RNA Secondary Structure Prediction Using Hierarchical Folding

by

Hosna Jabbari

B.Sc., The University of Victoria, 2005

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Computer Science)

The University Of British Columbia

August, 2007

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 analyze 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, based on work of S. Zhao, 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.

We also report on the experimental evaluations of HFold and present thorough analyses of the results. We show that if the input structure to the algorithm is correct, running the algorithm results in 16% accuracy improvement on average over the accuracy of the true pseudoknot free structures. However if the input structure is not correct, the accuracy improvement is not significant. If the first 10 suboptimal foldings are given as input to our algorithm instead of just the minimum free energy structure (MFE), the prediction accuracy improves significantly over the accuracy of the MFE structures. This improvement is even more when the number of suboptimal foldings as input to our algorithm increases. The comparison of the energy of the structures predicted by HFold on the true pseudoknot free structures with the energy of the true structures calculated using a different method with the same energy model shows that the energy model may be the cause for the cases for which HFold predicts structures far from the true structures. Our experimental result provides some ways in which the hierarchical folding hypothesis might need to be refined.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor, professor Anne Condon, for her endless support and encouragements. She introduced me to a new and interesting branch of science that I did not know about before. I cannot thank her enough for all she has done for me.

I would also like to thank Dr. Will Evans for being my second reader and for his helpful comments.

Many thanks go to people of Beta lab for making Beta such a lovely place to work.

Last, but not least, I would like to thank my parents, my sisters and my wonderful husband for believing in me through all stages of my studies. Without their love and support I could not have done anything.

Hosna Jabbari.

To my best friend and wonderful husband,

Majid.

Chapter 1

Introduction

1.1 RNA Secondary Structure

RNA molecules aid in translation and replication of the genetic code, catalyze cellular processes, and regulate the expression level of genes [7]. 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 (see Fig. 1.1). State-of-the-art prediction algorithms, such as Mfold [18] or RNAfold [11] find the structure with *minimum free energy* (MFE) from the set of all possible pseudoknot free secondary 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 [24, 27], and in viral RNA [6]. Examples include simple H-type pseudoknots, with two interleaved stems, which are essential for certain catalytic functions and for ribosomal frameshifting [2], as well as kissing hairpins, which are essential for replication in the coxsackie B virus [19].

Unfortunately, MFE pseudoknotted secondary structure prediction is NP-hard [1, 13, 14], 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 [1, 8, 21, 22, 26], 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* – the class of structures handled, 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 [1] and Dirks and Pierce [8] but are in the class handled by the $\Theta(n^6)$ algorithm of Rivas and Eddy [22]. (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* [16, 25], 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" [25]. (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 [29]. 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.

This work is a continuation of the M.Sc. work of Zhao [30], in which she presented a novel and efficient algorithm to predict RNA secondary structures, in a manner consistent with a natural formalization of the hierarchical folding hypothesis. She defined the problem of predicting the secondary structure as follows: given a sequence S and a pseudoknot free secondary structure G (a set of base pairings), find a pseudoknot free secondary structure G' (a set of base pairings 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. [28] – those structures which can be partitioned into two pseudoknot free secondary structures G and G'. There is no known way to solve the hierarchical-MFE prediction for the class of bi-secondary structures. Instead, Zhao suggested a solution with respect to a subclass of the bi-secondary structures, which she called *density-2* structures, explained in Section 2.

This is quite a general class, 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 [20] showed that, out of a set of over 1,100 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.

The main contributions of this work are:

- refining and fixing the original recurrences presented in the work of Zhao [30],
- defining and proving four lemmas, that help establish corrections of Zhao's recurrences,
- implementing the HFold algorithm, and
- throughly analyzing its performance on a dataset of 70 strands.

We published a preliminarily version of this work in [12].

In Chapter 2, we present some useful background information and notations pertaining to RNA structure prediction. In Chapter 3, we summarize 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, and present our lemmas.

In Chapter 4, we describe the results of our experimental analysis. Our experimental evaluation of HFold shows that, when provided with the true pseudoknot free substructure for the input sequence, HFold adds pseudoknots which, on average, improve the accuracy (measured as fraction of correctly predicted bases) by 16%. However, HFold does not significantly improve accuracy when given as input a computational prediction of the MFE pseudoknot free secondary structure G, since HFold cannot correct errors in G.

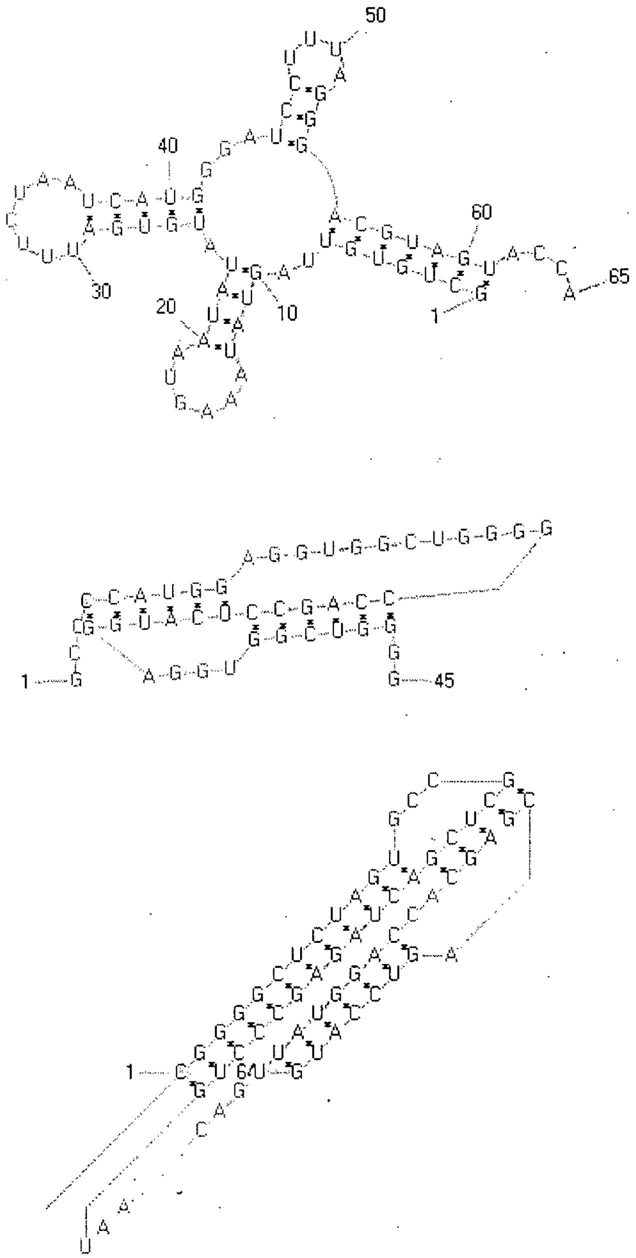


Figure 1.1: A pseudoknot free structure (top), an H-type pseudoknotted structure (center) and a kissing hairpin (bottom).
Figures were generated by PseudoViewer [10].

Chapter 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 certain pairs of bases, where each base may pair with at most one other base. 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. The pair $i.j$ denotes that the base indexed i is paired with the base indexed j. 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 (see Fig. 2.1).

2.1 Notation

We use the following notations when describing our algorithms. Throughout, we use R to denote a secondary structure.

- $bp_R(i)$: 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 $i.j$, $i'.j'$, and $i < i' < j < j'$, we say that pair $i.j$ crosses pair $i'.j'$ (and $i'.j'$ crosses $i.j$).
- **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 the given structure, it is called pseudoknot free secondary structure.

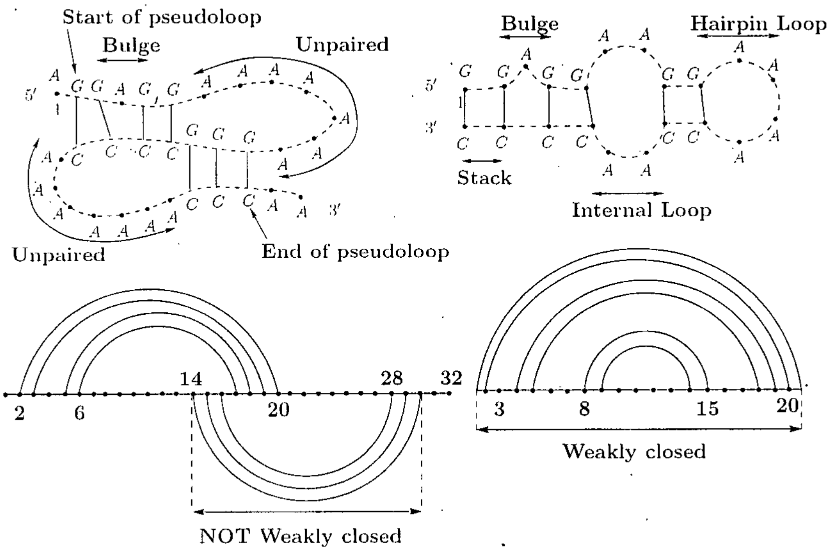


Figure 2.1: An H-type pseudoknotted structure (left) and a pseudoknot free structure (right) in graphical (top) and arc diagram (bottom) formats.

- **cover**: Let G be a pseudoknot free secondary structure. Base pair $i.j$ *covers* base k if $i \leq k \leq j$ 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.
- **region** $[i, j]$: Sequence of indices between i and j inclusive.
- **disjoint regions**: Two regions $[i, j]$ and $[i', j']$ are disjoint if no index is in both regions, i.e. $j < i'$ or $j' < i$.

2.2 Region and Loop Classification and Related Definitions

Following is the terminology used to refer to loops and other standard elements in a secondary structure. These definitions are mostly taken from and illustrated in the work of Rastegari and Condon [20]. 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.

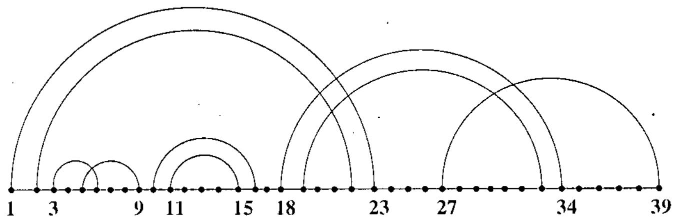


Figure 2.2: Pseudoknot

- **empty**$(R, [i, j])$ with respect to secondary structure R: 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 with at least two bases, $[i, j]$, 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) [20].

- **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 $[i, j]$.

- **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) $i' \leq i < j \leq j'$, and
(2) $[i'+1, i-1]$ and $[j+1, j'-1]$ are weakly closed regions.
"Directly banded in" is transitive: $i.j \preceq i'.j' \preceq i''.j''$ implies $i.j \preceq i''.j''$.

- **band**: Let us consider a maximal chain of $\preceq$. The minimum (maximum) base pair in the maximal chain is the band's inner (outer) closing pair. If $i.j$ is the outer and $i'.j'$ the inner closing pair of a band, then $[i, i']$ and $[j', j]$ are the band's regions. For example there are 3 bands in Fig. 2.2: $[1, 2] \cup [22, 23]$, $[18, 19] \cup [33, 34]$ and $[27, 27] \cup [39, 39]$.

i and i' are the borders of $[i, i']$ band region, j and j' are the borders of $[j', j]$ band region. We refer to i and j as the left and the right border 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.

- **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 child of $[i'', j'']$. For example, in Fig 2.2, the three bands $[1, 2] \cup [22, 23]$, $[18, 19] \cup [33, 34]$ and $[27, 27] \cup [39, 39]$ are associated with closed region $[1, 39]$.

- **unpaired bases associated with closed region** $[i, j]$: 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 Fig. 2.2, the unpaired bases associated with closed region $[1, 39]$ are 17, 20, 21, 24-26, 28-32, and 35-38.

- **base pairings associated with closed region** $[i, j]$: are the base pairings in $[i, j]$ but not in any closed region or band region which are subregions of $[i, j]$.

- **hairpin loop** (or **hairpin**): 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 Fig. 2.1 contains 4 unpaired bases.

- **internal loop**: An internal loop, sometimes called interior loop, contains two closing base pairs, and all bases between them are unpaired. The tuple (i, i', j', j), with $i + 1 < i' < j' < j - 1$, defines an internal loop if $i.j$ and $i'.j'$, and $[i+1, i'-1]$ and $[j'+1, j-1]$ are empty regions.

- **stacked loop**: A stacked loop, also called stacked pair, contains two consecutive base pairs. The tuple (i, 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.

- **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 region) children, or a pseudoknotted child. 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 associated with 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. For example, in the structure of Fig. 2.2, $[2,22]$ shows a multiloop that spans a band. Note that 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 closing base pairs of the bands associated with $[i,j]$, are members of a pseudoloop. 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 subregion of any other region.

- **closed region associated with pseudoloop**: We say that closed region $[i',j']$ is *associated with* pseudoloop $[i,j]$, 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 Fig. 2.2, closed regions $[3,9]$ and $[10,16]$ are associated with pseudoloop $[1,39]$ but closed region $[11,15]$ is *not* associated with pseudoloop $[1,39]$.

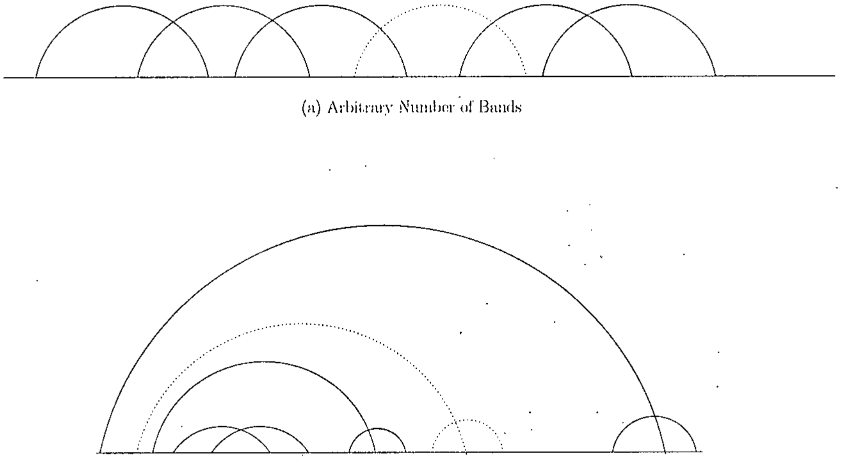
2.3 Bi-secondary and Density-2 Structures

- **Bi-secondary structures**: Witwer et al. [28] 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:

$$density(L) = \max_{i \leq k \leq j} (\#B(L,k)) \tag{2.1}$$

The density of a structure, R, is 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



(a) Arbitrary Number of Bands

(b) Arbitrary Depth of Bands

Figure 2.3: Density-2 structures

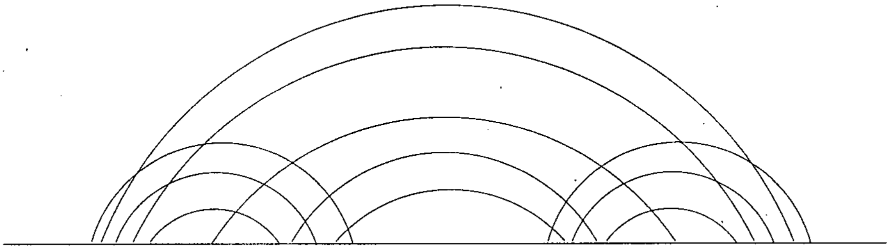


Figure 2.4: Bi-secondary structure that is not density-2.

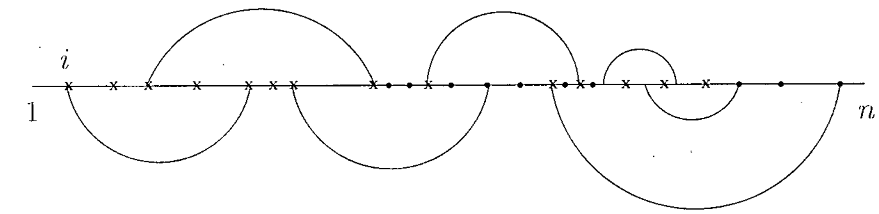


Figure 2.5: The figure shows a density-2 secondary structure for a sequence of length n. R_{ij} is the structure restricted to the region $[i, j]$ (for $i < j \leq n$) and G_{ij} is that part of R_{ij} above the horizontal line. The circle dots show positions of j where R_{ij} is a prefix of a density-2 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.

most 2. Figure 2.3 illustrates density-2 secondary structures. Figure 2.4 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 starts 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,
    3. the rightmost border of any band associated with $R_{ij} - G_{ij}$ 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).

