

Thermodynamics of RNA structures by Wang-Landau sampling

Feng Lou¹, Peter Clote^{1,2,3,*}

¹Laboratoire de Recherche en Informatique (LRI), Université Paris-Sud XI, bât. 490, 91405 Orsay cedex, France.

²Department of Biology, Boston College, Chestnut Hill, MA 02467, USA.

³Digiteo Chair, Laboratoire d'Informatique (LIX), Ecole Polytechnique, 91128 Palaiseau, France.

Received on Jan 8, 2010; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

ABSTRACT

Motivation: Thermodynamics-based dynamic programming RNA secondary structure algorithms have been of immense importance in molecular biology, where applications range from the detection of novel selenoproteins using EST data, to the determination of microRNA genes and their targets. Dynamic programming algorithms have been developed to compute the minimum free energy secondary structure and partition function of a given RNA sequence, the minimum free energy and partition function for the hybridization of two RNA molecules, etc. However, the applicability of dynamic programming methods depends on disallowing certain types of interactions (pseudoknots, zig-zags, etc.), as their inclusion renders structure prediction an NP-complete problem. Nevertheless, such interactions have been observed in X-ray structures.

Results: A non-Boltzmannian Monte Carlo algorithm was designed by Wang and Landau to estimate the density of states for complex systems, such as the Ising model, that exhibit a phase transition. In this paper, we apply the Wang-Landau (WL) method to compute the density of states for secondary structures of a given RNA sequence, and for hybridizations of two RNA sequences. Our method is shown to be much faster than existent software, such as RNAsubopt. From density of states, we compute the partition function over all secondary structures and over all pseudoknot-free hybridizations. The advantage of the WL method is that by adding a function to evaluate the free energy of arbitrary pseudoknotted structures and of arbitrary hybridizations, we can estimate thermodynamic parameters for situations known to be NP-complete. This extension to pseudoknots will be made in the sequel to this paper; in contrast, the current paper describes the Wang-Landau algorithm applied to pseudoknot-free secondary structures and hybridizations.

Availability: The Wang-Landau RNA hybridization web server is under construction at <http://bioinformatics.bc.edu/clotelab/>.

Contact: clote@bc.edu

1 INTRODUCTION

RNA is an important biomolecule, now known to play both an *information carrying* role, as well as a *catalytic* role. Indeed, the

genomic information of retroviruses, such as the hepatitis C and human immunodeficiency viruses, is encoded by RNA rather than DNA, while the peptidyl transferase reaction, arguably the most important enzymatic reaction responsible for life, is catalyzed not by a protein, but rather by RNA [56]. It has recently emerged that RNA plays a wide range of previously unsuspected roles in many biological processes, including *retranslation* of the genetic code (selenocysteine insertion [6], ribosomal frameshift [3]), transcriptional and translational gene regulation [34, 36], temperature sensitive conformational switches [13, 52], chemical modification of specific nucleotides in the ribosome [45], regulation of alternative splicing [9], etc.

A secondary structure for a given RNA nucleotide sequence a_1, \dots, a_n is a set S of base pairs (i, j) , such that a_i, a_j forms either a Watson-Crick or GU (wobble) base pair, and such that there are no *base triples* or *pseudoknots* in S .¹ For example, the secondary structure of Y RNA² with EMBL access code AAPY01489510/220-119 is displayed in panels (a,b) of Figure 1, while panels (c,d) of the same figure depict the pseudoknotted structure of the Gag/pro ribosomal frameshift site of mouse mammary tumor virus [53]. In conventional dot-bracket notation, this latter structure is given as follows, where it should be noted that two kinds of bracket are needed due to the pseudoknot:

```
AAAAAACUUGUAAAGGGGCAGUCCCCUAGCCCCGCUCAAAAGGGGGAUG
.....((((([[[[[[.))))).....]]]]]]].
```

It is computationally intractable to compute the minimum free energy tertiary structure of RNA; indeed, determining the optimal pseudoknotted structure is NP-complete [35]. In contrast, by disallowing pseudoknots, secondary structure prediction is algorithmically tractable; there are dynamic programming algorithms to compute the minimum free energy structure for a single RNA molecule, as well as for the hybridization of two or more RNA molecules. In particular, such methods can be loosely grouped into two types of algorithm – those that use (i) a *stochastic context free grammar*

¹ A base triple in S consists of two base pairs $(i, j), (i, \ell) \in S$ or $(i, j), (k, j) \in S$. A pseudoknot in S consists of two base pairs $(i, j), (k, \ell) \in S$ with $i < k < j < \ell$.

² According to [48], one of the functions of Y RNA is to bind to certain misfolded RNAs, including 5S rRNA, as part of a quality control mechanism.

*to whom correspondence should be addressed

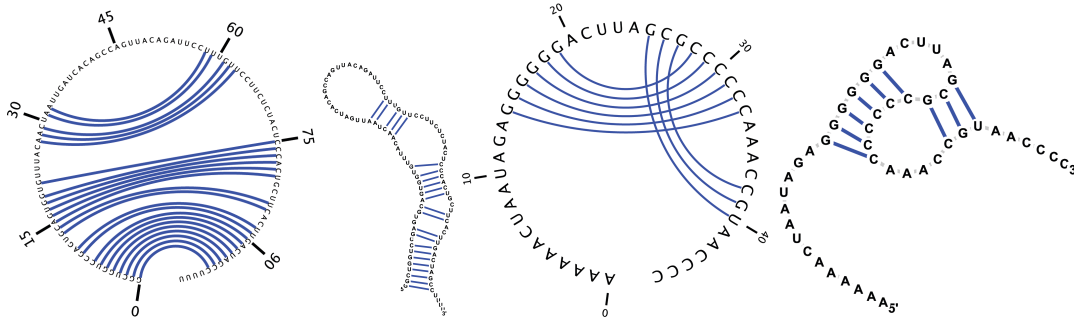


Fig. 1. (a,b) Pseudoknot-free secondary structure of Y RNA with EMBL access code AAPY01489510/220-119, depicted in panel (a) in Feynman circular form, and in panel (b) in classical form. (c,d) Pseudoknotted structure for the Gag/pro ribosomal frameshift site of mouse mammary tumor virus, depicted in panel (c) in Feynman circular form, and in panel (d) in classical form. Images produced with software jviz [57] from structures taken respectively from Rfam [25] and Pseudobase [53].

