A Comparative Study on Formal Grammars for Pseudoknots

Yuki Kato yuuki-ka@is.aist-nara.ac.jp
Hiroyuki Seki seki@is.aist-nara.ac.jp
Tadao Kasami kasami@empirical.jp

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 630-0192, Japan

Keywords: RNA secondary structure prediction, pseudoknot, formal grammar

1 Introduction

Much attention has been paid to RNA secondary structure predicition based on context-free grammar (cfg) since cfg can represent stem-loop structure by its derivation tree. Especially, techniques based on CKY (Cocke-Kasami-Younger) algorithm have been widely investigated [1]. Pseudoknots play an important role in RNA functions such as ribosomal frameshifting and splicing. A database (PseudoBase) for RNA pseudoknots has been constructed [9]. Unfortunately, it is known that cfg cannot represent pseudoknot structure and a few grammars have been proposed to represent pseudoknots [5, 8]. However, the relation among the expressive (generative) power of these grammars and/or other grammars in formal language theory beyond cfg has not been clarified.

The authors have proposed a class of grammars called *multiple context-free grammars* [3, 7]. In this research, we identify grammars for RNA secondary structure [5, 8] as subclasses of mcfg and also clarify the inclusion relation among the class of languages generated by these grammars.

2 Multiple Context-Free Grammars

A *multiple context-free grammar* (*mcfg*) is a 5-tuple $G = (N, T, F, P, S)$, where N is a finite set of nonterminals, T a finite set of terminals, F a finite set of functions, P a finite set of productions of the form $A \to f[A_1, \ldots, A_k]$, and S the start symbol. For each $A \in N$, a positive integer $dim(A)$ is specified and A derives $dim(A)$-tuples of terminal sequences. Every function $f \in F$ is a total function $(T^*)^{d_1} \times \cdots \times (T^*)^{d_k} \to (T^*)^{d_0}$ for given positive integers d_i $(0 \le i \le k)$ where each component of f is defined as the concatenation of some components of arguments and constant sequences. For example, $PN_1[(x_1, y_1), (x_2, y_2)] = (x_1x_2, y_1y_2)$.

Let $\overset{*}{\Rightarrow}$ denote the least relation satisfying: **(L1)** $A \to \alpha \in P$ implies $A \overset{*}{\Rightarrow} \alpha$, and **(L2)** $A \to f[A_1, \ldots, A_k] \in P$, $A_i \overset{*}{\Rightarrow} \alpha_i$ $(1 \le i \le k)$ implies $A \overset{*}{\Rightarrow} f[\alpha_1, \ldots, \alpha_k]$. The language generated by an mcfg G is $L(G) = \{w \mid S \overset{*}{\Rightarrow} w\}$. A language L is a *multiple context-free language* (*mcfl*) if there exists an mcfg G such that $L = L(G)$.

Consider the follwing productions of mcfg in order to represent a pseudoknot.

$S \to J[A]$	$J[(x, y)] = xy$
$A \to PN_1[A, A]$	$PN_1[(x_1, y_1), (x_2, y_2)] = (x_1x_2, y_1y_2)$
$A \to BP_{au}[A] \mid BP_{ua}[A] \mid BP_{cg}[A] \mid BP_{gc}[A]$	$BP_{au}[(x, y)] = (ax, yu), BP_{ua}[(x, y)] = (ux, ya), \ldots$
$A \to UP_a^{1,L}[A] \mid UP_a^{1,R}[A] \mid \cdots$	$UP_a^{1,L}[(x, y)] = (ax, y)$, $UP_a^{1,R}[(x, y)] = (xa, y), \ldots$
$A \to BF_1[A, A] \mid BF_2[A, A]$	$BF_1[(x_1, y_1), (x_2, y_2)] = (x_1, y_1x_2y_2), \ldots$
$A \to (\varepsilon, \varepsilon)$	

The structure of a pseudoknot can be represented by arc depiction in which arcs cross (see Fig.1). The sequence `aggaacccugaccucgcaucag` can be generated by the above productions.

Figure 1: An arc depiction of a pseudoknot.

