Bioinformatics Advance Access published May 11, 2006

BIOINFORMATICS

SCARNA: Fast and Accurate Structural Alignment of RNA Sequences by Matching Fixed-length Stem Fragments

Yasuo Tabei [a]*, Koji Tsuda [b], Taishin Kin [b], Kiyoshi Asai [ab]

[a] Department of Computational Biology, Graduate School of Frontier Science, University of Tokyo, CB04 Kiban-tou 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan, [b] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: The functions of non-coding RNAs are strongly related to their secondary structures, but it is known that a secondary structure prediction of a single sequence is not reliable. Therefore, we have to collect *similar* RNA sequences with a common secondary structure for the analyses of a new non-coding RNA without knowing the exact secondary structure itself. Therefore, the sequence comparison in searching *similar* RNAs should consider not only their sequence similarities but their potential secondary structures. Sankoff's algorithm predicts the common secondary structures of the sequences, but it is computationally too expensive to apply to large-scale analyses. Because we often want to compare a large number of cDNA sequences or to search similar RNAs in the whole genome sequences, much faster algorithms are required.

Results: We propose a new method of comparing RNA sequences based on the structural alignments of the fixed-length fragments of the stem candidates. The implemented software, SCARNA (Stem Candidate Aligner for RNAs), is fast enough to apply to the long sequences in the large-scale analyses. The accuracy of the alignments is better or comparable to the much slower existing algorithms.

Availability: The web server of SCARNA with graphical structural alignment viewer is available at http://www.scarna.org/

Contact: scarna@m.aist.go.jp

Supplementary information: The data and the supplementary information are available at http://www.ncrna.org/papers/SCARNA/.

1 INTRODUCTION

One of the important foundation of biological sequence analyses is comparing the sequences by the alignment with similarity scores. For the analyses of non-coding RNAs, we want to compare the nucleotide sequences for many purposes, such as finding the relatives of a new non-coding RNA in the genome, classification of the cDNA sequences and so on. The standard sequence comparison methods are not accurate enough for RNA sequences, however, because the secondary structures play important role in the functions and the evolutions of non-coding RNAs (Eddy, 2001). The alignments and the similarity scores of RNA sequences should consider both the primary sequences and the secondary structures.

Therefore, it is natural to try to find the common secondary structures in order to align two RNA sequences. Secondary structure prediction for a single sequence of length n without considering pseudoknots requires $O(n^2)$ in memory and $O(n^3)$ in time for computation (Nussinov *et al.*, 1978; Zuker and Stiegler, 1981). The structural RNA alignment is computationally so expensive even if the pseudoknots are ignored. The Sankoff's algorithm (Sankoff, 1985), which simultaneously allows the solution of the structure prediction and alignment problem, requires $O(n^4)$ in memory and $O(n^6)$ in time for a pair of sequences of length n. Such an algorithm is applicable only for short RNAs and not for all of the functional RNA sequences. By restricting the distances of the base pairs in the primary sequences it can be reduced to $O(n^4)$ in time (Havgaard *et al.*, 2005; Hofacker *et al.*, 1994) but it is still impractical for long sequences. In order to compare the RNA sequences without aligning them, a kernel method on Stochastic Context Free Grammar (SCFG) is proposed (Kin *et al.*, 2002).

Another way of finding the common secondary structures is to use *stem-based* representations, where the structure of an RNA sequence is represented by a number of sets of continuous base pairs. For constructing a stem-based representation, one approach is to use the predicted secondary structures (Karklin *et al.*, 2005). The predictions are not always accurate, however, that approach has a high risk of using totally wrong secondary structures. A more robust approach is to use a number of *stem candidates* derived by the simple Watson-Crick base-pair rule or by scanning the base-pair probability matrix (McCaskill, 1990). Those representations may contain many false stems, which have to be excluded in the end. Selecting the correspondence of the potential stems in two RNA sequences is also a combinatorial problem. DP (Dynamic Programming) can solve the problem, but it requires time complexity of $O(m^6)$ for the number of potential stems m even if pseudoknots are not considered.

Perriquet *et al.* (2003) proposed a fast heuristic stem-based algorithm implemented in CARNAC. It first determines anchor regions that are highly conserved in given RNA sequences and then seeks a set of the other stems that have minimum folding energy. Bafna *et al.* (2005) also proposed a stem-based method, which is similar to Sankoff's algorithm, implemented in RNAscf. It differs also in the way of finding structurally conserved anchors, which is based on secondary structure similarities. Ji *et al.* (2004) proposed a graph theoretic approach which finds conserved stems of multiple RNA sequences, implemented in comRNA.

