RNA Partial Degradation Problem: motivation, complexity, algorithm

Jacek Blazewicz[†‡], Marek Figlerowicz[‡], Marta Kasprzak[†‡],
Martyna Nowacka[‡] and Agnieszka Rybarczyk[*†]

Abstract

Studies conducted during the last decade unexpectedly revealed several new biological functions of RNA molecules. The involvement of RNA in many complex processes requires highly effective systems controlling its accumulation. In this context, the mechanisms of degradation appear as one of the most important factors influencing RNA activity. Here, we present our first attempt to describe the RNA degradation process using bioinformatics methods. Based on the obtained data, we propose a formulation of a new problem, called RNA Partial Degradation Problem (RNA PDP) and the algorithm that is capable of reconstructing an RNA molecule using the results of biochemical analysis of its degradation. In addition, we present the results of biochemical experiments and computational tests.

Key words: RNA degradation, nonenzymatic hydrolysis, computational complexity, branch-and-cut algorithms.

[†] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland; phone: +48 61 8790790; fax: +48 61 8771525; E-mail: jblazewicz@cs.put.poznan.pl, mkasprzak@cs.put.poznan.pl, arybarczyk@cs.put.poznan.pl

[‡] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland; phone: +48 61 8528503; fax: +48 61 8520532; E-mail: jblazewicz@cs.put.poznan.pl, marekf@ibch.poznan.pl, mkasprzak@cs.put.poznan.pl, martn@ibch.poznan.pl.

[*] To whom correspondence should be directed. E-mail: arybarczyk@cs.put.poznan.pl.

1 Introduction

Studies conducted during the last decade clearly demonstrated that RNA molecules play much more important role in all biological systems than it has been earlier expected. They are not only templates for protein synthesis (mRNA), adaptors translating information encoded in nucleotide sequence into amino acid chain (tRNA), skeletons joining multi-enzymatic complexes (rRNA, snRNA) but also regulators of numerous processes including the gene expression process. Recently, several different classes of regulatory RNA molecules have been identified. Usually, they bind with proteins and function as specific probes targeting enzymes into RNA or DNA molecules. Depending on the localization and composition of ribonucleoprotein complexes they can participate in RNA degradation, translation repression or in the modification of genomic DNA.

The level of RNA accumulation is regulated by maintaining balance between transcription and degradation pathways. Consequently, the stability of RNA molecules is one of the major factors shaping the composition of cellular transcriptome. In eukaryotic cells, the degradation of mRNA is initiated by the removal of a poly(A) tail by deadenylases. Next, the RNA molecule is digested by 3'-5' exonucleases in exosome. In addition, DCP2 protein-mediated decapping causes mRNA to be a proper substrate for 5'-3' exoribonucleases (Eulalio et al. 2007). Many of the enzymes and cofactors involved in RNA degradation participate also in RNA-processing or maturation (Houseley et al. 2009). Although all RNA molecules present in a given cell are exposed to the same RNA degradation machinery some of them exist and function for a longer time than other. The half-life of mRNAs can range from hours, to seconds (Lorentzen et al. 2006). Unfortunately, at present our knowledge about different factors affecting RNA degradation is very limited. Accordingly, one cannot predict the stability of individual RNAs. Some short-living mRNAs have been shown to carry adenosine and uracil rich cis-acting elements (ARE) usually located in their 3' UTRs. These molecules undergo rapid degradation by the ARE-mediated mRNA decay (AMD) (Houseley et al. 2009). To avoid errors in RNA biogenesis or function, aberrant or nonfunctional RNAs are preferentially degraded by numerous quality control systems. One of them, nonsense-mediated decay (NMD), enables degradation of mRNAs carrying nonsense mutations (e.g. premature translation termination codon). Quality control systems are also involved in the repression of function or in degradation of viral and parasitic RNAs (Doma et al. 2007).

In contrast with mRNA, the degradation pathways of relatively stable molecules like tRNA and rRNA, accounting for more than 90% of total cellular RNA, have been poorly characterized. In bacteria, the degradation of stable RNAs is usually associated with starvation (Deutscher 2003). Recent reports have shown that rRNA and tRNAs can undergo cleavage in response to oxidative stress or under developmental regulation (Thompson et al. 2008). Now, the most important question concerns biological function of the cleavage products. There are some evidences suggesting that they may function as translation inhibitors or signaling molecules (Kierzek 1992, Zhang et al. 2009).

Better understanding of RNA degradation is indispensable not only for broadening our knowledge on the physiological functions of RNA molecules. It is also necessary if one wishes to modulate a half-life of any RNA molecule and in this way regulate its biological activity, e.g. by influencing the stability of mRNA one can affect the level of protein accumulation. There is a large number of factors that can affect RNA stability. Generally, they can be divided into two groups: (i) factors acting in trans - all cellular factors; and (ii) factors acting in cis - RNA primary, secondary and tertiary structure.

Although many RNAs are similarly accessible by the degradation machinery, some molecules are more stable than the others. It suggests that the primary, secondary and tertiary structure of RNA molecules significantly affect their degradation.

Unfortunately, currently available experimental methods do not allow for studying all aspects of this process. Instead, experiments in model *in vitro* systems can be performed in well-defined conditions.

In this paper, we present our first attempt to describe the RNA degradation process using bioinformatics methods. Our studies were focused on cis acting factors affecting RNA stability and involved two artificial RNA molecules. Both of them were designed in such a way that they contained well defined unstable regions (Bibillo et al. 1999, Bibillo et al. 2000, Kierzek 2001). The undertaken biochemical and bioinformatics analyses confirmed the predicted pattern of RNA degradation. Based on the obtained data, we proposed a formulation of a new problem, called RNA Partial Degradation Problem (RNA PDP) and the algorithm that is capable of reconstructing RNA molecule using results of biochemical analysis of its degradation. By solving RNA PDP problem one can determine the location of cleavage sites within analyzed RNA molecule having as an input the lengths of fragments being the degradation products. As a result one can identify the unstable regions of RNA molecules which are the most susceptible to the degradation.

The organization of the paper is as follows. Section 2 discusses the combinatorial model and gives the strong NP-completeness proof of the decision version of RNA PDP problem which is equivalent to a non-existence of a polynomial-time exact algorithm for the problem in question (Garey et al. 1979). Section 3 presents the exact algorithm for RNA PDP based on the branch-and-cut idea. In Section 4 the results of the biochemical experiment and computational tests are given, while Section 5 points out the directions for further research.

