Thermodynamics of RNA structures by Wang-Landau sampling

Authors: Feng Lou, Peter Clote

Persistent link: http://hdl.handle.net/2345/bc-ir:103575

This work is posted on eScholarship@BC, Boston College University Libraries.

Published in Bioinformatics, vol. 26, no. 12, pp. i278-i286, June 2010

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Thermodynamics of RNA structures by Wang–Landau sampling

Feng Lou¹ and Peter Clote¹,²,³,*

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 and ³Digiteo Chair, Laboratoire d’Informatique (LIX), Ecole Polytechnique, 91128 Palaiseau, France

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 expressed sequence tag (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 nondeterministic polynomial time (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 article, 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 existing 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 article; in contrast, such methods can be loosely grouped into two types of algorithm— those that use (i) a stochastic context free grammar to compute a covariation model and (ii) 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 (Nawrocki et al., 2009) and Pfam (Knudsen and Hein et al., 2003). Examples of thermodynamics-based algorithms are the programs mfold (Zuker and Stiegler, 1981), UNAFOLD (Markham and Zuker, 2008), RNAfold (Hofacker et al., 1994), RNastructure (Mathews et al., 2004). Thermodynamics-based algorithms for hybridization of two structures are given in UNAFOLD (Dimitrov and Zuker, 2004), RNAfold (Bernhart et al., 2006; Mückstein et al., 2006), while the UNIPACK software considers hybridization of three or more RNA molecules. (Dirks et al., 2006)

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 (Weinger et al., 2004). 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 (Böck et al., 1991), ribosomal frameshift (Bekaert et al., 2003), transcriptional and translational gene regulation (Lim et al., 2003; Mandal et al., 2003), temperature-sensitive conformational switches (Chowdhury et al., 2003; Tucker and Breaker, 2005), chemical modification of specific nucleotides in the ribosome (Omer et al., 2000), regulation of alternative splicing (Cheah et al., 2007), etc.

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

© The Author(s) 2010. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
2007). 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 Banerjee et al. (1993) for experimental data supporting this view. Composing the minimum free-energy pseudoknotted structure for a given RNA sequence is NP-complete Lyngso and Pedersen (2000) for the Turner nearest neighbor energy model. 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 and (iii) heuristic methods. Examples of exact algorithms for pseudoknot structure prediction are the branch-and-bound algorithm of (Bon, 2009) and the method using tree-width decomposition of Zhao et al. (2008). Examples of algorithms that consider only pseudoknots of a particular class are found in the pioneering work of Rivas and Eddy (1999) and Lefebvre (1995), with subsequent refinements in Dirks and Pierce (2003); Reeder and Giegerich (2004) and Ren et al. (2005). Examples of heuristic approaches include Monte Carlo methods Metzler and Nebel (2008), genetic algorithms Abrahams et al. (1990) 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 article 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, advantage of the WL approach is not 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 \( s = a_1, \ldots, a_n \), 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: \((A,U), (U,A), (G,C), (C,G), (G,U)\) and \((U,G)\).
2. **Threshold requirement:** if \((i,j)\) belongs to \( S \), then \( j - i > \theta \).
3. **Non-existence of pseudoknots:** if \((i,j)\) and \((k,\ell)\) 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 \( \theta \), or minimum number of unpaired bases in a hairpin loop, is taken to be three. For any additional background on RNA and dynamic programming computation of secondary structures, see Clote and Backofen (2000) and the recent review Eddy (2004).

## 2 APPROACH

The non-Boltzmannian WL Monte Carlo algorithm was developed by Wang and Landau (2001a, b) 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 WL 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, \( p_{\text{wl}}(s) \) of visiting the state \( s \), whose energy is \( E \), is given by the uniform probability \( \frac{1}{Z} \) times the Boltzmann probability \( \frac{1}{Z} \exp(-E/RT) \); in contrast, for the WL algorithm, the expected frequency or stationary probability, of visiting state \( s \) is \( p_{\text{wl}}(s) = \frac{1}{\bar{E}Z} \), where \( \bar{E} \) 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 (2001a, b), Kou and Wong Kou et al. (2006a), 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 Chen and Xu (2006), 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 [Kou et al. (2006a)], 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 [Kou et al. (2006b)], as well as in protein structure prediction by the fragment assembly [Zhang et al. (2009)].

