Fast RNA structural comparison using coarse-grained base-pairing probabilities

Yuki Kato[1] ykato@is.naist.jp
Jakob Hull Havgaard[2] hull@rth.dk
Jan Gorodkin[2] gorodkin@rth.dk

[1] Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
[2] Center for non-coding RNA in Technology and Health (RTH), Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 3, 1870 Frederiksberg C, Denmark

Keywords: RNA secondary structure, base-pairing probability, comparative sequence analysis

1 Introduction

Non-coding RNAs (ncRNAs) have received considerable attention since they have significant roles in living cell such as regulation of gene expression. There is a strong correlation between the function of ncRNA and its structure, and thus prediction of RNA structures is of great importance. RNA structural alignment given two or more RNA sequences is a good approach to structural analysis when sequences that are expected to be homologous are available. This comparative approach boosts screening putative ncRNA regions with structural similarities from genomic sequences [3]. Since the comparative approach such as simultaneous folding and aligning RNA sequences is generally costly in run-time and memory for large genomes, we need to know the potential structured regions in advance as fast as possible. In this abstract, we present a preliminary technique to find structurally similar RNAs by comparing coarse-grained base-pairing probabilities of the RNAs.

2 Methods

Given two RNA sequences, we first compute each base-pairing probability matrix by a partition function-based method implemented by `RNAfold` [4] in the ViennaRNA package [5]. From the base-pairing probability matrix, we construct a binary vector that represents coarse-grained base-pairing probabilities. Finally, the two binary vectors are compared by a standard global alignment technique like the Needleman–Wunsch algorithm [6]. If the alignment score is high, these two RNAs are considered to be structurally similar.

3 Results

We created a data set of 46 families with 1,649 RNA sequences from Rfam 11.0 [2]. The main feature of this set is that structure conservation index (SCI) [8] of the sequences in each family is around 1, indicating that the structure is highly conserved in the family. To evaluate the proposed method, we performed all-against-all comparison of sequences in the form of binary vectors in the data set to have alignment scores. We then carried out clustering using the Markov cluster algorithm [7] that takes these scores as input. The clustering results show high specificity, meaning that any two sequences from distinct families are highly likely to be classified into different clusters by the clustering that uses our computed scores.

4 Discussion

This work deals only with global comparison between two relatively short sequences via binary vectors. In real application, we need to know the potential structured regions in genomic sequences, which can be inferred from local base-pairing probabilities [1]. We are now addressing the next step based on the technique presented in this abstract.

References

[1] Bernhart, S.H., Hofacker, I.L. and Stadler, P.F. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **22**, 614–615 (2006).

[2] Burge, S.W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **41**, D226–D232 (2013).

[3] Gorodkin, J. and Hofacker, I.L. From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Comput. Biol.* **7**, e1002100 (2011).

[4] Hofacker, I.L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188 (1994).

[5] Lorenz, R. *et al.* ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

[6] Needleman, S.B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

[7] van Dongen, S. Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. & Appl.* **30**, 121–141 (2008).

[8] Washietl, S., Hofacker, I.L. and Stadler, P.F. Fast and reliable prediction of noncoding RNAs. *P. Natl. Acad. Sci. USA* **102**, 2454–2459 (2005).