100000	50	10
[[[ ][ ]][ ][ ][ ]]	0.060000	50	10
[[[ ][ ]][ ][ ]]	0.060000	50	10
[ ][ ][ ][ ]	0.050000	50	10

SAM riboswitch with EMBL accession number AL939119.1/177986-178133. Shapes with low frequency are not displayed for reasons of space economy. Notice how the frequency of certain shapes depends on temperature ($25°\mathrm{C}, 37°\mathrm{C}, 50°\mathrm{C}$), and on the number *k* of base pairs fewer than that of the Nussinov-Jacobson optimal structure for this riboswitch, which has 55 base pairs. By "sat," we mean that there is no restriction on *k*; i.e., the sampling is general over all saturated structures, where the probability of sampling a *k*-saturated structure is $\widehat{Z}_k^T / \sum_k \widehat{Z}_k^T$. The shape frequencies for these saturated structures should be compared with those for general structures, presented in Table 1.

TABLE 3. SATURATED STRUCTURES FOR PSEUDOKNOT PKB 00152

secStr	*Frequency*	*Energy*
...(.(((((((((...).))))))))	0.455000	-3.500000
...(.(((((((.(.....))))))))	0.312000	-3.300000
((((....)))).((((...))))..	0.194000	-3.000000
...(.(((((((....)..)))))))	0.022000	-1.900000
(....(((((((((...).))))))))	0.005000	-0.600000
(....(((((((.(.....))))))))	0.005000	-0.400000
(...)(((((((((...).))))))).	0.002000	-0.300000
(...)(((((((.(.....)))))))).	0.002000	-0.100000
(....(((((((....)..)))))))	0.001000	1.000000
...(.((((((...(.....))))))))	0.001000	0.400000
...(.(.(((((.(.....))))))).)	0.001000	1.400000

One thousand saturated structures were sampled from the collection of all 111,014 saturated structures for the RNA sequence UGGUGCCCAUCAAC CAUAUGAUGGGA with secondary structure and annotated pseudoknot .(((.[[[[[[[)))...]]]]]]].. For each structure, the frequency and energy in kcal/mol are additionally indicated.

$$
\begin{aligned}
&\textstyle\sum_{1\le\ell<k}\sum_{i<x<j} Q^M_{i+1,x}(\ell)Q^{M1}_{x+1,j-1}(k+1-\ell)e^{-\beta a}\\
Q^M_{i,j}(k) = &\textstyle\sum_{i<x<j} e^{-\beta(x-i-1)c}Q^M_{x+1,j}(k) +\\
&\textstyle\sum_{1\le\ell<k}\sum_{i<x<j} Q^M_{i,x}(\ell)Q^B_{x+1,j}(k-\ell)e^{-\beta b} +\\
&Q^M_{i,j-1}(k)e^{-\beta c}\\
Q^{M1}_{i,j}(k) = &\; Q^{M1}_{i,j-1}(k)e^{\beta c} + Q^B_{i,j}(k)e^{\beta b}\\
Q_{i,i}(k) = 1, &\; Q^B_{i,i}(k) = 0, Q^M_{i,i}(k) = 0, Q^{M1}(k)_{i,i} = 0
\end{aligned}
$$

The difficulty of adding visibility predicates s, b, necessary to compute the partition function $Z_{i,j}(k)$ for saturated structures on $a[i, j]$ having exactly k base pairs can be illustrated in an attempt to define the multiloop partition function $Z^M_{i,j}(k, s, b)$ for k base pairs and visibility predicates s, b. The term

$$\sum_{i<x<j} e^{-\beta(x-i-1)c} Z^M_{x+1,j}(k, s, b)$$

contributes to the partition function $Z^M_{i,j}(k, s \cup \{a_{i+1}, \ldots, a_x\}, b)$, but there are other terms making it difficult to express how $Z^M_{i,j}(k, s, b)$ depends on previously computed terms $Z^M_{i,x}(\ell, s_0, b_0)$, $Z^B_{x+1,j}(k-\ell, s_1, b_1)$, etc. For this reason, after giving an overview of `RNAsat`, we present the pseudocode in Figures 10–14.

3.2. Notation

Given an RNA sequence $a_1, \ldots, a_n$ and fixed absolute temperature T, Boltzmann partition function values are computed using dynamic programming, where (essentially) two-dimensional arrays Z_s, Z'_s, Z^m_s, Z_m, Z_e are progressively filled along off-diagonals at increasing distance $d = 1, 2, \ldots$ from the principal diagonal. In this manner, when defining the value $Z[i, j]$, for $1 \le i < j \le n$, where Z collectively denotes any of Z_s, Z'_s, Z^m_s, Z_m, Z_e, the values of $Z[i', j']$ are already stored for all $1 \le i' < j' \le n$ such that

```
1.  for d = θ to n − 1
2.    for i = 0 to n − 1 {
3.      j = i + d // skip if j > n
4.      jCanBasePair = FALSE
5.      for r = i to j − θ − 1 {
6.        if r, j base pair {
7.          if r = i // CASE1: j base pair with i
8.            if segmentBasePair(i + 1, j − 1) = false // (i,j) is a hairpin
9.              Z'_s(i, j, 1) = e^{−hairpin(i,j)/RT}
10.             Z^m_s(i, j, 1, ?, 0) = Z_s(i, j, 1, ?, 0) = Z'_s(i, j, 1) · e^{−dangle(i,j)/RT}
11.           for all 1 < nBP ≤ maxBP(i, j)
12.             for all s ? P({A, C, G, U}) // (i,j) close a multi-loop
13.               for all 0 ≤ b ≤ θ + 1
14.                 Z'_s(i, j, nBP) += Z_m(i + 1, j − 1, nBP − 1, s, b) · e^{−(α1+α2)/RT} · e^{−dangle(j,i)/RT}
15.           for x = i + 1 to j − θ − 2 // stack, bulge or internal loop
16.             if segmentBasePair(i + 1, x − 1) = true then break
17.             for y = j − 1 downto x + θ + 1
18.               if segmentBasePair(i + 1, x − 1, y + 1, j − 1) = true then break
19.               Z'_s(i, j, nBP) += Z'_s(x, y, nBP − 1) · e^{−stack(i,x,y,j)/RT}
20.           Z^m_s(i, j, nBP, ?, 0) = Z_s(i, j, nBP, ?, 0) = Z'_s(i, j, nBP) · e^{−dangle(i,j)/RT}
21.         if i < r < j // CASE2: j base pair with intermediate r
22.           for all 1 ≤ nBP ≤ maxBP(i, j)
23.             if segmentBasePair(i, r − 1) = false
24.               s = nucl(i, r − 1)
25.               Z_s(i, j, nBP, s, 0) += Z_s(r, j, nBP, ?, 0)
26.               Z^m_s(i, j, nBP, s, 0) += Z_s(r, j, nBP, ?, 0) · e^{−(r−1−i)α3/RT}
27.             for all 0 ≤ n1 < maxBP(i, r − 1)
28.               n2 = nBP − n1
29.               for all s ? P({A, C, G, U})
30.                 for all 0 ≤ b ≤ θ + 1
31.                   s0 = s ? nucl(r − 1 − b − 1, r − 1)
32.                   Z_e(i, j, nBP, s0, 0) += Z_s(i, r − 1, n1, s, b) · Z_s(r, j, n2, ?, 0)
33.                   Z_e(i, j, nBP, s0, 0) += Z_e(i, r − 1, n1) · Z_s(r, j, n2, ?, 0)
34.                   Z_m(i, j, nBP, s0, 0) += Z^m_s(i, r − 1, n1) · Z_s(r, j, n2, ?, 0) · e^{−2α2/RT}
35.                   Z_m(i, j, nBP, s0, 0) += Z_m(i, r − 1, n1) · Z_s(r, j, n2, ?, 0) · e^{−α2/RT}
36.       }
37.     }
```

