



# Hybrid techniques based on solving reduced problem instances for a longest common subsequence problem

Christian Blum<sup>a,\*</sup>, Maria J. Blesa<sup>b</sup>

<sup>a</sup> Artificial Intelligence Research Institute (IIIA-CSIC), Campus UAB, Bellaterra, Spain

<sup>b</sup> Computer Science Department, Universitat Politècnica de Catalunya – BarcelonaTech, Barcelona, Spain

## ARTICLE INFO

### Article history:

Received 15 June 2017

Received in revised form 8 September 2017

Accepted 4 October 2017

Available online 18 October 2017

### Keywords:

Combinatorial optimization  
Longest common subsequences  
Integer linear programming  
Heuristic  
Hybrid algorithm

## ABSTRACT

Finding the longest common subsequence of a given set of input strings is a relevant problem arising in various practical settings. One of these problems is the so-called longest arc-preserving common subsequence problem. This NP-hard combinatorial optimization problem was introduced for the comparison of arc-annotated ribonucleic acid (RNA) sequences. In this work we present an integer linear programming (ILP) formulation of the problem. As even in the context of rather small problem instances the application of a general purpose ILP solver is not viable due to the size of the model, we study alternative ways based on model reduction in order to take profit from this ILP model. First, we present a heuristic way for reducing the model, with the subsequent application of an ILP solver. Second, we propose the application of an iterative hybrid algorithm that makes use of an ILP solver for generating high quality solutions at each iteration. Experimental results concerning artificial and real problem instances show that the proposed techniques outperform an available technique from the literature.

© 2017 Elsevier B.V. All rights reserved.

## 1. Introduction

In computer science terms, a *string* (or sequence)  $x$  of length  $l_x$  is a finite sequence of characters from a finite alphabet  $\Sigma$ . In fact, strings are popular data types for representing and storing information. Words and even complete texts, for example, may be stored in a computer in terms of strings. However strings are not only useful in fields such as information and text processing. They arise, in particular, in the field of computational biology. The reason is that most of the genetic instructions involved in the growth, development, functioning and reproduction of living organisms are stored by means of *deoxyribonucleic acid* (DNA) and *ribonucleic acid* (RNA) molecules, which are either double-stranded (DNA) or single-stranded (RNA) sequences of nucleotides. In short, each nucleotide is composed of a nitrogenous base, a five-carbon sugar (ribose or deoxyribose), and at least one phosphate group. Concerning RNA, each nucleotide has one of four different nitrogenous bases: guanine (G), uracil (U), adenine (A), and cytosine (C). As a consequence, any RNA molecule can be represented as a string of symbols from  $\Sigma = \{G, U, A, C\}$ , which is called the *primary structure* of a RNA molecule. The primary structure of a RNA molecule is

a simplified representation, because RNA molecules fold in space and different nucleotides bind together, for example, by means of hydrogen bonds. Generally, guanine (G) can only bind with cytosine (C) and uracil (U) can only bind with adenine (A). These hydrogen bonds are present in the so-called *secondary structure* of an RNA molecule; see Fig. 1a for an example.

For computer science purposes, the hydrogen bonds of the secondary structure of an RNA sequence  $x$  can be represented by a so-called *arc annotation set*  $P_x$ . In technical terms,  $P_x$  is an unordered set of pairs of positions of a string  $x$ .<sup>1</sup> Each pair  $(i_1, i_2) \in P_x$  represents an arc between positions  $i_1$  and  $i_2$  and is called an *arc annotation*. The only convention is that  $i_1 < i_2$  must hold for any arc  $(i_1, i_2) \in P_x$ . Finally,  $i_1$  is called the *left endpoint* of arc  $(i_1, i_2)$ , and  $i_2$  is called the *right endpoint*. A pair  $(x, P_x)$  is called an *arc-annotated sequence* [2] (or arc-annotated string). Given this definition, note that the secondary structure of an RNA sequence can conveniently be described by an arc-annotated sequence; see Fig. 1b for an example. In fact, arc-annotated sequences have been widely used for this purpose (see, for example, [3]). In particular, arc-annotated sequences have shown to be useful for the structural comparison of RNA sequences. One of the usual measures when comparing two (or more) sequences is the length of their *longest common subsequence*

\* Corresponding author.

E-mail addresses: [christian.blum@iia.csic.es](mailto:christian.blum@iia.csic.es) (C. Blum), [mjblesa@cs.upc.edu](mailto:mjblesa@cs.upc.edu) (M.J. Blesa).

<sup>1</sup> As a convention, the positions of a string  $x$  range from 1 to  $l_x$ .



Download English Version:

<https://daneshyari.com/en/article/6904195>

Download Persian Version:

<https://daneshyari.com/article/6904195>

[Daneshyari.com](https://daneshyari.com)