

Letter

CentroidAlign-Web: A Fast and Accurate Multiple Aligner for Long Non-Coding RNAs

Haruka Yonemoto ¹, Kiyoshi Asai ^{1,2} and Michiaki Hamada ^{1,2,*}

¹ Department of Computational Biology, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8561, Japan; E-Mails: yonemoto_haruka@cb.k.u-tokyo.ac.jp (H.Y.); asai@k.u-tokyo.ac.jp (K.A.)

² Computational Biology Research Center (CBRC), the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo Waterfront Bio-IT Research Building, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

* Author to whom correspondence should be addressed; E-Mail: mhamada@k.u-tokyo.ac.jp; Tel.: +81-3-3599-8783; Fax: +81-3-3599-8777.

Received: 20 November 2012; in revised form: 28 January 2013 / Accepted: 28 February 2013 / Published: 18 March 2013

Abstract: Due to the recent discovery of non-coding RNAs (ncRNAs), multiple sequence alignment (MSA) of those long RNA sequences is becoming increasingly important for classifying and determining the functional motifs in RNAs. However, not only primary (nucleotide) sequences, but also secondary structures of ncRNAs are closely related to their function and are conserved evolutionarily. Hence, information about secondary structures should be considered in the sequence alignment of ncRNAs. Yet, in general, a huge computational time is required in order to compute MSAs, taking secondary structure information into account. In this paper, we describe a fast and accurate web server, called CentroidAlign-Web, which can handle long RNA sequences. The web server also appropriately incorporates information about known secondary structures into MSAs. Computational experiments indicate that our web server is fast and accurate enough to handle long RNA sequences. CentroidAlign-Web is freely available from <http://centroidalign.ncrna.org/>.

Keywords: ncRNAs; rRNAs; multiple sequence alignment (MSA); structural alignment; secondary structure; consensus structures; web server

1. Introduction

Various non-coding RNAs (ncRNAs), especially long non-coding RNAs (lncRNAs/lincRNAs) [1], are emerging as new players in molecular biology, demonstrating potential roles in the mechanism of diseases, such as cancers [2]. In the ENCODE project, the number of ncRNAs, including lncRNAs, reported is more than 6,000 [3], and this is one of the most important research themes in the project. When analyzing the evolution and functions of ncRNAs, multiple sequence alignment (MSA) is an important first step. It is known that the secondary structures of many ncRNAs are strongly related to their functions, and so, not only primary (nucleotide) sequences, but also secondary structures of ncRNAs are evolutionarily conserved [2]; Hence, it is important to consider secondary structure explicitly when aligning RNA sequences. However, the computational cost of aligning RNA sequences, while considering secondary structures, is huge: the computational cost of the aligning of two RNA sequences is $O(L^6)$, where L is the length of RNA sequences (see [4]).

Currently, there are several web servers that can be used for aligning multiple RNA sequences and that consider secondary structures: PicXAA-Web [5]; R-coffee [6]; LocARNA [7]; FoldAlign (for aligning two sequences) [8]; StrAl Webservice [9]; MAFFT [10]; Dynalign [11], and so forth. However, due to the high computational demands of aligning RNA sequences while considering secondary structures, most existing web servers cannot handle long RNA sequences (e.g., rRNAs [12] or lincRNAs [1,13]).

We have developed a novel web server (called “CentroidAlign-Web”) for aligning multiple RNA sequences by extending CentroidAlign [14], which is a fast and accurate multiple aligner for RNA sequences that considers secondary structures. The features of CentroidAlign-Web are summarized as follows:

- CentroidAlign-Web can accept long RNA sequences, such as rRNAs. In order to handle those RNA sequences, we have reduced the time complexity of CentroidAlign by integrating the Rfold algorithm [15] into it (see the next section for details).
- Users can (optionally) give the secondary structure(s) of input sequences, if this information is available. For example, secondary structures of long RNA sequences from HIV-1 [16], HCV (hepatitis C virus) [17] and lincRNA (the steroid receptor RNA activator (SRA)) [18] have been recently determined by combining experimental techniques with computational approaches. This secondary structure information is useful for estimating multiple alignments.
- CentroidAlign-Web has an interface in which users can specify a region of the human genome (hg18) from which to extract a multiple alignment, and re-align that region using CentroidAlign. Because recent studies have suggested that re-alignment of genome sequence alignments reveals new non-coding RNAs [19], this function will be useful.

Computational experiments conducted in this study indicate that our web server is fast enough to compute a multiple alignment for long RNA sequences, and known secondary structure information can improve multiple alignments of RNA sequences. CentroidAlign-Web is freely available from <http://centroidalign.ncrna.org/>, and will be useful for research on non-coding RNAs.