FIG . 10. Algorithm to compute the Boltzmann partition function for saturated secondary structure having k base pairs, with respect to the Turner energy model.

```
38.       for all 1 < nBP ≤ maxBP(i, j)
39.         for all s ? P({A, C, G, U})
40.           if basepairWithSet(s, j) = false {
41.             for all 0 ≤ b < θ // general case
42.               Ze(i, j, nBP, s, b+ 1) += Ze(i, j − 1, nBP, s, b)
43.               Zm(i, j, nBP, s, b+ 1) += Zm(i, j − 1, nBP, s, b) · e^(−α3/RT)
44.               Zs(i, j, nBP, s, b+ 1) += Zs(i, j − 1, nBP, s, b)
45.               Z^m_s(i, j, nBP, s, b+ 1) += Z^m_s(i, j − 1, nBP, s, b) · e^(−α3/RT)
46.             if canBasePair(j − θ − 1, j) = false // special case
47.               s0 = s ? nucl(j − θ − 1, j − θ − 1)
48.               Ze(i, j, nBP, s0, θ+ 1) += Ze(i, j − 1, nBP, s, θ+ 1)
49.               Zm(i, j, nBP, s0, θ+ 1) += Zm(i, j − 1, nBP, s, θ+ 1) · e^(−α3/RT)
50.               Zs(i, j, nBP, s0, θ+ 1) += Zs(i, j − 1, nBP, s, θ+ 1)
51.               Z^m_s(i, j, nBP, s0, θ+ 1) += Z^m_s(i, j − 1, nBP, s, θ+ 1) · e^(−α3/RT)
52.           }
53.     }
```

FIG. 11. Continuation of algorithm from Figure 10, to compute the Boltzmann partition function for saturated secondary structure having k base pairs, with respect to the Turner energy model.

```
0.  samplingHelix(k,i,j,sample)
1.  va = random()
2.  current = 0
3.  addBasePair(i,j,sample)
4.  if k = 0 return // no more base pairs to add; end of sampling
5.  for x = i + 1 to j − θ − 1
6.    for y = x + θ + 1 to j
7.      if canBasePair(x, y) = true AND segmentBasepair(i, x, y, j) = false
8.        current += (stack(i,x,y,j) · Z'_s(x,y,k − 1)) / Z'_s(i,j,k)
9.        if va < current
10.         samplingHelix(k-1,x,y,sample)   // (i,j) belongs to a helix
11.         return
12. for s ? P({A, C, G, U})
13.   for b = 0 to θ + 1
14.     current += (e^((−dangle(j,i))/RT) · e^((−α1−α2)/RT) · Zm(i+1,j−1,k−1,s,b)) / Z'_s(i,j,k)
15.     if va < current
16.       samplingMultiLoop(k-1,i+1,j-1,s,b,sample,0)
            //(i,j) closes a multi-loop
17.       return
```

FIG. 12. Sampling algorithm for helix. This function is called only after having determined that i, j can base-pair.

$(j' - i') < (j - i)$. As described in Section 2.1, additional *visibility parameters* s, b are required to inductively account for all saturated structures. Each of s, b ranges over finitely many values; indeed $s \subseteq \{A, C, G, U\}$ and $0 \leq b \leq \theta + 1 = 4$. It follows that each of the arrays Z_s, Z'_s, Z^m_s, Z_m, Z_e is not two-dimensional as initially described, but rather of the form $Z[k, i, j, s, b]$. Note that in the particular case $s = \emptyset$ and $b = 0$, the secondary structures S considered in $Z_s(k, i, j, \emptyset, 0)$ are such that $(i, j) \in S$; i.e. i and j base-pair together in S.

The pseudocode description of the partition function algorithm in Figures 10 and 11 requires some explanation of notation used. The function $maxBP(i, j)$ returns the maximum number of base pairs possible in a secondary structure on $a_i, \ldots, a_j$, i.e. equal to the number of base pairs in the Nussinov-Jacobson optimal structure on $a_i, \ldots, a_j$. The function $nucl(i, j)$ returns the set $\{a_i, \ldots, a_j\}$; note that $nucl(i, j) \subseteq \{A, C, G, U\}$ is a set, not a multi-set or list. The function, *basepairWithSet(s,i)* returns `true` if the nucleotide a_i at position i can base-pair with a nucleotide in the set s, and `false` otherwise. This function is invoked when s is a set of visible nucleotides (i.e. external to any base pair).