In this paper, we propose an efficient pairwise alignment method based on fixed-length *stem fragments*. The fragments are made by

*to whom correspondence should be addressed

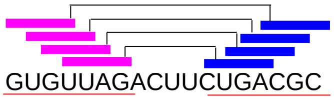


Fig. 1. Stem fragments. A stem candidate (marked by a red underline) is decomposed into 4 overlapping stem fragments. One fragment consists of a left component (red box) and a right component (blue box).

dividing the stem candidates in a number of overlapping fixed-length windows (Fig. 1). In order to align the stem fragments strictly, we need to adopt a computationally expensive algorithm. Instead, we decouple $5'$ and $3'$ parts of the stem fragments and use a pairwise alignment algorithm. The $5'$ parts (*left components*) and the $3'$ parts (*right components*) of the stem fragments are fixed-length subsequences of the stems that are complementary to each other. Because such a simple pairwise alignment does not guarantee the consistency of matches in both side (left and right), we get a certain number of mismatched components. In our approach, the common secondary structure is made only from the matches that include both left and right components. After the matched stem fragments are fixed, a pairwise alignment algorithm with affine gaps is used to make complete nucleotide alignment of two sequences.

To make our algorithm work on practical data, it is important to ensure the discovery of *true* stem fragments that belong to the common secondary structure. For high specificity in component matching, we designed the matching score of components based on various properties, e.g., sequence similarity, stacking energy and the distance to the complementary component. Unlike the other stem-based representations, the fixed-length fragments enable efficient computations of the matching score. The computation of matching scores of two variable-length sequences requires another alignment including gaps, which leads to an inefficient algorithm. One may think that it is difficult to take the stacking energy into account by fixed-length representations of the stems. We devised an engineered dynamic programming algorithm that includes the stacking energy in the score function.

In benchmarking experiments, we will show that our alignment accuracy is comparable to state-of-the-art methods, and the computational time is shorter by orders of magnitude.

2 METHOD

SCARNA takes two unaligned nucleotide sequences as the inputs and produces the alignment of the sequences based on the predicted common secondary structure. For efficiency, the stem fragments are first aligned, and the nucleotide-level alignment is made by a post-processing. In the following, our algorithm is explained step-by-step.

2.1 Extracting Stem Candidates

We start by representing the potential secondary structure of each RNA sequence by a set of overlapping stem fragments. To this aim, the base-pair probability matrix, is computed by means of McCaskill's algorithm (McCaskill, 1990). When the sequence has n bases, that matrix has $n \times n$ values, each of which represents the probability that the two bases form a base pair as a part of the whole secondary structure. When k is the fixed length of stem fragments (typically 2 to 5), the matrix is scanned by a counter-diagonal window of length k. If all the values in the window is larger than the threshold τ, that window is chosen as a part of a stem candidate, which is a stem fragment.

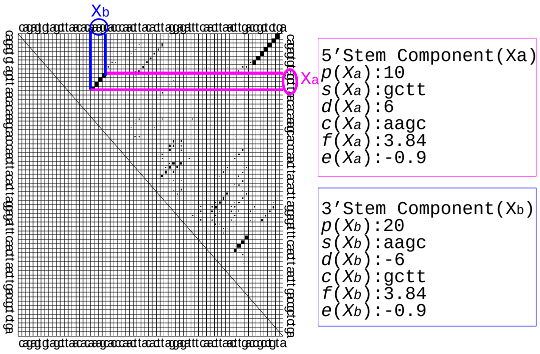


Fig. 2. An example of decoupling the selected stem fragment into two stem components, 5’(left) stem components and 3’(right) stem components. Base pairing probability matrix of tRNA D38114.1 taken from Rfam database (Griffiths-Jones *et al.*, 2003).

2.2 Properties of Stem Components

Each fixed-length stem fragment is decomposed into two stem components, $5'$ (left) component and $3'$ (right) component, both of which have the same fixed length (Fig. 2). A stem component X_a has the following properties.

- Position $p(X_a)$: the position of the leftmost base of the stem component X_a.
- Sequence $s(X_a)$: the sequence of the stem component X_a.
- Loop distance $d(X_a)$: the distance to the complementary stem component of the same stem fragment along the nucleotide sequence. $5'$ (left) stem components have positive distances and $3'$ (right) stem components have negative distances.
- Partner sequence $c(X_a)$: the sequence of the complementary stem component.
- Confidence score $f(X_a)$: the sum of the base-pair probabilities in the fragment.
- Stacking energy $e(X_a)$: the sum of stacking energy in the fragment.

In order to perform pairwise alignment, the stem components have to be ordered as a sequence. A Stem Component Sequence (SCS) is a sequence of all stem components sorted by their positions in the nucleotide sequence. Multiple stem components can take exactly the same position and their complementary components have different positions. Such components are sorted according to the distance to the complementary component (loop distance).

2.3 Matching Score of Stem Components

The matching score is used as the similarity measure of the stem components in the alignment. Because our goal is to capture both the structure similarities and the sequence similarities, the similarities of the corresponding base pairs and the differences of the two loop distances are combined. Because we want to align the stem candidates that have higher scores by means of free energy, the confidence scores and the stacking energy are also considered.

Let us describe the two SCSs to be aligned as $\{X_i\}_{i=1}^{n}$ and $\{Y_j\}_{j=1}^{m}$, where X_i and Y_j describe i-th and j-th stem components in the two nucleotide sequences, respectively. We denote by $[X_a, X_{a'}]$ that a left component X_a and a right component $X_{a'}$ form a stem fragment. Also, (X_i, Y_j) denotes a matched pair across the sequences in the alignment of SCSs.

