"Statistical sampling of RNA structures: Analysis and algorithmic improvements"

Historical approaches for the prediction of the conformation adopted by a macromolecule (RNA, Protein) rely on the assumption that the most probably observed conformation is the one that minimizes an energy (or pseudo-energy) function. Whereas this paradigm is at the core of successful predictive methods, it fails to capture the instability of certain biopolymers, or more generally its kinetics. Therefore, recent approaches favor a study of the properties of the Boltzmann ensemble, which consists in the set of all achievable conformations, weighted by a distribution based on the energy, the Boltzmann distribution. In the context of RNA, this paradigm shift has allowed for a significant improvement of existing methods for the problem of ab initio folding, and have shown useful for the rational design of silencing RNAs (Long et al 2007).

The evaluation of properties of interest on this ensemble is performed through a statistical sampling according to the Boltzmann distribution. Thanks to the nice algebraic properties of both the conformation space (secondary structures) and the energy function (Turner "additive" loop model), this sampling can be performed as a stochastic traceback. However, arbitrary sample sizes have been used so far, mainly because of the computational costs of these methods, whose algorithmic complexity was largely unknown. It is therefore crucial to improve the sampling methods in order to increase the coverage of the Boltzmann ensemble.

Firstly, we performed an analysis of the classical algorithm and find a worst-case and average-case complexities respectively in $\Theta(n^3+kn^2)$ and $\Theta(n^3+kn\sqrt{n})$. We adapted the so-called Boustrophedon strategy, introduced by Flajolet et al in the context of random generation, which consists in a non-trivial reordering of the potential partners for a base. This was shown to reduce the cost of the worst-case scenario to $\Theta(n^3+kn\log(n))$ arithmetic operations. Then, we proposed a modification of existing algorithms in order to perform a non-redundant sampling of RNA structures which, without any significant computational overhead, allowed for a more thorough coverage of the Boltzmann probability distribution.The algorithmic generality of the tools developed during this research should allow for a transposition to the dynamic-programming based tools, currently developed for the study of protein structure (PartiFold), and for an efficient generation of distinct suboptimal candidates.