PicXAA-R: Efficient Structural Alignment of Multiple RNA Sequences Using a Greedy Approach

Sayed Mohammad Ebrahim Sahraeian[1], Byung-Jun Yoon*[1]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA

Email: Sayed Mohammad Ebrahim Sahraeian - msahraeian@tamu.edu; Byung-Jun Yoon*- bjyoon@ece.tamu.edu;

*Corresponding author

Abstract

Background: Accurate and efficient structural alignment of non-coding RNAs (ncRNAs) has grasped more and more attentions as recent studies unveiled the significance of ncRNAs in living organisms. While the Sankoff style structural alignment algorithms cannot efficiently serve for multiple sequences, mostly progressive schemes are used to reduce the complexity. However, this idea tends to propagate the early stage errors throughout the entire process, thereby degrading the quality of the final alignment. For multiple protein sequence alignment, we have recently proposed PicXAA which constructs an accurate alignment in a non-progressive fashion.

Results: Here, we propose PicXAA-R as an extension to PicXAA for greedy structural alignment of ncRNAs. PicXAA-R efficiently grasps both folding information within each sequence and local similarities between sequences. It uses a set of probabilistic consistency transformations to improve the posterior base-pairing and base alignment probabilities using the information of all sequences in the alignment. Using a graph-based scheme, we greedily build up the structural alignment from sequence regions with high base-pairing and base alignment probabilities.

Conclusions: Several experiments on datasets with different characteristics confirm that PicXAA-R is one of the fastest algorithms for structural alignment of multiple RNAs and it consistently yields accurate alignment results, especially for datasets with locally similar sequences. PicXAA-R source code is freely available at: http://www.ece.tamu.edu/∼bjyoon/picxaa/.

Background

Increasing number of newly discovered non-coding RNAs (ncRNAs) with huge functional variety has revealed the substantial role that RNAs play in living organisms [1–3]. The function of ncRNAs is largely ascribed to their folding structure, which is often better conserved than their primary sequence. Therefore, it is important to consider this structural aspect in the comparative analysis of RNAs, and an accurate structural alignment algorithm can be helpful in decoding the function of ncRNAs and discovering novel ncRNA candidates.

To accurately align RNA sequences, one should take their secondary structure similarities into account, in addition to their sequence homologies. Simultaneous inference of both the consensus secondary structure and the alignment of RNA sequences is a computationally demanding task. Sankoff [4] proposed an algorithm for structural alignment of a set of unaligned RNA sequences. However, the high complexity of $O(L^{3N})$ in time and $O(L^{2N})$ in memory for N sequences of length L makes this algorithm impractical even for a small number of sequences. Hence, several studies have proposed various approximations to the Sankoff algorithm [5–19]. Algorithms such as Foldalign [5–7], Dynalign [8, 9], and Stemloc [10] employ several heuristics to impose constraints on the size or shape of substructures, thereby, reducing the search space. Murlet [12], RAF [13], PARTS [14], STRAL [15], LocARNA [16], CentroidAlign [17], and PMcomp [18] exploit probabilistic approaches by implementing base-pairing probabilities in a restricted Sankoff-style framework or employing the Needleman-Wunsch algorithm with structural scores. Although these variants of Sankoff's algorithm significantly reduce the time and memory complexities, they still cannot directly find the structural alignment of multiple sequences. Instead, these algorithms build up the multiple sequence alignment (MSA) by progressively combining pairwise structural alignments along a guide tree.

In addition to these Sankoff-style algorithms, several studies have recently investigated fast techniques to find the common structure of long RNA sequences. For example, MXSCARNA [20] progressively computes the pairwise structural alignment of a pair of stem candidates obtained from the base-pairing probability matrices. R-Coffee [21, 22] uses a library of input alignments to progressively compute the alignment by incorporating secondary structure information. LaRA [23] and MARNA [24] employ two different heuristic

approaches to compute all pairwise structure alignments and pass this information, as a primary library, to T-COFFEE [25], a progressive alignment technique. MAFFT-xinsi [26] uses a four-way consistency objective function to progressively build a structural alignment by combining pairwise alignments predicted by an external program.

Despite its computational efficiency, the progressive structural alignment approach tends to propagate the errors made in the early stages throughout the entire process, which may significantly degrade the quality of the final alignment. Even with the incorporation of additional heuristics, such as iterative refinement and consistency transformation, the fundamental shortcoming of progressive technique remains. A number of non-progressive structural alignment schemes have been proposed to address this problem [27–29]. RNASampler [27] predicts the common structure of multiple RNA sequences by probabilistically sampling aligned stems based on the stem conservation score. MASTER [28], another sampling approach, iteratively improves both sequence alignment and structure prediction by making small local changes using simulated annealing. Stemloc-AMA [29] employs sequence annealing to construct the multiple RNA alignment using the base alignment probabilities estimated by the Sankoff algorithm with structural considerations.

Recently, several studies have highlighted the effectiveness of the Maximum Expected Accuracy (MEA) approach for aligning biological sequences [30–36] and for predicting the consensus secondary structure of RNAs [12, 17, 20, 29, 37–39]. MEA tries to maximize the expected number of correctly aligned bases. This is especially useful for handling sequence analysis problems when the probability of the optimal alignment is low.

In this paper, we introduce PicXAA-R (**p**robabilist**ic** ma**x**imum **a**ccuracy **a**lignment of **R**NA sequences), a novel non-progressive algorithm that efficiently finds the maximum expected accuracy structural alignment of multiple RNA sequences. PicXAA-R greedily builds up the structural alignment from sequence regions with high local similarities and high base-pairing probabilities. To simultaneously consider both the local similarities among sequences and their conserved secondary structural information, we incorporate three types of probabilistic consistency transformations. These transformations modify both the inter-sequence pairwise base alignment probabilities and the intra-sequence base-pairing probabilities using the information from other sequences in the alignment. For a fast and accurate construction of the alignment, we propose an efficient two-step graph-based alignment scheme. In the first step, we greedily insert the most probable alignments of base-pairs with high base-pairing probability. In this way, we build up the skeleton of the alignment using the structure information of the RNA sequences. Next, we successively insert the most probable pairwise base alignments into the multiple structural alignment, as in

PicXAA [34], a multiple protein sequence alignment algorithm that we have recently proposed. This step can effectively grasp the local sequences similarities among the RNAs. Finally, we use a discriminative refinement step to improve the overall alignment quality in sequence regions with low alignment probability. Extensive experiments on several local alignment benchmarks clearly show that PicXAA-R is one of the fastest algorithms for structural alignment of multiple RNAs and it consistently yields accurate results in comparison with several well-known structural RNA alignment algorithms.

