W100–W106 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 11 May 2011*
doi:10.1093/nar/gkr290

CentroidHomfold-LAST: accurate prediction of RNA secondary structure using automatically collected homologous sequences

Michiaki Hamada[1,2,*], Koichiro Yamada[3], Kengo Sato[1], Martin C. Frith[2] and Kiyoshi Asai[1,2]

[1]Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064 and [3]Information and Mathematical Science Laboratory, Inc, 1-5-21, Ohtsuka, Bunkyo-ku, Tokyo 112-0012, Japan

Received February 6, 2011; Revised March 28, 2011; Accepted April 12, 2011

ABSTRACT

Although secondary structure predictions of an individual RNA sequence have been widely used in a number of sequence analyses of RNAs, accuracy is still limited. Recently, we proposed a method (called 'CentroidHomfold'), which includes information about homologous sequences into the prediction of the secondary structure of the target sequence, and showed that it substantially improved the performance of secondary structure predictions. CentroidHomfold, however, forces users to prepare homologous sequences of the target sequence. We have developed a Web application (CentroidHomfold-LAST) that predicts the secondary structure of the target sequence using automatically collected homologous sequences. LAST, which is a fast and sensitive local aligner, and CentroidHomfold are employed in the Web application. Computational experiments with a commonly-used data set indicated that CentroidHomfold-LAST substantially outperformed conventional secondary structure predictions including CentroidFold and RNAfold.

INTRODUCTION

Secondary structure prediction of a single RNA sequence is a classical and fundamental problem in bioinformatics which has been widely used in research. The importance of accurate predictions of RNA secondary structure has increased because of the recent discoveries of functional non-coding RNAs and the relation between secondary structures and function (1).

There are a number of studies of RNA secondary structure prediction (2–5). The most popular method is to predict the minimum free energy (MFE) structure, and this was used in Mfold (5), RNAfold (6) and RNAstructure (7). Another successful approach is to maximize the expected accuracy (MEA). In this approach, the entire distribution of possible structures is considered [see also (8) for the usefulness of MEA]. The MEA-based estimator used in CONTRAfold (2) and the γ-centroid estimator used in CentroidFold (3) use this method. Although several state-of-the-art algorithms have been proposed, the accuracy of secondary structure prediction seems to have peaked, and it is important to seek ways to improve secondary structure predictions.

In many cases, homologous sequences of the target RNA are available and using those sequences can improve the accuracy of secondary structure predictions. For example, it is well-known that common secondary structures improve the accuracy (9–11). However, what we would like to estimate is not a common secondary structure but a secondary structure for the specific target RNA sequence. Motivated by this idea, Hamada *et al*. (12) have proposed a new algorithm, called CentroidHomfold, which predicts a secondary structure of the target sequence by considering their homologous sequences. Although CentroidHomfold achieves much better accuracy than conventional secondary structure predictions, a drawback is that homologous sequences of the target RNA sequence must be prepared (which seems to be a hard task for most users).

We have, therefore, developed a Web application that predicts the secondary structure of a specific target sequence by employing 'automatically' collected homologous sequences. (The architecture is shown in Figure 1). To collect homologous sequences of the target sequence

*To whom correspondence should be addressed. Tel: +81-3-5281-5271; Fax: +81-3-5281-5331; Email: mhamada@k.u-tokyo.ac.jp

automatically, we used LAST, which is a sensitive and fast aligner (13,14), and a database of known RNA sequences derived from Rfam (15). Computational experiments in which a commonly-used data set was used (S151-Rfam data set; (2)) indicated that CentroidHomfold–LAST substantially outperformed conventional secondary structure predictors, such as CentroidFold (3), RNAfold (16), Simfold (17) and Sfold (18). A standalone pipeline is also available from our Web site.

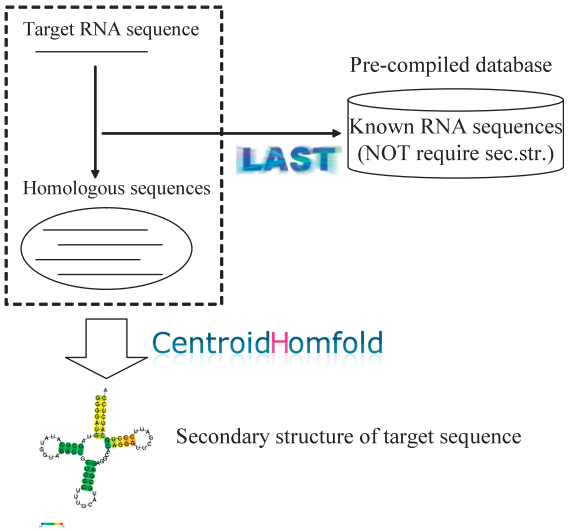

Figure 1. Architecture of the CentroidHomfold Web application (a.k.a. CentroidHomfold-LAST). The input is an RNA sequence and the output is a secondary structure of the input sequence. The Web application uses a pre-compiled database of known RNA sequences in order to find homologous sequences of the input sequence.

