

# A Hybrid Evolutionary Algorithm Based on Solution Merging for the Longest Arc-Preserving Common Subsequence Problem

Christian Blum<sup>1</sup> and Maria J. Blesa<sup>2</sup>

<sup>1</sup>Artificial Intelligence Research Institute (IIIA-CSIC)  
Campus UAB, Bellaterra, Spain  
christian.blum@iiia.csic.es

<sup>2</sup>Computer Science Department  
Universitat Politècnica de Catalunya (BarcelonaTech), Barcelona, Spain  
mjblesa@cs.upc.es

## Abstract

The longest arc-preserving common subsequence problem is an NP-hard combinatorial optimization problem from the field of computational biology. This problem finds applications, in particular, in the comparison of arc-annotated Ribonucleic acid (RNA) sequences. In this work we propose a simple, hybrid evolutionary algorithm to tackle this problem. The most important feature of this algorithm concerns a crossover operator based on solution merging. In solution merging, two or more solutions to the problem are merged, and an exact technique is used to find the best solution within this union. It is experimentally shown that the proposed algorithm outperforms a heuristic from the literature.

## 1 Introduction

In computer science, a *string* (or sequence)  $x$  of length  $l_x$  is defined as a finite sequence of characters from a finite alphabet  $\Sigma$ . A string is a data type used to represent and store information. Words in a specific language, for example, are stored in a computer in terms of strings. Even whole texts may be stored by means of strings. Apart from fields such as information and text processing, strings arise, in particular, in the field of computational biology. This is because most of the genetic instructions involved in the growth, development, functioning and reproduction of living organisms are stored in *Deoxyribonucleic acid* (DNA) and *Ribonucleic acid* (RNA) molecules, which are double-stranded (in the case of DNA) or single-stranded (in the case of RNA) sequences of nucleotides. Hereby, each nucleotide is composed of a nitrogenous base, a five-carbon sugar (ribose or deoxyribose), and at least one phosphate group. Nucleotides in the context of RNA have one of four different nitrogenous bases: guanine (G), uracil (U), adenine (A), and cytosine (C). Therefore, any RNA molecule can be represented as a string of symbols from  $\Sigma = \{G, U, A, C\}$ . Such a string is called the *primary structure* of an RNA molecule. However, RNA molecules generally fold in space, and different nucleotides bind together, for example, by means of hydrogen bonds. In simplified

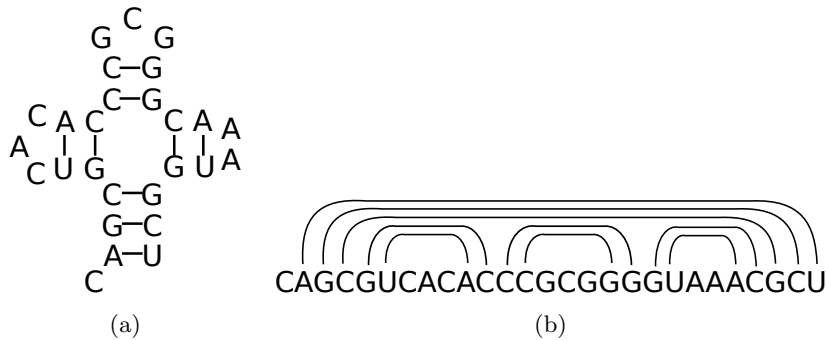


Figure 1: (a) Example of the secondary structure of an RNA molecule. (b) The corresponding arc-annotated sequence. The example is reproduced from [9].

terms, G can only bind with C and U can only bind with A. An example of the secondary structure of an RNA molecule is shown in Figure 1a.

In computer science terms, the hydrogen bonds of the secondary structure of an RNA sequence  $x$  can be represented by a so-called *arc annotation set*  $P_x$ , which is an unordered set of pairs of positions of  $x$ . Henceforth, the positions of a string  $x$  range from 1 to  $l_x$ . Each pair  $(i_1, i_2) \in P_x$  is called an *arc annotation* (or simply an *arc*) between positions  $i_1$  and  $i_2$ . As a convention, it must hold that  $i_1 < i_2$ . Moreover,  $i_1$  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* [5]. Note that the secondary structure of an RNA sequence can be described by an arc-annotated sequence. For an example see Figure 1b. In fact, arc-annotated sequences have been widely used for this purpose (see, for example, [4]). In particular, arc-annotated sequences have been useful in the context of the structural comparison of RNA sequences. An important way of comparing two (or more) sequences consists in computing their *longest common subsequence* (LCS). Given a sequence  $x$  over a finite alphabet  $\Sigma$ , sequence  $t$  is called a *subsequence* of  $x$ , if  $t$  can be produced from  $x$  by deleting characters. Given a set of input strings  $\{s_1, \dots, s_n\}$ , the problem of finding the longest common subsequence of all input strings is, in general, NP-hard [11]. The best techniques available nowadays for solving this problem are based on beam search [2].