to compute a covariation model, and (ii) those that use free energy parameters obtained from UV absorbance (optical melting) experiments, in order to determine the minimum free energy structure (i.e. thermodynamic-based algorithms). Examples of stochastic context-free grammars are the programs *Infernal* [43] and *Pfold* [30]. Examples of thermodynamics-based algorithms are the programs *mfold* [63], *UNAFOLD* [37], *RNAfold* [26], *RNAstructure* [39]. Thermodynamics-based algorithms for hybridization of two structures are given in *UNAFOLD* [19], *RNAcofold* [42, 4], while the *NUPACK* software considers hybridization of three or more RNA molecules. [20]. Such thermodynamics-based algorithms are particularly important, since the tertiary structure of RNA is believed to be largely determined by secondary structure, which acts as a scaffold for tertiary contacts; see [2] for experimental data supporting this view.³ Computing the minimum free energy pseudoknotted structure for a given RNA sequence is NP-complete [35] for the Turner nearest neighbor energy model.⁴ For that reason, pseudoknot structure prediction algorithms fall into three categories: (i) exponential time *exact* algorithms, (ii) dynamic programming algorithms that restrict pseudoknots to a particular class, (iii) heuristic methods. Examples of exact algorithms for pseudoknot structure prediction are the branch-and-bound algorithm of [7] and the method using tree-width decomposition of [62]. Examples of algorithms that consider only pseudoknots of a particular class are found in the pioneering work of [50] and [33], with subsequent refinements in [21, 47, 49]. Examples of heuristic approaches include Monte Carlo methods [41], genetic algorithms [1], and a simple, yet elegant algorithm called *ProbKnot* (D.H. Mathews, to appear) that appears to be the state-of-the-art method according to recent benchmarking studies. Finally, it is beyond the scope of this paper to provide additional background on algorithms for RNA structural alignment, motif detection, or tertiary structure prediction.

As will be shown later, by Wang-Landau (WL) Monte Carlo methods, we can obtain essentially the same results as by dynamic programming computation of the partition function from *UNAFOLD*

and *RNAcofold*; however, the advantage of the WL approach is that by extending the energy evaluation function for a given structure or hybridization, we can estimate the partition function for arbitrary pseudoknotted structures, known to be an NP-complete problem.

Before proceeding, we formally define a secondary structure as follows. Given an RNA sequence $s = a_1, \dots, a_n$, a secondary structure S on s 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 $i = k$.

For steric reasons, 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, see the text [14] and the recent review [22].

2 APPROACH

The non-Boltzmannian WL Monte Carlo algorithm was developed by Wang and Landau [55, 54] to estimate the density of states and partition function for complex systems, such as the Ising model, that exhibit a phase transition. While the Metropolis-Hastings Monte Carlo algorithm samples low energy states, the Wang-Landau algorithm is designed to visit states uniformly across all energies in a discrete energy landscape. Indeed, for the Metropolis-Hastings algorithm, the expected frequency, or *stationary probability*, $p_{mc}^*(x)$ of visiting the state x , whose energy is E , is given by the uniform probability $\frac{1}{g(E)}$ times the Boltzmann probability $p_{mc}^*(x) = \frac{e^{-E/RT}}{Z}$, where $g(E)$ is the number of states having energy E , and the partition function $Z = \sum_z e^{-E(z)/RT}$; in contrast, for the WL

³ There is some controversy about the extent to which RNA secondary structure constrains the tertiary structure. See [12] for more on this point.

⁴ The minimum energy pseudoknotted structure can be computed by maximum weight matching in $O(n^3)$ time for the simple Nussinov energy model [51].

algorithm, the expected frequency, or stationary probability, of visiting state x is $p_{wl}^*(x) = \frac{1}{g(E) \cdot \mathcal{E}}$, where \mathcal{E} is the total number of distinct energies E (in the discrete case), or of energy bins (in the continuous case). It follows that non-Boltzmannian sampling strategies, such as that devised by Wang and Landau [55, 54], Kou and Wong [31], etc. are potentially useful in biopolymer folding, where one searches for a global energy minimum in a landscape having many local energy minima. Indeed in [10], Chen and Xu applied the WL algorithm for the structure prediction of helical transmembrane proteins, while the equi-energy sampling method of Kou and Wong [31], related to Monte Carlo with replica exchange, has been applied to estimate the density of states for lattice protein folding under the hydrophobic-hydrophilic (HP) energy model [32], as well as in protein structure prediction by fragment assembly [61].

In this paper, we apply the WL algorithm to compute the density of states and partition function for RNA secondary structure as well as for the hybridization of two RNA sequences. We begin by validating and benchmarking the WL method against the exhaustive method RNAsubopt [58], that enumerates all secondary structures of a given RNA sequence. Next, we compute the partition function over all secondary structures and over all pseudoknot-free hybridizations. We describe as well how to compute the partition function $Z(T)$ over all temperatures from 0° to 100° Celsius by performing two Wang-Landau computations, followed by convolution calculations. Although the computation of the partition function over all secondary structures and over all pseudoknot-free hybridizations can be done using existent software RNAfold [27] resp. RNAcofold [4], UNAFold [37], and a recently published method of Chitsaz et al. [11], the real advantage of our method is that by adding a function to evaluate arbitrary pseudoknotted structures and arbitrary hybridizations, we can approximately compute the partition function, heat capacity, melting temperature, etc. for a context known to be NP-complete [35].

The *density of states* is defined to be the *absolute frequency* function for energy; i.e. density of states $g(e)$ counts the number of states having energy e . In the context of RNA secondary structure, a *state* is a secondary structure for an arbitrary but fixed RNA sequence s . In [16], Cupal et al. described the first efficient algorithm, running in $O(m^2 n^3)$ time, to compute the density of states for an RNA sequence of length n , where energy is discretized into m bins. The program of Cupal et al. is no longer available, since it has been superseded by the program RNAsubopt, developed by Wuchty et al. [58], which enumerates all secondary structures, whose free energy is within a user-defined bound above the minimum free energy. Though not documented, the RNAsubopt program additionally admits the option -D, which, instead of outputting structures, outputs only the number of secondary structures in each energy bin above the minimum free energy (bin size is 0.1 kcal/mol).

3 METHODS

Monte Carlo algorithms have been implemented by a number of groups, in order to study RNA kinetics of folding. In particular, KinFold, developed by Flamm et al. [23], computes the *mean first passage time* (MFPT) of folding, by using a variant of the Gillespie algorithm in an event-driven simulation with a choice of Metropolis-Hastings and Kawasaki dynamics. In [28, 59], a similar time-driven Monte Carlo simulation program, KineFold, is described to compute kinetically determined pseudoknotted structure for a given RNA sequence. Danilova et al. [17] describe

```

1. procedure Metropolis-Hastings( )
2.    $T = T_{hi}$ 
3.    $x = \text{initial state}$ 
4.   while ( $T > T_{lo}$ ) {
5.     repeat M times {
6.       choose random neighbor  $y \in N_x$ 
7.       if ( $E(x) \leq E(y)$ ) then
8.          $x = y$ 
9.       else
10.        choose random  $z \in (0, 1)$ 
11.        if ( $z < \frac{e^{-E(y)/RT} / N_x}{e^{-E(x)/RT} / N_y}$ ) then  $x = y$ 
12.      }
13.     $T = T * 0.9$ 
14.  }
15.  return  $x$ 