2. Materials and Methods

2.1. CentroidAlign

CentroidAlign [14] is a fast and accurate aligner for multiple RNA sequences. In contrast to usual MSA tools for DNA/protein sequences (e.g., ClustalW [20] or ProbCons [21]), CentroidAlign can consider (common) secondary structures among input RNA sequences when aligning RNA sequences (cf. Figure A1). Because secondary structures of RNAs are often conserved in their evolution, it is important to consider secondary structures in multiple alignments of RNA sequences. However, considering a common secondary structure in a multiple alignment (this kind of alignment is often called “structural” alignment) entails a huge computational cost (cf. [4]). CentroidAlign reduces the computational costs by several heuristic techniques, factorizing a probability distribution of structural alignments (given by, e.g., the Sankoff model [4]) into (i) a probability distribution of secondary structures (given by, e.g., the McCaskill model [22]) and (ii) a probability distribution of (usual) alignments (given by, e.g., the ProbCons model [21]) (b-2 in Figure A1). This approximation leads to an algorithm based on a base-pairing probability matrix (BPPM) for each RNA sequence (a BPPM gives the (marginal) probability of every base-pair with respect to a probability distribution of secondary structures) and an aligned-base probability matrix (ABPM) for every pair of RNA sequences (an ABPM gives the (marginal) probability of every aligned base-pair with respect to a probability distribution of alignments). Both matrices include information about the ambiguity of secondary structures and alignments. The result is that the time complexity of the pairwise alignment step in CentroidAlign is $O(L^3 + c^2dL^2) \approx O(L^3)$, where L is the length of input sequences and both c and d are constants independent of L .

Moreover, we have integrated the probabilistic consistency transformation (PCT) of the alignment probability matrix [21] into the proposed estimator. Finally, the extension to multiple alignment is conducted by a progressive alignment algorithm similar to CONTRAlign [23].

Note that CentroidAlign employs an estimator based on maximum expected accuracy (MEA), which has been successfully applied in much software in the field of bioinformatics; see the review by Hamada and Asai [24] for details. In CentroidAlign, the sum-of-pair scores (SPS) [25] is optimized for predicting multiple alignments of RNA sequences (cf. c and d in Figure A1).

2.2. Rfold

The Rfold algorithm, which was proposed in [15], computes a BPPM for a given RNA sequence. In the computation of the BPPM, Rfold can use the maximum distance of base-pairs in a predicted secondary structure, which enables it to handle longer RNA sequences. The time complexity of Rfold is $O(w^2L)$, where w is the maximum size (span) of base-pairs, while the time complexity of algorithms that compute a full BPPM (such as the McCaskill algorithm [22]) is $O(L^3)$, where L is the length of the RNA sequence.

2.3. Dataset Utilized in Computational Experiments

Table 1 shows a summary of the dataset used in this study. RNA families whose length is more than 800 are taken from seed alignments in the Rfam 11.0 database (August 2012) [2]. Note that those seed alignments give high-quality benchmark datasets, because they are manually curated MSAs, which take into consideration (consensus) secondary structures.

Table 1. Datasets used in this study. Each family is taken from the Rfam 11.0 database [2]. “Num”, “Average length (nt)” and “MPI” mean the number of sequences in the family, the average length of sequences and the mean pairwise identity of sequences in each family, respectively.

Dataset name	Accession	Num	Average length(nt)	MPI(%)	Description
SSU_rRNA_eukarya	RF01960	84	1791.20	80.00	Eukaryotic small subunit ribosomal RNA
SSU_rRNA_bacteria	RF00177	93	1524.50	80.00	Bacterial small subunit ribosomal RNA
SSU_rRNA_archaea	RF01959	19	1480.50	81.00	Archaeal small subunit ribosomal RNA
Sacc_telomerase	RF01050	13	1189.50	70.00	Saccharomyces telomerase
snR86	RF01272	5	998.40	69.00	Small nucleolar RNA snR86
RUF21	RF01825	5	691.80	65.00	RNA of unknown function 21

3. Results and Discussion

3.1. CentroidAlign Web Application (CentroidAlign-Web)

Usage of the server is quite simple. Users can paste sequences in FASTA format (<http://www.ebi.ac.uk/help/formats.html#fasta>) into a text area or upload a FASTA file, then click on the “submit” button (Figure 1). The server responds with a multiple alignment (Figure 2)(See Table 2 and Figure 3 for computational time of our web server). The resulting format is multiple alignment format (MAF) or clustalW. By expanding “Options” in the interface, users can adjust several internal parameters of the web server (see Table 3 for the detailed parameters). There are three major advantages in this web server. (1) The maximum distance between the two bases of a base-pair can be specified (by users) in order to reduce computational cost for computing BPPMs (which is the most time-consuming part of CentroidAlign) (see Section 3.1.1.). This option ensures that the alignment finishes in a practical amount of time, even if users’ query sequences are relatively long (e.g., rRNAs); (2) Users can utilize secondary structural information for alignment. An example of the required format is given on the help page (<http://centroidalign.ncrna.org/help.html>). When the structures of users’ query sequences are experimentally determined, the probabilities of positions at which bases make a pair should be 1 and, otherwise, 0 (cf. Section 3.1.2.). Using actual (not predicted) probabilities should enable more accurate alignment of structured RNAs. (3) Users can extract an MAF region (from the hg18 17way MULTIZ alignment) by specifying chromosome, start position, end position and strand. The sequences in the multiple alignment are realigned by CentroidAlign.