Because the stem components has a same fixed length, the sequence similarity is calculated in linear time by substitution probabilities using RIBOSUM (Klein and Eddy, 2003). Denote by $R(X_i, Y_i)$ the sum of RIBOSUM scores.

If we take all of those scores into account, the matching score $s(i, j)$ of two corresponding stem components (X_i, Y_j) can be written as

$$\begin{aligned} s(i,j) &= R(X_i, Y_j) + \eta_1 \left(f(X_i) + f(Y_j)\right) \\ &- \eta_2 (e(X_i) + e(Y_j)) - \eta_3 \sqrt{d(X_i) - d(Y_j)}. \end{aligned} \quad (1)$$

Because the confidence scores and the stacking energy are mutually correlated, and because we have to control the importance of the terms, the parameters η_1, η_2 and η_3 are used. The term of η_3 encourages the stems with similar loop distances to match.

2.4 Consistency in Alignment of Stem Components

Before explaining the DP algorithm for the alignment of stem components, we discuss on the consistency conditions for the stem components in the alignments. We discuss first on the stem components of a single nucleotide sequence, and then on each match of the stem components of two nucleotide sequences.

2.4.1 Consistency in a single SCS The major difference of the alignment of stem components from a pairwise sequence alignment is that only small number of the stem components are *selected* to be included in the alignment. For each nucleotide sequence, a large number of combinations of stem components mutually contradict and should not be included in the same alignment.

If two stem fragments do not overlap in the nucleotide sequence, there are three types of positions. If the two stem fragments $[X_a, X_{a'}]$ $[X_b, X_{b'}]$ do not overlap each other, they are:

- *parallel* if $p(X_a) < p(X_{a'}) < p(X_b) < p(X_{b'})$ or $p(X_b) < p(X_{b'}) < p(X_a) < p(X_{a'})$.
- *nested* if $p(X_a) < p(X_b) < p(X_{b'}) < p(X_{a'})$ or $p(X_b) < p(X_a) < p(X_{a'}) < p(X_{b'})$.
- *pseudoknotted* if $p(X_a) < p(X_b) < p(X_{a'}) < p(X_{b'})$ or $p(X_b) < p(X_a) < p(X_{b'}) < p(X_{a'})$.

All those three types, including pseudoknotted positions, are permitted in the alignment of SCSs.

If two stem fragments overlap in the nucleotide sequence, we can also find the three cases. Two stem fragments $[X_a, X_{a'}]$ and $[X_b, X_{b'}]$ are:

- *r-continuous* if X_a overlaps X_b, $X_{a'}$ overlaps $X_{b'}$ and satisfy

  $$r = p(X_b) - p(X_a) = p(X_{a'}) - p(X_{b'}). \quad (2)$$

  If two overlapping stem fragments appear in a same alignment, they should be a part of a longer stem and r-continuous (Fig. 3). A pair of *stacked* fragments are 1-continuous fragments.

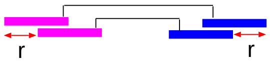


Fig. 3. An r-continuous pair of stem fragments. For any two stem fragments that are a part of a longer stem, the left and right components of the two stem fragments are shifted by r bases in opposite direction. If the margin r is different in each side, the fragments are ill-continuous.

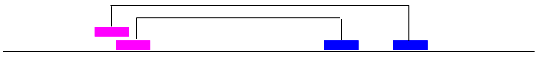


Fig. 4. An example of contradictory overlap of stem fragments. The left components overlap while the right components do not. In this case, the secondary structure cannot be uniquely determined at the nucleotide level.

- *ill-continuous* if X_a overlaps X_b and $X_{a'}$ overlaps $X_{b'}$, but (2) does not hold.
- *contradictory* if only one side (left or right) of the components overlap each other but the other side do not overlap (Fig. 4).

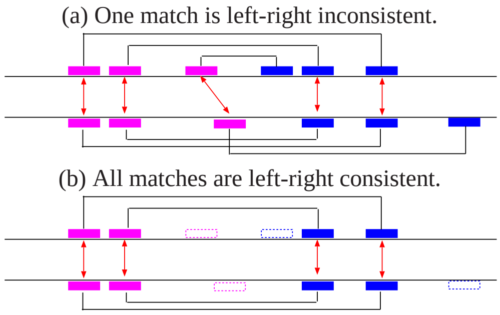
If two of the stem fragments of a nucleotide sequence appear in the alignment, they should not overlap at all or be r-continuous. The overlapping stem components are carefully treated in the SCS alignment that the left and right components of stem fragments satisfy r-continuous condition as explained later. Therefore, ill-continuous stem fragments never appear in our alignments. The non-overlapping stem components, however, may have overlapping complementary stem components because the left and right components of the stem fragments are separately aligned. Therefore, contradictory overlaps do happen in our alignments.

