Structural bioinformatics

Memory efficient folding algorithms for circular RNA secondary structures

Ivo L. Hofacker¹⁺ and Peter F. Stadler²¹³

¹Institute for Theoretical Chemistry, University of Vienna, Währingerstr. 17, A-1090 Vienna, Austria,
²Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics,
University of Leipzig, HärTELstrasse 16-18, D-04107 Leipzig, Germany and ³The Santa Fe Institute,
1399 Hyde Park Road, Santa Fe, NM, USA

Received on November 5, 2005; revised on December 3, 2005; accepted on January 23, 2006
Advance Access publication February 1, 2006
Associate Editor: Thomas Lengauer

ABSTRACT
Background: A small class of RNA molecules, in particular the tiny genomes of viroids, are circular. Yet most structure prediction algorithms handle only linear RNAs. The most straightforward approach is to compute circular structures from ‘internal’ and ‘external’ substrates separated by a base pair. This is incompatible, however, with the memory-saving approach of the Vienna RNA Package which builds a linear RNA structure from shorter (internal) structures only.

Result: Here we describe how circular secondary structures can be obtained without additional memory requirements as a kind of ‘pre-processing’ of the linear structures.

Availability: The circular folding algorithm is implemented in the current version of the Vienna RNA Package which can be downloaded from http://www.tbi.univie.ac.at/RNA/

Contact: ivo@tbi.univie.ac.at

1 INTRODUCTION

Most RNA molecules are linear. Circular single stranded RNAs, on the other hand, occur only in a few cases. The most prominent class are the ‘genomes’ of viroids, see (Flores et al., 2004; Tabler and Tsagris, 2004) for recent reviews. A related example is the circular RNA genome of Hepatitis Delta virus which contains a viroid-like domain, see e.g. (Gudima et al., 2004; Wadkins and Been, 2002) and the references therein. In addition, alternative splicing may lead to circular RNAs from intrinsic sequences. This appears to be a general property of nuclear group I introns (Nielsen et al., 2003) and was also observed during RNA splicing in Halobacterium volcanii (Salgía et al., 2003). Circularized C/D box snoRNAs were recently reported in Pyrococcus furiosus (Starostina et al., 2004). Circular nucleic acids, furthermore, have been investigated in the context of in vitro selection experiments (Kong et al., 2002), and they appear as intermediates in a sequencing strategy for the UTRs of RNA viruses.

While structure prediction of these fairly rare circular RNAs may appear as a rather esoteric topic, most of the examples above have functional secondary structures. Indeed, viroids were among the first RNAs for which secondary structures have been studied systematically (Steiger et al., 1984), see also Repsilber et al. (1999) for more recent work. Since viroid RNAs are short (≈200–400 nt), we have to expect significant differences between the folds of linear and circular sequences. We will demonstrate in Figure 1 that this is indeed the case.

It is therefore worthwhile to develop circular variants of at least the most common RNA folding tools; indeed algorithms for computing minimum energy folding and the computation of suboptimal structure of circular RNAs are implemented in Michael Zuker’s mfold package (Zuker, 1989, 2003). These algorithms, in fact, treat linear RNAs as exceptional variants of the circular ones. In contrast, the Vienna RNA Package http://www.tbi.univie.ac.at/RNA/ (Hofacker et al., 1994; Hofacker, 2003), optimizes the memory requirements for linear RNAs; this approach saves approximately a factor of 2 in memory as well as some CPU time. Circular RNAs, however, are non-trivial to handle in this framework. In this contribution we demonstrate how circular RNA folding can be implemented efficiently as a kind of ‘pre-processing’ step of the forward recursion and as a corresponding ‘pre-processing’ step for the the backtracking part of the folding algorithms without requiring significant additional resources or a redesign of the optimized recursion for the linear RNA case.

Circular RNA folding can therefore be included into the Vienna RNA Package without duplicating the code or compromising the efficiency of the current implementations.

This contribution is organized as follows: We briefly recall the RNA folding algorithms as implemented in the Vienna RNA Package. We then discuss the extension of the minimum free energy folding approach to circular RNAs and describe how the same ideas apply to the computation of the partition function.