```

Fig. 2. Pseudocode for Metropolis-Hastings algorithm with simulated annealing [29].

```

1. procedure WangLandau(s)
2.    $S = \emptyset$  // empty initial structure
3.    $c = \exp(1)$  // initial modification factor
4.   while  $c > 1 + \epsilon$  {
5.     for all energies bins  $e$ :  $g(e) = 1$ 
6.     while  $h$  is not flat {
7.       for  $i = 1$  to NumSteps
8.         choose random  $T \in \mathcal{N}(S)$  of  $S$ 
9.          $e_0 = \text{bin}(E(S))$ ;  $e_1 = \text{bin}(E(T))$ 
10.        choose random  $z \in (0, 1)$ 
11.        if  $z < \frac{g(e_0)}{g(e_1)}$ 
12.           $S = T$ 
13.           $e = e_1$ 
14.        else //  $S$  remains unchanged
15.           $e = e_0$ 
16.           $g(e) = c \cdot g(e)$  // update d.o.s.
17.           $h(e) = h(e) + 1$  // update histogram
18.        }
19.       $c = \sqrt{c}$  // reduce modification factor
20.    }
21.  return relative density of states  $g$ ,
  where  $g(i) = g(i) / \sum_j g(j)$ 

```

Fig. 3. Pseudocode for Wang-Landau algorithm, as applied to RNA secondary structure density of states computation. In line 8, $\mathcal{N}(S)$ denotes the collection of immediate neighbors of structure S ; i.e. those obtained by adding or removing a single base pair. In line 16, d.o.s. abbreviates density of states.

the RNAkinetics web server used to study the kinetics of the folding transitions of a growing RNA molecule, as in the case of transcriptional folding.

We now begin by providing background definitions and describing the Wang-Landau algorithm.

WANG-LANDAU

The WL algorithm [55, 54] was designed in order to compute the *density of states* and *partition function*, neither of which can be computed directly by classical Monte Carlo methods, such as the Metropolis-Hastings algorithm, simulated annealing, replica exchange, etc.

Recall the definition of Markov chain. Let $Q = \{1, \dots, n\}$ be a finite set of states, let $\pi = (p_1, \dots, p_n)$ be the distribution for initial state, and let $P = (p_{i,j})$ be a matrix of transition probabilities, satisfying $\sum_j p_{i,j} = 1$ for all i . A (first-order, time-homogeneous) *Markov chain* $M = (Q, \pi, P)$ is a stochastic process, whose state q_t at time t is a random variable determined by

$$\begin{aligned} \Pr[q_0 = i] &= \pi_i, \\ \Pr[q_{t+1} = j | q_t = i] &= p_{i,j}. \end{aligned}$$

Define $p_i(t) = \Pr[q_t = i]$ and $p_{i,j}^{(t)} = \Pr[q_t = j | q_0 = i]$. Clearly, the (i, j) th entry of the t th power P^t of P equals $p_{i,j}^{(t)}$; moreover, by time-homogeneity it follows that $p_{i,j}^{(t)} = \Pr[q_{t_0+t} = j | q_{t_0} = i]$, for all t_0 . The *stationary probability* of state i is defined by $\lim_t p_i(t) = p_i^*$, provided the limit exists. It is a classical result that every finite, aperiodic, irreducible Markov chain has an *equilibrium* distribution of *stationary probabilities*; see the text of Clote and Backofen [14] for a new, self-contained proof of this result. A Markov chain with state set Q and stationary probabilities p_1^*, \dots, p_n^* is *reversible*, if for all $i, j \in Q$, $p_i^* p_{i,j} = p_j^* p_{j,i}$.

Figure 2 presents pseudocode for the classical Metropolis-Hastings Monte Carlo algorithm with simulated annealing [40, 29], which implements a random walk on the Markov chain whose transition probabilities $p_{i,j}$ of moving from state x_i to x_j is given by

$$\begin{aligned} p_{i,j} &= P(x_i \rightarrow x_j) = \min \left(1, \frac{\exp(-E(x_j)/RT)/Z}{\exp(-E(x_i)/RT)/Z} \cdot \frac{\mathcal{N}(x_j)}{\mathcal{N}(x_i)} \right) \\ &= \min \left(1, \frac{\exp(\frac{-(E(x_j)-E(x_i))}{RT})}{\mathcal{N}(x_i)} \right). \end{aligned} \quad (1)$$

where $\mathcal{N}(x_i)$ is the set of immediate neighbors of state x_i , and $\mathcal{N}(x_j)$ is the set of immediate neighbors of state x_j ; i.e. $\mathcal{N}(x_i)$ is the set of states that can be reached by a single move from state x_i . It can be proved that the stationary probabilities for this Markov chain are given by the Boltzmann probabilities $p_i^* = \frac{e^{-E(x_i)/RT}}{Z}$, as shown in [14].

In contrast, Figure 3 presents pseudocode for the WL algorithm, which implements a random walk on the Markov chain whose transition probabilities $p_{i,j}$ of moving from state x_i to x_j are given by

$$\begin{aligned} p_{i,j} &= P(x_i \rightarrow x_j) = \frac{1}{\mathcal{N}(x_i)} \cdot \min \left(\frac{g(E(x_i))}{g(E(x_j))}, 1 \right) \\ &= P(x_j \rightarrow x_i) = \frac{1}{\mathcal{N}(x_j)} \cdot \min \left(\frac{g(E(x_j))}{g(E(x_i))}, 1 \right). \end{aligned} \quad (2)$$

In this case, the stationary probability of state x_i is given by $\frac{g(E(x_i))}{\sum_x g(E(x))}$.

The mathematical justification for applying the Metropolis-Hastings Monte Carlo method [40] to determine the minimum energy conformation of a biopolymer [46, 8, 18] depends on two facts: (i) every finite, irreducible, aperiodic Markov chain has a *stationary probability* distribution, (ii) if the Markov chain is *reversible*, a situation called *detailed balance* by the physics community, then the stationary distribution of the Markov chain corresponding to the Metropolis-Hastings algorithm is the *Boltzmann distribution*, defined by $P(x) = \frac{\exp(-E(x)/RT)}{Z}$, where $E(x)$ is the energy of state (i.e. conformation) x , R is the universal gas constant 1.986 cal/mol, T is absolute temperature, and the *partition function* Z is defined by $\sum_x \exp(-E(x)/RT)$, where the sum is taken over all states x in the Markov chain. As temperature T approaches zero, the Boltzmann probability of the minimum energy state approaches 1, in the case of a unique minimum energy state, or more generally $1/m$, in the case of m distinct minimum energy states. See [14] for details.

In contrast to the Metropolis-Hastings algorithm, which performs a random walk on the Markov chain of states (secondary structures), the Wang-Landau algorithm performs a random walk on the *energy space* of the Markov chain of states (secondary structures), where the stationary probability of visiting energy e_i is proportional to $\frac{1}{g(e_i)}$, then the histogram of energies encountered in the random walk will be flat.