2.4.2 Consistency of matches of stem components Next we discuss on the matches of the stem components from two nucleotide sequences. For stem fragments $[X_a, X_{a'}]$ and $[Y_b, Y_{b'}]$ from two SCSs $\{X_i\}_{i=1}^{n}$ and $\{Y_j\}_{j=1}^{m}$, if X_a matches Y_b in the alignment, $X_{a'}$ should also match $Y_{b'}$. Let us define such a match as a *left-right consistent* match. Because the left component and the right component of each stem fragment are aligned separately, left-right consistency is not guaranteed in general. However, left-right consistent matches occur frequently because the matching scores encourage the matches of stem components that have similar loop distances.

2.4.3 Removing inconsistent matches The left-right inconsistent matches are removed after the SCS alignment (Fig. 5). If any two of the stem components of a same SCS appear in the SCS alignment and their complementary components overlap (i.e. contradictory overlap), those complementary components do not appear together in the alignment because the alignment of complementary components are controlled to be either non-overlapping or r-continuous. Therefore, the contradictory overlaps of the stem fragments are removed just by removing the left-right inconsistent matches of the components (Fig. 6).

2.5 Alignment Algorithm for Stem Components

The alignment of SCSs is computed by two DP matrix, $M(i, j)$ and $G(i, j)$. $M(i, j)$ is the best score up to a pair of X_i and Y_j


(a) One match is left-right inconsistent.

(b) All matches are left-right consistent.

Fig. 5. Removing the inconsistent matches. One of matches of the left components is left-right inconsistent because the corresponding right components do not match (a). By removing the components concerning the left-right inconsistent match (boxes by dotted lines), remaining matches represent the estimated common secondary structure (b).

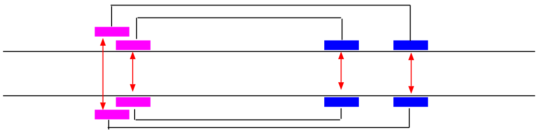


Fig. 6. Contradictory overlap and left-right consistency. The two pairs of right components (blue boxes) may appear together in the SCS alignment, but the two pairs of left components (red boxes) cannot appear together because they overlap without satisfying the r-continuous condition. Therefore, one of the matches in right components becomes left-right inconsistent.

given that X_i matches Y_j and $G(i,j)$ is the best score given that X_i mismatches Y_j.

The updates to derive $M(i,j)$ and $G(i,j)$ are described as

$$M(i,j) \quad = \quad max \begin{cases} M(\alpha_i,\beta_j) \\ \quad +(\delta_R(X_i)+\delta_R(Y_j)) \\ \quad +\eta_4(\delta_f(X_i)+\delta_f(Y_j)) \\ \quad -\eta_5(\delta_e(X_i)+\delta_e(Y_j)) \\ M(p_i,q_j)+s(i,j) \\ G(p_i,q_j)+s(i,j) \end{cases} \tag{3}$$

$$G(i,j) \quad = \quad max \begin{cases} M(i-1,j) \\ M(i,j-1) \\ G(i-1,j) \\ G(i,j-1) \end{cases} \tag{4}$$

with the initial conditions; $M(0,0) = 0, M(\cdot,0) = M(0,\cdot) = G(0,0) = G(\cdot,0) = G(0,\cdot) = -\infty$.

α_i is the index (smaller than i) of the component which is 1-continuous to X_i, β_j is that of Y_j. p_i is the index (smaller than i) of the nearest component which does not overlap with X_i, q_j is that of Y_j. η_4 and η_5 are control parameters.

In a simple DP for pairwise alignment, DP matrices depends on the adjacent elements. In our algorithm, however, $M(i,j)$ is derived from the remote elements denoted as $M(\alpha_i,\beta_j)$, $M(p_i,q_j)$ and $G(p_i,q_j)$. That ensures the adjacent matches of stem components in DP being either 1-continuous or non-overlapping (Fig. 7).

The updates (3) takes the maximum of three arguments. The first argument treats the case for continuous stems longer than the fixed length of stem fragments and ensures that the adjacent matches are 1-continuous. The indices α_i and β_j are determined such that $[X_{\alpha_i}, X_{\alpha_i'}]$ and $[Y_{\beta_j}, Y_{\beta_j'}]$ are 1-continuous fragments of $[X_i, X_{i'}]$ and $[Y_j, Y_{j'}]$, respectively. Since α_i and β_j do not always exist, the first argument in (3) is taken into account only if both α_i and β_j are available. Only the incremental parts of sequence similarities, the confidence scores and the stacking energy should be included in those continuous matches because 1-continuous matches share base pairs except one. The symbols, $\delta_R(X_i)$, $\delta_f(X_i)$ and $\delta_e(X_i)$, correspond to the incremental differences for X_i-X_{α_i} of RIBOSUM scores, confidence scores and stacking energy respectively(See Fig. 1 in supplement material for deltails).