## 1.1 Contribution of this Work

The longest common subsequence problem in the context of arc-annotated sequences—the *longest arc-preserving common subsequence* (LAPCS) problem—has first been introduced in [6, 5]. In particular, in the same works it was shown that the most general case of the problem (without any restrictions on the arcs) is NP-hard. In the meanwhile, five different variants of the problem—that is, restrictions of the general problem—have been studied in the related literature, and efficient algorithms were developed for three of these variants.<sup>1</sup> However, as far as we know, only one algorithm that is applicable to the most general case has been proposed so far (see [9]). In this work, we first phrase the LAPCS problem in the form of an integer linear program (ILP) [12]. Then we make use of this ILP in the context of a simple evolutionary algorithm based on *solution merging*, where two or more solutions are merged and the best solution in this union is derived by means of an exact technique.

<sup>1</sup>More details are given in Section 2.

Aggarwal et al. [1] originally suggested such an approach, labeled *optimized crossover*, for the independent set problem. We provide an extensive experimental comparison of the proposed algorithm in comparison to the heuristic from the literature and to the application of a randomized multi-start heuristic.

## 1.2 Outline of the Paper

The remainder of this paper is structured as follows. In Section 2, a technical description of the tackled problem is provided. In subsequent sections—see Section 3 and Section 4—we describe a heuristic from the literature and phrase an ILP model for the tackled problem. Next, the proposed algorithm is outlined in Section 5. Finally, an extensive experimental evaluation on artificial and real problem instances is provided in Section 6, and an outlook to future work is given in Section 7.

## 2 The LAPCS Problem

Given two input sequences  $x$  and  $y$ , we define the set of possible assignments  $A$  as the set of all  $a_{i,j}$ —where  $i \in \{1, \dots, l_x\}$  and  $j \in \{1, \dots, l_y\}$ —such that  $x[i] = y[j]$ . That is  $A$  consists of all  $a_{i,j}$  such that the letter at position  $i$  of  $x$  is equal to the letter at position  $j$  of  $y$ . A valid common subsequence of the two input sequences  $x$  and  $y$  can then be represented by a subset  $S \subseteq A$  that fullfills the following conditions:

- **Common subsequence condition:** For any two assignments  $a_{i,j}, a_{k,l} \in S$  (where  $a_{i,j} \neq a_{k,l}$ ) it must hold that either  $i < k$  and  $j < l$  or  $i > k$  and  $j > l$ .

In order to translate such a solution into the corresponding common subsequence, the assignments in  $S$  have to be ordered from small to large indeces, either according to the first or the second index. Then, the letters corresponding to the assignments must be joined in this order.

A solution  $S$  that fullfills the common subsequence condition is called *arc-preserving* if the arcs induced by the solution are preserved:

- **Arc preservation condition:** for any two assignments  $a_{i,j}, a_{k,l} \in A$  (where  $a_{i,j} \neq a_{k,l}$  and  $i < k$ ) it must hold that  $(i, k) \in P_x \Leftrightarrow (j, l) \in P_y$ .

Given two arc-annotated input strings  $x$  and  $y$ , the LAPCS problem consists in finding a solution  $S \subseteq A$  that fullfills both the common subsequence and the arc preservation condition and is of maximal cardinality. Note that such a mapping corresponds to the longest arc-preserving common subsequence of  $x$  and  $y$ . In [5, 6] it was shown that this problem is, in general, NP-hard. However, the structure of the arc annotation in the context of RNA sequences, for example, is in practise likely to satisfy some constraints. Concerning a string  $x$ , the following types of constraints have been considered in the literature:

1. No two arcs may share an endpoint. That is,  $\forall (i_1, i_2), (i_3, i_4) \in P_x$  it must hold that  $i_1 \neq i_4$  and  $i_2 \neq i_3$ .
2. Crossing arcs do not exist. That is,  $\forall (i_1, i_2), (i_3, i_4) \in P_x$  it must hold that  $i_3 \leq i_1 \leq i_4 \Leftrightarrow i_3 \leq i_2 \leq i_4$ .

3. Nesting arcs do not exist. That is,  $\forall (i_1, i_2), (i_3, i_4) \in P_x$  it must hold that  $i_1 \leq i_3 \Leftrightarrow i_2 \leq i_4$ .
4. No arcs exist. That is,  $P_x = \emptyset$ .

Based on these four restrictions concerning the arc annotation, input strings are characterized into UNLIMITED (no constraints), CROSSING (constraint 1), NESTED (constraints 1 and 2), CHAIN (constraints 1, 2, and 3), and PLAIN (constraint 4). Different versions of the LAPCS problem can therefore be denoted as follows: LAPCS( $\cdot, \cdot$ ) where each of the two dots must be replaced by the characterization of the arc annotation of the first and the second input string. For example, in problem LAPCS(UNLIMITED,NESTED), the arc annotation of the first input string is characterized as UNLIMITED, and the one of the second one as NESTED. It is well known that LAPCS(PLAIN,PLAIN), for example, can be solved in polynomial time with the dynamic programming algorithm by Smith and Waterman [13]. In this paper, however, we deal with the most general version of the problem, LAPCS(UNLIMITED,UNLIMITED), which—as mentioned above—was shown to be NP-hard. For simplicity reasons we refer to this problem version simply as LAPCS.