2 Problem formulation and complexity

To analyze the degradation process dependent on cis acting elements (RNA structure) the following two types of experiments have been carried out: (i) involving multi-labeled RNA - ^{32}P labeled nucleotides were randomly introduced along the RNA molecule; and (ii) involving single labeled RNA - ^{32}P was introduced at the 5' terminus of the tested RNA. The first type of experiment was carried out to visualize all fragments generated during degradation. In this case the information of exact position of the fragments within analyzed molecule is missing. The second type of experiment was carried out to visualize only these degradation products which contained labeled 5' end. In this case their location within tested molecule is known. Thus, we see that two collections of fragments are created. Each fragment is represented by its length. For further mathematical considerations we will use the lengths as specified parameters of fragments. Let $D = \{d_1, \ldots, d_k\}$ be a multiset of fragments (lengths) obtained during the multi-labeled RNA degradation and let $Z = \{z_1, \ldots, z_n\}$ be a set of fragments (lengths) coming from the single-labeled RNA degradation. Furthermore, we will distinguish between two types of cleavage sites: primary and secondary. **To classify any fragment as a product of the primary cleavage we should observe it during a single-labeled RNA analysis. In addition, the same fragment and its complement corresponding to the remaining portion of the full length molecule have to be observed during multi-labeled RNA analysis (See Figure 3).** 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 lengths of the fragments define the distances between cleavage sites. As the effect of primary cleavages, all distances between the primary cleavage sites are acquired. In the case of secondary cleavages a pair of lengths will be obtained, which sum is equal to one of the primary length. A primary fragment is assumed to cleave at most once. i.e. it may produce two secondary fragments or none. It is easy to see that (in the ideal case) multiset D is composed of all pairwise distances between the primary cleavage sites and ends of the molecule, together with all fragment lengths obtained from secondary cleavages. Set Z contains (in the ideal case) positions of all primary cleavage sites and also it contains these secondary fragments which 5' end is labeled. Our aim is to reconstruct the coordinates of primary and secondary cleavage sites within input RNA molecule, basing on the data gathered from the biochemical experiment.

With respect to the above description of the problem, we propose its mathematical model that will serve as a background for the complexity analysis and for the construction of the algorithm that solves the problem.

The mathematical formulation of the RNA Partial Degradation Problem is presented below. 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 Partial Degradation Problem — 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:

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

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

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

(4) $T = \{t_i = (p_a, p_b, p_c) :\ \ p_a, p_c \in P_1 \cup \{0, L\}\ \ \wedge\ \ p_b \in P_2\ \ \wedge\ \ p_a < p_b < p_c$

$\wedge\ \ d'_{i1} = p_b - p_a\ \ \wedge\ \ d'_{i2} = p_c - p_b\ \ \wedge\ \ \{d'_{i1}, d'_{i2}\} = D'_i\},$

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

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

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

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

(9) $|D'| + |Z'| \leq k + n + C$?

Here, the general version of the problem is assumed, where missing elements, i.e. *false negatives*, in D and Z are allowed. Π_{RNAPDP} is the decision version of the optimization problem, in which the total number of false negatives is minimized. In the formulation, D' and Z' are the complete counterparts of the sets, and C is the error bound. Assigning C to 0 we get the *ideal problem* with no errors allowed.

In this section, we prove strong NP-completeness of problem Π_{RNAPDP} by a pseudo-polynomial transformation from problem Numerical Matching With Target Sums, cited below (of course, the strong NP-completness of Π_{RNAPDP} implies strong NP-hardness of the search version of the RNA PDP). This guides a choice of the proper solution strategy for the latter: either an exact exhaustive search algorithm or a polynomial-time heuristic strategy, not guaranteed to find an exact solution.)

Problem 2. Numerical Matching With Target Sums (Π_{NM}) — decision version.

Instance: Disjoint sets X and Y, each containing q elements, sizes $s(x_i)$ and $s(y_i)$ for every $x_i \in X$ and $y_i \in Y$, and a target vector $[u_1, ..., u_q]$; $s(x_i)$, $s(y_i)$, and u_i being positive integers, $1 \leq i \leq q$.

Question: Can $X \cup Y$ be partitioned into q disjoint sets $W_1, ..., W_q$, each containing exactly one element from X and one element from Y, such that $\sum_{w \in W_i} s(w) = u_i$, $1 \leq i \leq q$?

This problem is strongly NP-complete (Garey et al. 1979).

The transformation proposed below in its first step (modification of Π_{NM}) uses some ideas of our former proof for a distinct problem of the Simplified Partial Digest (SPDP) of DNA molecules (Blazewicz et al. 2005) (cf. also Waterman 1995, Blazewicz et al. 2001, Blazewicz et al. 2007). For the moment, the dependence between somewhat similar problems: RNA PDP and PDP as well as SPDP, is not yet found out, being an interesting question for further studies.

The first stage of proving strong NP-completeness of problem Π_{RNAPDP} consists in a simple modification of problem Π_{NM}. We are interested in a variant with the ranges of variables shifted by some added values. The modified ranges will have several properties, in particular, they will be disjoint. Initially, the ranges of values of variables $s(x_i)$, $s(y_i)$, and u_i, $1 \leq i \leq q$, can be written as follows:

$$\begin{aligned} s(x_i) &\in \langle x_{\text{L}}, x_{\text{R}} \rangle, \\ s(y_i) &\in \langle y_{\text{L}}, y_{\text{R}} \rangle, \\ u_i &\in \langle u_{\text{L}}, u_{\text{R}} \rangle, \end{aligned}$$

where the variables with index L mean the smallest values and the ones with index R mean the largest values in the respective collections. These ranges are arbitrary, however, we assume here that they satisfy few obvious conditions:

$$\begin{aligned} u_{\text{R}} &\leq x_{\text{R}} + y_{\text{R}}, \\ u_{\text{L}} &\geq x_{\text{L}} + y_{\text{L}}, \\ u_{\text{R}} &> x_{\text{R}}, \\ u_{\text{R}} &> y_{\text{R}}. \end{aligned}$$

The above assumptions do not change the computational complexity of problem Π_{NM} — if one of them is not satisfied, the problem becomes easy because the answer is obviously "no".