Energy contributions for different types of loops (helix, bulge, interior loop and multi-loop) are defined as follows. For $1 \leq i < j \leq n$, $hairpin(i, j)$ denotes the energy contribution of a hairpin loop on RNA subsequence $a_i, \ldots, a_j$; i.e. a_i and a_j form a closing base pair of the loop region $a_{i+1}, \ldots, a_{j-1}$, where the latter are not base-paired. The energy contribution for a stacked base pair, bulge or interior loop is denoted

```
0.  samplingMultiLoop(k,i,j,s,b,sample,lastHelix)
1.  va = random()
2.  current = 0
3.  if b = θ + 1 // previous b can take two values
4.    for b' = θ to θ + 1
5.      current += e^(-α3/RT) · Zm(i,j-1,k,s,b') / Zm(i,j,k,s,b)
6.      if va < current
7.        samplingMultiLoop(k,i,j-1,s,b',sample,lastHelix)
8.        return
9.      s' = remove(j - θ - 1, s)
10.     if b' = θ + 1 AND s' ≠ s
11.       current += e^(-α3/RT) · Zm(i,j-1,k,s',b') / Zm(i,j,k,s,b)
12.       samplingSingleLoop(k,i,j-1,s',b',sample,lastHelix)
13.       return
14. else if b > 0
16.   current += e^(-α3/RT) · Zm(i,j-1,k,s,b') / Zm(i,j,k,s,b)
17.   if va < current
18.     samplingMultiLoop(k,i,j-1,s,b-1,sample,lastHelix)
19.     return
20. else // b = 0
21.   for r = i to j - θ - 1
22.     if canBasePair(r, n) = true
23.       if lastHelix=false AND segmentBasePair(i, r - 1) = true
24.         for s' ? P(s)
25.           for b' = 0 to θ + 1
26.             for k1 = 0 to k - 1
27.               k2 = k - k1
28.               current += e^(-α2/RT) · Zm(i,r-1,k1,s',b') · e^(-dangle(r,j)/RT) · Z's(r,j,k2) / Zm(i,j,k,s,b)
29.               if va < current // multi-loop sampling is NOTterminated
30.                 samplingMultiLoop( k1,i,r-1,s',b',sample,false)
31.                 samplingHelix( k2,r,j,sample)
32.                 return
33.               current += e^(-2·α2/RT) · Z^m_s(i,r-1,k1,s',b') · e^(-dangle(r,j)/RT) · Z's(r,j,k2) / Zm(i,j,k,s,b)
34.               if va < current // sampling last two helices in multi-loop
35.                 samplingMultiLoop( k1,i,r-1,s',b',sample,true)
36.                 samplingHelix( k2,r,j,sample)
37.                 return
38.       else if lastHelix=true AND compatible(i, r - 1, s) = true
39.         current += e^(-(α2 + α3·(r - i))/RT) · e^(-dangle(r,j)/RT) · Z's(r,j,k) / Zm(i,j,k,s,b)
40.         if va < current
41.           samplingHelix(k,r,j,sample)
42.           return
```

FIG . 13. Sampling algorithm for multi-loop.

by $stack(i_1, i_2, j_1, j_2)$, where (i_1, j_2) is the exterior closing base pair, and (i_2, j_1) is the interior closing base pair. Finally, the energy associated with a multi-loop is computed by the affine function $\alpha_1 + i \cdot \alpha_2 + n \cdot \alpha_3$, where i is the number of base pairs in the multi-loop and n the number of unpaired nucleotides. Both $5'$ and $3'$ dangles are taken into account by our algorithm, and we denote the energy contribution by $dangle(i, j)$. Specifically, if $i < j$, then the function $dangle(i, j)$ returns the energy of the $3'$ dangle of $i + 1$ on i plus the energy of the $5'$ dangle of $j - 1$ on j. Additionally, if $i < j$, then by convention, the function $dangle(j, i)$ returns the energy of the $5'$ dangle of $i - 1$ on i plus the energy of the $3'$ dangle of $j + 1$ on j. Note that energy values of $3'$ dangles are higher than those of $5'$ dangles. Our manner of handling dangles is similar to that of Hofacker et al. (1994) in `RNAfold`.

Although a single array Z suffices to compute the partition function with respect to the Nussinov-Jacobson energy model, this is no longer the case for the Turner model. In the latter case, for each k, i, j, s, b, we must distinguish the partition function according to the last secondary structure element encountered (i.e. hairpin, stem, bulge, internal loop, multi-loop or exterior loop). Each of these 5 conditions requires a distinct array Z.

- Z_s: Secondary structures such that the first and the last paired nucleotides base-pair together. An example is `...(***)..`, where `.` denotes an unpaired position, `***` denotes any valid substructure. The subscript s in Z_s suggests "stem".

```
0.  samplingExteriorLoop(k,i,j,s,b,sample)
1.  va = random()
2.  current = 0
3.  if b = θ + 1 // previous b can take two values
4.    for b' = θ to θ + 1
5.      current += Z_e(i,j-1,k,s,b') / Z_e(i,j,k,s,b)
6.      if va < current
7.        samplingExteriorLoop(k,i,j-1,s,b',sample)
8.        return
9.      s' = remove(j - θ - 1, s)
10.     if b' = θ + 1 AND s' ≠ s
11.       current += Z_e(i,j-1,k,s',b') / Z_e(i,j,k,s,b)
12.       samplingSingleLoop(k,i,j-1,s',b',sample)
13.       return
14. else if b > 0
16.   current += Z_e(i,j-1,k,s,b') / Z_e(i,j,k,s,b)
17.   if va < current
18.     samplingExteriorLoop(k,i,j-1,s,b-1,sample)
19.     return
20. else // b = 0
21.   for r = i to j - θ - 1
22.     if canBasePair(r, j) = true
23.       if segmentBasePair(i, r - 1) = true
24.         for s' ? P(s)
25.           for b' = 0 to θ + 1
26.             for k1 = 0 to k - 1
27.               k2 = k - k1
28.               current += Z'_e(i,r-1,k1,s',b') · e^(-dangle(r,j)/RT) · Z'_s(r,j,k2) / Z_e(i,j,k,s,b)
29.               if va < current
30.                 samplingExteriorLoop( k1,i,r-1,s',b',sample,false)
31.                 samplingHelix( k2,r,j,sample)
32.                 return
33.       else if compatible(i, r - 1, s) = true
34.           current += e^(-dangle(r,j)/RT) · Z'_s(r,j,k) / Z_e(i,j,k,s,b)
35.           if va < current
36.             samplingHelix(k,r,j,sample)
37.             return
```

FIG. 14. Sampling algorithm for exterior loop.