Methods

PicXAA-R extends the idea of PicXAA, the multiple sequence alignment that maximizes the expected number of correctly aligned bases, to the structural alignment of RNA sequences. PicXAA-R uses a greedy approach that builds up the alignment from sequence regions with high local similarities and high base-pairing probabilities. Thus, it avoids the propagation of early stage alignment errors, usually observed in progressive techniques. The algorithm employs a probabilistic framework by realizing both the inter-sequence base alignment probabilities and the intra-sequence base-pairing probabilities. The following subsections provide an overview of the proposed algorithm.

Preliminary

To align m RNA sequences in a set $\mathbf{S} = \{s_1, \cdots, s_m\}$, we need to compute the following probabilities.

- $P_a(x_i \sim y_j|\mathbf{x}, \mathbf{y})$: For each pair sequence $\mathbf{x}, \mathbf{y} \in \mathbf{S}$, $P_a(x_i \sim y_j|\mathbf{x}, \mathbf{y})$ is the probability that bases $x_i \in \mathbf{x}$ and $y_j \in \mathbf{y}$ are matched in the true (unknown) alignment. We can compute the posterior pairwise alignment probabilities using the pair hidden Markov model (PHMM) [40].
- $P_b(x_i \sim x_j|\mathbf{x})$: For each sequence $\mathbf{x} \in \mathbf{S}$, $P_b(x_i \sim x_j|\mathbf{x})$ is the probability that two bases $x_i, x_j \in \mathbf{x}$ form a base-pair. We can exploit different approaches, such as the McCaskill algorithm [41] or the CONTRAfold model [39], to compute the base-pairing probabilities.

We use these probabilities in the following probabilistic structural alignment scheme.

Consistency transformation

Here, we use three types of probabilistic consistency transformations to modify the pairwise base alignment probabilities and base-pairing probabilities using the information from other sequences in the alignment.

This modification makes these posterior probabilities suitable for constructing a consistent and accurate structural alignment.

Inter-sequence probabilistic consistency transformation for base alignment probabilities.

In the first consistency transformation, we incorporate the information from other sequences in the alignment to improve the estimation of pairwise base alignment probabilities. The motivation of this transformation is that all the pairwise alignments induced from a given MSA should be consistent with each other. This means that if position x_i ($\in \mathbf{x}$) aligns with position z_k ($\in \mathbf{z}$) in the $\mathbf{x}-\mathbf{z}$ alignment, and if z_k aligns with position y_j ($\in \mathbf{y}$) in the $\mathbf{z}-\mathbf{y}$ alignment, then x_i must align with y_j in the $\mathbf{x}-\mathbf{y}$ alignment. We can thus utilize the "intermediate" sequence $\mathbf{z}$ to improve the $\mathbf{x}-\mathbf{y}$ alignment by making it consistent with the alignments $\mathbf{x}-\mathbf{z}$ and $\mathbf{z}-\mathbf{y}$.

Based on this motivation, we introduced an enhanced probabilistic consistency transformation in PicXAA [34], which improves the original transformation proposed by Do *et al.* [30]. The enhanced transformation modifies the alignment probability for a base-pair $x_i \sim y_j$, by incorporating the alignment probability between x_i and z_k and that between z_k and y_j. This transformation can be written as:

$$P'_a(x_i \sim y_j|\mathbf{S}) \simeq \frac{\sum_{\mathbf{z}\in\mathbf{S}} \sum_{z_k} P_a(x_i \sim z_k|\mathbf{x},\mathbf{z}) P_a(z_k \sim y_j|\mathbf{z},\mathbf{y}) P(\mathbf{x}\diamond\mathbf{z}) P(\mathbf{y}\diamond\mathbf{z})}{\sum_{\mathbf{z}\in\mathbf{S}} P(\mathbf{x}\diamond\mathbf{z}) P(\mathbf{y}\diamond\mathbf{z})}.$$

where $P(\mathbf{x}\diamond\mathbf{z})$ represents the probability that $\mathbf{x}$ and $\mathbf{z}$ are homologous, defined as:

$$P(\mathbf{x}\diamond\mathbf{z}) \triangleq \frac{1}{|\bar{a}|} \sum_{x_i\sim z_k\in\bar{a}} P_a(x_i \sim z_k|\mathbf{x},\mathbf{z}),$$

where $\bar{a}$ is the optimal pairwise alignment of $\mathbf{x}$ and $\mathbf{z}$.

This transformation improves the consistency of the $\mathbf{x}-\mathbf{y}$ alignment with other pairwise alignments in the MSA, by incorporating information only from homologous sequences. In this way, we can obtain more probabilistically consistent estimate of the posterior alignment probabilities, which helps enhance the quality of the final MSA.

Intra-sequence probabilistic consistency transformation for base-pairing probabilities.

In the second transformation, we incorporate the pairwise alignment information to the structural formation of the sequences. This transformation exploits this observation that the base-pairings in each sequence should be consistent with the pairwise base alignments induced from a given structural alignment.

This means that if positions $y_j \sim y_{j'}$ form a base-pair in $\mathbf{y}$, where x_i ($\in \mathbf{x}$) aligns with y_j ($\in \mathbf{y}$) and $x_{i'}$ ($\in \mathbf{x}$) aligns with $y_{j'}$ ($\in \mathbf{y}$), then $x_i \sim x_{i'}$ must form a base-pair in $\mathbf{x}$. Thus, we can utilize the base alignment information to improve the estimation of the $x_i \sim x_{i'}$ base-pairing probability.

Based on this observation, Kiryu *et al.* [12] introduced a transformation for base-pairing probabilities, which was modified later in [42] as:

$$P'_b(x_i \sim x_{i'}|\mathbf{x}) = \alpha P_b(x_i \sim x_{i'}|\mathbf{x}) + \frac{1-\alpha}{m-1} \sum_{\mathbf{y} \in \mathbf{S}-\{\mathbf{x}\}} \sum_{j<j'} P_a(x_i \sim y_j|\mathbf{x},\mathbf{y}) P_b(y_j \sim y_{j'}|\mathbf{y}) P_a(x_{i'} \sim y_{j'}|\mathbf{x},\mathbf{y})$$

where $\alpha \in [0,1]$ is a weight parameter between the target sequence $\mathbf{x}$ and rest of sequences. This transformation assumes that all sequences $\mathbf{y} \in \mathbf{S} - \{\mathbf{x}\}$ are homologous to the given sequence $\mathbf{x}$. However, when we have a set of distantly related sequences in $\mathbf{S}$, this assumption does not necessarily hold.

To address this problem, here, we modify this transformation by improving the base-pairing probability using the information just from the closely related sequences to the given sequence $\mathbf{x}$. Therefore, like the *inter-sequence consistency transformation*, we explicitly consider the relative significance of each sequence $\mathbf{y} \in \mathbf{S} - \{\mathbf{x}\}$ in improving the base-pairing probabilities in $\mathbf{x}$.

