Monte Carlo Tree Search for Molecule Design

Mehyar MLAWEH

Email: mehyar.mlaweh@dauphine.eu

Supervisor: Prof. Tristan Cazenave

Email: tristan.cazenave@lamsade.dauphine.fr

Abstract

The design of RNA sequences that fold into predefined three-dimensional structures, known as the inverse RNA folding problem, is a fundamental challenge in computational biology. This problem is inherently difficult due to the vast combinatorial sequence space and the need to simultaneously satisfy multiple constraints, including thermodynamic stability, structural accuracy, and biological functionality. As a result, RNA design has been shown to be computationally hard, motivating the development of advanced heuristic and stochastic optimization approaches.

This PhD research focuses on the development and analysis of Monte Carlo Tree Search (MCTS) and related stochastic search algorithms for de novo RNA design at the tertiary structure level. The goal is to explore how such methods can efficiently navigate the high-dimensional sequence space while integrating feedback from modern structure prediction tools. The work is positioned at the intersection of combinatorial optimization, machine learning, and structural bioinformatics, with the broader objective of enabling reliable inverse folding for complex RNA structures.

In addition to Monte Carlo-based approaches, recent work has explored alternative metaheuristics, including Bee Colony Optimization algorithms, applied to the RNA tertiary inverse folding problem. Preliminary results demonstrate that this approach can generate valid RNA sequences for relatively short molecules (typically below 100 nucleotides). However, its performance degrades as sequence length increases, highlighting scalability challenges. Furthermore, this approach does not explicitly account for pseudoknots, which correspond to complex structural motifs involving multi-chain or non-nested base-pairing interactions. As a consequence, it remains limited in its ability to model more realistic and biologically relevant RNA tertiary structures.

The current research direction addresses these limitations through a multi-objective optimization framework. Rather than focusing solely on structural matching, the proposed approach aims to simultaneously optimize several criteria, including folding accuracy, physical plausibility, and geometric consistency of the resulting three-dimensional structures. In particular, the work emphasizes maintaining the physics and geometry of the 3D correspondences associated with generated de novo RNA sequences, ensuring that predicted conformations are both structurally accurate and physically meaningful.

By combining Monte Carlo Tree Search, multi-objective optimization, and structure-aware evaluation, this thesis seeks to advance the state of the art in RNA inverse folding at the tertiary level. The expected contributions include more scalable algorithms, improved handling of complex structural features such as pseudoknots, and enhanced integration of physical constraints. Ultimately, this research may support applications in synthetic biology, RNA therapeutics, and de novo molecular design.