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 $i.j$; we assume that for all but a small number of cases, $e_H(i,j)$ depends only on the length of the loop, and the two paired bases i and j on the loop.
- $e_S(i,j)$: 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. We define the free energy of a strand S to be the minimum free energy of the strand, with respect to all structures R.

2.4.2 Pseudoknotted energy model

- $BE(i,i',j',j)$: The total energy of band $[i,i'] \cup [j',j]$ 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.
- $e_{stP}(i,j)$: defines the energy of stacked pairs in a band, and its value is $e_S(i,j) * 0.83$.
- $e_{intP}(i,r,r',j)$: defines the internal loop that spans a band, and its value is $e_{int}(i,r,r',j) * 0.83$.

We define energy of multiloops that span a band to be similar to pseudoknot free multiloops.

The energy of an exterior pseudoloop is calculated as: energy of bands plus $P_b{*}m + P_{ps}{*}k + P_{up}{*}u + P_s$, where m is the number of the bands, k is the number of closed subregions, 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 2.1 summarizes the energy constants and functions used in our energy model for pseudoknotted structure.

Table 2.1: Energy parameters.

Name	Description	Value (Kcal/mol)
P_s	exterior pseudoloop initiation penalty	9.6
P_{sm}	penalty for introducing pseudoknot inside a multiloop	15.0
P_{sp}	penalty for introducing pseudoknot inside a pseudoloop	15.0
P_b	band penalty	0.2
P_{up}	penalty for unpaired base in a pseudoloop	0.1
P_{ps}	penalty for closed subregion inside a pseudoloop	0.1
$e_H(i,j)$	energy of a hairpin loop closed by $i.j$	
$e_S(i,j)$	energy of stacked pair closed by $i.j$	
$e_{stP}(i,j)$	energy of stacked pair that spans a band	$e_S(i,j) \times 0.83$
$e_{int}(i,r,r',j)$	energy of a pseudoknot free internal loop	
$e_{intP}(i,r,r',j)$	energy of internal loop that spans a band	$e_{int}(i,r,r',j) \times 0.83$
a	multiloop initiation penalty	3.4
b	multiloop base pair penalty	0.4
c	penalty for unpaired base in a multiloop	0
a'	penalty for introducing a multiloop that spans a band	3.4
b'	base pair penalty for a multiloop that spans a band	0.4
c'	penalty for unpaired base in a multiloop that spans a band	0

Chapter 3

Algorithm

3.1 A Useful Property 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. Roughly, the lemmas show how to calculate the band borders for a given region.

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 $l \in [i+1, j]$ be such that*

1. *l is paired in G' (but not in G),*
2. *l is not in any closed proper subregion of $[i, j]$ with respect to $G \cup G'$, and*
3. *$\exists r : i \leq r < l < bp_G(r) \leq j$.*

Let

$$\begin{aligned} b_{(i,l)} &= \min\{k | i \leq k < l < bp_G(k)\}, \text{ and} \\ b'_{(i,l)} &= \max\{k | i \leq k < l < bp_G(k)\}. \end{aligned}$$

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 $b'_{(i,l)}.bp_G(b'_{(i,l)})$.

Note that restriction (3) ensures that $b_{(i,l)}$ and $b'_{(i,l)}$ are well defined. The bottom left part of Fig. 2.1 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

Throughout the proof we assume that $l < bp_{G'}(l)$, but the proof can easily be modified to handle the case when $bp_{G'}(l) < l$ by substituting $bp_{G'}(l).l$ instead of $l.bp_{G'}(l)$ everywhere in the 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 $bp_{G'}(l) \in [i, j]$, since if it is not, then $[i, j]$ is not a closed region of $G \cup G'$.

We claim that for any r satisfying restriction (3), $l.bp_{G'}(l)$ crosses $r.bp_G(r)$.

If $l.bp_{G'}(l)$ does not cross $r.bp_G(r)$, then it must be that $i \leq r < l < bp_{G'}(l) < bp_G(r)$. Also, we must have one of the following cases:

- $l.bp_{G'}(l)$ does not cross another base pair. Then region $[l, bp_{G'}(l)]$ is a closed subregion of $[i, j]$ with respect to $G \cup G'$, and thus, l is in a closed subregion of $[i, j]$ with respect to $G \cup G'$, which is a contradiction.

- $l.bp_{G'}(l)$ crosses another base pair, say $m.bp_G(m)$. Then $l.bp_{G'}(l)$ is in a band associated with a pseudoloop. We claim that, this pseudoloop is a subregion of $[r, bp_G(r)]$. Otherwise, it must also be that $r.bp_G(r)$ is in a band associated with the pseudoloop. Moreover, the bands containing $r.bp_G(r)$, $m.bp_G(m)$ and $l.bp_{G'}(l)$ must all be distinct, based on the definition of band and the fact that $l.bp_{G'}(l)$ is in $[r, bp_G(r)]$, and $m.bp_G(m) \in G$ and thus, cannot cross $r.bp_G(r)$. A line drawn vertically at position m or at position l must cross all three bands. This is because $m.bp_G(m)$ crosses $l.bp_{G'}(l)$ which is in $[r, bp_G(r)]$. Therefore, $G \cup G'$ must have density 3, contradiction.

Therefore, our assumption is incorrect and $l.bp_{G'}(l)$ crosses $r.bp_G(r)$.

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 $r.bp_G(r)$ that $l.bp_{G'}(l)$ crosses. We have $i \leq b_1 \leq b_2 < l < bp_G(b_2) \leq bp_G(b_1) \leq j$.

Now we prove that $b_{(i,l)} = b_1$. Since $i \leq b_1 < l < bp_G(l)$ it must be that $b_{(i,l)} \leq b_1$, 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. By same argument as used above, we conclude that $b_{(i,l)}.bp_G(b_{(i,l)})$ crosses $l.bp_{G'}(l)$. Thus, $b_{(i,l)}.bp_G(b_{(i,l)})$ is the outer base pair of the band, 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 $[i, j]$ be a region such that $l \in [i, j]$ and $bp_{G'}(l) \notin [i, j]$. Let*

$$\begin{aligned}
b_{(i,l)} &= \min\{k | i \leq k < l < bp_G(k)\} \cup \{\infty\}, \\
b'_{(i,l)} &= \max\{k | i \leq k < l < bp_G(k)\} \cup \{-1\}, \\
B_{(l,j)} &= \max\{bp_G(k) | k < l < bp_G(k) \leq j\} \cup \{-1\}, \text{ and} \\
B'_{(l,j)} &= \min\{bp_G(k) | k < l < bp_G(k) \leq j\} \cup \{\infty\}.
\end{aligned}$$

Then, if $[i, j]$ is a weakly closed region of G, then either all four of these quantities have finite, positive values, in which case the structure $G \cup G'$ contains a band with outer base pair $b_{(i,l)}.B_{(l,j)}$ and inner base pair $b'_{(i,l)}.B'_{(l,j)}$ (i.e. $bp(b_{(i,l)}) = B_{(l,j)}$ and $bp(b'_{(i,l)}) = B'_{(l,j)}$), or none of the four of these quantities have finite, positive values, in which case l is not covered by a base pair of G_{ij}.

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

For region $[i,j] = [1,23]$, which is weakly closed in G, in Fig. 2.2, we have $b_{(1,23)} = 1$, $b'_{(1,23)} = 2$ and $B'_{(1,23)} = 22$ and $B_{(1,23)} = 23$.

Proof

We first show that if $[i,j]$ is a weakly closed region of G, then either all four of these quantities have finite, positive values, or none of the four of these quantities have finite, positive values.

Throughout the proof we assume that $l < bp_{G'}(l)$, but the proof can easily be modified to handle the case when $bp_{G'}(l) < l$. Let us consider $\text{Cover}(G,l) = (r, bp_G(r))$. Note that $\text{Cover}(G,l)$ may be $(-1,-1)$ if l is not covered in G. Based on the definition of weakly closed region, either $\text{Cover}(G,l) \in [i,j]$ or $\text{Cover}(G,l) \notin [i,j]$. Therefore, we either have $r < i < l < j < bp_G(r)$, in which case, none of the four quantities have finite positive values, or we have $i \leq r < l < bp_G(r) \leq j$, in which case $r.bp_G(r)$ crosses $l.bp_{G'}(l)$, and therefore all four quantities have finite positive values.

Now we claim when $[i,j]$ is weakly closed, $\text{Cover}(G,l) = (r, bp_G(r))$, and $r.bp_G(r) \in [i,j]$, the structure $G \cup G'$ contains a band with outer base pair $b_{(i,l)}.B_{(l,j)}$ and inner base pair $b'_{(i,l)}.B'_{(l,j)}$. To prove this claim we first prove that $bp_G(b_{(i,l)}) = B_{(l,j)}$, and $bp_G(b'_{(i,l)}) = B'_{(l,j)}$.
Assume $bp_G(b_{(i,l)}) \neq B_{(l,j)}$. Then, since both $b_{(i,l)}.bp_G(b_{(i,l)})$ and $bp_G(B_{(l,j)}).B_{(l,j)}$ are in G, then they cannot cross, and one must be nested in the other. If $b_{(i,l)}.bp_G(b_{(i,l)})$ is in $[bp_G(B_{(l,j)}), B_{(l,j)}]$, then we have $i \leq bp_G(B_{(l,j)}) < b_{(i,l)} < l$, which is contradiction to the definition of $b_{(i,l)}$. Therefore $b_{(i,l)}.bp_G(b_{(i,l)})$ cannot be in $[bp_G(B_{(l,j)}), B_{(l,j)}]$. With a similar argument we can show that $bp_G(B_{(l,j)}).B_{(l,j)}$ cannot be in $[b_{(i,l)}, bp_G(b_{(i,l)})]$. Thus, we must have $bp_G(b_{(i,l)}) = B_{(l,j)}$. The proof to show that $bp_G(b'_{(i,l)}) = B'_{(l,j)}$ is similar to what came above.

Now we prove that a band containing $b_{(i,l)}.B_{(l,j)}$, also contains $b'_{(i,l)}.B'_{(l,j)}$. Assume the bands containing $b_{(i,l)}.B_{(l,j)}$ and $b'_{(i,l)}.B'_{(l,j)}$ are distinct. Then, a vertical line drawn at l crosses 3 distinct bands. Therefore the structure of $G \cup G'$ has density 3, contradiction. Thus, one band contains both $b_{(i,l)}.B_{(l,j)}$ and $b'_{(i,l)}.B'_{(l,j)}$.

Based on the definition of $b'_{(i,l)}$, there is no other base pair $m.bp_G(m)$ in G, such that $b'_{(i,l)} < m < l < bp_G(m) < B'_{(l,j)}$. Therefore, $b'_{(i,l)}.B'_{(l,j)}$ is the inner base pair of the band.

We can easily show that if $b_{(i,l)}.B_{(l,j)}$ is not the outer base pair of the band, then we must have a base pair $m.bp_G(m) \in [i,j]$, such that $m < b_{(i,l)} < B_{(l,j)} < bp_G(m)$. It is clear that $m.bp_G(m)$ must be contained in the same band as $b_{(i,l)}.B_{(l,j)}$. Thus, contradiction. Therefore $b_{(i,l)}.B_{(l,j)}$ is the outer base pair of

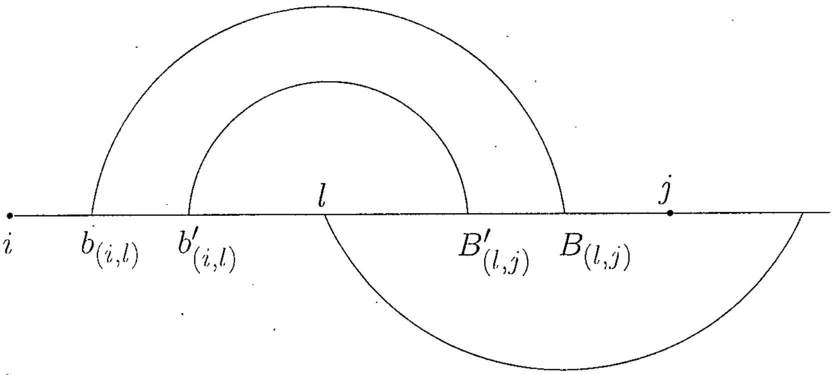


Figure 3.1: Illustration of band borders in Lemma 3 and Lemma 4

the band. □

Lemma 3 *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 $l \in [i,l]$ and $bp_{G'}(l) \notin [i,l]$. Let*

$$\begin{aligned} b_{(i,l)} &= \min\{k | i \leq k < l < bp_G(k)\} \cup \{\infty\}, \text{ and} \\ b'_{(i,l)} &= \max\{k | i \leq k < l < bp_G(k)\} \cup \{-1\}. \end{aligned}$$

Then, either both or neither of $b_{(i,l)}$ and $b'_{(i,l)}$ have finite positive values.

For example, for region $[17,19]$ in Fig. 2.2, we have $b_{(17,19)} = 18$, $b'_{(17,19)} = 19$.

Proof

Let us consider Cover$(G,l) = (r, bp_G(r))$. Note that Cover(G,l) may be $(-1,-1)$ if l is not covered in G. If $r < i$ then there must be no base pair k such that $i \leq k < l < bp_G(k)$. Otherwise, since both $r.bp_G(r)$ and $k.bp_G(k)$ are in G, we must have $r < k < l < bp_G(k) < bp_G(r)$. Thus, we must have Cover$(G,l) = (k, bp_G(k))$ for one such k, which is contradiction. Therefore, $b_{(i,l)} = \infty$ and $b'_{(i,l)} = -1$.

Let us assume $i \leq r$, Cover$(G,l) = (r, bp_G(r))$, $l = j$ and $bp_{G'}(l) = i-1$. Then, $r.bp_G(r)$ must cross $l.bp_{G'}(l)$, since $i \leq r$ and based on the definition of cover, we must have $r < l < bp_G(r)$. We claim that both $b_{(i,l)}$ and $b'_{(i,l)}$ have finite positive values. Moreover, $b'_{(i,l)} = r$.
Assume $b'_{(i,l)} \neq r$. Then, $b'_{(i,l)}.bp_G(b_{(i,l)})$ must be in $[r, bp_G(r)]$ (based on the definition of $b'_{(i,l)}$). Therefore we have $r < b'_{(i,l)} < l < bp_G(b'_{(i,l)}) < bp_G(r)$, contradiction. Thus, $b'_{(i,l)} = r$. Now, clearly $b_{(i,l)} \neq \infty$. We claim that if

$b_{(i,l)} \neq b'_{(i,l)}$, then $b'_{(i,l)}.bp_G(b'_{(i,l)})$ is in $[b_{(i,l)}, bp_G(b_{(i,l)})]$. Otherwise, the bands containing $b_{(i,l)}.bp_G(b_{(i,l)})$, $b'_{(i,l)}.bp_G(b'_{(i,l)})$ and $l.bp_{G'}(l)$ must be distinct. A vertical line drawn vertically at position l crosses 3 bands, thus $G \cup G'$ has density 3, contradiction. □

Lemma 4 *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 $[l, j]$ be a region such that $l \in [l, j]$ and $bp_{G'}(l) \notin [l, j]$. Let*

$$\begin{aligned} B_{(l,j)} &= \max\{bp_G(k) | k < l < bp_G(k) \leq j\} \cup \{-1\}, \text{ and} \\ B'_{(l,j)} &= \min\{bp_G(k) | k < l < bp_G(k) \leq j\} \cup \{\infty\}. \end{aligned}$$

Then, either both or neither of $B_{(l,j)}$ and $B'_{(l,j)}$ have finite positive values.

For example, for region $[9, 11]$ in Fig. 2.2, we have $B'_{(9,11)} = B_{(9,11)} = 9$.

Proof

Proof is very similar to proof of Lemma 3. □

3.2 High Level Description of HFold

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) [15, 25]. 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.

Here we outline our hierarchical fold algorithm. We first briefly review key ideas of dynamic programming algorithm which predicts the energy of the MFE pseudoknot free secondary structure for a fixed sequence $S = s_1 s_2 \ldots s_n$ [18]. 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 first two rows of the recurrence for $W_{i,j}$ below.

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

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

where $V_{i,j}$ is the free energy of the MFE structure for $s_i \ldots s_j$ that contains $i.j$. The recurrence for $V_{i,j}$ can in turn be expressed in terms of the energies of the hairpin loop $(e_H(i,j))$, an internal loop, or a multiloop closed by $i.j$, and multiloops. If $i \ge j$, $V_{i,j}$ 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 2.1, 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 $i \ge j$, 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:

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

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 2.1.