### 3 Existing Heuristic for LAPCS

As far as we know, the only heuristic from the literature that is applicable to the most general version of the LAPCS problem was described in [9], and works as follows. First, the dynamic programming algorithm by Smith and Waterman is applied to input strings  $x$  and  $y$ , disregarding the arc annotations. The result is a mapping  $S \subseteq A$  that—most probably—violates some of the arc preservation constraints. In order to *repair* this invalid solution, the following is done. First a graph  $G$  is constructed as follows. A vertex  $v$  is introduced for each assignment  $a_{i,j} \in S$ . Two vertices  $v$  (corresponding to an assignment  $a_{i,j} \in S$ ) and  $v'$  (corresponding to an assignment  $a_{k,l} \in S$  with  $i < k$ ) are connected by an edge if either  $(i, k) \in P_x$  or  $(j, l) \in P_y$ , but not both. In other words, two vertices are connected by an edge if they represent a violation of the arc preservation constraints. Note that in order to repair  $S$  by removing as few assignments from  $S$  as possible, we can solve the *maximum independent set* (MIS) problem in  $G$ , and remove all assignments from  $S$  that correspond to vertices that are not in the optimal solution to the MIS problem. In our implementation we used CPLEX 12.6 to solve the MIS problem in all cases.

### 4 An ILP Model for LAPCS

The LAPCS problem can be stated in terms of an integer linear program (ILP) in the following way. For each  $a_{i,j} \in A$  is introduced a binary variable  $z_{i,j}$ . The set of all binary variables is denoted by  $Z$ . We say that two variables  $z_{i,j} \neq z_{k,l}$  (where  $i \leq k$ ) are *in conflict*, if setting both variables to one violates (1) the common subsequence condition, (2) the arc preservation condition, or both. In technical terms, two variables  $z_{i,j} \neq z_{k,l}$  (where  $i \leq k$ ) are in conflict, if at least one of the following holds:

1.  $j \geq l$
2. Either  $(i, k) \in P_x$  or  $(j, l) \in P_y$ , but not both at the same time.

The LAPCS problem can then be rephrased as the problem of selecting a maximal number of non-conflicting variables from  $Z$ . Given these notations, the ILP is stated as follows.

$$\begin{aligned} & \max \sum_{z_{i,j} \in Z} z_{i,j} & (1) \\ \text{subj. to:} & & \\ & z_{i,j} + z_{k,l} \leq 1 \quad \forall z_{i,j} \neq z_{k,l}, i \leq k \text{ in conflict} & (2) \\ & z_{i,j} \in \{0, 1\} \quad \text{for } z_{i,j} \in Z & (3) \end{aligned}$$

Hereby, constraints (2) ensure that selected variables are not in conflict.

## 5 The Hybrid EA with Solution Merging

The proposed hybrid EA, henceforth labelled HYB-EA, is pseudo-coded in Algorithm 1. In the context of this algorithm, valid solutions to the problem are subsets of the complete set  $Z$  of variables introduced in the context of the ILP model. If a solution  $S$  contains a variable  $z_{i,j}$ , this means that the variable must be given value one in order to produce the corresponding solution. The main loop of the EA is executed while the CPU time limit is not reached. It consists of the following actions. First, the best-so-far solution  $S_{\text{bsf}}$  is initialized to  $\emptyset$ . Then, at each iteration, first, the set of variables representing a set of merged solutions is initialized with the best-so-far solution  $S_{\text{bsf}}$ . Then, a number of  $n_{\text{sols}}$  solutions is probabilistically constructed in function `GenerateRandomSolution`( $d_{\text{rate}}, l_{\text{size}}, Z$ ) in line 6 of Algorithm 1. The variables contained in these solutions are added to  $S'$ . Afterwards, solution merging is applied to  $S'$ , that is, an ILP solver is applied to find the best valid solution that can be built from the variables in  $S'$  (see function `ApplySolutionMerging`( $t_{\text{max}}, S'$ ) in line 9 of Algorithm 1). Parameter  $t_{\text{max}}$  is a time limit for the ILP solver. In particular, the output of this function is the best solution found by the ILP solver within  $t_{\text{max}}$  seconds. Note that for applying an ILP solver to  $S' \subseteq Z$ , all the appearances of  $Z$  in the ILP model of Section 4 have to be replaced with  $S'$ . In case  $S'_{\text{opt}}$  is better than the current best-so-far solution  $S_{\text{bsf}}$ , solution  $S'_{\text{opt}}$  is stored as the new best-so-far solution (line 10). The output of the algorithm is the best-so-far solution  $S_{\text{bsf}}$ .

