SIMPLE ALIGNMENT CONSTRAINTS FOR EFFICIENT ALIGNMENT OF RNA SEQUENCES USING FAMILY-SPECIFIC MODELS

Byung-Jun Yoon

Dept. of Electrical & Computer Engineering
Texas A&M University
Collge Station, TX 77843-3128, USA
Email: bjyoon@ece.tamu.edu

ABSTRACT

In this extended abstract, we present a simple method for finding alignment constraints that can be used for efficient alignment of RNAs based on family-specific models, such as profile context-sensitive HMMs (profile-csHMMs) and covariance models (CMs). The alignment constraints are established based on the alignment positions predicted by a profile-HMM. Application of the proposed constraints to the profile-csHMM based structural RNA alignment method significantly improved the average alignment speed without degrading the alignment accuracy.

1. INTRODUCTION

Many noncoding RNAs (ncRNAs) are known to conserve the base-paired secondary structure as well as the primary sequence [1]. Therefore, when aligning RNA sequences, it is important to consider both structure and sequence similarities to obtain an accurate and biologically meaningful alignment [2]. Conservation of the secondary structure gives rise to pairwise base correlations in the RNA sequence. A typical problem of most RNA alignment algorithms is the high computational cost that results from analyzing these correlations. For example, the optimal alignment algorithm (called the CYK algorithm) for covariance models (CMs) [3] has a computational complexity of $O(L^3)$, where L is the length of the RNA to be aligned. Simultaneous RNA structure prediction and alignment algorithms have an even higher complexity, where the computational cost for pairwise folding and alignment of RNAs amounts to $O(L^6)$ [4]. The high computational cost limits the utility of RNA alignment algorithms in practical applications, and there have been extensive research efforts to develop heuristic methods for making these algorithms faster without sacrificing the accuracy [5, 6, 7, 8, 9].

In this extended abstract, we propose a novel method for finding effective sequence alignment constraints that can improve the computational efficiency of RNA alignment based on family-specific models, such as profile-csHMMs (profile context-sensitive HMMs) [10] and covariance models [3].

2. MOTIVATION

Let us assume that we have constructed a profile-csHMM (or a CM) to represent a set of RNAs that belong to the same family. The constructed model can be used to search for similar RNAs in a sequence database or predicting the secondary structure of a homologous RNA with an unknown structure. How can we make the alignment algorithm for these models faster?

One popular approach is to restrict the search space for finding the optimal alignment, thereby reducing the overall computation. For example, consider aligning two RNAs $\mathbf{x} = x_1 x_2 \cdots x_{L_1}$ and $\mathbf{y} = y_1 y_2 \cdots y_{L_2}$. A recent version of Dynalign [9], a pairwise RNA structure prediction and alignment algorithm, assumes that the bases x_m and y_n in the respective RNAs can be aligned to each other only if they satisfy the following condition:

$$\left| \frac{L_2}{L_1} m - n \right| \leq M. \tag{1}$$

The parameter M is used to specify the maximum distance between the alignable bases. This constraint restricts the solution space for the optimal alignment between $\mathbf{x}$ and $\mathbf{y}$, making the algorithm significantly faster compared to the unconstrained one. A more recent implementation of Dynalign [6] takes a more principled approach, where they try to estimate the possible alignment positions based on the sequence similarity between the RNAs. In [6], a general pair-HMM is used to estimate the set of (m, n), whose "co-incidence probability" $P(x_m \leftrightarrow y_n | \mathbf{X}, \mathbf{Y})$ exceeds a certain threshold λ:

$$\mathcal{S} = \Big\{ (m,n) \Big| P(x_m \leftrightarrow y_n | \mathbf{x}, \mathbf{y}) > \lambda \Big\}. \tag{2}$$

The estimated set $\mathcal{S}$ is used to constrain the final alignment [6]. Although these methods have been proposed to speed up simultaneous RNA folding and alignment algorithms, we can take a similar approach to expedite RNA alignment algorithms based on family-specific models.

However, when we are interested in a specific RNA family, it will be more appropriate to establish the alignment constraints based on the member sequences in the given family. In this case, it is desirable to use a family-

Table 1. Average sensitivity (SN) and positive predictive value (PPV) of the predicted RNA alignments.

RNA FAMILIES	PROFILE-CSHMM						PSTAG	
	M-CONSTRAINT		ORIGINAL		PROPOSED			
	SN (%)	PPV (%)	SN (%)	PPV (%)	SN (%)	PPV (%)	SN (%)	PPV (%)
CORONA_PK3	95.5	95.7	95.7	96.5	94.8	96.0	94.6	95.5
HDV_RIBOZYME	94.5	95.1	94.5	95.3	94.2	95.9	94.1	95.6
TOMBUS_3_IV	95.9	96.4	95.9	96.4	96.8	97.4	97.4	97.4
FLAVI_PK3	94.6	96.5	94.5	96.4	94.5	96.8	N/A	N/A

Table 2. Average CPU time for finding an RNA alignment.

