BMC Bioinformatics

RESEARCH Open Access

Fast and accurate clustering of noncoding RNAs using ensembles of sequence alignments and secondary structures

Yutaka Saito[1], Kengo Sato[1,2,3], Yasubumi Sakakibara[1*]

From The Ninth Asia Pacific Bioinformatics Conference (APBC 2011)
Inchon, Korea. 11-14 January 2011

Abstract

Background: Clustering of unannotated transcripts is an important task to identify novel families of noncoding RNAs (ncRNAs). Several hierarchical clustering methods have been developed using similarity measures based on the scores of structural alignment. However, the high computational cost of exact structural alignment requires these methods to employ approximate algorithms. Such heuristics degrade the quality of clustering results, especially when the similarity among family members is not detectable at the primary sequence level.

Results: We describe a new similarity measure for the hierarchical clustering of ncRNAs. The idea is that the reliability of approximate algorithms can be improved by utilizing the information of suboptimal solutions in their dynamic programming frameworks. We approximate structural alignment in a more simplified manner than the existing methods. Instead, our method utilizes *all possible* sequence alignments and *all possible* secondary structures, whereas the existing methods only use *one optimal* sequence alignment and *one optimal* secondary structure. We demonstrate that this strategy can achieve the best balance between the computational cost and the quality of the clustering. In particular, our method can keep its high performance even when the sequence identity of family members is less than 60%.

Conclusions: Our method enables fast and accurate clustering of ncRNAs. The software is available for download at http://bpla-kernel.dna.bio.keio.ac.jp/clustering/.

Background

Noncoding RNAs (ncRNAs) serve a variety of cellular functions depending on their primary sequences and secondary structures [1]. A group of ncRNAs sharing sequence and structural homology is annotated as one family, and included in the database [2]. Recently, high-throughput transcriptome sequencing has uncovered tens of thousands of ncRNAs that lack significant homology to known families [3,4]. Thus, evaluating homology *among* these unannotated transcripts, that is, *clustering* has become an important task to identify novel ncRNA families [5,6].

Accurate clustering of ncRNAs needs a reliable similarity measure that takes into account primary sequences and secondary structures. Given a pair of sequences without known structures, the Sankoff algorithm [7] simultaneously predicts their sequence alignment and consensus secondary structure (*i.e.*, structural alignment); thus, the obtained alignment score can be a suitable choice for a similarity measure. However, the original Sankoff algorithm is too time-consuming to deal with an all-against-all comparison of many sequences required in clustering procedures. To address this problem, similarity measures based on the approximation of the Sankoff algorithm have been proposed, and shown to be applicable to hierarchical clustering [8-10]. Each method has its own heuristics to reduce the huge dynamic programming matrix used in the

* Correspondence: yasu@bio.keio.ac.jp
[1]Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan
Full list of author information is available at the end of the article

Sankoff algorithm. Will *et al.* [8] have developed LocARNA that precludes unsure secondary structures including low-probability base pairs. Torarinsson *et al.* [9] have developed FOLDALIGNM based on the FOLDALIGN program [11] that dynamically excludes low-scoring sequence alignments by means of length-dependant thresholds. Sato *et al.* [10] have developed Stem kernel that employs heuristics similar to LocARNA, but further precludes secondary structures including any bifurcation.

Although the approximate Sankoff-style algorithms have enabled similarity measures based on structural alignment, the quality of clustering results has not been so high. In the previous studies [6,8,12], resultant clusters in a hierarchical tree were quite unclear, requiring additional verification or manual inspection. This was partly because of the diversity within one ncRNA family. Most ncRNA families have only less than 60% identity at the primary sequence level [2], and cannot be correctly aligned without taking into account secondary structures [13]. The approximate Sankoff-style algorithms seemed to be degraded by discarding the secondary structures in the excluded portion of the dynamic programming matrix.

To improve the reliability of the approximate Sankoff-style algorithms, we focus attention on the information of suboptimal structural alignments. Among the existing methods, LocARNA and FOLDALIGN calculate the similarity based on the score of *one optimal* structural alignment. This means that these methods ignore the scores of suboptimal structural alignments, and only use *one optimal* sequence alignment and *one optimal* secondary structure. In contrast, Stem kernel sums up the scores of structural alignments allowed in the approximate Sankoff-style algorithm, incorporating a *subset of* sequence alignments and a *subset of* secondary structures. Thanks to this strategy, Stem kernel gives comparable clustering results to LocARNA, while employing the more reduced dynamic programming matrix. These observations suggest the possibility that we can design a more reliable similarity measure by utilizing *all possible* sequence alignments and *all possible* secondary structures. This is not trivial because if we naively try to incorporate all possible structural alignments, it will require the full-size dynamic programming matrix used in the original Sankoff algorithm with the prohibitive computational cost.

In this paper, we describe a new similarity measure for the hierarchical clustering of ncRNAs. We approximate the problem of structural alignment by the two separate problems: the prediction of sequence alignment, and the prediction of secondary structure for each sequence. For this purpose, the Sankoff algorithm for structural alignment is approximated by the combination of the Smith-Waterman (SW) algorithm [14] for sequence alignment, and the McCaskill algorithm [15] for secondary structures. The approximation allows to obtain all possible sequence alignments from the SW algorithm, and all possible secondary structures from the McCaskill algorithm, much faster than obtaining all possible structural alignments from the original Sankoff algorithm. We first describe a similarity measure using the scores of all possible sequence alignments between two RNAs. Then, we design a scoring function for these sequence alignments using all possible secondary structures of each of the two RNAs. We start from a scoring function that measures the similarity between two secondary structures using the state of base pairing at each position. The proposed scoring function is defined as an expectation of this scoring function over all possible secondary structures of each of the two RNAs.

