SQU Journal For Science, 15(2010) 1-8

An Approach to Analyze the Ambiguity in RNA Structure

Shailendra Singh and *Amardeep Singh

**Department of Computer Science and Engineering, PEC University of Technology, Chandigarh- 160012, India, Email: shailendra_sing@yahoo.com, Department of Computer Engineering, University College of Engineering (UCoE), Punjabi University, Patiala, India, Email: amardeep-dhiman@yahoo.com.*

طريقة لتحليل عدم الوضوح في شكل ال RNA

سيلاندرا سنج و أماريديب سنج

خلاصة: ان DNA وال RNA هما من أهم الجزيئات في الخلية الإنسانية حيث تحتل ال RNA المرتبة الثانية من حيث اهمية الحامض النووي في خلية الأنسان وهي تمثل عامل الوصل مابين ال DNA وتكوين البروتين. هنالك العديد من أنواع ال RNA المتواجدة في مختلف الخلايا ولكل منها مهمة محددة. إن تخزين ومن ثم استعمال المعلومات الجينية بواسطة الخلية يعتمد أساساً على المظهر الخارجي وشكل ال RNA . إن ال RNA تلعب دوراً أساسياً في الخلايا لذ فإن فهم تركيبة ال RNA يمكن استغلالها لمعرفة العلاقة السائدة لكل أو لبعض نوع سلسلة النيوكلوتايدات. هنالك عدة طرق وأنواع من التحاليل لشكل ال RNA والتي يمكن أن تؤدي لمعلومات قد تساعد لمعرفة عمل ال RNA والمهام التي تؤديها وذلك بدرجات متفاوتة من التفاصيل. في هذه الورقة تمت دراسة ومناقشة أنواع متعددة من الأشكال الثانوية ل RNA وقد تم إختيار أشكال محددة لدراسة الاحتمالات المختلفة والتي تبين وضوح الرؤية لشكل ال RNA.

ABSTRACT: DNA and RNA are two very important bio-molecules of the human cell. RNA is the second major form of nucleic acid in human cells that plays an intermediary role between DNA and functional protein. Several classes of RNA²s are found in cells, each with a / its distinct function. Understanding of storage and utilization of a cell's genetic information is based on the structure of RNA. However, many experimental results have shown that RNA plays another greater role in the cells. RNA sequences which contain signals at the structure level can be exploited to detect functional motifs common to all, or a portion of, those sequences. Different types of analysis of a structure can provide functional information in different degrees of detail. This paper discusses various types of RNA secondary structure representation and which structure can be adopted as appropriate for a probabilistic approach that avoids ambiguity.

KEYWORDS: Secondary structure, Stochastic context-free grammar (SCFG), Derivation tree.

1. Introduction

RNA is a biological polymer consisting of monomers called nucleotides. Each nucleotide consists of a (ribose) sugar, a phosphate group and a base. There are four main types of bases: Adenine (A), Cytosine (C), Guanine (G), and Uracile (U). The base-paired structure formed by the Watson-Crick base-pairs A-U and C-G and the wobbling base-pair G-U can be divided into loops, also known as 'structure elements'. A loop is a formation of a base-pair that encloses a chain of nucleotides or other base-pairs. RNA primary structure is commonly represented by a string, S, over the alphabet $\sum$ = {A, G, C, U}. RNA is mostly involved in the biological machinery that expresses the genetic information from DNA to RNA. Information is encoded in RNA by the linear arrangement of the four different constituent nucleotides. RNA molecules perform a number of critical functions. Many of these functions are related to protein synthesis. Some RNA molecules bring genetic information from a cell's chromosomes to its ribosome's, where proteins are assembled.

The RNA plays a very important role in bio cells. Determining RNA shapes has gained considerable importance in the last decade because it is essential for researchers to know the shape of a molecule, in order to understand its role within a cell. A lot of work has been done in the structural analysis of RNA in the bioinformatics field, but there exist a large number of challenging problems like the analysis of ambiguities in RNA structure, prediction of structure, and predictions of functions performed (Hiroshi, *et al.* 2005 and, Jizhen, *et al.* 2006). The structure of an RNA molecule is closely related to its function (Yinglei, S. *et al.* 2004). For this reason, predicting the secondary structure of an RNA molecule based on its primary sequence has been of interest to many researchers. Since RNA structure is essentially governed by base pairing of nucleotides, many computational methods and algorithms have been proposed for finding the "optimal base pairing" of RNA in an efficient manner (Keum, Y. S. 2006, Mount, D. W. 2004, and Rafael, G. 2006).

