Tabu Search Algorithm for RNA Degradation Problem

Agnieszka Rybarczyk, Marta Kasprzak, Jacek Blazewicz
Institute of Computing Science
Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland
and
Institute of Bioorganic Chemistry
Polish Academy of Sciences
Noskowskiego 12/14, 61-704 Poznan, Poland
Email: arybarczyk@cs.put.poznan.pl
Email: mkasprzak@cs.put.poznan.pl
Email: jblazewicz@cs.put.poznan.pl

***Abstract*—In the last few years, there has been a great interest in the RNA (ribonucleic acid) research due to the discovery of the role that RNA molecules play in biological systems. They do not only serve as a template in protein synthesis or as adaptors in translation process but also influence and are involved in the regulation of gene expression. It was demonstrated that most of them are produced from larger molecules due to enzyme cleavage or spontaneous degradation. In this work, we would like to present our recent results concerning the RNA degradation process. In our studies, we used artificial RNA molecules designed according to the rules of degradation developed by Kierzek and co-workers. On the basis of the results of their degradation, we have proposed the formulation of the RNA Partial Degradation Problem (RNA PDP) and we have shown that the problem is strongly NP-complete. We would like to propose a new efficient heuristic algorithm based on tabu search approach which allows us to reconstruct the cleavage sites of the given RNA molecule.**

Keywords–RNA degradation; tabu search; computational complexity.

I. INTRODUCTION

For the last few years, there has been an increased interest in RNA because of its involvement in controlling gene expression. The sequencing projects and analysis of higher eukaryotic genomes have revealed that in contrast to prior expectations, only a small fraction of the genetic material codes for proteins. It has been demonstrated that the vast majority of genomes of complex organisms are transcribed into an abundance of non-protein coding RNA molecules that perform not only housekeeping but also regulatory functions in the cell. It became evident that not only large RNAs (e.g., messenger (mRNA) or transfer RNA (tRNA)) are responsible for proper functioning of living organisms. There exist plenty of smaller RNAs, called small regulatory RNA, which are involved in the cellular processes. In contrast to housekeeping RNAs (like tRNA, rRNA (ribosomal RNA), tmRNA (transfer-messenger RNA), snRNA (small nuclear RNA)), which are constitutively expressed, regulatory RNAs exhibit the expression pattern dependent on organism's developmental stage, cell differentiation or the interaction with the external environment [1].

Small RNAs are 19-28 nucleotide sequences, which are produced through enzyme cleavage or spontaneous degradation of large RNA molecules. The nature of the latter mechanism is now heavily investigated by the researchers [2].

Numerous studies demonstrated that all components of the degradation machinery affect the RNA molecules in the same manner, but some RNAs apparently are more stable than others e.g., in mammalian cells c-fos mRNA lasts 15 minutes while β–globuline over 24 hours [3]. It has been noted that there exist many sequence elements, which can influence the mRNA stability. Those sequences act through stimulating or inhibiting the degradation of mRNA and, e.g., in mammalian cells were identified as rich in adenosine and uridine elements (AURE). The mRNA stability is strongly dependent on the location of AURE elements within the molecule. It is suggested that the spatial structure combined with the sequence motifs decides that the degradation process performs differently in case of distinct RNAs.

In this paper, we would like to report our recent results concerning the degradation of RNA molecules. In our studies we have used artificial RNA molecules, which were designed in such a way that they should be unstable, according to the rules developed by Kierzek and co-workers [4]. On the basis of the results of their degradation, we have proposed a formulation of a new problem, called RNA PDP and we have shown that the problem is strongly NP-complete [5]. Here, we would like to propose a new efficient heuristic algorithm based on tabu search approach that allows to reconstruct the cleavage sites of the given RNA molecule.

The organization of the paper is as follows. In Section 2, the combinatorial model of the decision version of RNA PDP problem is presented. Section 3 introduces the tabu search algorithm for RNA PDP problem. In Section 4, the results of the computational tests are given, while Section 5 points out the directions for further research.