We demonstrate that our method can achieve the best balance between the computational cost and the quality of the clustering among the existing methods. In particular, our method can keep its high performance even when the sequence identity of family members is less than 60%.

Methods

In this section, we propose a new method for measuring the similarity between two RNA sequences without known structures. The proposed method is applied to the hierarchical clustering of ncRNAs with the weighted pair group method with averaging (WPGMA) algorithm. Given a set of sequences, we calculate an all-against-all similarity matrix using our method. Then, we derive the distance matrix by one minus the similarity, and obtain the cluster tree by the WPGMA algorithm.

The idea of our similarity measure is to approximate the Sankoff algorithm for structural alignment by the combination of the SW algorithm for sequence alignment, and the McCaskill algorithm for secondary structures. This approximation allows to utilize the ensembles of *all possible* sequence alignments and *all possible* secondary structures separately from each of the two algorithms. First, we describe a similarity measure using the scores of all possible sequence alignments between two RNAs. Next, we design a scoring function for these alignments using all possible secondary structures of each of the two RNAs.

Ensemble of all possible sequence alignments

To measure the similarity between two RNAs, one common approach is to perform pairwise alignment, and to calculate its alignment score. The Sankoff algorithm simultaneously models sequence alignments and secondary structures, and is extremely time-consuming. Therefore, we first approximate the Sankoff algorithm by the SW algorithm that only models sequence alignments apart from secondary structures. Although this is a strong approximation, we attempt to improve the reliability by utilizing *all possible* sequence alignments rather than *one optimal* sequence alignment.

For an RNA sequence $\mathbf{x}$, we denote its length by $|\mathbf{x}|$. For each position $1 \le i \le |\mathbf{x}|$ in $\mathbf{x}$, we denote the nucleotide by $x_i \in$ {A, C, G, U}.

For two sequences, $\mathbf{x}$ and $\mathbf{y}$, let $\Pi_{\mathbf{xy}}$ be the set of all possible sequence alignments in the SW algorithm. Let $\pi_{\mathbf{xy}}$ denote one particular sequence alignment in $\Pi_{\mathbf{xy}}$.

We calculate the similarity between $\mathbf{x}$ and $\mathbf{y}$ by accumulating the alignment score of $\pi_{\mathbf{xy}}$ over $\Pi_{\mathbf{xy}}$. For this purpose, we employ local alignment (LA) kernel [16] defined as follows:

$$K(\mathbf{x},\mathbf{y}) = \sum_{\pi_{\mathrm{xy}} \in \Pi_{\mathrm{xy}}} e^{\beta \mathrm{Score}(\pi_{\mathrm{xy}})}, \tag{1}$$

where $\beta \ge 0$ is a parameter, and Score($\pi_{\mathbf{xy}}$) is the alignment score of $\pi_{\mathbf{xy}}$ under a given scoring scheme (gap penalties and match scores). In practice, we take the logarithm of LA kernel, and similarity values are normalized to range from 0 to 1:

$$K_n(\mathbf{x},\mathbf{y}) = \frac{\log K(\mathrm{x,y})}{\sqrt{\log K(\mathrm{x,x}) \log K(\mathrm{y,y})}}. \tag{2}$$

LA kernel (1) can be computed by the variant of the SW algorithm as follows:

```
Initialization:
    for $i \in \{0,\dots,|\mathbf{x}|\}$ and $j \in \{0,\dots,|\mathbf{y}|\}$ do
        $M(i,0) = I_X(i,0) = I_Y(i,0) = T_X(i,0) = T_Y(i,0) = 0$
        $M(0,j) = I_X(0,j) = I_Y(0,j) = T_X(0,j) = T_Y(0,j) = 0$
    end for
Iteration:
    for $i \in \{1,\dots,|\mathbf{x}|\}$ and $j \in \{1,\dots,|\mathbf{y}|\}$ do
        $M(i,j) = e^{\beta S_{\mathrm{xy}}}(i,j)(1 + I_X(i-1,j-1) + I_Y(i-1,j-1) + M(i-1,j-1))$
        $I_X(i,j) = e^{\beta g} M(i-1,j) + e^{\beta d} I_X(i-1,j)$
        $I_Y(i,j) = e^{\beta g}(M(i,j-1) + I_X(i,j-1)) + e^{\beta d} I_Y(i,j-1)$
        $T_X(i,j) = M(i-1,j) + T_X(i-1,j)$
        $T_Y(i,j) = M(i,j-1) + T_X(i,j-1) + T_Y(i,j-1)$
    end for
Termination:
    $K(\mathbf{x},\mathbf{y}) = 1 + T_X(|\mathbf{x}|,|\mathbf{y}|) + T_Y(|\mathbf{x}|,|\mathbf{y}|) + M(|\mathbf{x}|,|\mathbf{y}|)$
```

where the parameters g and d are the penalties for gap opening and gap extension, respectively, and $S_{\mathbf{xy}}(i, j)$ is a scoring function for matching the i-th position in $\mathbf{x}$ and the *j-th* position in $\mathbf{y}$. The design of $S_{\mathbf{xy}}(i, j)$ impacts the performance of the resulting similarity measure, and will be described later.

At this point, we note that our method can take into account all possible sequence alignments in $O(|\mathbf{x}||\mathbf{y}|)$ time. If we use the exact Sankoff algorithm instead, it takes prohibitive $O(|\mathbf{x}|^3|\mathbf{y}|^3)$ time, which is not practical. In the case of the approximate Sankoff-style algorithms employed in the existing methods, all possible sequence alignments cannot be incorporated to the reduced dynamic programming matrix. Therefore, LA kernel based on the SW algorithm is an efficient way to deal with the ensemble of all possible sequence alignments.

