

Efficient Tree Grammatical Modeling of RNA Secondary Structures from Alignment Data

Takashi Takakura¹, Hiroki Asakawa², Shinnosuke Seki¹,
Satoshi Kobayashi¹

Keywords: RNA secondary structure, RNA secondary structure alignment, Tree adjoining grammar, Probabilistic Grammar

1 Introduction

Recent advances of research on non-coding RNAs (nc-RNAs) reveal that they are more numerous and important than previously thought ([8][5], etc.). In order to find new members of nc-RNA families, there have been proposed various powerful grammatical tools which can model both primary sequences and secondary structure information ([7][4][9][2], etc.). In order to develop a full automatic grammatical system for finding new members of nc-RNA families, we select a tree adjoining grammar (TAG) as a model of RNAs. In this manuscript, we will give an efficient linear time algorithm which constructs probabilistic tree adjoining grammar (PTAG) models of RNA secondary structures from alignment data.

2 Methods

We first define a class of grammars, called alignment TAGs, which can be used to model RNAs³. Then, the problem of modeling RNAs from secondary structural alignment data is formulated as a decision problem whether there exists an alignment TAG which can appropriately represent a given RNA alignment data.

In order to solve this problem, we introduce a *virtual* RNA secondary structure which can be obtained by taking union of all base pairing information contained in the given RNA alignment data. Then, the modeling problem above is mathematically reduced to the problem of deciding whether the obtained virtual RNA secondary structure can be generated by a *universal* TAG or not, where a universal TAG is defined as a TAG whose adjunct trees contain only one nonterminal symbol.

Based on an efficient parsing technique of a universal tree grammar using structural information ([1]), we will give a linear time algorithm for solving this problem, which leads us to an efficient algorithm for constructing probabilistic TAG model from RNA alignment data. The work by Asakawa ([1]) also provides a theoretical characterization of the class of RNA secondary structures which can be modeled by TAGs, which is not discussed by Condon, Davy and Tarrant ([3]).

3 Preliminary Experiments

We constructed a PTAG from the alignment data of “corona_pk3” ([6]), which contains 61 RNA secondary structures. A randomly drawn 30 sequences were used to construct a PTAG,

¹Graduate School of University of Electro-Communications, E-mail: satoshi@cs.uec.ac.jp

²Hewlett Packard Japan, Ltd.

³In this manuscript, by TAG, we mean a special TAG, called TAG_{RNA}, which was proposed by the authors for a grammatical model of RNAs.

and the rest (31 sequences) were used for testing the classification accuracy of the obtained PTAG. The average length of “corona_pk” RNA sequences is 62. Therefore, in order to make negative RNA test samples, we collected 30 RNA sequences of similar length from Rfam Database and adjusted their lengths to 62 in a appropriate manner. The accuracy of the obtained PTAG was high, which is illustrated in Figure 1. Figure 1 represents the distribution of log-odds ratios. The horizontal and vertical axes represent log-odds scores and the number of sequences, respectively. The distribution of log-odds ratios for negative RNA test samples are shown in dashed lines. Although the number of sequences of this experiment is still small⁴, the effectiveness of the proposed efficient construction method of PTAGs was shown.

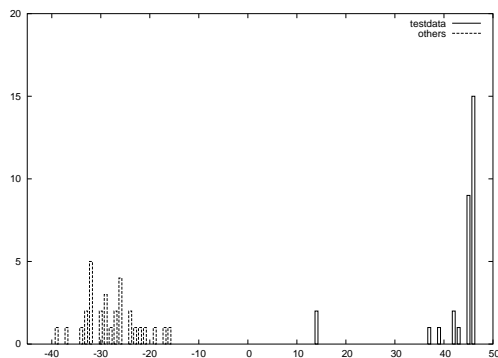


Figure 1: Distribution of Log-odds Ratios

References

- [1] Asakawa, H. 2003. Efficient parsing of universal tree grammars using structural information. master’s thesis, Graduate School of University of Electro-Communications. (in Japanese)
- [2] Cai, L., Malmberg, R. L. and Yunzhou, W. 2003. Stochastic modeling of RNA pseudoknotted structures: a grammatical approach. *Bioinformatics* 19:i66-i73.
- [3] Condon, A., Davy, B., and Tarrant, F. 2004. Classifying RNA pseudoknotted structures. *Theoretical Computer Science* 320:35-50.
- [4] Eddy, S. R. and Durbin, R. 2002. RNA sequence analysis using covariance models. *Nucleic Acids Res.* 22:2079–2088.
- [5] Eddy, S. R. 2002. Computational genomics of noncoding RNA genes. *Cell* 109:137–140.
- [6] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna A., and Eddy, S. R. 2003. Rfam: an RNA family database. *Nucleic Acids Res.* 31:439–441.
- [7] Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., sjolander, K., Underwood, R. C. and Haussler, D. 1994. stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.* 22:5112-5120.
- [8] Storz, G. 2002. An expanding universe of noncoding RNAs. *Science* 296:1260–1263.
- [9] Uemura, Y., Hasegawa, A., Kobayashi, S. and Yokomori, T. 1999. Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science* 210:277–303.

⁴More efficient TAG parsing algorithm might be necessary in order to test the accuracy of the obtained PTAGs from large set of long RNA sequences. Currently, we use an $O(n^5)$ time PTAG parsing algorithm for the accuracy test experiment.