2 FOLDING LINEAR RNA MOLECULES

The energy model for RNA folding is based upon carefully measured energy parameters (Mathews et al., 1999, 2004) for the loops of the RNA secondary structure [i.e. the cycles of the unique minimum cycle basis (Leydold and Stadler, 1998, http://www.combinatorics.org/)]. The energy of a loop depends on the sequence near the base pairs that are part of the loop, the length of the loop and on its type. From the biophysical point of view one distinguishes hairpin loops, stacked base pairs, bulges, true interior loops and multi(branched) loops. From an algorithmic point of view one can treat bulges, stacked pairs and true interior loops as subtypes of interior loops.
Circular RNA folding

Fij, free energy of the optimal substructure on the subsequence x[i, j] subject to the constraint that i and j form a basepair.

Mij, free energy of the optimal substructure on the subsequence x[i, j] subject to the constraint that x[i, j] is part of a multiloop and has

3. FOLDING ALGORITHMS FOR CIRCULAR RNA

A straightforward way of dealing with circular RNA molecules is to compute Cij and Mij also for the subsequences of the form x[i, n]+1[i, i]. This is implemented in the mfold package (Zuker, 2003) and described e.g. in Zuker (1989). The disadvantage of this approach is, however, that it doubles the memory requirements (and also the CPU requirements, because more matrix entries need to be computed).

As an alternative, we propose here to extend the linear folding algorithms in such a way that the circular molecules are handled as a kind of post-processing of the arrays that are computed in the linear case. This is not only memory efficient but also allows us to assess the structural differences between linear and circular sequences with just a single run of the forward recursions. (Recall that the backtracking step for minimum energy folding is fast: O(n) compared with the O(n^2) steps for filling the arrays.)

The key observation is that the only difference between the linear and the circular case is the energy of the loop that contains x_i and x_{i+1}. In the linear case, there is no energy contribution associated with the ‘exterior’ loop, while it has to be scored like any other loop in the

\begin{align}
F_{ij} &= \min \left\{ F_{i+1,j}, \min_{k \in [1,j]} C_{ik} + C_{k+1,j} \right\} \\
C_{ij} &= \min \left\{ \mathcal{H}(i,j), \min_{k \in [1,j]} C_{ik} + I(i,j,k,l), \min_{k \in [1,j]} M_{i+1,j} + M_{0,j} \right\} \\
M_{ij} &= \min \left\{ \min_{k \in [1,j]} (a - i - 1)c + C_{a+1,j} + b, \min_{k \in [1,j]} M_{i+1,j} + b, M_{i+1,j} + c \right\} \\
M_{ij} &= \min \left\{ M_{i,j+1} + c, C_{ij} + b \right\}.
\end{align}

These recursions are directly derived from the structure decomposition shown in Figure 2. The corresponding recursions for the partition function are obtained by replacing minimum operations with sums and additions with multiplications (McCaskill, 1990).

The computation of the minimum free energy structure requires to store only the arrays F, C and M. In addition, the full M^2 array is required for the more elaborate backtracking procedure of the RNAsubopt program (Wuchty et al., 1999) which produces all RNA secondary structures within a given energy interval above the ground state. Similarly, uniqueness of the decomposition is necessary for partition function algorithms, see Section 4.

We consider an RNA sequence x of length n. Hairpin loops are uniquely determined by their closing pair k, l. The energy of a hairpin loop is

\[\mathcal{H}(k,l) = \mathcal{H}(x_k, x_{k+1}, \ell_1, x_{l-1}, x_l),\]

where \(\ell_1\) is the length of the loop (expressed as the number of its unpaired nucleotides). Each interior loop is determined by the two base pairs enclosing it. Its energy is tabulated as

\[I(k,l; p, q) = I(x_k, x_{k+1}, \ell_1, x_{p-1}, x_p, x_q, x_{q+1}, \ell_2, x_{l-1}, x_l),\]

where \(\ell_2\) is the length of unpaired strand between k and p and \(\ell_2\) is the length of the unpaired strand between q and l. Symmetry of the energy model dictates \(I(k,l; p, q) = I(q, p; l, k)\). If \(\ell_1 = \ell_2 = 0\) we have a (stabilizing) stacked pair, if only one of \(\ell_1\) and \(\ell_2\) vanish we have a bulge. For multiloops, finally we have an additive energy model of the form \(M = a + b \times \beta + c \times \ell\) where \(\ell\) is the length of multiloop (again expressed as the number of unpaired nucleotides) and \(\beta\) is the number of branches, not counting the branch in which the closing pair of the loop resides. As described here, the multiloop energy is independent of the sequence of the closing pair. Sequence dependence is introduced, however, by the dangling end contributions which we briefly discuss at the end of Section 3.