RNA FAMILIES	PROFILE-CSHMM			PSTAG
	M-CONSTRAINT	ORIGINAL	PROPOSED	
	TIME (SEC)	TIME (SEC)	TIME (SEC)	TIME (SEC)
CORONA_PK3	9.37	0.71	0.23	19.65
HDV_RIBOZYME	10.30	1.03	0.13	158.77
TOMBUS_3_IV	6.99	0.35	0.07	193.06
FLAVI_PK3	13.31	3.96	0.35	N/A

specific model (e.g., profile-HMM), for finding the constraints, rather than using a general model (e.g., pair-HMM) that applies to all RNAs. However, the problem of this approach is that, in many practical situations, we may not have enough number of sequences in the given family for reliably estimating the model parameters. Although the alignment constraint in (2) is expected to work well when we can accurately estimate the base alignment probabilities, this approach is not suitable when only a handful of RNAs are available for training the model, since accurate estimation of the probabilities will not be possible. Then how can we find effective sequence alignment constraints for a family-specific model when we have only a limited number of sequences in the reference RNA family?

Unlike the alignment probability, the predicted alignment positions in an optimal sequence alignment are not very sensitive to small parameter changes. As a result, two different HMMs with reasonably similar parameters often yield nearly identical alignment results. This motivates us to exploit the predicted alignment positions for establishing the constraints, instead of using the base alignment probabilities.

3. PROPOSED METHOD

Based on the previous motivation, we propose a simple method for finding effective alignment constraints for fast and accurate alignment of RNAs using profile-csHMMs. The proposed method computes the alignment constraints based on alignment positions predicted by a sequence-based alignment. Assume that we have a set of homologous RNAs, based on which we construct a profile-csHMM. Given a new RNA (the "target RNA") with an unknown structure, we want to predict its optimal alignment to the model using a dynamic programming algorithm, called the SCA (sequential component adjoining) algorithm [10]. Our goal is to find good sequence alignment constraints that can improve the speed of the SCA algorithm. The proposed method can be summarized as follows:

1. Build a profile-HMM based on the same reference RNA family.

2. Find the best alignment between the profile-HMM and the target RNA.

3. Find all matched positions that have at least γ consecutive matches before and after them. These positions will be fixed when predicting the optimal structural alignment using the profile-csHMM.

4. For regions that are bounded by fixed alignment positions, estimate the maximum distance between the aligned bases. This will be used to restrict the alignable positions when finding the final structural alignment.

4. RESULTS

To demonstrate the idea, we applied the proposed constraints to the profile-csHMM based structural alignment method [10]. For our experiments we chose four RNA families with pseudoknot structures from Rfam [11]. Based on these RNA families, we performed cross-validation experiments using the following methods: (i) profile-csHMM + proposed constraints, (ii) profile-csHMM with the "M-constraint" defined in (1), (iii) profile-csHMM (original implementation in [10]), and (iv) pair stochastic tree adjoining grammars (PSTAG) [12]. Table 1 shows the average sensitivity and PPV (positive predictive value) of the predicted structural alignment and Table 2 shows the average CPU time for finding an alignment. As we can clearly see from Table 1 and Table 2, using the proposed constraints significantly improved the alignment speed without degrading the alignment accuracy. More details and further analysis of the proposed method can be found in our manuscript recently submitted to the EURASIP Journal on Bioinformatics and Systems Biology [13].

5. CONCLUSION

In this work, we presented a simple method for finding effective alignment constraints that can make RNA structural alignment based on profile-csHMMs faster. Although our focus was on profile-csHMM based RNA alignment, the proposed scheme can also be applied to CM-based methods.

6. REFERENCES

[1] S. R. Eddy, "Non-coding RNA genes and the modern RNA world", *Nature Reviews Genetics*, vol. 2, pp. 919-929, 2001.

[2] B.-J. Yoon and P. P. Vaidyanathan, "Computational identification and analysis of noncoding RNAs - Unearthing the buried treasures in the genome", *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 64-74, Jan. 2007.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*, Cambridge Univ. Press, Cambridge, UK, 1998.

[4] D. Sankoff, "Simultaneous solution of the RNA folding, alignment, and protosequence problems," *SIAM Journal on Applied Mathematics*, vol. 45, pp. 810-825, 1985.

[5] R. D. Dowell and S. R. Eddy, "Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints," *BMC Bioinformatics*, 7:400, 2006.

[6] A. O. Harmanci, G. Sharma, and D. H. Mathews, "Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign," *BMC Bioinformatics*, 8:130, 2007.

[7] J. H. Havgaard, E. Torarinsson, and J. Gorodkin, "Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix," *PLoS Computational Biology*, 3(10):e193, 2007.

[8] I. Holmes, "Accelerated probabilistic inference of RNA structure evolution," *BMC Bioinformatics*, 6:73, 2005.

[9] A. V. Uzilov, J. M. Keegan, D. H. Mathews, "Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change," *BMC Bioinformatics*, 7: 173, 2006.

[10] B.-J. Yoon and P. P. Vaidyanathan, "Structural alignment of RNAs using profile-csHMMs and its application to RNA homology search: Overview and new results," *IEEE Transactions on Automatic Control (Joint Special Issue on Systems Biology with IEEE Transactions on Circuits and Systems: Part-I)*, vol. 53, pp. 10-25, Jan. 2008.

[11] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy and A Bateman, "Rfam: annotating non-coding RNAs in complete genomes," *Nucleic Acids Res.*, vol. 33, pp. D121-D124, 2005.

[12] H. Matsui, K. Sato, and Y. Sakakibara, "Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures", *Bioinformatics*, vol. 21, pp. 2611-2617, 2005.

[13] B.-J. Yoon, "Efficient alignment of RNAs with pseudoknots using sequence alignment constraints," *EURASIP Journal on Bioinformatics and Systems Biology*, submitted.