For the computer science community, the primary structure of bio-molecules is just a very long string of commands forming long programs written in any computer programming language. This long program in the form of a long string is to be processed by compilers, translators etc. using regular expressions, grammars, and similar other techniques. According to Noam Chomsky, the Context Free Grammar (CFG) has great importance in the Linguistic field, Computer Science, Engineering and Bioinformatics. It is a more powerful class of formal grammars than the regular grammar. CFGs are often used to define the syntax of programming languages (Byung-Jun, Y. and Vaidyanathan, P.P. 2007). A CFG, also called a Type 2 Grammar, is similar to a regular grammar, but it permits a greater variety of production rules. The other methods used for the analysis of RNA structure are the free energy based model and conditional log-linear models (CLLMs). CLLMs are a generalization of grammar based models (Dowell, R.D. and Eddy S. R. 2004). According to the evaluation done by its authors, it has accuracies that are better than those of the current probabilistic and physics based models. One purpose of this paper is to present an effective method for analyzing the ambiguity in RNA structure and estimating a stochastic context-free grammar to model a family of RNA sequences (Yuki, K., Hiroyuki, S. and Tadao, K. 2003).

2. Analysis of RNA structure

The importance of grammars in compliers is well-known. The grammars are useful tools to model character sequences and, in a certain way, these tools are useful to model molecular biological sequences (Yan, D. and Yulei, Z. 2005). Many bioinformatics problems can be reformulated in terms of formal languages, producing the corresponding grammar from the available data. Among several utilities contributed by grammars, the main contribution is the ability to test by derivations if a sequence is syntactically correct, i.e. if it belongs to a determined language. A derivation can be represented as a tree-like structure known as a 'derivation tree'. This tree reflects the syntactical structure of a sequence. It is possible that for a given sequence there may be more than one derivation tree. In this case, we say that the grammar is ambiguous. In ambiguous grammar, complexity of the derivation rises given that the number of possible trees grows exponentially with the length of the

sequence to be derived. Stochastic syntactic analysis algorithms for the class of stochastic context free grammars (SCFG) have been proposed and their application has been demonstrated in pattern classification problems.

3. Context free grammar for RNA

Type-2 grammars, or CFGs, are 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. The grammar is a major tool for a parser to build a parse tree to check if the given string is a valid sentence. The whole leaves of a parse tree constitute a sentence of the language defined by the grammar. As the name 'context-free grammar' implies, the non-terminals on the left-hand side of a production rule do not consider the context in which it is situated.

For example, one of the applications of productions as shown in Figure 1 can generate the RNA sequence "**AGCGUCAGUGACUUGAUGC**U" by the following derivation, and the equivalent derivation tree is shown in Figure 7.

3.1 Productions

$$
\begin{aligned}
P = \{ \quad & S_0 \rightarrow S_1, & & S_7 \rightarrow AS_8U \\
& S_1 \rightarrow AS_2U, & & S_8 \rightarrow GS_9U \\
& S_2 \rightarrow GS_3C, & & S_9 \rightarrow US_{10} \\
& S_3 \rightarrow CS_4G, & & S_{10} \rightarrow GS_{11} \\
& S_4 \rightarrow GS_5U, & & S_{11} \rightarrow AS_{12} \\
& S_5 \rightarrow US_6A & & S_{12} \rightarrow C \quad \} \\
& S_6 \rightarrow CS_7G
\end{aligned}
$$

Figure 1. Set of production rules 'P'

Figure 1 shows a set of production rules P that generates an RNA sequence for a certain restricted structure, in which $S_0, S_1 \ldots\ldots S_{12}$ are non-terminals. A, G, C and U are terminals. Beginning with the start symbol S_0, any production with S_0 left of the arrow can be chosen to replace S_0. If the production $S_0 \rightarrow S_1$ is selected, then the symbol S_1 replaces S_0. This derivation step is written as $S_0 \rightarrow S_1$, where the arrow signifies application of a production. Next, if the production $S_1 \rightarrow A\ S_2\ U$ is selected, the derivation step is $S_1 \rightarrow A\ S_2\ U$. Continuing with the same procedure of replacing left-hand side with the right-hand side of an appropriate production, we obtain the following derivation terminating with the desired sequence:

3.2 Derivation

$S_0 \rightarrow S_1$

$\rightarrow AS_2U \qquad (S_1 \rightarrow AS_2U)$

$\rightarrow AGS_3CU \quad (S_2 \rightarrow GS_3C)$

$\rightarrow AGCS_4GCU \quad (S_3 \rightarrow CS_4G)$

$\rightarrow AGCGS_5UGCU \quad (S_4 \rightarrow GS_5U)$

$\rightarrow AGCGUS_6AUGCU \quad (S_5 \rightarrow US_6A)$

$\rightarrow AGCGUCS_7GAUGCU \quad (S_6 \rightarrow CS_7G)$

$\rightarrow$AGCGUCAS_8UGAUGCU ($S_7 \rightarrow$ AS_8U)

$\rightarrow$AGCGUCAGS_9UUGAUGCU ($S_8 \rightarrow$GS_9U)

$\rightarrow$ AGCGUCAGUS_{10}UUGAUGCU **(**$S_9 \rightarrow$ US_{10}**)**

$\rightarrow$ AGCGUCAGUGS_{11}UUGAUGCU **(**$S_{10} \rightarrow$ GS_{11}**)**

$\rightarrow$ AGCGUCAGUGAS_{12}UUGAUGCU **(**$S_{11} \rightarrow$ AS_{12}**)**

$\rightarrow$ AGCGUCAGUGACUUGAUGC **(**$S_{12} \rightarrow$ C**)**

4. Different secondary structure for RNA

RNA secondary structures can be displayed in different kinds of representations. Depending on the use of the RNA molecules, specific representations are more or less useful. The bracket notation as shown in Figure 2 is a text-based representation. The structure has been reflected in a string of dots and brackets. Dots denote non-bonding bases and a pair of brackets indicates a base-pair. A more convenient representation, which expands in all directions in a plane and thus is closer to a spatial representation, is the squiggle plot as shown in Figure 3. It is the most appropriate plot to easily describe the approximate spatial structure of RNA. Base-pairs are given as two bases connected through either a straight line (Watson-Crick base-pairs) or a circle indicating the so-called wobbling base-pair, G-U.

Considering RNA in a more theoretical way, the representations as trees or as arc-annotated sequences are well-accepted. In recent years, tree-representations of RNA secondary structures have occurred in the literature, and algorithmic applications on trees are performed successfully. Arc-annotated sequences focus on representing sequences as straight lines. Arcs indicate base-pairings. This kind of representation is used in this paper mainly due to its beneficial representation of single base and base-pair operations. A similar representation to the arc-annotated sequence is the drawing of this sequence on a circle as shown in Figure 5. Arcs are plotted as curved lines inside this circle. The mountain plot as shown in Figure 6 is useful for large RNAs. Plateaus represent unpaired regions, and the heights of these mountains are determined by the number of base-pairs in which the partial sequences are embedded. Figure 7 shows a derivation tree for a given sequence and Figure 8 shows the appropriate way representation of a sequence.