In this paper, we consider the Markov chain, whose states are the secondary structures of a given RNA sequence, and for which permissible local moves correspond to the addition or removal of a single base pair [23]. Although detailed balance holds for the Metropolis-Hastings algorithm in Figure 2, it does not necessarily hold for the Metropolis algorithm, obtained by replacing line

$$11. \text{ if } (z < \frac{e^{-E(y)/RT}/N_y^x}{e^{-E(x)/RT}/N_x^y}) \text{ then } x = y$$

by

$$11. \text{ if } (z < \frac{e^{-E(y)/RT}}{e^{-E(x)/RT}}) \text{ then } x = y$$

Indeed for the case of RNA secondary structures, detailed balance does not hold in this situation, since if we define the stationary probability p_i^* for state x_i to be the Boltzmann probability $p_i^* = \frac{\exp(-E(x_i)/RT)}{Z}$, and the transition probabilities given by equation (1), then it is not always the case that $p_i^* \cdot p_{i,j} = p_j^* \cdot p_{j,i}$. For instance, the empty structure $S = \dots$ on the 10-mer GGGGGCCCCC has 18 immediate neighbors, one of which is $T = (\dots)$. The structure T has 11 immediate neighbors, one of which is the empty structure S . Letting $x_i = S$ and $x_j = T$, we have $E(x_i) = 0$ kcal/mol, $E(x_j) = 2.70$ kcal/mol, ensemble free energy is $-RT \ln(Z) = -3.96$, hence $Z = \exp(3.96/RT)$ where $T = 310^\circ$ Celsius so $Z = 621.5$, and we have stationary probabilities $p_i^* = \frac{1}{621.5} = 0.00161$, $p_j^* = \frac{0.012456}{621.5} = 0.00002$, $p_{i,j} = \frac{0.012456}{18}$, and $p_{j,i} = \frac{1}{11}$. We compute that

$$\begin{aligned} p_i^* \cdot p_{i,j} &= 0.00161 \cdot 0.012456/18 = 692.01 \times 10^{-6} \\ p_j^* \cdot p_{j,i} &= 0.00002 \cdot 1/11 = 1.82 \times 10^{-6}. \end{aligned}$$

Summarizing, in the Metropolis algorithm (with modified line 11), reversibility of a Markov chain depends on the permissible local moves, while in the Metropolis-Hastings algorithm (with line 11 as in Figure 2), reversibility is always ensured. In the case at hand, if every secondary structure is an immediate neighbor of every secondary structure, then in the Metropolis algorithm, transition probabilities would be

$$\begin{aligned} p_{i,j} &= P(x_i \rightarrow x_j) = \min \left(1, \frac{\exp(-E(x_j)/RT)/Z}{\mathcal{N} \cdot \exp(-E(x_i)/RT)/Z} \right) \\ &= \min \left(1, \frac{\exp(\frac{-(E(x_j)-E(x_i))}{RT})}{\mathcal{N}} \right). \end{aligned} \quad (3)$$

where \mathcal{N} is the number of secondary structures. In this case, an easy computation shows that the Markov chain is reversible. Despite the non-reversible nature of the Markov chain corresponding to the Metropolis algorithm, whose states are the secondary structures of a given RNA sequence, and whose local moves consist of the addition or removal of a single base pair, it has been standard practice to apply the Metropolis algorithm in this case [23, 28, 59, 17]. For that reason, we do not hesitate to apply the Wang-Landau algorithm for the study of RNA secondary structure formation.

Note that in Figure 3, the Wang-Landau computes the relative density of states, defined by $g(i) = N(e_i)/N$, where $N(e_i)$ is the number of states having energy e_i , and N is the total number of states. In the case of RNA secondary structures, it is simple to compute the total number of secondary structures by dynamic programming, given as follows. Given an RNA sequence of length n , let $BP_{i,j} = 1$ if positions i, j can form a Watson-Crick or wobble pair, otherwise let $BP_{i,j} = 0$. Let $\theta = 3$ denote the minimum number of unpaired bases in a hairpin loop. Letting $N_{i,j}$ denote

the number of secondary structures on subsequence $[i, j]$ of the given RNA sequence, we have that $N_{i,j} = 0$ if $j < i + 3$, and otherwise

$$N_{i,j} = N_{i,j-1} + \sum_{k=i}^{j-\theta-1} BP_{k,j} \cdot N_{i,k-1} \cdot N_{k+1,j-1}.$$

It follows that the total number of secondary structures is then $N_{1,n}$. From the relative density of states computed by Wang-Landau algorithm, we compute the absolute density of states by

$$g(e_i) = g(e_i) \cdot N.$$

For fixed temperature T for which the Wang-Landau computation was done, we can compute the partition function $Z(T) = \sum_S \exp(-E(S)/RT)$ by

$$Z(T) = \sum_E g(E) \cdot \exp(-E/RT). \quad (4)$$

In their original article [55, 54], Wang and Landau mentioned that in the case of the Ising model, equation (4) allows one to compute the partition function at any desired temperature T from the density of states. Unfortunately, this is no longer the case for the Turner nearest neighbor model [60] of RNA secondary structure, since the free energy parameters for stacked base pairs, hairpins, bulges, internal loops, etc. all depend on temperature. We can nevertheless proceed by computing the density of states for free energy at $T = 37^\circ$ Celsius, and the density of states for enthalpy (assumed to be temperature independent), and then by convoluting these values, we obtain the density of states for free energy at any desired temperature.

PARTITION FUNCTION FOR A SINGLE RNA

The left panel of Figure 4 displays the relative density of states for the free energy of secondary structures of the 45 nt flavivirus capsid hairpin (cHP) with EMBL access code AB010982/1-45. The right panel of the same figure displays the sum of squared differences between the density of states and the best fitting normal distribution resp. extreme value distribution. The cHP is a conserved RNA hairpin structure in the capsid coding region of flavivirus genomes. Note that the relative density of states, or energy histogram, is approximately normal. In Clote et al. [15] it is rigorously proved that the relative density of states is asymptotically normal; specifically, it is shown that the limit, as n approaches infinity, of the relative density of states for an RNA sequence of length n is normal, where for the purpose of mathematical analysis it is assumed that any base can pair with any other base (homopolymer model) and that the energy of a secondary structure S is -1 times the number of base pairs in S (Nussinov energy model [44]).

PARTITION FUNCTION OF HYBRIDIZATION

Following the approach in program RNAcofold of Bernhardt et al. [5], we can modify the Wang-Landau program of Figure 3 in order to compute the density of states for all *hybridizations* of two RNA sequences, where both intermolecular and intramolecular base-pairing is allowed, provided that there are no pseudoknots.

In the case of the hybridization of two RNA secondary structures, the first of length n and the second of length m , we can compute the total number of hybridizations as follows. Given an RNA sequence $A = a_1, \dots, a_n$ of length n , and an RNA sequence $B = b_1, \dots, b_m$ of length m , let $HP_{i,j} = 1$ if positions a_i, b_j can hybridize, forming a Watson-Crick or wobble pair, otherwise let $HP_{i,j} = 0$. For $1 \leq i, j \leq n, 1 \leq k, \ell \leq m$, let $H_{i,j;k,\ell}$ denote the number of hybridizations of the subsequence a_i, \dots, a_j with b_k, \dots, b_ℓ . From equation (3), we can compute the number $NA_{x,y}$ resp. $NB_{x,y}$ of secondary structures on subsequence a_x, \dots, a_y of A resp. b_x, \dots, b_y of B . If $j < i$ or $\ell < k$, then defined $H_{i,j;k,\ell} = 0$; otherwise

define $H_{i,j;k,\ell}$ by

$$\begin{aligned} & H_{i,j-1;k,\ell-1} \cdot (1 + HP(j, \ell)) \\ & + \sum_{x=i}^{j-1} HP(x, \ell) \cdot H_{i,x-1;k,\ell-1} \cdot NA_{x+1,j} \\ & + \sum_{y=k}^{\ell-1} HP(j, y) \cdot H_{i,j-1;k,y-1} \cdot NB_{y+1,\ell} \end{aligned} \quad (5)$$

It follows that the total number of pseudoknot-free hybridizations is then $H_{1,n;1,m}$.⁵ The previous algorithm is clearly $O(n^4)$.