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 3.2 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

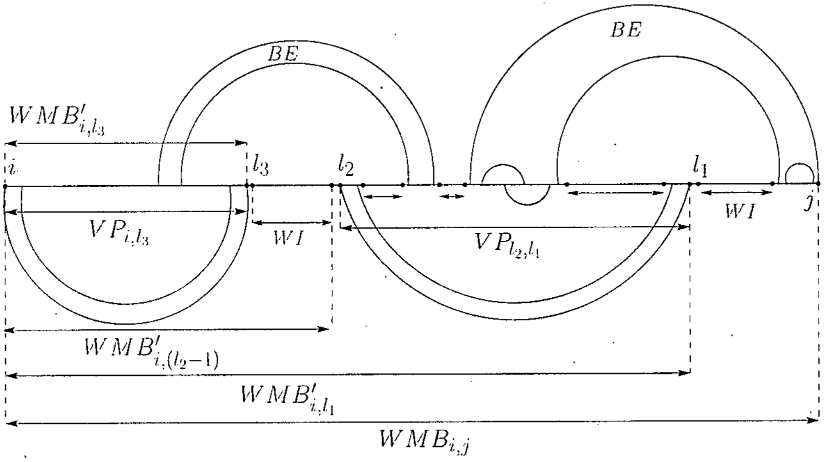


Figure 3.2: Illustration of how the *WMB* recurrence unwinds, to calculate $WMB_{i,j}$. Arcs above the horizontal line from i to j represent base pairs of G_{ij}, and arcs below the line represent base pairs of G'_{ij}. Case (a) of the *WMB* recurrence handles the overall structure whose energy is $WMB_{i,j}$, 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 (a) of the *WMB'* recurrence, with $l = l_2$ and terms to account for the lower right substructure labeled 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 (b) 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 (c) of the *WMB'* recurrence.

a band of G_{ij}), and weakly closed subregions, respectively. Further details on different recurrences can be found in Appendix A.

We now give the recurrence for $WMB_{i,j}$. 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 pseudoknotted. Otherwise, there are two cases, depending on whether j is paired in G or not. In case (a), 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 (a) 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 (b), j is not paired in G, and the recurrence is unwound by moving directly to a *WMB'* term. Thus, we have:

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

Complementing case (a) of the WMB recurrence, WMB' handles the case that the rightmost band is not in G, but is part of the structure G'. In the recurrence for WMB', case (a) 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 WMB' cases are determined using Lemmas 3 and 4. Case (b) is called when one iteration of $WMB_{i,j}$ or $WMB'_{i,j}$ case (a) 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 $cover(i) \neq cover(j)$, which avoids entering case (b) as the first iteration of WMB'. Cases (c) and (d) 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:

$$WMB'_{i,j} = \min$$

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

Figure 3.3 shows how all the recurrences call each other.

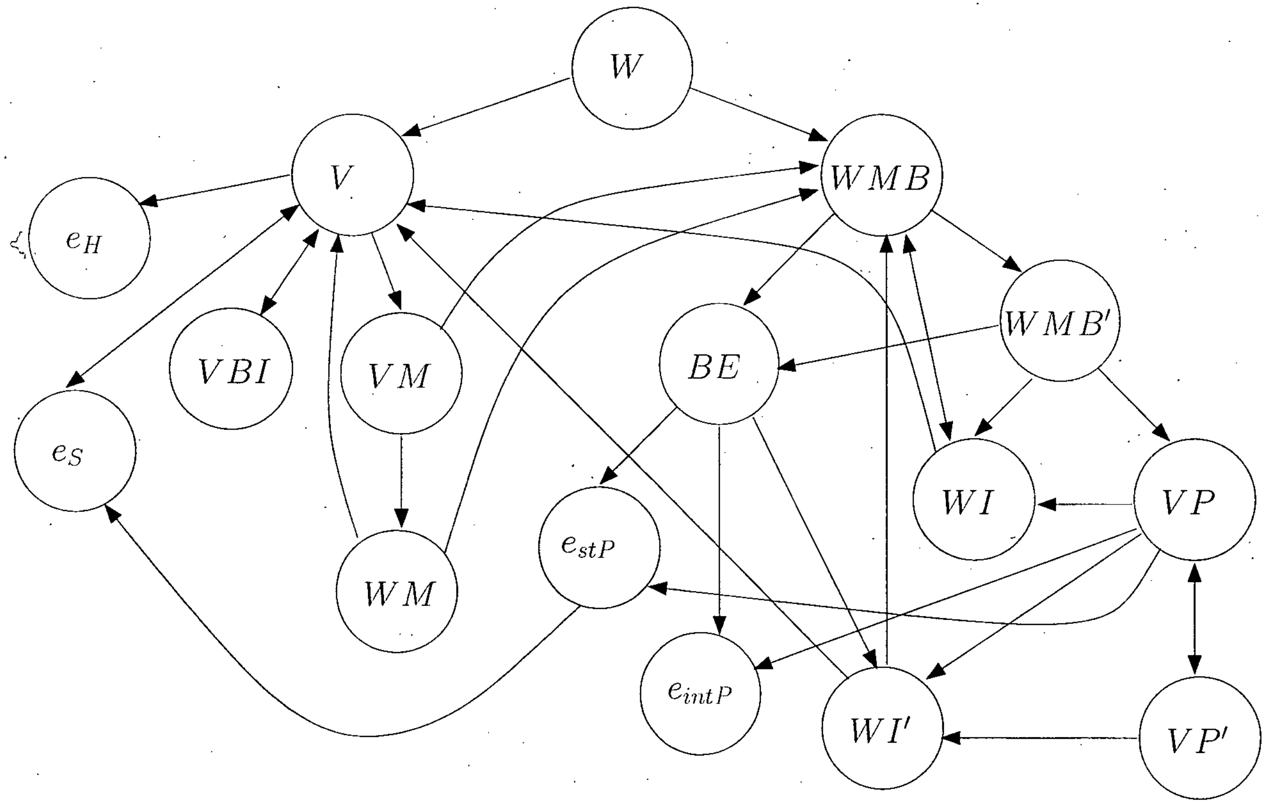


Figure 3.3: Visual illustration of recurrences in HFold.

Chapter 4

Results

The goals of our analyses were sevenfold:

- first, a baseline test to see if HFold finds pseudoknots when presented with the true pseudoknot free secondary structure for a sequence;
- second, to assess the accuracy of HFold, when using the predicted MFE pseudoknot free structure for a sequence as input;
- third, to see whether by taking the best HFold output, taken over several runs when suboptimal pseudoknot free structures are used as input, it is possible to obtain significant improvements in accuracy over simply running HFold with MFE structure as input;
- fourth, to see whether by taking the minimum free energy HFold output, taken over several runs when suboptimal pseudoknot free structures are used as input, it is possible to obtain significant improvements in accuracy over simply running HFold with MFE structure as input;
- fifth, to see whether by taking a substructure of the MFE structure in which bases are paired with high confidence we can gain any significant improvement in accuracy;
- sixth, to see whether by taking the MFE structure and peeling away the stems from inside and/or outside stems that close hairpin loops it is possible to gain better prediction accuracy than with MFE alone;
- and finally to compare the energy of the true structure, calculated with a separate method using the same energy model, with the energy of the structure output by HFold, when it is presented with the true pseudoknot free structure.

We implemented the HFold algorithm in C++, using the energy model described in Section 2.4. Our implementation builds on the SimFold algorithm of Andronescu [3]. All experiments were run on a SUSE Linux 10.1 with two Intel 2.40 GHz processors and 1 GB of RAM. We performed several analyses on a large dataset. In the following sections we present our thorough analyses.

4.1 Accuracy

In our analyses, we measure the accuracy of a predicted structure R for sequence S with true structure T as follows. Each base (position) i of S gets a score of 1 if $bp_R(i) = bp_T(i)$, and gets a score of 0 otherwise. The accuracy is the total score over all bases of S, divided by the length of the sequence. Thus, the accuracy lies between 0 and 1, with 1 indicating perfect accuracy.

4.2 Dataset

Our dataset of 70 sequences includes 7 with pseudoknot free and 63 with pseudoknotted structures. These 63 structures include 6 sequences with kissing interactions, 6 with H-type pseudoknots with nested structures and 1 sequence with density-3 structure. Tables 4.1, 4.2 and 4.3 show the sequences used in this work. The length of these sequences varies from 26 to 214 bases. The sequences include the following types: viral ribosomal frame shifting, viral ribosomal readthrough, sequences with high affinity to HIV-1-RT, mRNA, tmRNA, viral 3' UTR, ribozomes, viral RNA, signal recognition particle RNA, small nuclear RNA and tRNA [5, 6, 9, 23, 27]. Since there are no known energy parameters for non-canonical base pairs and loops with less than 3 unpaired bases inside in our energy model, we removed the non-canonical base pairings and the loops with less than 3 unpaired bases from the structures of our dataset. We partitioned each pseudoknotted structure in our dataset into two pseudoknot free structures G_{big} and G_{small}. Structure G_{big} was created by removing the minimum number of base pairs needed to get a pseudoknot free structure from the input pseudoknotted structure, and structure G_{small} consists of the removed base pairs. We call each of these pseudoknot free structures a *true* pseudoknot free secondary structure for the corresponding sequence.

4.3 Accuracy when input is the true pseudoknot free structure

We first test the accuracy of HFold, when presented with a sequence S whose true structure is pseudoknotted, and the corresponding true pseudoknot free secondary structures G_{big}. The average accuracy of the G_{big} structures obtained from pseudoknotted structures in our dataset is 77%. Figure 4.1 plots the accuracy of HFold(S, G_{big}) versus the accuracy of the G_{big} structures, for each sequence in our dataset. The data points in the lower triangle of Fig. 4.1 show the cases that HFold improved the accuracy of the original structure and the points on the upper triangle show the cases in which HFold added incorrect base pairings and thus decreased the accuracy of the final structure in comparison to the accuracy of the original structure. The average accuracy of the structure obtained *after* running HFold, namely HFold(S, G_{big}), is 77% averaged over 63

Sequence IDs and lengths	Sequence type	Ref.
Homo sapiens(164), Arabidopsis thaliana(164)	snRNA	[9]
RR4640(55), RK5280(67), RM8530(77), RG6241(66), RA1662(76)	tRNA	[23]

Table 4.1: Pseudoknot free sequences used for secondary structure prediction. In order, the columns provide (1) sequence ID, as found in the database or paper from which we obtained the sequence, and their length in parenthesis; (2) type of sequence; and (3) the reference from which the sequence was obtained.

structures (not including the pseudoknot free structures), with 6 cases achieving perfect accuracy.

Since G_{big} of the pseudoknot free structures is the pseudoknot free structures themselves, the input structures to HFold include 7 cases with perfect accuracy (presented in Fig. 4.1 with circle); from those cases 2 remain the same after running HFold (i.e. HFold$(S, G_{big}) = G_{big}$) but some extra base pairings get added to the rest of G_{big} for the pseudoknot free structures, such that the average accuracy after running HFold on the pseudoknot free structures is 92%. Note that HFold does not add any pseudoknots to the pseudoknot free structures. Figure 4.2 shows an example of before and after structures for a pseudoknot free structure.

There are 40 cases presented in Fig. 4.1 marked with 'x' whose accuracy is between 44% and 80%. Of those cases, 34 are the result of the high penalty for introducing a pseudoknot. In all of these 34 cases, the expected output structures need addition of stems of size less than or equal to 7 bases, and the pseudoknot initiation penalty is much higher than the free energy of the stem. That is why HFold does not add these stems and thus the accuracy is not perfect. In 16 of these cases HFold adds extra base pairings to the input structure but in the wrong places and in the rest of the cases it does not add anything. Figure 4.3 shows an example of the before and after structures for one of these cases.

There are 3 cases among the above mentioned 40 cases for which we removed non-canonical base pairs and loops with less than 3 unpaired bases inside; in 2 of those cases, HFold achieves higher accuracy (in one case it achieves perfect accuracy) when presented with the original G_{big} (that is, the G_{big} structure with non-canonical base pairings and small loops in place), while in the 3rd case the accuracy of the HFold prediction does not change when given the original G_{big}

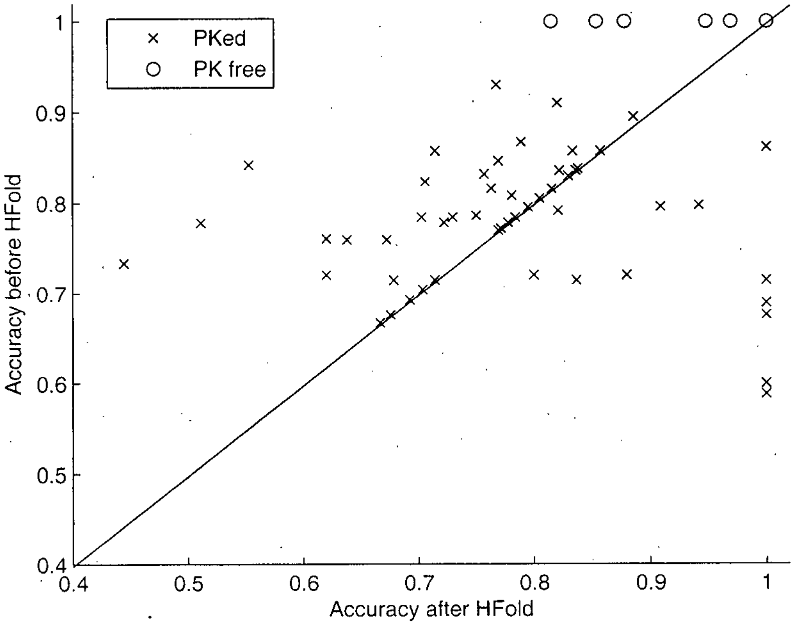


Figure 4.1: Accuracy of HFold when the input is the true pseudoknot free secondary structure (G_{big}), versus the accuracy of G_{big} structures. The horizontal axis represents the accuracy of structures predicted by HFold while the vertical line represents the accuracy of the input structures. Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

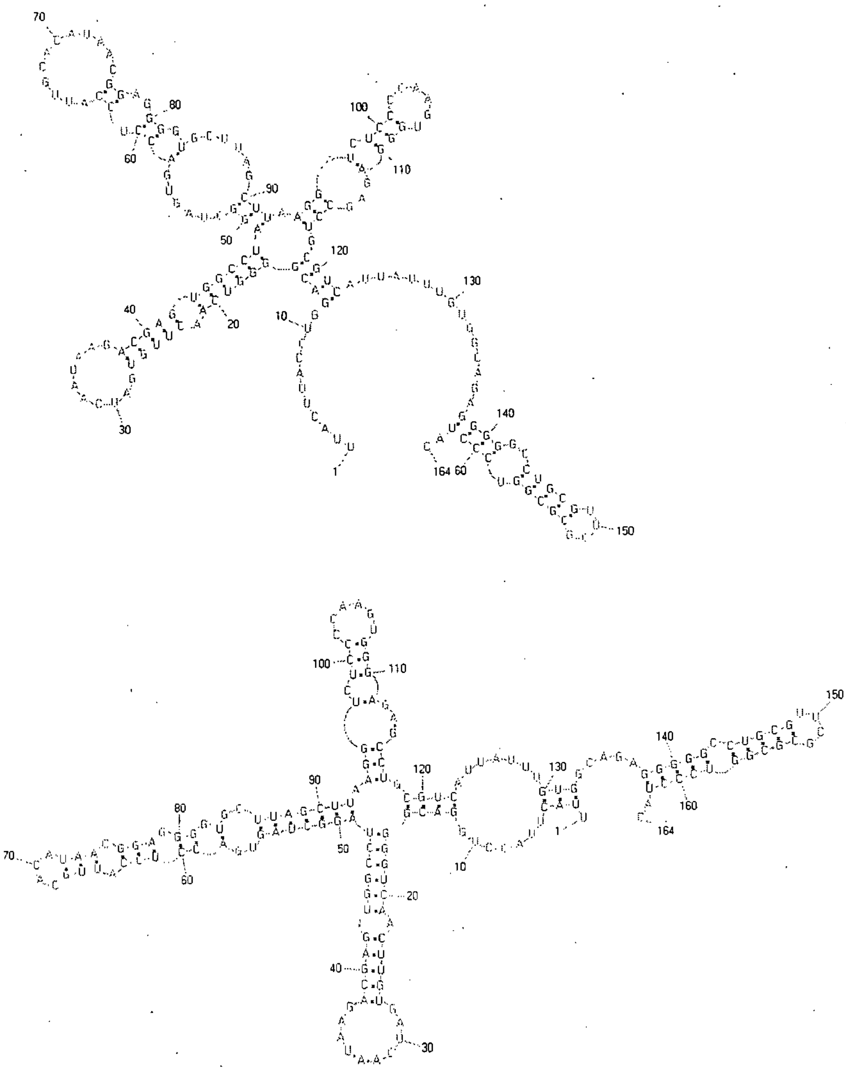


Figure 4.2: A pseudoknot free structure given as input to HFold (top) and the corresponding HFold pseudoknot free result (bottom). The overall structure remains pseudoknot free after HFold, but more base pairs are added to the loops. For example, the large internal loop starting at 1 and ending at 164 on the top figure is broken into two smaller loops in the bottom figure. Similarly the loops of the branch starting at 51 and ending at index 90 are broken into smaller loops.