Figure 1. Input page of CentroidAlign-Web, in which RNA sequences are given in the FASTA format. Additionally, by using the interface: (i) users can give secondary structures for parts of input RNA sequences; (ii) users can specify a region of the human genome (hg18); (iii) users can utilize the Rfold algorithm to compute base-pairing probability matrices (BPPMs), with a user-given maximum distance for base-pairs.

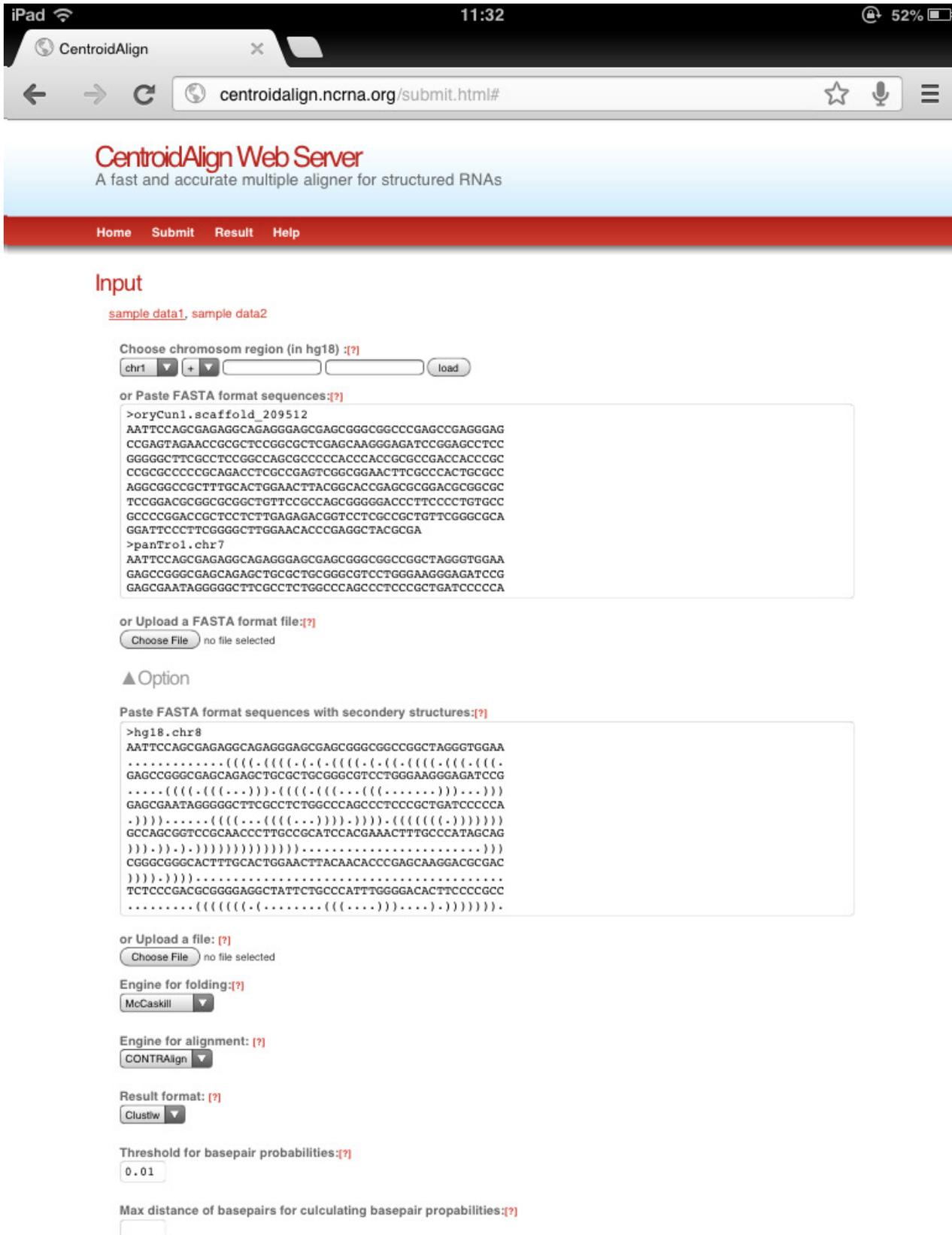
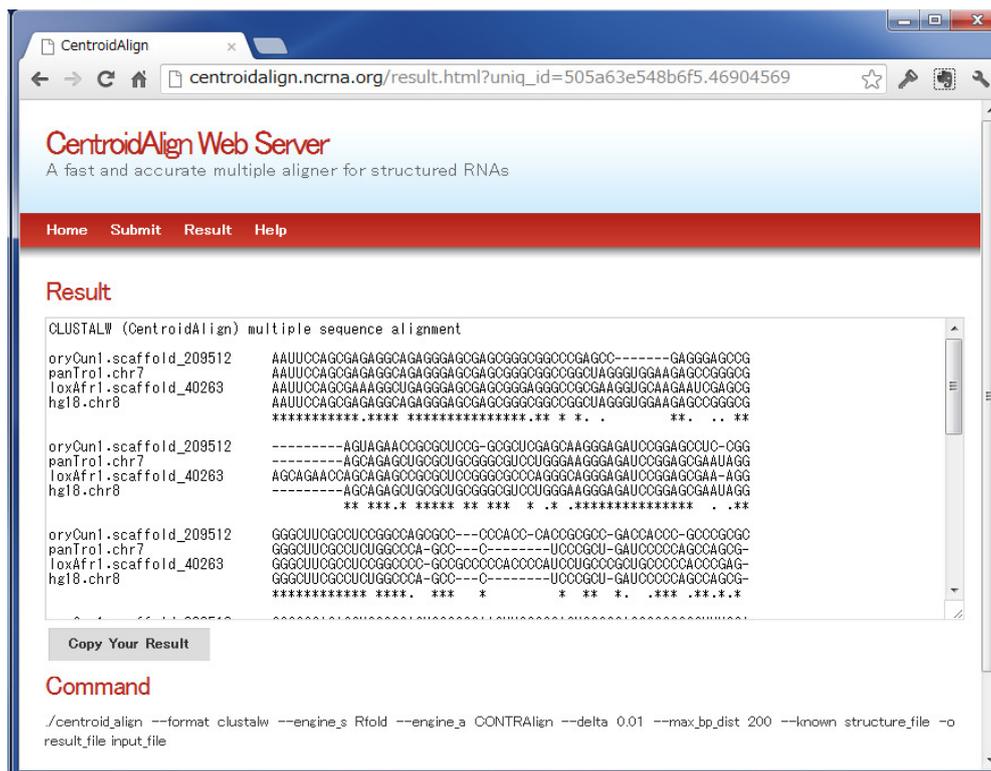


