

# Comparing RNA sequences based on the alignment of stem candidates

Kiyoshi Asai,<sup>1</sup> Yasuo Tabei,<sup>2</sup> Taishin Kin<sup>3</sup>

**Keywords:** RNA, secondary structure, stem, alignment

## 1 Introduction.

Recent study on comparative genomics revealed that the conserved regions between genomes contain large amount of non-coding sequences. One of the important challenge is to discover as many candidates of non-coding RNAs as possible with high accuracy by sequence information analyses.

Comparison of the two sequences is the most important foundation of all sequence analyses. That is necessary for the annotation of the functional elements by the help of sequence search in the database, clustering of the sequences and comparing the two genome sequences. In most cases, the two sequences are aligned by dynamic programming (DP) using the scores based on a substitution matrix and a gap penalty function. The substitution matrix can be replaced by position dependent score matrices if the local property of the sequence are known. The local property can be the secondary structures, positional distributions of amino acids and so on.

In order to compare two RNA sequences, however, it is not easy to align RNA sequences because the standard alignment algorithms based on the simple DP are unable to treat the base pair restrictions in the secondary structures. If the two sequences have the known common secondary structure, it is possible to align them by the correspondence of the secondary structures. If the common secondary structure is unknown, however, the problem is much more difficult. Because the predictions of the secondary structure are not accurate in general, it is not desirable to use each predicted structure for further analyses. Alignment based methods in RNA sequence comparison with secondary structure restrictions are generally computationally expensive, even if they ignore the pseudoknot structures.

There are several attempt to combine the sequence comparison and the estimation of the common secondary structures with lower computational costs [1, 2]. The authors have also proposed a kernel method on a SCFG for RNA sequence analyses [3].

However, for the purpose of sequence search in the whole genome, for example, a faster alignment algorithm that considers the secondary structure restrictions is required. We have implemented an alignment based algorithm that consider all the possible stem candidates in both sequences and that align those candidates in a simple manner.

## 2 The Algorithm

We have implemented DP algorithm that aligns the potential stem candidates of the sequences. Stem Candidate Aligner for RNA (Scarna) works as follows.

---

<sup>1</sup>Department of Computational Biology, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, Japan. & Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, Japan. E-mail: [asai@k.u-tokyo.ac.jp](mailto:asai@k.u-tokyo.ac.jp)

<sup>2</sup>Department of Computational Biology, University of Tokyo. E-mail: [tabei@cb.k.u-tokyo.ac.jp](mailto:tabei@cb.k.u-tokyo.ac.jp)

<sup>3</sup>Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology. E-mail: [taishin@cbrc.jp](mailto:taishin@cbrc.jp)

- (a) Calculation of Stem Candidate Sequence (SCS) for each RNA sequence
- (b) Alignment of two SCSs
- (c) Alignment of remaining (loop) regions
- (d) Removing inconsistent stems

A Stem Candidate (SC) corresponds to one of the two strands of a potential stem and consists of (1) the position in the sequence, (2) the string, (3) the distance to the stem partner (reverse complementary string) in the sequence, (4) the string of the partner, (5) the score of the stem. For example, if 'gcua' that begins from position 11 potentially forms a stem with 'uagc' that begins from position 23, there are two SCs corresponding to this potential stem, (11, *gcua*, 8, *uagc*, 0.53) and (23, *uagc*, -8, *gcua*, 0.53). Because the stacking energy of the stem candidate can be calculated without knowing the information of other stems, the score of the stem are assigned based on that energy. A Stem Candidate Sequence (SCS) is a sequence of all possible fixed-length SCs, sorted by their positions.

Selecting the correspondence of the potential stems means a structural alignment in RNA sequences. The strict solution requires the complexity of Pair Stochastic Context Free Grammars (PSCFGs), but we apply much simpler DP in Scarna. The SCSs are simply aligned by a two dimensional DP. The match score of the two corresponding SCs are defined based on the string similarity, the difference of the distance to the stem partner, and the scores of the both stems based on stacking energy. Longer stems are treated as the consecutive matches of the fixed length stems. Because the 5' strand and the 3' strand of each stem are aligned separately in SCS alignment, the inconsistent stems are removed as a post-process.

### 3 Discussions

The time complexity of Scarna's algorithm is the product of the length of stem candidate sequences. If we restrict the distance of the stem candidates to fixed length, the lengths of the stem candidate sequences are linear to the length of the sequence, and the time complexity is about  $O(L^2)$  for the sequence length  $L$ . There is no guarantee of consistency of stem matches in this algorithm, but it gives the lower bound of the score because it is guaranteed that no higher scoring consistent match exists. At the same time, pseudoknotted structures are accepted without paying any additional time costs.

Scarna web server is located in <http://www.scarna.org/> where the users can submit RNA sequences and display the predicted stem matches.

### References

- [1] Perriquet, O., Touzet, H. and Dauchet, M. 2003. Finding the common structures shared by two homologous RNAs, *Bioinformatics*, 19(1):108–116.
- [2] Hofacker, I.L., Bernhart, S.H.F and Stadler, P.F. 2004. Alignment of RNA Base Pairing Probability Matrices *Bioinformatics* 20: 2222-2227.
- [3] Kin, T., Tsuda, K. and Asai, K. 2002. Marginalized Kernels for RNA sequence Data Analysis, *Genome Informatics*, 13:112–122.
- [4] Sakakibara, Y. 2003. Pair hidden Markov models on tree structures, *Bioinformatics*, 19(Suppl.1):i232-i240.
- [5] Tsuda, K., Kin, T. and Asai, K. 2002. Marginalized kernel for biological sequences, *Bioinformatics*, 18(Suppl.1):S268-S275.