- Z'_s: Secondary structures such that the leftmost and rightmost nucleotides of the sub-sequence base pair together. Fields s and b are irrelevant in this case and so omitted. The well-balanced parenthesis expression with dots `(***)` is an example.
- Z_e: At least 2 hairpin loops occur in an exterior loop (unpaired nucleotides do not receive a penalty since they are external to any base pair). An example of this type of structure is given by `...(***)..(***)...`
- Z_s^m: Same as Z_s, except that a penalty for unpaired bases occurring in a multi-loop is added for each nucleotide occurring outside the stem. An example of this type of structure is given by `...(..(***).(***)...)`.
- Z_m: At least 2 hairpin loops appear in a multi-loop. In this case,a penalty is added for unpaired bases outside each stem.

It may seem that the case Z'_s is included in the case Z_s; however, values stored in $Z_s(i, j, k, \emptyset, 0)$ differ from those of $Z'_s(i, j, k)$ because of dangle energy contributions (see algorithm description).

In order to reduce space complexity, we use the function $segmentBasePair(i_1, j_1, i_2, j_2)$ which returns `true` if one can form a valid base pair for the segments $[i_1, j_1]$ and $[i_2, j_2]$ and `false` otherwise. More formally, $segmentBasePair(i_1, j_1, i_2, j_2)$ holds exactly when $(\exists x, y \in \{i_1, \ldots, j_1\} \cup \{i_2, \ldots, j_2\})$ such that $x < y$ and x, y can form a valid Watson-Crick or wobble pair.

The results of the $segmentBasePair$ function can be computed in $\mathcal{O}(1)$ time using a two-dimensional array $nbSingleBp$ which stores in a cell $nbSingleBp[i, j]$ the number of single base pairs (i.e. the number of secondary structures with one and only one base pair) over the subsequence $a_i, \cdots, a_j$. Indeed, one can

note that the number of basepairs than can be formed over the two intervals $[i_1, j_1]$ and $[i_2, j_2]$ is equal to $nbSingleBp[i_1, j_2] - nbSingleBp[i_1, i_2] - nbSingleBp[j_1, j_2] + nbSingleBp[j_1, i_2]$. Using this remark, we can compute this value and return `true` if and only if the result is positive (false otherwise). Note that when only two index are given, the function $segmentBasePair$ operates for a single interval $[i_1, j_1]$. The same array $nbSingleBp$ can be used for this task.

3.3. Algorithm description

When considering substring $a_i, \ldots, a_j$ of the input RNA sequence, the partition function algorithm, given in Figures 10 and 11, treats 3 distinct cases, according to whether i, j are base-paired, or r, j are base-paired for some intermediate $i < r \leq j - \theta - 1$, or j is not base-paired. CASE 1, treated in lines 7 to 20, considers all saturated secondary structures containing base pair (i, j), while CASE 2, treated in lines 21 to 36, considers those containing base pair (r, j), for some intermediate r, $i < r \leq j - \theta - 1$. CASE 3, treated in lines 38 to 52, considers those saturated secondary structures in which j does not base-pair with any position in $\{i, \ldots, j - \theta - 1\}$ – note that it can happen (later) that j base-pairs with a nucleotide in $\{1, \ldots, i - 1\}$ or $\{j + \theta + 1, \ldots, n\}$. This latter case is the most difficult, and indeed is the reason for having introduced the visibility parameters s, b in the first place.

In each block of of pseudocode, several possibilities are discussed. In Case 1, we consider the occurrence of a hairpin (lines 9–11), the closing of a multi-loop (lines 12–14) and any stacked base pairs, bulges and internal loops (lines 15–19). The `break` statements at lines 16 and 18 prevent useless loops. Indeed, if the predicate $segmentBasePair(i, x, y, j)$ is `true`, then $segmentBasePair(i, x', y', j)$ is also `true` for all $x \geq x'$ and $y' \leq y$. In Case 2, extension with unpaired bases on the left is examined in lines 23 to 26, while concatenation of helices appearing in a multi-loop or an exterior loop is handled in lines 27–36. Note that in this case lines 32–33 correspond to an initialization of the helical sequence, while lines 34–35 add a new stem to an already existent helical sequence. In Case 3, before any extension with an unpaired nucleotide on the right, we distinguish the "general case" (lines 41–45), such that $0 \leq b \leq \theta$, from the special case (lines 46–51), where $b = \theta + 1$. Any extension requires us to check that the rightmost nucleotide cannot already base-pair, before forming a base pair with the rightmost position; i.e. that $(j - \theta - 1, j)$ cannot base-pair.

3.4. Complexity

As earlier mentioned, s and b are of constant size (s can take at most 16 values, and b can take at most 5 values). The partition function algorithm given in Figures 10 and 11 runs in time $\mathcal{O}(n^5)$ with space requirement of $\mathcal{O}(n^3)$. In practice, our implementation requires roughly 1 minute and 250 MByte to run for an input RNA sequence of 100 nucleotides.

4. BOLTZMANN SAMPLING OF SATURATED SECONDARY STRUCTURES

The Boltzmann partition function can be used to weight each secondary structure with respect to its energy contribution in the ensemble of all possible secondary structures. In analogy to Ding and Lawrence (2003), who gave the first correct and mathematically rigorous sampling algorithm for RNA structures, we proceed as follows.