4.1 Dot-bracket representation

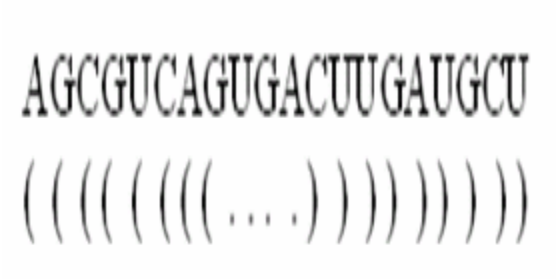

Figure 2. Dot-bracket representation

4.2 Squiggle plot

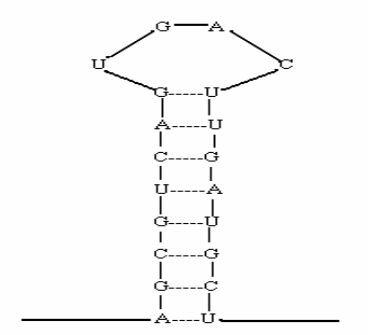

Figure 3. Squiggle plot

4.3 Arc-annotated sequence

Figure 4. Arc-annotated sequence

4.4 Circle representation

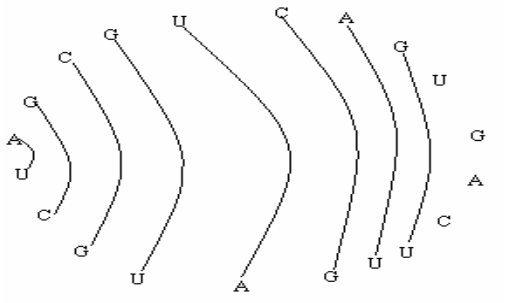

Figure 5. Circle representation

4.5 Mountain plot representation

Figure 6. Mountain plot representation

4.6 Derivation Tree Representation

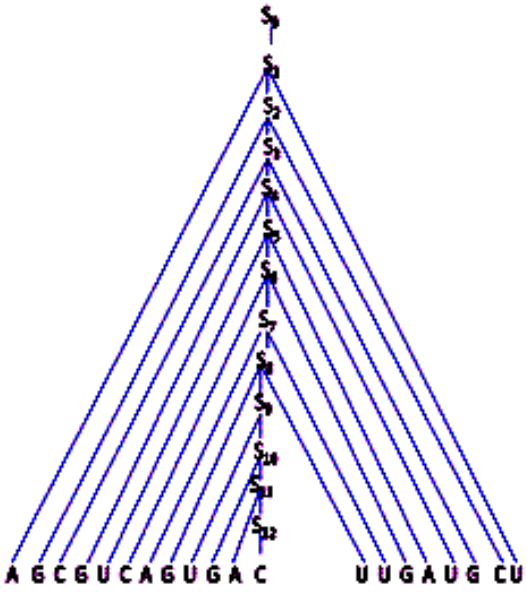

Figure 7. Derivation tree representation

4.6 Most Appropriate Way Representation

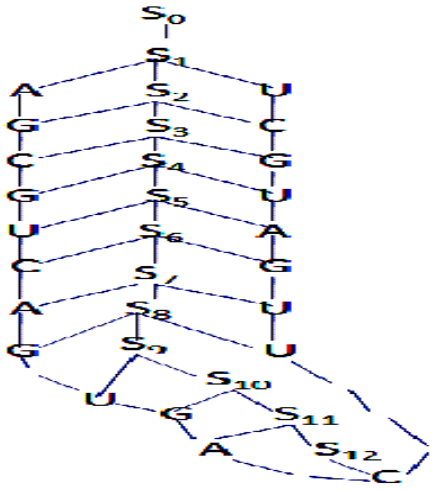

Figure 8. Most appropriate way representation

5. Assignment of probability on productions

A SCFG extends the definition of context free grammars by associating a probability to every production in the grammar. Consequently every string that the grammar can generate is assigned a probability which is equal to the product of the probabilities of the productions used in the string's derivation. The probability of a parse tree can be calculated as a product of the probabilities of the production instances in the tree. There are various methods used to determine such probabilities, and using one such method, the assignment of probabilities is as shown in Table 1. To derive the trained grammar, the initial grammar was designed by using some prior knowledge about the RNA family.

Table 1. Probabilities for the Type 2 grammar, with uniform distribution placed over each set of the same type of production.

Category of Productions	Productions	Probabilities
C#1	$S_0 \rightarrow S_1$	0.3000
C#2	$S_1 \rightarrow A S_2 U$	0.0250
C#2	$S_2 \rightarrow G S_3 C$	0.0250
C#2	$S_3 \rightarrow C S_4 G$	0.0250
C#2	$S_4 \rightarrow G\ S_5 U$	0.0250
C#2	$S_5 \rightarrow U S_6 A$	0.0250
C#2	$S_6 \rightarrow C S_7 G$	0.0250
C#2	$S_7 \rightarrow A S_8\ U$	0.0250
C#2	$S_8 \rightarrow G S_9 U$	0.0250
C#3	$S_9 \rightarrow U S_{10}$	0.0666
C#3	$S_{10} \rightarrow G S_{11}$	0.0666
C#3	$S_{11} \rightarrow A S_{12}$	0.0666
C#4	$S_{12} \rightarrow C$	0.3000

