

# The Use of Context-Sensitive Grammar For Modeling RNA Pseudoknots

Keum-Young Sung

Division of Computer Science and Electronic Engineering  
Handong University  
Pohang, South Korea  
kysung@handong.edu

**Abstract** - In this study, a context-sensitive grammar is suggested to model various forms of RNA secondary structures, especially pseudoknots. Comparing with a conventional context-free grammar used to model secondary structures of RNA sequences, the use of context-sensitive grammar gives us an advantage of more natural representation of pseudoknots. The suggested grammar directly reflects the appearance characteristic of each form of RNA secondary structure, i.e., hairpins, internal loops, double helixes, and bulge loops. An augmented transition network and the Java programming language are suggested to implement the suggested context-sensitive grammar.

**Keywords:** Context-Sensitive Grammar, Pseudoknots, RNA sequence.

## 1 Introduction

The recognition and prediction of RNA secondary structure, especially RNA pseudoknots, plays an important role in protein synthesis, e.g., ribosomal frameshifting, an infectious or tumor virus, a mutation of HIV, etc [1, 2, 3]. Various forms of context-free grammars have been used to recognize and model RNA sequence consisting of base pairing between a secondary loop structure and complimentary bases outside the loop. Conventionally context-free grammars have been used to identify the secondary structure of RNA molecules from the given nucleotide sequence when we consider an RNA sequence as a string (or a valid sentence) of a programming language. There are several types of pseudoknots, i.e., an interior loop, a bulge loop, a hairpin loop, and so on as illustrated in Figure 1.

The context-free grammar has been used to define the syntactic definition of a programming language. The grammar is a major tool for a parser to build a parse tree to check if the given string is a valid sentence [9]. The whole leaves of a parse tree constitute a sentence of the language defined by the grammar. It consists of a starting

non-terminal, production rules, a set of non-terminals, and a set of terminals. Left hand side (LHS) and right hand side (RHS) comprise a production rule as shown below.  
LHS\_Nonterminal  $\rightarrow$  a Sequence of Grammar Symbols

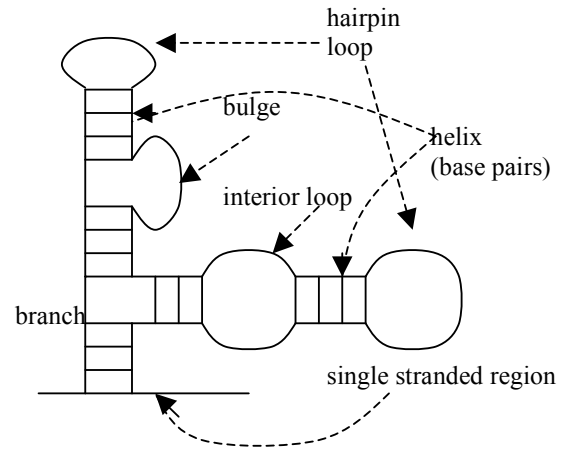
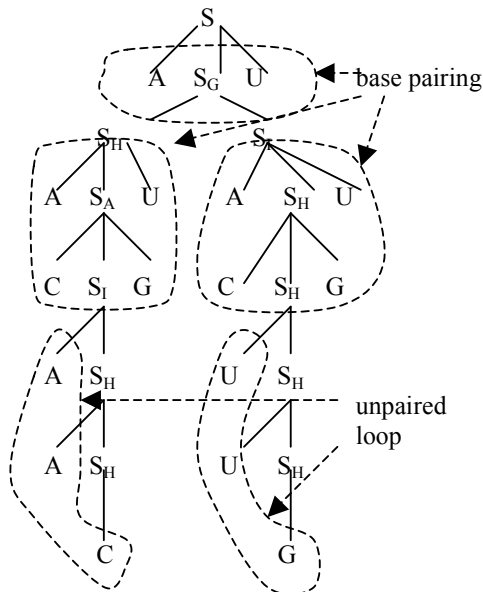