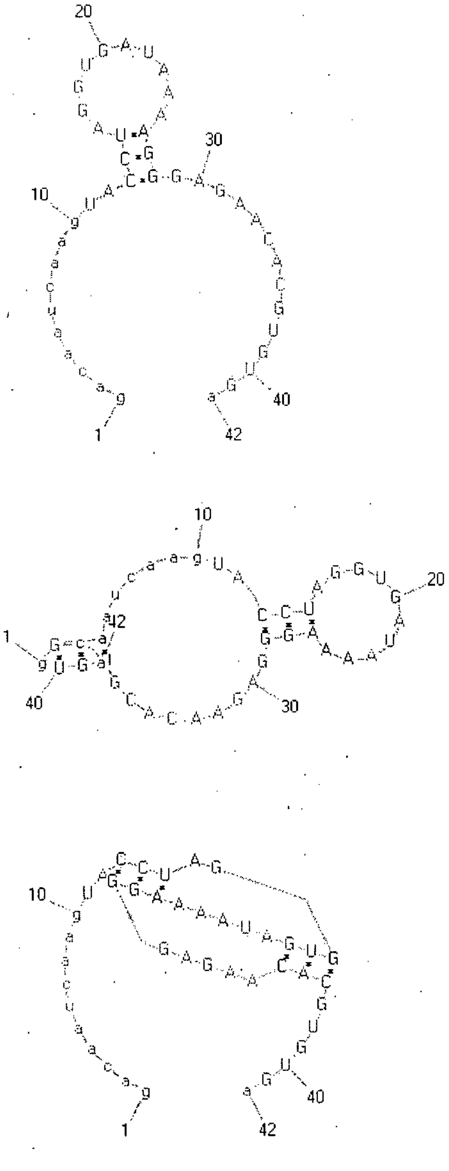


Figure 4.3: G_{big} structure given as input to HFold (top), the corresponding HFold result (center) and the true pseudoknotted structure (bottom). HFold adds 3 pseudoknot free stacked base pairs at 2.40, 3.39, and 4.38, while the true pseudoknotted structure has pseudoknotted base pairs at 18.36, 19.35, and 20.34.

as input.

In the remaining 3 cases of the above mentioned 40 cases, HFold's prediction is pseudoknotted but the pseudoknotted structure either has more crossing base pairs than the original structure or it has pseudoknots in different places. High pseudoknot initiation penalty may be the cause behind this problem except when HFold is adding pseudoknot free base pairs inside or outside the loops (similar to the structure of Fig. 4.2). For example if the true structure includes several H-type pseudoknots, HFold predicts interleaved bands instead. This is because for every separate pseudoknot, the energy value is penalized by one pseudoknot initiation penalty, while for a big interleaved pseudoloop it is only penalized once. In addition, extra bands stabilize the energy value. Figure 4.4 shows an example of before and after structures for one of these cases.

In all the cases where HFold's prediction is different from the true structure, the minimum free energy of HFold's result is lower than the energy of the true structure.

We also ran HFold, when presented with a sequence S and the corresponding true pseudoknot free secondary structure G_{small} as its input. Figure 4.5 plots the accuracy of HFold(S, G_{small}) versus the accuracy of the G_{small} structures. In this case, the average accuracy of the G_{small} structures not including the pseudoknot free structures is 64%. The average accuracy of the HFold output, HFold(S, G_{small}), not including the pseudoknot free structures is 80%, with 19 cases achieving perfect accuracy.

Interestingly, when the better of the two accuracies obtained by using HFold on G_{big} and G_{small} is taken for each sequence, the average accuracy, over all sequences is 88% including the pseudoknot free structures and 87% not including them.

4.4 Accuracy when input is the MFE pseudoknot free structure

More realistically, we need to be able to predict the secondary structure of a sequence without knowing the true pseudoknot free secondary structure for the sequence. A natural approach is to run HFold on the predicted MFE secondary structure for the sequence. We use SimFold [3] to produce the pseudoknot free secondary structure G, which becomes the input to HFold. Figure 4.6 plots the accuracy for each data sequence versus the accuracy of the input structures. The average accuracy is 59%, which is exactly the same as the average accuracy of the MFE pseudoknot free structures.

Figure 4.6 shows 14 points with accuracy between 18% to 31%. In all of these cases, the MFE pseudoknot free structures bear little resemblance to the true structures, and HFold does not add any pseudoknotted base pairs to them. In 4 out of these 14 cases, the low accuracy of the MFE pseudoknot free structures is explained by stems that are "shifts" of stems in the true structure. In all those cases HFold does not add any base pairs to the input structure. Figure

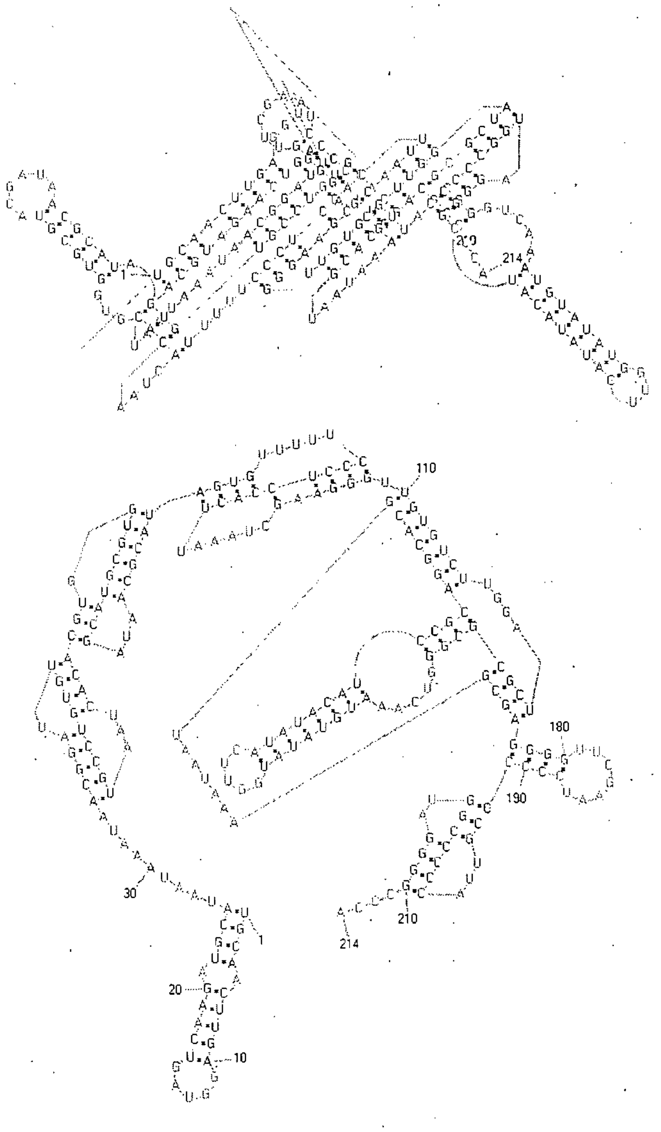


Figure 4.4: Predicted pseudoknotted structure (top) and the true pseudoknotted structure (bottom). HFold predicts interleaved bands, while the true pseudoknotted structure has 5 separate H-type pseudoknots.

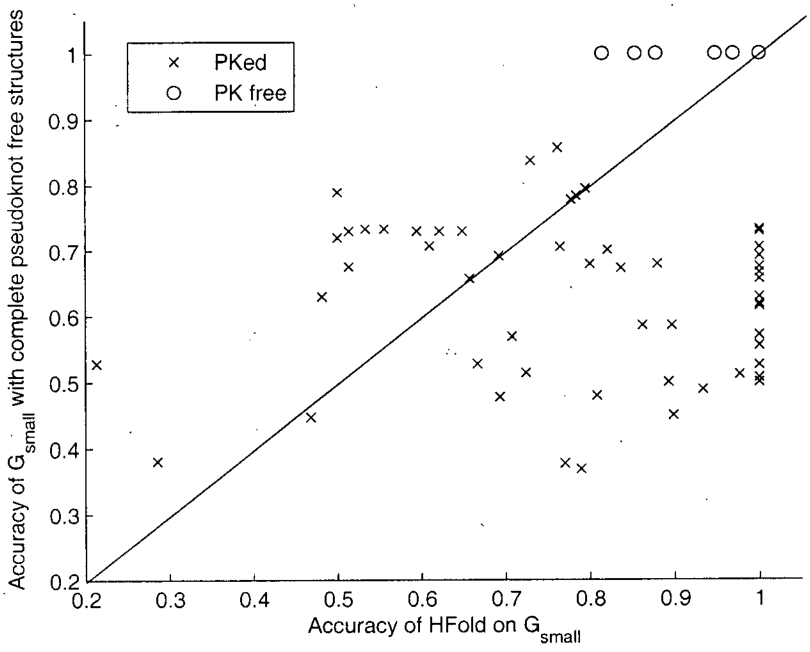


Figure 4.5: Accuracy of HFold when the input is the true pseudoknot free secondary structure (G_{small}) (horizontal axis), versus the accuracy of G_{small} structures (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

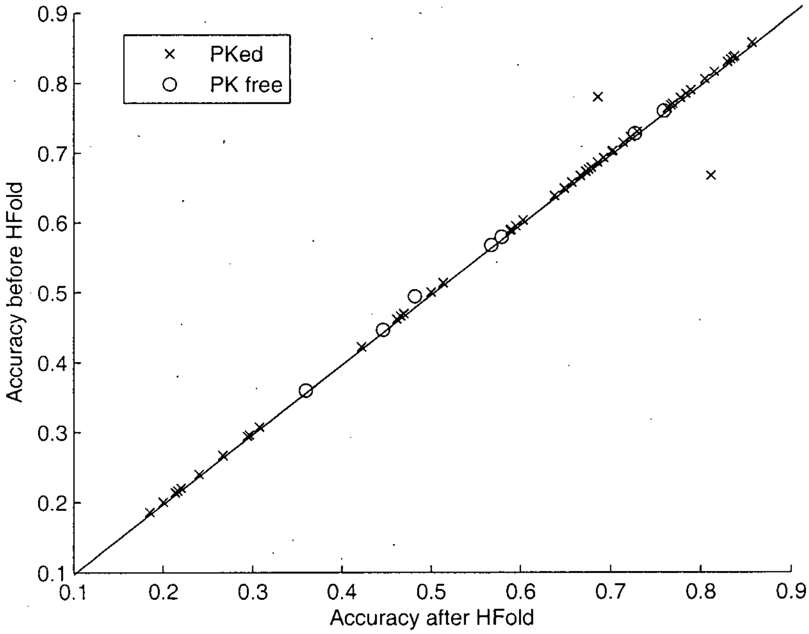


Figure 4.6: Accuracy of HFold when the input is the MFE structure (horizontal axis), versus the accuracy of the MFE structures (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

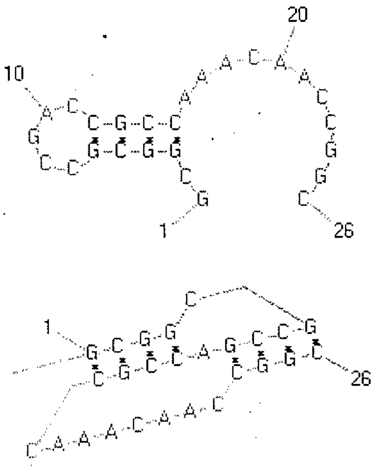


Figure 4.7: MFE structure (top), true pseudoknotted structure (bottom). The true pseudoknotted structure has 4 stacking base pairs at 1.14, 2.13, 3.12, 4.11 and another stem of size 4 at 6.26, 7.25, 8.24, and 9.23, while in the MFE structure the first stem is a shift of the true structure (3.15, 4.14, 5.13, and 6.12). In this cases HFold does not add any base pairs to the input structure.

4.7 shows an example of one of those cases.

There are only two data points presented in Fig. 4.6 that show significant change of accuracy before and after running HFold. In both cases, HFold predicts pseudoknots. In one case the predicted pseudoknotted structure is very similar to the true structure since the MFE structure for the corresponding sequence is very similar to the G_{big} structure of the sequence. In the other case, the true pseudoknotted structure includes one H-type pseudoknot with only 3 pseudoknotted base pairs, but because of the high pseudoknot initiation penalty, HFold predicts one pseudoloop of 3 interleaved bands, thus lowering the accuracy.

4.5 Best accuracy, taken over suboptimal folds

Since HFold does not improve accuracy, on average, when using the MFE pseudoknot free predictions as the input, we also measured the best accuracy obtainable using HFold with suboptimal structures as input.

For each sequence in our dataset, we used SimFold to calculate the 25 lowest-energy structures for that sequence. (We chose the 25 lowest-energy structures because, when we compared the 1000 lowest-energy structures from SimFold with the true pseudoknot free structures for each sequence in our dataset, in

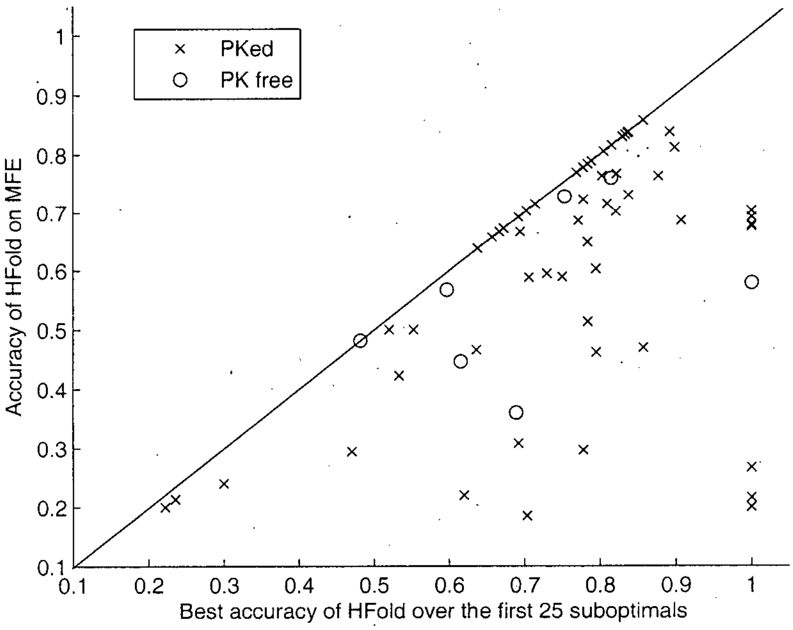


Figure 4.8: Accuracy of HFold on the best-of-25 (horizontal axis), versus the accuracy of the MFE structures (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

only 39 cases was the true structure found among these 1000 structures and in 29 of the 39 cases the true structure was among the 25 lowest-energy structures.) We then ran HFold on all 25 structures, and calculated the accuracy of the HFold output on each run. We did the same with the 10 lowest energy structures. Then we took the best accuracy among the 10 lowest-energy structures, and the best accuracy among the 25 lowest-energy structures.

When averaged over all sequences, the best-of-10 accuracy is 70% with 4 cases with perfect accuracy and the best-of-25 is 74% with 8 cases with perfect accuracy. These accuracies are a significant improvement over the 59% average accuracy obtained using the MFE structures. Figure 4.8 shows the accuracy results of the best-of-25 versus the accuracy of the MFE structures.

