International Journal of Information Technology and Knowledge Management
January-June 2012, Volume 5, No. 1, pp. 107-110

ANALYZING THE AMBIGUITY IN RNA STRUCTURE USING PROBABILISTIC APPROACH

Rajesh Kumar Varun[1] and Sunil Gupta[2]

ABSTRACT: RNA is the second major form of nucleic acid in human cells that play intermediary role between DNA and functional protein. Several classes of RNA's are found in cells, each with distinct function. Understanding of storage and utilization of a cell's genetic information is based on the structure of RNA. Many experimental results have shown that RNA plays a greater role in the cells. RNA sequences contains signals at the structure level can be exploited to detect functional motifs common to all or a portion of those sequence. Different types of analysis of a structure can provide functional information in different degrees of detail. In this paper various types of RNA secondary structure representation has been discussed and in which appropriate structure has been adopted for probabilistic approach that shows un-ambiguity.

Keywords: Secondary Structure, Stochastic Context-free Grammar, 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 mainly four 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 $\Sigma = \{A, G, C, U\}$. 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 no. 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 protein are assembled.

According to Noam Chomsky the Context free grammar (CFG) has very much importance in Linguistic field, Computer Sc. and Engineering and in 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 [1]. A CFG is also called Type 2 Grammar similar to a regular grammar but permits a greater variety of production rules. One purpose of this paper is to present an effective method for estimating a stochastic context-free grammar to model a family of RNA sequences [2]. Determining RNA shapes has gained considerable importance in the last decade because knowing the shape of the molecule is essential for researchers to understand its role within a cell. The RNA plays a very important character in bio cells. A lot of work has been done in structural analysis of RNA in bioinformatics field but there exist a large number of challenging problems. However structural analysis of RNA is still a challenging problem in bioinformatics field [3], [4]. The structure of an RNA molecule is closely related to its function [5]. 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 [6][8][9]. Such Algorithm are typically called RNA folding algorithm.

2. STRUCTURAL ANALYSIS FOR RNA

The importance of grammars in compliers is known to everyone. The grammars are useful tools to model character sequences and in a certain way these tools are useful to model molecular biological sequences [10]. 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 derivation tree. This tree reflects the syntactical structure of a sequence. It is possible that for a given sequence there may be more than on derivation tree. In this case, we say that the grammar is ambiguous. In ambiguous grammar, complexity for the

[1,2] Assistant Professor Department of Information Technology Northern India Engineering College, New Delhi
E-mail: errajesh.pec@gmail.com

derivation rises given that the possible trees grow 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. 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 does not consider the context in which it is situated.

For example, one of the applications of productions in Fig. 1 can generate the RNA sequence "AGCGUCAGUGACUU GAUGCU" by the following derivation and the equivalent derivation tree is shown in Fig. 7.

3.1. Productions

$P = \{S_0 \rightarrow S_1,$ $S_7 \rightarrow AS_8U$

$S_1 \rightarrow AS_2U,$ $S_8 \rightarrow GS_9U$

$S_2 \rightarrow GS_3C,$ $S_9 \rightarrow US_{10}$

$S_3 \rightarrow CS_4G,$ $S_{10} \rightarrow GS_{11}$

$S_4 \rightarrow GS_5U,$ $S_{11} \rightarrow AS_{12}$

$S_5 \rightarrow US_6A$ $S_{12} \rightarrow C\}$

$S_6 \rightarrow CS_7G$

Figure 1: Set of Production Rules P

Fig.1 shows set of productions rules P that generates RNA sequence for a certain restricted structure, in which S_0, S_1……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 steps 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 similar derivation steps and replacing it with the right hand side of an appropriate production, we obtain the following derivation terminating with the desired sequence:

3.2. Derivation

$S_0 \rightarrow S_1$

$\rightarrow AS_2U \quad (S_1 \rightarrow AS_2U)$

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

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

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

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

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

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

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

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

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

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

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

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 (Fig. 2) is a text-based representation. The structure is re?ected 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 (Fig. 3). It is the most prominent 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 RNAs 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 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 mainly used in this paper due to its bene?cial 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 (Fig. 5) Arcs are plotted as curved lines inside this circle. The mountain plot (Fig. 6) is useful for large RNAs. Plateaus represent unpaired regions,the heights of these mountains are determined by the number of base-pairs in which the partial sequences are embedded. Fig. 7 shows derivation tree for a given sequence and Fig. 8 shows appropriate way representation of sequence.

4.1. Dot-Bracket Representation

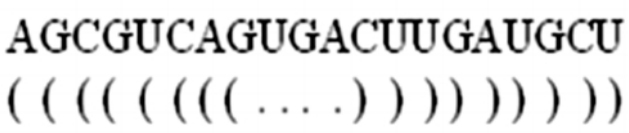

Figure 2:

4.2. Squiggle Plot

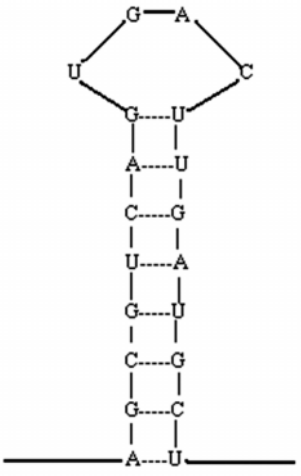

Figure 3:

4.3. Arc-Annotated Sequence

Figure 4:

4.4. Circle Representation

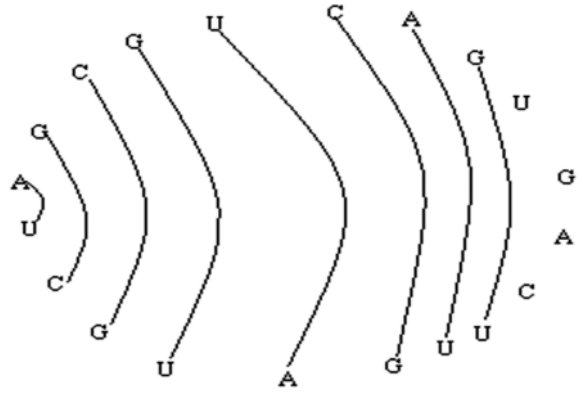

Figure 5:

4.5. Mountain Plot Representation

Figure 6

4.6. Derivation Tree Representation

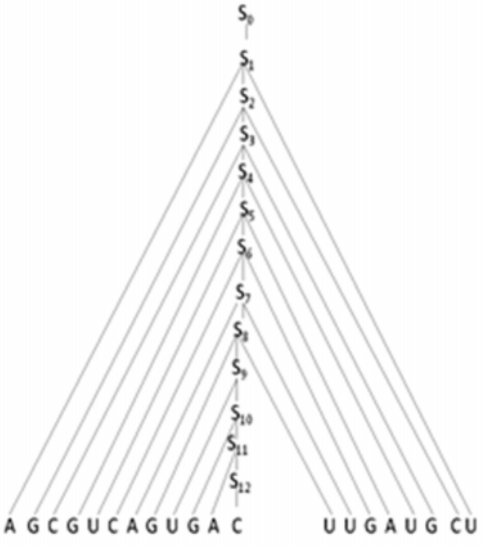

Figure 7:

4.6. Most Appropriate Way Representation

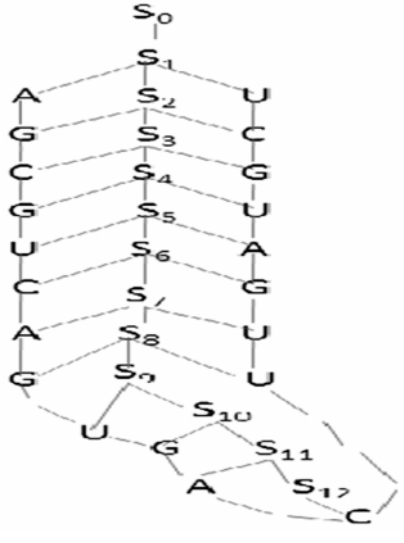

Figure 8:

5. ASSIGNING PROBABILITY OVER GRAMMER

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. One such method of assigning the probabilities is shown in Fig. 7.

To derive the trained grammar, we designed the initial grammar by using some prior knowledge about the RNA family.

Table1
Probabilities for the Type 2 Grammar, we Placed Uniform Distribution over Each Set of Same Type Production

Category of Probabilities ns	Productions	Productions
C#1	S0 → S1	1.000
C#2	S1→ AS2U	0.125
C#2	S2 → GS3C	0.125
C#2	S3 → CS4G	0.125
C#2	S4 → G S5U	0.125
C#2	S5 → US6A	0.125
C#2	S6 → CS7G	0.125
C#2	S7 → AS8 U	0.125
C#2	S8 → GS9U	0.125
C#3	S9 → US10	0.333
C#3	S10 → GS11	0.333
C#3	S11 → AS12	0.333
C#4	S12 → C	1.000

6. CONCLUSION

A detailed understanding of the functional 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 than the actual sequence . One of the problems with CFGs is that it generally has an ambiguity in the grammar that results more than one parse tree for a sequence, and alternative parse tree reflect alternative secondary structure, a grammar often gives several possible secondary structure 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 a (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 for the variations of RNAs like mRNA.

REFERENCES

[1] Byung-Jun Yoon and P.P. Vaidyanathan V, "Computational Identification and Analysis of Noncoding RNAs", IEEE Signal Processing Magazine, pp. 64-74, Jan 2007.

[2] Yuki Kato, Hiroyuki Sek, Tadao Kasami, "A Comparative Study on Formal Grammars for Pseudoknots", Proceedings of Genome Informatics, 14, pp.470-471, 2003.

[3] Jizhen Zhao, Liming Cai, and Russell L. Malmberg, "Learning the Parameters of Stochastic Grammar Models for RNA Structures with Pseudoknots", IEEE Transactions, pp. 170-175, 2006.

[4] Hiroshi Matsui, Kengo Sato, Yasubumi Sakakibara, "Pair Stochastic Tree Adjoining Grammars for Aligning and Predicting Pseudoknot RNA Structures", Journal of Bioinformatics, 21, No. 11, pp. 2611-2617, 2005.

[5] Yinglei Song et. al., "Tree Decomposition Based Fast Search of RNA Structures Including Pseudoknots in Genomes", Proceedings of IEEE Computational Systems Bioinformatics Conference, 2004.

[6] Rafael García, "Prediction of RNA Pseudoknotted Secondary Structure using Stochastic Context Free Grammars", CLEI Electronic Journal, 9 No. +2, December 2006.

[7] Keum Y. Sung, "Recognition and Modeling of RNA Pseudoknots Using Context-Sensitive Pattern Matching", International Conference on Hybrid Information Technology (ICHIT), 2006.

[8] Michael Brown, Charles Bilson, Santa Cruz, "RNA Pseudoknot Modeling Using Intersection of SCFG with Application to Database Searches", 1995.

[9] Robin D. Dowell, Sean R Eddy, "Efficient Pairwise RNA Structure Prediction and Alignment Using Sequence Alignment Constraints", Journal of BMC Bioinformatics, 2006.

[10] Yan Dang, Yulei Zhang, "Statistical Parser for RNA Secondary Structure Prediction", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.

[11] Saad Mneimneh, "On the Approximation of Optimal Structures for RNA-RNA Interaction", IEEE Transactions on Computational Biology and Bioinformatics, 2007.