Figure 1. Various Kinds of Pseudoknots

Suppose a context-free grammar defined as follows:

$$\begin{aligned}
 S &\rightarrow S_A \mid S_G \\
 S_A &\rightarrow A S_A U \mid U S_A A \mid A S_G U \mid U S_G A \\
 &\quad \mid S_I S_H \mid S_H S_I \mid \text{PIN} \\
 S_G &\rightarrow G S_G C \mid C S_G G \mid G S_A C \mid C S_A G \\
 &\quad \mid G S_H S_I C \mid C S_H S_I G \mid \text{PIN} \\
 S_H &\rightarrow A S_H U \mid U S_H A \mid A S_A U \mid U S_A A \\
 &\quad \mid S_I \mid \text{PIN} \\
 S_I &\rightarrow G S_I C \mid C S_I G \mid C S_G G \mid G S_G C \\
 &\quad \mid S_H \mid \text{PIN} \\
 \text{PIN} &\rightarrow \text{LEFT\_SKEW} \mid \text{RIGHT\_SKEW} \\
 \text{LEFT\_SKEW} &\rightarrow S_I A \mid S_I U \mid S_I G \mid S_I C \\
 &\quad \mid \text{PIN} \mid \text{TERMINAL} \\
 \text{RIGHT\_SKEW} &\rightarrow A S_H \mid U S_H \mid G S_H \mid C S_H \\
 &\quad \mid \text{PIN} \mid \text{TERMINAL} \\
 \text{TERMINAL} &\rightarrow A \mid U \mid G \mid C
 \end{aligned}$$

For an RNA sequence, A A C A A C G U A C U U G G U U, a parse tree can be constructed in Figure 2.



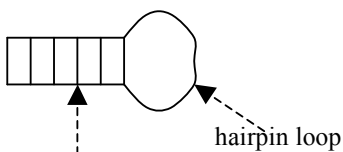
**Figure 2. A Parse Tree Construction of Pseudoknots**

The parse tree generated with a given grammar shows base pairing regions and unpaired hairpin regions. As the name, context-free grammar, implies, the non-terminals on the left-hand side of a production rule does not consider the context in which it is situated.

## 2 A Context-Sensitive Grammar for the Structure of Pseudoknots

The following illustrates RNA secondary structures and corresponding context-sensitive sequence of grammar symbols.

### 2.1 Hairpin

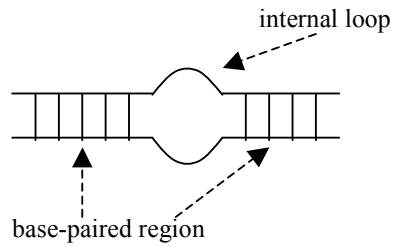


base-paired region  
 Hairpin ::= SequenceA HairPinLoop PairedSequenceA

**Figure 3. Hairpin**

The hairpin pseudoknot consists of a base-pairing region and a non-paired region. The nonterminals, SequenceA and PairedSequenceA comprise the base-pairing region of a hairpin structure.

### 2.2 Internal loop

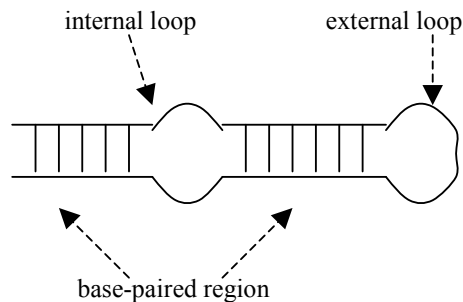


InternalLoop ::= Sequence SequenceA LoopSegmentA SequenceB Sequence PairedSequenceB LoopSegmentB PairedSequenceA Sequence

**Figure 4. Internal Loop**

In Figure 4, there are two base-pairing regions, i.e., one with SequenceA and PairedSequenceA, and another with SequenceB and PairedSequenceB. The two separate non-paired sequences, LoopSegmentA and LoopSegmentB, make an internal loop.

### 2.3 Double Helix



DoubleHelix ::= SequenceA LoopSegmentA SequenceB HairPinLoop PairedSequenceB LoopSegmentB PairedSequenceA

**Figure 5. Double Helix**

A double helix as shown in Figure 5 consists of an internal loop and an external loop (or a hairpin). There are two base-pairing in the given grammar, the pair with SequenceA and PairedSequenceA, and the pair with SequenceB and PairedSequenceB. The non-paired sequence, LoopSegmentA and LoopSegmentB constitutes a loop, and the HairPinLoop part makes an external loop.

## 2.4 Bulge Loop

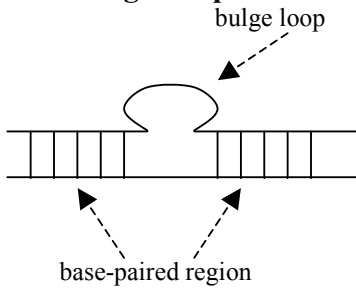


Figure 6. Bulge Loop

```
BulgeLoop ::=
Sequence SequenceA LoopSegmentA
SequenceB Sequence PairedSequenceB
Sequence PairedSequenceA
```

The symbol, LoopSegmentA, makes a bulge loop in the context of SequenceA and SequenceB. SequenceA and PairedSequenceA constitute an RNA pair. SequenceB and PairedSequenceB also make a base-pairing.