As shown in Fig. 4.8, 48 data points are in the lower triangle, showing improvement over the MFE structures. In many cases the improvement is simply due to the fact that the suboptimal structure is close to G_{big}. In 20 of the 48 cases, HFold adds nothing; in 15 out of 48 cases, it adds 1 or 2 pseudoknot

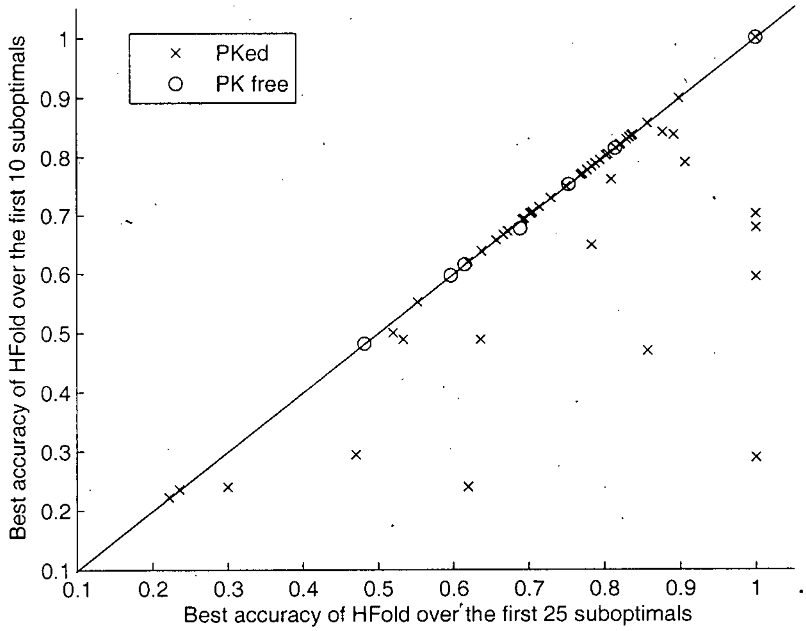


Figure 4.9: Accuracy of HFold on the best-of-25, versus the accuracy of the best-of-10. Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

free base pairs; but in the remaining 13 cases it perfectly identifies G_{small}, from which in 5 cases HFold adds extra base pairings, thus, not achieving 100% accuracy.

Figure 4.9 shows the improvements achieved using best-of-25 versus best-of-10. As shown in Fig. 4.9, there are only 10 data points in the lower triangle that show significant improvement on best-of-25 versus best-of-10, and the rest of the cases are largely similar. Since the difference between the average best accuracies of best-of-25 and best-of-10 is only 4%, but the number of structures is considerably lower and the improvement over accuracy of the MFE structures is significant even for best-of-10, biologists may prefer to obtain 10 lowest energy structures instead of just the MFE structure to get better prediction.

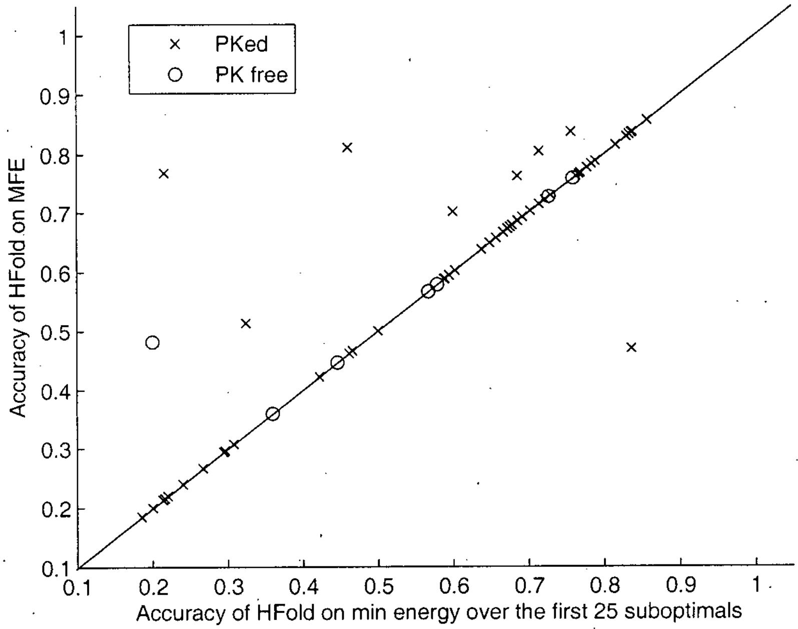


Figure 4.10: Accuracy of HFold on the minimum energy structure over the first 25 suboptimals, versus the accuracy of HFold on the MFE structure. Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

4.6 Accuracy of the lowest energy structure, taken over suboptimal folds

In this section, similar to Section 4.5, for each sequence we get the first 25 suboptimal structures, and then we run HFold on those structures. However, in this case we choose the final structure with the lowest energy and compare its accuracy with the accuracy of HFold on the MFE structure for the same sequence. Figure 4.10 shows the results for this case.

There are 9 data points in Fig. 4.10 not on the diagonal line. One of these data points is in the lower triangle, showing significant improvement over the accuracy of the MFE structure. In this case, the input structure is very similar to the corresponding G_{big} structure but has one extra loop, thus HFold correctly identifies the corresponding G_{small} structure and improves the accuracy of the overall structure.

For 5 of the 8 data points in the upper triangle, HFold correctly predicts the corresponding G_{small} structure but adds more pseudoknotted or pseudoknot free base pairs to lower the energy. An example of one of those cases is presented in Fig. 4.11. There are some stems in some of the input structures that are "shifts" of stems in the true structure; these shifts are also causes of low accuracies. In those cases HFold does not add any base pairs to the structure.

There is not much difference in the "energy" values of the predicted structures with lowest energy over the first 25 suboptimals versus the energy of the predicted structure when input structure is MFE, but some of the structures in the latter case are pseudoknotted, and thus have more similar structure to the true structure than MFE.

4.7 Peeling Techniques

One of the main problems with the MFE structures is that they typically contain long stems and short loops, making it difficult for HFold to add enough pseudoknotted base pairs into the small regions. One possible solution to obtain better input structures is to peel away the stems. This approach is also motivated by the process of hierarchical folding, where it has been shown that, upon formation of a pseudoknot, some base pairs at the ends of stems in the initially formed pseudoknot free structures may break [25]. For our testing purposes, we tried the following cases:

1. removing 1-3 base pairs from inside the stems that close hairpin loops (3 cases);
2. removing 1-3 base pairs from outside the stems that close hairpin loops (3 cases);
3. removing 1-3 base pairs from both inside and outside the stems that close hairpin loops (3 cases);
4. removing stems of size m when $1 \leq m \leq 3$ (3 cases), and
5. peeling enough base pairs from inside hairpin loops so that there are at least 6 unpaired bases inside each loop (1 case);

4.7.1 Accuracy when input is the MFE pseudoknot free structure with 1-3 base pairs removed from inside or outside the stems that close hairpin loops

For our first peeling technique, we removed 1-3 base pairs from inside each stem that closes a hairpin loop of the MFE structures and ran HFold on the resulting structures. Figures 4.12 shows the accuracy of the MFE structures when 2 base pairs are removed from inside each stem that closes a hairpin loop versus the accuracy of the MFE structures themselves, while Fig. 4.13 shows the accuracy of the predicted structure when the input structure is peeled versus the accuracy

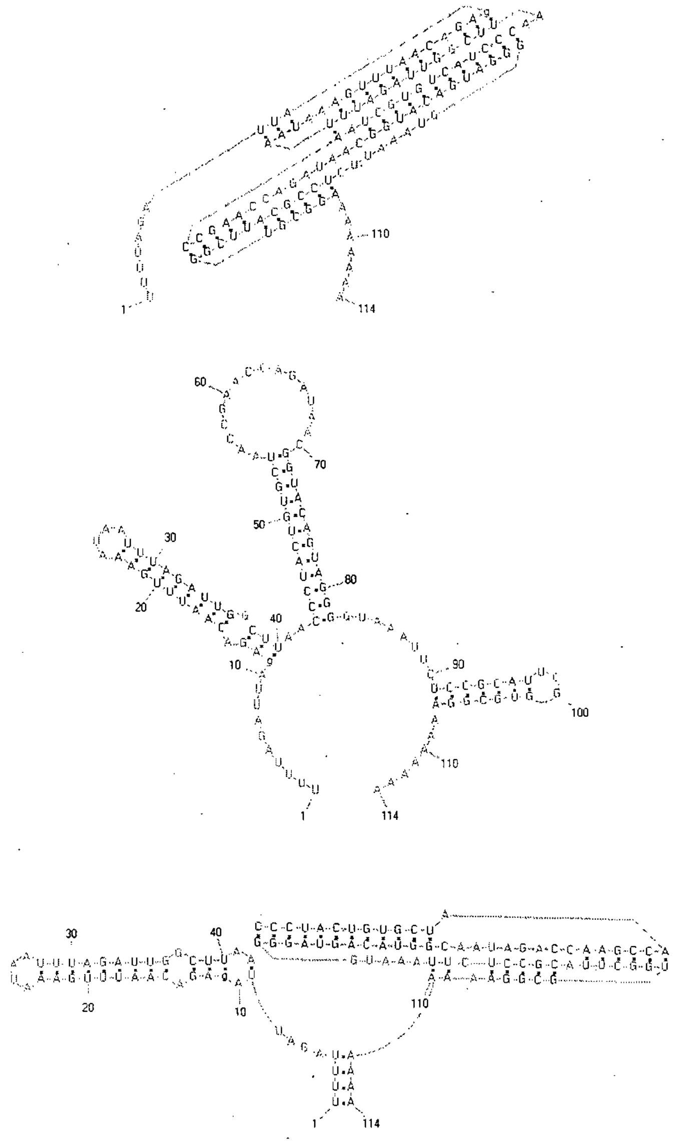


Figure 4.11: The pseudoknotted structure predicted by HFold (top), the corresponding MFE structure (center) and the true structure (bottom). Although HFold correctly predicts parts of the pseudoknot in the structure on the top, the accuracy of the structure predicted by HFold is lower than the accuracy of the corresponding MFE structure, because there are some extra pseudoknotted base pairs predicted by HFold. Note that the MFE structure has all the correct pseudoknot free sub-structures (except the stem closing at 1.114). High pseudoknot initiation penalty, and low band penalty are the main causes for such predictions by HFold.

of the MFE structures. Since the rest of the figures are similar we only present a detailed discussion on this case.

Generally speaking, as shown in Fig. 4.12 when the accuracy of the MFE structures is lower than 55%, peeling increases the accuracy of the structure. This is because the MFE structure in these cases has little resemblance to the true structure. However, as shown in Fig. 4.13, in all but one case, HFold adds the peeled base pairs back and creates the original MFE structure. There is only one data point on the lower triangle showing improvement over the MFE structure. In this case the peeled structure is similar to G_{big}, thus HFold identifies the corresponding G_{small} structure correctly, but since it adds more base pairs to lower the energy of the whole structure the accuracy of the predicted structure is not perfect.

There are some data points shown in Fig. 4.13 whose accuracies do not change after running HFold. This is because the stems in their corresponding MFE structure are shifts of the stems of their G_{big} structure, thus removing a base pair from inside each stem that closes a hairpin loop does not change the accuracy of the whole structure. By removing more base pairs from inside the structures, we obtain more cases where the accuracies before and after running HFold are different. For example, when we remove only 1 base pair from inside each stem that closes a hairpin loop we find 10 data points with the same accuracies before and after HFold, while this number decreases to 6 when we remove 2 base pairs from inside each stem that closes a loop (presented in Fig. 4.13).

Similar results are achieved when 1, 2, or 3 base pairs are removed from outside each stem that closes a hairpin loop of the MFE structures and HFold is run given these structures as input.

4.7.2 Accuracy when input is the MFE pseudoknot free structure with 1-3 base pairs removed from both inside and outside each stem that closes a hairpin loop

We peeled away parts of stems from both ends of each hairpin loop. When we peeled only 1 base pair from each end of the stems, we ended up with 5 cases whose accuracies did not change before and after peeling (see Fig. 4.14). In those cases the MFE structure is very different from the true structure so that peeling one base from both ends of each loop does not change the accuracy. In all those cases HFold adds the peeled base pairs back and thus, the final predicted structure is the MFE structure. As shown in Fig. 4.14, generally speaking when the accuracy of the MFE structures is lower than 55%, peeling increases the accuracy of the structure. This is because the MFE structure in these cases has little resemblance to the true structure. Because of the same reason, HFold adds the peeled base pairs back to the structure to lower the energy of the whole structure (see Fig. 4.15).

There is only one case shown in Fig. 4.15 for which running HFold results

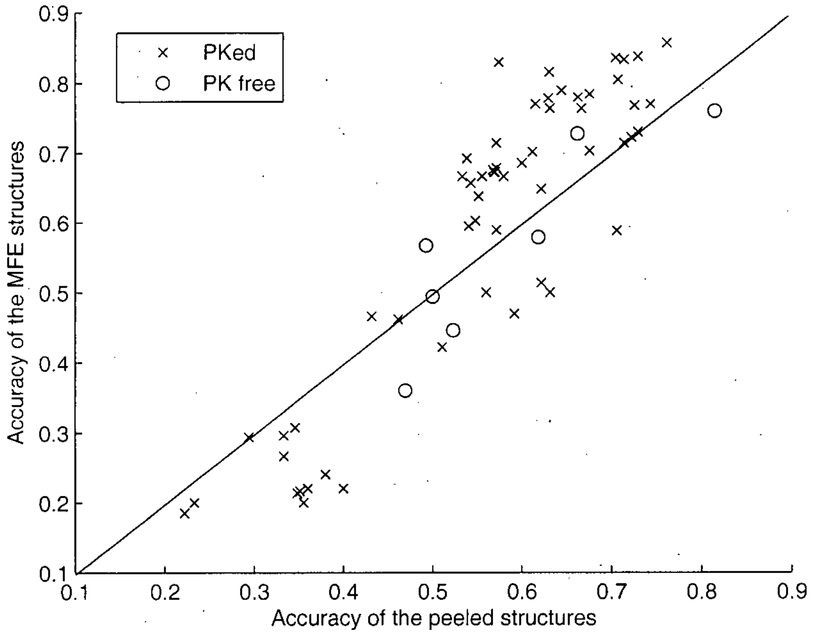


Figure 4.12: Accuracy of the MFE structures with 2 base pairs removed from inside each stem that closes a hairpin loop (horizontal axis), versus the accuracy of the MFE structures (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

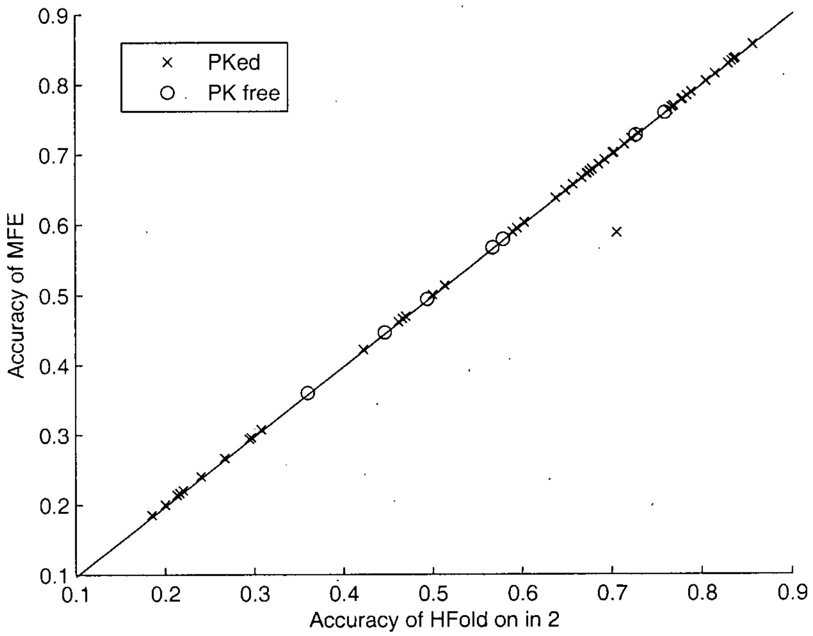


Figure 4.13: Accuracy of HFold when input is the MFE structures with 2 base pairs removed from inside each stem that closes a hairpin loop (horizontal axis), versus the accuracy of HFold when input structure is MFE (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

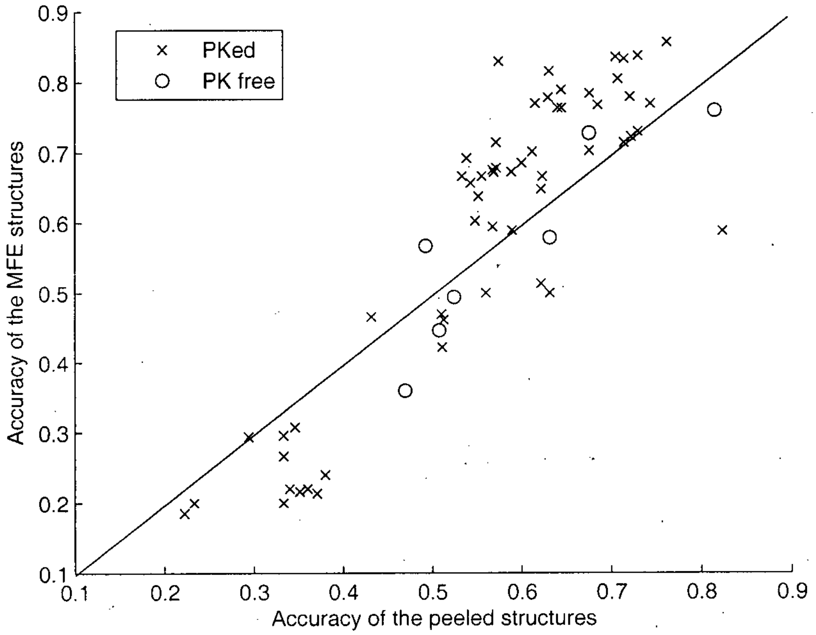


Figure 4.14: Accuracy of the MFE structures with 1 base pair removed from both ends of each stem (horizontal axis), versus the accuracy of the MFE structures (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