Let $\mathbf{Z} = \{\mathbf{y} \in \mathbf{S} - \{\mathbf{x}\} | \mathbf{x} \diamond \mathbf{y}\}$ be the set of sequences in $\mathbf{S} - \{\mathbf{x}\}$ that are homologous to $\mathbf{x}$. The notation $\mathbf{x} \diamond \mathbf{y}$ means $\mathbf{x}$ and $\mathbf{y}$ are homologous and functionally related to each other. Using only the relevant sequences, which are included in the set $\mathbf{Z}$, we define this transformation as:

$$P'_b(x_i \sim x_{i'}|\mathbf{x}) = \alpha P_b(x_i \sim x_{i'}|\mathbf{x}) + \frac{1-\alpha}{|\mathbf{Z}|} \sum_{\mathbf{y} \in \mathbf{Z}} \sum_{j<j'} P_a(x_i \sim y_j|\mathbf{x},\mathbf{y}) P_b(y_j \sim y_{j'}|\mathbf{y}) P_a(x_{i'} \sim y_{j'}|\mathbf{x},\mathbf{y})$$

The second term in the right hand side of the above equation can be also written as:

$$\frac{1-\alpha}{\sum_{\mathbf{y} \in \mathbf{S}-\{\mathbf{x}\}} \mathbf{I}\{\mathbf{x} \diamond \mathbf{y}\}} \sum_{\mathbf{y} \in \mathbf{S}-\{\mathbf{x}\}} \sum_{j<j'} P_a(x_i \sim y_j|\mathbf{x},\mathbf{y}) P_b(y_j \sim y_{j'}|\mathbf{y}) P_a(x_{i'} \sim y_{j'}|\mathbf{x},\mathbf{y}) \mathbf{I}\{\mathbf{x} \diamond \mathbf{y}\}$$

using the identity function $\mathbf{I}\{\cdot\}$, where $\mathbf{I}\{\mathbf{x} \diamond \mathbf{y}\} = 1$ if $\mathbf{y}$ is homologous to $\mathbf{x}$, and $\mathbf{I}\{\mathbf{x} \diamond \mathbf{y}\} = 0$ otherwise. In practice, we cannot judge with certainty whether two sequences are homologous or not. Thus, we describe this relationship probabilistically, using the expectation as: $\mathbf{E}\big[\mathbf{I}\{\mathbf{x} \diamond \mathbf{y}\}\big] = P(\mathbf{x} \diamond \mathbf{y})$, where $P(\mathbf{x} \diamond \mathbf{y})$ is the homology probability and can be estimated as described in the previous subsection. By replacing the identity functions with their expected values in the previous equation, we propose the following enhanced *intra-sequence probabilistic consistency transformation* as:

$$\begin{aligned} P'_b(x_i \sim x_{i'}|\mathbf{x}) &= \alpha P_b(x_i \sim x_{i'}|\mathbf{x}) \\ &+ (1-\alpha) \frac{\sum_{\mathbf{y} \in \mathbf{S}-\{\mathbf{x}\}} \sum_{j<j'} P_a(x_i \sim y_j|\mathbf{x},\mathbf{y}) P_b(y_j \sim y_{j'}|\mathbf{y}) P_a(x_{i'} \sim y_{j'}|\mathbf{x},\mathbf{y}) P(\mathbf{x} \diamond \mathbf{y})}{\sum_{\mathbf{y} \in \mathbf{S}-\{\mathbf{x}\}} P(\mathbf{x} \diamond \mathbf{y})} \end{aligned}$$

Probabilistic four-way consistency transformation for base alignment probabilities.

In the third consistency transformation, we incorporate the structural information to the pairwise alignments. This transformation is based on the same observation that motivated the *intra-sequence consistency transformation*; that is, the pairwise base alignments induced from a given structural alignment should be consistent with the base-pairings in the corresponding pair sequence. However, this time, we utilize the base-pairing information to improve the $\mathbf{x} - \mathbf{y}$ alignment.

Based on this motivation, Katoh and Toh introduced the four-way consistency transformation in [26] which was also latter implemented in [17]. We use this idea in a probabilistic fashion by incorporating the base alignment and the base-pairing probabilities as in [17]. This transformation is defined as:

$$\begin{aligned} P'_a(x_i \sim y_j|\mathbf{x},\mathbf{y}) &= \beta P_a(x_i \sim y_j|\mathbf{x},\mathbf{y}) + (1-\beta)\Bigg(\sum_{\substack{0<i'<i \\ 0<j'<j}} P_b(x_{i'} \sim x_i|\mathbf{x}) P_a(x_{i'} \sim y_{j'}|\mathbf{x},\mathbf{y}) P_b(y_{j'} \sim y_j|\mathbf{y}) \\ &+ \sum_{\substack{i<i'<|x| \\ j<j'<|y|}} P_b(x_i \sim x_{i'}|\mathbf{x}) P_a(x_{i'} \sim y_{j'}|\mathbf{x},\mathbf{y}) P_b(y_j \sim y_{j'}|\mathbf{y}) \Bigg) \end{aligned}$$

where $\beta \in [0,1]$ is a weight parameter.

Using the sparsity of alignment and pairing probability matrices, we can efficiently implement these three transformations successively. The *inter-sequence consistency transformation* has a complexity of $O(\mu^2 Lm^3)$, the *intra-sequence transformation* has a complexity of $O(\mu^3 Lm^2)$, and *the four-way consistency transformation* has a computational complexity of $O(\mu^4 Lm^2)$, where μ is the average number of non-zero elements per row (typically $1 \leq \mu \leq 5$ in real examples), m is the number of sequences, and L is the length of each sequence.

Constructing the structural alignment

To find a valid structural alignment of a set of RNA sequences, we propose a two-step greedy approach that builds up the alignment starting from those regions with higher base-pairing and base alignment probabilities. The proposed greedy scheme extends the idea of PicXAA [34] to multiple RNA alignments. In PicXAA, we construct the multiple protein sequence alignment by successively inserting the most probable pairwise residue alignment into the final alignment. In the proposed algorithm, we add another step before the greedy graph construction step of PicXAA to better incorporate the secondary structure information in RNAs. This two-step alignment construction approach, along with *intra-sequence*

consistency transformation and *four-way consistency transformation*, described in the previous subsection, helps PicXAA-R to effectively integrate both sequence and structural similarities to construct the final alignment. The proposed structural alignment approach is described in the following.

The greedy alignment approach we proposed in PicXAA [34] is conceptually similar to the one used in sequence annealing algorithms [29, 35, 36]. However, it should be noted that unlike sequence annealing, which greedily merges pairs of columns, we always add a single pairwise base alignment at a time, based on the consistency-transformed posterior alignment probabilities.

We represent the structural alignment as a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where, $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of directed edges. Each vertex $c^{(i)} \in \mathcal{V}$ corresponds to a column in the final alignment, and each directed edge $e = (c^{(i)}, c^{(j)}) \in \mathcal{E}$ implies that column $c^{(i)}$ precedes column $c^{(j)}$ in the given alignment. Each column $c^{(i)} \in \mathcal{V}$ consists of positions from different sequences that will appear in the same column in the final alignment.