Ensemble of all possible secondary structures

To design a scoring function $S_{\mathbf{xy}}(i, j)$ for LA kernel, we need secondary structures of $\mathbf{x}$ and $\mathbf{y}$. As mentioned above, the Sankoff algorithm models secondary structures simultaneously with sequence alignments which we have already modeled by the SW algorithm. Therefore, we next employ the McCaskill algorithm that only models secondary structures apart from sequence alignments. Although this is an additional approximation, we attempt to improve the reliability by utilizing *all possible* secondary structures rather than *one optimal* secondary structure.

For an RNA sequence x, let $\Theta_{\mathbf{x}}$ be the set of all possible secondary structures. Let $\theta_{\mathbf{x}}$ denote one particular secondary structure in $\Theta_{\mathbf{x}}$. We represent a secondary structure as a set of binary variables $\theta_{\mathbf{x}} = \{\theta_{\mathbf{x}}(i, j)\}_{1 \le i < j \le |\mathbf{x}|}$, where $\theta_{\mathbf{x}}(i, j) = 1$ means that the i-th position and the *j-th* position in $\mathbf{x}$ form a base pair. For each position $1 \le i \le |\mathbf{x}|$ in $\mathbf{x}$, we represent the state of base-pairing using three kinds of binary variable: $L_{\mathbf{x}}(i) = \sum_{j:j>i} \theta_{\mathbf{x}}(i, j) = 1$ means that a base pair is formed with one of the downstream positions; $R_{\mathbf{x}}(i) = \sum_{j:j<i} \theta_{\mathbf{x}}(j, i) = 1$ means that a base pair is formed with one of the upstream positions; and $U_{\mathbf{x}}(i) = 1 - L_{\mathbf{x}}(i) - R_{\mathbf{x}}(i) = 1$ means that the position is unpaired. Given a fixed pair of secondary structures, $\theta_{\mathbf{x}}$ and $\theta_{\mathbf{y}}$, we can measure the similarity between the i-th position in $\mathbf{x}$ and the *j-th* position in $\mathbf{y}$ using their state of base pairing:

$W_{\mathbf{xy}}(i, j|\theta_{\mathbf{x}}, \theta_{\mathbf{y}}) = \alpha\, (L_{\mathbf{x}}(i)L_{\mathbf{y}}(j) + R_{\mathbf{x}}(i)R_{\mathbf{y}}(j)) + s(x_i, y_j)\, U_{\mathbf{x}}(i)U_{\mathbf{y}}(j)$, (3)

where $\alpha \ge 0$ is a weight parameter for structural similarity, and $s(x_i, y_j)$ is a substitution matrix for RNA sequences like the RIBOSUM 85-60 matrix [17]. This scoring function takes a non-zero value in three different cases: it takes α when both of the two positions form a base pair with one of their downstream positions, respectively; it takes α when both of the two positions form a base pair with one of their upstream positions, respectively; and it takes $s(x_i, y_j)$ when both of the two positions are unpaired.

The McCaskill algorithm defines a probability distribution $P(\theta_{\mathbf{x}}|\mathbf{x})$ over $\Theta_{\mathbf{x}}$. The binary variables $\theta_{\mathbf{x}}(i, j)$ and $\{L_{\mathbf{x}}(i), R_{\mathbf{x}}(i), U_{\mathbf{x}}(i)\}$ are converted to the probabilities by taking the expectation over $\Theta_{\mathbf{x}}$. For $\theta_{\mathbf{x}}(i, j)$, we obtain a base-pairing probability $P_{\mathbf{x}}(i, j)$ that the i-th and the *j-th* positions form a base pair:

$$P_{\mathrm{x}}(i,j) = \sum_{\theta_{\mathrm{x}}\in\Theta_{\mathrm{x}}} \theta_{\mathrm{x}}(i,j)P(\theta_{\mathrm{x}}\,|\,\mathbf{x}).$$

For $\{L_{\mathbf{x}}(i), R_{\mathbf{x}}(i), U_{\mathbf{x}}(i)\}$, we obtain three kinds of probability that the i-th position is paired with one of the downstream/upstream positions, or unpaired, respectively:

$$\begin{aligned}
P_{\mathrm{x}}^{L}(i) &= \sum_{\theta_{\mathrm{x}}\in\Theta_{\mathrm{x}}} L_{\mathrm{x}}(i)P(\theta_{\mathrm{x}}\,|\,\mathrm{x}) = \sum_{\theta_{\mathrm{x}}\in\Theta_{\mathrm{x}}}\sum_{j:j>i}\theta_{\mathrm{x}}(i,j)P(\theta_{\mathrm{x}}\,|\,\mathrm{x}) = \sum_{j:j>i}P_{\mathrm{x}}(i,j),\\
P_{\mathrm{x}}^{R}(i) &= \sum_{\theta_{\mathrm{x}}\in\Theta_{\mathrm{x}}} R_{\mathrm{x}}(i)P(\theta_{\mathrm{x}}\,|\,\mathrm{x}) = \sum_{\theta_{\mathrm{x}}\in\Theta_{\mathrm{x}}}\sum_{j:j<i}\theta_{\mathrm{x}}(j,i)P(\theta_{\mathrm{x}}\,|\,\mathrm{x}) = \sum_{j:j<i}P_{\mathrm{x}}(j,i),\\
P_{\mathrm{x}}^{U}(i) &= \sum_{\theta_{\mathrm{x}}\in\Theta_{\mathrm{x}}} U_{\mathrm{x}}(i)P(\theta_{\mathrm{x}}\,|\,\mathrm{x}) = 1 - P_{\mathrm{x}}^{L}(i) - P_{\mathrm{x}}^{R}(i).
\end{aligned}$$

