A Comparative Approach to RNA Pseudoknotted Structure Prediction Based on Multiple Context-Free Grammar

Hiroyuki Seki Nobuyoshi Mizoguchi Yuki Kato

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
Email: {seki,ykato}@is.naist.jp

1 Introduction

Multiple context-free grammar (mcfg) [10] is a natural extension of context free grammar (cfg) and inherits many good properties of cfg. For example, the class of languages generated by mcfg (called multiple context-free languages or *mcfl*) is a substitution closed full AFL and the membership problem for mcfl L is solvable in $O(n^e)$ time where n is the length of an input string and e is a constant determined by an mcfg that generates L.

Recently, formal language theory has been applied to structure prediction of biological sequences. For example, an RNA can be regarded as a string over four symbols (or bases) A, U, C, G, which is called a *primary sequence*. An RNA takes a folding structure called a *secondary structure*, which is made by base pairs such as A-U and C-G. There is close co-relation between the secondary structure of an RNA and its function, and so predicting the secondary structure of a given RNA primary sequence is an important problem. If the secondary structure consists of simple substructure called stem-loop only, then the structure can be modeled by a derivation tree of cfg. In this case, the prediction can be realized by a parsing algorithm for cfg. However, there is another important substructure called *pseudoknot*, which cannot be described by cfg. Hence, there have been a few studies on secondary structure prediction based on a grammar of which generative power is stronger than cfg [11, 7]. The authors have applied a parsing algorithm for mcfl to RNA secondary structure prediction [3, 4]. However, these methods need a fare amount of training data for parameter setting. Recently, we proposed a prediction method based on comparative sequence analysis, which does not require training data [6]. In the presentation, we would like to give a quick review of mcfg, followed by the experimental results on secondary structure prediction for eight families of real RNA sequences [6].

2 Related Work

Methods for secondary structure prediction are classified into two categories, namely, methods based on single sequence analysis and those based on comparative sequence analysis. Formal grammars are simple and natural models for describing topological constraints on base pairs. In particular, stochastic context-free grammar (scfg) is well-known as a formal model for pseudoknot free structures, and the optimal secondary structure can be predicted in $O(n^3)$ time where n is the length of an input sequence [1, 9]. For secondary structure including pseudoknots, several grammars of which expressive power is stronger than cfg have been used such as (extended) simple linear tree adjoining grammar [11], RNA pseudoknot grammar [7] and mcfg [3].

In comparative sequence analysis approaches, an algorithm takes as input several homologous sequences that are expected to fold into the same structure, and utilizes evolutionary information (characteristic substructures) conserved in input sequences to improve its prediction accuracy. For the pseudoknot free case, a grammatical approach has already been extended to comparative sequence analysis. For example, Pfold [5] is based on scfg parsing and uses base pair rate matrix to determine the probability parameters of scfg. For structures including pseudoknots, there have been a few non-grammatical approaches. For example, hxmatch [12] predicts a consensus pseudoknotted structure, using a maximum weighted matching algorithm and postprocessing the result to produce a bisecondary structure. ILM [8] uses a similar approach.

3 Experimental Results

We conducted experiments using a prototype prediction tool based on the proposed method, and compared the prediction performance of the proposed method with hxmatch [12], which is one of the best

Table 1: Prediction accuracy in F-measure (%)[6].

Family	# of sequences	Length	hxmatch	Proposed method
Corona_pk3	14	67	89.5	**97.1**
TMV_UPD-PK3	27	33	87.5	**88.9**
TMV_UPD-PK2	3	22	**100.0**	82.4
SBWMV2_UPD-PKb	3	29	**90.9**	81.9
UPD_PK2	6	22	**83.3**	66.7
UPSK	6	23	**60.0**	57.1
PK_IAV	32	48	N/A	**48.0**
Prion	167	57	**47.6**	0.0
Average			69.9	65.3

Note: hxmatch produced no prediction on PK_IAV.

comparative sequence analysis methods that can predict consensus pseudoknotted structures.

The data sets for experiments were taken from the Rfam database [2], where we selected eight families that have simple pseudoknotted structures. The number of sequences of these eight families is from 3 to 157 and the length is from 21 to 67. Table 1 shows the prediction results of hxmatch and the proposed method. As depicted in Table 1, the accuracy of the proposed method is more than 80% for the uppermost four out of the eight families and is almost comparable to hxmatch.

References

[1] S. R. Eddy and R. Durbin, "RNA sequence analysis using covariance model," *Nucleic Acids Research*, vol. 22, no. 11, pp. 2079–2088, 1994.

[2] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: Annotating non-coding RNAs in complete genomes," *Nucleic Acids Research*, vol. 33, pp. D121–D141, 2005.

[3] Y. Kato, H. Seki, and T. Kasami, "RNA pseudoknotted structure prediction using stochastic multiple context-free grammar," *IPSJ Transactions on Bioinformatics*, vol. 47, no. (SIG 17(TBIO 1)), pp. 12–21, 2006.

[4] Y. Kato, T. Akutsu and H. Seki, "A grammatical approach to RNA-RNA Interaction Prediction," *Pattern Recognition*, vol. 42, pp. 531–538, 2009.

[5] B. Knudsen and J. Hein, "Pfold: RNA secondary structure prediction using stochastic context-free grammars," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3423–3428, 2003.

[6] N. Mizoguchi, Y. Kato, and H. Seki, "A grammar-based approach to RNA Pseudoknotted Structure Prediction for Aligned Squences," 1st IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2011), pp. 135–140, 2011.

[7] E. Rivas and S. R. Eddy, "The language of RNA: a formal grammar that includes pseudoknots," *Bioinformatics*, vol. 16, no. 4, pp. 334–340, 2000.

[8] J. Ruan, G. D. Stormo, and W. Zhang, "An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots," *Bioinformatics*, vol. 20, no. 1, pp. 58–66, 2004.

[9] Y. Sakakibara, M. Brown, R. Hughey, K. I. S. Mian, Sjölander, R. C. Underwood, and D. Haussler, "Stochastic context-free grammars for tRNA modeling," *Nucleic Acids Research*, vol. 22, pp. 5112–5120, 1994.

[10] H. Seki, T. Matsumura, M. Fujii, and T. Kasami, "On multiple context-free grammars," *Theoretical Computer Science*, vol. 88, pp. 191–229, 1991.

[11] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori, "Tree adjoining grammars for RNA structure prediction," *Theoretical Computer Science*, vol. 210, pp. 277–303, 1999.

[12] C. Witwer, I. L. Hofacker, and P. F. Stadler, "Prediction consensus RNA secondary structures including pseudoknots," *IEEE Trans. Computational Biology and Bioinformatics*, vol. 1, no. 2, pp. 66–77, 2004.