Figure 2. Results page of CentroidAlign-Web. The output of an multiple sequence alignment (MSA) is in either ClustalW format or multi-FASTA format. The complete command line information is also provided on the results page. Users can copy the result to the clipboard for use in the next analysis: e.g., common secondary structure prediction using CentroidAlifold [26] (<http://www.ncrna.org/centroidfold>) with the predicted multiple alignment from CentroidAlign-Web.



The screenshot shows a web browser window with the URL `centroidalign.ncrna.org/result.html?uniq_id=505a63e548b6f5.46904569`. The page title is "CentroidAlign Web Server" and the subtitle is "A fast and accurate multiple aligner for structured RNAs". The navigation menu includes "Home", "Submit", "Result", and "Help". The main content area is titled "Result" and displays a ClustalW multiple sequence alignment for four sequences: `oryCun1.scaffold_209512`, `panTrol.chr7`, `loxAfr1.scaffold_40263`, and `hg18.chr8`. The alignment shows conserved regions with asterisks indicating high similarity. Below the alignment is a "Copy Your Result" button. At the bottom, the "Command" section shows the command line used for the job: `/centroid_align --format clustalw --engine_s Rfold --engine_a CONTRAlign --delta 0.01 --max_bp_dist 200 --known structure_file -o result_file input_file`.

When submitting the job, the user is given a “Job ID” and a link to the results (multiple FASTA format or ClustalW format). Users can retrieve the results by using the Job ID at a later time. Then, users can copy the result to the clipboard and use it in the next analysis, for example, in common secondary structure prediction of the multiple alignment, using CentroidAlifold [26] <http://www.ncrna.org/centroidfold>. Additionally, a complete set of command line options can be obtained, which is useful for users of the command line version of CentroidAlign.

3.1.1. Incorporating the Rfold Algorithm into the Web Server

In CentroidAlign-Web, we incorporated the Rfold algorithm (cf. Section 2.2) to compute the BPPM for each RNA sequence in the input sequences. As a result, the total computational cost of CentroidAlign is reduced to $O(w^2L + c^2dL^2) \approx O(L^2)$, where w is the maximum length of base-pairs, L is the length of input sequences and both c and d are constants independent of L (Note that the computational cost of the original CentroidAlifold is $\approx O(L^3)$; see Section 2.1). This reduction of computational cost enables the prediction of MSAs for longer RNA sequences (e.g., ribosomal RNA sequences or lincRNAs), taking into account information about secondary structures.