RNA folding algorithms are based on decomposing the set of possible structures into sets of smaller structures. This decomposition can be chosen such that each possible structure appears in exactly one of the subcases. In the course of the ‘normal’ RNA folding algorithm for linear RNA molecules as implemented in the Vienna RNA Package (Hofacker et al., 1994; Hofacker, 2003) the following arrays, which correspond to different structural components in Figure 2, are computed for \(i < j\):

\[F_{ij}, \text{free energy of the optimal substructure on the subsequence } x[i, j].\]

\[C_{ij}, \text{free energy of the optimal substructure on the subsequence } x[i, j] \text{ subject to the constraint that } i \text{ and } j \text{ form a basepair.}\]

\[M_{ij}, \text{free energy of the optimal substructure on the subsequence } x[i, j] \text{ subject to the constraint that } x[i, j] \text{ is part of a multiloop and has at least one component, i.e. a sub-sequence that is enclosed by a base pair.}\]
circular case. Hence we have to distinguish the types of ‘exterior’ loops, Figure 3.

Exterior hairpin. If the exterior loop is a hairpin, then there is a base pair \( p, q \) such that both \( x[p, 1] \) and \( x[q + 1, n] \) are unpaired. The optimal energy of such a structure is
\[
F_{H} = \min_{p < q} \{ C_{pq} + \mathcal{H}(q, p) \}
\]
where \( \ell = p - 1 + (n - q + 1) \) is length of the hairpin loop and \( 1 \leq p < q \leq n \).

Exterior interior loop. In this case, the ‘exterior loop’ contains the closing pairs \( k, l \) and \( p, q \) of exactly two components. Thus
\[
F_{I} = \min_{k < q < l} \{ C_{pq} + C_{kl} + \mathcal{I}(q, p, l, k) \}
\]
where \( \ell_1 = n - q + k - 1 \) and \( \ell_2 = p - l - 1 \). In practice, the size \( \ell \) of an interior loop is limited to \( \ell \leq m \), typically \( m = 30 \). Thus \( F_{I} \) can be computed in \( O(n^2) \) time without additional memory requirements.

Exterior multiloop. Generalizing the approach for the interior loops, we can view an exterior multiloop as a multiloop with at least three branches on the sequence interval form 1 to \( n \). Starting from \( M^1 \) we compute the linear auxiliary array \( M^2 \) containing the optimal energy of \( x[k, n] \) given that the sequence interval is contained in a multiloop, has exactly two components, and starts with a base pair \( k, h \). We obtain
\[
M^2_{kn} = \min_{k < q < l < m} (M^1_{ku} + M^1_{vl} + a).
\]
This array requires only \( O(n) \) memory and can be computed in \( O(n^2) \) time. A multiloop with at least three components can now be constructed from a piece with at least one component at the beginning of the sequence and a piece that contains exactly two components (with first closing pair \( k + 1, v \), for some \( k < v < n - 2 \)):
\[
F^*_{M} = \min_{k < v < n - 2} \{ M^1_{k,v} + M^2_{v+1,n} + a \}.
\]
The multiloop case thus can be dealt with in quadratic time with only linear memory overhead.