When inserting a new pairwise base alignment, we should consider the following requirements to obtain a legitimate multiple RNA alignment:

- (Avoid Cycles) The alignment graph $\mathcal{G}$ should remain acyclic.
- (Left-Right Compatibility) In the first greedy step where we use structural information, we should consider left-right compatibility. That is, for any paired columns (c, c'), if column c appears in the left part of the stem in the final structure, then for each base $x_i \in c$ that pairs with some $x_{i'} \in c'$ of the same sequence $\mathbf{x}$, we should have $i < i'$.

Thus, while we build up the alignment graph, we satisfy the structural constraints and alignment constraints by verifying whether the new inserted pairwise base alignment keeps the graph acyclic and left-right compatible.

The two-step alignment construction approach is as follows:

Step 1-Structural skeleton construction

In the first alignment construction step, we greedily insert the most probable alignments of base-pairs with high base-pairing probability. To this aim, we define the ordered set $\mathbf{B}$ as

$$\mathbf{B} = \{(x_i, x_{i'}) | x_i, x_{i'} \in \mathbf{x},\, \mathbf{x} \in \mathbf{S},\, P_b'(x_i \sim x_{i'} | \mathbf{x}) > T_b\}.$$

Here, $\mathbf{B}$ is the ordered set of base-pairs whose transformed base-pairing probability is larger than a threshold T_b. The base-pairs in $\mathbf{B}$ are sorted in descending order according to their transformed base-pairing probability, $P'_b(x_i \sim x_{i'}|\mathbf{x})$. We successively pick the most confident base-pair $(x_i, x_{i'})$ from $\mathbf{B}$. For a selected base-pair, we look for the best match among the members of $\mathbf{B}$. That is, we seek for a pair $(y_j, y_{j'}) \in \mathbf{B}$ which belongs to another sequence $\mathbf{y}$ and satisfies the two compatibility conditions above in $\mathcal{G}$ while maximizing the following probability:

$$(y_j, y_{j'}) = \underset{(y_j, y_{j'}) \in \mathbf{B}}{\arg\max} \big(P_a(x_i \sim y_j|\mathbf{x}, \mathbf{y}) P_b(y_j \sim y_{j'}|\mathbf{y}) P_a(x_{i'} \sim y_{j'}|\mathbf{x}, \mathbf{y})\big).$$

For this pair $(y_j, y_{j'})$, we insert two pairwise alignments $(x_i \sim y_j)$ and $(x_{i'} \sim y_{j'})$ into the alignment graph $\mathcal{G}$. Figure 1A illustrates this process.

Upon inserting a new pair $p^* = (x_i, y_j)$ to $\mathcal{G}$, three scenarios may occur: (1) New column addition; (2) Extension of an existing column; or (3) Merging of two columns. The detailed description of the procedures needed for each case can be found in [34]. Later in this section, we provide a summary of those procedures. By successively inserting the most probable alignment for confident base-pairs, we construct the skeleton of the alignment enriched by structural information. Next, we complete this skeleton by greedily inserting highly probable base alignments.

Step 2-Inserting highly probable local alignments

In this step, we update the skeleton alignment obtained in the previous step by successively inserting the most probable pairwise base alignments into the multiple structural alignment, as in PicXAA [34].

Thus, we sort all remaining pairwise alignments (x_i, y_j) according to their transformed alignment probability $P'_a(x_i \sim y_j|\mathbf{x}, \mathbf{y})$ in an ordered set $\mathbf{A}$. We greedily build up $\mathcal{G}$ by repeatedly picking the most probable pair in $\mathbf{A}$, which is not processed yet, provided that it is compatible with the current alignment. Again, insertion of any pair $p^* = (x_i, y_j)$ to $\mathcal{G}$ will result in one of the scenarios of new column addition, extension of an existing column, or merging of two columns.

Here, we briefly discuss these three cases (For detailed description see [34]):

1. **New column addition**: We insert a new compatible vertex $c^* = \{x_i, y_j\}$ in $\mathcal{G}$ if neither x_i nor y_j belongs to some existing column in $\mathcal{G}$. Figure 1B illustrates this process.

2. **Extending an existing column**: If only one of the bases in p^*, let say x_i, belongs to some vertex $c \in \mathcal{V}$, we should add the other base y_j to the same vertex c. Figure 1C illustrates this process.

3. **Merging two vertices**: When $x_i \in c_1$ and $y_j \in c_2$ belong to two different vertices $c_1, c_2 \in \mathcal{V}$, we merge the vertices c_1 and c_2. Figure 1D illustrates this process.

After updating the graph as described above, we prune $\mathcal{G}$ to avoid redundant edges, thereby improving the computational efficiency of the construction process.

Upon finishing the two-step graph construction, we use the obtained alignment graph $\mathcal{G}$ to find the multiple alignment. We use the depth-first search algorithm to order the vertices in $\mathcal{V}$ in an ordered set $\mathcal{A} = (v_1, v_2, \cdots, v_n)$ such that there is no path from v_i to v_j in $\mathcal{G}$ for any $i > j$. In the resulted ordered set $\mathcal{A}$, each member corresponds to a column in the alignment, and putting them together gives the alignment. Further details of the graph construction and alignment process can be found in [34]. An illustrative example for the graph construction process using PicXAA-R can be found in Figure 2.

Discriminative refinement

As the final step, we apply a refinement step to improve the alignment quality in sequence regions with low alignment probability. We employ the iterative refinement strategy based on the discriminative-split-and-realignment technique that was introduce in PicXAA [34]. We repeat the following steps successively for each sequence $\mathbf{x} \in \mathbf{S}$:

1. Find $\mathbf{S_x} \subset \mathbf{S}$, the set of similar sequences to $\mathbf{x}$ using the k-means clustering.
2. Align $\mathbf{x}$ with the profile of sequences in $\mathbf{S_x}$.
3. Perform the profile-profile alignment of $\mathbf{S'_x} = \mathbf{S_x} \cup \mathbf{x}$ and $\mathbf{S} - \mathbf{S_x}$.

This refinement strategy, takes advantage of both the intra-family similarity as well as the inter-family similarity, thereby improving the alignment quality in low similarity regions without breaking the confidently aligned bases.

Results and Discussion

We use four different benchmark datasets: BRAliBase 2.1 [43], Murlet [12], BraliSub [44], and LocalExtR [44] to assess the performance of PicXAA-R on different alignment conditions. The first two are general datasets not specially designed for local RNA alignment testing while the last two datasets are designed to verify the alignment accuracy for locally similar RNAs.