MATERIALS AND METHODS

CentroidHomfold

CentroidHomfold, which was originally proposed by Hamada *et al.* (12), takes a target RNA sequence and its homologous sequences as the input. Then, it estimates a secondary structure of the target sequence using the information in the homologous sequences.

Basically, CentroidHomfold adopts the γ-centroid estimator (3), which is based on the MEA with respect to base pairs. As a probability distribution of secondary structures for the target sequence, CentroidHomfold employs a marginalized and averaged distribution of probability distributions of structural alignments (cf. the Sankoff model (19)) between the target sequence and each of the homologous sequences (see Supplementary Figure S1).

However, it requires huge computational effort to consider the probability distributions of the structural alignments. CentroidHomfold, therefore, factorizes the distributions into three parts: (i) the distribution of the secondary structures of the target sequence (given by the McCaskill model (20), for example), (ii) the distribution of the pairwise alignments between the target sequence and each homologous sequence (given by the CONTRAlign model (21), for example) and (iii) the distributions of the secondary structures of each homologous sequence (see Figure 2). By this factorization, the computational cost becomes much smaller than for the non-approximated method, so CentroidHomfold is

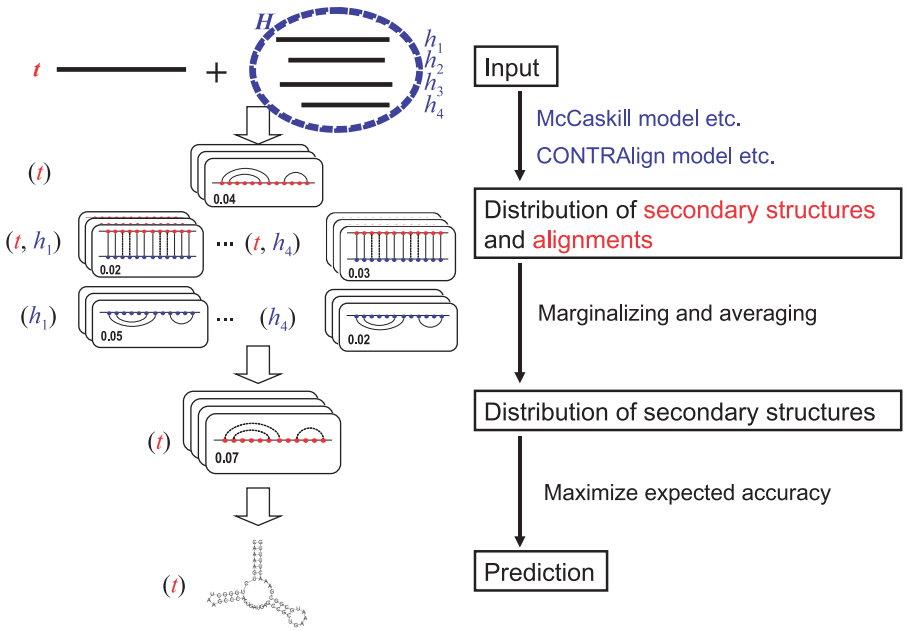

Figure 2. Illustration of the CentroidHomfold algorithm proposed in Hamada *et al.* (12), where t is a target RNA sequence and $H = \{h_1, h_2, h_3, h_4\}$ is a set of homologous sequences of t. CentroidHomfold considers all the secondary structures of the target and homologous sequences, as well as all the pairwise alignments between the target sequence and each homologous sequence. This algorithm can be viewed as an approximation of the γ-centroid estimator (3) with a marginalized and averaged distribution of distributions of structural alignments between the target sequence and each of homologous sequences. See also Supplementary Figure S1.

Table 1. Adjustable parameters in the Web application, each parameter can be altered in the 'Advanced options' control

Parameter name	Description	Possible	Default
Model (secondary structure)	Probabilistic model of secondary structures	BL, CONTRAfold, McCaskill[a]	BL
Model (alignment)	Probabilistic model of pairwise alignments	CONTRAlign, ProbCons[b]	CONTRAlign
Gamma[c]	Adjust tradeoff between Sensitivity and PPV	2^n ($n = -5, \ldots, 10$)	8
Non-canonical base-pairs	Allow non-canonical base pairs in the prediction	Yes, No	No
Database	Database of RNA sequences	Rfam.seed.X, Rfam.full.X (X = 99, 95, 90)[d]	Rfam.full.99
E-value[e]	*E*-value when searching homologous sequences	0.1,0.01,0.001,0.0001	0.01
No. of homologous[f]	Maximum number of homologous sequences	10, 20, 30, 40, 50	30