II. PROBLEM FORMULATION

To analyze the degradation process dependent on RNA structure, we carried out two types of experiments [5]: involving multi-labeled and single-labeled RNA. The aim of the first type of experiment was to visualize all fragments generated during degradation. In this case, the exact location of the fragments within analyzed RNA molecule is missing. The second type of experiment was carried out to visualize only those fraction of degradation products, which contained labeled 5' end of the RNA molecule. In this case, their location within the tested molecule is known. In this way, two collections of fragments are created. Each fragment is represented by its

 ISBN: 978-1-61208-409-1

length.

We will denote as $D = \{d_1, \ldots, d_k\}$ the multiset of fragments (lengths) obtained during the multi-labeled RNA degradation and as $Z = \{z_1, \ldots, z_n\}$ the set of fragments (lengths) coming from the single-labeled RNA degradation. Moreover, we will distinguish between two types of cleavage sites: primary and secondary. Primary cleavage sites occur only within input RNA molecule of the full length while secondary cleavage sites only within lengths obtained as a result of primary cleavages. The computational phase of the reconstruction of the cleavage sites of the given RNA molecule is strongly NP–hard [5]. Hence, there is a need for developing an efficient heuristic.

The mathematical formulation of the RNA PDP is presented below [5]. P_1 stands for the set of primary cleavage sites in the solution and P_2 for the set of secondary ones.

Problem 1: RNA PDP — decision version (Π_{RNAPDP}) .
Instance: Multiset $D = \{d_1, ..., d_k\}$ and set $Z = \{z_1, ..., z_n\}$ of positive integers, positive integer L, constant $C \in \mathbb{Z}^+ \cup \{0\}$.
Question: Do there exist sets P_1 and P_2 such that:

$$P_1 \cup P_2 = P = \{p_1, ..., p_m\}, \ \forall_{p_i \in P} \ 0 < p_i < L, \tag{1}$$

$$D \subseteq D', \ D' \supseteq R = \{p_i - p_j : \ p_i, p_j \in P_1 \cup \{0, L\} \ \wedge \ p_i > p_j\}, \tag{2}$$

$$D' \setminus R = \bigcup_{i=1}^{|T|} D'_i, \tag{3}$$

$$T = \{t_i = (p_a, p_b, p_c) : \ p_a, p_c \in P_1 \cup \{0, L\} \ \wedge \ p_b \in P_2 \ \wedge \ p_a < p_b < p_c \ \wedge \ d'_{i1} = p_b - p_a \ \wedge \ d'_{i2} = p_c - p_b \ \wedge \ \{d'_{i1}, d'_{i2}\} = D'_i\}, \tag{4}$$

$$\forall_{t_i, t_j \in T, \ t_i = (p_{ia}, p_{ib}, p_{ic}), \ t_j = (p_{ja}, p_{jb}, p_{jc})} \quad i \neq j \rightarrow \{p_{ia}, p_{ic}\} \neq \{p_{ja}, p_{jc}\}, \tag{5}$$

$$Z \subseteq Z', \ Z' \subseteq P \cup \{L\}, \ Z' \supseteq P_1 \cup \{L\}, \tag{6}$$

$$Z' \setminus [P_1 \cup \{L\}] = \{p_b : \ (p_a, p_b, p_c) \in T \ \wedge \ p_a = 0\}, \tag{7}$$