We compared PicXAA-R with several well-known RNA sequence alignment algorithms: ProbConsRNA 1.10 [30], MXSCARNA 2.1 [20], CentroidAlign [17], and MAFFT-xinsi 6.717 [26]. Among these techniques, ProbConsRNA uses only the sequence level information while the others take advantage of structural information. We picked these methods as they are among the fastest structural RNA aligners which yield high accuracy. There exists several other aligners such as RAF 1.00 [13], Murlet [12], Stemloc-AMA [29], LaRA 1.3.2 [23], M-LocARNA [16], and R-Coffee [21], which have much more complexity than MAFFT-xinsi (in some cases they are near 60 times slower) while their accuracy is usually worse or at least comparable to MAFFT-xinsi. Thus, the most complex algorithm that we compare our algorithm with will be the state-of-the-art technique, MAFFT-xinsi.

All the experiments have been performed on a 2.2GHz Intel Core2Duo system with 4GB memory. On all datasets we use two measurement to evaluate the performance of each alignment scheme: (1) *sum-of-pairs score* (SPS), which represents the percentage of correctly aligned bases; (2) *structure conservation index* (SCI) [45] that measures the degree of conservation of the consensus secondary structure for a multiple alignment. The SCI score is defined as $\text{SCI} = \frac{E_A}{\bar{E}}$ where E_A is the minimum free energy of the consensus MSA as computed by RNAalifod [46] and $\bar{E}$ is the average minimum free energy of all single sequences in the alignment as computed by RNAfold [47].

On Murlet dataset, in addition to the SPS and SCI scores, we measure sensitivity $\text{SEN} = \text{TP}/(\text{TP} + \text{FN})$, Positive Predictive Value $\text{PPV} = \text{TP}/(\text{TP} + \text{FP})$, and Matthews correlation coefficient (MCC):

$$\text{MCC} = \frac{\text{TP} \times \text{TN} + \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}.$$

where true positive (TP) indicates the number of correctly predicted base-pairs, true negative (TN) is the number of base-pairs correctly predicted as unpaired, false negative (FN) is the number of not predicted true base-pairs, and false positive (FP) is the number of incorrectly predicted base-pairs.

In each table the total computational time for each algorithm is also reported in seconds.

Throughout the experiment we use the parameter setting of $\alpha = 0.4$, $\beta = 0.1$, and $T_b = 0.5$. These parameters are optimized manually using small datasets. Besides, we use McCaskill algorithm [41] to compute the base-pairing probabilities and RNAalifold [46] to find the induced consensus structure of the computed alignment.

Results on BRAliBase 2.1

First, we evaluated the accuracy of PicXAA-R using the BRAliBase 2.1 alignment benchmark. Wilm *et al.* [43] has developed BRAliBase 2.1 based on hand-curated seed alignments of 36 RNA families taken from Rfam 7.0 database [48]. BRAliBase 2.1 contains in total 18,990 aligned sets of sequences each consists of 2, 3, 5, 7, 10, or 15 sequences (categorized into k2, k3, k5, k7, k10, and k15 reference sets) with average pairwise sequence identities ranging from 20% to 95%.

Table 1 summarizes the SPS and SCI scores along with the running time of each algorithm. As we see, MAFFT-xinsi has the highest average scores while it is two times slower than PicXAA-R. In comparison with other techniques PicXAA-R has similar scores which usually gets better as the number of sequences increases (k10 and k15).

To more clearly compare these techniques, we provided the average SPS and SCI scores as a function of the average percent identity on k5, k7, k10, and k15 reference sets in Figure 3. As shown in this figure, for sequence identities less than 60% PicXAA-R outperform all the other schemes in terms of both scores except for MAFFT-xinsi which is two times slower than PicXAA-R. This observation shows that the proposed greedy approach can efficiently and effectively construct the alignment for low identity sequence sets. This was expected as in lower sequence identities the proposed greedy alignment construction approach can effectively detect local structural similarities.

Results on BraliSub and LocExtR

The BraliBase 2.1 benchmark is not designed for local alignment testing and has reference alignments with just up to 15 sequences. Thus, Wang *et al.* [44] designed two types of datasets to verify the potential of RNA sequence aligners in dealing with local similarities in the alignment set: (1) BraliSub, the subsets of BraliBase 2.1 with high variability (containing 232 reference alignments); (2) LocalExtR, an extension of BraliBase 2.1 consisting total of 90 large-scale reference alignments categorized into k20, k40, k60, and k80 reference sets receptively with 20, 40, 60, and 80 sequences in each alignment.

Tables 2 and 3 summarize the performance measures on these datasets. As we can see, MAFFT-xinsi has the best accuracy but it is 2.5 times slower than PicXAA-R in BraliSub dataset and four times slower than PicXAA-R in LocExtR dataset. Besides, PicXAA-R outperforms MXSCARNA with average 6-7% in terms of SPS and SCI scores. It also outperforms CentroidAlign by average 1-2% in both scores.

These results confirm that PicXAA-R can efficiently yield an accurate structural alignment for a set of large number of locally similar RNAs.

Results on Murlet dataset

Murlet dataset [12] consists of 85 alignments of 10 sequences obtained from the Rfam 7.0 database [48]. This dataset includes 17 families and there are five alignments for each family. The mean pairwise sequence identities varies from 40% to 94%. Table 4 shows the results on this dataset. We observe that PicXAA-R yields comparable accuracy with MAFFT-xinsi while PicXAA-R has much less complexity. In comparison with CentroidAlign, we have similar SPS and better SCI scores, while we are 3% better in terms of SEN score and 2% worse in terms of PPV score. However, for MCC score which compromises between sensitivity and specificity PicXAA-R outperforms CentroidAlign by 0.8%.

Computational complexity analysis

Figure 4 shows the average CPU time for different algorithms as a function of the number of sequences in the alignments in BraliSub and LocExtR datasets. As we see, the complexity of MAFFT-xinsi grows much faster than other algorithms as the number of sequences increase, while the complexity of PicXAA-R smoothly grows with number of sequences. We also see that PicXAA-R stands between MXSCARNA and CentroidAlign in terms of CPU time. However, as shown in the previous subsections, we outperform both these techniques in datasets consisting sequences with local similarity and low pairwise identity.

Conclusions

In this paper, we proposed PicXAA-R, a probabilistic structural RNA alignment technique based on a greedy algorithm. Using a set of probabilistic consistency transformations, including a novel *intra-sequence consistency transformation*, we incorporate the folding and alignment information of all sequences to enhance both the posterior base-pairing and base alignment probabilities. We utilize these enhanced probabilities as the building blocks of the two-step greedy scheme which builds up the alignment starting from sequence regions with high local similarity and high base-pairing probability. As shown in several experiments, PicXAA-R can efficiently yield highly accurate structural alignment of ncRNAs. This performance is more vivid for datasets consisting sequences with local similarities and low pairwise identities. To the best of our knowledge, PicXAA-R is the fastest structural alignment algorithm after MXSCARNA among all the current RNA aligners while it significantly outperforms MXSCARNA on local datasets like BraliSub and LocExtR. High speed implementation of PicXAA-R as well as its accuracy makes it a practical tool for structural alignment of large number of ncRNAs with low sequence identity which is very helpful for novel ncRNA prediction.