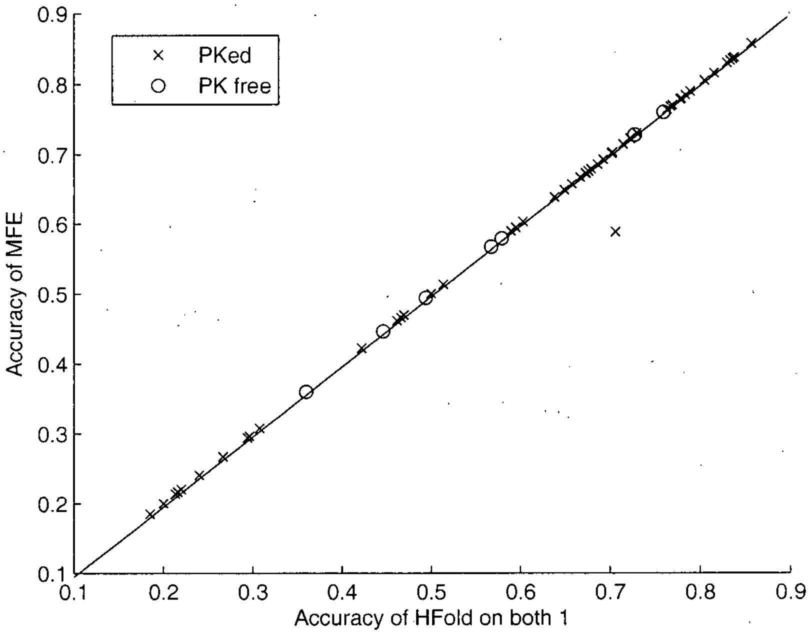


Figure 4.15: Accuracy of HFold when input is the MFE structures with 1 base pair removed from both ends of each stem (horizontal axis), versus the accuracy of HFold when input structure is MFE (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

in improvement in accuracy; in this case the peeled structure is similar to G_{big}, thus, HFold identifies the corresponding G_{small} structure correctly but adds more base pairs to lower the energy of the structure, resulting in a not perfect accuracy.

Removing more than 1 base pair from both ends of the stems results in fewer cases with the same accuracy before and after peeling; that is because the peeled structures have fewer base pairs. But since the MFE structures in general have little resemblance to the true structures, HFold adds the peeled base pairs back and thus, the average accuracy of the predicted structures is the same as the average accuracy of the MFE structures.

4.7.3 Accuracy when input is the MFE pseudoknot free structure with stems of size 1-3 removed

Since the MFE structures typically have long stems and small loops, we decided to remove the stems of size 1-3 and then give those structures as input to HFold. Since none of the MFE structures have any stems of size 1, peeling away stems of size 1 does not change any of the structures and thus, the final predicted structures are the MFE results.

When removing stems of size 2 or 3, there is little change in the structure of some of our data points. However since HFold cannot fix the errors of the input structure, it does not find any structure with lower energy than the MFE structure and thus puts the removed base pairs back. In only 1 case, the input structure is similar to G_{big} and thus HFold finds a pseudoknotted structure which is very similar to the true structure but with some extra base pairs.

4.7.4 Accuracy when input is the MFE pseudoknot free structure with loosened loops

Finally, we decided to loosen up the loops of the MFE structures so that HFold could possibly add extra base pairings to the structure and create pseudoknotted structures. Since we noticed from the HFold results in Section 4.4 that if the true pseudoknotted structure needs less than a stack of size 6 (that is 6 consecutive base pairs) the structure typically will not form, due to the large pseudoknot initiation penalty, we decided to make room for at least 6 bases inside the loops. After removing enough base pairs from inside the loops of the MFE structures, we ran HFold on them and obtained the accuracy of the results.

It is interesting to note that the average accuracy in both cases (before and after running HFold) is almost the same (59% and 58.5% respectively) and is also the same as the average accuracy of the MFE structures.

In only 1 case HFold predicts the pseudoknot as it is in the true structure but it adds some extra base pairings. In this case the MFE structure is very similar to the corresponding G_{big} structure but has overcrowded loops; that is why loosening the loops helps HFold find the true pseudoknotted structure. In the rest of the cases, HFold adds the peeled base pairs back and results in the MFE structure.

4.8 Base Pair Confidence

Structure formation can be viewed as a probabilistic process, with the likelihood of a given secondary structure (for a fixed sequence) being a function of the energy of that structure. The lower the free energy, the more likely the structure; thus, the MFE structure is the most likely to form.

Energy-based methods have also been developed to predict base pairing probabilities of pseudoknot free secondary structures at thermodynamic equilibrium. Mathews [15] found that, roughly speaking, the higher the probability that a base pair is predicted to be in the equilibrium structure, the more likely it is to occur in the true structure. One possible way to improve the accuracy may be to use base pair probability as a confidence measure – that is, base pairs are kept only if their probability is higher than a threshold value.

Based on the results presented in Mathews [15], we chose different thresholds starting from base pair probabilities of 75%, and increased the threshold by steps of 5% to the maximum of 95% threshold. In each case, we chose only the base pairs whose probabilities are higher than the given threshold, and gave these structures as input to HFold. Figure 4.16 shows the accuracy comparison plot for predicted structures when base pair confidence is at least 75% and the MFE structures.

As shown in Fig. 4.16 there are only two points that are not on the diagonal line. Both of these two cases show improvement over accuracy of the MFE structures. It is worth mentioning that in both of these cases the structure predicted by HFold when the input structure is base pairs with confidence above 75% achieves higher accuracy than the structure predicted by HFold when the input structure is the MFE structure. In one of these cases the improvement is 11% and in the other case it is 8%.

The results are similar to Fig. 4.16 when the input structure has base pair confidence of at least 80% and 85%. When the confidence level reaches 90% all the data points except one are on the diagonal line. This one case is one of the two cases we discussed when the threshold was 75%.

It is clear that the higher the threshold, the fewer base pairs are chosen for the input structure. For example, when the threshold is 90%, there are no base pairs chosen in 23 cases out of 70 cases, thus, the input structure to HFold is empty and HFold gives the MFE structure for those cases. For all the different threshold values the average accuracy is 59%, which is exactly the same as the HFold accuracy when the input structures are the MFE structures.

4.9 Energy Comparison

Using a separate method, with the same energy model, we calculated the energy of the "true" pseudoknotted structures, and compared those values with the energies of the structures predicted by HFold when the input was G_{big}. Figure 4.17 shows the result of our comparison. All of the data points, except one case, are either on the diagonal line, or in the upper triangle. As expected,

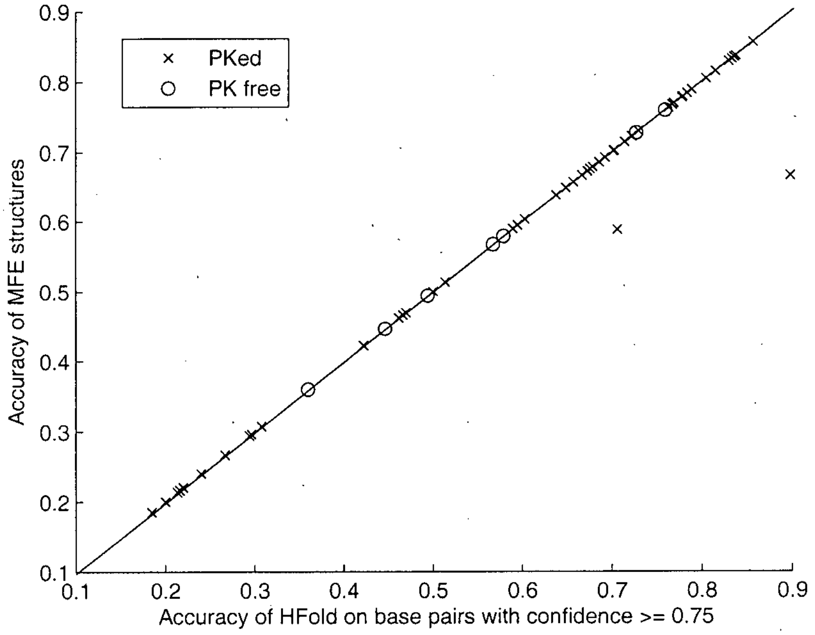


Figure 4.16: Accuracy of HFold when the input is the pseudoknot free structures with base pair confidence above 75% (horizontal axis), versus the accuracy of the MFE structures (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

HFold's results have the same energy as the "true" structure or the "true" structures have higher energy than the HFold's results using our energy model. The only case that HFold results in a structure with higher energy value than the true structure is when the true structure is a density-3 structure which cannot be handled by HFold.

Figure 4.18 shows the comparison of the energy of G_{big} structure for a sequence versus the energy of the MFE structure for the same sequence. In many cases, the energy differences are large, indicating the energy model is poor. Ideally, when the energy model is good, all the data points should be on the diagonal line or very close to it, such that running HFold on the MFE prediction would result in the true structures.

It is interesting to note that for two of the pseudoknot free structures (presented by circles in the figure) the MFE energy value is much lower than the energy of the corresponding G_{big} structures; this can be an indication of the fact that even the pseudoknot free energy model is far from being correct. This is because large loops are broken into small loops to lower the energy in the MFE structures (see Fig. 4.2). This is the main reason for not getting the perfect accuracy for all cases when the true pseudoknot free structures are given to HFold as input structures.

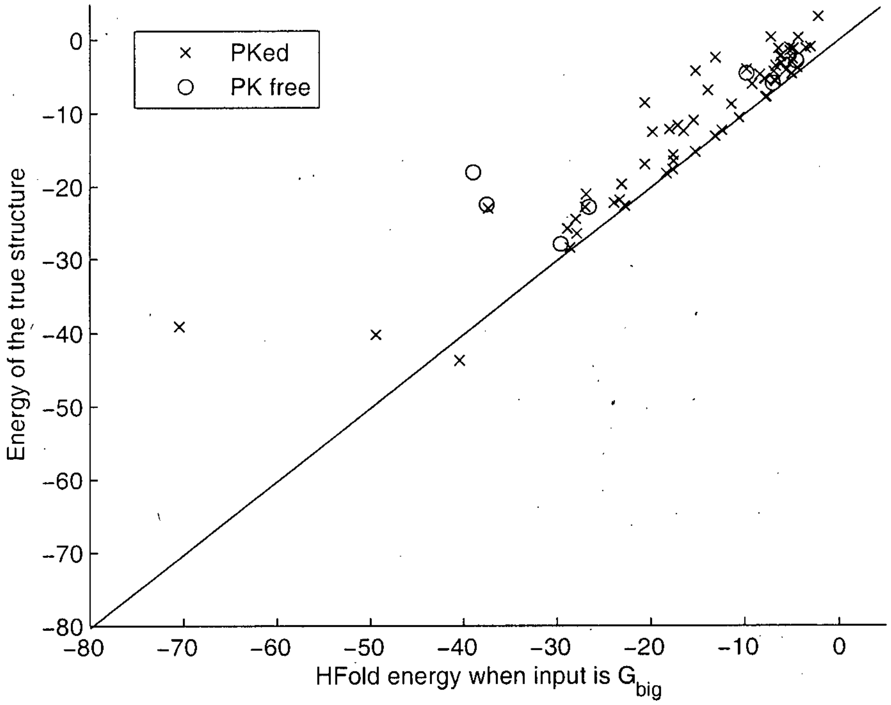


Figure 4.17: Comparison of HFold energy and true structure energy. The horizontal axis represents the energy of structures predicted by HFold when the input is G_{big} while the vertical axis represents the energy of the "true" structures. Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

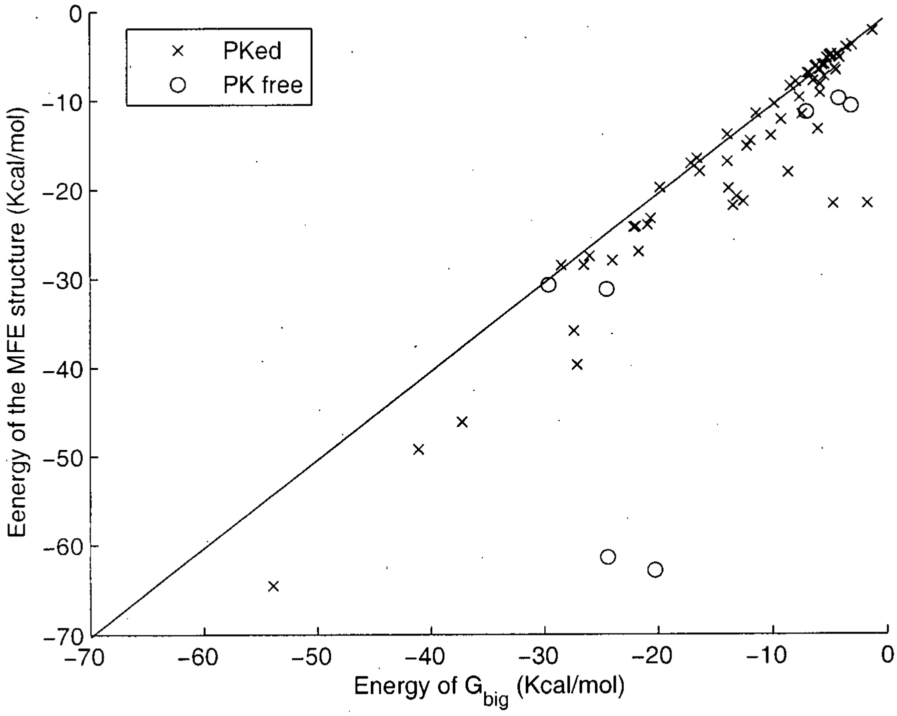


Figure 4.18: Comparison of energy of G_{big} (horizontal axis) and the energy of the MFE structure (vertical axis). Pseudoknotted and pseudoknot free structures are presented by 'x' and 'O', respectively.

Sequence IDs and lengths	Sequence type	Str.	Ref.
BLV(27), BYDV-NY-RPV(27),	viral ribosomal frame shifting	H	[27]
CABYV(27), FIV(35),		H	
PLRV-S(26), PRRSV-LV(59),		H	
SRV1gag/pro(37),		H	
BChV(26), BWYV(26), EIAV(35),		H	
PEMV(28), PLRV-W(26),		H	
MMTVgag/pro(34),		H	
PRRSV-16244B(58),		H	
SARS-CoV(69)		H*	
AKV-MuLV(50), GaLV(49),	viral ribosomal readthrough	H	[27]
Cas-Br-E-MuLv(50), SNV(50)		H	
BaEV(50), Mo-MuLV(50), FeLV(50)		H	
1.1(37), 2.9(34), 2.7a(36),	with high affinity to HIV-1-RT	H	[5]
1.7(37), 2.1b(39), 2.10(37),		H	
1.6(37), 2.12(37), 2.6b(42),		H	
1.3a(37), 2.4a(37), 2.11(37),		H	
1.17(37), 1.8(39), 1.9b(37),		H	
2.5a(41), 2.2b(42), 2.3a(45)		H	
Ec-S15(67), Hs-PrP(45),	mRNA	H	[27]
Bt-PrP(45), T4-gene32(28)		H	
Ec-PK1(30), Bp-PK2(80)	tmRNA	H	[27]
LP-PK1(30), Ec-PK4(52)		H	
Tt-LSU-P3/P7(65),	Ribozymes	H*	[27]
HDV-It_ag(89)		H*	
BVDV_IRES(73), HCV_Ires(56)	viral RNA	H*	[27]
TMV-L(38), CSFV_IRES(76),		H	
turnip yellow mosaic virus(86),		H	
tobacco mosaic virus(214)		H*	

Table 4.2: Sequences with H-type pseudoknots used for secondary structure prediction.
In order, the columns provide (1) sequence ID, as found in the database or paper from which we obtained the sequence, and their length in parenthesis; (2) type of sequence; (3) structure of sequence (H-type pseudoknot = H, H-type pseudoknots with nested structures = H*); and (4) the reference from which the sequence was obtained.