We design a scoring function $S_{\mathbf{xy}}(i, j)$ by taking the expectation of (3) over $\Theta_{\mathbf{x}}$ and $\Theta_{\mathbf{y}}$:

$$\begin{aligned}
S_{\mathrm{xy}}(i,j) &= \sum_{\theta_{\mathrm{x}}\in\Theta_{\mathrm{x}}}\sum_{\theta_{\mathrm{y}}\in\Theta_{\mathrm{y}}} W_{\mathrm{xy}}(i,j\,|\,\theta_{\mathrm{x}},\theta_{\mathrm{y}})P(\theta_{\mathrm{x}}\,|\,\mathrm{x})P(\theta_{\mathrm{y}}\,|\,\mathrm{y})\\
&= \alpha(P_{\mathrm{x}}^{L}(i)P_{\mathrm{y}}^{L}(j) + P_{\mathrm{x}}^{R}(i)P_{\mathrm{y}}^{R}(j)) + s(x_i, y_j)P_{\mathrm{x}}^{U}(i)P_{\mathrm{y}}^{U}(j).
\end{aligned} \tag{4}$$

The proposed method is obtained by combining the normalized LA kernel (2) with the scoring function (4).

It should be noted that our method can take into account all possible secondary structures in $O(|\mathbf{x}|^3 + |\mathbf{y}|^3)$ time, thanks to the McCaskill algorithm. Just as in all possible sequence alignments, the exact Sankoff algorithm results in $O(|\mathbf{x}|^3|\mathbf{y}|^3)$ time, and the existing methods cannot incorporate all possible secondary structures. Our method requires $O(|\mathbf{x}||\mathbf{y}|) + O(|\mathbf{x}|^3 + |\mathbf{y}|^3)$ time in total, which is more efficient than the exact Sankoff algorithm. Therefore, our strategy that combines the SW algorithm and the McCaskil algorithm allows to utilize the ensemble information with the reasonable computational cost.

Variations of the proposed method

The scoring function (4) proposed in this study is similar to the scoring function used in BPLA kernel [18,19]. BPLA kernel is a prediction method that we previously developed for detecting new members of known ncRNA families. Although BPLA kernel was not applied to clustering problems in our previous study, we here clarify its relation to the proposed method. The scoring function used in BPLA kernel is defined as follows:

$$\begin{aligned}
S_{\mathrm{xy}}^{\mathrm{BPLA}}(i,j) &= \alpha\left(\sqrt{P_{\mathrm{x}}^{L}(i)P_{\mathrm{y}}^{L}(j)} + \sqrt{P_{\mathrm{x}}^{R}(i)P_{\mathrm{y}}^{R}(j)}\right) + s(x_i, y_j)\sqrt{P_{\mathrm{x}}^{U}(i)P_{\mathrm{y}}^{U}(j)}\\
&= \alpha(C^{L}P_{\mathrm{x}}^{L}(i)P_{\mathrm{y}}^{L}(j) + C^{R}P_{\mathrm{x}}^{R}(i)P_{\mathrm{y}}^{R}(j)) + s(x_i, y_j)C^{U}P_{\mathrm{x}}^{U}(i)P_{\mathrm{y}}^{U}(j),
\end{aligned} \tag{5}$$

where $C^L = 1/\sqrt{P_x^L(i)P_y^L(j)}$, $C^R = 1/\sqrt{P_x^R(i)P_y^R(j)}$, and $C^U = 1/\sqrt{P_x^U(i)P_y^U(j)}$. Therefore, the scoring function (5) can be regarded as a variation of the proposed scoring function (4) with the additional coefficients C^L, C^R, and C^U. These coefficients take large values when the probabilities $P_{\mathrm{x}}^{\cdot}(i)$ and $P_{\mathrm{y}}^{\cdot}(j)$ are small. That is, BPLA kernel emphasizes the contribution of low-probability (unsure) secondary structures compared to the proposed method. In the next section, we experimentally verify this theoretical implication; the proposed method outperforms BPLA kernel.

Because of the resemblance between the scoring functions, (4) and (5), we set the parameters of the proposed method as used in BPLA kernel: $\alpha = 1.0$, $\beta = 0.1$, $g = -27$, and $d = -0.1$

Results and discussion

In this section, we examine the performance of the proposed method in the hierarchical clustering of ncRNAs.

Dataset and experimental system

We compared our method with the state-of-the-art methods developed for the hierarchical clustering of ncRNAs: LocARNA v1.5.2 [8], FOLDALIGN v2.1.1 [11], and Stem kernel v216c [10]. We also performed the experiments with CLUSTALW v1.83 [20] and LA kernel by setting $\{P_{\mathrm{x}}^{L}(i) = 0, P_{\mathrm{x}}^{R}(i) = 0, P_{\mathrm{x}}^{U}(i) = 1\}$ in our method (4).

We can summarize our method and the existing methods as follows. Our method utilizes *all possible* sequence alignments and *all possible* secondary structures. LocARNA and FOLDALIGN only use *one optimal* sequence alignment and *one optimal* secondary structures. Stem kernel utilizes a *subset of* all possible sequence alignments and a *subset of* all possible secondary structures. CLUSTALW and LA kernel ignore secondary structures; CLUSTALW only uses *one optimal* sequence alignment, while LA kernel utilizes *all possible* sequence alignments.

We created a dataset as summarized in Table 1. This dataset was collected from the BRAliBASE benchmark v2.1 [13], which includes multiple alignments of a broad range of ncRNA families established in the Rfam database [2]. We treated each multiple alignment as a reference cluster, and each ncRNA sequence in a multiple alignment as a member sequence. The reference clusters were divided into four categories according to their sequence identity: 20–39%, 40–59%, 60–79%, and 80–

Table 1 Summary of the dataset

	20–39%	40–59%	60–79%	80–99%
#clusters	13	21	34	36
#members	3.2	5.0	3.8	4.6
Length	138	130	111	102

#clusters: number of reference clusters; each reference cluster represents a different ncRNA family.