In the following we will describe in detail the remaining component of the algorithm: the probabilistic construction of solutions in function `GenerateRandomSolution`( $d_{\text{rate}}, l_{\text{size}}, Z$ ). First, a common subsequence of  $x$  and  $y$  is—without regarding the arc preservation constraints—probabilistically generated as follows. For this purpose let us first introduce for each letter  $a \in \Sigma$  the subset  $Z_a \subseteq Z$  of variables which correspond to letter  $a$ . A solution construction starts with an empty solution  $S = \emptyset$ , and the first step consists in generating the set of variables  $C \subseteq Z$  that serve as options to be added to  $S$ . More specifically, the initial set  $C$  is generated in order to contain for each letter  $a \in \Sigma$  the variable  $z_{i,j} \in Z_a$  (if any) such that  $i \leq k$  and  $j \leq l, \forall z_{k,l} \in Z_a$ . Moreover, options  $z_{i,j} \in C$  are given a weight value  $w(z_{i,j}) := \frac{i}{l_x} + \frac{j}{l_y}$ , which is a known greedy function for longest common subsequence problems (see, for example, [7, 8]). At each construction step, exactly one variable is chosen from  $C$  and added to  $S$ . For doing so, first, a value  $r$  is chosen uniformly at random from  $[0, 1]$ . In case  $r \leq d_{\text{rate}}$ , where  $d_{\text{rate}}$  is a parameter of the algorithm, the variable  $z_{i,j} \in C$

---

**Algorithm 1** HYB-EA for the LAPCS problem

---

```
1: input: strings  $x$  and  $y$  over alphabet  $\Sigma$ , values for parameters  $n_{\text{sols}}$ ,  $d_{\text{rate}}$ ,  $l_{\text{size}}$ , and  $t_{\text{max}}$ 
2:  $S_{\text{bsf}} := \emptyset$ 
3: while CPU time limit not reached do
4:   for  $i = 1, \dots, n_{\text{sols}}$  do
5:      $S' := S_{\text{bsf}}$ 
6:      $S := \text{GenerateRandomSolution}(d_{\text{rate}}, l_{\text{size}}, Z)$ 
7:      $S' := S' \cup S$ 
8:   end for
9:    $S'_{\text{opt}} := \text{ApplySolutionMerging}(t_{\text{max}}, S')$ 
10:  if  $|S'_{\text{opt}}| > |S_{\text{bsf}}|$  then  $S_{\text{bsf}} := S'_{\text{opt}}$ 
11: end while
12: output:  $S_{\text{bsf}}$ 
```

---

with the smallest weight value is deterministically chosen. Otherwise, a candidate list  $L \subseteq C$  of size  $\min\{l_{\text{size}}, |C|\}$  containing the options with the lowest weight values is generated and exactly one variable  $z_{i,j} \in L$  is then chosen uniformly at random and added to  $S$ . Note that  $l_{\text{size}}$  is another parameter of the solution construction process. Finally, the set of options  $C$  for the next construction step is generated. Being  $z_{i,j}$  the last variable that was added to  $S$ ,  $C$  contains for each letter  $a \in \Sigma$  the variable  $z_{r,s} \in Z_a$  (if any) with the lowest weight value  $w(z_{r,s})$  calculated as  $w(z_{r,s}) := \frac{r-i}{l_x-i} + \frac{s-j}{l_y-j}$ . The solution construction is finished when the set of options is empty.

However, note that a solution  $S$  constructed in the way as described above does not necessarily respect all arc preservation constraints. Therefore, the same *repair mechanism* is utilized as in the heuristic from Section 3 in order to transform  $S$  into a valid LAPCS solution.

## 6 Experimental Evaluation

Summarizing, the following techniques are included in the experimental evaluation: (1) the heuristic described in Section 3 (HEURISTIC), (2) the hybrid EA (HYB-EA), and (3) HYB-EA without the application of solution merging. This last algorithm—henceforth denoted by MS-HEUR—is basically a multi-start heuristic that constructs randomized solutions (and applies the repair procedure to them) until it runs out of computation time. Comparing HYB-EA with MS-HEUR will enable us to measure the contribution of solution merging. The three above-mentioned algorithms were implemented in ANSI C++ using GCC 4.7.3, without the use of any external libraries. In addition, the ILP models in the context of HYB-EA were solved with the ILP solver IBM ILOG CPLEX v12.6 in one-threaded mode. The experimental evaluation has been performed on a cluster of PCs with Intel(R) Xeon(R) CPU 5670 CPUs of 12 nuclei of 2933 MHz and at least 40 Gigabytes of RAM. Note that we also tried to apply CPLEX to the complete ILP models for each problem instance. However, the models were too large, even in the case of the smallest problem instances.

The remainder of this section is organized as follows. First, the set of benchmark instances is described. Second, the tuning experiments that were conducted in order to determine a proper setting for the parameters of HYB-EA are outlined. Finally, an exhaustive experimental evaluation is presented.

Table 1: Characteristics of the real-life instances. All 20 RNA sequences, together with their secondary structure, were downloaded from the RNase P Database [3].

Instance name	First String			Second string		
	RNA	Lenght	Arcs	RNA	Lenght	Arcs
Real.1	<i>Allochromatium vinosum</i>	369	119	<i>Haemophilus influenza</i>	377	124
Real.2	<i>Bacteroides thetaiotaomicron</i>	361	121	<i>Porphyromonas gingivalis</i>	398	131
Real.3	<i>Halococcus morrhuae</i>	475	154	<i>Haloferax volcanii</i>	433	142
Real.4	<i>Klebsiella pneumoniae</i>	383	127	<i>Escherichia coli</i>	377	124
Real.5	<i>Methanococcus jannaschii</i>	252	75	<i>Archaeoglobus fulgidus</i>	229	67
Real.6	<i>Methanosarcina barkeri</i>	371	115	<i>Pyrococcus abyssi</i>	330	100
Real.7	<i>Mycoplasma genitalium</i>	384	119	<i>Mycoplasma pneumoniae</i>	369	112
Real.8	<i>Saccharomyces kluyveri</i>	336	90	<i>Schizosaccharomyces octosporus</i>	281	71
Real.9	<i>Serratia marcescens</i>	378	125	<i>Shewanella putrefaciens</i>	354	115
Real.10	<i>Streptomyces bikiniensis</i>	398	135	<i>Streptomyces lividans</i>	405	138