We modify the initial ranges of the variables in the following way (where "$\leftarrow$" assigns the right-hand side value to the left-hand side variable):

$$\begin{aligned} s(x_i) &\leftarrow s(x_i) + u_{\text{R}}, & 1 \leq i \leq q, \\ s(y_i) &\leftarrow s(y_i) + 2x_{\text{R}} + 2u_{\text{R}}, & 1 \leq i \leq q, \\ u_i &\leftarrow u_i + 2x_{\text{R}} + 3u_{\text{R}}, & 1 \leq i \leq q. \end{aligned}$$

The new ranges have the following form ($1 \leq i \leq q$):

$$\begin{aligned} s(x_i) &\in \langle x_{\text{L}} + u_{\text{R}}, x_{\text{R}} + u_{\text{R}} \rangle, \\ s(y_i) &\in \langle y_{\text{L}} + 2x_{\text{R}} + 2u_{\text{R}}, y_{\text{R}} + 2x_{\text{R}} + 2u_{\text{R}} \rangle, \\ u_i &\in \langle u_{\text{L}} + 2x_{\text{R}} + 3u_{\text{R}}, 4u_{\text{R}} + 2x_{\text{R}} \rangle. \end{aligned}$$

They are visualized in Figure 1.

As a result we have:

Lemma 1. (Blazewicz et al. 2005) *Problem* Π_{NM} *and its version with the ranges of variable values shifted as above, are equivalent.*

The variables from problem Π_{NM} increased by the proposed values have the following useful proper-

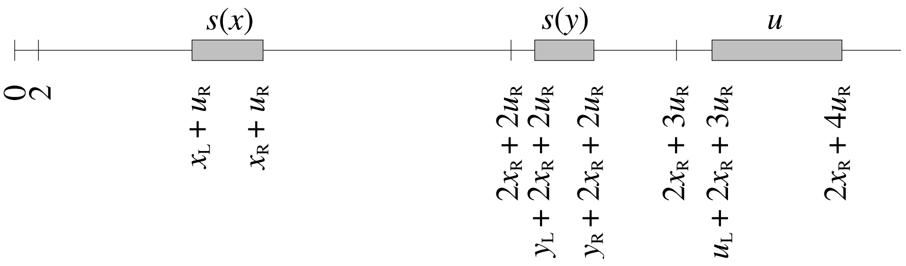

Figure 1: The ranges of values of variables $s(x_i)$, $s(y_i)$, and u_i, $1 \leq i \leq q$, after the modification.

ties.

Lemma 2. (Blazewicz et al. 2005) *None of the modified u_i, $1 \leq i \leq q$, can be equal to some $s(x_j)$ or to a sum of any $s(x_j)$ and $s(x_k)$.*

Lemma 3. (Blazewicz et al. 2005) *None of the modified u_i, $1 \leq i \leq q$, can be equal to some $s(y_j)$ or to a sum of any $s(y_j)$ and $s(y_k)$.*

Now, we can define the transformation from problem Π_{NM} to problem Π_{RNAPDP}, which is given below.

The transformation

Given an instance of problem Π_{NM}, the corresponding instance of Π_{RNAPDP} is constructed as follows.

(1) Shift the ranges of numbers in problem Π_{NM} as specified above, i.e.

$$\begin{aligned} s(x_i) &\leftarrow s(x_i) + u_{\text{R}}, & 1 \leq i \leq q, \\ s(y_i) &\leftarrow s(y_i) + 2x_{\text{R}} + 2u_{\text{R}}, & 1 \leq i \leq q, \\ u_i &\leftarrow u_i + 2x_{\text{R}} + 3u_{\text{R}}, & 1 \leq i \leq q. \end{aligned}$$

From now on all the variables have these modified values, if not stated otherwise.

(2) Create set $Z = \{\sum_{j=0}^{i} u_j : 0 \leq i \leq q\}$, where u_0 is a new variable of value 1. Add values $s(x_i)$ and $s(y_i)$, $1 \leq i \leq q$, to (initially empty) multiset D. Also add to D multiset $\{z_i^+ - z_j^+ : z_i^+, z_j^+ \in Z \cup \{0\} \ \wedge \ z_i^+ > z_j^+\}$. Now $n = q + 1$, $k = 2q + \binom{q+2}{2}$.

(3) Assign values to L and C: $L = \max(Z)$, $C = 0$.

Lemma 4. *The proposed transformation can be computed in time bounded by a polynomial in the length of the instance of Π_{NM} (Len_{NM}) and the maximal number appearing in this instance (Max_{NM}).*

Proof. Len_{NM} is $\text{O}(q\lceil \log \text{Max}_{\text{NM}} \rceil)$. In the first step of the transformation we make $\text{O}(\text{Len}_{\text{NM}})$ operations. New values of the variables do not change Len_{NM} and Max_{NM} substantially: q is not changed, new $\mathrm{Max}_{\mathrm{NM}}$ is up to 6 times larger than previously. Filling set Z requires $\mathrm{O}(\mathrm{Len}_{\mathrm{NM}}{}^2)$ operations. Filling multiset D requires $\mathrm{O}(\mathrm{Len}_{\mathrm{NM}}{}^2\lceil\log \mathrm{Len}_{\mathrm{NM}}\rceil)$ operations. Step (3) is $\mathrm{O}(\mathrm{Len}_{\mathrm{NM}}\lceil\log \mathrm{Len}_{\mathrm{NM}}\rceil)$. Taking the above functions together, we have $\mathrm{O}(\mathrm{Len}_{\mathrm{NM}}{}^2\lceil\log \mathrm{Len}_{\mathrm{NM}}\rceil)$ as the complexity of the proposed transformation. □

Because the instance size does not decrease exponentially and the greatest number in the instance does not increase exponentially, this transformation is pseudo-polynomial. We can now prove the following main theorem.

Theorem 1. *RNA Partial Degradation Problem (Π_{RNAPDP}) is strongly NP-complete.*

Proof. Lemma 4 proves that the proposed transformation is pseudo-polynomial. It remains to prove that the transformation is correct, i.e. that for every instance I of problem Π_{NM}, $I \in \mathrm{Y}_{\Pi_{\mathrm{NM}}}$ if and only if $\tau(I) \in \mathrm{Y}_{\Pi_{\mathrm{RNAPDP}}}$, where Y means the set of instances of a problem with answer "yes" and τ means the transformation. The first step of the transformation slightly modifies problem Π_{NM}, but both versions are equivalent (see Lemma 1). In the following, the shifted ranges of values of the variables are used.