4.1. Notation

Here we define additional functions required for the sampling algorithm, depicted in Figures 12 and 13. The function $remove(i, s)$, where $i \in \mathbb{N}$ and $s \subseteq \{A, U, G, C\}$, returns the subset $s \setminus \{a_i\}$ of s which does not contain the nucleotide at position i. The function $compatible(i, j, s)$ returns `true` if the set of all nucleotides present in $a_i, \ldots, a_j$ equals s, and returns `false` otherwise. The power set function $\mathcal{P}(s)$ returns the collection of all subsets of s, and function `addBasePair(i,j,sample)` adds the base pair (i, j) to a secondary structure (called $sample$ below). All other functions have been previously defined. The

Z-tables, i.e., Z_e, Z_m, etc., used in our sampling algorithm are the same as those used in our partition function algorithm. Additionally, in the case of exterior loop, we define a new array Z'_e such that $Z'_e(i,j,k,s,b) = Z_s(i,j,k,s,b) + Z_e(i,j,k,s,b)$, for all i,j,k,s and b. This array stores the values of the partition function for helical sequences belonging to an exterior loop, regardless of the number of helical sequences (one or more instead of at least two).

4.2. Algorithm description

The quadratic time sampling algorithm works by calling itself recursively. We begin with an empty secondary structure, denoted $sample$, and we stipulate the number k of base pairs that the sample should ultimately contain. Since the Boltzmann probability at temperature T for the collection of saturated secondary structures having k base pairs equals the relative density of states $\rho_k^T = Z_k^T / \sum_k Z_k^T$, the value of k is obtained by sampling. Moreover, $Z_k^T = \sum_{s,b} Z_k^T[k,1,n,s,b]$, and hence initial values of s, b are obtained by sampling —say s_0, b_0. Given the visibility constraint s_0, b_0, in the sample to be produced, either $(1,n) \in sample$, or $(r,n) \in sample$ for some intermediate $1 < r \leq n - \theta - 1$, or n is not base-paired in $sample$. Each of these alternatives determines visibility parameters, which must then agree with the initially determined values s_0, b_0. The structure $sample$ is then progressively filled by recursive calls of the sampling algorithm. At the end of the sampling algorithm, the completed structure $sample$ having k base pairs is returned.

Subfunctions are designed to sample a specific (or a set of specific) secondary structure elements and to decide which ones will be called recursively. Three such subfunctions are needed. They are respectively associated with the sampling of base pairs in helices (Fig. 12), decomposition of helix sequences in multi-loop (Fig. 13) and decomposition of helix sequences in exterior loop (Fig. 14). All functions follow the same approach. First, a random number in the real interval $(0,1)$ is generated according to the uniform distribution (line 1 in Figs. 12 and 13) and then this random number is used together with (previously computed, stored) partition function values to choose the next case according to the Boltzmann probability.

The sampling function `samplingHelix` (Fig. 12) is the only function to write a base pair in the sample structure (line 3). The recursive call in the case of a stem structure is developed in lines 5–10, and the occurrence of a multi-loop is discussed in lines 12–17.

The function `samplingMultiLoop` (Fig. 13) discusses the value of b and its possible ancestors. In lines 3–13, we assume that the upper bound for b is reached. The distance d from the rightmost nucleotide to the rightmost base-paired nucleotide is thus *at least* $\theta + 1$. We consider two possible values for its ancestor, namely $\theta + 1$ ($d > \theta + 1$) and θ ($d = \theta + 1$). When $1 \leq b \leq \theta$ (lines 14–19) this discussion does not pertain since there is only one possibility (i.e., $b - 1$).

When $b = 0$ (lines 20–36), the rightmost base is paired. This means that we need to extract the rightmost helix from the helical sequence of the multi-loop. Thus, we search for a valid cut point r (lines 21 and 22) and recursively call the helix sampling algorithm on the rightmost helix and recursively call the multi-loop sampling algorithm on the left part. Once the cut point r has been chosen, we take care of the left string status (line 23). If this segment can base-pair (lines 24–37), then the left part contains *at least* one helix and hence the function calls itself recursively on this segment while `samplingHelix` is called on the right part. Lines 28–32 treat the case of a left part which contains at least two helices. This implies that the multi-loop sampling is not terminated and then field `minipage` keeps the `false` value for the recursive sampling (line 30). In lines 33–37, we treat the case of left and right parts containing one and only one helix. Otherwise (lines 38–42), the left string cannot base-pair and the helix is the last one in the multi-loop. Before calling `samplingHelix` on the right side and terminating the sampling of the multi-loop we must check the correctness of this configuration (line 38). Hence we check first that more than one stem will appear in the multi-loop and next that the argument s is compatible with the visibility set of the left string.

When sampling an exterior loop (Fig. 14), the principles are identical to those already described for function `samplingMultiLoop`. We need only to omit the energy contributions from dangle and to use array Z_s, Z_e, and Z'_e instead of Z_s^m and Z_m. This implies a simplification when sampling an exterior loop that contains at least two helices (lines 28–32).

4.3. Correctness of the sampling algorithm

To prove the correctness of the sampling algorithm, we need to demonstrate that a saturated structure can be decomposed into smaller saturated structures. This is done in the following Lemma 4.3. In this section, we

consider a fixed RNA sequence $a_1, \ldots, a_n$, and use the notation $[i, j]$ to denote the interval $\{i, i+1, \ldots, j\}$ of integral values between i and j.