In this article, 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 [Wuchty et al. (1999)], 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 \( Z(T) \) at all temperatures from 0°C to 100°C by performing two WL 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 the existent software UNAFold (Hofacker, 2003), respectively, RNAcofold (Bernhart et al., 2006). UNAFold (Markham and Zuker, 2008) and a recently published method of Chitsaz et al., 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 [Lyngsø and Pedersen (2000)].

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 structure is a secondary structure for an arbitrary but fixed RNA sequence \( x \). In [Cupal et al. (1996)], described the first efficient algorithm, running in \( O(n^2m^3) \) time, to compute the density of states for an RNA sequence for a length \( m \), where energy is discretized into \( m \) bins. The program of Cupal et al. (1996) is no longer available, since it has been superseded by the program RNAsubopt, developed by Wuchty et al. (1999), 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 \( \rightarrow D \), which, instead of outputting structures, outputs only the number of secondary structures in each energy bin above the minimum free energy (bin size 0.1 kcal/mol).

3 METHODS
Monte Carlo algorithms have been implemented by a number of groups, to study RNAkines of folding. In particular, KinesFold, developed by Flam et al. (2000), 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 Isambert and Sigga (2000) and Xyaphumhume et al. (2005) a similar time-driven Monte Carlo simulation program, KinesFold, is described to compute kinetic determinations pseudoknotted structure for a given RNA sequence. Dandova et al. (2006) describe the RNAkinetics web server used to compute 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 WL algorithm.

---

1. procedure Metropolis-Hastings( )
2. \( T = T_{0} \)
3. \( x = \) initial state
4. while \( (T > T_{0}) \) {
5. repeat \( M \) times {
6. choose random neighbor \( y \in \mathcal{N}_{x} \)
7. if \( (E(x) \leq E(y)) \) then
8. \( x = y \)
9. else
10. choose random \( e(0,1) \)
11. if \( ( \frac{e}{T_{0}} \leq \frac{E(y) - E(x)}{T_{0}} ) \) then \( x = y \)
12. }
13. \( T = T 	imes 0.9 \)
14. }
15. return \( x \)

Fig. 2. Pseudocode for Metropolis-Hastings algorithm with simulated annealing (Kirkpatrick et al., 1983).

3.1 WL
The WL algorithm, (Wang and Landau, 2001a, b) 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 \( Q = \{1, \ldots, n\} \) be a finite set of states, let \( r = (r_{1}, \ldots, r_{n}) \) be the distribution for initial state, and let \( P = (p_{ij}) \) be a matrix of transition probabilities, satisfying \( \sum r_{i} = 1 \) for all \( i \). A (first-order, time-homogeneous) Markov chain \( M = (Q, P, \pi) \) is a stochastic process, whose state \( q_{i} \) at time \( t \) is a random variable determined by

\[
P(r_{0} = i | q_{0} = 1) = r_{i}
\]

Define \( p_{ij}(t) = P(q_{t} = j | q_{t-1} = i) \) and \( p_{ij}^{(0)} = P(q_{0} = j | q_{0} = i) \). Clearly, the \( (j, i) \)-th entry of the \( t \)-th power \( P^{t} \) of \( P \) equals \( p_{ij}^{(t)} \); moreover, by time-homogeneity it follows that \( p_{ij}^{(t)} = p_{ij}^{(1)} \), for all \( t \). The stationary probability of state \( i \) is defined by \( \lim t p_{ij}(t) = \pi_{i}^{*} \), 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 Cote and Backofen (2000) for a new, self-contained proof of this result. A Markov chain with state set \( Q \) and stationary probabilities \( \pi_{1}^{*}, \ldots, \pi_{n}^{*} \) is reversible, if for all \( i, j \in Q \), \( p_{ij}^{(t)} = p_{ji}^{(t)} \).