Authors contributions

Conceived the algorithm: SMES, BJY. Implemented the algorithm and performed the experiments: SMES.

Analyzed the results: SMES, BJY. Wrote the paper: SMES, BJY.

Acknowledgements

This work was supported in part by Texas A&M faculty start-up fund.

References

1. Eddy SR: **Non-coding RNA genes and the modern RNA world**. *Nat. Rev. Genet.* 2001, **2**:919–929.
2. Storz G: **An expanding universe of noncoding RNAs**. *Science* 2002, **296**:1260–1263.
3. Costa FF: **Non-coding RNAs: lost in translation?** *Gene* 2007, **386**:1–10.
4. Sankoff D: **Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems**. *SIAM Journal on Applied Mathematics* 1985, **45**(5):810–825.
5. Gorodkin J, Stricklin SL, Stormo GD: **Discovering common stem-loop motifs in unaligned RNA sequences**. *Nucleic Acids Res.* 2001, **29**:2135–2144.
6. Havgaard JH, Lyngso RB, Stormo GD, Gorodkin J: **Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%**. *Bioinformatics* 2005, **21**:1815–1824.
7. Havgaard JH, Torarinsson E, Gorodkin J: **Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix**. *PLoS Comput. Biol.* 2007, **3**:1896–1908.
8. Mathews DH, Turner DH: **Dynalign: an algorithm for finding the secondary structure common to two RNA sequences**. *J. Mol. Biol.* 2002, **317**:191–203.
9. Mathews DH: **Predicting a set of minimal free energy RNA secondary structures common to two sequences**. *Bioinformatics* 2005, **21**:2246–2253.
10. Holmes I: **Accelerated probabilistic inference of RNA structure evolution**. *BMC Bioinformatics* 2005, **6**:73.
11. Dowell RD, Eddy SR: **Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction**. *BMC Bioinformatics* 2004, **5**:71.
12. Kiryu H, Tabei Y, Kin T, Asai K: **Murlet: a practical multiple alignment tool for structural RNA sequences**. *Bioinformatics* 2007, **23**:1588–1598.
13. Do CB, Foo CS, Batzoglou S: **A max-margin model for efficient simultaneous alignment and folding of RNA sequences**. *Bioinformatics* 2008, **24**:68–76.
14. Harmanci AO, Sharma G, Mathews DH: **PARTS: probabilistic alignment for RNA joinT secondary structure prediction**. *Nucleic Acids Res.* 2008, **36**:2406–2417.
15. Dalli D, Wilm A, Mainz I, Steger G: **STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time**. *Bioinformatics* 2006, **22**:1593–1599.
16. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R: **Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering**. *PLoS Comput. Biol.* 2007, **3**:e65.
17. Hamada M, Sato K, Kiryu H, Mituyama T, Asai K: **CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score**. *Bioinformatics* 2009, **25**:3236–3243.
18. Hofacker IL, Bernhart SH, Stadler PF: **Alignment of RNA base pairing probability matrices**. *Bioinformatics* 2004, **20**:2222–2227.
19. Anwar M, Nguyen T, Turcotte M: **Identification of consensus RNA secondary structures using suffix arrays**. *BMC Bioinformatics* 2006, **7**:244.

20. Tabei Y, Kiryu H, Kin T, Asai K: **A fast structural multiple alignment method for long RNA sequences**. *BMC Bioinformatics* 2008, **9**:33.

21. Wilm A, Higgins DG, Notredame C: **R-Coffee: a method for multiple alignment of non-coding RNA**. *Nucleic Acids Res.* 2008, **36**:e52.

22. Moretti S, Wilm A, Higgins DG, Xenarios I, Notredame C: **R-Coffee: a web server for accurately aligning noncoding RNA sequences**. *Nucleic Acids Res.* 2008, **36**:W10–13.

23. Bauer M, Klau GW, Reinert K: **Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization**. *BMC Bioinformatics* 2007, **8**:271.

24. Siebert S, Backofen R: **MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons**. *Bioinformatics* 2005, **21**:3352–3359.

25. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment**. *J. Mol. Biol.* 2000, **302**:205–217.

26. Katoh K, Toh H: **Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework**. *BMC Bioinformatics* 2008, **9**:212.

27. Xu X, Ji Y, Stormo GD: **RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment**. *Bioinformatics* 2007, **23**:1883–1891.

28. Lindgreen S, Gardner PP, Krogh A: **MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing**. *Bioinformatics* 2007, **23**:3304–3311.

29. Bradley RK, Pachter L, Holmes I: **Specific alignment of structured RNA: stochastic grammars and sequence annealing**. *Bioinformatics* 2008, **24**:2677–2683.

30. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment**. *Genome Res.* 2005, **15**:330–340.

31. Roshan U, Livesay DR: **Probalign: multiple sequence alignment using partition function posterior probabilities**. *Bioinformatics* 2006, **22**:2715–2721.

32. Paten B, Herrero J, Beal K, Birney E: **Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment**. *Bioinformatics* 2009, **25**:295–301.

33. Do C, Gross S, Batzoglou S: **CONTRAlign: Discriminative Training for Protein Sequence Alignment**. In *Proceedings of the Tenth Annual International Conference on Computational Molecular Biology (RECOMB): 2-5 April 2006; Venice, Italy.* 2006:160–174.

34. Sahraeian SM, Yoon BJ: **PicXAA: greedy probabilistic construction of maximum expected accuracy alignment of multiple sequences**. *Nucleic Acids Res.* 2010, **38**:4917–4928.

35. Schwartz AS, Pachter L: **Multiple alignment by sequence annealing**. *Bioinformatics* 2007, **23**:e24–29.

36. Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L: **Fast statistical alignment**. *PLoS Comput. Biol.* 2009, **5**:e1000392.

37. Lu ZJ, Gloor JW, Mathews DH: **Improved RNA secondary structure prediction by maximizing expected pair accuracy**. *RNA* 2009, **15**:1805–1813.

38. Kiryu H, Kin T, Asai K: **Robust prediction of consensus secondary structures using averaged base pairing probability matrices**. *Bioinformatics* 2007, **23**:434–441.

39. Do CB, Woods DA, Batzoglou S: **CONTRAfold: RNA secondary structure prediction without physics-based models**. *Bioinformatics* 2006, **22**:e90–98.

40. Durbin R, Eddy SR, Krogh A, Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press 1998.

41. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure**. *Biopolymers* 1990, **29**:1105–1119.

42. Hamada M, Sato K, Kiryu H, Mituyama T, Asai K: **Predictions of RNA secondary structure by combining homologous sequence information**. *Bioinformatics* 2009, **25**:i330–338.

43. Wilm A, Mainz I, Steger G: **An enhanced RNA alignment benchmark for sequence alignment programs**. *Algorithms Mol Biol* 2006, **1**:19.

