JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 2, 2008
© Mary Ann Liebert, Inc.
Pp. 139–163
DOI: 10.1089/cmb.2007.0198

Novel and Efficient RNA Secondary Structure Prediction Using Hierarchical Folding

HOSNA JABBARI, ANNE CONDON, and SHELLY ZHAO

ABSTRACT

Algorithms for prediction of RNA secondary structure—the set of base pairs that form when an RNA molecule folds—are valuable to biologists who aim to understand RNA structure and function. Improving the accuracy and efficiency of prediction methods is an ongoing challenge, particularly for pseudoknotted secondary structures, in which base pairs overlap. This challenge is biologically important, since pseudoknotted structures play essential roles in functions of many RNA molecules, such as splicing and ribosomal frameshifting. State-of-the-art methods, which are based on free energy minimization, have high run-time complexity (typically $\Theta(n^5)$ or worse), and can handle (minimize over) only limited types of pseudoknotted structures. We propose a new approach for prediction of pseudoknotted structures, motivated by the hypothesis that RNA structures fold hierarchically, with pseudoknot-free (non-overlapping) base pairs forming first, and pseudoknots forming later so as to minimize energy relative to the folded pseudoknot-free structure. Our HFold algorithm uses two-phase energy minimization to predict hierarchically formed secondary structures in $O(n^3)$ time, matching the complexity of the best algorithms for pseudoknot-free secondary structure prediction via energy minimization. Our algorithm can handle a wide range of biological structures, including kissing hairpins and nested kissing hairpins, which have previously required $\Theta(n^6)$ time.

Key words: computational molecular biology, RNA, secondary structure.

1. INTRODUCTION

THE RNA MOLECULES aid in translation and replication of the genetic code, catalyze cellular processes, and regulate the expression level of genes (Dennis, 2002). Structure is key to the function of RNA molecules, and so methods for predicting RNA structure from the base sequence are of great value. Currently, prediction methods focus on secondary structure—the set of base pairs that form when the RNA molecule folds. There has been significant success in prediction of *pseudoknot-free* secondary structures, which have no crossing base pairs (for an example of a pseudoknot-free structure, see Fig. 1). State-of-the-art prediction algorithms, such as Mfold (Mathews et al., 1999) or RNAfold (Hofacker et al., 1994) find the structure with *minimum free energy* (MFE) from the set of all possible pseudoknot-free secondary

Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada.

structures. The energy of a structure is estimated as the sum of energies of *loops* that form when the molecule folds, where the loop energy values are provided by biologists.

While many small RNA secondary structures are pseudoknot free, pseudoknots do arise frequently in biologically important RNA molecules, both in the cell (Staple and Butcher, 2005; van Batenburg et al., 2001), and in viral RNA (Deiman and Pleij, 1997). Examples include simple H-type pseudoknots, with two interleaved stems, which are essential for certain catalytic functions and for ribosomal frameshifting (Alam et al., 1999), as well as kissing hairpins, which are essential for replication in the coxsackie B virus (Melchers et al., 1997).

Unfortunately, MFE pseudoknotted secondary structure prediction is NP-hard (Akutsu, 2000; Lyngsø, 2004; Lyngsø and Pedersen, 2000), even for a simple energy model that depends on base pairs but not on unpaired bases. Polynomial-time MFE-based approaches to pseudoknotted structure prediction have been proposed (Akutsu, 2000; Dirks and Pierce, 2003; Reeder and Giegerich, 2004; Rivas and Eddy, 1999; Uemura et al., 1999), with respect to various sum-of-loops energy models for pseudoknotted structures, which find the MFE structure for a given input sequence, from a restricted class of structures. A class of structures can be defined by specifying allowable patterns of interleaving among base pairs. For example, Mfold and RNAfold handle the class of pseudoknot-free secondary structures; we provide more examples later. We say that a structure *R can be handled* by a given algorithm if R is in the class of structures over which the algorithm optimizes.

Algorithms for MFE pseudoknotted secondary structure prediction trade off run-time complexity and *generality*—that is, the class of structures over which the algorithms optimize. For example, kissing hairpins are not in the class of structures handled by the $\Theta(n^5)$ algorithms of Akutsu (2000) and Dirks and Pierce (2003) but are in the class handled by the $\Theta(n^6)$ algorithm of Rivas and Eddy (1999). (We note that, even when the true structure R for a sequence is handled by an algorithm, the algorithm still may not correctly predict R, because correctness depends not only on the generality of the algorithm but also on the energy model and energy parameters.)

Our work is motivated by two limitations of MFE-based algorithms for pseudoknotted secondary structure prediction: they have high time complexity and ignore the folding pathway from unfolded sequence to stable structure. Several experts have provided evidence for, and support, the *hierarchical folding hypothesis* (Mathews, 2006; Tinoco and Bustamante, 1999), which is succinctly stated by Tinoco and Bustamante as follows: "An RNA molecule [has] a hierarchical structure in which the primary sequence determines the secondary structure which, in turn, determines its tertiary folding, whose formation alters only minimally the secondary structure." (These and other authors consider the initially-formed secondary structure to be pseudoknot-free, and refer to base pairs that form pseudoknots as part of the tertiary structure. However, here we refer to all canonical base pairs, namely A-U, C-G, and G-U, as secondary structure.) We note that while the hierarchical folding hypothesis is a common assumption, some counter examples have been reported, notably formation of the structure of a subdomain of the Tetrahymena thermophila group I intron ribozyme (Wu and Tinoco, 1998). However, even in this case, 15 of the 19 base pairs in the initially-formed pseudoknot-free secondary structure are retained upon formation of tertiary structure, and the 4 missing base pairs lie at the ends of stems.

In this paper, we present a novel and efficient algorithm to predict RNA secondary structures, in a manner consistent with a natural formalization of the hierarchical folding hypothesis. We consider the problem of predicting the secondary structure as follows: given a sequence S and a pseudoknot-free secondary structure G (a set of base pairs), find a pseudoknot-free secondary structure G' (a set of base pairs disjoint from G) for S, such that the free energy of $G \cup G'$ is less than or equal to the free energy of $G \cup G''$ for all pseudoknot-free structures $G'' \neq G'$.

As with algorithms for MFE pseudoknotted secondary structure prediction, algorithms for hierarchical-MFE secondary structure prediction may handle a restricted class of structures. That is, the type of structure formed by $G \cup G'$ may have restricted patterns of interleaving among base pairs. Since both G and G' are pseudoknot-free, the most general class of structures that could be handled by an algorithm for hierarchical-MFE secondary structure prediction would be the *bi-secondary* structures of Witwer et al. (2004)—those structures which can be partitioned into two pseudoknot-free secondary structures G and G'. There is no known efficient method to solve the hierarchical-MFE prediction for the class of bi-secondary structures. Instead, we suggest a solution with respect to a subclass of the bi-secondary structures, which we call *density-2* structures, explained in Section 2.

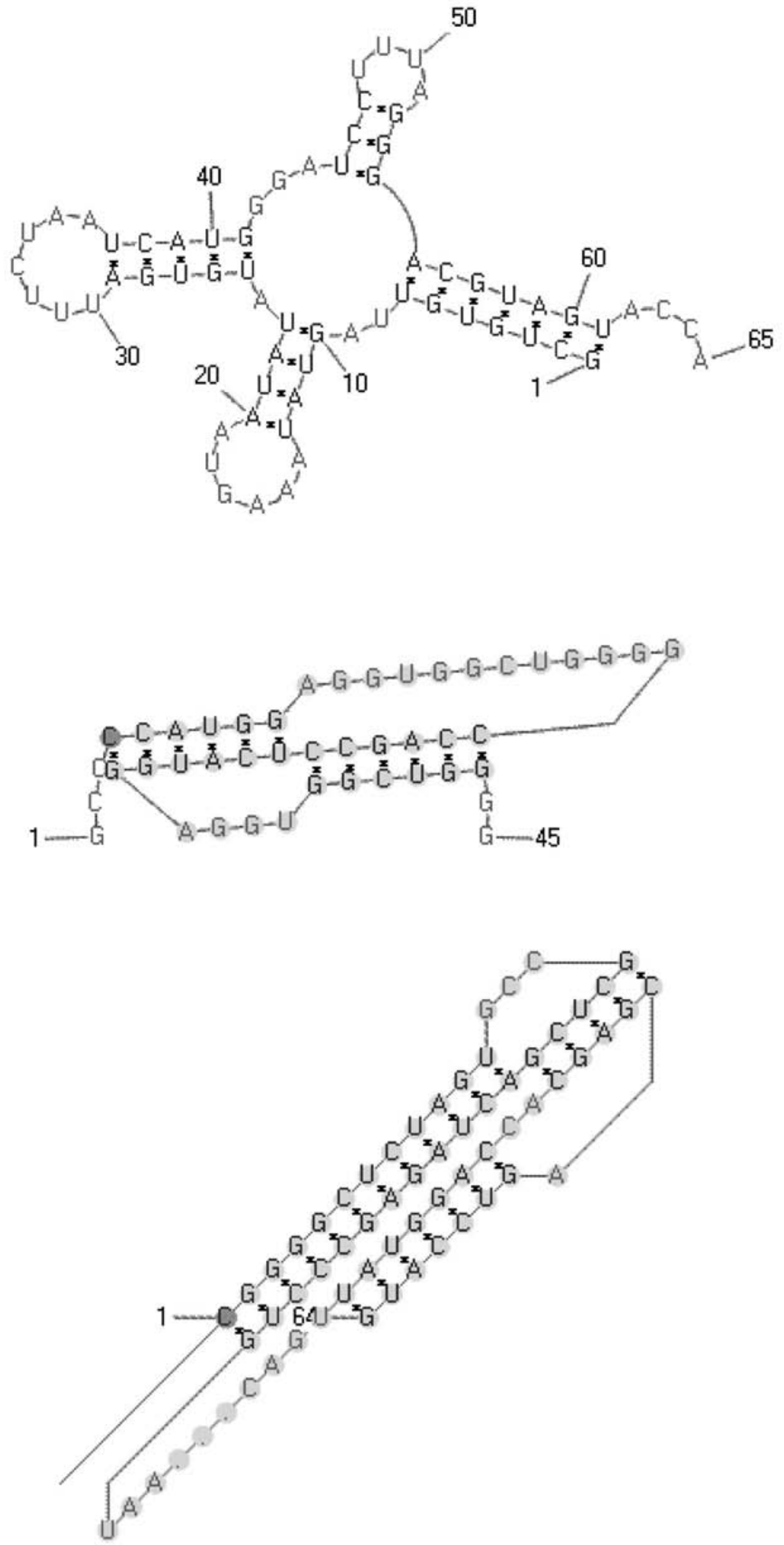


FIG. 1. A pseudoknot-free structure (**top**), an H-type pseudoknotted structure (**center**) and a kissing hairpin (**bottom**). Figures were generated by PseudoViewer (Han et al., 2002).

The class of density-2 structures is quite general, including H-type pseudoknots and kissing hairpins, as well as structures containing nested instances of these structural motifs. The only known algorithm for predicting MFE nested kissing hairpins, that of Rivas and Eddy, requires $\Omega(n^6)$ time. Rastegari and Condon (2007) showed that, out of a set of over 1100 biological structures, all but nine are density-2 (when isolated base pairs are removed), and six of these nine are also not in the class handled by Rivas and Eddy's algorithm.

In Section 2, we present some useful background information and notations pertaining to RNA structure prediction. In Section 3, we provide a theoretical basis for the correctness of our HFold algorithm. In Section 4, we present HFold, a dynamic programming algorithm that solves the hierarchical-MFE secondary structure prediction problem for the class of density-2 secondary structures in $O(n^3)$ time and $O(n^2)$ space. We start by a high level description of HFold and proceed with details of different recurrences in our dynamic programming approach. We then present our conclusion and future work in Section 5.

2. BACKGROUND ON RNA SECONDARY STRUCTURE

An RNA molecule is a sequence of nucleotides, or bases, of which there are four types: Adenine (A), Guanine (G), Cytosine (C), and Uracil (U). The molecule has chemically distinct ends, called the $5'$ and $3'$ ends. We model an RNA molecule as a sequence over the alphabet $\{A, C, G, U\}$, with the left end of the sequence being the $5'$ end. Throughout, n denotes the length of an RNA sequence. We index the bases consecutively from the $5'$ end starting from 1, and refer to a base by its index.

When an RNA molecule folds, bonds may form between canonical pairs of bases, where each base may pair with at most one other base. The canonical base pairs, which form the secondary structure, are the Watson-Crick pairs A-U and C-G, as well as the wobble pair G-U (Fig. 2). A *secondary structure* R is a set of pairs $i.j$, $1 \leq i < j \leq n$, such that no index occurs in more than one pair and pair of bases indexed i and j are canonical. The pair $i.j$ denotes that the base indexed i is paired with the base indexed j.

