Chapter 24
Structural Alignment of RNAs with Pseudoknots

Thomas K. F. Wong
The University of Hong Kong, Hong Kong

S. M. Yiu
The University of Hong Kong, Hong Kong

ABSTRACT

Non-coding RNAs (ncRNAs) are found to be critical for many biological processes. However, identifying these molecules is very difficult and challenging due to the lack of strong detectable signals such as opening read frames. Most computational approaches rely on the observation that the secondary structures of ncRNA molecules are conserved within the same family. Aligning a known ncRNA to a target candidate to determine the sequence and structural similarity helps in identifying de novo ncRNA molecules that are in the same family of the known ncRNA. However, the problem becomes more difficult if the secondary structure contains pseudoknots. Only until recently, many of the existing approaches could not handle structures with pseudoknots. This chapter reviews the state-of-the-art algorithms for different types of structures that contain pseudoknots including standard pseudoknot, simple non-standard pseudoknot, recursive standard pseudoknot, and recursive simple non-standard pseudoknot. Although none of the algorithms is designed for general pseudoknots, these algorithms already cover all known ncRNAs in both Rfam and PseudoBase databases. The evaluation of the algorithms also shows that the approach is useful in identifying ncRNA molecules in other species, which are in the same family of a known ncRNA.

INTRODUCTION

A non-coding RNA (ncRNA) is a functional RNA molecule that is not translated into a protein. There are many different types of ncRNAs such as tRNAs, rRNAs, snoRNAs, microRNAs, and siRNAs. These RNA molecules have been found to be involved in many biological processes such as gene regulation, chromosome replication and RNA modification (Frank and Pace, 1998; Nguyen *et al.*,

DOI: 10.4018/978-1-60960-491-2.ch024

2001; Yang *et al.*, 2001). Some are found to be related to cancers and other diseases as well. Similar to proteins, ncRNAs also appear to form a highly structured network that regulates gene expression and translation in the cell (Esquela-Kerscher and Slack, 2006). The number of ncRNAs within the human genome was underestimated before, but recently some databases reveal over 212,000 ncRNAs (He *et al.*, 2007) and more than 1,300 ncRNA families (Griffiths-Jones *et al.*, 2003). Data accumulated on ncRNAs and their families show that ncRNAs may be as diverse as protein molecules (Eddy, 2001).

Identifying ncRNAs is an important problem in the system biological studies. However, this process is very difficult and challenging. Although it is known that some ncRNAs do have promoters and terminators, it is generally believed that ncRNA genes do not contain signals such as open reading frames and ribosome binding sites, which can be easily detected (Argaman *et al*., 2001). Many different computational approaches have been proposed to solve this problem. There are few possible approaches to identify ncRNAs along the genome. Since it is known that the secondary structure of an ncRNA molecule usually plays an important role in its biological functions, for example, the hairpin structures for miRNA precursors and cloverleaf structures for tRNAs, some researches attempted to identify ncRNAs by considering the stability of secondary structures formed by the substrings of a given genome (Le *et al.*, 1990). However, this method is not effective because a random sequence with high GC composition also allows for an energetically favorable secondary structure (Rivas and Eddy, 2000).

Another promising method is the comparative approach. The idea is to make use of some known ncRNAs and try to identify ncRNA candidates along the genome. Along this direction, some authors (Lowe and Eddy, 1997; Nawrocki *et al.*, 2009) use a set of ncRNAs from the same family to train a model (e.g. covariance model). Then, they employ this model to scan the genome and identify potential regions that are ncRNA candidates of that family. The information to be captured from the known ncRNAs depends on how the model is defined. However, in some cases, there are not enough known members in a given family to reliably train a model.

Since the primary sequence and the secondary structure of ncRNA are evolutionary conserved, the ncRNAs of the same family share similar sequence and structure. Another approach is to use a known ncRNA and identify the regions along the genome whose sequence and structure are similar to that of the ncRNA. The resulting regions are the potential ncRNAs candidates of the same family. The key of this approach is to compute the structural alignment between the folded ncRNA (*query*) and the unfolded region (*target*). The unfolded sequence will be folded and aligned simultaneously to the folded ncRNA. The alignment score represents their sequence and structural similarity. The methods like PHMMTS-based method (Sakakibara, 2003), RSEARCH (Klein and Eddy, 2003) and FASTR (Zhang *et al.*, 2005) belong to this category.

However, these methods do not support *pseudoknots*. Given two base pairs at positions (i,j) and (i_0,j_0), where $i<j$ and $i_0<j_0$, pseudoknots are base pairs crossing each other, i.e. $(i<i_0<j<j_0)$ or $(i_0<i<j_0<j)$, as shown in Figure 1. A structure without pseudoknot is regarded as *regular*. Secondary structures including pseudoknots are found in some telomerases (Chen & Greider, 2005), and self-splicing introns (Adams *et al.*, 2004). Pseudoknot structures appear to be key players in some catalytic reactions as well (Dam *et al.,* 1992). For example, pseudoknot structure in the human telomerase RNA is conserved in all vertebrates and is essential for telomerase activity (Chen & Greider, 2005). Structural alignment of ncRNAs containing pseudoknots is believed to be NP-complete (which means the problem cannot be solved in polynomial time) and the problem is computationally demanding. Therefore, researchers started to develop approaches for specific types