

Comparison of pseudoknotted RNA secondary structures by topological centroid identification and tree edit distance

Mathematical Bioinformatics WANG Feiqi

Comparison of RNA structures is one of the most crucial analysis for elucidating their individual functions and promoting medical applications. Since it is widely accepted that their functions and structures are strongly correlated, various methods for RNA secondary structure analysis have been proposed due to the difficulty in predicting RNA three dimensional (3D) structure directly from its sequence. However, there are few methods dealing with RNA secondary structures with a specific and complex partial structure called pseudoknot in spite of its significance to biological process, which is a big obstacle for analyzing their functions.

Here, we propose a novel tree representation of pseudoknotted RNA secondary structures by topological centroid identification and their comparison methods based on the tree edit distance. In the proposed method, a given graph representing an RNA secondary structure is transformed to a tree rooted at one of the vertices constituting the topological centroid which is identified by removing cycles with peeling processing for the graph. When comparing tree-represented RNA secondary structures collected from a public database using the tree edit distance and functional gene groups defined by Gene Ontology (GO), the proposed method showed better clustering results according to their GOs than canonical RNA sequence-based comparison. In addition, we also report a case that the combination of the tree edit distance and the sequence edit distance shows a better classification of the pseudoknotted RNA secondary structures.