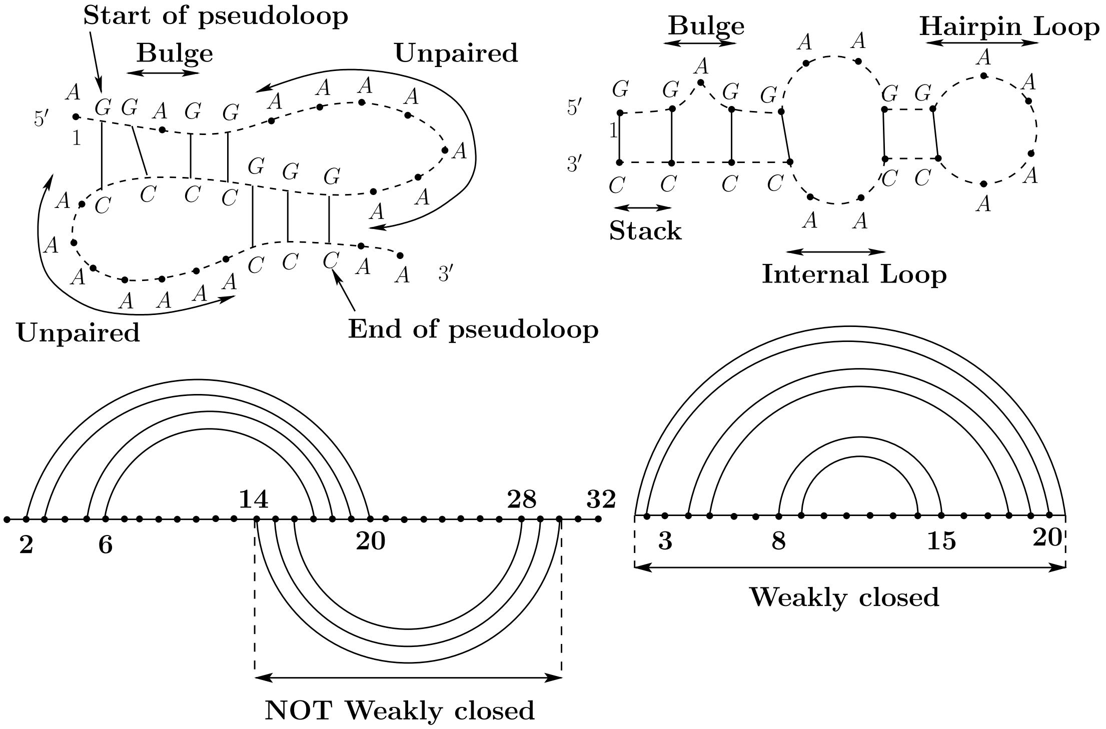


FIG. 2. An H-type pseudoknotted structure (**left**) and a pseudoknot-free structure (**right**), in graphical (**top**) and arc diagram (**bottom**) formats.

We use the following notation when describing our algorithms. These definitions are mostly taken from and illustrated in the work of Rastegari and Condon (2007). Throughout, definitions are with respect to a fixed secondary structure R. Generally we use R to refer to a structure that may be pseudoknotted (that is, contains at least one pseudoknotted base pair), and use G to refer to a structure that we know to be pseudoknot-free.

2.1. Notation

Notation is as follows:

- $bp_R(i)$: We let $bp_R(i)$ denote the index of the base that is paired with base i in R, if any; otherwise $bp_R(i) = 0$.
- **paired**(R, i): true if and only if i is paired in the structure R.
- **cross:** if $i.j$ and $i'.j'$ are in R, and $i < i' < j < j'$, we say that pair $i.j$ crosses pair $i'.j'$ from the left (and $i'.j'$ crosses $i.j$ from the right).
- **pseudoknotted base pair:** We say that $i.j$ is a pseudoknotted base pair if for some other base pair $i'.j'$ in structure R, $i.j$ crosses $i'.j'$. We also refer to i and j as *pseudoknotted* base indices.
- **pseudoknot-free secondary structure:** If there are no pseudoknotted base pairs in a given structure, it is a pseudoknot-free secondary structure.
- **cover:** Base pair $i.j$ *covers* base k if $i < k < j$ and there is no other base pair $i'.j' \in G$ with $i < i' < k < j' < j$. In this case, we denote $i.j$ by $cover(G, k)$. Otherwise $cover(G, k) = (-1, -1)$.
- **isCovered**(G, k): true if and only if some base pair of G covers k.

2.2. Region and loop classification and related definitions

- **region** $[i, j]$: Sequence of indices between i and j inclusive, where i and j are called the left and right borders of the region respectively.
- $G_{i,j}$: The set of base pairs of G contained in region $[i, j]$, i.e., $G \cup [i, j] \times [i, j]$.
- **disjoint regions:** Two regions $[i, j]$ and $[i', j']$ are disjoint if no index is in both regions, i.e., $j < i'$ or $j' < i$.
- **empty**$(R, [i, j])$: true if region $[i, j]$ contains no base pair in R. Formally, $\forall k, i \leq k \leq j, \overline{paired(R, k)}$.
- **weakly closed region:** A region is weakly closed if no base pair connects a base in the region to a base outside the region. Formally, $[i, j]$ is weakly closed if and only if for all $k \in [i, j]$, either $bp_R(k) \in [i, j]$ or $bp_R(k) = 0$. *Weakly closed*$(R, [i, j])$ is true if and only if $[i, j]$ is a weakly closed region of R.
- **closed region:** A weakly closed region $[i, j]$, with at least two bases, is closed if it cannot be partitioned into two smaller weakly closed regions. Formally, $[i, j]$ is closed if and only if $i < j$, $[i, j]$ is weakly closed, and for all $l \in [i, j-1]$, neither $[i, l]$ nor $[l+1, j]$ is weakly closed. Note that if $[i, j]$ is closed then both i and j must be paired (although not necessarily with each other).
- **pseudoknot-free closed region:** A closed region $[i, j]$ that does not contain any pseudoknotted base pairs.
- **pseudoknotted closed region:** A closed region $[i, j]$ of a structure R such that $i.bp_R(i)$ and $bp_R(j).j$ are pseudoknotted base pairs. We refer to indices i and j as the left and right borders of the pseudoknotted region $[i, j]$.

  Note: closed regions must be either pseudoknot-free or pseudoknotted.
- **directly banded in:** For a pseudoknotted base pair $i.j$, we say $i.j$ is *directly banded in* base pair $i'.j'$ and write $i.j \preceq i'.j'$ if:
  (1) $i' < i < j < j'$, and
  (2) $[i'+1, i-1]$ and $[j+1, j'-1]$ are weakly closed regions.
- **band:** Let $i.j$ and $i'.j'$ be the first and the last base pairs in a maximal chain of $\preceq$. Then, $[i, i'] \cup [j', j]$ is a band. We call $[i, i']$ and $[j', j]$ the band regions, and call $i'.j'$ and $i.j$ the inner and outer base pairs of the band respectively. For example, there are three bands in Figure 3: $[1, 2] \cup [22, 23]$, $[18, 19] \cup [33, 34]$ and $[27, 27] \cup [39, 39]$.

  We refer to i and j as the left and the right borders of the band respectively.
- **inside a band:** A region $[i, j]$ is inside a band $[i_1, i'_1] \cup [j'_1, j_1]$, if either $i_1 < i \leq j < i'_1$ or $j'_1 < i \leq j < j_1$ is true.

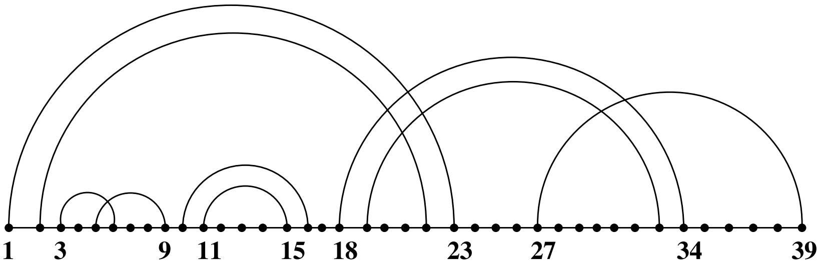


FIG. 3. Pseudoknot.

- **band associated with closed region:** We say that band $[i, i'] \cup [j', j]$ is *associated with* closed region $[i'', j'']$ if $[i, i']$, and thus $[j', j]$, are subregions of $[i'', j'']$ but are not subregions of any closed subregions of $[i'', j'']$. For example, in Figure 3, the three bands $[1, 2] \cup [22, 23]$, $[18, 19] \cup [33, 34]$ and $[27, 27] \cup [39, 39]$ are associated with closed region $[1, 39]$, whereas $[3, 3] \cup [6, 6]$ and $[5, 5] \cup [9, 9]$ are associated with closed region $[3, 9]$.
- **unpaired bases associated with closed region** $[i, j]$: These are the unpaired bases in $[i, j]$ but not in any closed region or band region which are subregions of $[i, j]$. For example, in the structure of Figure 3, the unpaired bases associated with closed region $[1, 39]$ are 17, 20, 21, 24-26, 28-32, and 35-38.
- **base pairs associated with closed region** $[i, j]$: These are the base pairs in $[i, j]$ but not in any closed region or band region which are subregions of $[i, j]$.
- **hairpin loop** (or **hairpin):** A hairpin loop contains a base pair $i.j$ and the bases in $[i + 1, j - 1]$ that are all unpaired. Formally, the tuple (i, j) defines a hairpin loop in a secondary structure if i and j are paired and $[i + 1, j - 1]$ is an empty region. $i.j$ is called the closing base pair of the hairpin loop. The hairpin marked in Figure 2 contains four unpaired bases.
- **internal loop:** An internal loop, sometimes called interior loop, contains two base pairs $i.j$ and $i'.j'$, and the bases in $[i, i'] \cup [j', j]$ that are all unpaired. Formally, the tuple (i, i', j', j), with $i < i' < j' < j$, defines an internal loop if $i.j$ and $i'.j'$, and $[i + 1, i' - 1]$ and $[j' + 1, j - 1]$ are empty regions. $i.j$ and $i'.j'$ are called the closing base pairs of the internal loop.
- **stacked loop:** A stacked loop, also called stacked pair, contains two consecutive base pairs. The tuple $(i, i + 1, j - 1, j)$ defines a stacked pair if $i.j$ and $(i + 1).(j - 1)$ are in R. A *stem* or *helix* is made of consecutive stacked loops.

  Note that, in fact, a stacked loop is also a special case of an internal loop, with no unpaired bases on either side.
- **bulge loop:** A bulge loop, or simply bulge, is a special case of an internal loop, which has no unpaired base on one side, and at least one unpaired base on the other side. Formally, the tuple (i, i', j', j), with $i < i' < j' < j$, defines a bulge loop if $i.j$ and $i'.j'$, and either $i' = i + 1$ and $[j' + 1, j - 1]$ is an empty region or $[i + 1, i' - 1]$ is an empty region and $j' = j - 1$.
- **spans a band:** There are two types of internal loops, stacked loops and bulge loops; those for which the closing base pair, $i.j$, is not pseudoknotted and those for which $i.j$ is pseudoknotted. In the latter case, we say that the loop spans a band.
- **multi-branched loop:** There are two types of multi-branched loops, or multiloops, depending on whether or not they span a band:
  (1) Let $[i, j]$ be a closed region which is not pseudoknotted, and has at least two closed subregions, or a pseudoknotted subregion. Then the unpaired bases and base pairs associated with $[i, j]$ form a multiloop.
  (2) Let $i.j$ be a pseudoknotted base pair and $i'.j' \preceq i.j$, where at least one of the (weakly closed) regions $[i + 1, i' - 1]$ and $[j' + 1, j - 1]$ is not empty. Then the unpaired bases and base pairs in the band region $[i, i'] \cup [j', j]$ comprise a multiloop that spans a band.

  For both types of multiloop, we say that $i.j$ is the closing base pair of the multiloop. Each closed subregion of $[i, j]$ is called a *branch* of the corresponding multiloop.
- **pseudoloop:** Let $[i, j]$ be a pseudoknotted closed region. Then the unpaired bases and base pairs associated with $[i, j]$, together with the inner and outer base pairs of the bands associated with $[i, j]$,

form a pseudoloop of region $[i, j]$. The base pairs $i.bp(i)$ and $bp(j).j$ are the closing base pairs of the pseudoloop. The pseudoloop is an *exterior pseudoloop* if region $[i, j]$ is not a subregion of any other region.