The minimum free energy structure of the folded circular molecule is therefore
\[
F^* = \min \{ F^*_{H}, F^*_{I}, F^*_{M} \}.
\]
Backtracking. Backtracking is straightforward with this approach: First we determine whether the optimal ‘exterior loop’ is a hairpin \((F_1 = F_{li})\), an interior loop \((F_1 = F_1)\) or a multiloop \((F_1 = F_{lk})\). Depending on the result we determine either

1. \(p,q\) such that \(F_{li} = C_{pq} + \mathcal{H}(q,p)\), or
2. \(k,l\) and \(p,q\) such that
   \[ F_1 = C_{pq} + 2u + \mathcal{I}(q,p,l,k), \]
3. \((a)\) \(k\) such that \(M_k = M_{i,k} + M_{k+1,n} + a\), and then
   \((b)\) \(u\) such that \(M_{k,n} = M_{k,a} + M_{a+1,n}\).

The next step already follows the normal backtracking procedure of the linear folding problem.

Dangling ends. The Vienna RNA Package implements three different models for handling the so-called dangling-end contributions that arise when an unpaired nucleotide stacks with an adjacent base pair. These contributions can be (1) ignored, (2) taken into account for every combination of adjacent bases and base pairs or (3) a more complex model can be used in which the unpaired base can stack with at most one base pair. The latter model strictly speaking violates the secondary structure model in that an unpaired base \(b\) between two base pairs \((x_p, x_{p+1})\) and \((x_q, x_{q+1})\) has three distinct states with different energies: \(x_b\) does not stack to its neighbors, \(x_b\) stacks to \(x_{p+1}\), or \(x_{q+1}\). (The secondary structure model, in contrast, distinguishes only unpaired from paired bases.) The folding algorithm then minimizes over these possibilities. In cases (1) and (2) one can absorb the dangling-end contributions in the loop energies. In case (3), however, they have to be treated explicitly, which is done in the forward recursions already for all cases with the exception of the dangling end contribution reaching across the ‘gap’ \(1 - n\). The cases unpaired \(x_b\) stacks to paired \(x_p\) and unpaired \(x_b\) stacks to paired \(x_q\) have to be treated separately, adding two additional subcases to the multi-loop recursion above. Even more subclasses are needed if one wants to allow also for co-axial stacking of helices in the multiloop.

An important observation about the recursions (2–5) is that each possible secondary structure is counted exactly once, i.e. the recursions are non-redundant. This is important when one is interested in enumerating structures as, for example, in the RNAsubopt program. This property is also crucial for the partition function calculations discussed in the next section. For the purpose of energy minimization, however, it is not necessary. One can therefore replace Equation (4) by

\[
M^2_{kn} = \min_{k \leq a \leq n} \{M_{kn} + M_{a+1,n}\},
\]  

and reinterpret \(M^2\) as the contribution of segments with at least two branches in a multiloop. As a consequence, the \(M^2\) array does not need to be stored and the memory requirements of the minimum free energy folding are the same as in the linear case up to a the auxiliary array \(M^2\) of size \(n\).

### 4 PARTITION FUNCTION

It is straightforward to translate the recursions (2–5) into recursions for the partition function because they already provide a partition of the set of all secondary structures that can be formed by the sequence \(x\). In the following we suppress the factor \(1/RT\) in the Boltzmann factors of the energy parameters, i.e. we assume that the energy parameters are already scaled relative to the thermal energy. Equation (2–5) then become

\[
Z_{kn}^{M_2} = \sum_{a} Z_{kn}^{M_1} Z_{a+1,n}^{M_1}
\]

\[
Z_{H}^{M} = \sum_{p,q} Z_{pq}^{B} e^{-\mathcal{H}(q,p)}
\]

\[
Z_{I}^{M} = \sum_{k < q < q'} Z_{kq}^{B} e^{-\mathcal{I}(k,l,p,q)}
\]

\[
Z_{M} = \sum_{k} Z_{k-l,n}^{M} Z_{k+1,n}^{M}
\]

\[
Z'_{M} = Z_{H}^{M} + Z_{I}^{M} + Z_{M}^{M}.
\]

The probability \(P_{kl}\) of a base pair \(kl\) can be represented, in the simplified version of the Nussinov algorithm (Nussinov et al., 1978), as

\[
P_{kl} = P_{kl}^{0} + \sum_{p < k, q < l} P_{pq}^{0} \frac{Z_{kl,p+1,q+1} e^{-x}}{Z_{pq}^{B}}.\]