Let us assume, that an instance of problem Π_{NM} gives a positive answer. It means, that there is a partition of $X \cup Y$ such that every disjoint subset W_i, $1 \leq i \leq q$, contains one x_j and one y_l, and $s(x_j)+s(y_l) = u_i$, for some $j, l \in \langle 1, q\rangle$. Then, the solution of problem Π_{RNAPDP} in the instance after the transformation can be constructed by assigning to sets P_1 and P_2 the following values: $P_1 \leftarrow Z \setminus \{L\}$, $P_2 \leftarrow \{s(x_j) + 1 + \sum_{l=1}^{i-1} u_l \colon 1 \leq i \leq q \quad \wedge \quad x_j \in W_i\}$. Such solution always satisfies all 9 constraints from the definition of problem Π_{RNAPDP}. Set P will be composed of elements being in the range $(0, L)$ (constraint 1). The construction of multiset D in the second step of the transformation first adds all elements of $D' \setminus R$ and next all elements of R. Elements of multiset R can be ordered to cover P_1, what follows from the construction of R (all interpoint distances between pairs of values from $P_1 \cup \{0, L\}$). Elements of multiset $D' \setminus R$ can be paired into $\{d'_{i1}, d'_{i2}\}$ on the base of the solution of Π_{NM}: $d'_{i1} = s(x_j)$, $d'_{i2} = s(y_l)$, where $x_j, y_l \in W_i$, what follows from the construction of D and P_2 (constraints 2-5). All secondary cleavage sites in these pairs from $D' \setminus R$ are different and cover P_2 (constraint 8). Set Z, constructed by the transformation is equal to $P_1 \cup \{L\}$ and does not contain any element of P_2, and because no element of P_2 comes from a cleavage of a primary segment beginning in 0, $Z = Z'$ (constraints 6-7). Constructed multiset D is also complete with reference to the corresponding solution P (i.e. $D = D'$), thus finally, constraint 9 from the definition of Π_{RNAPDP} is also satisfied.

Now let us assume, that an instance of problem Π_{RNAPDP} (after the transformation) has answer "yes". In such a case we have sets P_1 and P_2 satisfying all 9 constraints from the definition of Π_{RNAPDP}. We know, that constructed sets Z and D do not contain errors (because $C = 0$), so $Z = Z'$ and $D = D'$. Set $Z \setminus \{L\}$ indicates q cleavage sites of a molecule of length L, which define segments of lengths u_i, $0 \leq i \leq q$. The instance of Π_{RNAPDP} contains in multiset D $2q$ values less than any u_i, $i > 0$, among them q values from the range of $s(x)$ and q values from the range of $s(y)$ (see Fig. 1). The addition of any of these "segments" to the set of cleavage sites produced by $Z \setminus \{L\}$ would cause the addition of at least one new cleavage site. We know, that they can be only secondary cleavage sites of primary segments not beginning in 0, because otherwise they would increase the error pool (constraints 6-7). The positive answer of problem Π_{RNAPDP} forces such pairing of the segments that they cover all (except the one beginning in 0) primary segments of the lowest level, i.e. the segments of lengths u_i, $1 \leq i \leq q$. These segments cannot cover primary segments of a higher level (i.e. primary segments containing other primary segments), because no two lengths from the ranges of $s(x)$ and $s(y)$ sum up to such big value. The properties of the shifted ranges of $s(x)$ and $s(y)$ guarantee, as the only possible assignment, always one element belonging to $s(x)$ and one to $s(y)$ in every segment u_i, $1 \leq i \leq q$ (see Lemmas 2-3). The solution of problem Π_{NM} is the set of pairs $W_i = \{x_j, y_l\}$ such that $s(x_j)$ and $s(y_l)$ fill up the i^{th} segment of length u_i, $1 \leq i \leq q$. In such a case $P_1 = Z \setminus \{L\}$, P_2 equals to all newly added cleavage sites and all not yet used elements of D cover (following the construction) all interpoint distances between pairs of values from $P_1 \cup \{0, L\}$. This ends the proof. □

The proposed transformation of an instance of problem Π_{NM} to an instance of problem Π_{RNAPDP} is illustrated by the following example.

Example 1. Let the example instance of problem Π_{NM} be:

$$\begin{aligned}
&q = 3,\\
&X = \{x_1, x_2, x_3\},\\
&Y = \{y_1, y_2, y_3\},\\
&s(x_1) = 2, s(x_2) = 3, s(x_3) = 5,\\
&s(y_1) = 4, s(y_2) = 5, s(y_3) = 6,\\
&u_1 = 7, u_2 = 9, u_3 = 9.
\end{aligned}$$

After shifting the ranges of the variables we get the values:

$$\begin{aligned}
&s(x_1) = 11, s(x_2) = 12, s(x_3) = 14,\\
&s(y_1) = 32, s(y_2) = 33, s(y_3) = 34,\\
&u_1 = 44, u_2 = 46, u_3 = 46.
\end{aligned}$$

The construction of the instance of problem Π_{RNAPDP} ends with the following result:

$$\begin{aligned}
&Z = \{1, 45, 91, 137\},\\
&D = \{11, 12, 14, 32, 33, 34, 1, 45, 91, 137, 44, 90, 136, 46, 92, 46\},\\
&n = 4, k = 16, L = 137, C = 0.
\end{aligned}$$

The feasible solution for the instance after the transformation (i.e. of problem Π_{RNAPDP}) is shown in Figure 2, with $P_1 = \{1, 45, 91\}$ and $P_2 = \{12, 57, 105\}$.

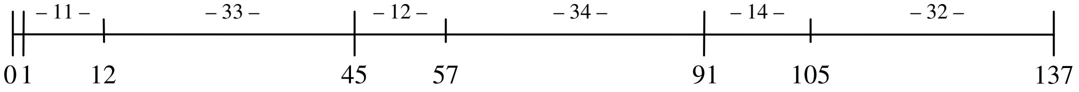

Figure 2: A solution of problem Π_{RNAPDP} for the example instance.

