

Murlet: A Practical Multiple Alignment Tool for Structural RNA Sequences

Hisanori Kiryu^{1,2}

kiryu-h@aist.go.jp

Yasuo Tabei³

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

Taishin Kin¹

kin-taishin@aist.go.jp

Kiyoshi Asai^{1,3}

asai@k.u-tokyo.ac.jp

- ¹ Computational Biology Research Center, The National Institute of Advanced Industrial Science and Technology (AIST), Tokyo Waterfront Bio-IT Research Building 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan
- ² Graduate School of Information Sciences, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan
- ³ Department of Computational Biology, Faculty of Frontier Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

Keywords: RNA secondary structure, multiple alignment

1 Introduction

Structural RNA genes show unique evolutionary patterns to conserve their secondary structures, that should be taken into account for constructing accurate multiple alignments of RNA genes. The Sankoff algorithm is a natural alignment algorithm which supports the base pair covariance effect in the alignment model. However, its extremely high computational cost forbids applying the algorithm to most of RNA sequences. We propose an efficient algorithm for multiple sequence alignment of structural RNA sequences. Our algorithm is a variant of the Sankoff algorithm [1] with an efficient scoring system that considerably reduces the time and space requirements without the degradation of alignment quality. Our algorithm first computes the match probability matrix that measures the alignability of each position pair between sequences, and the base pairing probability matrices for each sequence. Then these probabilities are combined to score the alignment by the Sankoff algorithm. We show that both the alignment quality and the accuracy of the consensus secondary structure prediction from the alignment are the highest among the examined alignment programs. The algorithm is implemented as the software “Murlet”.

2 Method

2.1 The Model

We consider the consensus secondary structure annotation \mathcal{S} for each pairwise alignment \mathcal{A} of length L , which consists of sequences x and y of lengths L_x and L_y , respectively.

$$\mathcal{S} = \mathcal{S}_{\mathcal{A}} = \{\mathcal{L}, \mathcal{P}\}$$

$$\mathcal{L} = \{i \in \mathcal{C} \mid \text{column } i \text{ does not form any base pair}\}$$

$$\mathcal{P} = \{(i, j) \in \mathcal{PC} \mid \text{columns } (i, j) \text{ form a base pair}\}$$

where the set of match columns \mathcal{C} is the set of alignment columns without gap characters, and $\mathcal{PC} = \{(i, j) \in \mathcal{C} \times \mathcal{C} \mid 1 \leq i < j \leq L\}$ is the set of pairs of match columns \mathcal{C} . We consider only the cases that

all base pairs are formed within the match columns. We also ignore pseudo-knotted structures. We assign a score e_L to each loop column $i \in \mathcal{L}$ and a score e_S to each column pair $(i, j) \in \mathcal{P}$

$$e_L(x_i, y_i) = \gamma_L p^{(a)}(x_i, y_i) q^{(b)}(x_i) q^{(b)}(y_i) \quad (1)$$

$$\begin{aligned} e_S(x_i, y_i, x_j, y_j) &= \gamma_S p^{(a)}(x_i, y_i) p^{(a)}(x_j, y_j) \\ &\quad \times p^{(b)}(x_i, x_j) p^{(b)}(y_i, y_j) \\ &\quad \times \exp(s(x_i, y_i, x_j, y_j)) \end{aligned} \quad (2)$$

where x_i and y_i represent the sequence positions of two sequences x and y , aligned at the alignment column i . $s(x_i, y_i, x_j, y_j)$ is an element of the base pair substitution matrix. γ_L and γ_S are the constant coefficients. The match probability $p^{(a)}(k, l)$ is the posterior probability that sequence positions k and l will be matched in an alignment and is calculated using the standard pair hidden Markov model (PHMM) of sequence alignment [2]. $p^{(b)}(k, l)$ is the base pairing probability of sequence positions k and l in the same sequence which is calculated by the McCaskill algorithm [3]. $q^{(b)}(k)$ is the loop probability at position k .

For each alignment \mathcal{A} and its consensus structure candidate \mathcal{S} , the alignment score $z = z(\mathcal{A}, \mathcal{S})$ is defined as the sum of loop match scores e_L and stem match scores e_S .

$$z = \sum_{i \in \mathcal{L}} e_L(x_i, y_i) + \sum_{(i, j) \in \mathcal{P}} e_S(x_i, y_i, x_j, y_j)$$

The maximum value $z = z_{\max}$ among all alignments and secondary structures is calculated by the Sankoff algorithm. The corresponding alignment \mathcal{A} is our alignment result.

Because the loop match score e_L and the base pair match score e_S are both proportional to the match probability $p^{(a)}$, we can restrict the $L_x \times L_y$ DP matrix to a smaller region out of which there are no positions (k, l) with match probability $p^{(a)}(k, l)$ larger than the specified threshold. For sufficiently low threshold value, the restriction of DP region does not change the alignment result, hence is considered to be an *exact* reduction method. If two sequences are highly similar, the match probabilities concentrate along a specific diagonal in the DP matrix and the reduction of DP region is quite significant.

3 Results

We show that Murelet performs the best in terms of the secondary structure prediction and the alignment accuracy among the examined multiple alignment programs. Moreover, Murelet can align ten sequences of the SRP_euk_arch family of mean length 291, while the memory consumption of other Sankoff-based programs such as Stemloc and PMMulti blow up for sequences above 100 nucleotides and these programs cannot align sequences above 200 nucleotides in realistic memory and time.

References

- [1] Sanko, D., Simultaneous solution of the RNA folding, alignment and protosequence problems, *SIAM J. Appl. Math.*, 45(5):810–825, 1985.
- [2] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., Biological sequence analysis, *Cambridge University Press*, 1998.
- [3] McCaskill, J. S., The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers*, 29(6-7):1105–1109, 1990.