## 3 Augmented Transition Network for Context-Sensitive Grammar

An augmented transition network (ATN) [8] as a parser is suggested to implement the suggested context-sensitive grammar. The example pseudocode for recognizing a hairpin is as follows:

```
public class Hairpin
{
    public Hairpin(String GivenSequence)
    {
        Sequence = GivenSequence;
    }
    while (remaining_sequence > 0)
    {
        ChoppedSeq = ChoppingSequence(Sequence)
        ReversedChoppedSeq = Reverse(Sequence)
        BasePair(ChoppedSeq);
        // search a sequence to be matched
        Loop(ChoppedSeq);
        // search a loop that is not paired
        BasePair(ReversedChoppedSeq);
        // search another sequence to be paired
    }
    private String Sequence;
    private RNASequence SequenceA;
    BasePairing BasePair;
    LoopStructure Loop;
}
```

The above class invokes pre-defined class functionalities in the sequence of given context that is necessary to recognize a hairpin, which is indicated as comments.

Figure 7 depicts the order of processing based on the given context:

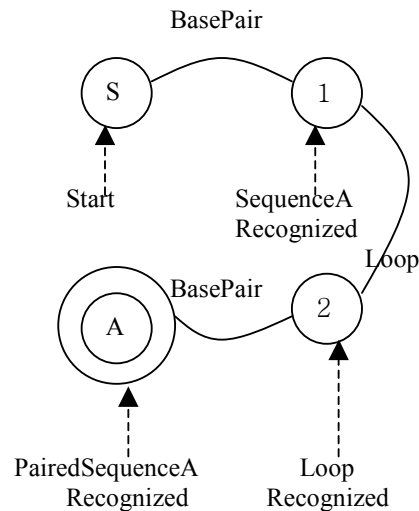


Figure 7. An Augmented Transition Network

## 4 Conclusions

The suggested use of context-sensitive grammar gives more expressiveness and understandability to represent RNA secondary structure than using context-free grammar. However, the suggested technique suffers from the similar weakness of the conventional context-free grammar. The searching time of base-pairing in a long RNA sequence becomes prohibitively formidable, and wrong base-pairing may be predicted when a variety of pseudoknots are mixed in an RNA sequence. The further study for overcoming these problems includes the following:

- The reduction of processing time to perform base-pairing especially in a long RNA sequence;
- More speedy and exact recognition of a loop in a mixed chain of paired and non-paired nucleotides;
- Refinement of the given grammar for more compact representation of an RNA secondary structure; and
- Detailed visualization of pseudoknots with Java applet for internet use.

The last difficulty with this study is to construct a parser to implement the suggested context-sensitive grammar because there is no known automatic parser generator based on a context-sensitive grammar. Even with all the difficulties related to context-sensitivity applied to RNA sequences, we can describe and model the configuration of RNA pseudoknots more naturally and precisely than using context-free grammar.

## 5 References

- [1] Jan Liphardt, et al., Evidence for an RNA Pseudoknot Loop-Helix Interaction Essential for Efficient – 1 Ribosomal Frameshifting, *Journal of Molecular Biology*, 288, 1999, 321-335.
- [2] Changyu Wang, et al., An RNA Pseudoknot, an Essential Structure of the Ribosome Entry Site Located within the C Virus 5' Noncoding Region, *RNA*, 1, 1995, 526-537.
- [3] Campbell, Mitchell, and Reece, *Biology Concepts and Connections*, Second Edition, (Addison Wesley, 1997).
- [4] Y. Sakakibara, et al., The Application of Stochastic Context-Free Grammars to Folding, Aligning, and Modeling Homologous RNA sequences, Technical Report UCSC-CRL-94-14, 1993.
- [5] Y. Schabes and R. C. Waters, Lexicalized Context-Free Grammars, In 21st Meeting of the Association for Computational Linguistics (ACL'93), June 1993, 1993, 121-129.
- [6] S. Kobayashi and T. Yokomore, Modeling RNA Secondary Structure Using Tree Grammars, In *Proceedings of Genome Informatics Workshop V*, University Academy Press, 1994, 29-38.
- [7] L. Grate, et al., RNA Modeling Using Gibbs Sampling and Stochastic Context-Free Grammars, In *ISMB-94*, AAAI/MIT Press, 1994.
- [8] P. H. Winston, *Artificial Intelligence*, Third Edition, Addison Wesley, 1993.
- [9] R. W. Sebesta, *Concepts of programming languages, fifth edition* (Addison Wesley, 2002).