This solution can be easily translated to a feasible solution of problem Π_{NM}:

$$W_1 = \{x_1 + y_2\} \quad (11 + 33 = 44 \ \rightarrow \ 2 + 5 = 7),$$
$$W_2 = \{x_2 + y_3\} \quad (12 + 34 = 46 \ \rightarrow \ 3 + 6 = 9),$$
$$W_3 = \{x_3 + y_1\} \quad (14 + 32 = 46 \ \rightarrow \ 5 + 4 = 9). \ \square$$

Corollary 1. *RNA Partial Degradation Problem in its decision version without any errors allowed (i.e. with $C = 0$) is strongly NP-complete.*

Corollary 2. *RNA Partial Degradation Problem in its search version is strongly NP-hard.*

Following this result, we will propose in the next section an exhaustive search algorithm which finds first cleavage sites, however, at the expense of the computational time.

3 The exact algorithm

In this section, we introduce the branch-and-cut algorithm that works for the case of RNA PDP problem (search version) with false negatives (i.e. missing fragments in D and Z). Its aim is to find the coordinates of primary and secondary cleavage sites in a given RNA molecule, taking into account false negatives. The proposed algorithm builds and searches through the solution tree with leaves corresponding to elements of the solution space of the problem. The algorithm is implemented in C programming language and runs in 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, and also maximum permissible number of errors ($maxErr = C$). In addition, the following structures are defined: multiset R of all pairwise distances between points in $P_1 \cup \{0, L\}$, where each distance is described by a pair of points, and sorted (in non-increasing order) list B of all secondary fragment lengths. These structures do not contain any element at the beginning of the algorithm. The output of the algorithm is set $P_1 \cup P_2$ which will contain reconstructed primary and secondary cleavage sites in RNA molecule.

The aim of the preliminary step of the algorithm is to check whether Z is consistent with D. If not, D is increased by elements of Z missing there. The output of this step is an updated multiset D and the current number of corrected errors ($currErr$). If $currErr \leq maxErr$ is not satisfied, the solution does not exist.

The main algorithm consists of two parts. First part was designed to find the coordinates of primary cleavage sites including false negatives (see Algorithm 1 and 2). In this part P_1 is recognized. If a given cleavage site cannot be assigned due to high $currErr$, the current branch is cut off and the algorithm backtracks. First, each of the elements of Z is checked whether it can be classified as a primary cleavage site as shown in pseudo-code below (Algorithm 1). The number of immediate successors of the node is equal to two, because the element of Z can be added to set P_1 as a primary cleavage site or denoted as a secondary cleavage site and added to list B. Next, the algorithm attempts to reconstruct the missing primary cleavage sites basing on D in a very similar way as for Z (Algorithm 2). Each element of $D \setminus \{L\}$ which is not equal to any of the distances between reconstructed so far primary cleavage sites in P_1 including ends of the molecule and does not belong to list B is considered as a potential primary cleavage site. Every such element can be placed either on the left or right side of the RNA molecule and inserted into P_1 or added to the list of secondary fragments B. Finally, P_1 will contain reconstructed primary cleavage sites. The main procedures of the first part, given in pseudo-code are shown below as Algorithm 1 and 2, respectively. The presented pseudo-code works for the case of ideal data and gives an overview of the algorithm. The notation $\Delta(z_{max}, P_1 \cup \{0, L\})$ denotes the multiset of distances between point z_{max} and all points in set $P_1 \cup \{0, L\}$. The procedure stops when the stopping criterion is satisfied, defined as follows. If set $Z \setminus \{L\}$ becomes empty, then the procedure ends with reconstructed primary cleavage sites in P_1.

Algorithm 1 Reconstruction of primary cleavage sites basing on set Z (the case of data without errors).

```
procedure ReconstructFromZ($Z, D, P_1, B$)
    if $Z \setminus \{L\}$ is empty then
        return
    end if
    $z_{max} \leftarrow$ Maximum element in $Z \setminus \{L\}$
    if $\Delta(z_{max}, P_1 \cup \{0, L\}) \subseteq D$ then
        ReconstructFromZ($Z \setminus \{z_{max}\}, D \setminus \Delta(z_{max}, P_1 \cup \{0, L\}),$
                                $P_1 \cup \{z_{max}\}, B$)
    else
        ReconstructFromZ($Z \setminus \{z_{max}\}, D \setminus \{z_{max}\}, P_1, B \cup \{z_{max}\}$)
    end if
    return
end procedure
```

Algorithm 2 Reconstruction of primary cleavage sites basing on set D (the case of data without errors).

```
procedure ReconstructFromD($D, P_1, B$)
    if $D \setminus \{L\}$ is empty then
        return
    end if
    $d_{max} \leftarrow$ Maximum element in $D \setminus \{L\}$
    if $\Delta(d_{max}, P_1 \cup \{0, L\}) \subseteq D$ then
        ReconstructFromD($D \setminus \Delta(d_{max}, P_1 \cup \{0, L\}), P_1 \cup \{d_{max}\}, B$)
    else if $\Delta(L - d_{max}, P_1 \cup \{0, L\}) \subseteq D$ then
        ReconstructFromD($D \setminus \Delta(L - d_{max}, P_1 \cup \{0, L\}),$
                                $P_1 \cup \{L - d_{max}\}, B$)
    else
        ReconstructFromD($D \setminus \{d_{max}\}, P_1, B \cup \{d_{max}\}$)
    end if
    return
end procedure
```

As a result of the first part of the algorithm the solution set P_1 will contain all reconstructed primary cleavage sites.

The second part of the algorithm was designed to reconstruct the coordinates of the secondary cleavage sites including false negatives. In this part it is checked whether there exists the pair of elements in list B of secondary fragment lengths, which sum is equal to one of the primary fragment lengths. The secondary cleavage site can also be reconstructed when one of the elements of the pair is missing, but only if $currErr < maxErr$. Additionally, only one of the elements in the pair and only within primary fragment starting with the first nucleotide of the RNA of length L, can correspond to a secondary cleavage site obtained during experiment with labeled 5' termini of the RNA molecule. If a pair could not be found, then the current branch is cut off and the algorithm backtracks. The number of immediate successors of the node is equal to a number of pairs which can be composed of the first (i.e. the longest) element of the current list B and any other element of B (including possible missing fragment if $currErr < maxErr$) in such way that these two elements sum up to a length of a primary fragment within current P_1. The main procedure of the second part consists of the presented steps in pseudo-code shown as Algorithm 3. The presented pseudo-code works for the case of ideal data and gives an overview of the algorithm. The procedure stops when the stopping criterion is satisfied, defined as follows. If list B becomes empty then the procedure ends with reconstructed secondary cleavage sites in P_2. If a solution is not found because of the higher number of errors than assumed, then the algorithm backtracks to Part I. A new reconstruction of primary cleavage sites is searched for.