- **closed region associated with pseudoloop:** We say that closed region $[i', j']$ is *associated with* pseudoloop of region $[i, j]$, if $[i', j']$ is a closed proper subregion of $[i, j]$ but not a subregion of any closed subregion of $[i, j]$. For example, in Figure 3, closed regions $[3, 9]$ and $[10, 16]$ are associated with pseudoloop $[1, 39]$ but closed region $[11, 15]$ is *not* associated with pseudoloop $[1, 39]$.
- **inside a pseudoloop:** We say that the structure $R_{i,j}$ is inside a pseudoloop if $[i, j]$ is a proper weakly closed subregion of a pseudoloop but not a subregion of any closed subregion of the pseudoloop.

2.3. Bi-secondary and density-2 structures

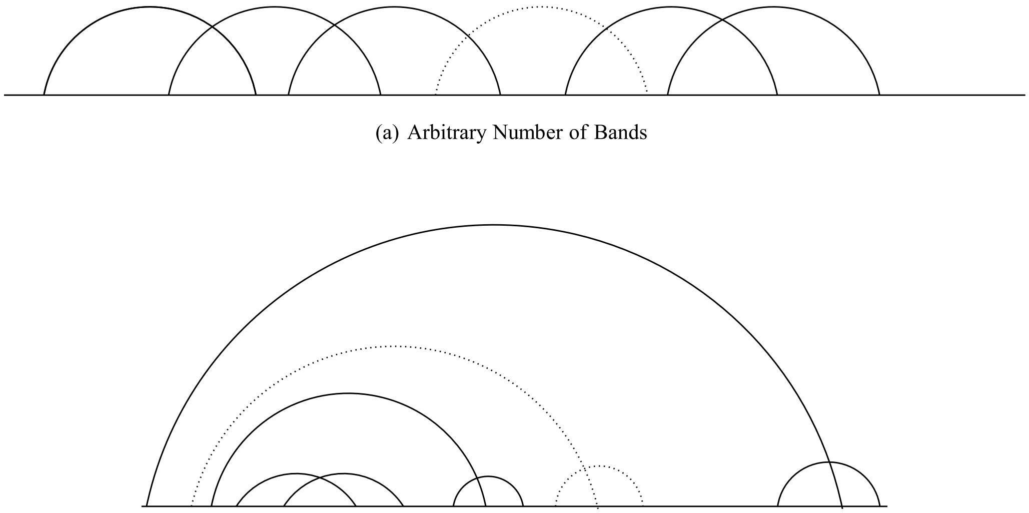
The bi-secondary and density-2 structures are as follows:

- **bi-secondary structures:** Witwer et al. (2004) introduced a definition of "bi-secondary structure," which is a union of two disjoint pseudoknot-free secondary structures. The pseudoknotted secondary structures we can handle in our algorithm are a subset of the bi-secondary structures.
- **density:** We define density as follows: Let L be a pseudoloop and $i.bp(i)$ and $bp(j).j$ be the closing base pairs of L. We say a band $[i_1, i_1'] \cup [j_1', j_1]$ crosses k if $i_1 \leq k \leq j_1$. Let $\#B(L, k)$ be the number of bands associated with L that cross k. Then the density of L is:

$$density(L) = \max_{i \leq k \leq j} (\#B(L, k)) \tag{1}$$

The density of a structure, R, is the maximum density of L over all pseudoloops L of R. We say R is a density-2 structure if the density of R is at most 2. Figure 4 illustrates density-2 secondary structures. Figure 5 shows a bi-secondary structure that is not a density-2 structure.

- **prefix:** Let G_{ij} be a pseudoknot-free structure over region $[i, j]$. Let R_{ij} be a density-2 structure over region $[i, j]$ containing G_{ij}. We say that R_{ij} is a **prefix of a density-2 pseudoloop with respect to** G_{ij}, if i is the left border of the first (leftmost) band associated with a pseudoloop of R_{ij}, and j is either
  1. the right border of a closed region associated with the pseudoloop,
  2. the right border of the pseudoloop starting at i,


(a) Arbitrary Number of Bands

(b) Arbitrary Depth of Bands

FIG. 4. Density-2 structures.

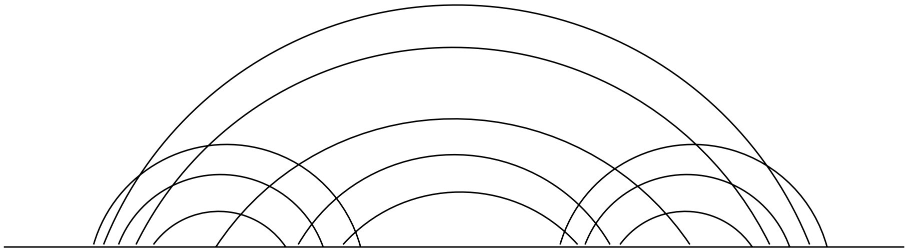


FIG. 5. Bi-secondary structure that is not density-2.

3. the rightmost border of any band of $R_{ij} - G_{ij}$ except the first band, or
4. an unpaired base associated with the pseudoloop that is not inside the first two bands (and not inside any closed subregion).

See Figure 6 for an example of a pseudoloop and indications of prefixes of pseudoloop.

2.4. *Energy model*

Computational methods for predicting the secondary structure of an RNA or DNA molecule are based on models of the free energy of loops. The parameters of these models are driven in part by current understanding of experimentally determined free energies, and in part by what can be incorporated into an efficient algorithm. The free energy of a loop depends on temperature; throughout we assume that the temperature is fixed.

2.4.1. Pseudoknot-free energy model. We first summarize the notation used to refer to the free energy of pseudoknot-free loops, along with some standard assumptions that are incorporated into loop free energy models. We refer to a model that satisfies all of our assumptions as a standard free energy model. This model is somewhat simpler than that underlying Mfold and Simfold, but our algorithm can be extended to their more detailed model.

- $e_H(i, j)$: gives the free energy of a hairpin loop closed by $i.j$.
- $e_S(i, i+1, j-1, j)$: gives the free energy of a stacked pair that consists of $i.j$ and $(i+1).(j-1)$.
- $e_{int}(i, i', j', j)$: gives the free energy of an internal loop or bulge with exterior pair $i.j$ and interior pair $i'.j'$.

The free energy of a multiloop with k branches and u unpaired bases is $a + bk + cu$, where a, b, c are constants.

The free energy of a sequence S with respect to a fixed secondary structure R is the sum of the free energies of the loops of R. Sometimes when the strand S is fixed, it is convenient to refer simply to the free energy of the structure R.

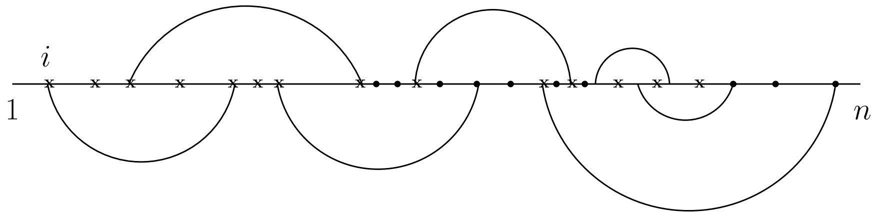


FIG. 6. A density-2 secondary structure for a sequence of length n. Recall that for $i < j \leq n$, R_{ij} is the structure restricted to the region $[i, j]$, and G_{ij} is that part of R_{ij} above the horizontal line. The black circle dots show positions of j where R_{ij} is a prefix of a density-2 pseudoloop with respect to G, and the crosses show positions of j where R_{ij} is not a prefix of a density-2 pseudoloop with respect to G.

TABLE 1. ENERGY PARAMETERS

Name	*Description*
P_s	Exterior pseudoloop initiation penalty
P_{sm}	Penalty for introducing pseudoknot inside a multiloop
P_{sp}	Penalty for introducing pseudoknot inside a pseudoloop
P_b	Band penalty
P_{up}	Penalty for unpaired base in a pseudoloop
P_{ps}	Penalty for closed subregion inside a pseudoloop
$e_H(i, j)$	Energy of a hairpin loop closed by $i.j$
$e_S(i, i+1, j-1, j)$	Energy of stacked pair closed by $i.j$
$e_{stP}(i, i+1, j-1, j)$	Energy of stacked pair that spans a band
$e_{int}(i, r, r', j)$	Energy of a pseudoknot-free internal loop
$e_{intP}(i, r, r', j)$	Energy of internal loop that spans a band
a	Multiloop initiation penalty
b	Multiloop base pair penalty
c	Penalty for unpaired base in a multiloop
a'	Penalty for introducing a multiloop that spans a band
b'	Base pair penalty for a multiloop that spans a band
c'	Penalty for unpaired base in a multiloop that spans a band

2.4.2. Pseudoknotted energy model. The pseudoknotted energy model is as follows:

- $BE_{i,i'}$: The total energy of band $[i, i'] \cup [bp(i'), bp(i)]$ is the sum of the energies of its loops. If a band has no loops, i.e., consists of just one base pair, we define its energy to be 0.
- $e_{stP}(i, i+1, j-1, j)$: defines the energy of stacked pairs in a band.
- $e_{intP}(i, r, r', j)$: defines the energy of internal loop that spans a band.

We define energy of multiloops that span a band to be the same as pseudoknot-free multiloops.

The energy of an exterior pseudoloop is the energy of bands plus $P_b \times m + P_{ps} \times k + P_{up} \times u + P_s$, where m is the number of the bands, k is the number of closed subregions, and u is the number of unpaired bases. If the pseudoknot is inside a multiloop or a pseudoloop, P_s is replaced by P_{sm} or P_{sp}, respectively.

Let $R_{i,j}$ be a prefix of a pseudoloop. The energy of $R_{i,j}$ is the sum of the energies of all loops within $R_{i,j}$ plus a penalty for each band and each unpaired base in $[i, j]$ associated with the pseudoloop of which $R_{i,j}$ is a prefix.

Table 1 summarizes the energy constants and functions used in our energy model for pseudoknotted structure.

3. PROPERTIES OF DENSITY-2 STRUCTURES

As will become clearer later, the reason that the HFold algorithm works for density-2 structures is because of the following lemmas, which are key for efficient decomposition of energies in the recurrences. In this section we first present six lemmas which are used in showing correctness of our algorithm. The lemmas are admittedly technical. The reader may want to skip this section until the recurrences in Section 4 are fully absorbed. The lemmas identify the borders of the band for a given region for each case in the recurrences.

Lemma 1. *Let G and G' be disjoint, pseudoknot-free, secondary structures, such that $G \cup G'$ is a density-2 secondary structure, and let i, j be the start and end of a pseudoloop of $G \cup G'$. Let $l \in [i, j]$ be such that*

1. l is paired in G' (but not in G), and
2. $bp_{G'}(l) < bp_G(j) < l < j$.

Let

$$b_{(i,l)} = \min\{k | i \le k < l < bp_G(k)\}, \textit{ and}$$

$$b'_{(i,l)} = \max\{k | i \le k < l < bp_G(k)\}.$$

Then, the structure $G \cup G'$ contains a band with outer base pair $b_{(i,l)}.bp_G(b_{(i,l)})$ and inner base pair $b'_{(i,l)}.bp_G(b'_{(i,l)})$.

The bottom left part of Figure 2 illustrates Lemma 1, showing the borders of the band whose arcs cross the base pair involving base $l = 14$. If $[i, j]$ is the region $[2, 30]$, then $b_{(2,14)} = 2$ and $b'_{(2,14)} = 6$.

Proof. Since i is the start of a pseudoloop and j is the end of the pseudoloop, $[i, j]$ must be a pseudoknotted closed region of $G \cup G'$. Restriction (1) implies that $bp_{G'}(l) \in [i, j]$, since if it is not, then $[i, j]$ is not a closed region of $G \cup G'$.

Based on restriction (2) and the definition of crossing base pairs, we have $bp_{G'}(l).l$ crosses $bp_G(j).j$.

Let $b_1.bp_G(b_1)$, and $b_2.bp_G(b_2)$ be the outer and the inner base pairs of the band containing $bp_G(j).j$, respectively. We have $i \le b_1 \le b_2 < l < bp_G(b_2) \le bp_G(b_1) \le j$.

Now we prove that $b_{(i,l)} = b_1$. Since $i \le b_1 < l < bp_G(b_1)$ it must be that $b_{(i,l)} \le b_1$, by the definition of $b_{(i,l)}$. If $b_{(i,l)} < b_1$, then we have $b_{(i,l)} < b_1 < bp_G(b_1) < bp_G(b_{(i,l)})$, since G is pseudoknot-free. We show that $bp_{G'}(l).l$ crosses $b_{(i,l)}.bp_G(b_{(i,l)})$. If $bp_{G'}(l).l$ does not cross $b_{(i,l)}.bp_G(b_{(i,l)})$, then it must be that $i \le b_{(i,l)} < bp_{G'}(l) < l < bp_G(b_{(i,l)})$. Based on restriction (2) and that G is pseudoknot-free we must have $i \le b_{(i,l)} < bp_{G'}(l) < bp_G(j) < l < j < bp_G(b_{(i,l)})$. But this contradicts the fact that j is the right border of the closed region $[i, j]$. Therefore, our assumption is incorrect and $bp_{G'}(l).l$ crosses $b_{(i,l)}.bp_G(b_{(i,l)})$.