By considering the number of hybridizations to be the same as the number of secondary structures of a chimeric sequence, formed by concatenating A, B to form $c_1, \dots, c_{n+m} = a_1, \dots, a_n, b_1, \dots, b_m$, we have an $O(n^3)$ algorithm, as follows. For $1 \leq i, j \leq n+m$, if $j < i$ or $(1 \leq i, j \leq n, j-i \leq \theta = 3)$, then $N_{i,j} = 0$, while if $1 \leq i \leq n, n+1 \leq j \leq n+m$, then $N_{i,j} = 1$; otherwise $N_{i,j}$ is equal to

$$N_{i,j-1} + \sum_{k=i}^{j-1} BP_{k,j} \cdot N_{i,k-1} \cdot N_{k+1,j-1}.$$

It follows that the total number of hybridizations is then $N_{1,n}$.

We now describe how to compute the *melting temperature* T_M of hybridization.

1. Compute number of structures for each of 5 species (temperature independent): $\mathcal{S}(A), \mathcal{S}(B), \mathcal{S}(AA), \mathcal{S}(BB), \mathcal{S}(AB)$.
2. For (Celsius) temperature $T \in \{0^\circ, \dots, 100^\circ\}$, compute *relative density of states* $f(A, T), f(B, T), f(AA, T), f(BB, T), f(AB, T)$ for each species by Wang-Landau.
3. For (Celsius) temperature $T \in \{0^\circ, \dots, 100^\circ\}$, compute partition functions $Z(A, T), Z(B, T), Z(AA, T), Z(BB, T), Z(AB, T)$ by

$$Z(T) = \sum_E g(E) \cdot e^{-\frac{E}{RT}}$$

where *absolute density of states* $g(E)$ is relative density times number of structures. For instance

$$g(AB, T)(E) = f(AB, T)(E) \cdot \mathcal{S}(AB).$$

4. Following Dimitrov and Zuker [19], for (Celsius) temperature $T \in \{0^\circ, \dots, 100^\circ\}$, compute ensemble free energy $\Delta G(A, T), \Delta G(B, T), \Delta G(AA, T), \Delta G(BB, T), \Delta G(AB, T)$. This involves the following.

a Redundancy correction:

$$\begin{aligned} Z_{AA} &= Z_{AA} - Z_A^2 \\ Z_{BB} &= Z_{BB} - Z_B^2 \\ Z_{AB} &= Z_{AB} - Z_A \cdot Z_B \end{aligned}$$

b Symmetry correction:

$$\begin{aligned} Z_{AA} &= \frac{Z_{AA}}{2} \\ Z_{BB} &= \frac{Z_{BB}}{2} \end{aligned}$$

c Temperature-dependent chemical equilibrium constants:

$$\begin{aligned} K_A &= \frac{Z_{AA}}{Z_A^2} \\ K_B &= \frac{Z_{BB}}{Z_B^2} \\ K_{AB} &= \frac{Z_{AB}}{Z_A \cdot Z_B} \end{aligned}$$

⁵ In the literature, various types of hybridization are allowed. In Dimitrov-Zuker, no intramolecular structure is allowed, while in Bernhardt et al. pseudo-knot free hybridizations are allowed with intramolecular structure.

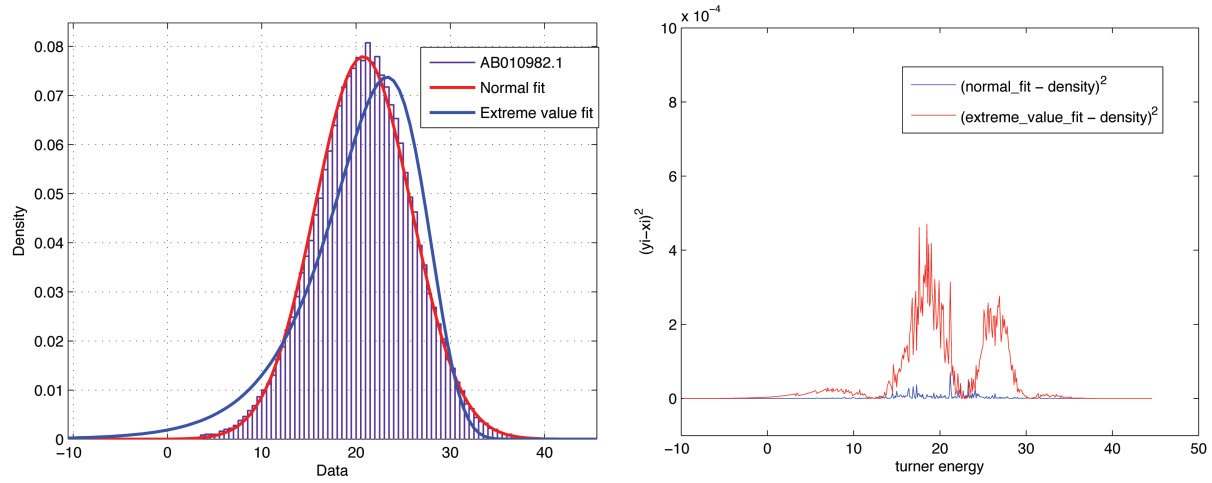


Fig. 4. (Left) Density of states for free energy of secondary structures of the 45 nt flavivirus capsid hairpin (cHP) with EMBL access code AB010982.1 and sequence AUGAACCAACC AACGAAAAAG GACGGGAAAA CCGUCUAUCA AUAUG. Overlaid on the graph is the best fitting normal distribution and the best fitting extreme value distribution. (Right) Sum of squared differences between the density of states and the best fitting normal distribution resp. extreme value distribution. The x -axis of both panels depicts free energy in kcal/mol.