Figure 2 presents pseudocode for the classical Metropolis–Hastings Monte Carlo algorithm with simulated annealing (Kirkpatrick et al., 1983; Metropolis et al., 1953), which implements a random walk on the Markov chain whose transition probabilities \( p_{ij} \) of moving from state \( x_{i} \) to \( x_{j} \) is given by

\[
p_{ij} = P(x_{i} \rightarrow x_{j}) = \min \left\{ 1, \frac{\exp(-E(x_{i})/RT) / Z(x_{i})}{\exp(-E(x_{j})/RT) / Z(x_{j})} \right\} = \min \left\{ 1, \frac{\exp(-E(x_{i})/RT)}{\exp(-E(x_{j})/RT)} \right\} \frac{Z(x_{i})}{Z(x_{j})}
\]

where \( Z(x_{j}) \) is the set of immediate neighbors of state \( x_{j} \), and \( Z(x_{j}) \) is the set of immediate neighbors of state \( x_{i} \); i.e. \( Z(x_{j}) \) is the set of states that can be
1. procedure WangLandau(s)
2. \( S = \emptyset \) // empty initial structure
3. \[ e = \exp(1) \] // initial modification factor
4. while \( e > 1 \times e \) {
5. for all energies bins \( e_i \): \( g(e) = 1 \)
6. while \( h \) is not flat {
7. for \( i = 1 \) to \( \text{NumSteps} \) {
8. choose random \( T \in N(S) \) of \( S \)
9. \( \epsilon_0 = \text{bin}(E(S)); \epsilon_1 = \text{bin}(E(T)) \)
10. choose random \( \epsilon \in (0,1) \)
11. if \( \epsilon < \frac{g(\epsilon_0)}{g(\epsilon_1)} \)
12. \( S = T \)
13. \( \epsilon = \epsilon_1 \)
14. else // \( S \) remains unchanged
15. \( \epsilon = \epsilon_0 \)
16. \( g(e) = c \cdot g(e) \) // update d.o.s.
17. \( h(e) = h(e)+1 \) // update histogram
18. }
19. \( \epsilon = \sqrt{\epsilon} \) // reduce modification factor
20. }
21. return relative density of states \( g \), where \( g(i) = g(i)/\sum g(j) \)

Fig. 3. Pseudo code for WL algorithm, as applied to RNA secondary structure density of states computation. In line 8, \( N(S) \) 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.

reached by a single move from state \( s_i \). It can be proved that the stationary probabilities for this Markov chain are given by the Boltzmann probabilities \( p^*_j = \exp(-E_j/RT) \) as shown in Clote and Backofen (2000).

In contrast, Figure 3 presents pseudo code for the Markov chain, which implements a random walk on the Markov chain whose transition probabilities \( p_{ij} \) of moving from state \( s_i \) to \( s_j \) are given by

\[
p_{ij} = P(s_i \rightarrow s_j) = \frac{1}{N(s_i)} \min \left( \frac{g(\epsilon_0)}{g(\epsilon_1)}, 1 \right)
= \frac{1}{N(s_i)} \min \left( \frac{g(\epsilon_0)}{g(\epsilon_1)}, 1 \right).
\]

In this case, the stationary probability of state \( s_j \) is given by \( \sum_{i} p_{ij} \).

The mathematical-justification for applying the Metropolis-Haings Monte Carlo method (Metropolis et al., 1953) to determine the minimum energy conformation of a biopolymer (Bradley et al., 2005; Das and Baker, 2007; Orz et al., 1998) depends on two facts: (i) every finite, irreducible, aperiodic Markov chain has a stationary probability distribution and (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-Haings algorithm is the Boltzmann distribution, defined by \( P(s) = \exp(-E(s)/RT) \), where \( E(s) \) is the energy of state \( i \) (i.e. conformation) \( s \), \( R \) is the universal gas constant 1.980 cal/molK, \( T \) is absolute temperature, and the partition function \( Z \) is defined by \( Z = \sum \exp(-E(s)/RT) \), where the sum is taken over all states \( s \) 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 Clote and Backofen (2000) for details.

In contrast to the Metropolis-Haings algorithm, which performs a random walk on the Markov chain of states (secondary structures), the WL 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 \( \exp(-E_i/RT) \), then the histogram of energies encountered in the random walk will be flat.