Table 2. Computational results for various ratios of known secondary structures: **(a)** no secondary structure information is given; **(b)** (resp. **(c)** and **(d)**) 25% (resp. 50% and 75%) of secondary structures for input sequences are given. The “SPS” columns show the sum-of-pairs-score of a predicted multiple alignment [25]. In CentroidAlign-Web, we utilized the Rfold model [15] (where the maximum size (span) of base-pairs is set to 300) for a model of RNA secondary structures; we utilized the CONTRAlign model [23] for a model of pairwise alignments (see Figure A1). In PicXAA-R, we conducted a standalone version of PicXAA-R (version 1.0) and the default parameters were utilized. A Linux OS machine with a 3.33 GHz Intel(R) Xeon(R) CPU W5590 processor and 32 GByte of memory was used in this experiment. See Table 1 for detailed information about the datasets used.

ID	Num	Average length (nt)	CentroidAlign								PicXAA-R	
			(a) 0%		(b) 25%		(c) 50%		(d) 75%		SPS	Time(s)
			SPS	Time(s)	SPS	Time(s)	SPS	Time(s)	SPS	Time(s)	SPS	Time(s)
RF01960	84	1791.2	0.9173	6546.08	0.9164	5924.78	0.9179	5317.99	0.9208	4687.98	0.9121	8945.82
RF00177	93	1524.5	0.9560	5874.18	0.9576	5314.39	0.9589	4782.40	0.9572	4228.57	0.9548	7544.52
RF01959	19	1480.5	0.9800	569.66	0.9816	474.21	0.9817	364.89	0.9821	251.17	0.9786	508.96
RF01050	13	1189.5	0.8848	273.87	0.8840	219.46	0.8861	165.06	0.8926	111.81	0.8764	155.42
RF01272	5	998.4	0.8975	80.06	0.9049	64.75	0.9152	48.97	0.9286	34.31	0.8913	35.58
RF01825	5	691.8	0.8319	44.80	0.8467	36.38	0.8261	27.73	0.8572	18.98	0.8236	12.53

Figure 3. Computational time of CentroidAlign and PicXAA-R for RNA sequences with various length. In this experiment, five random RNA sequences were utilized for each dataset. The black and red lines correspond to CentroidAlign and PicXAA(-R), respectively. PicXAA did not finish within three days for the length of 20,000 nt.

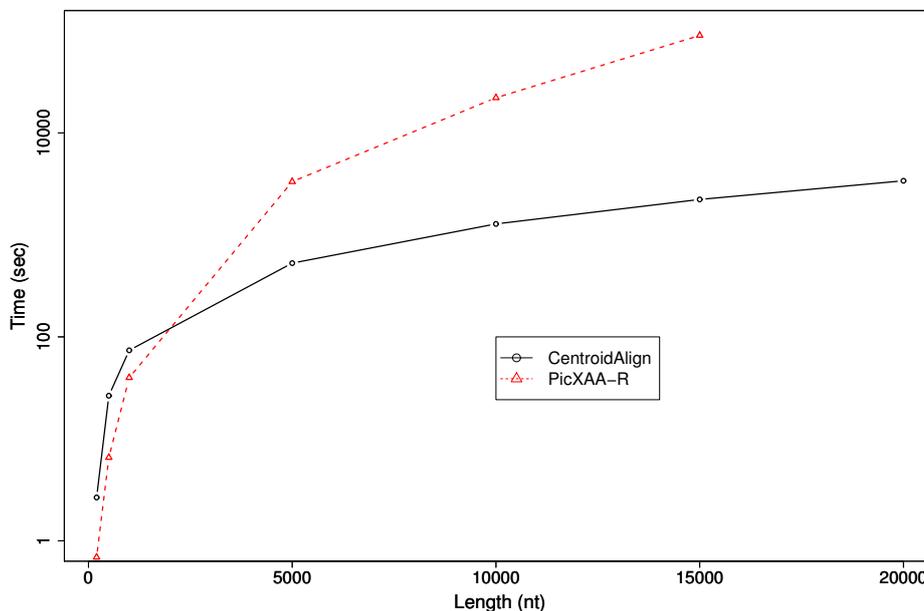


Table 3. Adjustable parameters in CentroidAlign-Web. Each parameter can be altered in the “Options” control.

Parameter name	Description	Possible	Default
Engine for folding	Probabilistic model of secondary structures	McCaskill, CONTRAfold, Rfold ¹	Rfold
Engine for alignment	Probabilistic model of pairwise sequence alignments	CONTRAlign, ProbCons ²	CONTRAlign
Result format	Output format	ClustalW, MFA	ClustalW
Threshold for base-pair probabilities	Threshold for base-pairing probabilities	0 to 1	0.01
Max distance of base-pairs	The maximum distance of base-pairs	More than 0	300

¹ CONTRAfold and McCaskill are probability distributions of secondary structures of RNA sequences proposed in [22,27], respectively; ² CONTRAlign and ProbCons are probability distributions of pairwise alignments proposed in [21,23], respectively; ³ If the length of RNA sequences is long, users should specify this value in order to reduce the computational cost.