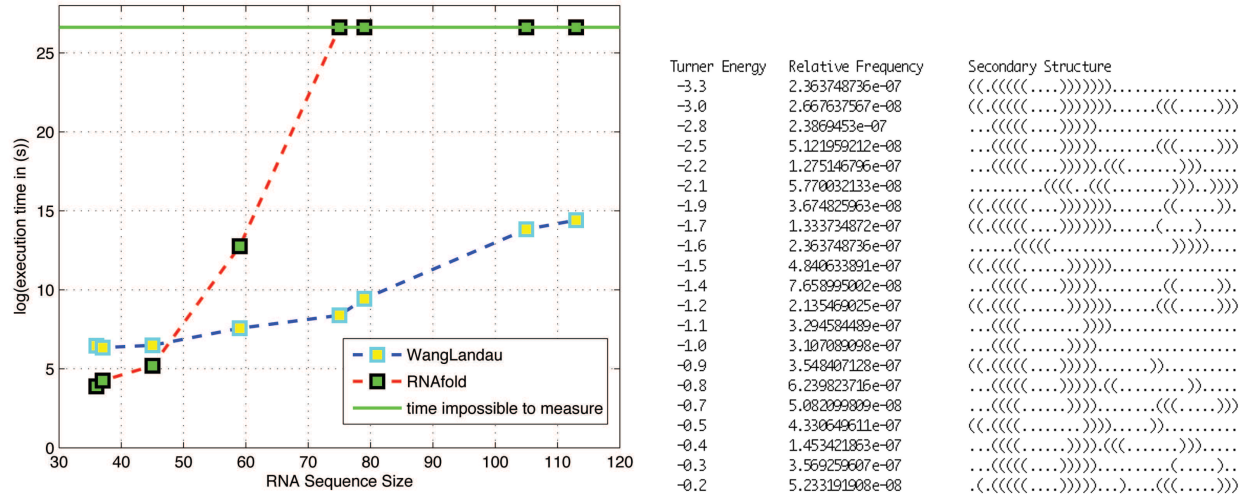


Fig. 5. (Left) Comparison of execution times of WL and program RNAsubopt-D [58], in computing density of states. Since the program of Cupal et al. [16] is no longer publicly available, and is superseded by RNAsubopt (private correspondence from I. Hofacker), we computed the execution time in seconds as a function of $\log n$, where n is RNA sequence size. Horizontal green line is slightly above the value of $\exp(25)$ seconds, or equivalently one day. It appears that for sequences of length 46 nt or greater, the WL method is more efficient than RNAsubopt. (Right) Sample output of WL method on sequence CUGCUUUGAGGACAAAGAGAAUAAAGACUUAUGUU, after 17, 402, 000 WL MonteCarlo steps, where the value of ϵ in line 4 of Figure 3 is defined to be 0.001. The leftmost column contains the energy bin, the middle column contains the relative frequency in the WL sampling run, and the rightmost column contains the lowest energy secondary structure in the associated energy bin. Though our WL program allows the user to modify bin size, the default energy bin size (here) is 0.1 kcal/mol; empty bins, where no structure has yet been sampled, are not displayed. The lowest energy structure sampled by the WL method is $((.(((.(.))))))$ with energy -3.3 kcal/mol, which is identical to the minimum free energy structure, as computed by RNAfold. Only a portion of the output is displayed. In particular, the largest energy of any sampled structure is $+48.8$ kcal/mol; in that energy bin the least energy structure is $.(.(.(.).)((.)(.)).((.)(.)))$.

d Temperature-dependent concentration (number) of molecules A and B:

$$\begin{aligned} 2 \cdot K_A \cdot N_A^2 + K_{AB} \cdot N_A \cdot N_B + N_A - N_A^0 &= 0 \\ 2 \cdot K_B \cdot N_B^2 + K_{AB} \cdot N_A \cdot N_B + N_B - N_B^0 &= 0 \end{aligned}$$

where N_A^0, N_B^0 are given and K_A, K_B, K_{AB} are obtained from previous step. Values N_A and N_B are gotten by using, for example, Newton's method for solving two nonlinear functions; due to issues of numerical instability, Markham uses binary search (p. 43 of [38]).

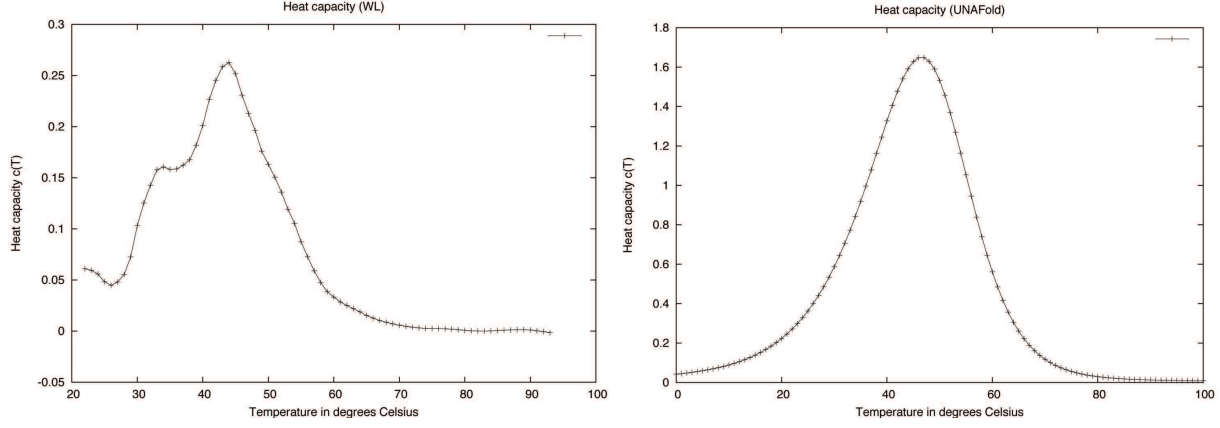


Fig. 6. Computation of heat capacity $c_p(T)$ for the toy sequence 5'-AGCGA-3', hybridized to its reverse complement 3'-UCGCU-5'. (Left) Graph generated by WL method described in this paper. (Right) Graph generated by the program UNAFold [37].

e Letting $Z(A, B, AB, AA, BB)$ equal the following expression

$$\frac{N_A^0! N_B^0!}{N_A, N_B, N_{AB}, N_{AA}, N_{BB}} \cdot Z_A^{N_A} \cdot Z_B^{N_B} \cdot Z_{AB}^{N_{AB}} \cdot Z_{AA}^{N_{AA}} \cdot Z_{BB}^{N_{BB}}$$

it follows that the total partition function Z satisfies

$$Z = \sum_{N_A, N_B, N_{AB}, N_{AA}, N_{BB}} Z(A, B, AB, AA, BB)$$

which can be approximated by term $Z(A, B, AB, AA, BB)$ where $N_A, N_B, N_{AB}, N_{AA}, N_{BB}$ obtained as previously explained. The chemical potential μ_X for each species X is the partial derivative $\frac{\partial -RT \ln Z}{\partial N_X}$ of ensemble free energy with respect to number of molecules of X , hence

$$\mu_A = \frac{-RT \partial \ln Z(A, B, AB, AA, BB)}{\partial N_A}$$

so

$$\mu_A = -RT \ln(Z_A) + RT \ln\left(\frac{N_A}{N_A^0}\right)$$

$$\mu_B = -RT \ln(Z_B) + RT \ln\left(\frac{N_B}{N_B^0}\right)$$

$$\mu_{AB} = -RT \ln(Z_{AB}) + RT \ln\left(\frac{N_{AB}}{N_A^0 \cdot N_B^0}\right)$$

$$\mu_{AA} = -RT \ln(Z_{AA}) + RT \ln\left(\frac{N_{AA}}{N_A^0 \cdot N_A^0}\right)$$

$$\mu_{BB} = -RT \ln(Z_{BB}) + RT \ln\left(\frac{N_{BB}}{N_B^0 \cdot N_B^0}\right)$$