6. Conclusion

A detailed understanding of the functions and interactions of RNA requires knowledge of their structures. For many RNA molecules, the secondary structure is highly important to the correct function of the RNA, often more so than the actual sequence. One of the problems with CFGs is that they generally have an ambiguity in the grammar that results in more than one parse tree for a sequence, and alternative parse trees reflect alternative secondary structures. Thus a grammar often gives several possible secondary structures for one RNA sequence. The SCFG is used to overcome the problem of ambiguity. One of the advantages of a SCFG is that it can provide the most likely parse tree. If the grammar and their probabilities are carefully designed, the correct secondary structure will appear as the most likely parse tree among the alternatives. The grammar itself may be a valuable tool for representing a RNA family or domain. For (long-chain) RNA there are exponentially many possible structures which may be assigned to RNA, but assigning the correct one can only be done on the basis of a probability distribution. However, the most challenging future problem is to model a family of longer RNA sequences, and also the variations of RNAs like mRNA, tRNA, and siRNA.

6. References

BYUNG-JUN, Y. and VAIDYANATHAN, P.P. 2007. Computational identification and analysis of noncoding RNAs. *IEEE Signal Processing Magazine,* **24(1):** 64-74.

DOWELL, R.D. and EDDY S.R. 2004. Evaluation of Several Lightweight Stochastic Context Free Grammars for RNA Secondary Structure Prediction. *BMC Bioinformatics,* **5:** 71-78.

HIROSHI, M., KENGO, S. and YASUBUMI S. 2005. Pair Stochastic Tree Adjoining Grammars in Yan, D. and Yulei, Z. 2005Aligning and Predicting Pseudoknot RNA Structures. *Journal of Bioinformatics*, **21**: 2611-2617.

HOPCROFT, J.E. and ULLMAN, J.D. 1979. *Introduction to Automata Theory, Languages and Computation*. Addison Wesley.

JIZHEN, Z., LIMING, C. and RUSSELL L.M. 2006. Learning the Parameters of Stochastic Grammar Models for RNA Structures with Pseudoknots. *IEEE International Conference on Granular Computing 2006, Atlanta, May 10-12:* 170-175.

KEUM, Y.S. 2006. Recognition and Modeling of RNA Pseudoknots Using Context-Sensitive Pattern Matching. *International Conference on Hybrid Information Technology (ICHIT-2006)*, Korea, November 09-11, 2006, **1:** 660-665.

MOUNT, D.W. 2004. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York USA.

RAFAEL, G. 2006. Prediction of RNA Pseudoknotted Secondary Structure using Stochastic Context Free Grammars. *CLEI Electronic Journal*, **9(2):** 221-228.

ROBIN, D.D. and SEAN, R.E. 2006. Efficient Pairwise RNA Structure Prediction and Alignment Using Sequence Alignment Constraints. *Journal of BMC Bioinformatics*, **7:** 400-437.

SAAD, M. 2007. On the Approximation of Optimal Structures for RNA-RNA Interaction. *IEEE Transactions on Computational Biology and Bioinformatics*, **6(4):** 682-688.

SEARLS, D.B. 2002. The Language of Genes. *Nature*, **420:** 211-217.

WOODS, D.A. and BATZOGLOU, S. 2006. CONTRAfold: RNA Secondary Structure Prediction Without Physics-Based models. *Bioinformatics* **22(14):** 90-98.

YAN, D. and YULEI, Z. 2005. Statistical Parser For RNA Secondary Structure Prediction. *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics,* Guangzhou, 18-21 August 2005.

YINGLEI, S. 2004. Tree Decomposition Based Fast Search of RNA Structures Including Pseudoknots in Genomes. *Proceedings of IEEE Computational Systems Bioinformatics Conference,* Stanford, August 16-19, 2004.

YUKI, K., HIROYUKI, S. and TADAO, K. 2003. A Comparative Study on Formal Grammars for Pseudoknots, *Proceedings of Genome Informatics*, **14:** 470-471.

Received 15 March 2009
Accepted 17 May 2010