HFold: RNA Pseudoknotted Secondary Structure Prediction Using Hierarchical Folding

Hosna Jabbari[1], Anne Condon[1], Ana Pop[2], Cristina Pop[2], and Yinglei Zhao[1]

[1] Dept. of Computer Science, U. of British Columbia
{hjabbari,condon}@cs.ubc.ca
[2] Dept. of Electrical and Computer Engineering, U. of British Columbia

Abstract. Improving the accuracy and efficiency of computational RNA secondary structure prediction is an important challenge, particularly for pseudoknotted secondary structures. We propose a new approach for prediction of pseudoknotted structures, motivated by the hypothesis that RNA structures fold hierarchically, with pseudoknot free pairs forming initially, and pseudoknots forming later so as to minimize energy relative to the initial pseudoknot free structure. Our HFold (Hierarchical Fold) algorithm has $O(n^3)$ running time, and can handle a wide range of biological structures, including nested kissing hairpins, which have previously required $\Theta(n^6)$ time using traditional minimum free energy approaches. We also report on an experimental evaluation of HFold.

Keywords: RNA, Secondary Structure Prediction, Folding Pathways, Pseudoknot, Hierarchical Folding.

1 Introduction

RNA molecules aid in translation and replication of the genetic code, catalyze cellular processes, and regulate the expression level of genes [1]. 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). State-of-the-art prediction algorithms, such as Mfold [2] or RNAfold [3] find the structure with *minimum free energy* (MFE) from the set of all possible pseudoknot free secondary structures.

While many small RNA secondary structures are pseudoknot free, pseudoknots do arise frequently in biologically-important RNA molecules. Examples include simple H-type pseudoknots, with two interleaved stems, which are essential for certain catalytic functions and for ribosomal frameshifting [4], as well as kissing hairpins, which are essential for replication in the coxsackie B virus [5]. Unfortunately, MFE pseudoknotted secondary structure prediction is NP-hard [6,7]. Polynomial-time MFE-based approaches to pseudoknotted structure prediction have been proposed [6,8,9] which find the MFE structure for a given input

R. Giancarlo and S. Hannenhalli (Eds.): WABI 2007, LNBI 4645, pp. 323–334, 2007.

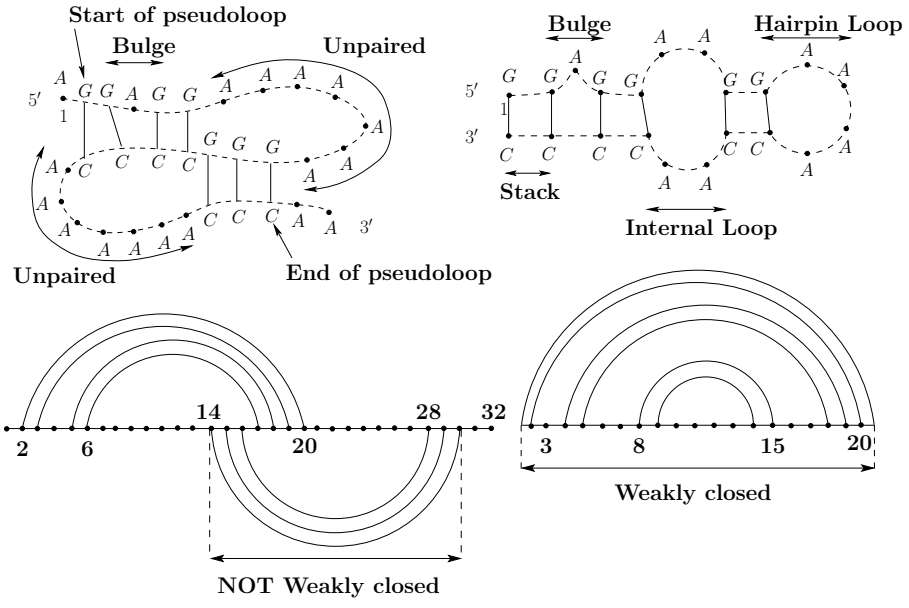

Fig. 1. An H-type pseudoknotted structure (left) and a pseudoknot free structure (right) in graphical (top) and arc diagram (bottom) formats

sequence, from a restricted class of structures. 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 handled by $\Theta(n^5)$ algorithms [6,8], but can be handled in $\Theta(n^6)$ time [9]. (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 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* [10,11], 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" [10]. (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, in this paper 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 [12]. However, even in this case, 15 of the 19 base pairs in the