Total free energy satisfies

$$F = \mu_A \cdot N_A + \mu_B \cdot N_B + \mu_{AA} \cdot N_{AA} + \mu_{BB} \cdot N_{BB} + \mu_{AB} \cdot N_{AB}$$

which simplifies to

$$F = \mu_A \cdot N_A^0 + \mu_B \cdot N_B^0$$

f Normalize the ensemble free energy in terms of energy per mole of solute:

$$\Delta G = \frac{\mu_a \cdot N_A^0 + \mu_b \cdot N_B^0}{\max(N_A^0, N_B^0)}$$

5. Determine *heat capacity* as a function of temperature by

$$C_p(T) = \frac{\partial \Delta H}{\partial T} = -T \frac{\partial^2 \Delta G}{\partial T^2}$$

by computing the second partial of a fitting parabola determined by $2m+1$ evenly spaced points, using the approximation for $\frac{\partial^2 \Delta G}{\partial T^2}$ given by

$$\frac{30}{m(m+1)4m^2(2m+3)\delta T^2} \sum_{-m \leq i \leq m} (3i^2 - m(m+1)) \Delta G(T_0 + i\delta T)$$

In a post-processing step, smooth the heat capacity curve by computing a running average. The melting temperature $T_M(C_p)$ is computed by determining the temperature at which heat capacity achieves a maximum.

4 DISCUSSION

The left panel of Figure 5 displays the run time of the WL method, compared with that of RNAsubopt from the Vienna RNA package, while the right panel of the same figure shows sample output from our WL program. Figure 5 clearly shows the advantage of WL over existent methods in computing the density of states for both single RNA molecules and for hybridization complexes of two RNA molecules. The left and right panels of the Figure 6 depict the heat capacity computed by WL method (left) and the program UNAFold (right). Melting temperature, which is usually defined as the temperature at which half of the molecules are single-stranded while the other half are hybridized, is determined as that temperature where heat capacity achieves its maximum. The program UNAFold does not allow any intramolecular structure (base pairing between two nucleotides of the same structure), a feature that our WL method permits, as does the RNACofold program. While it is clear that additional work must be done to improve heat capacity computation with the WL method, the melting temperature T_M computed by WL agrees reasonably well with that computed by $O(n^3)$ methods UNAFold, RNACofold, and the recent $O(n^6)$ method of Chitsaz et al. [11], each of which methods admits slightly different interactions.

We now describe how to approximately compute the partition function $Z(T)$ over all secondary structures and over all pseudoknot-free hybridizations, simultaneously over all temperatures from 0° to 100° Celsius, by performing two WL computations, followed by a computation of the convolution of enthalpy relative

frequency with free energy relative frequency. Similar computations using existent methods require over 100 cubic time computations.

- Compute the relative density of states p_h for free energy using Wang-Landau with temperature $T = -273^\circ$ Celsius (absolute zero Kelvin). It follows that p_h is the relative density of states for enthalpy. Due to the fundamental thermodynamic relation

$$\Delta G = \Delta H - T\Delta S \quad (6)$$

where T is absolute temperature in Kelvin, and ΔG , ΔH , ΔS respectively denote the change in free energy, enthalpy, and entropy.

- Compute the relative density of states p_g for free energy using Wang-Landau with temperature $T = 37^\circ$ Celsius (310 Kelvin).
- From equation (6), we have that

$$\Delta S = \frac{\Delta H - \Delta G}{T}.$$

- Given arbitrary absolute temperature T , compute the relative density of states for free energy at temperature T by the following pseudocode, representing a kind of *convolution* of p_g with p_h .
 1. for all z initialize $p(z) = 0$
 2. for x ranging over *enthalpy* bins
 3. for y ranging over *free energy* bins
 4. $z = \frac{x-y}{T}$
 5. $p(z) += p_h(x) * p_g(y)$
- Compute the absolute density of states $g(z) = p(z) \cdot N$, where N is the total number of secondary structures, computed by equation (3).

By this method, one can approximate the partition function $Z(T)$ for all temperatures from 0° to 100° in Celsius, by performing two Wang-Landau sampling runs, respectively at temperatures -273° and 37° Celsius, and then to repeatedly perform a fast convolution. The method just described, which involves two WL computations, together with convolution computations, has until now not worked well in practice, for certain technical reasons. This direction needs further exploration.

Another issue concerning any sampling method is the required time to obtain reasonably good estimates of the quantity in question. In the case of RNA kinetics, computations of mean first passage time (MFPT) to reach the minimum free energy structure take inordinate amounts of time, when using Metropolis-Hastings Monte Carlo methods, which are *time-driven* simulations. For this reason, the program KinFold [24] uses an *event-driven* simulation, where time is incremented by an exponentially distributed random variable. It may be possible to use similar ideas to increase efficiency of our WL program, which should further improve the accuracy in the computation of heat capacity. Finally, we intend to implement a new energy evaluation function, that allows arbitrary pseudoknots, zig-zags, etc. using energy parameters from the recent dissertation of Bon [7]. This will allow us to estimate the partition function, ensemble free energy, heat capacity, melting temperature, etc. for a context known to be NP-complete.

5 CONCLUSION

In this paper, we have implemented the Wang-Landau algorithm to compute the relative density of states for RNA secondary structures and hybridizations. Separately computing the number of structures and hybridizations, we obtain the absolute density of states, which then yields the partition function, and thence, in the case of hybridization, the melting temperature. The WL method is much faster than existent software RNAsubopt in computing the density of states, but could not be benchmarked with the binning method of Cupal et al. which runs in $O(m^2n^3)$ time, for length n sequence and m energy bins, since the latter software is no longer available, being superseded by RNAsubopt-D. In preliminary tests, we obtain roughly the same melting temperature for duplex RNA, as that computed by existent methods; however, the real advantage of the WL method is that there is no restriction on types of allowed interaction, unlike the situation with dynamic programming approaches that disallow pseudoknots, zig-zags, etc.

ACKNOWLEDGEMENTS

We would like to thank Ivo Hofacker, for informing us of the status of the program of Cupal et al. [16], and for explaining the undocumented option -D of RNAsubopt, which computes the density of states from the enumeration of all secondary structures within a certain energy range of the minimum free energy. As well, thanks to three anonymous reviewers for their suggestions.

Funding: Research of P. Clote and F. Lou was funded by the Digeo Foundation, in the form of a Digeo Chair of Excellence to P. Clote. Additionally, P. Clote was partially by the National Science Foundation under grants DMS-0817971 and DBI-0543506. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- [1] J.P. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.*, 18:3035–3044, 1990.
- [2] A.R. Banerjee, J.A. Jaeger, and D.H. Turner. Thermal unfolding of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry*, 32:153–163, 1993.
- [3] M. Bekaert, L. Bidou, A. Denise, G. Duchateau-Nguyen, J.P. Forest, C. Froidevaux, I. Hatin, J.P. Rousset, and M. Termier. Towards a computational model for -1 eukaryotic frameshifting sites. *Bioinformatics*, 19:327–335, 2003.
- [4] S. H. Bernhart, H. Tafer, U. Muckstein, C. Flamm, P. F. Stadler, and I. L. Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms. Mol. Biol.*, 1(1):3, 2006.
- [5] S.H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P.F. Stadler, and I.L. Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, 1(1):3, 2006.
- [6] A. Böck, K. Forschhammer, J. Heider, and C. Baron. Selenoprotein synthesis: An expansion of the genetic code. *Trends Biochem. Sci.*, 16:463–467, 1991.
- [7] Michael Bon. *Prédiction de structures secondaires d'ARN avec pseudo-noeuds*. PhD thesis, Ecole Polytechnique, 2009. Ph.D. dissertation in Physics.
- [8] P. Bradley, K. M. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, September 2005.
- [9] M. T. Cheah, A. Wachter, N. Sudarsan, and R. R. Breaker. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, 447(7143):497–500, May 2007.
- [10] Z. Chen and Y. Xu. Structure prediction of helical transmembrane proteins at two length scales. *J. Bioinform. Comput. Biol.*, 4(2):317–333, April 2006.