See Figure 4. Here \(P_{kl}^{0}\) is the probability of that \(kl\) is a closing pair contained in the exterior loop. This is the only term that differs from the linear case. For the full energy model we can use the same logic, but we need to consider the individual loop types separately. In detail we obtain (MacCaskill, 1990):

\[
P_{kl}^{0} = P_{kl}^{0} + \sum_{p < k, q < l} P_{pq}^{0} \frac{Z_{kl,p+1,q+1} e^{-x}}{Z_{pq}^{B}} \left( e^{-\mathcal{H}(q,p,k,l)} + \left( \sum_{p < k < c < l} Z_{p,c}^{B} Z_{c+1,n}^{M} \right) e^{-\mathcal{I}(q,p-1,k,l)} + \left( \sum_{i \leq c < q} Z_{c,i}^{M} Z_{i+1,n}^{M} \right) e^{-\mathcal{I}(k-1,q-1)} \right).
\]

The first term covers the case where \(p, q, k, l\) delimit an interior loop. The remaining three terms cover the multi-loop case with the three sub-cases that \(kl\) delimits the most \(3',\) the most \(3'\) or an intermediate branch, respectively.

The contribution \(P_{kl}^{0}\) covers the cases in which the basepair \(kl\) is part of the ‘exterior’ loop. In the linear case we have simply

\[
P_{kl}^{0} = \frac{Z_{1,k-1}^{1} Z_{l+1,n}^{1}}{Z_{kn}^{M}}.
\]
In the circular case we have to consider the three possible loop types for the ‘exterior’ loop separately. This yields:

\[
P_{kl} = \frac{Z_M}{Z_{MC}} \left( \mathcal{H}(l,k) \right)_{\text{hairpin}} \right. \\
+ \sum_{p,q} Z_{pq}^l e^{-\frac{I(p,q,l)}{a}} \sum_{l' \in l, k' \in k} Z_{l'k'}^p e^{-\frac{I(p,q,l)}{c}} \left( \text{Interior left} \right) \\
+ \sum_{l \in l, p \in p} Z_{l'k'}^p e^{-\frac{I(p,q,l)}{c}} \left( \text{Interior right} \right) \\
+ \sum_{p \in p, l \in l} Z_{pq}^l e^{-\frac{I(p,q,l)}{c}} \left( \text{Multi middle} \right) \\
+ \sum_{p \in p, l \in l} Z_{pq}^l e^{-\frac{I(p,q,l)}{c}} \left( \text{Multi left} \right) \\
+ \sum_{p \in p, l \in l} Z_{pq}^l e^{-\frac{I(p,q,l)}{c}} \left( \text{Multi right} \right)
\]

For given \(k\) and \(l\) this expression can be evaluated in linear time without additional memory requirements. It follows that the base pairing probability matrix \(P\) for the case of circular RNAs can be computed with a constant additional factor in CPU time and negligible additional memory requirements.

5. CONCLUDING REMARKS

Circular RNA folding is being added as an additional feature to the Vienna RNA Package. The energy minimization is already available via cvs, the implementation of the circular version of RNAalifold (Hofacker et al., 2002) is in progress. This tool computes the consensus structure of a set of aligned RNA sequences. Algorithmically, it is very similar to the energy minimization described above.

The main applications for these features are a more systematic analysis of viroid structures and circular snoRNAs. In conjunction with alignment algorithms for circular sequences (Gregor and Thomason, 1993; Maes, 1990) one can use circular RNAalifold to obtain consensus structures. The adiidot tool (Hofacker et al., 1998; Hofacker and Stadler, 1999) can be applied without changes to the problem of identifying evolutionarily conserved RNA secondary structure motifs in otherwise structurally variable RNA motifs. The circular version of RNAsubopt (Wuchty et al., 1999) will be of particular interest for a detailed understanding of the structural changes in viroid RNAs.

ACKNOWLEDGEMENTS

This work was supported in part by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung, Project No. P15893, and by the German DFG Bioinformatics Initiative BIZ-6/1-2.

Conflict of Interest: none declared.

REFERENCES


Nielsen,H. et al. (2003) The ability to form full-length intron RNA circles is a general property of nuclear group I introns. RNA, 9, 1464–1475.