3.1.2. BPPM for an RNA Sequence with a Secondary Structure

For an RNA sequence, x , with a (known) secondary structure, y , a BPPM for the sequence is given by:

$$p_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ form a base-pair in a given structure} \\ 0 & \text{otherwise} \end{cases}$$

instead of utilizing the BPPM calculated, e.g., by the McCaskill algorithm [22]. In this way, we can seamlessly incorporate information about secondary structures into computing multiple alignments in CentroidAlign.

3.2. Computational Experiments

In our computational experiments, we focused on relatively long RNA sequences in order to show that our web server can handle longer sequences than existing software. (Note that there exists several studies that extensively evaluate among multiple aligners (including CentroidAlign) for short (~ 500 nt) RNA sequences with low sequence similarities [14,28], indicating that CentroidAlign achieved good performance to those datasets.) We have tested six RNA families (from the Rfam 11.0 database [2]), whose average length is relatively long (from 700 to 1,800 bases). The largest dataset contains 84 RNA sequences, whose average length is around 1,800. See Table 1 for the details of the datasets. We conducted our computational experiments on a Linux machine with a 3.33 GHz Intel(R) Xeon(R) CPU W5590 processor and 32 GByte of memory. Note that the current version of CentroidAlign-Web is also implemented on a machine with the same specification.

The results are shown in column (a) in Table 2. In the experiments, Rfold with the maximum size of base-pairs set to be 300 was employed for the probability distribution of secondary structures (in other words, Rfold was employed to calculate the base-pairing probability matrix of secondary structures, setting the maximum distance between base-pairs to 300 nt); CONTRAlign was utilized for the probability distribution of pairwise alignments. The computational time for the largest dataset (RF01960, which contains 84 sequences with an average length of 1,791) is a few hours; for a moderately sized dataset (RF01959, which contains 19 sequences with an average length of 1,190), the computational time is less than 300 s. We compared CentroidAlign with PicXAA(-R) (version 1) [5,28], which is one of the fastest multiple aligners for RNA sequences, wherein the information of secondary structures is taken into account. For a larger dataset (e.g., RF01960 and RF00177), CentroidAlign was faster than PicXAA, and SPSs of CentroidAlign were consistently better than those of PicXAA among all dataset (Table 2). In addition, in Figure 3, we show the computational time of multiple sequence alignment for five (random) sequences with various lengths up to 20,000. This result indicated that, for longer RNA sequences, CentroidAlign is much faster than PicXAA (e.g., CentroidAlign took 40 minutes for five sequences of 15,000 nt, while PicXAA took more than 1 day), which is one of the advantages of our web sever.

Finally, in order to examine whether information about secondary structures improves the accuracy of MSAs of RNA sequences, we conducted computational experiments using known secondary structures. The ratio of known secondary structures in the input sequences was 25%, 50% or 75% (corresponding

to columns (b), (c) and (d), respectively, in Table 2). The secondary structures are given by mapping consensus secondary structures (in seed alignments) to the RNA sequence. Table 2 shows that the information about secondary structures (slightly) improved the accuracy of multiple alignments, which indicates the usefulness of known secondary structures in MSAs. The use of secondary structure information seems to have more impact on datasets RF01272 and RF01825, which correspond to RNA families with lower primary sequence identity (see MPI values in Table 1), compared to datasets containing sequences with higher (>70%) identity (Table 2), indicating that the importance of secondary structures in RNA families with low sequence conservation.

3.3. Future Work

We are planning to incorporate biochemical experimental information (such as SHAPE) into the web server, because such information can be used to determine secondary structure [29] by employing a recently developed method that enables the updating of the BPPM according to experimental information [30].

Recent studies have clearly indicated the importance of lincRNAs [13]. Not only lincRNAs are longer than conventional non-coding RNAs (such as snoRNAs and miRNAs), but also most lincRNAs exhibit low sequence similarity. We therefore plan to apply our Web Server to the detailed analysis of lincRNAs (such as SRA [18] and HOTAIR [31]), which might lead to important biological findings.