## 6.1 Benchmark Instances

Two sets of benchmark instances were generated. The first set, labelled SET1, consists of artificial problem instances. Each of these instances consists of two artificially generated RNA strings of length  $n \in \{100, 200, \dots, 900, 1000\}$ . The probability of each letter and each position was chosen to be  $1/4$ . Moreover, for each input string we randomly generated a number of  $n_{\text{arcs}} \in \{n/10, n/5, n/2\}$ . Hereby, it was taken care that all  $n_{\text{arcs}}$  arcs were different. For each combination of  $n$  and  $n_{\text{arcs}}$  we randomly generated 30 problem instances. This makes a total of 900 problem instances.

For the second benchmark set, labelled SET2, we downloaded arc-annotated RNA sequences from the RNase P Database [3]. In total we assembled 10 problem instances, whose characteristics are described in Table 1. Moreover, the secondary structures of the RNA sequences involved in instances Real.1 and Real.8 are exemplary shown in Figure 3.

## 6.2 Algorithm Tuning

The automatic configuration tool *irace* [10] was used for tuning the parameters of HYB-EA. The following parameters of HYB-EA were considered for tuning: ( $n_{\text{sols}}$ ) the number of solution constructions per iteration ( $d_{\text{rate}}$ ) the determinism rate, ( $l_{\text{size}}$ ) the candidate list size, and ( $t_{\text{max}}$ ) the maximum time in seconds allowed for solution merging (at each call of the solution merging procedure). In particular, HYB-EA was tuned separately for each input string length, which—after initial experiments—seemed to have a greater influence on the behavior of the algorithm than the number of arcs. For each  $n \in \{100, 200, \dots, 900, 1000\}$  we randomly generated two tuning instances for each of the three values of  $n_{\text{arcs}}$ . This makes a total of six tuning instances for each  $n$ . The tuning process for each  $n$  was given a budget of 1000 runs of HYB-EA, where each run was given a computation time limit of  $n/10$  CPU seconds. Finally, the following parameter value ranges were considered concerning the four parameters of HYB-EA:

- $n_{\text{sols}} \in \{5, 10, 20\}$
- $d_{\text{rate}} \in \{0.0, 0.3, 0.5, 0.7, 0.9\}$ , where a value of 0.0 means that the selection of the assignment to be added to the partial solution under construction is always done randomly from the candidate list, while a value of 0.9 means that solution constructions are nearly deterministic.

Table 2: Results of tuning HYB-EA with irace.

$n$	$n_{\text{sols}}$	$d_{\text{rate}}$	$l_{\text{size}}$	$t_{\text{max}}$
100	10	0.3	2	5.0
200	5	0.7	3	1.0
300	5	0.7	2	5.0
400	5	0.7	3	10.0
500	5	0.3	2	20.0
600	5	0.7	2	5.0
700	5	0.5	2	20.0
800	5	0.7	2	5.0
900	5	0.5	2	5.0
1000	5	0.7	2	5.0

- $l_{\text{size}} \in \{1, 2, 3, 4\}$
- $t_{\text{max}} \in \{1.0, 5.0, 10.0, 20.0\}$  (in seconds).

The tuning runs with irace produced the configurations of HYB-EA as shown in Table 2. The following trends can be observed. Apart from  $n = 100$ , the number of solution constructions is always set to five. This is because the smaller  $n_{\text{sols}}$ , the smaller is the ILP model that has to be solved by CPLEX in the context of solution merging at each iteration of the algorithm. Moreover, the smaller the ILP model, the more efficient is CPLEX in solving such a model. The values of  $d_{\text{rate}}$  are consistently between 0.3 and 0.7, whereas the values of  $l_{\text{size}}$  are consistently set to two or three. Finally, the settings of  $t_{\text{max}}$  seem somewhat erratic. However, this is due to the fact that the application of CPLEX in solution merging is very efficient and stops, most of times, much below 5 CPU seconds. Therefore, it is only of importance that the value of  $t_{\text{max}}$  is at least set to 5.0.

### 6.3 Numerical Results

The results concerning the artificial problem instances from SET1 are presented in Table 3. Each row provides the results of HEURISTIC, MS-HEUR and HYB-EA in terms of the average solution quality obtained for the 30 problem instances of the corresponding combination of  $n$  and  $n_{\text{arcs}}$ . All techniques were applied exactly once to each problem instance. The computation time limit used for MS-HEUR and HYB-EA was  $n/10$  CPU seconds. The column with heading **time** shows the average computation time for the 30 problem instances in the case of HEURISTIC, and the average time at which the best solution of a run was found, in the case of MS-HEUR and HYB-EA. The best result of each table row is marked by a lightgrey background. The following observations can be made:

- HEURISTIC is very fast. Its application to any of the problem instances requires less than one CPU second. Moreover, the results of HEURISTIC are always better than the results of MS-HEUR.
- HYB-EA is, by far, the best algorithm in the comparison. It obtains the best result for each combination of  $n$  and  $n_{\text{arcs}}$ . In particular, HYB-EA always improves over MS-HEUR. This means that the solution merging component is an essential part of HYB-EA. This is remarkable as the application of CPLEX to the original problem instances



Table 3: Experimental results concerning the artificial problem instances from SET1.