Since $G \cup G'$ is a density-2 structure, there can be no other band, except the band containing $bp_G(j).j$, that crosses $bp_{G'}(l).l$ from the right. If there is another band, say B', different from the band containing $bp_G(j).j$ crossing $bp_{G'}(l).l$ from the right then a vertical line drawn at position l crosses 3 bands, and thus, $G \cup G'$ has density at least 3, which is a contradiction. Thus our assumption of $b_{(i,l)} \neq b_1$ does not hold, and $b_{(i,l)} = b_1$. Similarly we can show that $b'_{(i,l)} = b_2$. ■

Lemma 2. *Let G and G' be disjoint, pseudoknot-free, secondary structures, such that $G \cup G'$ is a density-2 secondary structure. Let l be paired in G' (but not in G) and let $[i, l]$ be a region such that $l < bp_{G'}(l)$. Let*

$$b_{(i,l)} = \min\{k | i \le k < l < bp_G(k) < bp_{G'}(l)\} \cup \{\infty\}, \textit{ and}$$

$$b'_{(i,l)} = \max\{k | i \le k < l < bp_G(k) < bp_{G'}(l)\} \cup \{-1\}.$$

Then, either both or neither of $b_{(i,l)}$ and $b'_{(i,l)}$ have finite positive values. In the former case $b_{(i,l)}.bp_G(b_{(i,l)})$ and $b'_{(i,l)}.bp_G(b'_{(i,l)})$ are the outer and the inner base pairs of a band that crosses $l.bp_{G'}(l)$ from the left in structure $G \cup G'$, respectively.

Figure 7 illustrates the notation used in Lemma 2.

For example, for region $[1, 18]$ in Figure 3, we have $b_{(1,18)} = 1$, $b'_{(1,18)} = 2$.

Proof. If there is no base pair, $k.bp_G(k)$ in G such that we have $i \le k < l < bp_G(k) < bp_{G'}(l)$, then there is no base pair crossing $l.bp_{G'}(l)$ from the left, and thus, there is no band in G that crosses $l.bp_G(l)$. Therefore, $b_{(i,l)} = \infty$ and $b'_{(i,l)} = -1$. It is easy to show that we cannot have the case in which $b_{(i,l)} \neq \infty$ but $b'_{(i,l)} = -1$ (or similarly $b_{(i,l)} = \infty$ but $b'_{(i,l)} \neq -1$), since if we did, we must have at least one base pair, $k.bp_G(k)$ in G such that $i \le k < l < bp_G(k) < bp_{G'}(l)$ and thus, both $b_{(i,l)}$ and $b'_{(i,l)}$ must have positive values, which is a contradiction. Thus, either both or neither of $b_{(i,l)}$ and $b'_{(i,l)}$ have finite positive values.

Otherwise, let $b_1.bp_G(b_1)$ and $b_2.bp_G(b_2)$ be the outer and the inner base pair of the band that crosses $l.bp_{G'}(l)$ from the left, respectively. We have $i \le b_1 \le b_2 < l < bp_G(b_2) \le bp_G(b_1)$. By definition

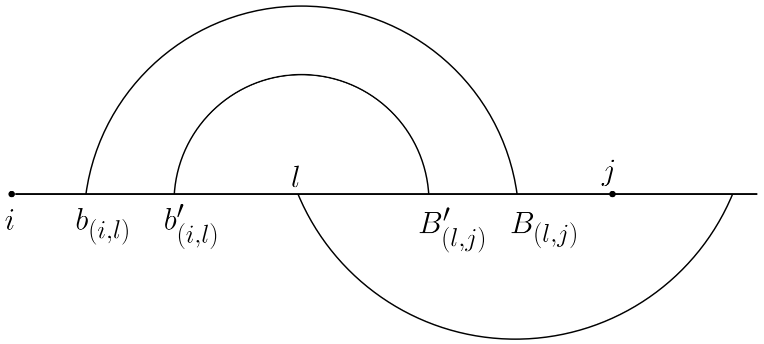


FIG. 7. Illustration of band borders in Lemmas 2 and 4.

of $b_{(i,l)}$, we must have $b_{(i,l)} \leq b_1$. We claim $b_{(i,l)} = b_1$. If $b_{(i,l)} < b_1$, then since G is pseudoknot-free, we must have $b_{(i,l)} < b_1 < bp_G(b_1) < bp_G(b_{(i,l)})$. By definition, $b_{(i,l)}$ must cross $l.bp_{G'}(l)$. Since $G \cup G'$ is a density-2 structure, we cannot have more than one band crossing $l.bp_{G'}(l)$ from the left. Therefore $b_{(i,l)}.bp_G(b_{i,l})$ is in the same band as $b_1.bp_G(b_1)$, so $b_{(i,l)}.bp_G(b_{(i,l)})$ must be the outer base pair of the band, which is a contradiction. Thus, we must have $b_{(i,l)} = b_1$. Similarly, we can prove that $b'_{(i,l)} = b_2$. ■

Lemma 3. *Let G and G' be disjoint, pseudoknot-free, secondary structures, such that $G \cup G'$ is a density-2 secondary structure. Let i be paired in G' (but not in G) and let $i < bp_{G'}(i)$. Let*

$$b_{(i,bp_{G'}(i))} = \min\{k | i \leq k < bp_{G'}(i) < bp_G(k)\} \cup \{\infty\}, \textit{ and}$$

$$b'_{(i,bp_{G'}(i))} = \max\{k | i \leq k < bp_{G'}(i) < bp_G(k)\} \cup \{-1\}.$$

Then, either both or neither of $b_{(i,bp_{G'}(i))}$ and $b'_{(i,bp_{G'}(i))}$ have finite positive values. In the former case, $b_{(i,bp_{G'}(i))}.bp_G(b_{(i,bp_{G'}(i))})$ and $b'_{(i,bp_{G'}(i))}.bp_G(b'_{(i,bp_{G'}(i))})$ are the outer and the inner base pairs, respectively of a band that crosses $i.bp_{G'}(i)$ from the right in structure $G \cup G'$.

Proof. If there is no base pair, $k.bp_G(k)$ in G such that we have $i \leq k < bp_{G'}(i) < bp_G(k)$, then there is no base pair crossing $i.bp_{G'}(i)$, and thus, there is no band in G that crosses $i.bp_{G'}(i)$ from the right. Therefore, $b_{(i,bp_{G'}(i))} = \infty$ and $b'_{(i,bp_{G'}(i))} = -1$. As in the proof of Lemma 1 we can argue that either both or neither of $b_{(i,bp_{G'}(i))}$ and $b'_{(i,bp_{G'}(i))}$ have finite positive values.

Otherwise, let $b_1.bp_G(b_1)$ and $b_2.bp_G(b_2)$ be the outer and the inner base pairs of the band that crosses $i.bp_{G'}(i)$ from the right, respectively. We have $i \leq b_1 \leq b_2 < bp_{G'}(i) < bp_G(b_2) \leq bp_G(b_1)$. By definition of $b_{(i,bp_{G'}(i))}$, we must have $b_{(i,bp_{G'}(i))} \leq b_1$. We claim $b_{(i,bp_{G'}(i))} = b_1$. If $b_{(i,bp_{G'}(i))} < b_1$, then since G is pseudoknot-free, we must have $b_{(i,bp_{G'}(i))} < b_1 < bp_G(b_1) < bp_G(b_{(i,bp_{G'}(i))})$. By definition, $b_{(i,bp_{G'}(i))}$ crosses $i.bp_{G'}(i)$. Since $G \cup G'$ is a density-2 structure, we cannot have more than one band crossing $i.bp_{G'}(i)$ from the right. Therefore $b_{(i,bp_{G'}(i))}.bp_G(b_{i,l})$ is in the same band as $b_1.bp_G(b_1)$, so $b_{(i,bp_{G'}(i))}.bp_G(b_{(i,bp_{G'}(i))})$ must be the outer base pair of the band, which is a contradiction. Thus, we must have $b_{(i,bp_{G'}(i))} = b_1$. Similarly we can prove that $b'_{(i,bp_{G'}(i))} = b_2$. ■

Lemma 4. *Let G and G' be disjoint, pseudoknot-free, secondary structures, such that $G \cup G'$ is a density-2 secondary structure. Let j be paired in G' (but not in G) and let $bp_{G'}(j) < j$. Let*

$$B_{(bp_{G'}(j),j)} = \max\{bp_G(k) | k < bp_{G'}(j) < bp_G(k) \leq j\} \cup \{-1\}, \textit{ and}$$

$$B'_{(bp_{G'}(j),j)} = \min\{bp_G(k) | k < bp_{G'}(j) < bp_G(k) \leq j\} \cup \{\infty\}.$$

Then, either both or neither of $B_{(bp_{G'}(j),j)}$ and $B'_{(bp_{G'}(j),j)}$ have finite positive values. In the former case, $bp_G(B_{(bp_{G'}(j),j)}).B_{(bp_{G'}(j),j)}$ and $bp_G(B'_{(bp_{G'}(j),j)}).B'_{(bp_{G'}(j),j)}$ are the outer and the inner base pairs of a band that crosses $bp_{G'}(j).j$ from the left in structure $G \cup G'$, respectively.

Figure 7 illustrates the notation used in Lemma 4.

For example, for region [18, 34] in Figure 3, we have $B'_{(18,34)} = 22$, $B_{(18,34)} = 23$.

Proof. The proof is very similar to the proof of Lemma 3. ■

Lemma 5. *Let G and G' be disjoint, pseudoknot-free, secondary structures, such that $G \cup G'$ is a density-2 secondary structure. Let i be paired in G' (but not in G) and let $i < j$ such that region $[bp_{G'}(i) + 1, j]$ is weakly closed. Let*

$$b_{(i,j)} = \min\{k|i \leq k < j < bp_G(k)\} \cup \{\infty\},$$

$$b'_{(i,j)} = \max\{k|i \leq k < j < bp_G(k)\} \cup \{-1\},$$

Then, either both or neither of $b_{(i,j)}$ and $b'_{(i,j)}$ have finite positive values. In the former case, $b_{(i,j)}.bp_G(b_{(i,j)})$ and $b'_{(i,j)}.bp_G(b'_{(i,j)})$ are the outer and the inner base pairs of a band in $G \cup G'$ that crosses $i.bp_{G'}(i)$ from the right.

Note that each term may be either infinity or −1 to account for the cases when there is no such band border.

For region $[i, j] = [1, 26]$ in Figure 3, we have $b_{(1,26)} = 1$, $b'_{(1,26)} = 2$.

Proof. Since region $[bp_{G'}(i) + 1, j]$ is weakly closed, there are no base pairs in the region $[bp_{G'}(i) + 1, j]$ crossing $i.bp_{G'}(i)$, thus, $b_{(i,bp_{G'}(i))} = b_{(i,j)}$ and $b'_{(i,bp_{G'}(i))} = b'_{(i,j)}$. The rest of the proof is similar to the proof of Lemma 3. ■

Lemma 6. *Let G and G' be disjoint, pseudoknot-free, secondary structures, such that $G \cup G'$ is a density-2 secondary structure. Let j be paired in G' (but not in G) and let $i < j$ be such that region $[i, bp_{G'}(j) - 1]$ is weakly closed. Let*

$$B_{(i,j)} = \max\{bp_G(k)|k < i < bp_G(k) \leq j\} \cup \{-1\}, \text{ and}$$

$$B'_{(i,j)} = \min\{bp_G(k)|k < i < bp_G(k) \leq j\} \cup \{\infty\}.$$

Then, either both or neither of $B_{(i,j)}$ and $B'_{(i,j)}$ have finite positive values. In the latter case, $bp_G(B_{(i,j)}).B_{(i,j)}$ and $bp_G(B'_{(i,j)}).B'_{(i,j)}$ are the outer and the inner base pairs of a band in $G \cup G'$ that crosses $bp_{G'}(j).j$ from the left.

Note that each term may be either infinity or −1 to account for the cases when there is no such band border.

For region $[i, j] = [17, 34]$ in Figure 3, we have $B'_{(17,34)} = 22$ and $B_{(17,34)} = 23$.