Lemma 1. *Assume that $S_{i,j}$ is a saturated secondary structure on RNA subsequence $a_i, \ldots, a_j$, and that $S_{i,j}$ has k base pairs. Then exactly one of the following cases holds.*

1. *Nucleotides i and j base-pair and enclose a saturated secondary structure with $k-1$ base pairs.*
2. *There exists an index $i < r \leq j - \theta - 1$ such that $(r, j) \in S_{i,j}$, and there exist two saturated secondary structures $S_{i,r-1}$ and $S_{r,j}$ having respectively k_1 and k_2 base pairs on $a_i, \ldots, a_{r-1}$ and $a_r, \ldots, a_j$ respectively, where $k = k_1 + k_2 + 1$ and $S_{i,j} = S_{i,r-1} \cup S_{r,j-1} \cup \{(r, j)\}$.*
3. *Position j) is not base-paired and the secondary structure $S_{i,j-1}$ restricted to $a_i, \ldots, a_{j-1}$ is saturated.*

Proof. Assume that nucleotides i and j base-pair with each other, and consider the secondary structure $S_{i+1,j-1}$ restricted to $a_{i+1}, \ldots, a_{j-1}$. If $S_{i+1,j-1}$ is not saturated, then for some $i < x < y < j$, the base pair (x, y) can be added to $S_{i+1,j-1}$. But then $S_{i,j} \cup \{(x, y)\}$ would be a valid secondary structure, contradicting local optimality of $S_{i,j}$.

Assume now that nucleotides i and j do not base-pair together. Then either the nucleotide at position j base-pairs with a nucleotide at position $i < r \leq j - \theta - 1$, or position j is unpaired with any position in the interval $[i, j]$. In the former case, $S_{i,j}$ can be decomposed into two distinct secondary structures $S_{i,r-1}$ and $S_{r,j-1}$ along with the base pair (r, j). If either of the substructures is not saturated, then for some x, y satisfying $i \leq x < y \leq r-1$ resp. $r \leq x < y \leq j-1$, the base pair (x, y) could be added to $S_{i,r-1}$ resp. $S_{r,j-1}$. But then $S_{i,j} \cup \{(x, y)\}$ would be a valid secondary structure, contradicting local optimality of $S_{i,j}$. In the latter case, if for $i \leq x < y \leq j-1$ the base pair (x, y) could be added to $S_{i,j-1}$, without violating the definition of secondary structure, then since $S_{i,j} = S_{i,j-1}$, the same is true for $S_{i,j}$, which contradicts local optimality of $S_{i,j}$. This concludes the proof. ■

Lemma 1 states that every saturated secondary structure can be decomposed into smaller saturated secondary structures. themselves saturated. This is the justification for a recursive sampling algorithm for saturated structures. While `RNAsat` executes correct sampling with respect to the Turner energy model, for readability, we present the proof of correctness for the simpler Nussinov-Jacobson model.[10] The following proposition gives the Boltzmann probabilities for each configuration described in Lemma 1.

Proposition 1. *Let $a_i, \ldots, a_j$ be a subword from the RNA sequence $a_1, \ldots, a_n$. Let $bp(i, j)$ denote the Nussinov-Jacobson energy associated with base pair (i, j); in particular, if a_i, a_j do not form a Watson-Crick or wobble pair, then $bp(i, j) = +\infty$. For all possible positive numbers of base pairs $0 < k \leq maxBP(1, n)$ and all possible visibility parameters s and b, we have the following.*

$$Z(k, i, j, s, b) = e^{-\frac{bp(i,j)}{RT}} \sum_{s', b'} Z(k-1, i+1, j-1, s', b') \tag{2}$$

$$+ \sum_{comp.\ k', r, s', b'} Z(k', i, r-1, s', b') \cdot e^{-\frac{bp(r,j)}{RT}} \sum_{s'', b''} Z(k-k'-1, r+1, j-1, s'', b'') \tag{3}$$

$$+ \sum_{comp.\ s', b'} Z(k, i, j-1, s', b'). \tag{4}$$

In the terms (3) and (4), the sum is taken over all values s', b' which are compatible with values s, b.

Proof. The proof is by simultaneous induction on k, s, b. By definition, we have

$$Z(k, i, j, s, b) = \sum_{S_{i,j}\ is\ loc.\ opt.} e^{\frac{-E(S_{i,j})}{RT}}$$

[10] This is similar to the presentation style of Hofacker et al. (2004), where details are given for the Nussinov-Jacobson model, although the authors implemented their algorithm for the Turner energy model.

where S is constrained by parameters k, s and b. Depending on the status of position j (base-paired or not), there are three cases. *(i)* (i,j) is a base pair, *(ii)* (r,j) is a base pair, for some $i < r \leq j-\theta-1$, *(iii)* j is not base-paired with any position in the interval $[i,j]$. This yields the following equation, where 's.t.' abbreviates "such that", and 'bp' abbreviates "base-pair".

$$Z(k,i,j,s,b) = \sum_{S_{i,j} \ s.t. \ (i,j) \ bp} e^{\frac{-E(S_{i,j})}{RT}} + \sum_{S_{i,j} \ s.t. \ (r,j) \ bp} e^{\frac{-E(S_{i,j})}{RT}} + \sum_{S_{i,j} \ s.t. \ j \ not \ bp} e^{\frac{-E(S_{i,j})}{RT}}$$

CASE 1. (i,j) is a base pair of $S_{i,j}$. Let $S_{i+1,j-1}$ denote the restriction of $S_{i,j}$ to $a_{i+1},\ldots,a_{j-1}$, obtained after removal of the extremal base pair (i,j). The energy $E(S_{i,j})$ of a secondary structure $S_{i,j}$ on $a_i,\ldots,a_j$ such that $(i,j) \in S_{i,j}$ can be decomposed as $E(S_{i,j}) = bp(i,j) + E(S_{i+1,j-1})$. Hence the first sum can be written as

$$\begin{aligned}\sum_{S_{i,j} \ s.t. \ (i,j) \ bp} e^{\frac{-E(S_{i,j})}{RT}} &= \sum_{S_{i+1,j-1}} e^{\frac{-(bp(i,j)+E(S_{i+1,j-1}))}{RT}} \\ &= e^{\frac{-bp(i,j)}{RT}} \sum_{S_{i+1,j-1}} e^{\frac{-E(S_{i+1,j-1})}{RT}}\end{aligned}$$

By Case 1 of Lemma 4.3.we know that $S_{i+1,j-1}$ is saturated. We then cluster such secondary structures $S_{i+1,j-1}$, each having $k-1$ base pairs, according to their visibility parameters s' and b'.

$$\begin{aligned}\sum_{S_{i,j} \ s.t. \ (i,j) \ bp} e^{\frac{-E(S_{i,j})}{RT}} &= e^{\frac{-bp(i,j)}{RT}} \sum_{s',b'} \sum_{S_{i+1,j-1}(s',b')} e^{\frac{-E(S_{i+1,j-1})}{RT}} \\ &= e^{\frac{-bp(i,j)}{RT}} \sum_{s',b'} Z(k-1,i+1,j-1,s',b')\end{aligned}$$

This completes the proof for CASE 2 of proposition 4.3.