$n$	# arcs	HEURISTIC		MS-HEUR		HYB-EA	
		result	time	result	time	result	time
100	10	55.73	< 1	51.80	3.70	58.87	1.17
	20	51.63	< 1	49.23	4.10	57.00	1.52
	50	42.63	< 1	42.57	3.84	50.07	2.26
200	20	113.77	< 1	104.33	7.74	120.57	7.36
	40	104.13	< 1	97.67	11.06	114.97	7.22
	100	87.10	< 1	84.20	6.19	101.70	7.66
300	30	170.43	< 1	158.70	12.87	178.77	13.18
	60	156.87	< 1	149.80	11.55	171.47	12.08
	150	132.40	< 1	128.17	15.45	151.67	13.82
400	40	229.20	< 1	198.27	17.05	239.53	24.43
	80	210.93	< 1	185.23	17.37	228.97	27.74
	200	175.63	< 1	159.77	18.10	202.60	28.26
500	50	289.20	< 1	235.90	19.25	298.43	32.93
	100	263.80	< 1	222.83	22.02	285.83	33.89
	250	220.43	< 1	192.57	21.52	253.13	39.81
600	60	346.63	< 1	310.50	26.86	360.10	32.62
	120	317.93	< 1	290.03	26.07	343.70	44.54
	300	266.03	< 1	247.00	25.24	304.43	43.58
700	70	404.53	< 1	343.33	25.42	420.77	46.32
	140	369.97	< 1	320.97	28.49	398.37	51.93
	350	308.57	< 1	275.17	31.20	352.50	57.58
800	80	461.17	< 1	408.50	34.41	478.97	59.30
	160	424.57	< 1	380.97	35.10	458.10	61.30
	400	353.57	< 1	326.33	36.55	403.73	65.53
900	90	520.87	< 1	435.33	39.06	537.53	72.56
	180	477.90	< 1	409.20	44.19	513.90	71.06
	450	398.97	< 1	350.83	48.57	453.20	75.98
1000	100	578.30	< 1	507.60	42.37	599.63	76.23
	200	531.33	< 1	473.40	43.07	570.93	78.28
	500	443.33	< 1	404.23	35.12	505.37	84.39

was not viable at all. Nevertheless, in the context of solution merging, mathematical programming plays an important role for the success of the HYB-EA algorithm.

In order to study the magnitude of the improvement of HYB-EA over MS-HEUR, we show the percentage improvement of HYB-EA over MS-HEUR—by means of boxplots—for each combination of  $n$  and  $n_{\text{arcs}}$  in the three graphics of Figure 2. The x-axis of each graphic ranges from  $n = 100$  to  $n = 1000$ . These graphics show, first, that the improvement of HYB-EA over MS-HEUR is generally around 15 – 20%. Moreover, it can be observed that the improvements become bigger with a growing number of arcs.

In a second set of experiments we applied the three algorithms considered in this work to the set of 10 real problem instances (SET2). The results are provided in Table 4, in