Proof. The proof is very similar to the proof of Lemma 5. ■

4. THE HFold ALGORITHM

HFold is a method for prediction of pseudoknotted RNA secondary structure that integrates MFE-based prediction with folding pathway considerations in a novel way. The method is motivated by the hypothesis that pseudoknotted RNA secondary structures form in a hierarchical fashion, with a pseudoknot-free structure forming first and additional pseudoknot-forming base pairs that are added later (possibly with minor rearrangements of the initial pseudoknot-free structure) (Mathews, 2004; Tinoco and Bustamante, 1999). HFold works by taking as input a sequence of bases, S, and a pseudoknot-free secondary structure G, and finding a second pseudoknot-free structure G' which minimizes the energy of $G \cup G'$ (i.e., HFold$(S, G) = G \cup G'$). Like MFE methods, HFold handles only a restricted class of structures, but this class is quite general (density-2 structures). The method has two potential advantages over MFE-based secondary structure predication. First, HFold's hierarchical folding principle may model biological folding just as well, or better, than does the MFE structure formation hypothesis, at least on biological structures. Second, HFold has $O(n^3)$ running time, making it significantly more efficient than MFE-based methods that require $\Omega(n^5)$ time or more to predict biologically-important pseudoknotted structures.

4.1. High level description of HFold

Before providing the detailed recurrences for HFold, we first give a high level overview of the algorithm. We start by briefly reviewing key ideas of the dynamic programming algorithm which predicts the energy of the MFE pseudoknot-free secondary structure for a fixed sequence $S = s_1 s_2 \ldots s_n$ (Mathews et al., 1999). Let $W_{i,j}$ be the energy of the MFE pseudoknot-free secondary structure for the subsequence $s_i s_{i+1} \ldots s_j$. If $i \geq j$, $W_{i,j} = 0$, since the subsequence is empty. Otherwise, it must either be that $i.j$ is a base pair in the MFE structure for $s_i \ldots s_j$, or that the MFE structure can be decomposed into two independent subparts. These two cases correspond to the two rows of the following recurrence for $W_{i,j}$.

$$W_{i,j} = \min \begin{cases} V_{i,j}, \\ \min_{i \leq r < j} W_{i,r} + W_{(r+1),j}, \end{cases}$$

where $V_{i,j}$ is the free energy of the MFE structure for $s_i \ldots s_j$ that contains $i.j$. If $i \geq j$, $V_{i,j}$ is set to be ∞. Otherwise, $i.j$ closes a hairpin, an internal loop, or a multiloop in the MFE structure for $s_i \ldots s_j$. Thus, $V_{i,j}$ can be expressed as the minimum of the free energies attainable in three cases:

$$V_{i,j} = \min \begin{cases} e_H(i,j), \\ \min_{r,r'} e_{int}(i,r,r',j) + V_{r,r'}, \\ VM_{i,j} \end{cases}$$

where $e_H(i,j)$ and $e_{int}(i,r,r',j)$ are as given in Table 1, and $VM_{i,j}$ is the energy of a MFE structure for $s_i \ldots s_j$ in which $i.j$ closes a multiloop.

We extend the definition of $W_{i,j}$ for the hierarchical folding algorithm as follows. Let G be a given pseudoknot-free structure for S. If some arc of G covers i or j, then $W_{i,j} = \infty$. If $i \geq j$, then $W_{i,j} = 0$. Otherwise we define $W_{i,j}$ to be the energy of the MFE secondary structure $G_{ij} \cup G'_{ij}$ for the strand $s_i \ldots s_j$, taken over all choices of G'_{ij} which is pseudoknot-free, disjoint from G_{ij}, and such that $G_{ij} \cup G'_{ij}$ is density-2. In this case, $W_{i,j}$ satisfies the following recurrence:

$$W_{i,j} = \min \begin{cases} V_{i,j}, \\ \min_{i \leq r < j} W_{i,r} + W_{(r+1),j}, \\ WMB_{i,j} + P_s \end{cases}$$

where the first two cases are the same as for pseudoknot-free cases and the last case is specific to pseudoknotted structures. P_s is the pseudoknot initiation penalty, given in Table 1.

The third row of this recurrence accounts for the case when the optimal secondary structure $G_{ij} \cup G'_{ij}$ includes pseudoknotted base pairs and cannot be partitioned into two independent substructures for two regions $[i,r]$ and $[r+1,j]$, for some r. Such a structure must contain a chain of two or more successively-overlapping bands, which must alternate between G_{ij} and G'_{ij}, possibly with nested substructures interspersed throughout. Figure 8 provides an example, and shows how the recurrence for WMB, given below, unwinds when the example structure is the MFE structure.

In order to calculate the energies of substructures in such a structure in the recurrences, we use three additional terms: BE, VP, and WI. Roughly, these account for energies of bands spanned by base pairs of G_{ij}, regions enclosed by pseudoknotted base pairs of G'_{ij} (excluding part of those regions that are within a band of G_{ij}), and weakly closed subregions, respectively.

We now give the recurrence for $WMB_{i,j}$. Figure 9 illustrates different cases of WMB recurrence. As the base case, we set $WMB_{i,j} = +\infty$ if $i \geq j$, since if $i \geq j$ the structure is empty, and thus cannot be

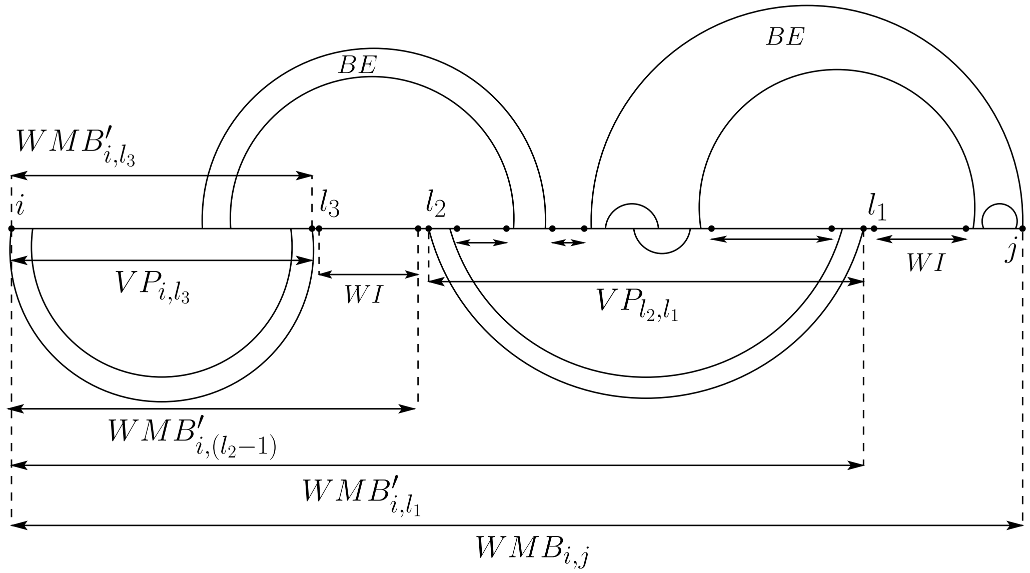


FIG. 8. Illustration of how the *WMB* recurrence unwinds, to calculate $WMB_{i,j}$. Arcs above the horizontal line from i to j represent base pairs of G_{ij}, and arcs below the line represent base pairs of G'_{ij}. Case (1) of the *WMB* recurrence handles the overall structure whose energy is $WMB_{i,j}$, with $l = l_1$, with terms to account for energies of the right upper band (BE) and right lower closed subregion ($WI_{(l_1+1),(bp_G(b'_{i,l_1})-1)}$) as well as the remaining prefix (WMB'_{i,l_1}). The term WMB'_{i,l_1} is handled by case (1) of the WMB' recurrence, with $l = l_2$ and terms to account for the lower right substructure labelled VP_{l_2,l_1}, the upper left band (BE), and the remaining "prefix" of the overall structure ($WMB'_{i,(l_2-1)}$). $WMB'_{i,(l_2-1)}$ is then handled by case (2) of the WMB' recurrence, with $l = l_3$, and terms to account for $WI_{(l_3+1),(l_2-1)}$ and WMB'_{i,l_3}. Finally, the WMB'_{i,l_3} term is handled by end case (3) of the WMB' recurrence.

pseudoknotted. Otherwise, there are two cases, depending on whether j is paired in G or not. In case (1), j is paired in G. Then, in the MFE structure, some base l with $bp(j) < l < j$ must be paired in G', causing $bp(j).j$ to be pseudoknotted. We minimize the energy over all possible choices of l (note that l must be unpaired in G, since it will be paired in G', which is disjoint from G).

By Lemma 1, once l is fixed, the inner base pair of the band whose outer base pair is $bp(j).j$ is also determined. The $P_b + BE$ term in case (1) of the recurrence accounts for the energy of the band, a *WI* term accounts for a weakly closed region that is in the band, and the remaining energy is represented by the WMB' term. In case (2), j is not paired in G, and the recurrence is unwound by moving directly to a WMB' term. Thus, we have:

$$WMB_{i,j} = \begin{cases} (1)\ P_b + \min\limits_{\substack{bp_G(j)<l<j \\ bp_G(l)=0}} (BE_{b_{(i,l)},b'_{(i,l)}} + WMB'_{i,l} + WI_{(l+1),(bp_G(b'_{(i,l)})-1)}), & \text{if } 0 < bp(j) < j \\ \\ (2)\ WMB'_{i,j} & \end{cases}$$

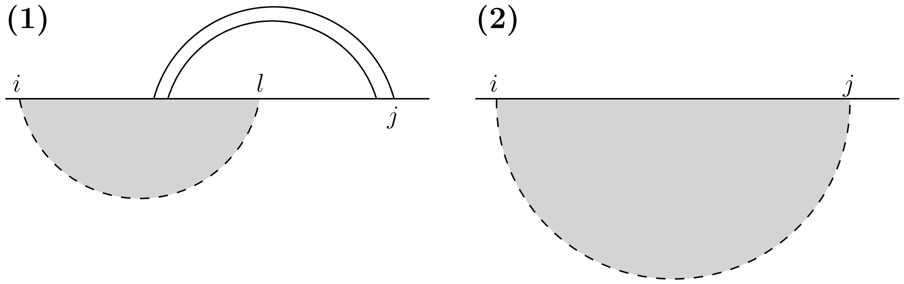


FIG. 9. Illustration of cases for $WMB_{i,j}$.

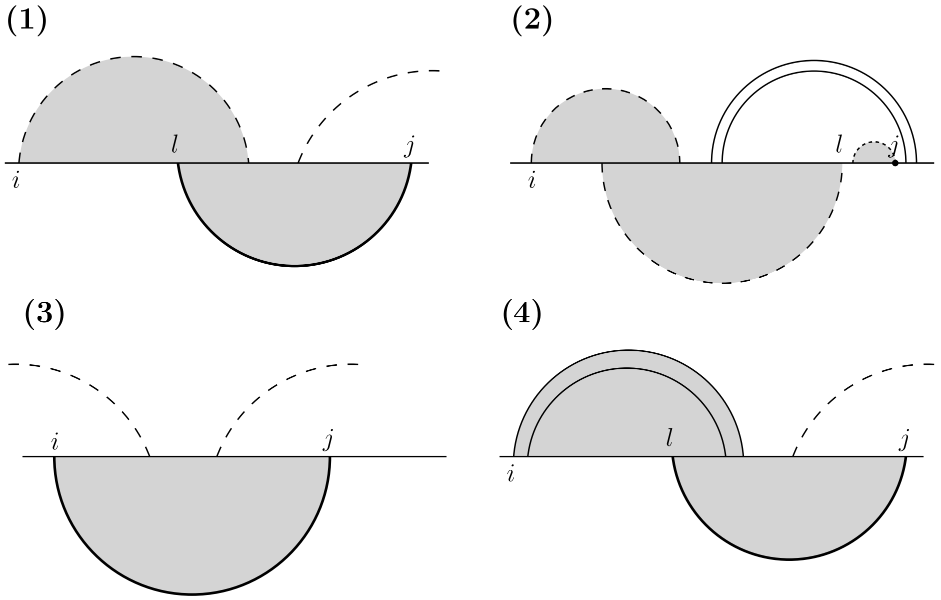


FIG. 10. Illustration of cases for $WMB'_{i,j}$.