- [11] H. Chitsaz, R. Salari, S. C. Sahinalp, and R. Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–i373, June 2009.
- [12] S. S. Cho, D. L. Pincus, and D. Thirumalai. Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. *Proc. Natl. Acad. Sci. U.S.A.*, 106(41):17349–17354, October 2009.
- [13] S. Chowdhury, C. Ragaz, E. Kreuger, and F. Narberhaus. Temperature-controlled structural alterations of an RNA thermometer. *J. Biol. Chem.*, 278(48):47915–47921, November 2003.
- [14] P. Clote and R. Backofen. *Computational Molecular Biology: An Introduction*. John Wiley & Sons, 2000. 279 pages.
- [15] P. Clote, E. Kranakis, D. Krizanc, and B. Salvy. Asymptotics of canonical and saturated RNA secondary structures. *J. Bioinform. Comput. Biol.*, 7(5):869–893, October 2009.
- [16] J. Cupal, I. Hofacker, and P. Stadler. Dynamic programming algorithm for the density of states of RNA secondary structures. In R. Hofstädt, T. Lengauer, M. Löffler, and D. Schomburg, editors, *Computer Science and Biology 96 (Proceedings of the German Conference on Bioinformatics)*, pages 184–186. Univ. Leipzig, 1996.
- [17] L. V. Danilova, D. D. Pervouchine, A. V. Favorov, and A. A. Mironov. RNAkinetics: a web server that models secondary structure kinetics of an elongating RNA. *J. Bioinform. Comput. Biol.*, 4(2):589–596, April 2006.
- [18] R. Das and D. Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, 104(37):14664–14669, September 2007.
- [19] R. A. Dimitrov and M. Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophysical Journal*, 87:215–226, 2004.
- [20] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, 49(1):65–88, 2007.
- [21] R. M. Dirks and N. A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24(13):1664–1677, 2003.
- [22] S. R. Eddy. How do RNA folding algorithms work? *Nat. Biotechnol.*, 22(11):1457–1458, November 2004.
- [23] C. Flamm, W. Fontana, I. L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.
- [24] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of multistable RNA molecules. *RNA*, 7(2):254–265, February 2001.
- [25] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res.*, 31(1):439–441, 2003.
- [26] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of rna secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.
- [27] I. L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31:3429–3431, 2003.
- [28] H. Isambert and E. D. Siggia. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. U.S.A.*, 97(12):6515–6520, June 2000.
- [29] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
- [30] Bjarne Knudsen and Jotun Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, 31(13):3423–3428, 2003.
- [31] S. Kou, Q. Zhou, and W. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.*, 34:1581–1652, 2006.
- [32] S. C. Kou, J. Oh, and W. H. Wong. A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling. *J. Chem. Phys.*, 124(24):244903, June 2006.
- [33] F. Lefebvre. An optimized parsing algorithm well-suited to rna folding. In AAAI press, editor, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 222–230, 1995.
- [34] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel. Vertebrate microRNA genes. *Science*, 299(5612):1540, 2003.
- [35] R. B. Lyngso and C. N. Pedersen. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.*, 7(3-4):409–427, 2000.
- [36] M. Mandal, B. Boese, J. E. Barrick, W. C. Winkler, and R. R. Breaker. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*, 113(5):577–586, 2003.
- [37] N. R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, 453:3–31, 2008.
- [38] N. R. Markham. Algorithms and software for nucleic acid sequences, 2006.
- [39] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, 101:7287–7292, 2004.
- [40] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.
- [41] D. Metzler and M. E. Nebel. Predicting RNA secondary structures with pseudoknots by MCMC sampling. *J. Math. Biol.*, 56(1-2):161–181, January 2008.
- [42] U. Mückstein, H. Tafer, J. Hackermüller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.
- [43] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, May 2009.
- [44] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences, USA*, 77(11):6309–6313, 1980.
- [45] A. D. Omer, T. M. Lowe, A. G. Russell, H. Ebhardt, S. R. Eddy, and P. P. Dennis. Homologues of small nucleolar RNAs in Archaea. *Science*, 288:517–522, 2000.
- [46] A. R. Ortiz, A. Kolinski, and J. Skolnick. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.*, 277:419–448, 1998.
- [47] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC. Bioinformatics*, 5:104, August 2004.
- [48] K. M. Reinisch and S. L. Wolin. Emerging themes in non-coding RNA quality control. *Curr. Opin. Struct. Biol.*, 17(2):209–214, April 2007.
- [49] J. Ren, B. Rastegari, A. Condon, and H. H. Hoos. Hotknots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11(10):1494–1504, October 2005.
- [50] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.
- [51] J. E. Tabaska, R. E. Cary, H. N. Gabow, and G. D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14:691–699, 1998.
- [52] B. J. Tucker and R. R. Breaker. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, 15(3):342–348, June 2005.
- [53] F. H. Van Batenburg, A. P. Gulyaev, and C. W. Pleij. Pseudobase: structural information on RNA pseudoknots. *Nucleic Acids Res.*, 29(1):194–195, January 2001.
- [54] F. Wang and D. P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64(056101):056101(1)–056101(16), 2001.
- [55] F. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, 86(10):2050–2053, Keywords =, 2001.
- [56] J. S. Weinger, K. M. Parnell, S. Dorner, R. Green, and S. A. Strobel. Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nat. Struct. Mol. Biol.*, 11:1101–1106, 2004.
- [57] K. C. Wiese, E. Glen, and A. Vasudevan. Jvarkit: a Java tool for RNA secondary structure visualization. *IEEE. Trans. Nanobioscience*, 4(3):212–218, September 2005.
- [58] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–164, 1999.
- [59] A. Xayaphoummine, T. Bucher, and H. Isambert. Kinfold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.*, 33(Web):W605–W610, July 2005.
- [60] T. Xia, Jr. J. SantaLucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–35, 1999.
- [61] J. Zhang, S. C. Kou, and J. S. Liu. Biopolymer structure simulation and optimization via fragment regrowth Monte Carlo. *J. Chem. Phys.*, 126(22):225101, June 2007.
- [62] J. Zhao, R. L. Malmberg, and L. Cai. Rapid ab initio prediction of RNA pseudoknots via graph tree decomposition. *J. Math. Biol.*, 56(1-2):145–159, January 2008.
- [63] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, 1981.