ON PATH-CONTROLLED GRAMMARS AND PSEUDOKNOTS

Jiří Koutný
Doctoral Degree Programme (4), FIT BUT
E-mail: xkoutn11@stud.fit.vutbr.cz

Supervised by: Alexander Meduna
E-mail: meduna@fit.vutbr.cz

Abstract: This paper discusses path controlled grammars—context-free grammars with a root-to-leaf path in their derivation trees restricted by a control language. First, it introduces a close relationship between some pseudoknots and path controlled grammars generating them in an intuitive way. Then, it discusses pseudoknot-like structures and its relationship to grammars with several controlled paths.

Keywords: path controlled grammars, pseudoknots

1 INTRODUCTION

The investigation of context-free grammars with controlled paths represents an important trend in today's formal language theory (see [2], [6], [9], and [10]). In [9], path-controlled grammars are introduced as an attempt to increase the generative power of context-free grammar without changing the basic formalism and without loosing some basic properties of the class of context-free languages. Consider a context-free grammar G and a context-free language R. A string w generated by G belongs to the language defined by G and R if there is a derivation tree t for w in G such that there exists a path p of t described by R.

A pseudoknot is introduced as the turnip yellow mosaic virus (see [13]) and it is a nucleic acid secondary structure with two or more stem-loop structures such that half of one stem is inserted between the two halves of another stem. Although pseudoknots form knot-shaped three-dimensional patterns, they are not true topological knots. The biological significance of pseudoknots rely on RNA molecules that form pseudoknots (see [4]). The fundamental problem in pseudoknot theory in relation to formal language theory is identification of a pseudoknot—membership problem in terms of theoretical computer science. It is well-known that the general problem of predicting lowest free energy structures with pseudoknots is NP-complete (see [7] and [8]).

The main goal of this paper is to demonstrate some typical pseudoknots generated by path controlled grammars (see [9]) for which membership problem is decidable in a polynomial time (see [10]).

2 PRELIMINARIES

This paper assumes that the reader is familiar with the graph theory (see [1]) and the theory of formal languages (see [11]) including the theory of regulated rewriting (see [3]).

For an alphabet V, V^* denotes the free monoid (generated by V under the operation concatenation), ε is the unit of V^*, and $V^+ = V^* - \{\varepsilon\}$. A subset $L \subseteq V^*$ is a *language* over V. For $x \in V^*$, x^R is mirror image of x.

A *context-free grammar* is a quadruple $G = (V,T,P,S)$ where V is a total alphabet, $T \subseteq V$ is a terminal alphabet, P is a finite set of rules of the form $p : A \rightarrow x$ where p is unique label, $A \in V - T$, $x \in V^*$, and $S \in V - T$ is the starting symbol. For the conciseness, we use the notation $A \rightarrow B|C \in P$ in usual meaning—$A \rightarrow B \in P$ and $A \rightarrow C \in P$. A grammar $G = (V,T,P,S)$ is *linear*, if and only if for all $p : A \to \; x \in P$, $x \in T^*(V-T)T^* \cup T^*$. A derivation step in G is defined for $u, v \in V^*$ and $p : A \to x \in P$ as $uAv \Rightarrow uxv\,[p]$. In the standard manner, we introduce the relations $\Rightarrow^i$, $\Rightarrow^+$, and $\Rightarrow^*$ (see [11]). The language of context-free, linear grammar G is called *context-free language*, *linear language*, respectively, and it is defined as $L(G) = \{x \in T^* \mid S \Rightarrow^* x\}$. The families of linear languages and context-free languages are denoted by **LIN** and **CF**, respectively.

Let $G = (V, T, P, S)$ be a context-free grammar and $x \in T^*$. Let ${}_G\triangle(x)$ denote the set of the derivation trees with frontier x in G. Let $t \in {}_G\triangle(x)$. A *path* of t is any nonempty sequence of the nodes with the first node equals the root of t, the last node equals a leaf of t, and there is an edge in t between each two consecutive nodes of the sequence. Let s be a sequence of the nodes of t, then $word(s)$ denotes the string obtained by concatenation of all labels of the nodes of s in order from left to right.

3 DEFINITIONS

Since, in general, restrictions placed upon a path is a restriction placed upon a derivation tree, we use a slightly modified but equivalent formulation of the definitions stated in [9] and [10]. Consequently, aforementioned modifications allow us to study all derivation-tree-based restrictions (levels, paths, cuts) using the same terminology.