Algorithm 3 Reconstruction of secondary cleavage sites (the case of data without errors).

```
$R = \{r_{ij} = p_j - p_i : p_i, p_j \in P_1 \cup \{0, L\} \wedge p_j > p_i \wedge 0 \leq i < j \leq |P_1| + 1\}$

procedure ReconstructSecondarySites($B, R, P_2$)
    if $B$ is empty then
        return
    end if
    for $k := 2$ to $|B|$ do
        if $\exists r_{ij} \in R : \{b_1 + b_k\} = r_{ij}$ then
            ReconstructSecondarySites($B \setminus \{b_1, b_k\}, R \setminus \{r_{ij}\},$
                                            $P_2 \cup \{r_{0i} + b_1\}$)
        end if
    end for
    return
end procedure
```

The whole algorithm stops after finding a first feasible solution of the considered problem. The algorithm was also extended to cover the case of errorneous data.

The example below illustrates the case with false negatives.

Example 2 presents in detail the action of the proposed algorithm.

Example 2. Let our instance of the considered problem Π_{RNAPDP} be: $D = \{19, 21, 23, 24, 25, 26, 43, 45, 51, 66\}$, $Z = \{24, 43, 45, 66\}$, $L = 66$ and $C = maxErr = 5$ (because the fragment of length 15 is missing in both Z and D, the corresponding error is counted twice). Exemplary reconstruction of cleavage sites can be as follows. As a result of the reconstruction of primary cleavage sites from Z in Part I of the algorithm we will obtain: $P_1 = \{43, 45\}, B = \{24\}$ and $currErr = 1$. Next, elements of $D \setminus \{L\}$ that do not belong to set $P_1 \cup B$ and are not equal to any distances between elements of P_1 and endes of the molecule are selected as potential primary cleavage sites. These elements are: 19, 25, 26, 51. The reconstruction of the primary cleavage sites from $D \setminus \{L\}$ will result in: $P_1 = \{15, 43, 45\}$, $B = \{26, 25, 24, 19\}$ and $currErr = 5$. In Part II of the algorithm the secondary cleavage sites are reconstructed. The solution set will be $P_1 \cup P_2 = \{15, 24, 41, 43, 45\}$.

4 Experimental results

In this section results of the biochemical experiments and 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 experimental and randomly generated data were prepared. The biochemical analysis of RNA degradation was conducted using two artificial molecules RNA-A_{108} and RNA-B_{66}. The secondary structures of both molecules were designed on the base of earlier described rules of nonenzymatic degradation (Kierzek 1992, Bibillo et al. 1999, Bibillo et al. 2000, Kierzek 2001) and the program mfold (See Figure 3).

RNA-A_{108} and RNA-B_{66} molecules were obtained and labeled as described in the literature (Dutkiewicz et al. 2005). Briefly, both RNAs were synthesized by *in vitro* transcription involving chemically synthesized DNA as templates. The resultant transcripts were labeled in two different ways; either a single ^{32}P atom was incorporated at their 5' terminus (single-labeled RNA) or numerous ^{32}P atoms were introduced along the molecules (multi-labeled RNA). To prepare single-labeled RNA, a transcription product was purified and subjected to reaction with [γ-^{32}P]ATP and T4 polinucleotide kinase. To obtain multi-labeled RNA, [α-^{32}P]UTP was added to the in vitro transcription reaction mixture. All reactions were conducted with the MEGAshortscript^{TM} kit (Ambion) and resultant transcripts were purified as described earlier (Dutkiewicz et al. 2005).

RNA degradation experiments were carried out according to procedure proposed for studying nonenzymatic RNA hydrolysis (Kierzek 1992, Bibillo et al. 1999, Bibillo et al. 2000, Kierzek 2001). Reaction mixture containing 0.1 pmol of labeled RNA, 50 mM Tris-HCl (pH 7.5), 2 mM EDTA, 1 mM spermidine, 50 mM NaCl and 0.1% PVP solution was incubated at 37°C for 1, 2, 4 and 6 hours. The reaction products were separated in a 16% denaturating polyacrylamide gel and visualised by autoradiography. **The locations of cleavage sites were determined by applying standard biochemical methods, it means by the comparison of the length of degradation products with the length of the RNA markers. The RNA markers were generated using typical procedures. In order to obtain the first set of RNA markers the examined RNA was single-labeled at the 5' end and subjected to random hydrolysis. As a result, we obtained so called ladder (marked with L in Figure 3) i.e. a mixture of labeled RNA which lengths ranged between a single nucleotide and a full length RNA. To produce the second set of markers, the examined RNA was single-labeled at the 5' end and subjected to cleavage by RNase T1. It cuts RNA after G, thus we obtained a mixture of RNAs labeled at the 5' end and having G at the 3' end (marked with T1 in Figure 3).** The results of the experiment are shown in Figure 3. Their analysis revealed that primary and secondary cleavages occurred during the first hour of incubation. For this reason we considered in further computational tests only those fragments that arose at T_1. During the degradation of single-labeled RNA-A_{108} the following 5'radiolabeled fragments were generated: 30,