44. Wang S, Gutell RR, Miranker DP: **Biclustering as a method for RNA local multiple sequence alignment**. *Bioinformatics* 2007, **23**:3289–3296.

45. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs**. *Proc. Natl. Acad. Sci. U.S.A.* 2005, **102**:2454–2459.

46. Hofacker IL, Fekete M, Stadler PF: **Secondary structure prediction for aligned RNA sequences**. *J. Mol. Biol.* 2002, **319**:1059–1066.

47. Hofacker IL: **Vienna RNA secondary structure server**. *Nucleic Acids Res.* 2003, **31**:3429–3431.

48. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic Acids Res.* 2005, **33**:D121–124.

Figures

Figure 1 - Graph constructing process

(A) Step 1-Structural skeleton construction: Adding a new base-pair $(x_i, x_{i'})$ and aligning that with its best match: $(y_j, y_{j'})$. (B-D) Step 2-Inserting highly probable local alignments: (B) Adding a new column (node) c^*. (C) Extending an existing column (node) c. (D) Merging two columns (nodes) c_1 and c_2 into a single column (node) c^*.

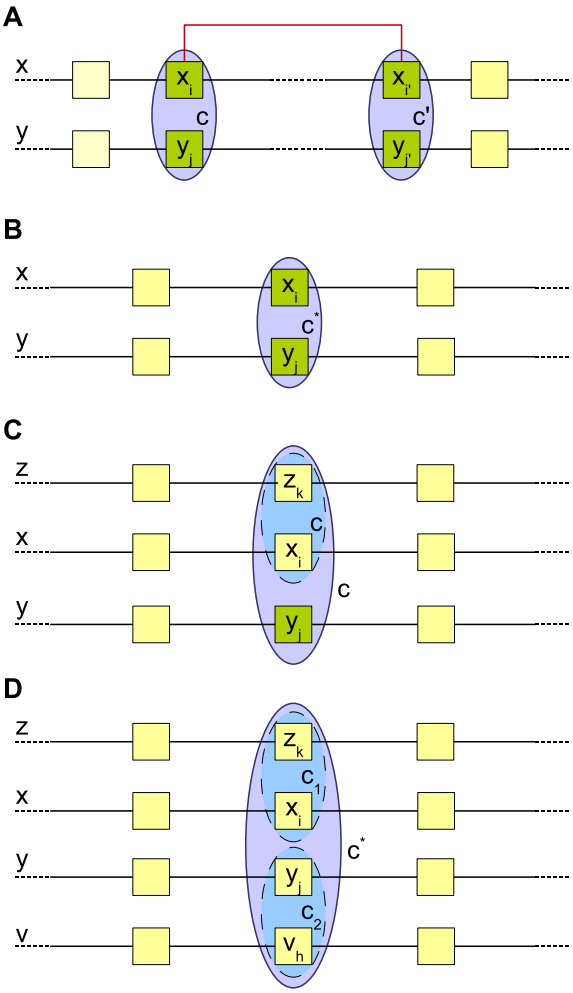


Figure 2 - An illustrative example for the graph construction process in PicXAA-R.

(A) The set of RNA sequences to be aligned. (B) The base-pairs are sorted according to their base-pairing probabilities. (C) The base alignments are sorted according to their transformed alignment probabilities. (D-K) Step 1- Structural skeleton construction: (D,E) Adding a new base-pair (x_2, x_5) and aligning that

with its best match (y_2, y_4). (F,G) Adding a new base-pair (y_1, y_5) and aligning that with its best match (v_1, v_6). (H,I) Extending nodes c_3 and c_4 by adding the base-pair (z_1, z_5) to its best match (y_1, y_5). (J,K) Adding a new base-pair (z_2, z_4) and aligning that with its best match (v_2, v_5). (L-R) Step 2- Inserting highly probable local alignments: (L) Extending the node c_3 by adding the base alignment (x_1, y_1). (M) Merging nodes c_1 with c_5 to include the base alignment (y_2, z_2) and merging nodes c_2 with c_6 to include the base alignment (x_5, z_4).(N) Adding a new node for the alignment (x_3, y_3). (O) Adding a new node for the alignment (z_3, v_3). (P) Merging nodes c_9 and c_{10} to include the alignment (x_3, v_3). (Q) Adding a new node for the alignment (x_4, v_4). (R) Extending the node c_3 by adding the base alignment (x_6, z_6). (S) The final alignment graph $\mathcal{G}$, which gives us the set $\mathcal{A}$ in a legitimate topological ordering. (T) The alignment obtained from $\mathcal{A}$.