Definition 1. A *tree-controlled* grammar, TC grammar for short, is a pair (G, R) where $G = (V, T, P, S)$ is a controlled grammar and $R \subseteq V^*$ is a control language. The *language that* (G, R) *generates under the path control by* R is denoted by ${}_{path}L(G, R)$ and defined by the following equivalence: For all $z \in T^*$, $z \in {}_{path}L(G, R)$ if and only if there exists a derivation tree $t \in {}_G\triangle(z)$ such that there is path p of t with $word(p) \in R$. Let $\mathbf{path\text{-}TC}(\mathbf{LIN}, \mathbf{LIN}) = \{{}_{path}L(G, R) \mid (G, R)$ is a TC grammar with linear grammar G and linear language $R\}$.

Example 1. Consider the TC grammar (G, R) that generates ${}_{path}L(G, R)$ where

$$G = (\{S, B, D, a, b, c, d\}, \{a, b, c, d\}, P, S),$$
$$P = \{S \to aSd, \quad S \to aBd, \quad B \to bBc, \quad B \to D, \quad D \to bc\},$$
$$R = \{S^n B^n Db \mid n \geq 1\}.$$

Clearly, ${}_{path}L(G, R) = \{a^k b^k c^k d^k \mid k \geq 1\} \notin \mathbf{CF}$.

Inspired by biology (see [13]), we just present some typical pseudoknots in the form of string representation and due to space restrictions, formal definition of general pseudoknot (see [5]) is omitted. Howerver, as opposed to biology where RNA is formed over finite alphabet (*A*denin, *G*uanin, *C*ytosin, and *U*racil), we generalize the pseudoknots over arbitrarily alphabet Σ. The pseudoknots are defined both as stem-only form as well as the form with arbitrarily string between the stems.

Definition 2. Let Σ be an alphabet. The following languages over Σ (see Figure 1) are pseudoknots.

1) $\{xyx^R y^R \mid x, y \in \Sigma^*\}$, $\{u_1 x u_2 y u_3 x^R u_4 y^R u_5 \mid x, y, u_i \in \Sigma^*, 1 \leq i \leq 5\}$,
2) $\{xyx^R zz^R y^R \mid x, y, z \in \Sigma^*\}$, $\{u_1 x u_2 y u_3 x^R u_4 z u_5 z^R u_6 y^R u_7 \mid x, y, z, u_i \in \Sigma^*, 1 \leq i \leq 7\}$,
3) $\{xyx^R zy^R z^R \mid x, y, z \in \Sigma^*\}$, $\{u_1 x u_2 y u_3 x^R u_4 z u_5 y^R u_6 z^R u_7 \mid x, y, z, u_i \in \Sigma^*, 1 \leq i \leq 7\}$,
4) $\{xyzx^R y^R z^R \mid x, y, z \in \Sigma^*\}$, $\{u_1 x u_2 y u_3 z u_4 x^R u_5 y^R u_6 z^R u_7 \mid x, y, z, u_i \in \Sigma^*, 1 \leq i \leq 7\}$.

4 RESULTS

In this section, we present some results related to pseudoknots generated by TC grammars with linear components that generate the language under path control.