Complementing case (1) of the WMB recurrence, WMB' handles the case that the rightmost band is not in G, but is part of the structure G'. Figure 10 illustrates different cases of WMB' recurrence. In the recurrence for WMB', case (1) is the complex case, accounting for the energy of the region spanned by the rightmost two bands using the $2P_b$, VP, and BE terms, and recursively calling WMB'. The band borders in the WMB' cases are determined using Lemmas 2 and 4. Case (2) is called when one iteration of $WMB_{i,j}$ or $WMB'_{i,j}$ case (1) is done and the rightmost substructure of the overall "prefix" up to position j is a weakly closed region. Note that $WI_{i,j} = +\infty$ when $cover(i) \neq cover(j)$, ensures that case (2) is not entered as the first iteration of WMB'. Cases (3) and (4) are end cases, where only one or two bands need to be accounted for, respectively and so no recursive call to WMB' is made. Thus we have:

$$WMB'_{i,j} = \min \begin{cases} (1)\ \ 2P_b + \min\limits_{\substack{i<l<b_{(i,j)} \\ isCovered(G_{ij},l)}} (BE_{b_{(i,l)},b'_{(i,l)}} + WMB'_{i,(l-1)} + VP_{l,j}), & \text{if } bp_G(j) = 0 \\ (2)\ \ \min\limits_{\substack{i<l<j \\ cover(l)=cover(j)}} (WMB'_{i,l} + WI_{(l+1),j}), & \text{if } bp_G(j) < j \\ (3)\ \ P_b + VP_{i,j}, & \\ (4)\ \ 2P_b + \min\limits_{i<l<bp_G(i)} (BE_{b_{(i,l)},b'_{(i,l)}} + WI_{(b'_{(i,l)}+1),(l-1)} + VP_{l,j}), & \text{if } 0 = bp_G(j) < bp_G(i) \end{cases}$$

Figure 11 shows how all the recurrences call each other.

4.2. Recurrences

In this section, we present the details of the HFold recurrences. Throughout this work, we will use the following notation:

- G: a pseudoknot-free structure.
- G': a pseudoknot-free structure that HFold adds to G.
- R: the complete pseudoknotted structure: $R = G \cup G'$.

Let $R_{i,j}$ be a minimum free energy (MFE) secondary structure for $[i, j]$, given a pseudoknot-free secondary structure $G_{i,j}$.

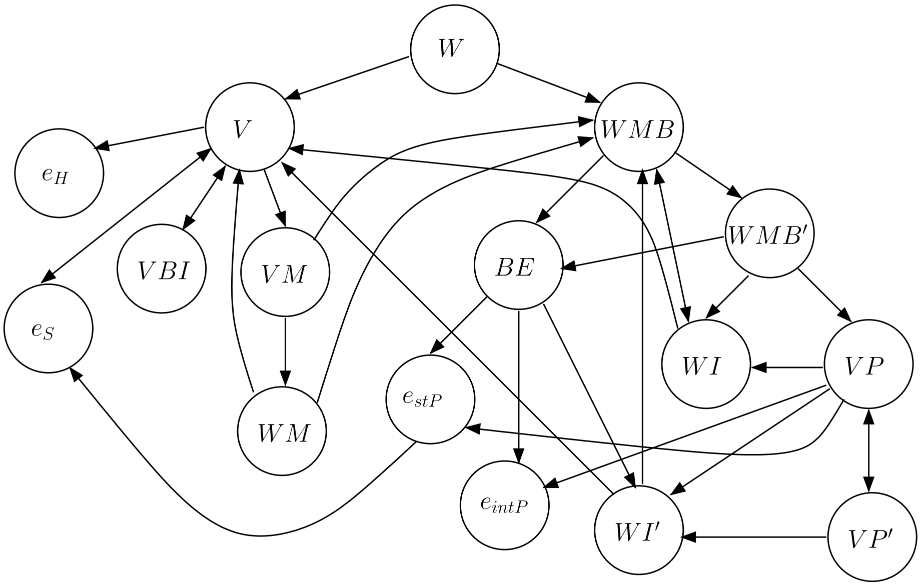


FIG. 11. Visual illustration of recurrences in HFold.

The energy value of each substructure type, for a given input sequence $S = s_1s_2, \ldots, s_n$ and the given pseudoknot-free secondary structure G, is stored in an array. In the next subsections, we describe how each is calculated. We illustrate each case with a figure, where we use the following notations in our figures:

- The normal black lines can be any arcs in $R_{i,j}$.
- The solid lines are for base pairs.
- The dotted lines connect bases that don't have to be paired.
- The clear shade within the arcs indicate that there are no additional base pairs within the arc.
- The shade within the arcs are unknown structures.

4.2.1. $W_{i,j}$. $W_{i,j}$ is the MFE of all structures $R_{i,j}$ over region $[i, j]$, if i and j are not covered in G, i.e. $\overline{isCovered(G, i)}$ and $\overline{isCovered(G, j)}$. Otherwise, $W_{i,j}$ is $+\infty$. Figure 12 illustrates different cases of W recurrence.

The base cases are as follows: $W_{i,j} = 0$, if $i \geq j$, since then the only possibility for structure $R_{i,j}$ is the empty structure; and $W_{i,j} = +\infty$, if *isCovered*(G, i), or *isCovered*(G, j).

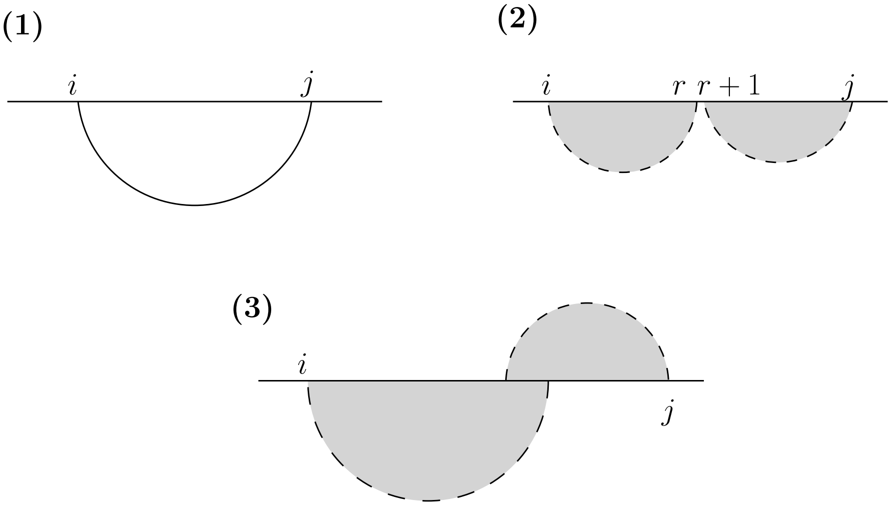


FIG. 12. Illustration of cases for $W_{i,j}$.

Otherwise, $W_{i,j}$ is given by the following recurrence:

$$W_{i,j} = \min \begin{cases} (1)\ \ V_{i,j} \\ (2)\ \ \displaystyle\min_{\substack{i \le r < j \\ \overline{isCovered(r)}}} (W_{i,r} + W_{(r+1),j}) \\ (3)\ \ WMB_{i,j} + P_s \end{cases} \tag{2}$$

The base cases indicate that W is being used only when the structure is an exterior structure and that there is no penalty for having unpaired bases at either end of the structure.

Case (1) handles the case that i pairs with j, i.e., $bp_{R_{i,j}}(i) = j$.

Case (2) handles the cases that $\exists r$, $i \le r < j$, $bp_{R_{i,j}}(i) \le r$ (i.e., i is either unpaired or paired with another base inside region $[i, r]$), and $bp_{R_{i,j}}(j) > r$ or $bp_{R_{i,j}}(j) = 0$ (i.e., j is either unpaired or paired with a base inside region $[r+1, j]$).

If $R_{i,j}$ does not fall into case (1) or (2), it must be that $[i, j]$ is a pseudoknotted closed region. This is an exterior pseudoknot because of the premise that i and j are not covered in G. In this case, we add a P_s penalty for introducing an exterior pseudoknot.

4.2.2. $WI_{i,j}$. $WI_{i,j}$ is the minimum free energy of all structures $R_{i,j}$, given that $[i, j]$ is weakly closed, and $R_{i,j}$ is inside a pseudoloop. Otherwise, $WI_{i,j}$ is $+\infty$. Figure 13 illustrates different cases of WI recurrence.

The base cases are as follows:

$WI_{i,j} = +\infty$, if $cover(i) \ne cover(j)$, since $[i, j]$ is not weakly closed.
$WI_{i,j} = P_{up}$, if $i = j$ and $bp_G(i) = 0$, since $[i, j]$ is an empty region, thus we give it the value for an unpaired base in a pseudoloop.
$WI_{i,j} = 0$, if $i > j$.

Otherwise, $WI_{i,j}$ is given by the following recurrence:

$$WI_{i,j} = \min \begin{cases} (1)\ \ V_{i,j} + P_{ps} & \text{if } i.j \in G, \text{ or } (bp_G(i) = 0 \text{ and } bp_G(j) = 0) \\ (2)\ \ \displaystyle\min_{i \le t < j} (WI_{i,t} + WI_{(t+1),j}) & \\ (3)\ \ WMB_{i,j} + P_{sp} + P_{ps} & \end{cases} \tag{3}$$

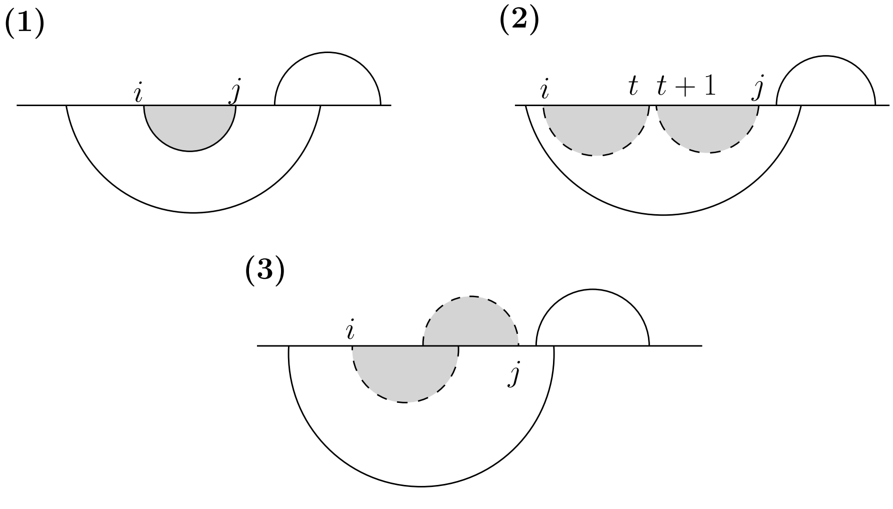


FIG. 13. Illustration of cases for $WI_{i,j}$. In case (3), the plotted structure from i to j could contain more than 2 bands (not illustrated).

Similar to $W_{i,j}$, case (1) handles the case that i pairs with j, i.e., $bp_{G_{i,j}}(i) = j$ or the case that neither is paired in G but they are paired in G'.

Case (2) handles the case that $\exists t$, $i \le t < j$, $bp_{R_{i,j}}(i) \le t$, and $bp_{R_{i,j}}(j) > t$ or $bp(R_{ij}, j) = 0$.

If $R_{i,j}$ does not fall into case (1) or (2), it must be that $paired(R_{i,j}, i)$ and $paired(R_{i,j}, j)$, and $bp_{R_{i,j}}(i) > bp_{R_{i,j}}(j)$, where $[i, j]$ is a pseudoknotted closed region in $R_{i,j}$. In this case, $R_{i,j}$ will be covered by case (3).

Since $WI_{i,j}$ is a structure inside a pseudoknot, but not inside a band, we add a P_{ps} penalty to case (1) and (3), and a P_{sp} penalty to case (3) for introducing a new pseudoknot inside a pseudoloop.

4.2.3. $WI'_{i,j}$. $WI'_{i,j}$ is the minimum free energy of all nonempty structures $R_{i,j}$, if $[i, j]$ is weakly closed with respect to G, given that $R_{i,j}$ is inside a band. Otherwise, $WI'_{i,j}$ is $+\infty$. Figure 14 illustrates different cases of WI' recurrence.

The base cases are as follows:

$WI'_{i,j} = +\infty$, if $[i, j]$ is not weakly closed with respect to G.
$WI'_{i,j} = +\infty$, if $i \ge j$, since empty$(R_{i,j}, [i, j])$.