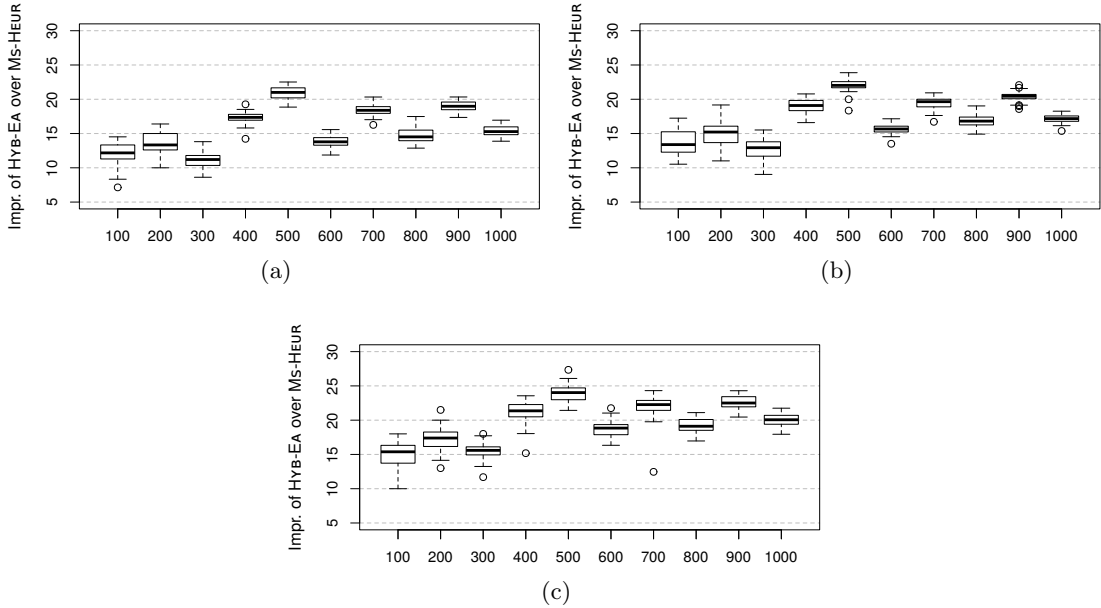


Figure 2: Improvement of HYB-EA over MS-HEUR (in percent). Each box shows the differences for the corresponding 30 instances. (a) Instances with  $n_{\text{arcs}} = n/10$ , (b) instances with  $n_{\text{arcs}} = n/5$ , (c) instances with  $n_{\text{arcs}} = n/2$ .

the following way. HEURISTIC was applied exactly once to each problem instance. The corresponding result can be found in columns with headings **result** and **time**. Both MS-HEUR and HYB-EA are applied 30 times to each problem instance, with a computation time limit of 30 seconds per application. In both cases we provide the best result obtained over 30 applications (column **best**), the average of the 30 results obtained (column **avg.**), and the average time at which the best solution of each run was found (column **time**). Again, the best result for each of the 10 instances is marked by a lightgrey background. Apart from one case (Real\_4), the results are consistent with our observations in the context of the artificial benchmark instances. In the case of Real\_4, Heuristic outperforms HYB-EA. This suggests that this problem instance has certain characteristics that are difficult to be discovered by the step-by-step way of constructing solutions as used in HYB-EA. Remember that HEURISTIC, in contrast, uses the dynamic programming algorithm by Smith and Waterman in order to generate the (possibly invalid) initial solution.

## 7 Conclusion

In this paper we proposed a simple, hybrid evolutionary algorithm for solving the so-called longest arc-preserving common subsequence problem. The most important feature of this algorithm is a crossover component based on solution merging. At each iteration, the best solution found so far is merged with randomly generated solutions, and a general purpose integer linear programming solver is used to find the best solution within the resulting set of assignments. The results show that the algorithm is superior to the only existing heuristic from the literature. Moreover, we have shown that the solution merging component is an

Table 4: Experimental results concerning the real problem instances from SET2.

inst.	HEURISTIC		MS-HEUR			HYB-EA		
	result	time	best	avg.	time	best	avg.	time
Real_1	238	< 1	181	178.30	22.34	247	237.70	22.20
Real_2	260	< 1	206	199.10	17.27	285	283.40	23.12
Real_3	265	< 1	222	218.80	16.67	288	280.30	26.79
Real_4	373	< 1	298	295.00	16.51	369	369.00	7.59
Real_5	152	< 1	133	131.10	13.87	175	174.30	14.96
Real_6	183	< 1	165	160.80	9.75	208	203.90	20.42
Real_7	316	< 1	248	240.90	12.12	328	326.60	17.89
Real_8	153	< 1	141	138.50	11.78	174	170.00	24.82
Real_9	285	< 1	231	223.60	13.05	299	297.50	22.04
Real_10	343	< 1	276	267.60	16.24	352	351.50	12.05

essential part of the algorithm.

In future work we will try to replace the probabilistic way of constructing solutions by a probabilistic version of the Smith and Waterman algorithm. In this way it might be possible to avoid situations such as the one for real-life instance Real\_4, where our algorithm was not able to outperform the existing heuristic.

## Acknowledgment

This work was funded by project TIN2012-37930-C02-02 (Spanish Ministry for Economy and Competitiveness, FEDER funds from the European Union) and project SGR 2014-1034 (AGAUR, Generalitat de Catalunya). Our experiments have been executed in the High Performance Computing environment managed by the RDLab at the Technical University of Barcelona (<http://rdlab.cs.upc.edu>) and we would like to thank them for their support.

## References

- [1] C. Aggarwal, J. Orlin, and R. Tai. Optimized crossover for the independent set problem. *Operations Research*, 45:226–234, 1997.
- [2] C. Blum, M. J. Blesa, and M. López-Ibáñez. Beam search for the longest common subsequence problem. *Computers & Operations Research*, 36(12):3178–3186, 2009.
- [3] J. W. Brown. The ribonuclease P database. *Nucleic Acids Research*, 27(1):314–314, 1999.
- [4] Jimmy Ka Ho Chiu and Yi-Ping Phoebe Chen. A comprehensive study of RNA secondary structure alignment algorithms. *Briefings in Bioinformatics*, page bbw009, 2016.
- [5] Patricia A. Evans. Finding common subsequences with arcs and pseudoknots. In Maxime Crochemore and Mike Paterson, editors, *Proceedings of CPM 1999 – 10th Annual Symposium on Combinatorial Pattern Matching*, volume 1645 of *Lecture Notes in Computer Science*, pages 270–280. Springer Berlin Heidelberg, 1999.

