# A Symbolic Representation for RNA Secondary Structures; towards the Construction of RNA Secondary Structures Data Base

A. Nakaya [1] A. Yonezawa [1] K. Yamamoto [2]

[1] Department of Information Sciences,
Faculty of Science, University of Tokyo.

[2] Laboratory of Clinical Microbiology and Immunology,
Bun'inn Hospital of University of Tokyo
backen@camille.is.s.u-tokyo.ac.jp

## 1 Introduction

It is known that a single stranded ribonucleic acids (ssRNA) take a variety of secondary structures. Recently a lot of new findings have been done for the functions of RNA molecules such as RNA editing and ribozyme. These functions are thought to have some relationship with the secondary structures of ssRNA molecules.

Some of these secondary structures of ssRNA molecules have been decided with the experimental results such as X-ray diffraction analysis, NMR and S1 nuclease digestion. Another method is the computer prediction of the secondary structures. Many studies have been made to predict the secondary structures of a given RNA sequence. The secondary structures obtained from the above experiments or computer programs is uniquely represented by a sequence of bases and hydrogen bonds. However, it is almost impossible to understand the secondary structure which consist of more than a few hundred of bases without proper representation. A lot of methods for the representation of ssRNA secondary structures have been reported. A planar graph is used for the representation. Each bases are represented as a character (i.e. A, G, C, and U) and connected with lines from 5' to 3'. Hydrogen bonds between the bases are represented by the lines which connect the paired bases. These method are enough to understand the structure for our eyes, but not suitable for the construction of the data base.

with the progress of genome sciences, a huge DNA sequence information including genomic RNA sequences has been accumulated. With the near future studies, almost all the DNA sequences which can be transcribed into ssRNA molecules will appear in front of us. The information of the secondary structures will be accumulated by experimental methods and computer algorithm. As a result of these progress, the construction of RNA secondary structures data base will be needed. For this case, the method represented on a plane is not suitable because of the memory size. We have newly developed a Symbolic representation for RNA secondary structures. We report its application for the construction of the data base.

## 2 Methods and Results

Here, we define the Secondary Structure Description Language (SSDL) which can represent a secondary structure without loss of information.

Definition:

<secondary structure> → <dangling bases> (D1)
| ([<5'bases>]<secondary structure>[<3'bases>]) (D2)
| <secondary structure>* (D3)
<dangling bases> → <base>* (D4)
<base> → A | C | G | U (D5)
<5'bases> → <base>* (D6)
<3'bases> → <base>* (D7)

Figs (b), (c) , (d) are the results obtained from the sequence shown in Fig. (a).

(a) 5'-
AAAGGGAAAGGGAAAGGGAAAAACCCAAAGGGAAAAACCCAAACCC
AAA

GGGAAAGGAAAAACCCAAAGGGAAAAACCCAAAGGGAAAAACCCAA
ACC
CAAACCCAAA-3'

(b)

```
5'-AAA
   ([GGG] AAA
          ([GGG] AAA
                ([GGG] AAAAA[CCC]) AAA
                ([GGG] AAAAA[CCC]) AAA
           [CCC]) AAA
          ([GGG] AAA
                ([GGG] AAAAA[CCC]) AAA
                ([GGG] AAAAA[CCC]) AAA
                ([GGG] AAAAA[CCC]) AAA
           [CCC]) AAA
    [CCC]) AAA-3'
```

(c)

```
5'-*
   ([*] *
          ([*] *
                ([*] **[*]) *
                ([*] **[*]) *
           [*]) *
          ([*] *
                ([*] **[*])*
                ([*] **[*]) *
                ([*] **[*]) *
           [*]) *
    [*]) *-3'
```

(d)
5'-((()())(()()())-3'