3 Inclusion Relation

Let $G = (N, T, F, P, S)$ be an arbitrary mcfg. Let $dim(G) = \max\{dim(A) \mid A \in N\}$. For a production $p : A_0 \to f[A_1, \ldots, A_k] \in P$, $rank(p) = k$ and $rank(G) = \max\{rank(p) \mid p \in P\}$. Let (m, r)-MCFL denote the class of languages generated by mcfg G with $dim(G) \leq m$ and $rank(G) \leq r$. Also, let m-MCFL=$\bigcup_{r\geq 1}(m, r)$-MCFL and MCFL=$\bigcup_{m\geq 1} m$-MCFL. The following shows inclusion relation among the classes of languages.

$$\text{MCFL} \supset \text{2-MCFL} \supset \text{(2,2)-MCFL} \supseteq \text{REL}^{[5]} \supset \text{HL}^{[4]} = \text{TAL}^{[2]} \supseteq \text{ESL-TAL}^{[8]} \supseteq (\text{SL-TAL}^{[8]} \cup \text{CFL})$$

REL is the class of languages generated by grammars introduced in [5]. Especially, $\{a_1^n a_2^n a_3^n a_4^n a_5^n \mid n \geq 0\} \in$MCFL\2-MCFL by Lemma 3.3 of [7], $L_{6,2} \in$2-MCFL\(2,2)-MCFL by Theorem 1 of [6], $\{a_1^m a_2^m b_1^n b_2^n c_1^m c_2^m d_1^n d_2^n \mid m, n \geq 0\} \in$REL\HL (or TAL) by Lemma 4.15 of [7], $\{a^n b^n c^n \mid n \geq 0\} \in$ SL-TAL\CFL by Proposition 2 of [8], $\{\sharp a_1^k b_1^k \sharp a_2^l b_2^l \sharp a_3^m b_3^m \sharp a_4^n b_4^n \sharp \mid k, l, m, n \geq 0\} \in$ CFL\SL-TAL by Proposition 1 of [8], and all regular languages and $\{a^n b^n \mid n \geq 0\}$ belong to CFL∩SL-TAL. Time complexity of recognition algorithms is as follows: $O(n^e)$ (e is the degree of the grammar) for MCFL [3, 7], $O(n^6)$ for (2,2)-MCFL [3, 7], $O(n^5)$ for ESL-TAL, $O(n^4)$ for SL-TAL [8], and $O(n^3)$ for CFL.

Acknowledgments The authors would like to express their thanks to Associate Professor S. Kanaya of Nara Institute of Science and Technology for his valuable discussions.

References

[1] Durbin, R., Eddy, S., Krogh, A., and Michison, G., *Biological Sequence Analysis*, Cambridge University Press, 1998.

[2] Joshi, A.K., Levy, L., and Takahashi, M. Tree adjunct grammars, *JCSS*, 10(1):136–163, 1975.

[3] Kasami, T., Seki, H., and Fujii, M., Generalized context-free grammars, multiple context-free grammars and head grammars, *Preprint of WG on Natural Language of IPSJ*, 87-NL-63-1, 1987.

[4] Pollard, C.J., Generalized phrase structure grammars, head grammars, and natural language, Ph. D. dissertation, Stanford University, 1984.

[5] Rivas, E. and Eddy, S., The language of RNA: A formal grammar that includes pseudoknots, *Bioinformatics*, 16(4):334–340, 2000.

[6] Rambow, O. and Satta, G., A two-dimensional hierarchy for parallel rewriting systems, *IRCS Report 94-02*, University of Pennsylvania, 1994.

[7] Seki, H., Matsumura, T., Fujii, M., and Kasami, T., On multiple context-free grammars, *TCS*, 88:191–229, 1991.

[8] Uemura, Y., Hasegawa, A., Kobayashi, S., and Yokomori, T., Tree adjoining grammars for RNA structure prediction, *TCS*, 210:277–303, 1999.

[9] Van Batenburg, F.H.D., Gultyaev, A.P., Pleij, C.W.A., Ng, J., and Oliehoek, J., PseudoBase: a database with RNA pseudoknots, *Nucleic Acids Research*, 28(1):201–204, 2000. `http://wwbio.leidenuniv.nl/~Batenburg/PKB.html`