In this article, 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 (Flamm et al., 2000). Although detailed balance holds for the Metropolis-Haings algorithm in Figure 2, it does not necessarily hold for the Metropolis algorithm, obtained by replacing line

\[
11. \text{if } (\epsilon < \frac{g(\epsilon_0)}{g(\epsilon_1)}) \text{ then } x = y
\]

by

\[
11. \text{if } (\epsilon < \frac{g(\epsilon_0)}{g(\epsilon_1)}) \text{ then } x = y
\]

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 \( s_i \) to be the Boltzmann probability \( p^*_i = \exp(-E_i/RT) \), and the transition probabilities given by Equation (1), then it is not always the case that \( p^*_j p_{ij} = p^*_j p_{ji} \). For instance, the empty structure \( s = \) . . . on the 10-mer GGUCCGCCCCC has 18 immediate neighbors, one of which is \( T = \) . . . . . . . . . . The structure \( T \) has 11 immediate neighbors, one of which is the empty structure \( S \). Letting \( s_i = S \) and \( s_j = T \), we have \( E(s_i) = 0 \) kcal/mol, \( E(s_j) = 2.70 \) kcal/mol, ensemble free energy is \( -kT \log Z = -9.36 \), hence \( Z = \exp(3.96/RT) \) when \( T = 310^\circ \), so \( Z = 0.8215 \) and we have stationary probabilities \( p^*_i = \frac{\exp(-E_i/RT)}{Z} = 0.00161, p^*_j = \frac{\exp(-E_j/RT)}{Z} = 0.00002 \) and \( p_{ij} = \frac{\exp(-E_j/RT)}{Z} \). We compute that

\[
p^*_j p_{ij} = 0.00161 \times 0.012456/18 = 6.92 \times 10^{-8}
\]

\[
p^*_i p_{ji} = 0.00002 \times 0.11 = 1.9 \times 10^{-8}
\]

Summarizing, in the Metropolis (and modified line 11), reversibility of a Markov chain depends on the permissible local moves, while in the Metropolis-Haings 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

\[
p_{ij} = P(s_i \rightarrow s_j) = \min \left( 1, \frac{\exp(-E_j/RT)}{Z} \right)
= \min \left( 1, \frac{\exp(-E_j/RT)}{\sum \exp(-E_k/RT)} \right).
\]

where \( N \) 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 a standard practice to apply the Metropolis algorithm in this case (Damilova et al., 2006; Flamm et al., 2000; Humbert and Sigga, 2000; Xantheoumanis et al., 2005). For that reason, we do not hesitate to apply the WL algorithm for the study of RNA secondary structure formation.

Note that in Figure 3, the WL computes the relative density of states, defined by \( g(i) = N(i)/N \), where \( N(i) \) 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 \( BP_{ij} \) = 1 if positions \( i,j \) form a Watson-Crick or wobble pair, otherwise \( BP_{ij} = 0 \). Let \( \theta = 3 \) denote the minimum number of unpaired bases in a hairpin loop. Letting \( N_j \) denote the number of secondary structures on subsequence \( [i,j] \) of the given RNA sequence, we have that \( N_j = \theta \times \gamma \times (i-j+3) + \) and otherwise

\[
N_j = N_{j-1} + \sum_{i=0}^{j-1} BP_{ij} N_{i+1-j} - N_{i+2-j}.
\]
It follows that the total number of secondary structures is then \( N_{ss} \). From the relative density of states computed by WL algorithm, we compute the absolute density of states by

\[
g(x) = N \cdot g(x).
\]

For fixed temperature \( T \) for which the WL computation was done, we can compute the partition function \( Z(T) \) by

\[
Z(T) = \sum_{E} g(E) \cdot \exp(-E/RT),
\]

Equation (4).

In their original article Wang and Landau (2001a, b) 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 neighborhood model Xia et al., 1999 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 any desired temperature, and by convoluting these values, we obtain the density of states for free energy at any desired temperature.

3.2 Partition function for a single RNA

Figure 4a displays the relative density of states for the free energy of secondary structures of the 45 nt flavivirus cHP with EMBL access code AB1010982/1-45 and sequence AUGAACAACC AACGAAAAAG GACGGGAAAA CCGUCUAUCA AUAUG. Overlaid on the graph is the best fitting normal distribution and the best fitting extreme value distribution. The x-axis of both panels depicts free energy in kcal/mol.