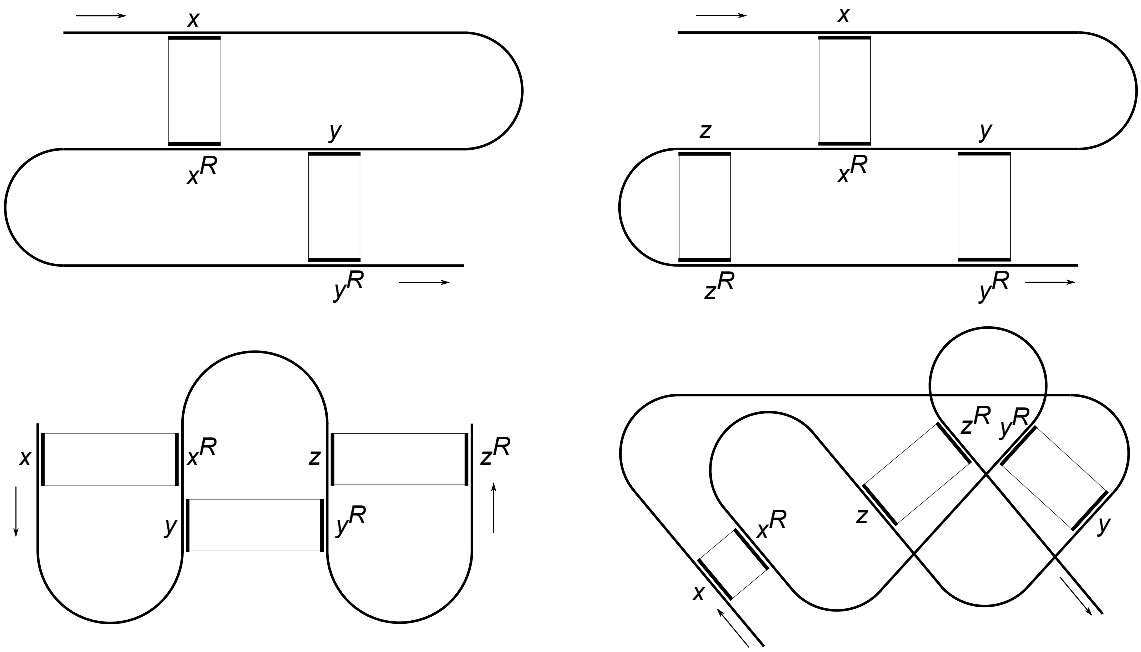

Figure 1: Pseudoknot examples, (top-left) $\{xyx^Ry^R|\ x,y \in \Sigma^*\}$, (top-right) $\{xyx^Rzz^Ry^R|\ x,y,z \in \Sigma^*\}$, (bottom-left) $\{xyx^Rzy^Rz^R|\ x,y,z \in \Sigma^*\}$, (bottom-right) $\{xyzx^Ry^Rz^R|\ x,y,z \in \Sigma^*\}$.

Theorem 1. $\{xyx^Ry^R|\ x,y \in \Sigma^*$ for some $\Sigma\} \in$ **path-TC**$(\mathbf{LIN},\mathbf{LIN})$.

Proof. Consider TC grammar (G,R) where

$$G = (\{S,A,B,A',B',C,D,U,V,a,b,0,1\},\{a,b,0,1\},P,S),$$
$$P = \{1:\ S \to aA \mid bB,$$
$$2:\ A \to aA \mid aB \mid 0C0 \mid 1D1,$$
$$3:\ B \to bB \mid bA \mid 0C0 \mid 1D1,$$
$$4:\ C \to 0C0 \mid 1D1 \mid A' \mid B',$$
$$5:\ D \to 1D1 \mid 0C0 \mid A' \mid B',$$
$$6:\ A' \to aA' \mid bB' \mid U,$$
$$7:\ B' \to bB' \mid aA' \mid V,$$
$$8:\ U \to a,$$
$$9:\ V \to b\}$$
$$R = \{Suvh(u^R)z|\ u \in \{A,B\}^*, v \in \{C,D\}^*\}, z \in \{Ua,Vb\}$$
where h is the morphism defined by $h(A) = A'$, $h(B) = B'$.

Explanation: Starting from S, (G,R) by 1 generates $w = aA$ or $w = bB$. Then, (G,R) repeatly uses 2, 3 to generate $w = xA$ or $w = xB$ where $x \in \{a,b\}^*$ with the derivation tree containing a path Su where $u \in \{A,B\}^*$. Next, (G,R) by 2, 3 generates C or D in a sentential form and thus $w = x0C0$ or $w = x1D1$ where $x \in \{a,b\}^*$ with the derivation tree containing a path SuC or SuD where $u \in \{A,B\}^*$, respectively. Then, (G,R) repeatly uses 4, 5 to generate $w = xyCy$ or $w = xyDy^R$ where $x \in \{a,b\}^*$, $y \in \{0,1\}^*$ with the derivation tree containing a path Suv where $u \in \{A,B\}^*$, $v \in \{C,D\}^*$. By 4, 5, (G,R) generates $w = xyA'y^R$ or $w = xyB'y^R$ where $x \in \{a,b\}^*$, $y \in \{0,1\}^*$ with the derivation tree containing a path $SuvA'$ or $SuvB'$ where $u \in \{A,B\}^*$, $v \in \{C,D\}^*$, respectively. Then, (G,R) uses 6, 7 to generate $w = xyx'A'y^R$ or $w = xyx'B'y^R$ where $x,x' \in \{a,b\}^*$, $y \in \{0,1\}^*$ with the derivation tree containing a path $Suvu'$ where $u \in \{A,B\}^*$, $v \in \{C,D\}^*$, $u' \in \{A',B'\}^*$, and the equivalence $u' = h(u^R)$ is ensured by the controlling language R. Next, (G,R) uses 6, 7 to generate $w = xyx'Uy^R$