4. Conclusions

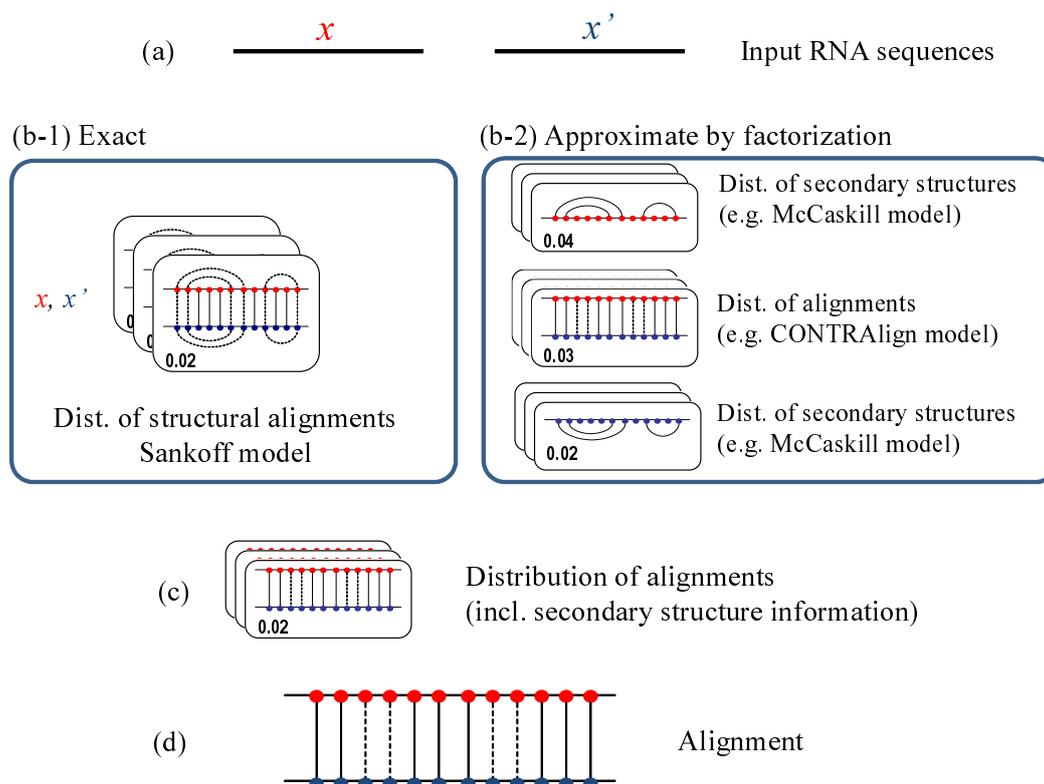
In this paper, we have introduced CentroidAlign-Web, a web server for predicting multiple alignments of long RNA sequences. We showed that the web server is capable of dealing with long RNA sequences, such as rRNAs, and that information about secondary structures can be used to improve the accuracy of multiple alignments. CentroidAlign-Web is freely available from <http://centroidalign.ncrna.org/>, which would be useful to researches of non-coding RNAs.

Acknowledgments

This work was supported in part by MEXT KAKENHI (Grant-in-Aid for Young Scientists (A): 24680031 to MH; Grant-in-Aid for Scientific Research (A): 30356357 to KA).

Appendix

Figure A1. Overview of the pairwise alignment step in the CentroidAlign algorithm [14]. **(a)** The input is two RNA sequences, (x, x') , to be aligned; **(b-1)** The exact algorithm of CentroidAlign considers a probability distribution of structural alignments between x and x' , which gives simultaneously the alignments between nucleotides and those between base-pairs (e.g., Sankoff model [4]); **(b-2)** The exact case can be approximated by factorizing the distribution of structural alignments into (i) a distribution of secondary structures of x (e.g., the CONTRAfold [27] or McCaskill [22] models); (ii) a distribution of pairwise alignments between x and x' (e.g., the CONTRAlign model [23]); and (iii) a distribution of secondary structures of x' ; **(c)** By marginalization of the distribution(s) in (b), we obtain a distribution of alignments (*) in which the information about secondary structures is included; **(d)** The best multiple alignment is estimated based on maximizing expected accuracy (MEA) [24] in which the SPS scores of predicted alignments are optimized with respect to the distribution (*) of pairwise alignments given in (c). It should be emphasized that the computational cost of the exact algorithm is $\approx O(L^6)$, while it is reduced to $\approx O(L^3)$ in the approximate algorithm, where L is the (maximum) length of two input sequences.



References

1. Volders, P.J.; Helsen, K.; Wang, X.; Menten, B.; Martens, L.; Gevaert, K.; Vandesompele, J.; Mestdagh, P. LNCipedia: A database for annotated human lncRNA transcript sequences and structures. *Nucl. Acids Res.* **2013**, *41*, D246–251.