Otherwise, $WI'_{i,j}$ is given by the following recurrence:

$$WI'_{i,j} = \min \begin{cases} (1)\ V_{i,j} + b' & \text{if } i.j \in G, \text{ or } (bp_G(i) = 0 \text{ and } bp_G(j) = 0) \\ (2)\ WI'_{(i+1),j} + c' & \text{if } bp_G(i) = 0 \\ (3)\ WI'_{i,(j-1)} + c' & \text{if } bp_G(j) = 0 \\ (4)\ \min\limits_{i \le t < j} (WI'_{i,t} + WI'_{(t+1),j}) & \\ (5)\ WMB_{i,j} + P_{sm} + b' & \end{cases} \tag{4}$$

Cases (2) and (3) handle free bases on each side of the sequence. The rest of the cases are similar to WI, with the only difference being that here in cases (1) and (5) we use b' as the penalty for introducing

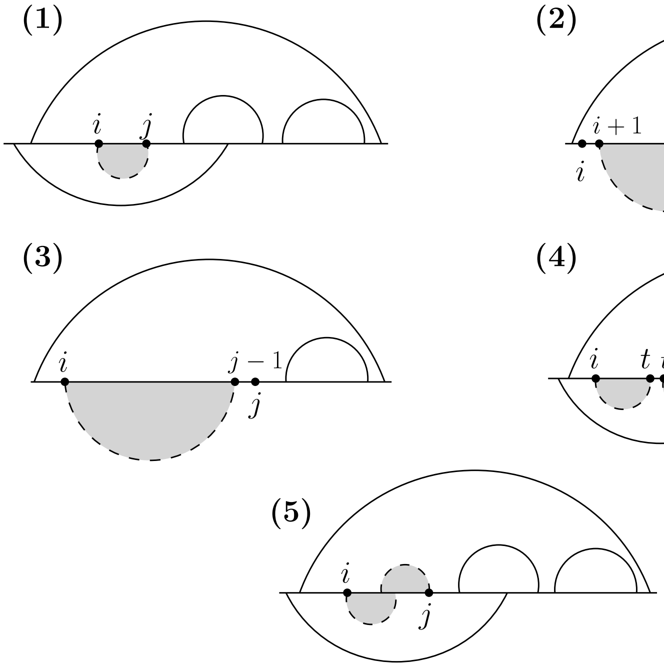


FIG. 14. Illustration of cases for $WI'_{i,j}$. In case (5), the plotted structure from i to j could contain more than 2 bands (not illustrated).

a new base pair in the structure instead of P_{ps}, since the base pair is not inside a pseudoloop, but rather in a multiloop that spans a band; for the same reason we used P_{sm} penalty instead of P_{sp} in case (5) for introducing a new pseudoknot inside a multiloop.

4.2.4. $VP_{i,j}$. $VP_{i,j}$ is the minimum free energy of all structures $R_{i,j}$, in which $bp_G(i) = bp_G(j) = 0$, bases i and j are paired in G', i.e., $bp_{G'}(i) = j$, and $i.j$ crosses a base pair of G. Here the energy of $R_{i,j}$ is the energy of all loops within $R_{i,j}$ that are not inside a band whose base pairs are in G and which crosses $i.j$. Otherwise, $VP_{i,j}$ is $+\infty$. Figure 15 illustrates different cases of VP recurrence.

The base case is as follows:

$$VP_{i,j} = +\infty \text{ , if } \begin{cases} i \geq j, \\ i.j \text{ does not cross any base pair of } G, \\ bp_G(i) > 0, \text{ or } bp_G(j) > 0 \end{cases} \tag{5}$$

Otherwise, $VP_{i,j}$ is given by the following recurrences:

$$VP_{i,j} = \min$$

$$\begin{cases} (1)\ WI_{(i+1),(B'_{(i,j)}-1)} + WI_{(B_{(i,j)}+1),(j-1)} & \text{if } \mathit{isCovered}(G,i), \text{ and } \overline{\mathit{isCovered}(G,j)} \\ (2)\ WI_{(i+1),(b_{(i,j)}-1)} + WI_{(b'_{(i,j)}+1),(j-1)} & \text{if } \overline{\mathit{isCovered}(G,i)}, \text{ and } \mathit{isCovered}(G,j) \\ (3)\ WI_{(i+1),(B'_{(i,j)}-1)} + WI_{(B_{(i,j)}+1),(b_{(i,j)}-1)} & \text{if } \mathit{isCovered}(G,i), \text{ and } \mathit{isCovered}(G,j) \\ \qquad + WI_{(b'_{(i,j)}+1),(j-1)} & \\ (4)\ e_{stP}(i, i+1, j-1, j) + VP_{(i+1)(j-1)} & \text{if } (bp_G(i+1) = 0, \text{ and } bp_G(j-1) = 0) \\ (5)\ \min\limits_{\substack{i<r<\min(B'_{(i,j)},b_{(i,j)}) \\ \max(b'_{(i,j)},B_{(i,j)})<r'<j}} (e_{intP}(i,r,r',j) + VP_{r,r'}) & \text{if } \mathit{cover}(G,i) = \mathit{cover}(G,r) \\ & \text{and } \mathit{cover}(G,j) = \mathit{cover}(G,r') \\ & \text{and empty}(G,[i+1,r-1]) \\ & \text{and empty}(G,[r'+1,j-1]) \\ (6)\ \min\limits_{\substack{i<r<\min(B'_{(i,j)},j) \\ bp_G(r)=0}} (WI'_{(i+1),(r-1)} + VP'_{r,(j-1)} + a' + 2b') & \\ (7)\ \min\limits_{\substack{\max(i,b'_{(i,j)})<r<j \\ bp_G(r)=0}} (VP'_{(i+1),r} + WI'_{(r+1),(j-1)} + a' + 2b') & \end{cases} \tag{6}$$

Cases (1), (2), and (3) handle the cases that there are no other base pairs in $[i, j]$ that cross the same band(s) that $i.j$ crosses. In these cases, we compute the energy between band borders. In these cases, the band borders $B_{(i,j)}$ and $B'_{(i,j)}$ are determined by Lemma 4 and $b_{(i,j)}$ and $b'_{(i,j)}$ are determined by Lemma 3.

Case (4) handles the case that base pairs $i.j$ and $(i+1).(j-1)$ form a stacked pair in $R_{i,j}$.

Case (5) handles the case that $i.j$ and $r.r'$ close an internal loop of $R_{i,j}$. In this case the band borders $B_{(i,j)}$ and $B'_{(i,j)}$ are determined by Lemma 4 and $b_{(i,j)}$ and $b'_{(i,j)}$ are determined by Lemma 3.

Cases (6) and (7) handle the similar condition to case (5) except that case (6) allows closed regions in the gap region $[i, r-1]$ and case (7) allows closed regions in the gap region $[r+1, j]$. In those cases $i.j$ does not close an internal loop, but rather closes a multiloop that spans a band. In these cases, the band border $B'_{(i,j)}$ is determined by Lemma 4 and $b'_{(i,j)}$ is determined by Lemma 3.

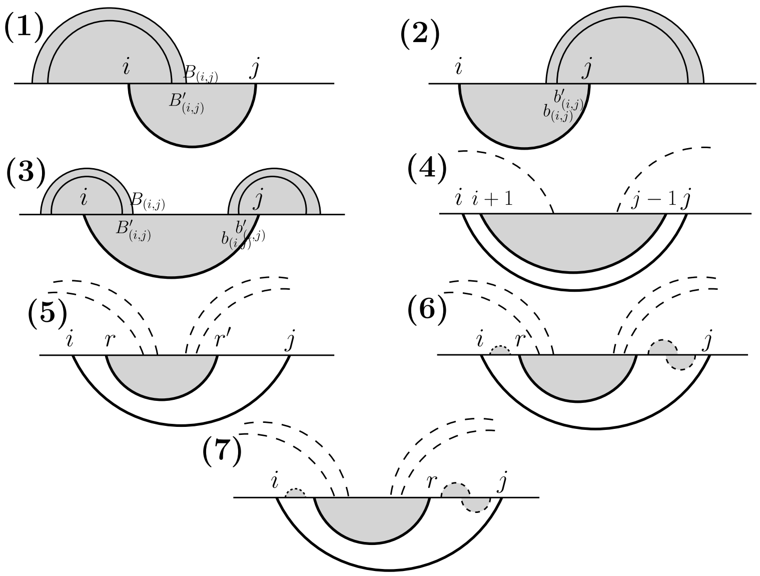


FIG. 15. Illustration of cases for $VP_{i,j}$.

Cases (6) and (7) can be combined into the following case:

$$\min_{\substack{i<r<\min(B'_{(i,j)},b_{(i,j)}) \\ \leq\max(b'_{(i,j)},B_{(i,j)})<r'<j}} (WI'_{(i+1),(r-1)} + VP_{r,r'} + WI'_{(r'+1),(j-1)} + P_{ps} + a' + 2b') \qquad (7)$$

Since the minimization is done over two parameters r and r', we should limit the size of region $[r, r']$ to keep the complexity of our algorithm to $O(n^3)$.

If $R_{i,j}$ does not fall into any case from (1) to (7), then there must exist $r.r'$ in G', with $i < r < r' < j$, and one base r (or r') inside the band region $[B'_{(i,j)}, B_{(i,j)}]$ (or $[b_{(i',j)}, b'_{(i,j)}]$) and the other base r' (or r) outside the band region $[B'_{(i,j)}, B_{(i,j)}]$ (or $[b_{(i,j)}, b'_{(i,j)}]$). Then $G \cup G'$ must have density at least 3, which is not allowed in our algorithm.

4.2.5. $VP'_{i,j}$. $VP'_{i,j}$ is the minimum free energy of all structures $R_{i,j}$ over region $[i, j]$, such that for some r, $i < r < j$, either $bp_{G'}(i) = r$ or $bp_{G'}(j) = r$, and either $i.r$ or $r.j$ crosses a base pair of G. Otherwise, $VP'_{i,j}$ is $+\infty$. Figure 16 illustrates different cases of VP' recurrence.

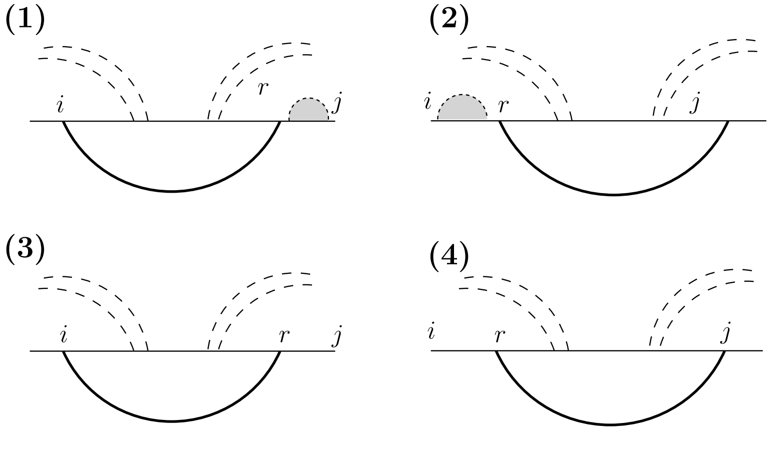


FIG. 16. Illustration of cases for $VP'_{i,j}$.

The base case is as follows:

$$VP'_{i,j} = +\infty, \text{ if } i \geq j \tag{8}$$

Otherwise, $VP'_{i,j}$ is given by the following recurrences:

$$VP'_{i,j} = \min \begin{cases} (1) & \min\limits_{\max(i, b'_{(i,j)}) < r < j} (VP_{i,r} + WI'_{(r+1),j}) \\ (2) & \min\limits_{i < r < \min(B'_{(i,j)}, j)} (WI'_{i,(r-1)} + VP_{r,j}) \\ (3) & \min\limits_{\max(i, b'_{(i,j)}) < r < j} (VP_{i,r} + c'(j-r)) \quad \text{if empty}(G, [r+1, j]) \\ (4) & \min\limits_{i < r < \min(B'_{(i,j)}, j)} (c'(r-i) + VP_{r,j}) \quad \text{if empty}(G, [i, r-1]) \end{cases} \tag{9}$$

In both cases (1) and (2), the energy of $R_{i,j}$ is the energy of all loops within $R_{i,j}$. In case (1), we have two components: the energy given by base pair $i.r$ which is covered by VP, and the energy given by the structure from base $r+1$ to base j, which is covered by WI'. Since only $VP_{i,j}$ uses $VP'_{i,j}$, the structure from $r+1$ to j is within a band (see case (6) of $VP_{i,j}$) and so is covered by WI'. Case (2) can be reasoned similarly, with reference to case (7) of $VP_{i,j}$ for use of WI'. Cases (3) and (4) are similar to cases (1) and (2) with the only difference that there is no base pairs in regions $[r+1, j]$ and $[i, r-1]$, respectively.

The band borders for cases (1) and (3) are determined by Lemma 5 and the band borders for cases (2) and (4) are determined by Lemma 6.

4.2.6. $V_{i,j}$. $V_{i,j}$ is the minimum free energy of all structures $R_{i,j}$ over region $[i, j]$, if $[i, j]$ is weakly closed or empty and $i.j$ forms a base pair of $R_{i,j}$. Otherwise, $V_{i,j}$ is $+\infty$.

This recurrence is identical to that used in pseudoknot-free algorithms (Mathews et al., 1999), so we omit the details here and for *VBI* in the next section.

4.2.7. $VBI_{i,j}$. $VBI_{i,j}$ is the minimum free energy of all structures $R_{i,j}$ over region $[i, j]$, if $[i, j]$ is weakly closed or empty, assuming $i.j$ closes a bulge or internal loop of $R_{i,j}$. Otherwise, $VBI_{i,j}$ is $+\infty$ (Mathews et al., 1999).