Sequence IDs and lengths	Sequence type	Str.	Ref.
HCV_229E(61)	viral ribosomal	K	[27]
CoxB3(68)	Viral 3' UTR	K	[27]
satRPV(72), Ni_VS(45), HDV-It_g(88)	Ribozymes	K D3	[27]
Coxsackie(114)	Viral and Phage RNA	K	[6]
Hs-SRP-pkn(47)	SRP RNA	K	[27]

Table 4.3: Sequences with kissing interactions and the density-3 sequence used for secondary structure prediction.
In order, the columns provide (1) sequence ID, as found in the database or paper from which we obtained the sequence, and their length in parenthesis; (2) type of sequence; (3) Structure of the sequence (kissing interactions = K, density-3 structure = D3); and (4) the reference from which the sequence was obtained.

Chapter 5

Conclusion and Future Work

In this work, we presented HFold, a fast, new dynamic programming algorithm that efficiently predicts RNA secondary structure including pseudoknots, based on the hierarchical folding hypothesis. HFold can predict kissing hairpins and pseudoloops with arbitrary number of bands. The algorithm had been proposed by Zhao [30]; this thesis focuses on an empirical analysis of its performance.

Our analysis shows that, when presented with the true pseudoknot free structure, G_{big} no significant improvement is obtained over the accuracy of the pseudoknot free structure alone. However when HFold is given G_{small} as input structure, 16% accuracy improvement is obtained on average over the accuracy of the pseudoknot free structure alone. This study also shows that using G_{small} as the input to HFold results in 80% accuracy with 19 cases of perfect accuracy. High pseudoknot initiation penalty and stem shifts are main causes of not achieving perfect accuracy for the rest of the cases.

Based on the analyses presented in this work, using only the MFE structures as input to HFold does not result in high accuracies as the MFE structures are usually overcrowded with long stems and small loops and thus, running HFold does not result in the addition of any extra base pairings. We showed here that since the MFE structures usually bear little resemblance to the "true" structures, peeling the stems away from inside and/or outside the loops does not improve the average accuracy. We took the peeling one step further and showed that neither removing stems of size m when $1 \leq m \leq 3$, nor loosening the loops to have at least 6 unpaired bases inside, changes the average accuracy of the results. Since the energy model using which the MFE structures are being formed does not consider pseudoknots, even using base pair probabilities is of no help in improving the accuracy of the results.

We also showed that using the best of 10 suboptimal structures improves the accuracy of prediction by 11%. The accuracy increases by another 4% if the first 25 suboptimals are used. In many cases the accuracy of the MFE structure is greater than or equal to the accuracy of the lowest energy structure predicted by HFold on the first 25 suboptimal structures. However, in numerous cases of the latter structures, HFold identifies the correct pseudoknotted base pairs. Note that adding more base pairs to lower the energy causes the accuracy of HFold result to be lower than the corresponding MFE structure.

This study showed that since pseudoknot free energy model predicts too many base pairs in the structure, peeling some base pairs away from the ends of stems that close hairpin loops in many cases improved the accuracy of the structure. However high pseudoknot initiation penalty in the pseudoknotted energy model made HFold avoid predicting pseudoknots by adding pseudoknot free base pairs when possible. We also found that because of the high pseudoknot initiation penalty and low band penalty, more bands are added to pseudoknotted structures to compensate for the energy.

The results of this study provide more insight to the hierarchical folding hypothesis. Our results suggest that with respect to the current energy model minor rearrangements of the pseudoknot free MFE structures do not often lead to the true pseudoknotted structures. We also showed that in some cases the pseudoknot free MFE structure contains shifted stems of the true structure; therefore without shifting the stems it is not possible for the MFE structure to fold into the true pseudoknotted structure.

Throughout this work, we used an index by index comparison of the base pairs to find the accuracy of prediction. However, in many examples the overall structure of the predicted structure was similar to the true structure, but more base pairs or shifts lowered the accuracy. We believe that using other notions of accuracy that consider structural similarities are more accurate that using an index by index comparison. Thorough analyses of the data based on different commonly used accuracy measures such as sensitivity, specificity, or positive predictive is underway.

Our findings raise three main questions:

1. whether the structures identified by the experimental methods are reliable,
2. whether the current energy model is precise enough, and
3. whether the hierarchical folding hypothesis is precise enough.

One of our future goals is to tune the parameters of the current energy model to improve the accuracy of the prediction for a wider sets of structures. One possible approach is using Andronescu's tuning method [4]. Another future work is to use a better energy model for pseudoknotted structures, such as that of Cao and Chen [5], 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 [17].

Bibliography

[1] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Disc. App. Math.*, 104(1-3):45–62, Aug 2000.

[2] S. L. Alam, J. F. Atkins, and R. F. Gesteland. Programmed ribosomal frameshifting: Much ado about knotting! *PNAS*, 96(25):14177–14179, Dec 1999.

[3] M. Andronescu, R. Aguirre-Hernàndez, A. Condon, and H. H. Hoos. RNA-soft: A suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res*, 31(13):3416–3422, Jul 2003.

[4] M. Andronescu, A. Condon, H.H. Hoos, D.H. Mathews, and K.P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *BMC Bioinformatics*, page to appear, July 2007.

[5] S. Cao and S. Chen. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res*, 34(9):2634–2652, 2006.

[6] B. A. L. M. Deiman and C. W. A. Pleij. Pseudoknots: A vital feature in viral RNA. *Seminars in Virol.*, 8(3):166–175, 1997.

[7] C. Dennis. The brave new world of RNA. *Nature*, 418(6894):122–124, Jul 2002.

[8] R. M. Dirks and N. A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24(13):1664–1677, Oct 2003.

[9] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33(Database issue), Jan 2005.

[10] K. Han, Y. Lee, and W. Kim. Pseudoviewer: automatic visualization of RNA pseudoknots. *Bioinformatics*, 18 Suppl 1, 2002.

[11] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):167–188, Feb 1994.

[12] H. Jabbari, A. Condon, A. Pop, C. Pop, and Y. Zhao. HFold: RNA pseudoknotted secondary structure prediction using hierarchical folding. In Raffaele Giancarlo and JSridhar Hannenhalli, editors, *LNBI*, Lecture Notes in Computer Science, page to appear. Springer, 2007.

[13] R. B. Lyngsø. Complexity of pseudoknot prediction in simple models. In Josep Díaz, Juhani Karhumäki, Arto Lepistö, and Donald Sannella, editors, *ICALP*, volume 3142 of *Lecture Notes in Computer Science*, pages 919–931. Springer, 2004.

[14] R. B. Lyngsø and C. N. Pedersen. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.*, 7(3-4):409–427, 2000.

[15] D. H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, Aug 2004.

[16] D. H. Mathews. Predicting RNA secondary structure by free energy minimization. *Theor. Chem. Acc.: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, pages 1–9, May 2006.

[17] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A*, 101(19):7287–7292, May 2004.

[18] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure,. *J. of Mol. Biol.*, 288(5):911–940, May 1999.

[19] W.J. Melchers, J.G. Hoenderop, H.J. Bruins Slot, C.W. Pleij, E.V. Pilipenko, V.I. Agol, and J.M. Galama. 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(1):686–696, Jan 1997.

[20] B. Rastegari and A. Condon. Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. *J. Comput. Biol.*, 14(1):16–32, 2007.

[21] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5, Aug 2004.

[22] E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots, Jul 1998.

[23] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*, 26(1):148–153, Jan 1998.

[24] D. W. Staple and S. E. Butcher. Pseudoknots: RNA structures with diverse functions. *PLoS Biol.*, 3(6), Jun 2005.

[25] I. Tinoco and C. Bustamante. How RNA folds. *J. Mol. Biol.*, 293(2):271–281, Oct 1999.

[26] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori. Tree adjoining grammars for RNA structure prediction. *Theor. Comput. Sci.*, 210(2):277–303, Jan 1999.

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

[28] C. Witwer, Ivo L. Hofacker, and P.F. Stadler. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(2):66–77, Apr 2004.

[29] M. Wu and I. Tinoco. RNA folding causes secondary structure rearrangement. *Proc. Natl. Acad. Sci. USA*, 95(20):11555–11560, Sep 1998.

[30] Y. S. Zhao. Efficient algorithm for RNA pseudoknotted secondary structure prediction given a pseudoknot free structure. Master's thesis, University of British Columbia, Department of Computer Science, September 2005.

[31] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, Jan 1981.

Appendix A

Recurrences

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

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

$R_{i,j}$ refers to the subset of R whose bases are in the region $[i, j]$. Let $G_{i,j} = G \cap \{i, j\} \times \{i, j\}$ i.e. the set of base pairs of G contained in region $[i, j]$. Let $R_{i,j}$ be a minimum free energy (MFE) secondary structure for $[i, j]$, given a pseudoknot free secondary structure $G_{i,j}$.

The energy value of each substructure type, for a given input sequence $S = s_1 s_2 ... s_n$ 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:

- G is presented by the portion of the figure above the base line.
- G' is presented by the portion of the figure under the base line.
- 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 gray shade within the arcs are unknown structures.

A.1 $W_{i,j}$

$W_{i,j}$ is the MFE of all valid 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 $+\infty$.

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

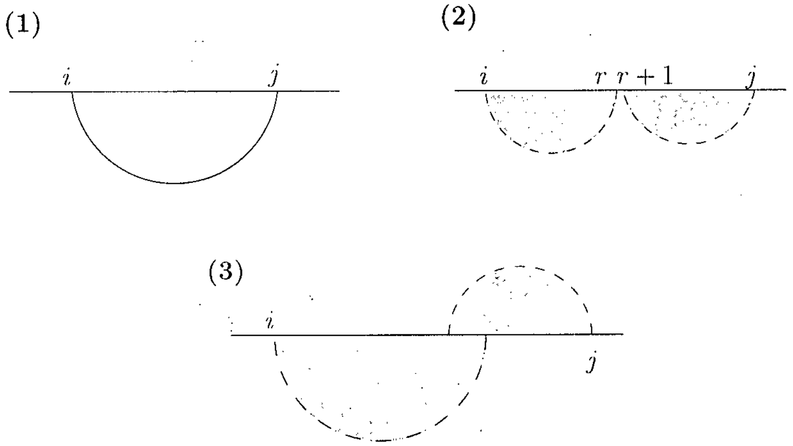


Figure A.1: Illustration of Cases for $W_{i,j}$.

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

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

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 $\exists r$, $i \leq r < j$, $bp_{R_{i,j}}(i) \leq r$ (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.

A.2 $WI_{i,j}$

$WI_{i,j}$ is the minimum free energy of all valid structures $R_{i,j}$, given that $[i, j]$ is weakly closed, and $R_{i,j}$ is inside a pseudoknot. Otherwise, $WI_{i,j}$ is $+\infty$.

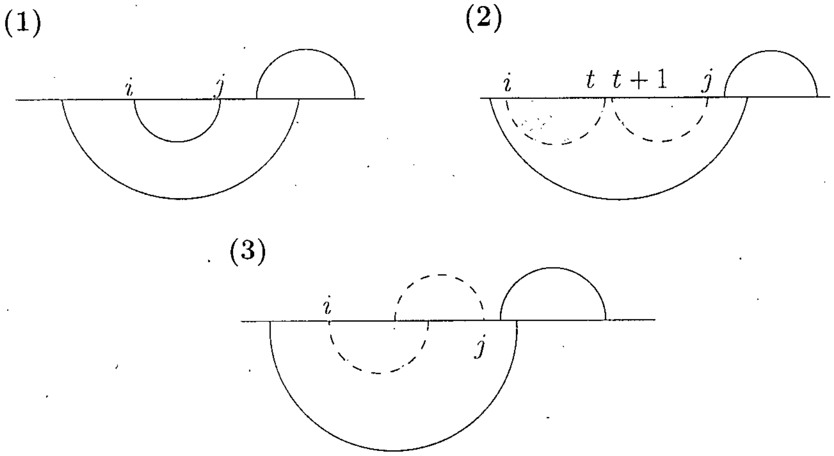


Figure A.2: Illustration of Cases for $WI_{i,j}$.

The base cases are as follows:

$WI_{i,j} = +\infty$, if $cover(i) \neq cover(j)$, 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:

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

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 are paired in G'.

Case (2) handles the cases that $\exists t$, $i \leq t < j$, $bp_{R_{i,j}}(i) \leq t$, 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 the structure inside a pseudoknot, but not inside a band, we add a P_{ps} penalty to case (1) and (3), and a P_{sm} penalty to case (3) for introducing a new pseudoknot inside a pseudoloop.

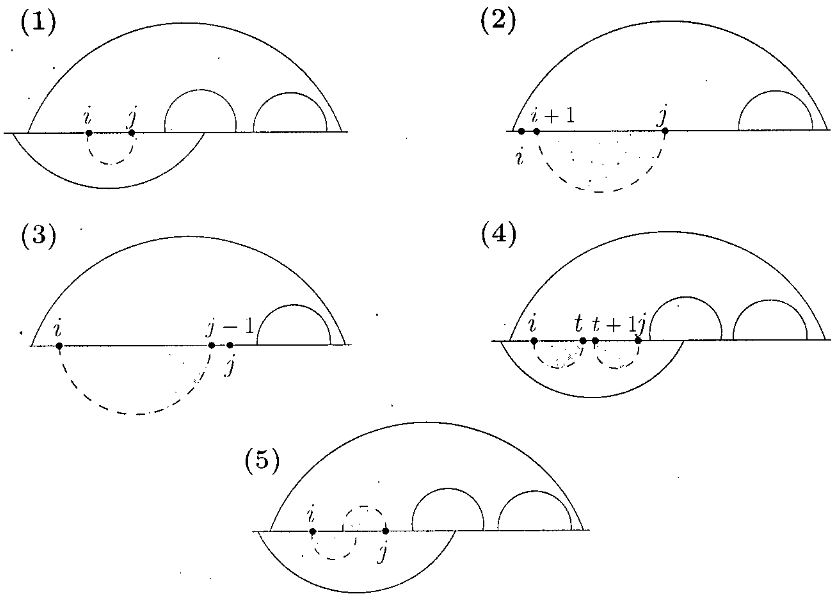


Figure A.3: Illustration of cases for $WI'_{i,j}$. In case (5), the plotted structure from i to j could contain more than 2 bands (not illustrated)

A.3 WI'_{ij}

$WI'_{i,j}$ is the minimum free energy of all valid nonempty structures $R_{i,j}$, if $[i,j]$ is weakly closed, given that $R_{i,j}$ is inside a band. Otherwise, $WI'_{i,j}$ is $+\infty$.

The base cases are as follows:
$WI'_{i,j} = +\infty$, if $[i,j]$ is not weakly closed.
$WI'_{i,j} = +\infty$, if $i \geq j$, since $\text{empty}(R_{i,j}, [i,j])$.

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

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

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 a new base pair in the structure instead of P_{ps}, since the base pair is not inside a pseudoloop, but rather in a band.

A.4 $VP_{i,j}$

$VP_{i,j}$ is the minimum free energy of all valid 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. Otherwise, $VP_{i,j}$ is $+\infty$.

The base case is as follows:

$$VP_{i,j} = +\infty \text{ , if} \begin{cases} i \geq j, \\ i.j \text{ does not cross any base pair of } G, \\ bp_G(i) > 0, \text{ or } bp_G(j) > 0 \end{cases} \tag{A.4}$$

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

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

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

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.

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 $R_{i,j}$.

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 close a multiloop that spans a band.

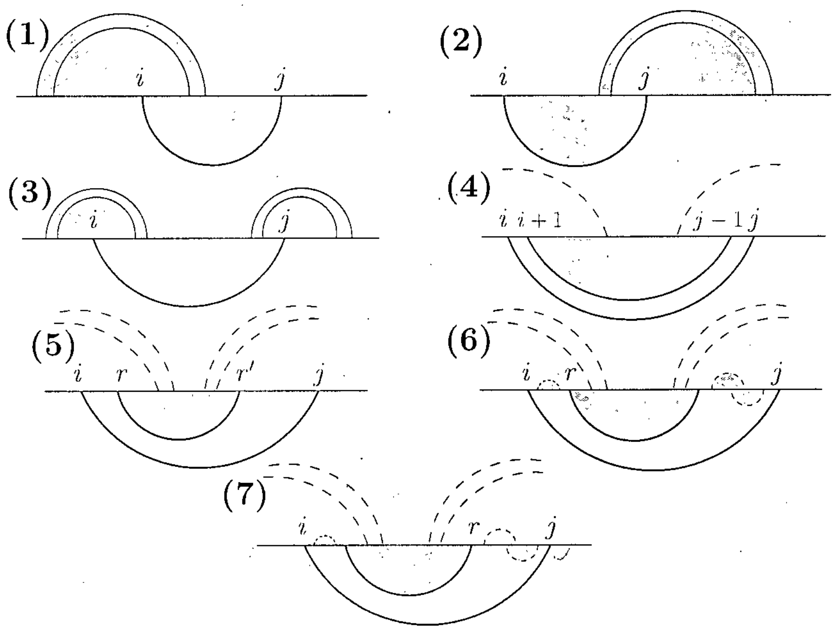


Figure A.4: Illustration of Cases for $VP_{i,j}$.