$$P_2 = \{p_b : \ (p_a, p_b, p_c) \in T\}, \tag{8}$$

$$|D'| + |Z'| \leq k + n + C \ ? \tag{9}$$

III. Tabu search algorithm

The heuristic algorithm that works for the case of RNAPDP problem with negative experimental errors (i.e., missing fragments in D and Z) presented in this section is based on tabu search metaheuristic, which is a kind of local search procedure. Its aim is to find the coordinates of primary and secondary cleavage sites in a given RNA molecule, taking negative errors into account. The algorithm is implemented in C programming language and runs in a Unix environment. The algorithm takes as an input the data containing fragment lengths obtained via the biochemical experiments, i.e., multiset D of k positive integers and set Z of n positive integers. The main algorithm consists of two parts: inner and outer tabu search method. The outer tabu search method was designed to find the coordinates of primary cleavage sites including negative errors. In this part, four kinds of moves are initially defined: add missing primary fragment to set Z, add missing secondary fragment to set D, prevent from considering element of Z as a primary cleavage site, consider element of Z as a primary cleavage site. The inner tabu search of the algorithm was designed to reconstruct the coordinates of the secondary cleavage sites including negative errors, basing on the results of the inner part. After performing a number of moves not leading to an improvement of the solution quality, the method is restarted, i.e., some randomly generated feasible solution becomes an initial solution. If a specified number of restarts is performed than the algorithm stops. By analyzing the obtained results, we noticed that the algorithm performs very efficiently and fast.

IV. Tabu search algorithm

In this section, results of the tests of the algorithm solving the RNA PDP problem in the case of erroneous data are presented. The algorithm has been tested on PC with Pentium(R) 4, 2.40 GHz processor and 1 GB RAM in Unix environment. As a testing set, a group of randomly generated data was prepared (See [5] for details). Table I summarizes exemplary average running time results for the proposed tabu search algorithm tested on the random instances. In the results, each entry corresponds to 100 instances of the same number of secondary and primary cleavage sites and number of negative errors, i.e., to 100 runs of the algorithm.

TABLE I. Average computational times for randomly generated erroneous instances. Secondary cleavage sites occur in 75% of all primary fragments, the number of the latter being equal to $\binom{r+2}{2}$, where r is the number of primary cleavage sites in the instance. Additionally, the input data set was separately tested with the number of missing fragment lengths ranging from 1 to 5.

No. of reconstructed primary cleavage sites	No. of reconstructed secondary cleavage sites	Average computational time [s] Number of negative errors				
		1	2	3	4	5
4	10	0.05	0.06	0.05	0.05	0.04
5	15	0.40	0.54	0.44	0.41	0.41
6	20	0.30	0.33	0.34	0.33	0.31
7	26	0.82	0.93	0.77	0.85	0.90
8	33	1.18	1.34	1.37	1.35	1.34
10	49	4.71	5.26	5.40	5.48	5.43
12	68	16.44	19.21	19.72	20.12	20.00
14	90	57.13	61.23	63.99	64.95	64.82

Tabu search algorithm was also tested on 3630 randomly generated instances without negative errors and was able to find optimal solution in 3578 cases (98.57% of the whole data).

V. Conclusion

The computational phase of the reconstruction of the cleavage sites in RNA PDP is a computationally hard problem if there are errors in the input data. Hence, the need of developing good polynomial time heuristics arises. In this work, an algorithm based on tabu search approach has been presented and its effectiveness has been tested. The computational tests confirmed high efficiency of the proposed algorithm. This algorithm may be very useful in practice as a tool that facilitates the analysis of the output of the biochemical experiment.

References

[1] M. Szymanski, M. Barciszewska, M. Zywicki, and J. Barciszewski, "Noncoding RNA transcripts," Journal of Applied Genetics, vol. 44, 2003, pp. 1–19.

[2] M. Nowacka and et al., "2D-PAGE as an effective method of RNA degradome analysis," Mol Biol Rep, vol. 39, 2011, pp. 139–146.

[3] M. Deutscher, "Degradation of Stable RNA in Bacteria," Journal Biol. Chem., vol. 278, 2003, pp. 45 041–45 044.

[4] R. Kierzek, "Nonenzymatic Cleavage of Oligoribonucleotides," Methods Enzymol., vol. 341, 2001, pp. 657–675.

[5] J. Blazewicz, M. Figlerowicz, M. Kasprzak, M. Nowacka, and A. Rybarczyk, "RNA Partial Degradation Problem: motivation, complexity, algorithm," J Comput Biol, vol. 18, 2011, pp. 821–834.

 ISBN: 978-1-61208-409-1