RNA Secondary Structural Alignment with Conditional Random Fields

Article *in* Bioinformatics · October 2005

DOI: 10.1093/bioinformatics/bti1139 · Source: PubMed

Kengo Sato
Keio University
90 PUBLICATIONS 1,535 CITATIONS

Yasubumi Sakakibara
Keio University
203 PUBLICATIONS 3,169 CITATIONS

RNA Structural Alignment with Conditional Random Fields

Kengo Sato[1] **Yasubumi Sakakibara**[1]
satoken@bio.keio.ac.jp yasu@bio.keio.ac.jp

[1] Keio University, Department of Biosciences and Informatics, 3–14–1 Hiyoshi, Kohoku-ku, Yokohama, 223–8522, Japan

Keywords: RNA secondary structure, structural alignment, conditonal random fields

1 Introduction

Computationally identifying non-coding RNA regions on the genome has much attention to be investigated. However, it is essentially harder than gene-finding problems for protein-coding regions because non-coding RNA sequences do not have a strong statistical signals. Since comparative sequence analysis is effective for non-coding RNA detection, efficient computational methods are expected for structural alignment of RNA sequences. [5] has proposed one of such algorithms, called pair hidden Markov models on tree structures (PHMMTSs), which can calculate a structural alignment of a binary tree and a sequence, and has applied PHMMTSs to aligning RNA secondary structures, that is, pairwise alignment to align an unfolded RNA sequence into an RNA sequence of known secondary structure.

To calculate structural alignments of RNA sequences requires some parameters, the substitution probability of base pairs and the state transition probability, which have much affect on the performance of structural alignments. There are some related works to estimating the parameters for aligning RNA secondary structures. For example, [3] have proposed a ribosomal RNA substitution matrix, called RIBOSUM, which is based on an analogous method to the BLOSUM matrices. However, since the RIBOSUM matrix is based on the maximum likelihood estimation by relative frequencies of RNA mutations, it requires a large number of high-quality structure-annotated alignments to avoid overfitting. Therefore, more effective methods for estimating the parameters for RNA structural alignment should be developed.

2 Method

We propose a novel approach for estimating the parameters for RNA structural alignment with Conditional Random Fields (CRFs). CRFs proposed by [4] have several advantages over traditional HMMs and stochastic grammars because CRFs can build discriminative and flexible state transition models from annotated training data. Therefore, our approach has a specific feature compared with previous methods in the sense that the parameters for structural alignment are estimated such that the model can discriminate between correct alignments given by the training data and incorrect alignments most likely.

3 Experimental Results

To confirm our method, we have done some experiments. A data set used for training and test in our experiments is extracted from the “SEED” sequences in Rfam RNA family database [2] which are biologically plausible. First, we have trained the parameters for structural alignment according to CRF’s training method, and then, structurally aligned RNA sequences with trained parameters. The result of them has been evaluated by specificity and sensitivity of predicted base pairs. Our experimental result clearly shows that the parameter estimation with CRFs can outperform some other existing methods for structural alignment of RNA sequences such as the EM algorithm, RIBOSUM [3] and the score matrix used by FOLDALIGN [1]. Further, we have done more practical experiments for predicting non-coding RNA regions by local alignment searches with a folded sequence of the target RNA on genome sequences. The result shows that structural alignment search based on CRFs is more efficient for predicting non-coding RNA regions than other scoring methods. These experimental results strongly support our discriminative method employing CRFs to estimate the score matrix parameters. This method will provide biologists explicit guidance for exploring non-coding RNA regions. In future, we would further improve the accuracy of predicting RNA secondary structures and non-coding RNA regions by employing more complex features for CRFs such as high-order dependencies of states, although only the first-order state transitions are used in our experiments.

References

[1] Gorodkin, J., Heyer, L. J., and Stormo, G. D. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Research*, 25(18):3724–3732, Sept. 1997.

[2] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. Rfam: an RNA family database. *Nucleic Acid Research*, 31(1):439–441, Jan. 2003. `http://www.sanger.ac.uk/Software/Rfam/`.

[3] Klein, R. J. and Eddy, S. R. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4(44), 2003.

[4] Lafferty, J., McCallum, A., and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[5] Sakakibara, Y. Pair hidden Markov models on tree structures. *Bioinformatics*, 19(Supplement 1):i232–i240, July 2003.

[6] Sato, K. and Sakakibara, Y. RNA secondary structural alignment with conditional random fields. *Bioinformatics*, 21 Suppl 2:ii237–ii242, Sep 2005.