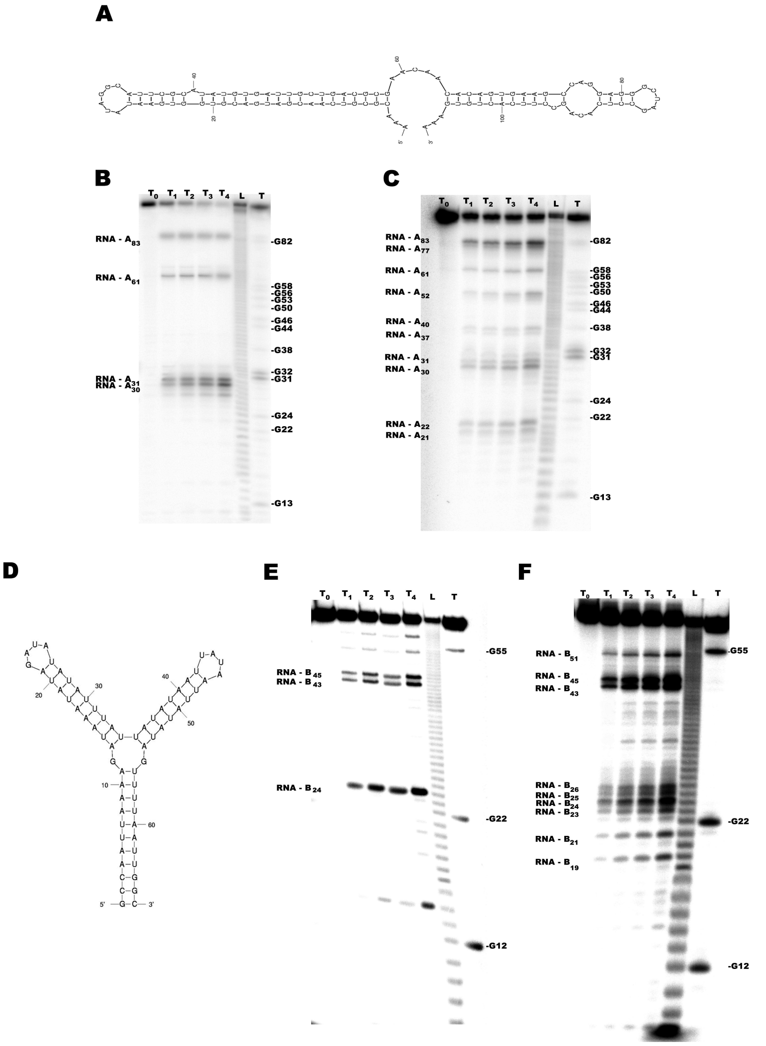

Figure 3: Secondary structure models of two artificial RNA molecules: (A) RNA-A_{108}, (D) RNA-B_{66}. Experimental results of RNA degradation: (B) RNA-A_{108} degradation with labeled 5' termini of the molecule (C) RNA-A_{108} degradation with labeled nucleotides (E) RNA-B_{66} degradation with labeled 5' termini of the molecule (F) RNA-B_{66} degradation with labeled nucleotides. Lanes: T_0 — reaction control; T_1–T_4 — incubated after 1, 2, 4, 6 h respectively; L — formamide ladder, T — limited hydrolysis by RNAse T1

31, 61, 83 and 108 nucleotide long. The degradation of single-labeled RNA-B_{66} led to the formation of 4 radiolabeled fragments: 24, 43, 45 and 66 nucleotides long. In the second experiment with multi-labeled RNA-A_{108} 11 fragments were detected: 21, 22, 30, 31, 37, 40, 52, 61, 77, 83 and 108. During similar experiment involving RNA-B_{66} we observed the formation of 10 products: 19, 21, 23, 24, 25, 26, 43, 45, 51 and 66.

The above data were used to perform computational tests. Although all of the instances gathered in the biochemical experiments had a high number of missing primary and secondary fragments, the algorithm was able to find the correct solution. It was able to determine the following primary cleavage sites in a case of RNA-A_{108}: 31, 83 and RNA-B_{66}: 15, 43, 45. The reconstructed secondary cleavage sites were as for RNA-A_{108}: 30, 52, 61, 71 and for RNA-B_{66}: 24, 41. Analyzing the obtained results, we noticed that the algorithm performs very fast for the real data. Table 1 summarizes running time results for the branch-and-cut algorithm tested on the above instances.

Since the real instances were solved very quickly, random instances with a higher number of primary and secondary cleavage sites as well as a defined number of false negatives were generated. The random erroneous data were obtained by generating sequences of random numbers, from a uniform distribution over interval [1, 2500]. First the primary cleavage sites were generated. The test data have been divided into four sets where secondary cleavage sites have occurred in respectively 25%, 50%, 75% and 100% of 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, each of the four input data sets was separately tested with the number of missing fragment lengths ranging from 1 to 5. The fragments were randomly removed from sets Z and D. The algorithm has been terminated after founding first feasible solution for the instance. Tables 2 – 5 present average results for the random instances, where each entry corresponds to 100 instances of the same number of secondary and primary cleavage sites and number of false negatives, i.e. to 100 runs of the algorithm.

Analyzing the obtained results we noticed, that algorithm performs quite efficiently and fast, despite its computational complexity which equals $O(r^2 2^o)$, where r is the number of primary cleavage sites and o stands for the number of secondary fragment lengths (a pair of secondary fragment lengths corresponds to one secondary cleavage site). The algorithm has been able to find a correct solution in all the considered examples.

5 Conclusions

In the paper, the new RNA Partial Degradation Problem has been formulated and solved. The mathematical formulation of the problem has been shown and its complexity has been analyzed. Since the problem has been proved to be strongly NP-complete, a branch-and-cut algorithm has been proposed. This algorithm gives very good results for erroneous data, especially for the real data obtained from the RNA degradation, and may be very useful in practice as a tool that facilitates the analysis of the output of the biochemical experiment.

In the RNA Partial Degradation Problem we have assumed that a primary fragment is cleaved at most once. Certainly, there could be in general several secondary cleavage sites in between when allowing for long enough RNA incubation. In fact, very long incubation should lead to complete RNA degradation into single nucleotides. However, the problem of the reconstruction of the whole history of an individual RNA from the full length molecule to single nucleotides is too complex. Thus, we have decided to focus our analysis on the preferred cleavage sites and this way simplify the general problem to make easier the reconstruction of the original molecule. The proposed formulation of the simplified problem is general enough to model the degradation of the analyzed RNA and at the same time is still tractable for the reader.

As a continuation of the research reported in this paper, one may consider the analysis of not only secondary but also further products of the spontaneous RNA degradation that occur after several hours. Consequently, we should expect that adding next-order cleavage sites into the search space will complicate the formulation of the problem and the searching algorithm.

Acknowledgements

The research has been supported by the Polish Ministry of Science grant N N519 314635 to M.K. and grant number PBZ-MNiI-2/1/2005 to M.F.

Disclosure Statement

No competing financial interests exist.

References

Bibillo, A., Figlerowicz, M. and Kierzek, R. 1999. The non-enzymatic hydrolysis of oligoribonucleotides. VI. The role of biogenic polyamines. *Nucleic Acids Res.* 27, 3931—3937.