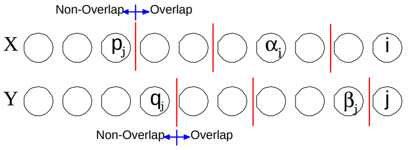


Fig. 7. Dependency in DP matrix computation. Shown are the two sequences of stem components. A group of components bounded by red poles have the same position in the original nucleotide sequence, but the different complementary components as their partners. The calculation of $M(i,j)$ depends on $M(\alpha_i,\beta_j)$, $M(p_i,q_j)$ and $G(p_i,q_j)$. The components X_{α_i} and Y_{β_j} form 1-continuous fragments with X_i and Y_j, respectively. The closest non-overlapping components from X_i and Y_j are denoted as X_{p_i} and Y_{q_j}, respectively.

The second and third arguments in (3) also take larger steps than a simple DP and ensure that the adjacent stem components have no overlap. X_{p_i} and Y_{q_j} are the closest components in SCSs that do not overlap with X_i and Y_j, respectively.

Because α_i and p_i for each i, β_j and q_j for each j, and the corresponding δ_R, δ_f and δ_e are calculated before the DP process in linear time, those calculations do not give any damage on time complexity of the algorithm.

2.6 Post-processing and Nucleotide Alignment

The inconsistency in the matches of stem components are removed as the post-process. In order to guarantee the consistency, it is sufficient to remove the left-right inconsistent matches of stem components as previously explained in section2.4.3. In other words, the matches of stem components whose complementary components do not match in the SCS alignment are removed as the post-process.

The nucleotide sequence alignment is intended to align remaining loop regions except the selected common stems represented by the consistent matches of stem components. It is simply implemented as a pairwise alignment with affine gaps of whole nucleotide sequences by adding a large value in DP matrix to the positions for the base pairs indicated by the SCS alignment. Those values are so large that the nucleotide alignment is forced to go through the positions and subtracted afterwards to get the score of the alignment.

3 RESULTS

In this section we show the performance of SCARNA for the alignment of RNA sequences by computational experiments on the benchmark dataset of tRNAs used by Gardner *et al.* (2005) and on the dataset from various families of non-coding RNA

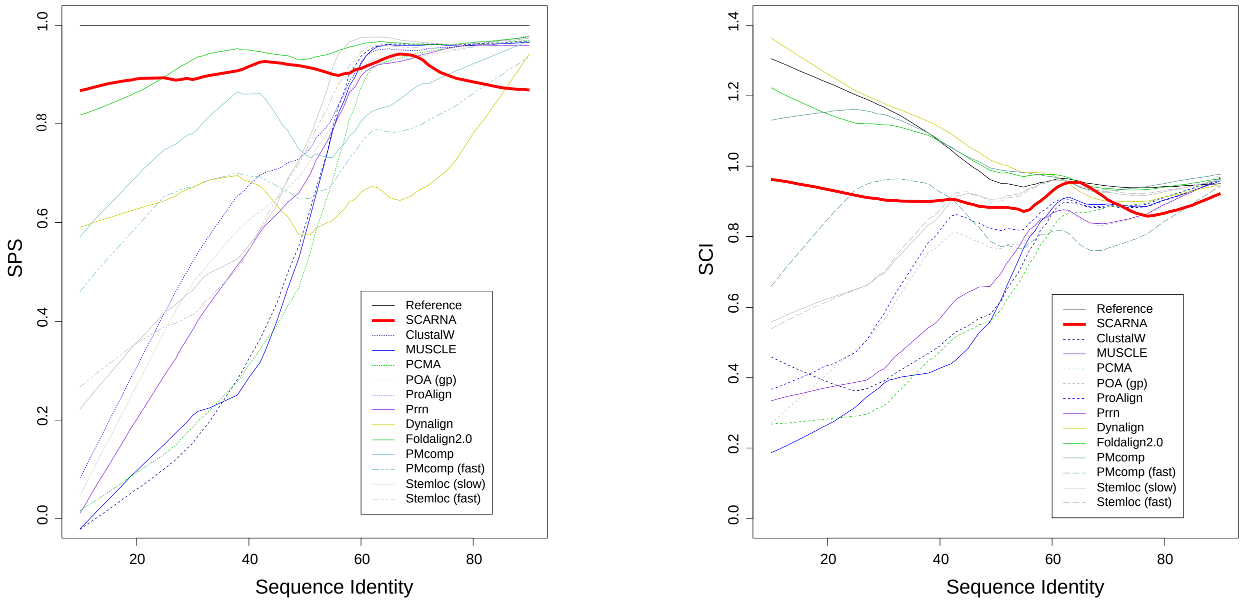


Fig. 8. SPS(left) and SCI(right) as functions of the sequence identity for Gardner's dataset of tRNAs. Lines are smoothed by lowess(local weighted regression) smoothing.

sequences in Rfam database (Griffiths-Jones *et al.*, 2003). It has been observed that SCARNA has a competitive performance to the other RNA structural alignment approaches using Sankoff's Algorithm (Sankoff, 1985).

We have evaluated the quality of the alignments by the sum-of-pairs score (SPS) and the structure conservation index (SCI) (Gardner *et al.*, 2005). The SPS is defined as the fraction out of all possible nucleotide pairs that are aligned both in the predicted alignment and in the alignment of the reference. The SPS provides a measure of the sensitivity of the prediction.

