

Multiple Alignment of RNAs by Maximizing the Sum of the Matching Probability of Stem Fragments

Yasuo Tabei^{1,2}

tabei@cb.k.u-tokyo.ac.jp

Hisanori Kiryu^{2,3}

kiryu-h@aist.go.jp

Taishin Kin²

taishin@cbrc.jp

Kiyoshi Asai^{1,2}

asai@k.u-tokyo.ac.jp

- ¹ Department of Computational Biology, Graduate School of Frontier Science, University of Tokyo, CB04 Kiban-tou 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan
- ² Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology
- ³ Graduate School of Information Sciences, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

Keywords: alignment, non-coding RNA

1 Introduction

Non-coding RNAs (ncRNAs) are RNA molecules which are transcribed but do not encode proteins. The functions of them seem relevant to the secondary structure rather than the sequence similarity. The ordinary tools considering only sequence similarity e.g. ClustalW [3] are unable to align the sequences which have low sequence identities. Therefore, the alignment for RNA sequences have to take the structural information into account.

The Sankoff's algorithm [10], which simultaneously allows the solution of the structure prediction and alignment problem, requires $O(n^{3m})$ in time and $O(n^{2m})$ in memory for m sequences of length n . Therefore, the algorithm is not applicable to long RNA sequences. Several variants of the Sankoff's algorithm have been proposed, which restrict the distances of the base pairs in the primary sequences and is applicable to pairwise alignment only [4, 5, 6, 8].

2 Method and Results

In recent year, maximum expected accuracy (MEA) alignment method, which is known as optimal accuracy alignment, has been proposed[7]. First, the method computes the posterior probability, $P(x_i \sim y_j \in a^* | x, y)$, that particular positions x_i and y_j of two sequences, x and y , respectively, will be matched in an alignment a^* , based on Pair-HMM's forward-backward algorithms. Then, the Needleman-Wunsch alignment with these posterior probabilities as substitution scores is done. ProbCons is one of the most accurate alignment software for multiple amino acid sequences using MEA alignment method [1]. MEA decoding algorithm also has a successful application to secondary structure prediction for RNA sequences [2].

In this article, we show an efficient multiple alignment method based on maximizing the sum of the matching probability of stem candidates. The method is an extension of the pairwise alignment method of our previous work to progressive multiple alignment [11]. The MEA alignment method is also used. However, computing the posterior probability, $P(x_i \sim y_k, x_j \sim y_l \in a^* | x, y)$, that particular positions x_i, x_j, y_k and y_l of two sequences x and y , respectively, will be matched in an alignment a^* is computationally intractable, because Pair-SCFG's inside-outside algorithm is $O(n^6)$ in time

and $O(n^4)$ in memory for a pair of sequences of length n . So, we devised an approximate posterior probability of $P(x_i \sim y_k, x_j \sim y_l | x, y)$ which is computed based on Pair-HMM's posterior probability, $P(x_i \sim y_k | x, y)$ and the base pairing probability, $P(x_i \diamond x_j | x)$, for each two sequences, x and y , which is computed by means of McCaskill's algorithm [9].

In benchmarking experiments, we will show that our alignment method is the most accurate in other state-of-the-art methods and the computational time is reasonable enough to be applicable to large scale analyses.

References

- [1] Do, C. B., Mahabhashyam, M. S., Brudno, M., and Batzoglou, S., ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res.*, 15(2):330–340, 2005.
- [2] Do, D. B., Woods, D. A., and Batzoglou, S., CONTRAfold:RNA secondary structure prediction without physics-based models, *Bioinformatics*, 22(14):e90–e98, 2006.
- [3] Thompson, J. D., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 27:2682–2690, 1999.
- [4] Havgaard, J. H., Lyngsø, R. B., Stormo, G. D., and Gorodkin, J., Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%, *Bioinformatics*, 21(9):1815–1824, 2005.
- [5] Hofacker, I. L., Bernhart, S. H., and Stadler, P. F., Alignment of RNA base pairing probability matrices, *Bioinformatics*, 20(14):2222–2227, 2004.
- [6] Holmes, I., A probabilistic model for the evolution of RNA structure, *BMC Bioinformatics*, 5(166), 2004.
- [7] Holmes, I. and Durbin, R., Dynamic programming alignment accuracy, *J.Comput.Biol.*, 5:493–504, 1998.
- [8] Holmes, I. and Rubin, G. M., Pairwise RNA structure comparison with stochastic context-free grammars, *Pac Symp Biocomput*, pages 163–174, 2002.
- [9] McCaskill, J. S., The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers*, 29:1105–1119, 1990.
- [10] Sankoff, D., Simultaneous solution of the RNA folding, alignment, and proto-sequence problems, *SIAM J. App. Math.*, 45:810–825, 1985.
- [11] Tabei, Y., Tsuda, K., Taishin, K., and Asai, K., SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments, *Bioinformatics*, 22:1723–1729, 2006.