Predicting and Learning RNA Secondary Structures

Aki HASEGAWA
haseg-ak@gold.cs.uec.ac.jp

Yasuo UEMURA
uemura-y@gold.cs.uec.ac.jp

Satoshi KOBAYASHI
satoshi@cs.uec.ac.jp

Takashi YOKOMORI
yokomori@cs.uec.ac.jp

Department of Computer Science and Information Mathematics,
University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo 182, Japan

Abstract

It is of great significance to develop an efficient software system for higher-level structural prediction in RNA/protein sequences. Speaking of RNA secondary structure prediction, it is inevitably required that a prediction system must have an ability to deal with so-called "pseudoknot" structures, one of the most typical and important constructs found in vivo, while no effective system is yet reported for predicting RNA secondary structures involving in pseudoknots.

We are developing prediction systems for RNA secondary structures that can handle pseudo-knots in an elegant manner, where the developing systems are constructed based on the following two ways.

Prediction System Using Tree Grammars : In the previous work[3], we developed a parsing algorithm for RNA modeling grammars, and applied it to some computational problems concerning RNA secondary structures. In this work, we prototype a prediction system for RNA secondary structures which is based on minimizing the free-energy change associated with base pairings and loop structures. One of the major advantages of this system is its capability of predicting structures which include pseudoknots.

We evaluate the system by presenting how long sequences it can process and how long and how much memory it takes to find the solution. We also make comparisons between biologically observed structures and prediction results.

Learning Secondary Structures Using Control Sets : We propose a learning system that acquires the common RNA secondary structures of the given set of RNA sequences.

In order to capture the common biological features of RNA secondary structures, it is possible to make use of derivational information obtained from parsing analysis by the prediction system described above.

More specifically, the key idea of our method is outlined as follows:

Initially we set up the most general tree grammar $G_0 = (C, A, \{S\})$, where every nonterminal node in each tree of C or A is labeled by exactly one symbol S, and A contains all types of uniquely labeled adjunct trees. Thus G_0 has no restriction on which adjunct trees to use at any configuration within the derivation.

Then, given RNA sequences are parsed by the prediction system (parser), to get the control words (i.e., words over the alphabet of labels of all trees).

Our learning system for RNA secondary structures consists of the following procedures:

1. Set up the most general tree grammar G_0,
2. Parse the given sequence with G_0 using the parser mentioned above,
3. Construct a set of control words,
4. Learn the control set.

In this manner, given a set of RNA sequences, the learning system produces a control set with which one can construct a prediction system together with G_0. In other words, one may have another type of prediction system using control set.

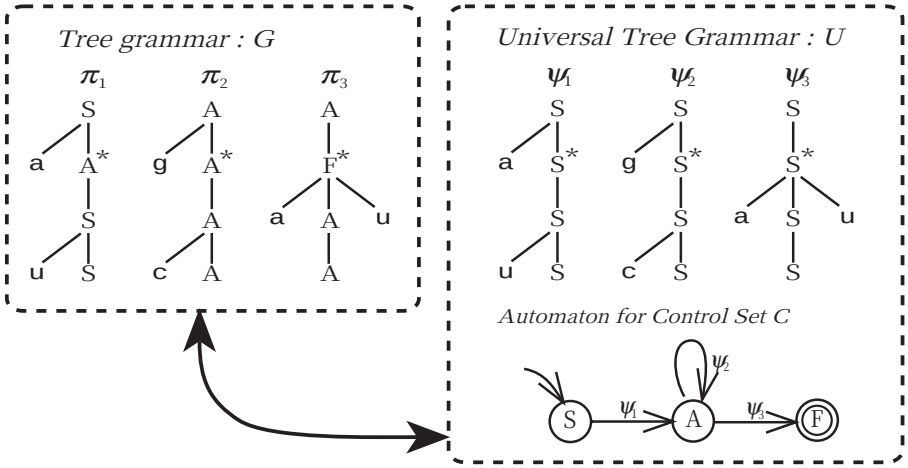

Figure 1: Tree Grammars with Control Set

Acknowledgements

This work was supported in part by Grants-in-Aid for Scientific Research No.08283103 from the Ministry of Education, Science and Culture, Japan.

References

[1] S. Kobayashi and T. Yokomori, "Modeling RNA Secondary Structures Using Tree Grammars," In *Proceedings of Genome Informatics Workshop V*, Universal Academy Press, pages 29-38, 1994.

[2] Y. Takada, "Grammatical Inference for Even Linear Languages Based on Control Sets," Information Processing Letters 28, pages 193-199, 1988.

[3] Y. Uemura, A. Hasegawa, S.Kobayashi and T.Yokomori, "Grammatically Modeling and Predicting RNA Secondary Structures," In *Proceedings of Genome Informatics Workshop VI*, Universal Academy Press, pages 67-76, 1995.