The structure conservation index (SCI) provides a measure of the conserved secondary structure information contained within the alignment (Washietl *et al.*, 2005). It is a derivative of the score calculated by the RNAalifold consensus folding algorithm (Hofacker *et al.*, 2002; Washietl and Hofacker, 2004) which is based upon the sum of the thermodynamic term and the covariance term. In contrast to the SPS, SCI is independent from a reference alignment. The SCI is close to zero if RNAalifold identifies no common RNA structure in the alignment, whereas a set of perfectly conserved structures has an SCI ≈ 1. The SCI points out the structural aspect of alignment accuracy and, therefore, a useful measure in addition to the SPS.

All the following tests were performed on a Linux machine with a AMD Opteron^{TM} Processor 850 2.4 GHz x 4 and 20GB RAM. The length of stem fragments k was set to 2. The threshold of base-pair probability τ was set to 0.0001. The control parameters, η_1, η_2, η_3, η_4 and η_5 were set to 3.7, 0.1, 3.1, 9.4 and 8.6, respectively. These parameters were used to all RNA familywise dataset. The command line options of other tools is listed on table1 in the supplement matrial.

3.1 Benchmark Dataset of tRNAs

Gardner's benchmark datasets (Gardner *et al.*, 2005) are composed of pairs of tRNA sequences that are classified by sequence identities. Though all the structural alignment programs are not able to align RNA sequences of more than 150 bases without any device, they can align those short tRNA sequences of 71.8 nucleotides in average. The sequences and the reference alignments for calculating the SPS were obtained from the Rfam database.

The experimental results are shown in Fig.8. The SPS and the SCI of SCARNA exceed those of sequence-based methods (e.g. ClustalW (Chenna *et al.*, 2003), MUSCLE (Edgar, 2004), PCMA (Pei *et al.*, 2003), POA (gp) (Lee *et al.*, 2002), ProAlign (Loytynoja and Sharlow, 2003) and Prrn (Gotoh, 1996)) and are comparable to those of structure-based methods (e.g. Dynalign (Mathews and Turner, 2002; Mathews, 2005), Foldalign2.0 (Havgaard *et al.*, 2005), PMcomp (Hofacker *et al.*, 2004) and Stemloc (Holmes and Rubin, 2002; Holmes, 2004, 2005)). While the sequence-based methods and structure-based ones have a dramatic divergence in relative performances below about 60% sequence identity, the SPS and the SCI of SCARNA do not come down. In particular, the SPSs of SCARNA outperform most of the structure-based methods in less than 50% sequence identity.

3.2 Benchmark Dataset of other non-coding RNAs

In order to evaluate our algorithm by longer non-coding RNAs, we made a benchmark dataset from 5S ribosomal RNA, 5.8S ribosomal RNA and Hammerhead ribozyme in the Rfam (Griffiths-Jones *et al.*, 2003). The collection of reliable sequences is a difficult task. We only used the 'seed' alignments of Rfam because the 'full' alignment includes computationally collected sequences. It is ovserved that even the 'seed' alignments includes questionable ones. We tried to filter out unreliable sequences whose alignments include inconsistend asignment of base pairs on gaps or very long gaps. We compared SCARNA to ClustalW and Foldalign2.0. The result is that the SPS and SCI of SCARNA can be compared to those of Foldalign2.0, favorably. (See supplement material, Fig.2, Fig.3 and Fig.4 for details).

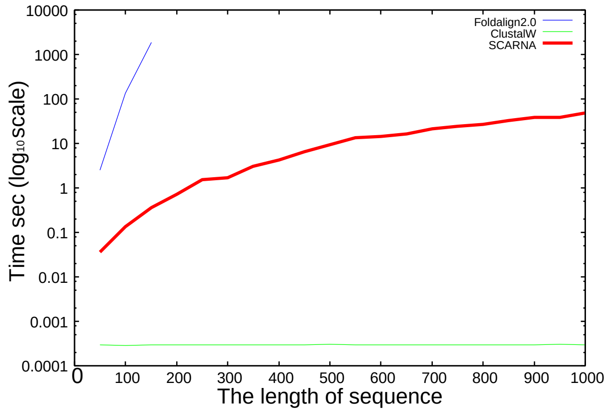


Fig. 9. Execution time comparison among SCARNA, ClustalW and Foldalign. The figure shows the execution time against the length of sequence.

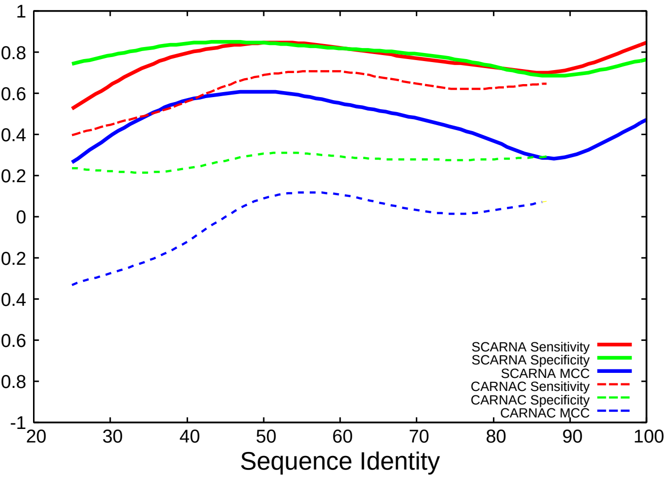


Fig. 10. Comparison between SCARNA and CARNAC concerning secondary structure prediction. Lines are smoothed by spline interpolation.

