Wikno, 16th–20th September 2025

INTEGER PROGRAMMING FRAMEWORK FOR RNA SECONDARY STRUCTURE PREDICTION

Olga Karelkina

Department of Decision Support in the Presence of Risk, Systems Research Institute PAS,
ul. Newelska 6, 01-447 Warsaw,
karelkin@ibspan.waw.pl

ABSTRACT

The prediction of RNA secondary structure is a fundamental task in bioinformatics and structural biology due to the crucial roles RNA molecules play in diverse biological processes. To carry out its functions, RNA must fold into a well-defined structure. RNA exhibits a naturally hierarchical organization. Its primary structure refers to the linear sequence of nucleotides, the secondary structure is formed through canonical base pairings (Watson-Crick-Franklin and Wobble), and the tertiary structure corresponds to the molecule's three-dimensional atomic arrangement. Since secondary structure contacts are typically stronger and form more rapidly than tertiary interactions, secondary structures can often be predicted independently and serve as a critical intermediate step toward solving the more complex problem of tertiary structure prediction.

Traditional computational methods for RNA secondary structure prediction, guided by a thermodynamic hypothesis and Turner's nearest-neighbor parameters [1], aim to identify the most stable conformation of an RNA molecule by minimizing its overall free energy. These methods decompose RNA structure into well-defined substructures – such as stems, hairpins, internal loops, bulges, and multibranch loops – and assign empirically derived free energy values to each component based on its sequence and structural context. Despite dynamic programming algorithms, such as those implemented in RNAstructure [2] or ViennaRNA [3], allow for finding the minimum free energy (MFE) structures efficiently, there is still a wide margin for improvement in predicting accuracy.

In this work, we present a novel energy-based integer programming (IP) framework to predict RNA secondary structure via loop decomposition from a single input sequence. We provide a formal mathematical definition of loops and integrate them into our optimization model via binary variables, which indicate whether to include a particular loop in the solution. The search space of all feasible loop decompositions is defined by a set of linear constraints. The integer programming model obtains optimal secondary structure by minimizing the sum of energies over all possible loops for a given RNA sequence. Since RNA molecules can populate an ensemble of structures, we also introduce an extended parameterized model that generates suboptimal structures to provide alternative conformations to the MFE structure. Additionally, we address scalability challenges by exploring the influence of initial feasible solutions on overall computation time.

The proposed IP model was benchmarked on the archive II dataset of sequences with experimentally determined reference secondary structures. It was implemented in Python 3 and solved with

a state-of-the-art optimizer that employs the branch-and-cut technique. Model's predictions were compared to references and structures produced by the dynamic programming methods using standard metrics, including Interaction Network Fidelity (INF) and F_β score. The results show that the IP-based approach generates biologically meaningful structures and, in some cases, outperforms dynamic programming algorithms in predictive accuracy. Therefore, the IP framework provides a valuable supplementary tool for identifying alternative secondary structures.

REFERENCES

[1] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner: *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure*, Proc. Natl. Acad. Sci. USA **101** (2004), 7287–7292.

[2] J.S. Reuter and D.H. Mathews: *RNAstructure: software for RNA secondary structure prediction and analysis*, BMC Bioinformatics **11** (2010), 1–10.

[3] R Lorenz, S.H. Bernhart, and C. et al. Höner: *ViennaRNA package 2.0*, Algorithms Mol Biol **6** (2011), 1–10.