4.2.8. VM_{ij}. VM_{ij} is the minimum free energy of all structures R_{ij} over region $[i, j]$, if $[i, j]$ is weakly closed or empty and $i.j$ closes a multiloop of R_{ij}. Otherwise, VM_{ij} is $+\infty$. Figure 17 illustrates different cases of *VM* recurrence. We can obtain a recurrence that calculates the loop cost as the sum of two subparts.

$$VM_{ij} = \min \begin{cases} (1) & \min\limits_{i+1 < h \leq j-1} (WM_{(i+1)(h-1)} + WM_{h(j-1)} + a + b) \\ (2) & WMB_{(i+1)(j-1)} + a + P_{sm} + b \end{cases} \tag{10}$$

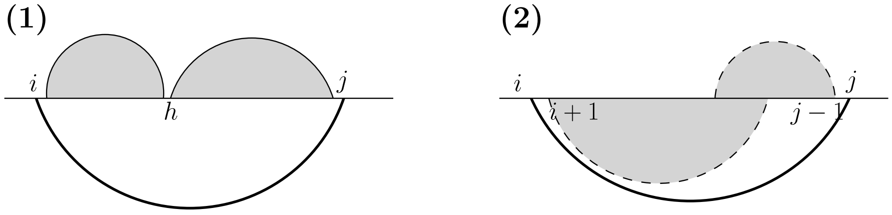


FIG. 17. Illustration of cases for VM_{ij}.

Case (1) is similar to recurrence for $W_{i,j}$, and case (2) handles the case that there is one pseudoknotted loops in the multiloop.

4.2.9. WM_{ij}. WM_{ij} is the minimum free energy of all structures R_{ij}, if $[i, j]$ is weakly closed, not empty, and i and j are on a multibranched loop.

The base case is as follows:

$WM_{ij} = +\infty$, if $i \geq j$.

Otherwise, WM_{ij} is given by the following recurrences:

$$WM_{i,j} = \min \begin{cases} (1) \;\; V_{i,j} + b, & \\ (2) \;\; WM_{(i+1),j} + c & bp_G(i) = 0 \\ (3) \;\; WM_{i,(j-1)} + c & bp_G(j) = 0 \\ (4) \;\; \min\limits_{1 \leq t \leq j} (WM_{i,t} + WM_{(t+1),j}) & \\ (5) \;\; WMB_{i,j} + P_{sm} + b & \end{cases} \tag{11}$$

Cases (1) to (4) are the same as in a pseudoknot-free structure, and case (5) handles the case of a pseudoknotted loop in the multiloop.

4.2.10. $BE_{i,i'}$. $BE_{i,i'}$ is the minimum free energy of the band $[i, i'] \cup [bp_G(i'), bp_G(i)]$, if $i \leq i' < bp_G(i') \leq bp_G(i)$. Figure 18 illustrates different cases of BE recurrence.

The base cases are as follows:

$BE_{i,i'} = +\infty$, if it is not the case that $i \leq i' < bp_G(i') \leq bp_G(i)$.
$BE_{i,i} = 0$, if $i < bp_G(i)$.

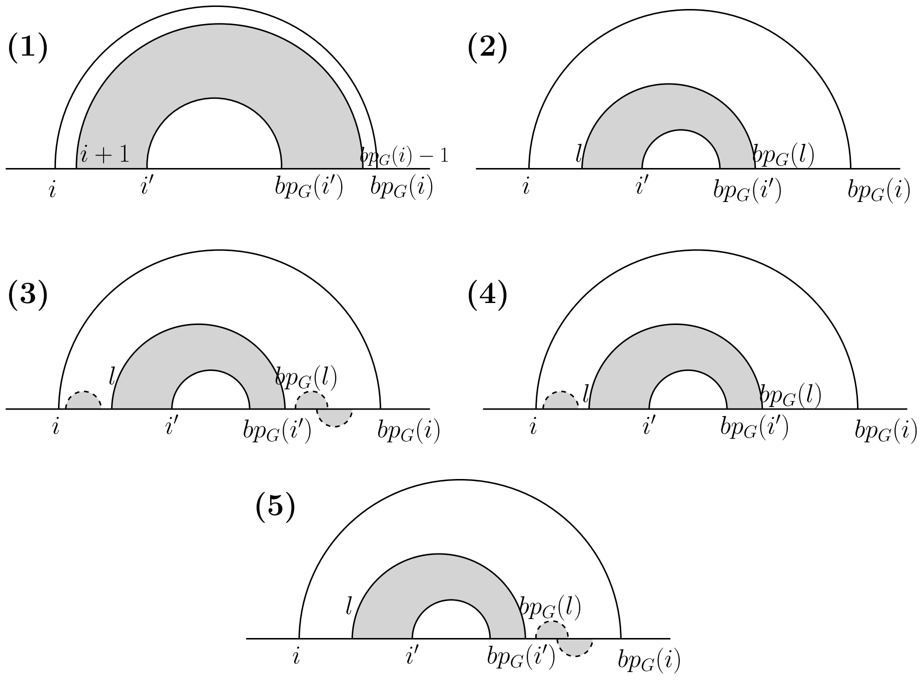


FIG. 18. Illustration of cases for $BE_{i,i'}$.

Otherwise,

$$
BE_{i,i'} = \min
$$

$$
\left\{
\begin{array}{ll}
(1)\ e_{stP}(i, bp_G(i)) + BE_{(i+1),i'} & \text{if } bp_G(i+1) = (bp_G(i) - 1) \\
(2)\ e_{intP}(i, l, bp_G(l), bp_G(i)) + BE_{l,i'} & \text{if } bp_G(l) > 0,\ \text{empty}(G, [i+1, l-1], \\
 & \text{and empty}(G, [bp_G(l) + 1, bp_G(i) - 1])), \\
 & i < l \leq i',\ \text{and } (bp_G(i') \leq bp_G(l) < bp_G(i)) \\
(3)\ WI'_{(i+1),(l-1)} + BE_{l,i'} & \text{if } bp_G(l) > 0,\ \text{weakly closed}(G, [i+1, l-1]) \\
\quad +\ WI'_{(bp_G(l)+1),(bp_G(i)-1)} + a' + 3b' & \text{and weakly closed}(G, [bp_G(l) + 1, bp_G(i) - 1]) \\
 & \text{and } i < l \leq i',\ \text{and } bp_G(i') \leq bp_G(l) < bp_G(i) \\
(4)\ WI'_{(i+1)(l-1)} + BE_{l,i'} & \text{if } bp_G(l) > 0,\ \text{weakly closed}(G, [i+1, l-1]) \\
\quad +\ a' + 2b' + c'(bp_G(i) - bp_G(l) + 1) & \text{and empty}(G, [bp_G(l) + 1, bp_G(i) - 1]) \\
 & \text{and } i < l \leq i',\ \text{and } bp_G(i') \leq bp_G(l) < bp_G(i) \\
(5)\ BE_{l,i'} + WI'_{(bp_G(l)+1),(bp_G(i)-1)} & \text{if } bp_G(l) > 0 \\
\quad +\ a' + 2b' + c'(l - i + 1) & \text{weakly closed}(G, [bp_G(l) + 1, bp_G(i) - 1]),\ \text{and} \\
 & \text{empty}(G, [i+1, l-1]) \text{ and} \\
 & i < l \leq i' \text{ and } bp_G(i') \leq bp_G(l) < bp_G(i)
\end{array}
\right. \tag{12}
$$

Case (1) handles the case that base pairs $i.bp_G(i)$ and $(i + 1).(bp_G(i) - 1)$ of G form a stacked loop in the band.

Case (2) handles the case that $i.bp_G(i)$ and $l.bp_G(l)$ are the base pairs of an internal loop of G.

Case (3) handles a similar situation as in case (2) except that there are other closed regions in both of the regions $[i, l]$ and $[bp_G(l), bp_G(i)]$.

Case (4) handles the case that the region $[bp_G(l), bp_G(i)]$ is empty, and so we must pay the unpaired base penalty c' for each unpaired base. In this case, the left side, $[i, l]$ must not be empty.

Case (5) is the same as case (4) except the left side is empty and the right is not empty.

5. CONCLUSIONS

In this work, we presented HFold, a new dynamic programming algorithm that efficiently predicts RNA secondary structure including pseudoknots in $O(n^3)$ time, based on the hierarchical folding hypothesis. HFold can predict kissing hairpins and pseudoloops with arbitrary number of bands.

In the conference version of this paper (Jabbari et al., 2007), we presented preliminary results on the accuracy obtained by HFold. Using parameters for pseudoknotted structures obtained from Dirks and Pierce (2003), HFold did not predict the desired results. However, we believe that it is still premature to draw conclusions since the parameters are poor. For this reason, we defer discussion of experimental results to future work.

Specifically, a short-term goal is to tune the parameters of the current energy model to improve the accuracy of the prediction for wider sets of structures. One possible approach is using Andronescu's tuning method (Andronescu et al., 2007). Another future work is to use a better energy model for pseudoknotted structures, such as that of Cao and Chen (2006), and obtain better energy parameters. This is also of great interest to us to evaluate the hierarchical folding hypothesis using computational methods in future.

We are not yet able to do a sound comparison of the prediction accuracy of HFold with MFE-based methods, since it would be important to ensure that the same energy model is used by both methods. Therefore, one of the main goals for our future work is to compare hierarchical and MFE algorithms implemented using the same energy model, at least for H-type pseudoknots.

Finally, we plan to incorporate other techniques to produce better input structures to HFold, such as information obtained from chemical modification data (Mathews et al., 2004).

REFERENCES

Akutsu, T. 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Disc. Appl. Math.* 104, 45–62.

Alam, S.L., Atkins, J.F., and Gesteland, R.F. 1999. Programmed ribosomal frameshifting: much ado about knotting! *Proc. Natl. Acad. Sci. USA* 96, 14177–14179.

Andronescu, M., Condon, A., Hoos, H.H., et al. 2007. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* 23, i19–i28.

Cao, S., and Chen, S. 2006. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.* 34, 2634–2652.

Deiman, B.A.L.M., and Pleij, C.W.A. 1997. Pseudoknots: a vital feature in viral RNA. *Semin. Virol.* 8, 166–175.

Dennis, C. 2002. The brave new world of RNA. *Nature* 418, 122–124.

Dirks, R.M., and Pierce, N.A. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* 24, 1664–1677.

Han, K., Lee, Y., and Kim, W. 2002. Pseudoviewer: automatic visualization of RNA pseudoknots. *Bioinformatics* 18, S321–S328. Suppl. 1.

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

Jabbari, H., Condon, A., Pop, A., et al. 2007. HFold: RNA pseudoknotted secondary structure prediction using hierarchical folding. *Algorithms Bioinform.*, 4645, 323–334.

Lyngsø, R.B. 2004. Complexity of pseudoknot prediction in simple models. *Lect. Notes Comput. Sci.* 3142, 919–931.

Lyngsø, R.B., and Pedersen, C.N. 2000. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* 7, 409–427.

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

Mathews, D.H. 2006. Predicting RNA secondary structure by free energy minimization. *Theor. Chim. Acta*, 116, 160–168.

Mathews, D.H., Disney, M.D., Childs, J.L., et al. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA* 101, 7287–7292.

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

Melchers, W.J., Hoenderop, J.G., Bruins Slot, H.J., et al. 1997. Kissing of the two predominant hairpin loops in the coxsackie B virus 3′ untranslated region is the essential structural feature of the origin of replication required for negative-strand RNA synthesis. *J. Virol.* 71, 686–696.

Rastegari, B., and Condon, A. 2007. Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. *J. Comput. Biol.* 14, 16–32.

Reeder, J., and Giegerich, R. 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinform.* 5.

Rivas, E., and Eddy, S.R. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285, 2053–2068.

Staple, D.W., and Butcher, S.E. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biol.* 3, e213.

Tinoco, I., and Bustamante, C. 1999. How RNA folds. *J. Mol. Biol.* 293, 271–281.

Uemura, Y., Hasegawa, A., Kobayashi, S., et al. 1999. Tree adjoining grammars for RNA structure prediction. *Theor. Comput. Sci.* 210, 277–303.

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

Witwer, C., Hofacker, I.L., and Stadler, P.F. 2004. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 66–77.

Wu, M., and Tinoco, I. 1998. RNA folding causes secondary structure rearrangement. *Proc. Natl. Acad. Sci. USA* 95, 11555–11560.

Address reprint requests to:
Hosna Jabbari
Department of Computer Science
University of British Columbia
201-2366 Main Mall
Vancouver V6T 1Z4, BC, Canada

E-mail: hjabbari@cs.ubc.ca