#members: average number of member sequences per reference cluster. Length: average length of sequences over all reference clusters. The dataset is divided by the sequence identity in a reference cluster.

99%. We sampled the dataset ten times from the BRAliBASE benchmark, and evaluated the average performance.

We produced three versions of dataset. First, we used ncRNA sequences without modification, and named them the "normal" dataset. Second, we concatenated random sequences to both ends of ncRNA sequences, and named them the "plus flanking regions" dataset. This dataset was intended to simulate the situation where we do not know the exact boundaries of unannotated transcripts. A random sequence was generated from a ncRNA sequence so that it had the quarter length and the same dinucleotide contents. Third, we added false reference clusters, each of which contains one random sequence, and named them the "plus unrelated sequences" dataset. This dataset was intended to simulate the situation where non-functional ncRNAs arise from transcriptional noises. Therefore, we evaluated whether a false reference cluster could be a resultant cluster with a single member. We used a quarter number of false reference clusters compared to true reference clusters. A random sequence was generated from a ncRNA sequence so that it had the same length and the same dinucleotide contents.

We evaluated the overall quality of the cluster tree by the ROC analysis proposed in [8]. (Note that we can obtain different resultant clusters from a cluster tree depending on a distance threshold to cut the branches.) Given a distance threshold, the number of true positives (*TP*) was defined as the number of sequence pairs that belong to the same reference cluster and are correctly assigned to the same resultant cluster. Analogously, the numbers of false positives (*FP*), true negatives (*TN*), and false negatives (*FN*) are defined, respectively, by counting the pairs from different reference clusters but the same resultant cluster, the pairs from different reference clusters and different resultant clusters, and the pairs from the same reference cluster but different resultant clusters. The ROC analysis was performed by plotting true positive rates $TP/(TP + FN)$ versus false positive rates $FP/(TN + FP)$ for different distance thresholds. The quality of the clustering was measured by the area under the ROC curve (AUC). We measured the total time for computing similarity matrices on a 2.53 GHz Intel Xeon processor.

Quality of the clustering

We first examined the quality of the clustering for the "normal" dataset (Figure 1). Our method achieved the better or comparable AUC to the existing methods in all the range of sequence identity. The accuracy of our method was especially remarkable in the sequence identity range below 60%, where the existing methods resulted in low AUC. This means that our method successfully grouped diverse member sequences in each reference cluster by detecting their remote homology.

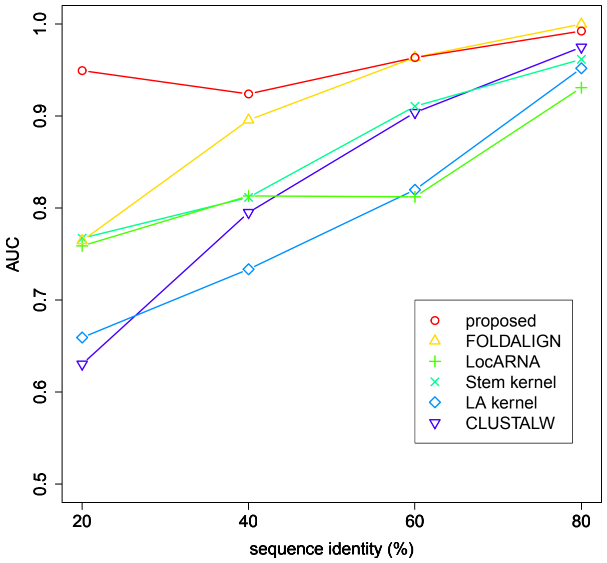

Figure 1 Quality of the clustering for the "normal" dataset. For each sequence identity range, the overall quality of the cluster tree is evaluated by the AUC.

Our results can be attributed to the design of each method. The AUC of CLUSTALW and LA kernel, which ignore secondary structures and only use sequence alignments, drastically fell down as the sequence identity decreased. LocARNA, FOLDALIGN, and Stem kernel, which consider secondary structures, kept the AUC relatively moderate in the low sequence identity range. However, their accuracy was still limited when the sequence identity was extremely low (20–39%) because these methods only use *one optimal* secondary structure or a *subset of* secondary structures. Our method, which utilizes *all possible* sequence alignments and *all possible* secondary structures, achieved the sufficiently high AUC in this region. These results suggest that our design of the similarity measure is effective for identifying a broad range of ncRNA families.

Figure 2 compares an example of the cluster tree between our method and FOLDALIGN in the sequence identity rage of 20–39%. As indicated by AUC, our method produced the more accurate cluster tree than FOLDALIGN, and reconstructed ncRNA families as compact clusters. Although the cluster tree of FOLDALIGN was largely consistent with the references in terms of its topology, boundaries of resultant clusters were quite unclear. In the actual application of hierarchical clustering, we need to choose a proper distance threshold for extracting clusters from a given tree. In this sense, the cluster tree of FOLDALIGN was not