CASE 2. For some intermediate position r, with $i < r \leq j-\theta-1$ $(r,j) \in S_{i,j}$. By Case 2 of Lemma 1, the substructures $S_{i,r-1}$ and $S_{r+1,j-1}$ are saturated. The energy $E(S_{i,j})$ of a secondary structure on $a_i,\ldots,a_j$, such that $(r,j) \in S_{i,j}$, can be decomposed as $E(S_{i,j}) = bp(r,j) + E(S_{i,r-1}) + E(S_{r+1,j-1})$.

$$\begin{aligned}\sum_{S_{i,j} \ s.t. \ (r,j) \ bp} e^{\frac{-E(S_{i,j})}{RT}} &= e^{\frac{-bp(j,r)}{RT}} \cdot \sum_{S_{i,r-1},S_{r+1,j-1}} e^{\frac{-(E(S_{i,r-1})+E(S_{r+1,j-1}))}{RT}} \\ &= e^{\frac{-bp(j,r)}{RT}} \cdot \sum_{S_{i,r-1}} e^{\frac{-(E(S_{i,r-1}))}{RT}} \cdot \sum_{S_{r+1,j-1}} e^{\frac{-(E(S_{r+1,j-1}))}{RT}}\end{aligned}$$

Let k' denote the number of base pairs of $S_{i,r-1}$, and s' and b' denote the visibility parameters of $S_{i,r-1}$. Note that s', b' must be compatible with s, b; i.e. since (r,j) is a base pair of $S_{i,j}$, it must be that $b = 0$ and $s = s' \cup \{a_{r-1-m} : 0 \leq m < b'\}$. Cluster the left substructures $S_{i,r-1}$ by (compatible) visibility parameters s', b'. By Lemma 1, the left substructure $S_{i,r-1}$ is saturated, hence the sum can be rewritten as follows.

$$\begin{aligned}\sum_{S_{i,j} \ s.t. \ (r,j) \ bp} e^{\frac{-E(S_{i,j})}{RT}} &= e^{\frac{-bp(r,j)}{RT}} \cdot \sum_{comp. \ k',r,s',b'} \sum_{S_{i,r-1}(k',r,s',b')} e^{\frac{-(E(S_{i,r-1}(k',r,s',b')))}{RT}} \cdot \\ \sum_{S_{r+1,j-1}} e^{\frac{-(E(S_{r+1,j-1}))}{RT}} &= \sum_{comp. \ k',r,s',b'} Z(k',r-1,s',b') \cdot \sum_{S_{r+1,j-1}} e^{\frac{-(E(S_{r+1,j-1}))}{RT}}\end{aligned}$$

Since (r,j) is a base pair, by induction we can apply Case 2 of this proposition to the right substructure $S_{r,j-1}$. This concludes the proof of Case 3.

$$\sum_{S_{i,j} \ s.t. \ (r,j) \ bp} e^{\frac{-E(S_{i,j})}{RT}} = \sum_{comp. \ k',r,s',b'} Z(k',r-1,s',b') \cdot e^{\frac{-bp(i,j)}{RT}} \cdot \sum_{s'',b''} Z(k_2-1,i+1,j-1,s'',b'').$$

CASE 3. In this case, the rightmost position j is not base-paired with any position in the interval $[i, j]$. Hence, $E(S_{i,j}) = E(S_{i,j-1})$, where $S_{i,j-1}$ denotes the secondary structure restricted $a_i, \ldots, a_{j-1}$.

$$\sum_{S_{i,j} \ s.t. \ j \ not \ bp} e^{\frac{-E(S_{i,j})}{RT}} = \sum_{S_{i,j-1}} e^{\frac{-E(S_{i,j-1})}{RT}}$$

Let s' and b' denote the visibility parameters of $S_{i,j-1}$. Note that these parameters must be compatible with s and b; i.e. since j is unpaired in $S_{i,j}$, it must be that $b = \max(b' + 1, \theta + 1)$, and $s = s' \cup \{a_{j-\theta-1}\}$ if $b' = \theta + 1$ while $s = s'$ if $0 \leq b' \leq \theta$. (This is the reason for line 10 in Figures 13 and 14.) Cluster the secondary structures $S_{i,j-1}$ by these compatible parameters s', b'. Lemma 4.3 ensures that these secondary structures are saturated, hence by induction we have the following.

$$\begin{aligned}\sum_{S_{i,j} \ s.t. \ j \ not \ bp} e^{\frac{-E(S_{i,j})}{RT}} &= \sum_{comp. \ s',b'} \sum_{S_{i,j-1}(s',b')} e^{\frac{-E(S_{i,j-1}(s',b'))}{RT}} \\ &= \sum_{comp. \ s',b'} Z(k, i, j-1, s', b')\end{aligned}$$

This concludes the proof of proposition 1. ■

Proposition 4.3 implies the correctness of the sampling algorithm for locally optimal secondary structures with respect to the Nussinov-Jacobson energy model. This explanation should suffice to explain the underlying ideas in the pseudocode for sampling with respect to the Turner energy model, given in Figures 12–14.

ACKNOWLEDGMENTS

We would like to thank Michael Zuker for remarks and for the suggestion to use `mfold` 2.3 energy parameters for our temperature-dependent computations. We are indebted to two anonymous referees for their helpful remarks and criticisms, and to Yann Ponty for an algorithmic suggestion. The work of both authors was partially supported by the NSF (grant DBI-0543506 to P.C.).

REFERENCES

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

Banerjee, A.R., Jaeger, J.A. and Turner, D.H. 1993. Thermal unfolding of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry* 32:153–163.