The AUGAACAACC AACGAAAAAG GACGGGAAAA CCGUCUAUCA AUAUG

3.3 Partition function of hybridization

Following the approach in program RNAcofold of Bernhart et al. (2006), we can modify the WL program of Figure 3 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 \( A = a_1, \ldots, a_n \) of length \( n \) and an RNA sequence \( B = b_1, \ldots, b_m \) of length \( m \), let \( H_{P,j} = 1 \) if positions \( a_j \) and \( b_j \) can hybridize, forming a Watson–Crick or wobble pair, otherwise let \( H_{P,j} = 0 \). For \( 1 \leq j \leq n \), \( 1 \leq k \leq m \), let \( H_{i,j,k} \) denote the number of hybridizations of the subsequence \( a_i, \ldots, a_j \) with \( b_k, \ldots, b_m \). From Equation (5), we can compute the number \( N_{A,B} \) of secondary structures on subsequence \( a_1, \ldots, a_n \) of \( A \), respectively, \( b_1, \ldots, b_m \) of \( B \). Then \( H_{i,j,k} \) is equal to \( N_{i,j,k} \), which is the density of states for the free energy at any desired temperature.

In the literature, various types of hybridization are allowed. In Dimitrov-Zuker (2004), no intramolecular structure is allowed, while in Bernhart et al. (2006) pseudoknot-free hybridizations are allowed with intramolecular structure.
We now describe how to compute the melting temperature $T_M$ of hybridization.

1. Compute number of structures for each of five species (temperature independent): $Z(A), Z(B), Z(AB), Z(AB',A)$ and $Z(AB',B)$.

2. For temperature $T \in [0\degree C, 100\degree C]$, compute relative density of states $f(A,T), f(B,T), f(AB,A,T), f(AB,B,T)$ and $f(AB,T)$ for each species by $W_L$.

3. For temperature $T \in [0\degree C, 100\degree C]$, compute partition functions $Z(A,T), Z(B,T), Z(AB,A,T)$ and $Z(AB,B,T)$ by

$$Z(T) = \sum g(E) e^{\beta E}$$

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

$$g(AB,T)(E) = f(AB,T)(E) \cdot Z(AB,T).$$

4. Following Dimitrov and Zuker (2004), for temperature $T \in [0\degree C, 100\degree C]$, compute ensemble free energy $\Delta G(A,T)$, $\Delta G(B,T), \Delta G(AB,A,T)$, $\Delta G(AB,B,T)$ and $\Delta G(AB,T)$. This involves the following:

(a) Redundancy correction:

$$Z_{AA} = Z_{AA} - Z_{AA}^0$$

$$Z_{AB} = Z_{AB} - Z_{AB}^0$$

$$Z_{BB} = Z_{BB} - Z_{BB}^0$$

(b) Symmetry correction:

$$Z_{AA} = \frac{Z_{AA}}{2}$$

$$Z_{BB} = \frac{Z_{BB}}{2}$$

(c) Temperature-dependent chemical equilibrium constants:

$$K_A = \frac{Z_{AA}}{Z_{AB}}$$

$$K_B = \frac{Z_{BB}}{Z_{AB}}$$

$$K_{AB} = \frac{Z_{AB}}{Z_{AA} Z_{BB}}$$

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

$$2 \cdot K_A \cdot N_0^0 + K_{AB} \cdot N_A \cdot N_B + N_A - N_0^0 \neq 0$$

$$2 \cdot K_B \cdot N_0^0 + K_{AB} \cdot N_A \cdot N_B + N_B - N_0^0 \neq 0$$

where $N_0^0$ is given and $K_A, K_B$, $K_{AB}$ are obtained from the 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 Markham, 2006).