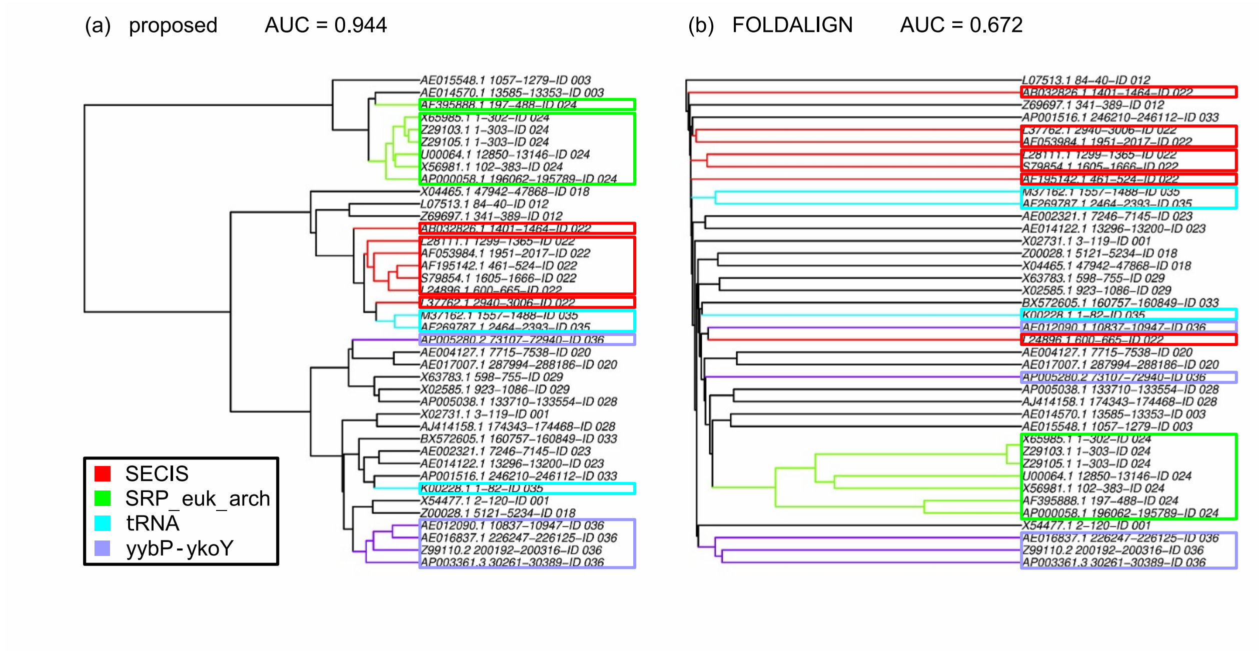
(a) proposed AUC = 0.944
AE015548.1 1057–1279–ID 003
AE014570.1 13585–13353–ID 003
AF395888.1 197–488–ID 024
X65985.1 1–302–ID 024
Z29103.1 1–303–ID 024
Z29105.1 1–303–ID 024
U00064.1 12850–13146–ID 024
X56981.1 102–383–ID 024
AP000058.1 196062–195789–ID 024
X04465.1 47942–47868–ID 018
L07513.1 84–40–ID 012
Z69697.1 341–389–ID 012
AB032826.1 1401–1464–ID 022
L28111.1 1299–1365–ID 022
AF053984.1 1951–2017–ID 022
AF195142.1 461–524–ID 022
S79854.1 1605–1666–ID 022
L24896.1 600–665–ID 022
L37762.1 2940–3006–ID 022
M37162.1 1557–1488–ID 035
AF269787.1 2464–2393–ID 035
AP005280.2 73107–72940–ID 036
AE004127.1 7715–7538–ID 020
AE017007.1 287994–288186–ID 020
X63783.1 598–755–ID 029
X02585.1 923–1086–ID 029
AP005038.1 133710–133554–ID 028
X02731.1 3–119–ID 001
AJ414158.1 174343–174468–ID 028
BX572605.1 160757–160849–ID 033
AE002321.1 7246–7145–ID 023
AE014122.1 13296–13200–ID 023
AP001516.1 246210–246112–ID 033
K00228.1 1–82–ID 035
X54477.1 2–120–ID 001
Z00028.1 5121–5234–ID 018
AE012090.1 10837–10947–ID 036
AE016837.1 226247–226125–ID 036
Z99110.2 200192–200316–ID 036
AP003361.3 30261–30389–ID 036
🟥 SECIS
🟩 SRP_euk_arch
🟦 tRNA
🟪 yybP-ykoY
(b) FOLDALIGN AUC = 0.672
L07513.1 84–40–ID 012
AB032826.1 1401–1464–ID 022
Z69697.1 341–389–ID 012
AP001516.1 246210–246112–ID 033
L37762.1 2940–3006–ID 022
AF053984.1 1951–2017–ID 022
L28111.1 1299–1365–ID 022
S79854.1 1605–1666–ID 022
AF195142.1 461–524–ID 022
M37162.1 1557–1488–ID 035
AF269787.1 2464–2393–ID 035
AE002321.1 7246–7145–ID 023
AE014122.1 13296–13200–ID 023
X02731.1 3–119–ID 001
Z00028.1 5121–5234–ID 018
X04465.1 47942–47868–ID 018
X63783.1 598–755–ID 029
X02585.1 923–1086–ID 029
BX572605.1 160757–160849–ID 033
K00228.1 1–82–ID 035
AE012090.1 10837–10947–ID 036
L24896.1 600–665–ID 022
AE004127.1 7715–7538–ID 020
AE017007.1 287994–288186–ID 020
AP005280.2 73107–72940–ID 036
AP005038.1 133710–133554–ID 028
AJ414158.1 174343–174468–ID 028
AE014570.1 13585–13353–ID 003
AE015548.1 1057–1279–ID 003
X65985.1 1–302–ID 024
Z29103.1 1–303–ID 024
Z29105.1 1–303–ID 024
U00064.1 12850–13146–ID 024
X56981.1 102–383–ID 024
AF395888.1 197–488–ID 024
AP000058.1 196062–195789–ID 024
X54477.1 2–120–ID 001
AE016837.1 226247–226125–ID 036
Z99110.2 200192–200316–ID 036
AP003361.3 30261–30389–ID 036

Figure 2 Comparison of the cluster trees between the proposed method and FOLDALIGN. For the sequence identity range of 20–39% in the “normal” dataset, an example of the cluster tree is shown with its AUC. In the leaf nodes, the strings such as “XXX-ID YYY” mean that the sequence XXX belongs to the reference cluster YYY. The four reference clusters that have more than two member sequences are colored, and their corresponding ncRNA families are noted.

sufficient for the practical use. In fact, the previous studies that employed clustering approaches required manual inspection to compensate for ambiguous cluster trees [6,8,12]. The cluster tree of our method was much more clear and easier to interpret than the existing methods. These results suggest that our method can reduce human labor costs of clustering approaches, and help to identify novel ncRNAs families.