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

$$\min_{\substack{i<r<min(B'_{(i,j)},b_{(i,j)'}) \\ \leq max(b'_{(i,j)},B_{(i,j)})<r'<j}} (WI'_{(i+1),(r-1)} + VP_{r,r'} + WI'_{(r'+1),(j-1)} + P_{ps} + a' + 2b') \tag{A.6}$$

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.

A.5 $VP'_{i,j}$

$VP'_{i,j}$ is the minimum free energy of all valid 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 $+\infty$.

The base case is as follows:

$$VP'_{i,j} = +\infty, \text{ if } i \geq j \tag{A.7}$$

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

$VP'_{i,j} = \min$

$$\begin{cases} (1) \min\limits_{max(i,b'_{(i,j)})<r<j} (VP_{i,r} + WI'_{(r+1),j}) & \\ (2) \min\limits_{i<r<min(B'_{(i,j)},j)} (WI'_{i,(r-1)} + VP_{r,j}) & \\ (3) \min\limits_{max(i,b'_{(i,j)})<r<j} (VP_{i,r} + c'(j-r)) & \text{if empty}(G, [r+1, j]) \\ (4) \min\limits_{i<r<min(B'_{(i,j)},j)} (c'(r-i) + VP_{r,j}) & \text{if empty}(G, [i, r-1]) \end{cases} \tag{A.8}$$

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.

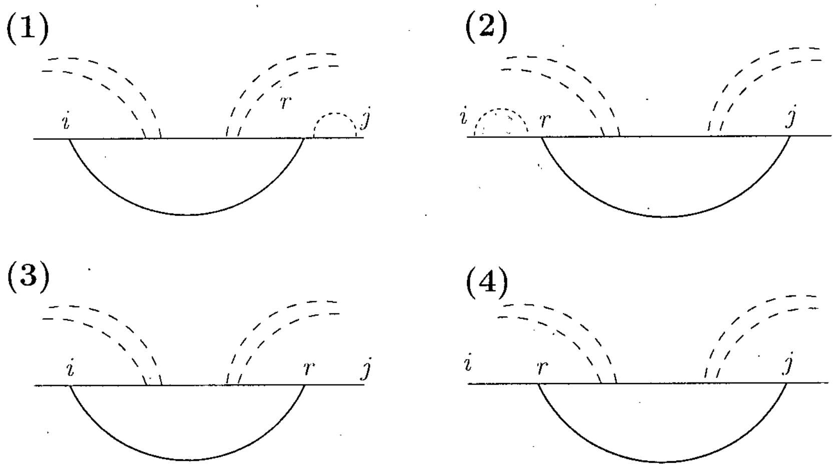


Figure A.5: Illustration of Cases for $VP'_{i,j}$.

A.6 $V_{i,j}$

$V_{i,j}$ is the minimum free energy of all valid 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 $+\infty$.

The base bases are as follows:

$$V_{i,j} = +\infty, \text{ if } \begin{cases} i \geq j \\ [i,j] \text{ is not weakly closed} \\ bp_{R_{i,j}}(i) > 0 \text{ and } bp_{R_{i,j}}(j) > 0 \text{ but } bp_{R_{i,j}}(i) \neq bp_{R_{i,j}}(j) \end{cases} \tag{A.9}$$

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

$$V_{i,j} = \min\{e_H(i,j), e_S(i,j) + V_{(i+1),(j-1)}, VBI_{i,j}, VM_{i,j}\} \tag{A.10}$$

This recurrence is identical to that used in pseudoknot free algorithms, so we omit the proof here and for *VBI* in the next section.

A.7 $VBI_{i,j}$

$VBI_{i,j}$ is the minimum free energy of all valid 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 $+\infty$.

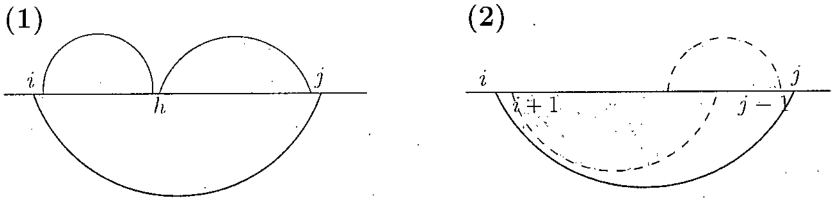


Figure A.6: Illustration of Cases for VM_{ij}.

The base case is as follows: $VBI_{i,j} = +\infty$, if $j - i \leq 1$ Otherwise,

$$VBI_{ij} = \min_{i<i'<j'<j} (e_{int}(i,i',j',j) + V_{i'j'}), \tag{A.11}$$

where all bases between $[i,i']$ and $[j',j]$ are unpaired.

$e_{int}(i,i',j',j)$ gives the free energy of an internal loop or bulge with exterior pair $i.j$ and interior pair $i'.j'$.
The complexity for computing VBI_{ij} can be done in $O(n^3)$ using [31].

A.8 VM_{ij}

VM_{ij} is the minimum free energy of all valid 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 $+\infty$.

We can obtain a recurrence that calculates the loop cost as the sum of two subparts.

$$VM_{ij} = \min \begin{cases} (1) \min\limits_{i+1<h\leq j-1} (WM_{(i+1)(h-1)} + WM_{h(j-1)} + a + b) \\ (2)\, WMB_{(i+1)(j-1)} + a + P_{sm} + b \end{cases} \tag{A.12}$$

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

A.9 WM_{ij}

WM_{ij} is the minimum free energy of all valid 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:

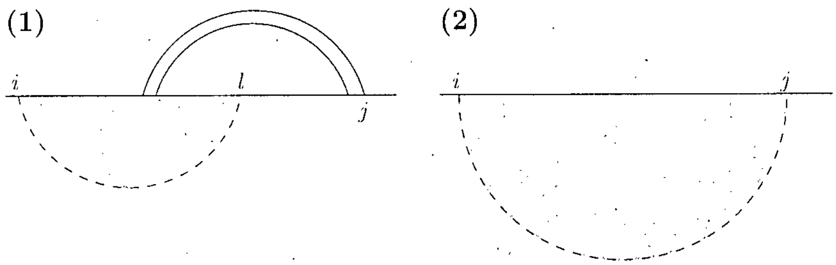


Figure A.7: Illustration of Cases for $WMB_{i,j}$.

$$WM_{i,j} \quad = \quad \min \begin{cases} (1)\, V_{i,j} + b, & \\ (2)\, WM_{(i+1),j} + c & bp_G(i) = 0 \\ (3)\, WM_{i,(j-1)} + c & bp_G(j) = 0 \\ (4)\, \min\limits_{1 \le t \le j} (\, WM_{i,t} + WM_{(t+1),j}) & \\ (5)\, WMB_{i,j} + P_{sm} + b & \end{cases} \tag{A.13}$$

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.

A.10 $WMB_{i,j}$

$WMB_{i,j}$ is the minimum free energy of all valid structures $R_{i,j}$ over region $[i,j]$, where $R_{i,j}$ is a density-2 pseudoloop. Otherwise, WMB is $+\infty$.

The base case is as follows:
$WMB_{i,j} = +\infty$, if $i \geq j$

Otherwise,
$WMB_{i,j} = \min$

$$\begin{cases} (1) P_b + \min\limits_{bp_G(j) < l < j} (BE(bp_G(j), bp_G(B'_{(l,j)}), B'_{(l,j)}, j) & \text{if } bp_G(j) > 0 \\ \qquad\qquad + WMB'_{i,l} + WI_{(l+1),(B'_{(l,j)}-1)}), & \\ (2)\, WMB'_{i,j} & \end{cases} \tag{A.14}$$

In case (1), j is paired in G. Then, in the MFE structure, some base l with $bp_G(j) < l < j$ must be paired in G', causing $bp_{R_{i,j}}(j).j$ 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 $bp_{R_{i,j}}(j).j$ 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 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.

A.11 $WMB'_{i,j}$

$WMB'_{i,j}$ is the minimum free energy of all valid structures $R_{i,j}$ over region $[i,j]$, where $R_{i,j}$ is a proper prefix of a density-2 pseudoloop. Otherwise, WMB' is $+\infty$.

The base case is as follows:
$WMB'_{i,j} = +\infty$, if $i \geq j$

Otherwise,
$WMB'_{i,j} = \min$

$$\begin{cases} (1) 2P_b + \min\limits_{\substack{i<l<min(j,b_{(i,j)}) \\ isCovered(G_{i,j},l)}} & \\ (BE(bp_G(B_{(l,j)}), bp_G(B'_{(l,j)}), B'_{(l,j)}, B_{(l,j)}) & \text{if } bp_G(j) = 0 \\ + WMB'_{i,(l-1)} + VP_{l,j}), & \\ (2) \quad \min\limits_{\substack{i<l<j \\ cover(l)=cover(j) \\ bp_G(l)=0}} \quad (WMB'_{i,l} + WI_{(l+1),j}), & \text{if } bp_G(j) < j \\ (3) P_b + VP_{i,j} & \\ (4) 2P_b + \min\limits_{i<l<bp_G(i)} (BE(i, b'_{(i,l)}, bp_G(b'_{(i,l)}), bp_G(i)) & \text{if } bp_G(j) = 0 \\ + WI_{(b'_{(i,l)}+1),(l-1)} + VP_{l,j}) & \text{and } bp_G(i) > 0 \end{cases} \tag{A.15}$$

Complementing case (1) of the WMB recurrence, WMB' handles the case that the rightmost band is not in G, but is part of the MFE structure G' (with respect to G). 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'.

Case (2) is called when one iteration of $WMB_{i,j}$ case (2) or $WMB'_{i,j}$ case (1) is done and we need to compute the energy of the remaining part of the structure inside an arc. Note that $WI_{i,j} = +\infty$ when $cover(i) = cover(j) = (-1,-1)$ such that entering case (2) as the first iteration of WMB' is not possible. For example, the total energy for the structure presented in Fig. A.11(2) is calculated in the following way: the energy of the rightmost band covering base j is calculated using WMB case (1); the energy of the remaining structure from i to j accounts for the energy of the closed subregion of $[i,j]$, $[l+1,j]$, and the remaining prefix ($WMB'_{i,l}$), which is handled by case (2) of the WMB' recurrence.

In cases (3) and (4), only one iteration of WMB' yields the result. Case (3) handles the case that i pairs with j in G', and $i.j$ is the left-most base pair in

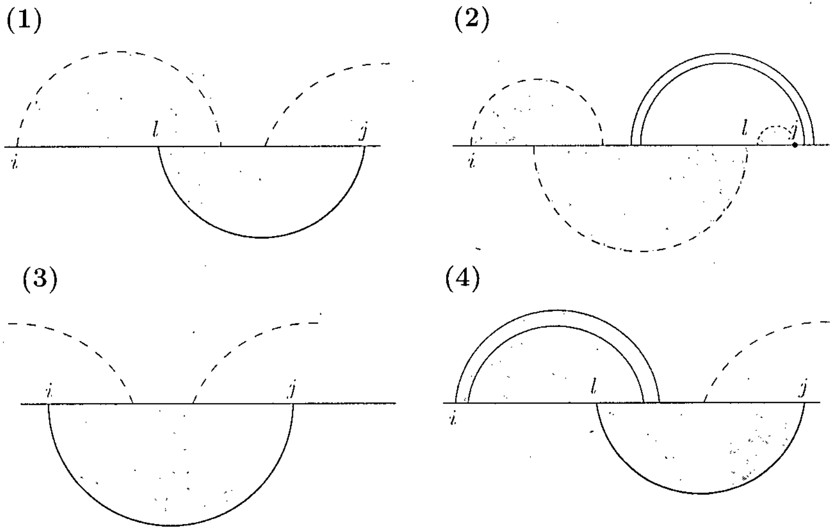


Figure A.8: Illustration of Cases for $WMB'_{i,j}$.

the structure. Since this is part of a pseudoknotted structure, it will be covered by VP.

Case (4) handles the case that there are only two bands in $R_{i,j}$, and the right band of the pseudoknotted loop is not in G.

A.12 $BE(i, i', j', j)$

$BE(i, i', j', j)$ is the minimum free energy of the band $[i, i'] \cup [j', j]$, if $i \leq i' < j' \leq j$, in which $bp_G(i) = j$ and $bp_G(i') = j'$.

The base cases are as follows: $BE(i, i', j', j) = +\infty$, if it is not the case that $i \leq i' < j' \leq j$

$BE(i, i, j, j) = 0$, if $i < j$

Otherwise, $BE(i, i'j', j) = \min$

$$\begin{cases}
(1) e_{stP}(i,j) & \\
+BE(i+1, i', j', j-1) & \text{if } bp_G(i+1) = (j-1) \\
(2) e_{intP}(i, l, bp_G(l), j) & \text{if } bp_G(l) > 0, \\
+BE(l, i', j', bp_G(l)) & (\text{empty}(G, [i+1, l-1]), \\
 & \text{and} \\
 & \text{empty}(G, [bp_G(l)+1, j-1])), \\
 & (i < l \leq i') \\
 & \text{and} \\
 & (j' \leq bp_G(l) < j) \\
(3)\, WI'_{(i+1),(l-1)} & \text{if } bp_G(l) > 0 \\
+BE(l, i', j', bp_G(l)) & (\text{weakly closed}(G, [i+1, l-1]) \\
+\, WI'_{(bp_G(l)+1),(j-1)} + a' + 2b' & \text{and} \\
 & \text{weakly closed}(G, [bp_G(l)+1, j-1])) \\
 & \text{and} \\
 & (i < l \leq i') \\
 & \text{and} \\
 & (j' \leq bp_G(l) < j) \\
(4)\, WI'_{(i+1)(l-1)} & \text{if } bp_G(l) > 0 \\
+BE(l, i', j', bp_G(l)) & (\text{weakly closed}(G, [i+1, l-1]) \\
+a' + 2b' + c'(j - bp(G, l) + 1) & \text{and} \\
 & \text{empty}(G, [bp_G(l)+1, j-1])) \\
 & \text{and} \\
 & (i < l \leq i') \\
 & \text{and} \\
 & (j' \leq bp_G(l) < j) \\
(5) a' + 2b' + c'(l - i + 1) & \text{if } bp_G(l) > 0 \\
+BE(l, i', j', bp_G(l)) & (\text{weakly closed}(G, [bp_G(l)+1, j-1]), \\
+\, WI'_{(bp_G(l)+1),(j-1)} & \text{and} \\
 & \text{empty}(G, [i+1, l-1])) \\
 & \text{and} \\
 & (i < l \leq i') \\
 & \text{and} \\
 & (j' \leq bp_G(l) < j)
\end{cases} \tag{A.16}$$

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

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

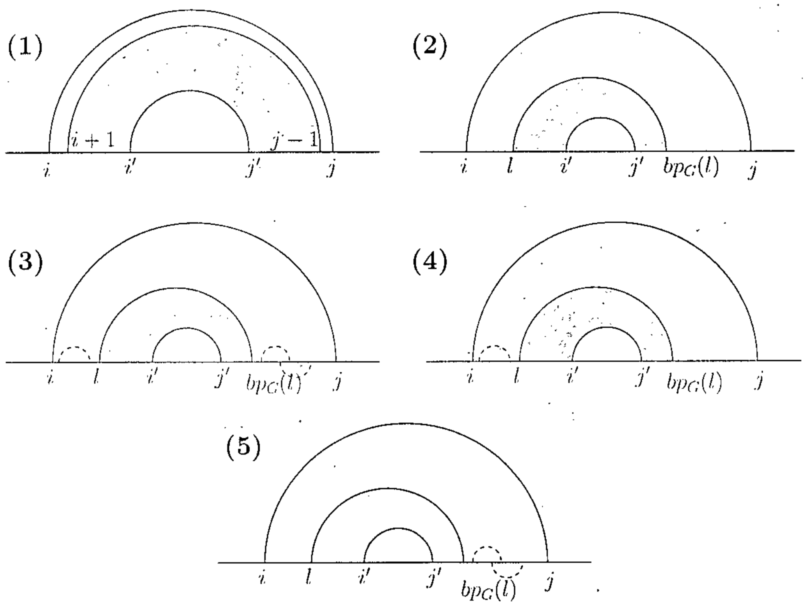


Figure A.9: Illustration of Cases for $BE_{i,i',j',j}$.

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 $[bp_G(l), j]$.

Case (4) handles the case that the region $[bp_G(l), j]$, 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.