Chapter 4
RNA Structure Comparison and Alignment

Kaizhong Zhang

Summary

We present an RNA representation scheme in which an RNA structure is described as a sequence of units, each of which stands for either an unpaired base or a base pair in the RNA molecule. With this structural representation scheme, we give efficient algorithms for computing the distance and alignment between two RNA secondary structures based on edit operations and on the assumptions in which either no bond-breaking operation is allowed or bond-breaking activities are considered. The techniques provide a foundation for developing solutions to the hard problems concerning RNA tertiary structure comparisons. Some experimental results based on real-world RNA data are also reported.

4.1 Introduction

Ribonucleic acid (RNA) is an important molecule, which performs a wide range of functions in biological systems. In particular, it is RNA (not DNA) that contains the genetic information of viruses such as HIV and thereby regulates the functions of these viruses. RNA has recently become the center of much attention because of its catalytic properties [68], leading to an increased interest in obtaining RNA structural information.

RNA molecules have two sets of structural information. First, the *primary structure* of an RNA molecule is a single strand made of the ribonucleotides A (adenine), C (cytosine), G (guanine) and U (uracil). Second, the ribonucleotide sequences fold over onto themselves to form double-stranded regions of base pairings, yielding higher order *tertiary structures*.

It is well known that the structural features of RNAs are important in the molecular mechanisms involving their functions. The presumption, of course, is that, corresponding to a preserved function, there exists a preserved molecular confirmation and therefore a preserved structure. The RNA *secondary structure* is a restricted subset of the tertiary structure, which plays an important role between primary structures and tertiary structures, since the problem of comparing and aligning the tertiary structures of RNA molecules is often intractable. Based on a reliable secondary structure alignment, the possible tertiary structure element alignments that are consistent with the secondary structure alignment can then be introduced.

A coarsely grained RNA secondary structure representation that uses the structural elements of hairpin loops, bulge loops, internal loops, and multibranched loops is proposed in [362, 363]. It has been shown that with this representation, a tree edit distance algorithm can be used to compare RNA secondary structures [363]. Similar ideas have also been used in [244, 245]. Those early works on RNA structure comparison used loops and stacked base pairs as basic units, making it difficult to define the semantic meaning in the process of converting one RNA structure into another. In another line of work, RNA comparison is basically done on the primary structures while trying to incorporate secondary structural information into the comparison [24, 86]. More recent work also uses the notion of arc-annotated sequences [115, 204].

In [447], edit distance, a similarity measure between two RNA secondary structures based on edit operations on base pairs and unpaired bases is proposed. This model has been extended from secondary structures to tertiary structures in [259, 448]. In this model, a base pair in one structure can be aligned only with a base pair in the other structure. Based on this model, algorithms have been developed for global and local alignment with affine gap penalty [72, 423]. In general, this is a reasonable model since in RNA structures when one base of a base pair changes, we usually find that its partner also changes so as to conserve the pairing relationship. However, occasionally a base pair in one structure should be aligned with unpaired bases in the other structure since a mutation of one base may forbid the pairing. In [205, 254] a refined model, which allows base-pair breaking (deleting the bond of the base pair) and base-pair altering (deleting one base and therefore the bond of the base pair), is proposed. In this chapter, we discuss these methods for comparing and aligning RNA structures.

4.2 RNA Structure Comparison and Alignment Models

In this section, we consider RNA structure comparison and alignment models. We first consider the RNA structure comparison model based on edit operations proposed in [259, 448] and the alignment model with gap initiation cost based on edit operations proposed in [423]. We then extend the edit