Next, we evaluated the quality of the clustering for the “plus flanking regions” dataset (Figure 3), and the “plus unrelated sequences” dataset (Figure 4). In both cases, we observed the same tendency as in the results for the “normal” dataset (Figure 1). Our method kept high accuracy in all the range of sequence identity, and achieved the best AUC in the sequence identity range below 60%. These results further support the effectiveness of our method in the practical situations that involve flanking regions and unrelated sequences.

Differences in the variations of the proposed method

As described in Methods, the proposed method has the theoretical advantage compared to BPLA kernel, which can be regarded as a variation of our method. To verify this point experimentally, we compare the proposed method and BPLA kernel using the scoring functions (4) and (5), respectively.

Figure 5 presents the experimental results. The proposed method achieved the slightly better AUC in the sequence identity range below 60%. These results are consistent with the fact that BPLA kernel emphasizes

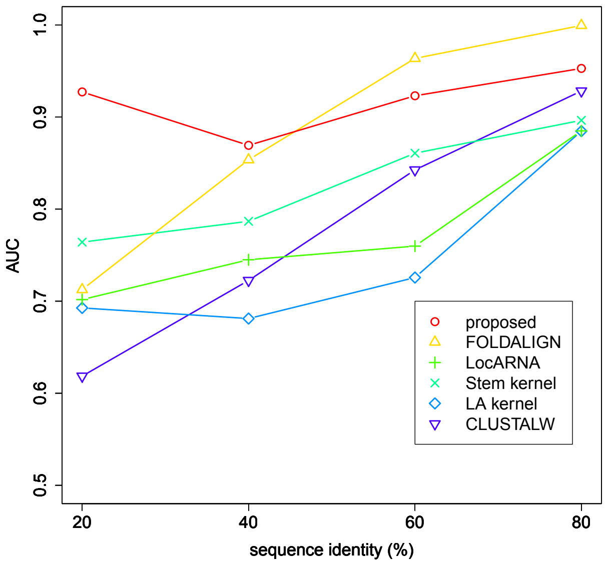

Figure 3 Quality of the clustering for the “plus flanking regions” dataset. For each sequence identity range, the overall quality of the cluster tree is evaluated by the AUC.

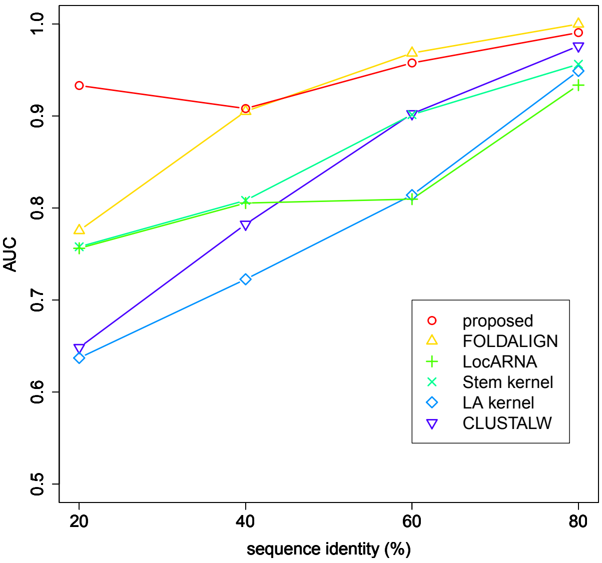

Figure 4 Quality of the clustering for the "plus unrelated sequences" dataset. For each sequence identity range, the overall quality of the cluster tree is evaluated by the AUC.

the contribution of unsure secondary structures compared to the proposed method. The proposed scoring function (4) has the theoretical justification as the expectation of the primitive scoring function (3) over all possible secondary structures. Our results provide an experimental verification of the superiority of the proposed scoring function.

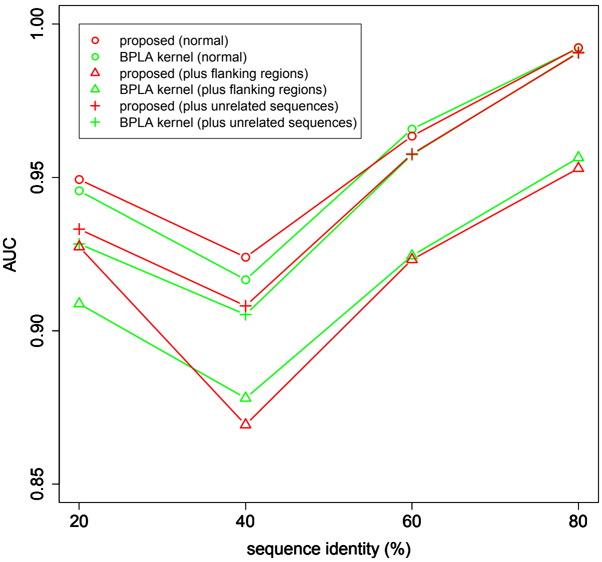

Figure 5 Differences in the variations of the proposed method. The proposed method is compared to BPLA kernel using three versions of the dataset. Note that BPLA kernel can be regarded as a variation of the proposed method.

Table 2 Computational cost of the similarity measures

Method	Computation time (s)		
	Normal	Plus flanking regions	Plus unrelated sequences
proposed	95	222	199
FOLDALIGN	71748	226066	167228
LocARNA	9704	64679	30287
Stem kernel	61	179	138
LA kernel	71	163	160
CLUSTALW	4	43	6

The total time for computing similarity matrices is shown for three versions of the dataset.