4 DISCUSSION

4.1 Computational Complexity

While the complexities of $O(n^3)$ in time and $O(n^2)$ in memory for secondary structure prediction of RNA sequences of length n are affordable in most cases, $O(n^6)$ and $O(n^4)$ in Sankoff's algorithm for structural alignment can hardly be accepted. The comparison of execution time of SCARNA with other methods (Fig. 9) shows its applicability to real sequences. SCARNA requires $O(m^2)$ in time and $O(m^2)$ in memory for the alignment of SCSs of length m. The length of the SCS for an RNA sequence depends on the length of the RNA sequence, the threshold τ for base-pair probabilities, and the fixed length k of stem fragments. The lengths of the SCSs are linear to the lengths of the RNA sequences for the reasonable k-s (See supplement material for detail, Fig.5). Therefore, the computational complexities of SCS alignment are evaluated as $O(n^2)$ in time and $O(n^2)$ in memory. The computation of base-pair probabilities for SCS building requires $O(n^3)$ in time and $O(n^2)$ in memory (See supplement material, Fig.6). For very long nucleotide sequences, however, it can be reduced to $O(n^2)$ and $O(n^2)$ by restricting the distance of base pairs to a fixed length. Therefore, SCARNA can be used for long sequences in large scale analyses enjoying $O(n^2)$ of computational time.

4.2 Secondary Structure Prediction

The accuracy of the nucleotide alignments by SCARNA depends on the predicted common secondary structures. The performance of the alignment by SCARNA suggests high accuracy of the predicted secondary structures. In order to evaluate the accuracy of the secondary structure predictions for the individual sequences directly, a post-processing of recovering the high-scoring base pairs that are consistent with the predicted common secondary structures has been tested.

We made datasets from RNA sequences in Rfam database by the fraction of base pairs. The datasets are filled with following conditions. (I)The percentage of base pairs are 25% to 50%. (Ⅱ)The sequences are between 80 to 150 nucleotides in length. (Ⅲ)Their reference alignments have more than 5% and less than 50% of gaps. Each dataset is classified into several subgroups of RNA sequences by their sequence identities.

After the alignment of SCSs and removing the inconsistent stem components, the base pairs with base-pair probabilities more than 0.95 have been recovered for the prediction of individual RNA sequences. Sn (Sensitivity), Sp (Precision) and MCC (Matthews Correlation Coefficient) (Matthews, 1975) have been used for the performance measures of secondary structure prediction. Sn is sometimes better recognized as *recall*. Sp is also known as *positive predictive value* (PPV). They are defined as

$$\begin{aligned} \text{Sn} &= \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad \text{Sp} = \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{MCC} &= \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}, \end{aligned}$$

where TP, TN, FN and FP are respectively the numbers of bases (NOT base pairs) that are correctly included, correctly excluded, incorrectly excluded and incorrectly included in the predicted base pairs. The bases that are predicted to form base pairs with wrong partners are coounted both in FP and in FN. MCC ranges from -1 for extremely inaccurate (TP=TN=0) to 1 for very accurate predictions(FP=FN=0). The results have been compared with the predictions of CARNAC (Perriquet *et al.*, 2003). CARNAC is one of the most accurate software for common secondary structures (Gardner and Giegerich, 2004). The result for the dataset of 50% to 75% of base pairs is shown in Fig. 10. It can be observed that the performance of SCARNA is stable with the change of sequence identities. The MCC of SCARNA outperforms CARNAC in all range of sequence identities. Resorting to anchoring approach based on sequence similarity, CARNAC had problems in both low and very high sequence identities. SCARNA has a quality of secondary structure prediction, which results in accurate nucleotide alignments. Another result for dataset of 25% to 50% of base pairs is showed in supplement material (Fig.7).

4.3 Ability to Capture Pseudoknots

The major drawback of the DP algorithm for SCS alignment in SCARNA is that the left-right consistency is not guaranteed. The lack of the consistency, however, becomes a merit for capturing pseudoknotted structures. SCARNA often finds pseudoknotted structures without paying any additional computational costs because the algorithm does not forbid two stem fragments having pseudoknotted positions (See supplement material, Fig.8 and Fig.9 for example of pseudoknot prediction), althogh McCaskill's algorithm does not consider the pseudoknots in the calculation of the base-pair probabilities and the probabilities for pseudoknotted base pairs may be underestimated. The DP algorithm for SCS alignment can be *improved* to be left-right consistent by using pair stochastic context free grammars (PSCFGs) and by paying expensive computational costs, but only for pseudoknot-free structures.

4.4 Local Alignment

Sicne our global alignment algorithm (see section 2.5) is extendable to local alienments, we are working on adding local alignment capability to SCARNA, which allows to search non-coding RNAs from genomic sequences based on the secondary structures as well as the sequence similarities.