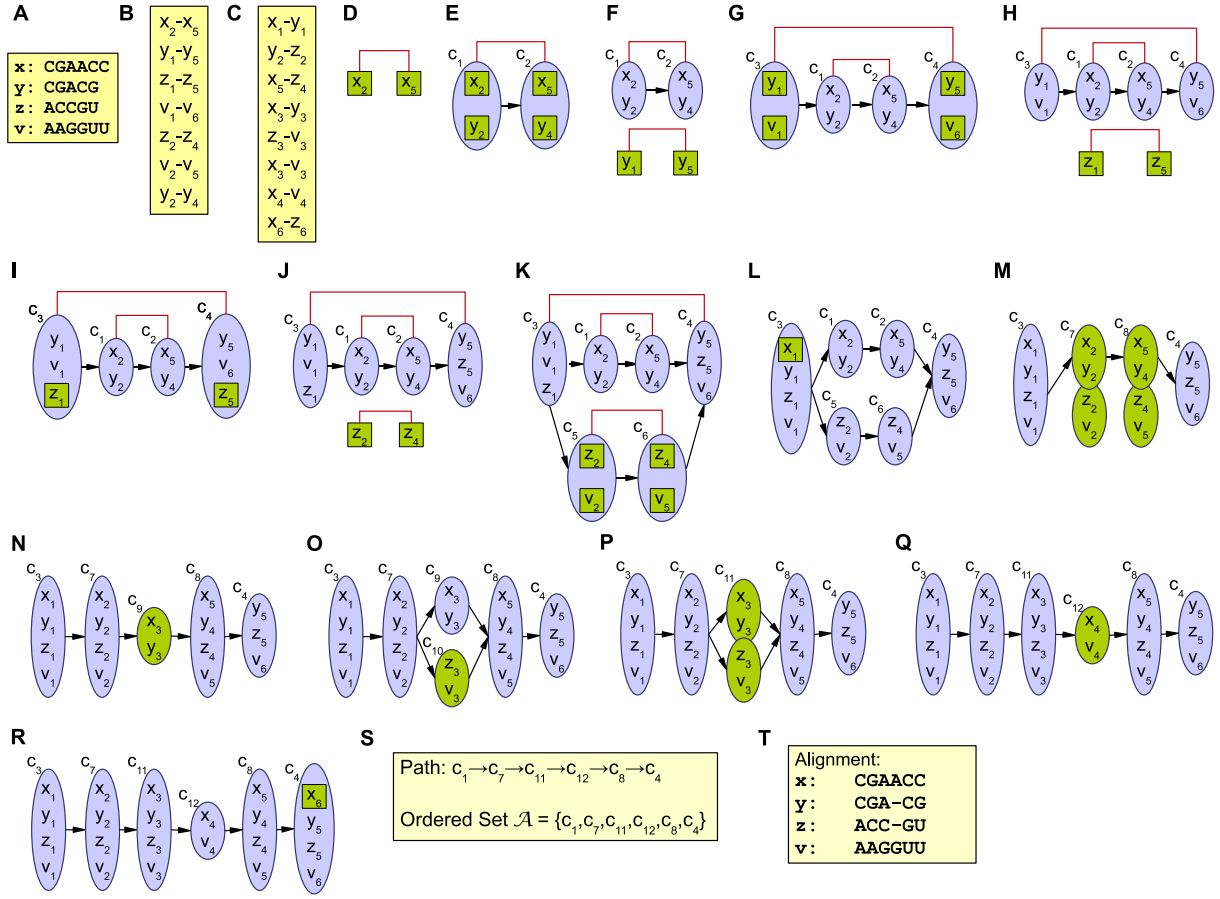


Figure 3 - Accuracy of alignment as a function of the average percent identity

Comparing the accuracy in terms of SPS and SCI scores versus the average percent identity of the alignments in k5, k7, k10, and k15 reference sets of BRAliBase 2.1.

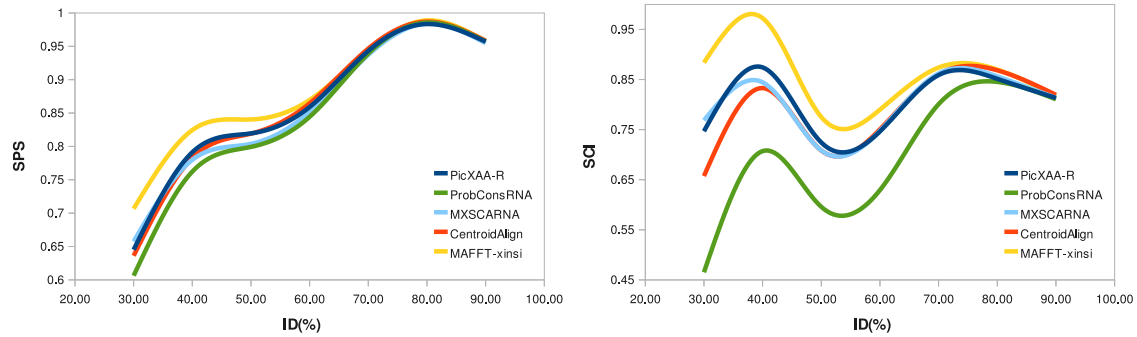
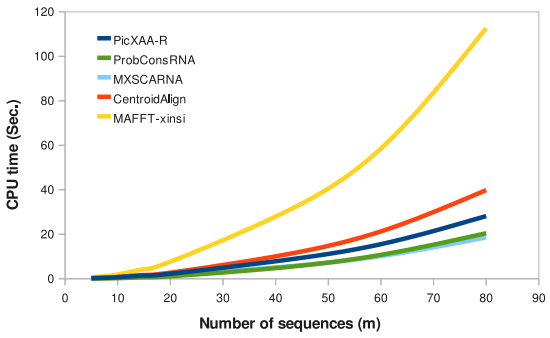


Figure 4 - Complexity analysis

Comparing the dependency of different algorithms to the number of sequences in the alignment. The average running time are shown for sequences in BraliSub and LocExtR datasets.



Tables

Table 1 - Performance evaluation on BRAliBase 2.1

Method	k2	k3	k5	k7	k10	k15	
	SPS/SCI	SPS/SCI	SPS/SCI	SPS/SCI	SPS/SCI	SPS/SCI	TIME
PicXAA-R	84.27 / 85.86	86.59 / 83.35	88.78 / 83.20	90.04 / 81.72	90.97 / 79.95	92.17 / 79.73	6502
ProbConsRNA	83.58 / 82.46	85.46 / 76.54	87.90 / 75.85	88.99 / 74.91	89.90 / 73.25	90.76 / 71.92	1444
MXSCARNA	85.02 / 90.67	86.57 / 85.56	88.43 / 83.44	89.40 / 80.89	90.17 / 78.34	91.26 / 77.18	6024
CentroidAlign	85.55 / 88.64	87.06 / 83.77	88.93 / 82.40	89.99 / 81.23	90.96 / 80.22	91.65 / 79.34	6443
MAFFT-xinsi	85.66 / 90.77	87.76 / 87.11	90.27 / 86.70	91.36 / 85.70	92.26 / 84.73	93.22 / 85.38	12386

Table 2 - Performance evaluation on BraliSub

Method	k5	k7	k10	k15	
	SPS/SCI	SPS/SCI	SPS/SCI	SPS/SCI	TIME
PicXAA-R	73.90 / 51.39	75.06 / 42.37	74.02 / 35.75	75.43 / 31.29	101
ProbConsRNA	70.59 / 34.94	70.18 / 28.45	68.73 / 24.03	66.53 / 18.29	35
MXSCARNA	70.77 / 46.30	69.93 / 35.95	68.58 / 27.91	69.75 / 17.79	84
CentroidAlign	74.23 / 47.26	74.39 / 39.13	74.51 / 35.59	72.92 / 29.14	106
MAFFT-xinsi	78.28 / 57.60	78.56 / 52.10	78.48 / 44.75	79.23 / 38.79	261

Table 3 - Performance evaluation on LocExtR

Method	k20	k40	k60	k80	
	SPS/SCI	SPS/SCI	SPS/SCI	SPS/SCI	TIME
PicXAA-R	71.46 / 17.43	77.52 / 16.08	80.19 / 11.00	82.51 / 10.73	999
ProbConsRNA	64.97 / 10.13	69.08 / 8.12	72.11 / 5.80	74.46 / 6.87	676
MXSCARNA	65.52 / 9.67	68.30 / 8.44	69.45 / 9.15	71.16 / 8.93	662
CentroidAlign	71.68 / 18.63	74.48 / 15.56	77.55 / 11.90	79.32 / 10.07	1359
MAFFT-xinsi	77.02 / 26.30	80.48 / 20.84	81.96 / 16.70	83.52 / 14.00	3791

Table 4 - Performance evaluation on Murlet dataset

Method	SPS	SCI	SEN	PPV	MCC	TIME
PicXAA-R	77.90	48.15	66.08	72.71	68.29	139
ProbConsRNA	76.26	37.47	56.79	78.12	65.10	40
MXSCARNA	74.67	44.28	64.06	74.58	68.37	120
CentroidAlign	77.99	47.80	63.08	74.88	67.48	146
MAFFT-xinsi	78.72	52.94	67.04	74.56	69.64	307