On the Use of Decision Diagrams for the Repetition-Free Longest Common Subsequence*

Matthias Horn[1], Marko Djukanovic[1], Christian Blum[2], and Günther R. Raidl[1]

[1] Institute of Logic and Computation, TU Wien, Vienna, Austria
{horn|djukanovic|raidl}@ac.tuwien.ac.at
[2] Artificial Intelligence Research Institute (IIIA-CSIC), Campus UAB, Bellaterra, Spain
christian.blum@iiia.csic.es

Abstract. The goal of the repetition-free longest common subsequence (RFLCS) problem is to find a longest sequence which is common to two input strings such that each character in the common subsequence appears at most once. In this work, the RFLCS problem is solved by transforming an instance to an instance of the maximum independent set (MIS) problem which is then solved by a mixed integer linear programming solver. To reduce the size of the underlying conflict graph of the MIS problem, a relaxed decision diagram is utilized.

Keywords: Decision Diagrams, Repetition-Free Longest Common Subsequence, Maximum Independent Set

The *longest common subsequence* (LCS) problem asks for the longest sequence which is common to a set of input strings. A subsequence is a string which can be obtained by possible deleting characters from another string. The problem has applications in bioinformatics, where strings often represents segments of RNA or DNA [5]. Other fields where the LCS problem appears are text editing, data compression, file comparison, and the production of circuits in field programmable gate arrays. An additional constraint which arises in certain real world scenarios is that each character may appear in the common subsequence at most once. This problem, denoted as the *repetition-free LCS* (RFLCS) problem, is usually considered for two input strings and is even then APX-hard [1].

This work builds upon the work of Blum et al. [3], where instances of the RFLCS problem are transformed to instances of the *maximum independent set* (MIS) problem. Hereby, an independent set of the underlying conflict graph of the MIS problem corresponds to a repetition-free common subsequence of the RFLCS instance. To solve the MIS problem the *integer linear programming* (ILP) solver CPLEX is applied. The performance of the ILP solver depends to a large extend on the size of the conflict graph. Therefore, in [3] the size of the conflict graph is reduced by filtering nodes based on lower and upper bounds.

In the last decade, *decision diagrams* (DDs) have been recognized as a powerful tool for combinatorial optimization problems. In particular *relaxed DDs* may

* This project is partially funded by the Doctoral Program "Vienna Graduate School on Computational Optimization", Austrian Science Foundation (FWF) Project No. W1260-N35.

provide compact representations discrete relaxations. Besides allowing for new interference techniques in constraint programming and novel branching schemes, they may also provide tight dual bounds. For a comprehensive survey see [2].

In this work, we compile relaxed *multivalued DDs* (MDDs) for the RFLCS problem. The advantages are twofold. First, with the aid of relaxed MDDs it is possible to reduce the size of the conflict graph even further, yielding performance improvement of the subsequently applied ILP solver. Second, if the ILP solver is not able to solve an instance to proven optimality within a given time limit then the compiled relaxed MDD may be able to provide a tighter upper bound as the ILP solver does. The relaxed MDDs are compiled with an adapted incremental refinement approach from [4] by incorporating also problem specific upper bounds of the RFLCS problem.

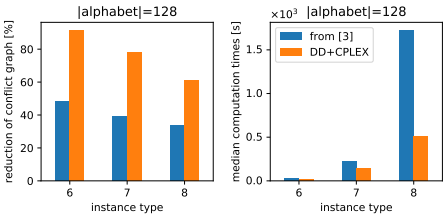

Fig. 1. Graph reduction [%] and median running times [s] for middle-sized instances.

Preliminary experimental results document the advantages of compiling relaxed MDDs of RFLCS instances in order to reduce the size of the conflict graphs. The subsequently applied ILP solver is able to solve 20.3% of the instances to proven optimality that could not be solved by [3]. Furthermore, for instances that the ILP solver can already solve to optimality, the overall median solving time can be decreased by using the relaxed MDD in advance see Figure 1. Finally, for the hardest instance classes, which cannot be solved to proven optimality, the compiled relaxed MDDs are able to provide on average tighter upper bounds.

References

1. S. S. Adi, M. D. Braga, C. G. Fernandes, C. E. Ferreira, F. V. Martinez, M.-F. Sagot, M. A. Stefanes, C. Tjandraatmadja, and Y. Wakabayashi. Repetition-free longest common subsequence. *Discrete Applied Mathematics*, 158(12):1315–1324, 2010.
2. D. Bergman, A. A. Cire, W.-J. van Hoeve, and J. N. Hooker. *Decision Diagrams for Optimization*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2016.
3. C. Blum, M. Djukanovic, A. Santini, H. Jiang, C.-M. Li, F. Manya, and G. R. Raidl. Solving longest common subsequence problems via a transformation to the maximum clique problem. Technical Report AC-TR-20-003, 2020. submitted to Computers and Operations Research.
4. A. A. Cire and W. V. Hoeve. Multivalued decision diagrams for sequencing problems. *Operations Research*, 61(6):1411–1428, 2013.
5. T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.