- [6] Patricia Anne Evans. *Algorithms and Complexity for Annotated Sequence Analysis*. PhD thesis, University of Victoria, 1999.
- [7] C. B. Fraser. *Subsequences and supersequences of strings*. PhD thesis, University of Glasgow, 1995.
- [8] K. Huang, C. Yang, and K. Tseng. Fast algorithms for finding the common subsequences of multiple sequences. In *Proceedings of the 2004 International Computer Symposium*, pages 1006–1011. IEEE press, 2004.
- [9] Tao Jiang, Guohui Lin, Bin Ma, and Kaizhong Zhang. The longest common subsequence problem for arc-annotated sequences. *Journal of Discrete Algorithms*, 2(2):257–270, 2004.
- [10] M. López-Ibáñez, J. Dubois-Lacoste, L. Pérez Cáceres, M. Birattari, and T. Stützle. The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43 – 58, 2016.
- [11] D. Maier. The complexity of some problems on subsequences and supersequences. *Journal of the ACM*, 25:322–336, 1978.
- [12] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization*. Wiley & Sons, 1988.
- [13] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

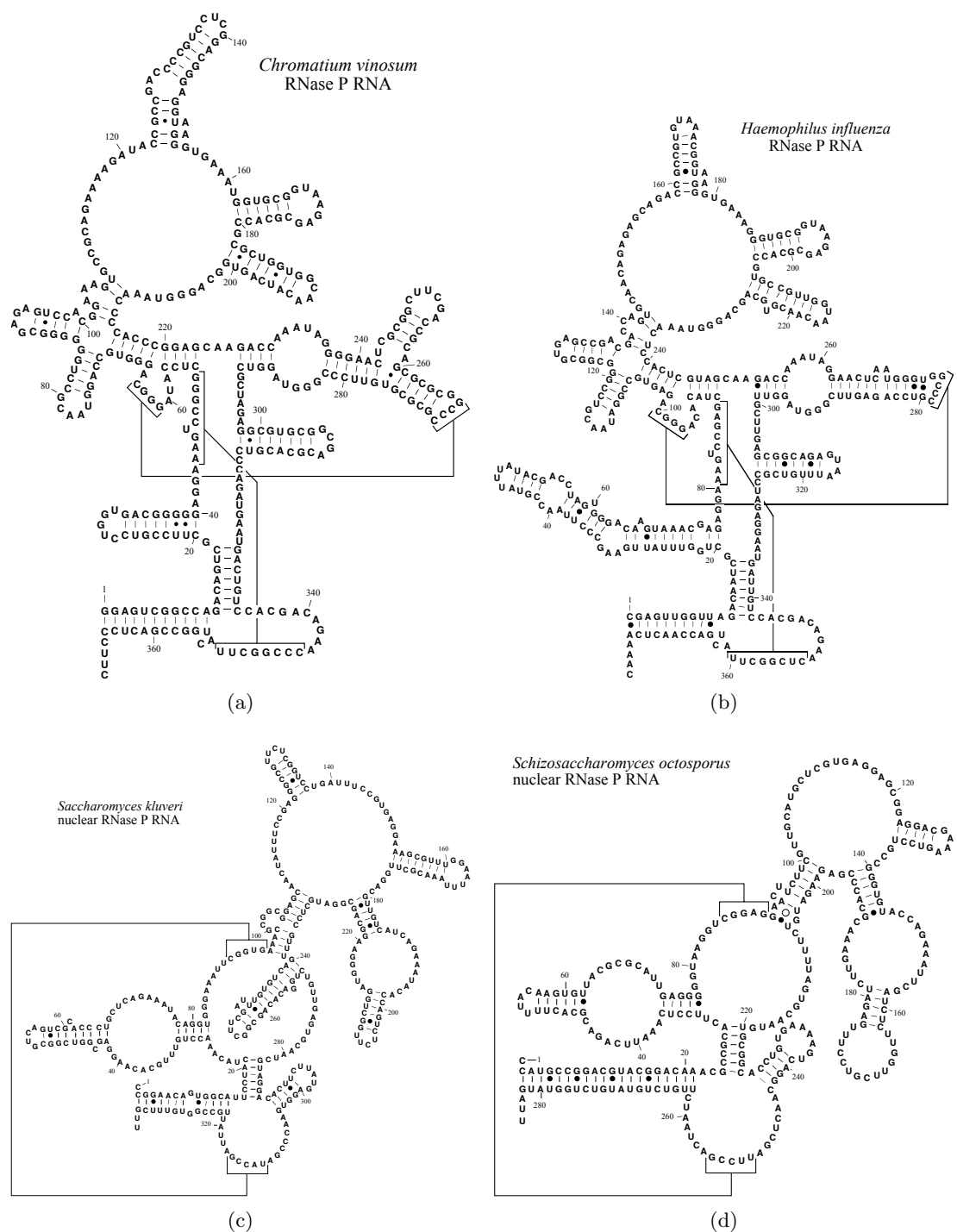


Figure 3: Secondary structure of the two RNA sequences involved in instances Real.1 (a and b) and Real.8 (c and d). All graphics were downloaded from the RNase P Database [3] (a) RNA of *Allochrochromatium vinosum*, (b) RNA of *Haemophilus influenza*, (c) RNA of *Saccharomyces kluyveri*, (d) RNA of *Schizosaccharomyces octosporus*.