or $w = xyx'Vy^R$ where $x, x' \in \{a,b\}^*$, $y \in \{0,1\}^*$ with the derivation tree containing a path $Suvu'U$ or $Suvu'V$, respectively, where $u \in \{A,B\}^*$, $v \in \{C,D\}^*$, $u' \in \{A',B'\}^*$, and $u' = h(u^R)$. Finally, (G,R) uses 8, 9 to generate $w = xyx^Ry^R \in T^*$ with the derivation tree containing a path $Suvu'Ua$ or $Suvu'Vb$ where $u \in \{A,B\}^*$, $v \in \{C,D\}^*$, $u' \in \{A',B'\}^*$ with $u = h(u^R)$. Thus, (G,R) generates ${}_{path}L(G,R) = \{w \mid w = xyx^Ry^R, x \in \{a,b\}^*, y \in \{0,1\}^*\}$ that forms the pseudoknot. Clearly, both G and R are linear. □

Using the same idea as in the proof of Theorem 1, we can demonstrate the following.

Theorem 2. $\{xyx^Rzz^Ry^R \mid x,y,z \in \Sigma^* \text{ for some } \Sigma\} \in$ **path-TC**(**LIN**, **LIN**).

Theorem 3. $\{xyx^Rzy^Rz^R \mid x,y,z \in \Sigma^* \text{ for some } \Sigma\} \in$ **path-TC**(**LIN**, **LIN**).

Proof. Due to space restrictions, TC grammars generating the pseudoknots stated in Theorems 2 and 3 that actually proves the theorems are omitted. However, the schemes of the derivation trees in corresponding TC grammars are sketched in Fig. 2 where the derivation trees of linear grammars that contain a path described by linear languages are presented. □

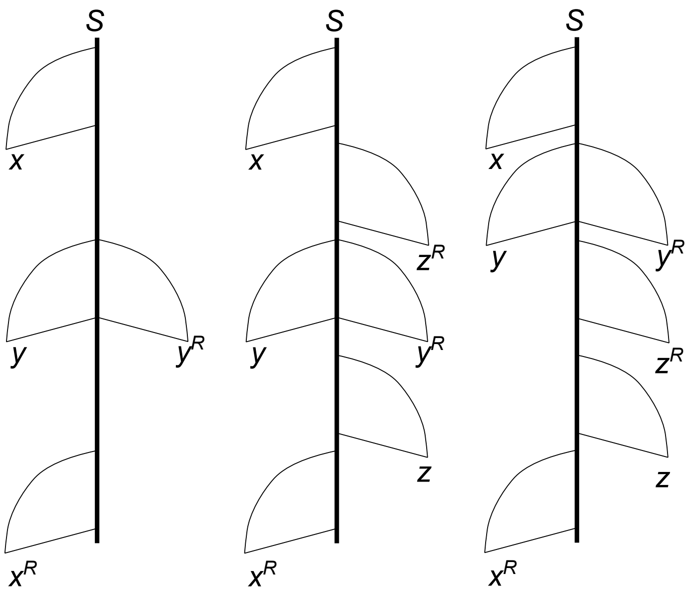

Figure 2: Schemes of the structure of the derivation trees of linear grammars that contain a path described by linear language, (left) $\{xyx^Ry^R \mid x,y \in \Sigma^* \text{ for some } \Sigma\}$, (middle) $\{xyx^Rzy^Rz^R \mid x,y,z \in \Sigma^* \text{ for some } \Sigma\}$, (right) $\{xyx^Rzz^Ry^R \mid x,y,z \in \Sigma^* \text{ for some } \Sigma\}$. Observe that the parts branched on the same level of the derivation tree (schematic view) are handled by the base linear grammar without use of the path control.

Corollary 4. The pseudoknots 1) through 3) introduced in Definition 2 belong to **path-TC**(**LIN**, **LIN**) both in stem-only form as well as in the form with arbitrarily string between the stems.

Open problem 1. Does it hold that $\{xyzx^Ry^Rz^R \mid x,y,z \in \Sigma^*\} \in$ **path-TC**(**LIN**, **LIN**)?