2. Gardner, P.P.; Daub, J.; Tate, J.; Moore, B.L.; Osuch, I.H.; Griffiths-Jones, S.; Finn, R.D.; Nawrocki, E.P.; Kolbe, D.L.; Eddy, S.R.; *et al.* Rfam: Wikipedia, clans and the “decimal” release. *Nucl. Acids Res.* **2011**, *39*, D141–D145.
3. ENCODE Project Consortium, Dunham, I.; Kundaje, A.; Aldred, S.F.; Collins, P.J.; Davis, C.A.; Doyle, F.; Epstein, C.B.; Frietze, S.; Harrow, J.; *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.
4. Sankoff, D. Simultaneous solution of the RNA folding alignment and protosequence problems. *SIAM J. Appl. Math* **1985**, *45*, 810–825.
5. Sahraeian, S.M.; Yoon, B.J. PicXAA-Web: A web-based platform for non-progressive maximum expected accuracy alignment of multiple biological sequences. *Nucl. Acids Res.* **2011**, *39*, 8–12.
6. Moretti, S.; Wilm, A.; Higgins, D.G.; Xenarios, I.; Notredame, C. R-Coffee: A web server for accurately aligning noncoding RNA sequences. *Nucl. Acids Res.* **2008**, *36*, W10–W13.
7. Smith, C.; Heyne, S.; Richter, A.S.; Will, S.; Backofen, R. Freiburg RNA Tools: A web server integrating INTARNA, EXPARNA and LOCARNA. *Nucl. Acids Res.* **2010**, *38*, W373–W377.
8. Havgaard, J.H.; Lyngso, R.B.; Gorodkin, J. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucl. Acids Res.* **2005**, *33*, W650–W653.
9. 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.
10. Katoh, K.; Toh, H. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinforma.* **2008**, *9*, 212.
11. Harmanci, A.O.; Sharma, G.; Mathews, D.H. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinform.* **2007**, *8*, 130.
12. Cole, J.R.; Wang, Q.; Cardenas, E.; Fish, J.; Chai, B.; Farris, R.J.; Kulam-Syed-Mohideen, A.S.; McGarrell, D.M.; Marsh, T.; Garrity, G.M.; Tiedje, J.M. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucl. Acids Res.* **2009**, *37*, D141–D145.
13. Bu, D.; Yu, K.; Sun, S.; Xie, C.; Skogerb, G.; Miao, R.; Xiao, H.; Liao, Q.; Luo, H.; Zhao, G.; *et al.* NONCODE v3.0: Integrative annotation of long noncoding RNAs. *Nucl. Acids Res.* **2012**, *40*, D210–D215.
14. 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.
15. Kiryu, H.; Kin, T.; Asai, K. Rfold: An exact algorithm for computing local base pairing probabilities. *Bioinformatics* **2008**, *24*, 367–373.
16. Watts, J.M.; Dang, K.K.; Gorelick, R.J.; Leonard, C.W.; Bess, J.W.; Swanstrom, R.; Burch, C.L.; Weeks, K.M. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **2009**, *460*, 711–716.
17. Pang, P.S.; Elazar, M.; Pham, E.A.; Glenn, J.S. Simplified RNA secondary structure mapping by automation of SHAPE data analysis. *Nucl. Acids Res.* **2011**, *39*, e151.
18. Novikova, I.V.; Hennelly, S.P.; Sanbonmatsu, K.Y. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucl. Acids Res* **2012**, *40*, 5034–5051.

19. Will, S.; Yu, M.; Berger, B. Structure-Based Whole Genome Realignment Reveals Many Novel Non-coding RNAs. In *RECOMB*; Chor, B., Ed.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7262, p. 341.
20. Thompson, J.D.; Gibson, T.J.; Higgins, D.G. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinform.* **2002**, doi:10.1002/0471250953.bi0203s00.
21. Do, C.B.; Mahabhashyam, M.S.; Brudno, M.; Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **2005**, *15*, 330–340.
22. McCaskill, J.S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **1990**, *29*, 1105–1119.
23. Do, C.; Gross, S.; Batzoglou, S. CONTRAlign: Discriminative Training for Protein Sequence Alignment. In *RECOMB*; Apostolico, A., Guerra, C., Istrail, S., Pevzner, P.A., Waterman, M.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3909, pp. 160–174.
24. Hamada, M.; Asai, K. A classification of bioinformatics algorithms from the viewpoint of maximizing expected accuracy (MEA). *J. Comput. Biol.* **2012**, *19*, 532–549.
25. Thompson, J.D.; Plewniak, F.; Poch, O. A comprehensive comparison of multiple sequence alignment programs. *Nucl. Acids Res.* **1999**, *27*, 2682–2690.
26. Hamada, M.; Sato, K.; Asai, K. Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucl. Acids Res.* **2011**, *39*, 393–402.
27. Do, C.B.; Woods, D.A.; Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**, *22*, e90–e98.
28. Sahraeian, S.M.; Yoon, B.J. PicXAA: Greedy probabilistic construction of maximum expected accuracy alignment of multiple sequences. *Nucl. Acids Res.* **2010**, *38*, 4917–4928.
29. Wan, Y.; Kertesz, M.; Spitale, R.C.; Segal, E.; Chang, H.Y. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.* **2011**, *12*, 641–655.
30. Hamada, M. Direct updating of an RNA base-pairing probability matrix with marginal probability constraints. *J. Comput. Biol.* **2012**, *19*, 1265–1276.
31. He, S.; Liu, S.; Zhu, H. The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol. Biol.* **2011**, *11*, 102.