Barrick, J.E., Corbino, K.A. Winkler, W.C. et al. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA* 101:6421–6426.

Bernhart, S.H., Tafer, H. Mückstein, U. et al. 2006. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol.* 1.

Böck, A., Forschhammer, K. Heider, J. et al. 1991. Selenoprotein synthesis: an expansion of the genetic code. *Trends Biochem. Sci.* 16:463–467.

Clote, P. 2005a. An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. *J. Comput. Biol.* 12:83–101.

Clote, P. 2005b. RNALOSS: a web server for RNA locally optimal secondary structures. *Nucleic Acids Res.* 33:W600–W604.

Clote, P. 2006. Combinatorics of saturated secondary structures of RNA. *J. Comput. Biol.* 13:1640–1657.

Clote, P., and Backofen, R. 2000. *Computational Molecular Biology: An Introduction*. John Wiley & Sons. New York.

Clote, P., Ferré, F. Kranakis, E. et al. 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* 11:578–591.

Dill, K.A., and Bromberg, S. 2002. *Molecular Driving Forces*. Garland Science, New York.

Dimitrov, R.A., and Zuker, M. 2004. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.* 87:215–226.

Ding, Y., Chan, C.Y. and Lawrence, C.E. 2004. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.* 32: W135–W141.

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

Eddy, S.R. 2004. How do RNA folding algorithms work? *Nat. Biotechnol.* 22:1457–1458.

Evers, D.J., and Giegerich, R. 2001. Reducing the conformation space in RNA structure prediction. *German Conf. Bioinform*.

Flamm, C., Hofacker, I.L. Stadler, P.F. et al. 2002. Barrier trees of degenerate landscapes. *Z. Phys. Chem.* 216:155–173.

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

Griffiths-Jones, S., Bateman, A. Marshall, M. et al. 2003. Rfam: an RNA family database. *Nucleic Acids Res.* 31:439–441.

Havgaard, J.H., Lyngsø, R. Stormo, G. et al. 2005. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics* 21:1815–1824.

Heider, J., Baron, C. and Böck, A. 1992. Coding from a distance dissection of the mRNA elements required for the incorporation of selenocysteine into protein. *EMBO J.* 11:3759–3766.

Hofacker, I.L., Priwitzer, B. and Stadler, P.F. 2004. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics* 20:186–190.

Tonoco, Jr, I., Borer, P.N. Dengler, B. et al. 1973. Improved estimation of secondary structure in ribonucleic acids. *Nat. New Biol.* 246:40–41.

Hofacker, I.L., Fontana, W. Stadler, P.F. et al. 1994. Fast folding and comparison of RNA secondary structures. *Monatsch. Chem.* 125:167–188.

Jaeger, J.A., Zuker, M. and Turner, D.H. 1990. Abstract melting and chemical modification of a cyclized self-splicing group I intron: similarity of structures in 1 M Na^{+}, in 10 mM Mg^{2+}, and in the presence of substrate. *Biochemistry* 29:10147–10158.

Lim, L.P., Glasner, M.E. Yekta, S. et al. 2003. Vertebrate microRNA genes. *Science* 299:1540.

Markham, N., and Zuker, M. 2005a. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* 33:W577–W581.

Markham, N.R., and Zuker, M. 2005b. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* 33:W577–W581.

Mathews, D.H. (2004). Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* 10:1178–1190.

Mathews, D.H., and Turner, D.H. 2002. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.* 317:191–203.

Mathews, D.H., Turner, D.H. and Zuker, M. 2000. Secondary structure prediction. In: Beaucage, S. Bergstrom, D.E. Glick, G.D. et al. editors, *Current Protocols in Nucleic Acid Chemistry*. John Wiley & Sons, New York, 11.2.1–11.2.10.

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

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

Moon, S., Byun, Y. Kim, H.-J. et al. 2004. Predicting genes expressed via -1 and $+1$ frameshifts. *Nucleic Acids Res.* 32:4884–4892.

Mückstein, U., Tafer, H. Hackermuller, J. et al. 2006. Thermodynamics of RNA-RNA binding. *Bioinformatics* 22:1177–1182.

Nussinov, R., and Jacobson, A. B. 1980. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proc. Natl. Acad. Sci. USA* 77:6309–6313.

Schattner, P., Decatur, W.A. Davis, C.A. Jr, et al. 2004. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.* 32:4281–4296.

Steffen, P., Voss, B. Rehmsmeier, M. et al. 2006. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 22:500–503.

Tuschl, T. 2003. Functional genomics: RNA sets the standard. *Nature* 421:220–221.

van Batenburg, F.H., Gultyaev, A.P. and Pleij, C.W. 2001. Pseudobase: structural information on rna pseudoknots. *Nucleic Acids Res.* 29(1):194–195.

Voss, B., Giegerich, R. and Rehmsmeier, M. 2006. Complete probabilistic analysis of RNA shapes. *BMC Biol.* 4.

Voss, B., Meyer, C. and Giegerich R. 2004. Evaluating the predictability of conformational switching in RNA. *Bioinformatics* 20:1573–1582.

Washietl, S., Hofacker, I.L. Lukasser, M. et al. 2005a. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* 23:1383–1390.

Washietl, S., Hofacker, I.L. and Stadler, P.F. 2005b. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* 102:2454–2459.

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

Xia, T., Jr., SantaLucia, J. Burkard, M.E. et al. 1999. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* 37:14719–14735.

Zheng, M., Wu, M. and I. Tinoco, Jr. 2001. Formation of a GNRA tetraloop in P5abc can disrupt an interdomain interaction in the Tetrahymena group I ribozyme. *Proc. Natl. Acad. Sci. USA* 98:3695–3700.

Zimm, B.H., and Bragg, J.K. 1959. Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* 31:526–535.

Zuker, M. 1986. RNA folding prediction: the continued need for interaction between biologists and mathematicians. *Lect. Math. Life Sci.* 17:87–124.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406–3415.

Zuker, M., and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9:133–148.

Address reprint requests to:
Dr. Rainer Sachs
970 Evans Hall
Berkeley, OA 94720

E-mail: sachs@math.berkeley.edu