5 CONCLUSION

We have proposed a new method for fast and accurate alignments of RNA sequences based on the potential common secondary structures. The method uses the fixed-length stem fragments as the representation of the secondary structures. The $3'$ components and the $5'$ components of the stem fragments are separately aligned by an engineered dynamic programming and the inconsistent matches are removed as the post-process. The base-pair probabilities, substitution probabilities as the base pairs, stacking energy are considered in the alignments. The method has been implemented as SCARNA, whose accuracies of the alignments have been shown to be much better than sequence-based methods and compatible to the computationally expensive structure-based methods. The high accuracies of SCARNA in the detections of common secondary structures also supports the performance. SCARNA is fast enough to align the sequences with more than 1000 nucleotides in length, which most of the structure-based methods are unable to handle. The computational complexity of the algorithm is $O(n^3)$ in time and $O(n^2)$ in memory for the length of sequence n. The time complexity can be reduced to $O(n^2)$ for long sequences by restricting the distance of the bases in the base pairs. Pseudoknotted structures are also found without paying extra computational costs.

ACKNOWLEDGMENT

This work is partially supported by Grant-in-Aid for Scientific Research on Priority Area "Comparative Genomics" of Ministry of Education, Culture, Sports, Science and Technology, by "Functional RNA Project" of Ministry of Economy, Trade and Industry, and by the grant from Institute for Bioinformatics Research Development. The authors thank Hisanori Kiryu, Michiaki Hamada, Tsuyoshi Kato and the other members of bioinformatics group for RNAs at National Institute of Advanced Industrial Science and Technology (AIST) and at Japan Biological Information Consortium (JBIC) for useful discussions.

REFERENCES

Bafna, V., Tang, H. and Zhang, S. (2005) Consensus folding of unaligned RNA sequences revisited. In *RECOMB*, pp. 172–187.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. and Thompson, J. D. (2003) Multiple sequence alignment with the clustal series of programs. *Nucl. Acids Res.*, **31**, 3497–3500.

Eddy, S. R. (2001) Non-coding RNA genes and the modern RNA world. *Nature Genetics*, **2**, 919–929.

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, **32**, 1792–1797.

Gardner, P., Wilm, A. and Washietl, S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucl. Acids Res.*, **33**, 2433–2439.

Gardner, P. P. and Giegerich, R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**.

Gotoh, O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.*, **264**, 823–838.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S. (2003) Rfam: an RNA family database. *Nucl. Acids Res.*, **31**, 439–441.

Havgaard, J. H., Lyngsø, R. B., Stormo, G. D. and Gorodkin, J. (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.

Hofacker, I., Bernhart, S. and Stadler, P. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.

Hofacker, I., Fekete, M. and Stadler, P. (2002) Secondary structure prediction for aligned RNA sequences. *J.Mol.Biol.*, **319**, 1059–1066.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, S., Tacker, M. and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, **125**, 167–188.

Holmes, I. (2004) A probabilistic model for the evolution of RNA structure. *BMC Bioinformatics*, **5**.

Holmes, I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, **6**.

Holmes, I. and Rubin, G. M. (2002) Pairwise RNA structure comparison with stochastic context-free grammars. *PSB*, pp. 163–174.

Ji, Y., Xu, X. and Stormo, G. D. (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, **20**, 1591–1602.

Karklin, Y., Meraz, R. F. and Holbrook, S. R. (2005) Classification of non-coding RNA using graph representations of secondary structure. In *Pac Symp Biocomput.*, pp. 4–15.

Kin, T., Tsuda, K. and Asai, K. (2002) Marginalized kernels for rna sequence data analysis. *Genome Informatics.*, **13**, 112–122.

Klein, R. and Eddy, S. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**.

Lee, C., Grasso, C. and Sharlow, M. F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.

Loytynoja, A. and Sharlow, M. C. (2003) A hidden markov model for progressive multiple alignment. *Bioinformatics*, **19**, 1505–1513.

Mathews, D. (2005) Predicting a set of Minimal free energy RNA secondary structures common to two sequences. *Bioinformatics*, **21**, 2246–2253.

Mathews, D. H. and Turner, D. H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, **317**, 191–203.

Matthews, B. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem.Biophys. Acta*, **405**, 442–451.

McCaskill, J. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Nussinov, R., Pieczenik, G., Griggs, J. R. and Kleitman, D. J. (1978) Algorithms for loop matchings. *SIAM J. App. Math.*, **35**, 68–82.

Pei, J., Sadreyev, R. and V.Grishin, N. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.

Perriquet, O., Touzet, H. and Dauchet, M. (2003) Finding the common structures shared by two homologous RNAs. *Bioinformatics*, **19**, 108–116.

Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. App. Math.*, **45**, 810–825.

Washietl, S. and Hofacker, I. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J.Mol.Biol.*, **342**, 19–30.

Washietl, S., Hofacker, I. and Stadler, P. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, **102**, 2454–2459.

Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Research*, **9**, 133–148.