Computational cost

Finally, we evaluated the computational cost of the similarity measures using three version of the dataset (Table 2). Our method was faster than LocARNA and FOLDALIGN by several orders of magnitude, and achieved the comparable computational cost to Stem kernel. Considering the high accuracy of our method (Figures 1, 2, 3, 4), we achieved the best balance between the computational cost and the quality of the clustering among the existing methods.

In the design of the proposed method, our idea was to improve the reliability of approximate algorithms by the information of suboptimal solutions in their dynamic programming frameworks. Among LocARNA and FOLDALIGN, which only use *one optimal* solution in their approximate Sankoff-style algorithms, there was a trade-off that LocARNA was faster but less accurate than FOLDALIGN (Figure 1, and Table 2). Stem kernel, which utilizes a *subset of* solutions in the more approximate Sankoff-style algorithm, partly improved this problem, being faster and more accurate than LocARNA. Our method, which utilizes *all possible* solutions in the combination of the Smith-Waterman algorithm and the McCaskill algorithm, successfully overcome the trade-off. These results suggest that our strategy is essential to enable fast and accurate clustering of ncRNAs.

Conclusions

We have described a new method for the hierarchical clustering of ncRNAs, which can be applied to the identification of novel ncRNA families. Our method can achieve the best balance between the computational cost and the quality of the clustering compared to the existing methods.

The performance of the clustering is determined by similarity measures based on the scores of structural alignment. The existing similarity measures, which only use *one optimal* structural alignment, suffer from the trade-off between time-consuming accurate algorithms and fast approximate algorithms. Our similarity measure, which is designed to utilize *all possible* sequence alignments and *all possible* secondary structures, have overcome this problem. The improvement is especially remarkable when the similarity among family members is not detectable at the primary sequence level.

In conclusion, our method enables fast and accurate clustering of ncRNAs, providing a promising way to explore the functional diversity of ncRNAs.

Acknowledgements
This work was supported by KAKENHI (Grant-in-Aid for Scientific Research) on Innovative Areas (No.221S0002) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. This work was also supported by KAKENHI on Priority Area "Comparative Genomics" (No.17018029), and by a grant from "Functional RNA Project" funded by the New Energy and Industrial Technology Development Organization (NEDO), Japan. KS was supported in part by Global COE program "Deciphering Biosphere from Genome Big Bang", and by KAKENHI for Young Scientists (B) (No.22700305) from MEXT, Japan.
This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 1, 2011: Selected articles from the Ninth Asia Pacific Bioinformatics Conference (APBC 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/12?issue=S1.

Author details
[1]Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan. [2]Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan. [3]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan.

Authors' contributions
Y Saito and KS developed the algorithm and wrote the code. Y Saito performed the experiments and drafted the manuscript. Y Sakakibara initiated and coordinated the project. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 15 February 2011

References

1. Eddy SR: **Non-coding RNA genes and the modern RNA world.** *Nat Rev Genet* 2001, **2(12)**:919-929.
2. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A: **Rfam: updates to the RNA families database.** *Nucleic Acids Res* 2009, **37(Database issue)**:D136-40.
3. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.** *Nat Biotechnol* 2010, **28(5)**:503-510.
4. Rederstorff M, Bernhart SH, Tanzer A, Zywicki M, Perfler K, Lukasser M, Hofacker IL, Hüttenhofer A: **RNPomics: defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles.** *Nucleic Acids Res* 2010, **38(10)**.
5. Shi Y, Tyson GW, DeLong EF: **Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column.** *Nature* 2009, **459(7244)**:266-269.
6. Weinberg Z, Perreault J, Meyer MM, Breaker RR: **Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis.** *Nature* 2009, **462(7273)**:656-659.
7. Sankoff D: **Simultaneous solution of the RNA folding, alignment, and proto-sequence problems.** *SIAM J Appl Math* 1985, **45(5)**:810-25.
8. 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(4)**:e65.
9. Torarinsson E, Havgaard JH, Gorodkin J: **Multiple structural alignment and clustering of RNA sequences.** *Bioinformatics* 2007, **23(8)**:926-932.
10. Sato K, Mituyama T, Asai K, Sakakibara Y: **Directed acyclic graph kernels for structural RNA analysis.** *BMC Bioinformatics* 2008, **9**:318.
11. Havgaard JH, Torarinsson E, Gorodkin J: **Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix.** *PLoS Comput Biol* 2007, **3(10)**:1896-1908.
12. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR: **Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes.** *Genome Biol* 2010, **11(3)**:R31.
13. Wilm A, Mainz I, Steger G: **An enhanced RNA alignment benchmark for sequence alignment programs.** *Algorithms Mol Biol* 2006, **1**:19.
14. Smith T, Waterman M: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195-7.
15. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29(6-7)**:1105-19.
16. Saigo H, Vert JP, Ueda N, Akutsu T: **Protein homology detection using string alignment kernels.** *Bioinformatics* 2004, **20(11)**:1682-9.
17. Klein RJ, Eddy SR: **RSEARCH: finding homologs of single structured RNA sequences.** *BMC Bioinformatics* 2003, **4**:44.
18. Morita K, Saito Y, Sato K, Oka K, Hotta K, Sakakibara Y: **Genome-wide searching with base-pairing kernel functions for noncoding RNAs: computational and expression analysis of snoRNA families in Caenorhabditis elegans.** *Nucleic Acids Res* 2009, **37(3)**:999-1009.
19. 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(13)**:1593-9.
20. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22)**:4673-80.

doi:10.1186/1471-2105-12-S1-S48
Cite this article as: Saito *et al.*: Fast and accurate clustering of noncoding RNAs using ensembles of sequence alignments and secondary structures. *BMC Bioinformatics* 2011 **12**(Suppl 1):S48.