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

$$Z(A,B,AB,AA,AB) = \frac{N_A^0 N_B^0 N_{AA} N_{AB} N_{BB}}{N_A N_B N_{AA} N_{AB} N_{BB}}$$

it follows that the total partition function $Z$ satisfies

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

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

$$\mu_A = -\frac{\partial \Delta G}{\partial N_A}$$

so

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

and

$$\mu_B = -\frac{RT \ln(Z(A, B, AB, AA, AB))}{N_B}$$

$$\mu_{AB} = -\frac{RT \ln(Z(A, B, AB, AA, AB))}{N_{AB}}$$

$$\mu_{AA} = -\frac{RT \ln(Z(A, B, AB, AA, AB))}{N_{AA}}$$

$$\mu_{BB} = -\frac{RT \ln(Z(A, B, AB, AA, AB))}{N_{BB}}$$

Total free energy satisfies

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

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:

$$\Delta G = \frac{\mu_A \cdot N_A^0 + \mu_B \cdot N_B^0}{\text{mass}(N_A^0, N_B^0)}$$

(5) Determine heat capacity as a function of temperature by

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

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

$$\frac{30}{m(m+1)(m+2)(m+3)} \sum_{i=1}^{m+1} (3^i - m(m+1)(i)\Delta G(T_i) + 8T_i).$$

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 Figure 5a displays the run time of the WL method, compared with that of RNAcofold from the Vienna RNA package, while the Figure 5b 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. Figure 6a and b depicts the heat capacity computed by the WL method (Fig. 5a) and the program RNAcofold (Fig. 5b). 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 RNAcofold does not allow any intramolecular structure (base pairing between 2 m 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
Fig. 5. (a) Comparison of execution times of WL and program RNAsubopt (Wuchty et al., 1999), in computing density of states. Since the program of Capal et al. (1996) is no longer publicly available, and is superceded by RNAsubopt (private correspondence from 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 a day. It appears that for sequences of length \( \geq 46 \) nt, the WL method is more efficient than RNAsubopt. (b) Sample output of WL method on sequence CUGCUUUGAGGACAAAGAGAAUAAAGACUUCAUGUU, after 17402000 WL Monte Carlo steps, where the value of \( \epsilon \) 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 ..(.(...).)((...)(...).((...)(...))).

Fig. 6. Computation of heat capacity \( c_P(T) \) for the toy sequence 5′-AGCGA-3′, hybridized to its reverse complement 3′-UCGCU-5′. (a) Graph generated by WL method described in this article. (b) Graph generated by the program UNAFold (Markham and Zuker, 2008).

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°C to 100°C, 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 WL with temperature \( T = 273°C \) (absolute zero Kelvin). It follows that \( p_H \) is the relative density of states for enthalpy. Due to the fundamental thermodynamic relation

\[
\Delta G = \Delta H - T \Delta S
\]

where \( T \) (K) is absolute temperature and \( \Delta G, \Delta H, \Delta S \), respectively, denote the change in free energy, enthalpy and entropy.
Thermodynamics of RNA structures by Wang–Landau sampling

• Compute the relative density of states $p_T$ for free energy using WL with temperature $T = 37°C (310K)$.

• 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 $p_T$ with $p_y$.

1. for all $i$ initialize $p(i)=0$
2. for $x$ ranging over enthalpy bins
3. for $y$ ranging over free energy bins
4. $z = \frac{1}{N} \sum_{i} (p_{\Delta H}(i)(p_{\Delta G}(y)))$
5. Compute the absolute density of states $p(z) = p(z)$, 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°C to 100°C, by performing two WL sampling runs, respectively, at temperatures $-37°C$ and $37°C$, 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 MPFT 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 (Flamm et al., 2001) 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 want to implement a new energy evaluation function, that allows arbitrary pseudoknots, zig-zags, etc. using energy parameters from the recent dissertation of Bon (Bon, 2009). 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 article, we have implemented the WL 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 existing software RNAsubopt in computing the density of states, but could not be benchmarked with the binning method of Cupal et al. (1996) which runs in $O(m^3n^3)$ time, for which 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 Michael Zuker, for a discussion on melting temperature and Ivo Hofacker, for informing us of the status of the program of Cupal et al. (1996), 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: Digitone Foundation (to P.C. and E.L.), in the form of a Digitone Chair of Excellence to P.C. National Science Foundation (grants DMS-0817971 and DBI-0543506 to P.C.).

Conflict of Interest: none declared.

REFERENCES