Bibillo, A., Figlerowicz, M., Ziomek, K. and Kierzek, R. 2000. The nonenzymatic hydrolysis of oligoribonucleotides. VII. Structural elements affecting hydrolysis. *Nucleosides Nucleotides Nucleic Acids* 19, 977—994.

Blazewicz, J., Burke, E., Kasprzak, M., Kovalev, A. and Kovalyov, M.Y. 2007. Simplified partial digest problem: enumerative and dynamic programming algorithms. *IEEE/ACM Trans. on Computational Biology and Bioinformatics* 4, 668—680.

Blazewicz, J., Formanowicz, P., Jaroszewski, M., Kasprzak, M. and Markiewicz, W. 2001. Construction of DNA restriction maps based on a simplified experiment. *Bioinformatics* 17, 398—404.

Blazewicz, J. and Kasprzak, M. 2005. Combinatorial optimization in DNA mapping — a computational thread of the Simplified Partial Digest Problem. *RAIRO — Operations Research* 39, 227–241.

Cieliebak, M. and Eidenbenz, S. 2004. Measurement errors make the partial digest problem NP-hard. *Lecture Notes in Computer Science* 2976, 379–390.

Deutscher, M.P. 2003. Degradation of Stable RNA in Bacteria. *Jurnal Biol. Chem.* 278, 45041–45044.

Doma, M.K. and Parker, R. 2007. RNA quality control in eukaryotes. *Cell Vol.* 131, 660–668.

Dutkiewicz, M. and Ciesiolka, J. 2005. Structural characterization of the highly conserved 98-base sequence at the 3' end of HCV RNA genome and the complementary sequence located at the 5' end of the replicative viral strand. *Nucleic Acids Res.* 33, 693–703.

Elbarbary, R.A., Takaku, H., Uchiumi, N., Tamiya, H., Abe, M., Takahashi, M., Nishida, H. and Nashimoto, M. 2009. Modulation of Gene Expression by Human Cytosolic tRNase Z L through 5'-Half-tRNA. *PLOS* 4, e5908.

Eulalio, A., Behm-Ansmant, I. and Izaurralde, E. 2007. P bodies: at the crossroads of post-transcriptional pathways. *Nat. Rev. Mol. Cell Biol.* 8, 9–22.

Garey, M.R. and Johnson, D.S. 1979. *Computers and Intractability. A Guide to the Theory of NP-Completeness.* W.H. Freeman and Company: San Francisco.

Houseley, J. and Tollervey, D. 2009. The Many Pathways of RNA Degradation. *Cell Vol.* 136, 763–776.

Kierzek, R. 1992. Hydrolysis of oligoribonucleotides: influence of sequence and length. *Nucleic Acids Res.* 20, 5073–5077.

Kierzek, R. 2001. Nonenzymatic Cleavage of Oligoribonucleotides. *Methods Enzymol.* 341Nonenzymatic Cleavage of Oligoribonucleotides., 657—75.

Lorentzen, E. and Conti, E. 2006. The exosome and the proteosome: nanocompartments for degradation. *Cell Vol.* 125, 651–654.

Thompson, D.M., Cheng, L., Green, P.J. and Parker, R. 2008. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA* 14, 2095–2103.

Waterman, M.S. 1995. Introduction to computational biology. Maps, sequences and genomes. *Chapman & Hall, London.*

Zhang, S., Li, S. and Kragler, F. 2009. The Phloem-Delivered RNA Pool Contains Small Noncoding RNAs and

Interfers with Translation. *Plant Physiology* 150, 378–387.

RNA molecule	No. of reconstructed primary cleavage sites	No. of reconstructed secondary cleavage sites	No. of false negatives	Average computational time [s]
RNA-A_{108}	2	4	3	0.01
RNA-B_{66}	3	2	5	0.01

Table 1: Computational results for erroneous instances based on the real data.

No. of reconstructed primary cleavage sites	No. of reconstructed secondary cleavage sites	Average computational time [s]				
		Number of false negatives				
		1	2	3	4	5
4	3	0.01	0.01	0.01	0.01	0.01
5	5	0.01	0.01	0.01	0.01	0.01
6	6	0.01	0.01	0.01	0.01	0.01
7	8	0.01	0.01	0.01	0.01	0.01
8	11	0.01	0.01	0.01	0.01	0.02
9	13	0.01	0.03	0.18	0.31	0.38
10	16	0.01	0.52	0.69	2.13	2.98

Table 2: Average computational times for randomly generated erroneous instances. Secondary cleavage sites occur in 25% of all primary fragments.

No. of reconstructed primary cleavage sites	No. of reconstructed secondary cleavage sites	Average computational time [s]				
		Number of false negatives				
		1	2	3	4	5
4	7	0.01	0.01	0.01	0.01	0.01
5	10	0.01	0.01	0.01	0.01	0.01
6	13	0.01	0.01	0.01	0.07	0.10
7	17	0.01	0.02	0.17	0.72	1.43
8	22	0.36	0.69	2.64	3.97	8.11

Table 3: Average computational times for randomly generated erroneous instances. Secondary cleavage sites occur in 50% of all primary fragments.

No. of reconstructed primary cleavage sites	No. of reconstructed secondary cleavage sites	Average computational time [s] Number of false negatives				
		1	2	3	4	5
4	10	0.01	0.01	0.01	0.01	0.01
5	15	0.01	0.01	0.01	0.01	0.02
6	20	0.01	0.02	0.05	0.18	0.34
7	26	0.11	0.65	2.18	5.04	12.10
8	33	0.66	3.01	10.17	16.96	25.38

Table 4: Average computational times for randomly generated erroneous instances. Secondary cleavage sites occur in 75% of all primary fragments.

No. of reconstructed primary cleavage sites	No. of reconstructed secondary cleavage sites	Average computational time [s]				
		Number of false negatives				
		1	2	3	4	5
4	14	0.01	0.01	0.01	0.01	0.01
5	20	0.01	0.01	0.02	0.03	0.07
6	27	0.03	0.16	0.45	0.76	2.26
7	35	0.30	1.87	6.80	12.07	18.62
8	44	7.37	19.64	31.03	41.51	47.48

Table 5: Average computational times for randomly generated erroneous instances. Secondary cleavage sites occur in 100% of all primary fragments.