5 CONCLUSION

We have demonstrated several typical pseudoknots used in biology represented by the strings. It is well-known that aforementioned pseudoknots do not belong to **CF**. Inspired by path-controlled grammars introduced in [9] which achieve several properties of context-free grammars, we have demonstrated that some pseudoknots belong to **path-TC**(**LIN**, **LIN**). As it clearly follows from Fig. 2, there are some other combinations of stem positions resulting in the language of pseudoknots-like strings in **path-TC**$(\mathbf{LIN},\mathbf{LIN})$ not mentioned in this paper, however, those structures do not belong to basic pseudoknots appearing in biology.

The open question is whether or not $\{xyzx^Ry^Rz^R \mid x,y,z \in \Sigma^*\}$ and other pseudoknot-like structures (e.g., $\{xyx^Rzy^Rwz^Rw^R \mid x,y,z,w \in \Sigma^*\}$ etc.) can be generated by TC grammars with linear components that generate the language under path control. To answer this question, Ogdens-like lemma should be established and used to disprove that those languages do belong to **path-TC**$(\mathbf{LIN},\mathbf{LIN})$. If they do not, it would mean either we need stronger components (e.g., **path-TC**$(\mathbf{CF},\mathbf{LIN})$) or we need to control more than one path (e.g., **n-path-TC**$(\mathbf{CF},\mathbf{LIN})$ or its variants, see [6]). Note that such kind of Ogdens lemma should be significantly stronger than Prop 8 (Pumping Lemma) in [9] since Ogdens lemma considers not only the substrings but also the positions (see [12]).

ACKNOWLEDGEMENTS

The author thanks Shinnosuke Seki and Szilard Zsolt Fazekas for their inspiration to put grammars with controlled path and pseudoknots into relation. This work was supported by the research plan MSM0021630528, the GA ČR grant GD102/09/H042, the BUT FIT grant FIT-10-S-2, and the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070).

REFERENCES

[1] A. Bondy. *Graph Theory*. Springer, 2010.

[2] M. Čermák, J. Koutný, and A. Meduna. Parsing based on n-path tree-controlled grammars. *Theoretical and Applied Informatics*, 2011(23):213–228, 2012.

[3] J. Dassow and Gh. Păun. *Regulated Rewriting in Formal Language Theory*. Springer, Berlin, 1989.

[4] C. W. Greider J. L. Chen. Functional analysis of the pseudoknot structure in human telomerase RNA. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 102, pages 8080–8085, 2005.

[5] Lila Kari and Shinnosuke Seki. On pseudoknot-bordered words and their properties. *Journal of Computer and System Sciences*, 75(2):113–121, 2009.

[6] J. Koutný, Z. Křivka, and A. Meduna. Pumping properties of path-restricted tree-controlled languages. In *7th Doctoral Workshop on Mathematical and Engineering Methods in Computer Science*, pages 61–69. Brno University of Technology, 2011.

[7] R. B. Lyngsø. Complexity of pseudoknot prediction in simple models. In Josep Díaz, Juhani Karhumäki, Arto Lepistö, and Donald Sannella, editors, *ICALP*, volume 3142 of *Lecture Notes in Computer Science*, pages 919–931. Springer, 2004.

[8] R. B. Lyngsø and Ch.N. S. Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*, 7(3-4):409–427, 2000.

[9] S. Marcus, C. Martín-Vide, V. Mitrana, and Gh. Păun. A new-old class of linguistically motivated regulated grammars. In *Walter Daelemans, Khalil Sima'an, Jorn Veenstra, Jakub Zavrel (Eds.): Computational Linguistics in the Netherlands 2000, Selected Papers from the Eleventh CLIN Meeting, Tilburg*, pages 111–125. Language and Computers - Studies in Practical Linguistics 37 Rodopi 2000, 2000.

[10] C. Martin-Vide and V. Mitrana. Decision problems on path-controlled grammars. *IJFCS: International Journal of Foundations of Computer Science*, 18, 2007.

[11] A. Meduna. *Automata and Languages: Theory and Applications*. Springer Verlag, 2005.

[12] W. Ogden. A helpful result for proving inherent ambiguity. *Mathematical Systems Theory*, 2(3):191–194, 1968.

[13] D. W. Staple and S. E. Butcher. Pseudoknots: RNA structures with diverse functions. *PLoS Biol*, 3(6):e213, 06 2005.