[a]BL, CONTRAfold, McCaskill are probability distributions of secondary structures of RNA sequences proposed in (21), (2) and (20), respectively.
[b]CONTRAlign and ProbCons are probability distributions of pairwise alignments proposed in (21) and (23), respectively.
[c]Equal to γ in the main text ('CentroidHomfold' section). Larger γ produces more base pairs in the prediction.
[d]Described in 'Known RNA sequences' section. See Supplementary Table S1 for the details of each database.
[e]Equal to *e* described in the main text ('CentroidHomfold–LAST Web application' section).
[f]Equal to *n* described in the main text ('CentroidHomfold–LAST Web application' section).

sufficiently practical for predicting the secondary structures of long RNA sequences. More precisely, computational cost with respect to time for CentroidHomfold scales as $O(nL^3)$, where n is the number of homologous sequences and L is the (maximum) length of those sequences (12).

In the current version of CentroidHomfold (in CentroidFold package version 0.0.9), the McCaskill model (implemented in the Vienna RNA package), the CONTRAfold model (implemented in CONTRAfold version 2.02) and the BL model (22) (the McCaskill model with Boltzmann likelihood parameters: http://www.cs.ubc.ca/labs/beta/Projects/RNA-Params/data/parameters_BLstar_Vienna.txt) can be employed as a probability distribution for secondary structures (cf. 'Model (secondary structure)' in Table 1). The ProbCons model (23) and the CONTRAlign model (21) can be used for the probability distribution of pairwise alignments (cf. 'Model (alignment)' in Table 1).

The parameter γ in CentroidHomfold ('Gamma' in Table 1) adjusts the balance between the sensitivity and the positive predictive value (PPV) of a predicted secondary structure. Larger γ produces more base pairs in the prediction.

Pseudo base-pairing probabilities in CentroidHomfold

In CentroidHomfold, the following *pseudo* base-pairing probability p_{ij} of a base pair (x_i, x_j) of a target RNA sequence x is computed:

$$p_{ij} = \alpha p_{ij}^{(s,x)} + \frac{1-\alpha}{|H|} \sum_{h \in H} \sum_{k<l} p_{ik}^{(a,x,h)} p_{jl}^{(a,x,h)} p_{kl}^{(s,h)}. \quad (1)$$

Here, $\alpha \in [0, 1]$ is a weighting parameter, $p_{ij}^{(s,x)} = p^{(s)}(\theta_{ij}^{x} = 1|x)$ is a base-pairing probability and $p_{ik}^{(a,x,h)} = p^{(a)}(\theta_{ik}^{xh} = 1|x, h)$ is an aligned base probability, where $p^{(s)}(\cdot|x)$ and $p^{(a,x,h)}(\cdot|x, h)$ denote the probability distribution of secondary structures of x and the probability distribution of pairwise alignments between x and h, respectively. It is easily seen that $p_{ij} \in [0, 1]$ (12). This (pseudo) probability is used as the reliability of a base pair in a graphical representation of an input RNA sequence in the Web application (cf. Figure 3B).

LAST

LAST (13,14) finds similar regions between sequences. LAST is sensitive and faster, because it uses variable-length (spaced) seeds realized by a suffix array. In LAST, the default settings and parameters, which are well-optimized for DNA and RNA sequences, were used (13). (We used LAST-128, downloaded from http://last.cbrc.jp, in this study.)

For a given (target) RNA sequence, a given database of RNA sequences and an *E*-value *e* (i.e. the expected number of alignments for a random database), a score threshold of significant alignments is determined using a method proposed by Sheetlin *et al*. (24). (The method is implemented in the `lastex` software in the LAST package.) It should be emphasized that this method enables us to choose a threshold for each input RNA sequence adaptively.

Known RNA sequences

To find homologous sequences of a specific RNA sequence, we utilized six databases of known RNA sequences, both of which were taken from the Rfam database (version 10.0) (15): (i) the sequences covered by Rfam full alignments, filtered to $<X$% identity (Rfam.full.X); (ii) the sequences covered by Rfam seed alignments, filtered to $<X$% identity (Rfam.seed.X). The `cd-hit-est` program (25) is employed to filter redundant sequences in both databases. Note that Rfam.full.99 includes a large number of predicted RNA sequences. See Supplementary Table S1 for a summary of the data.

Benchmark settings

In our computational experiments, we used the S-151 data set (2), which consists of 151 RNA sequences (each of which belongs to a different RNA family) with their reliable reference secondary structures. The smallest length, the largest length and average length of sequences in the data set is 23, 568 and 136.3, respectively. This data set has been used in several previous studies of RNA secondary structure prediction (2,3,17).

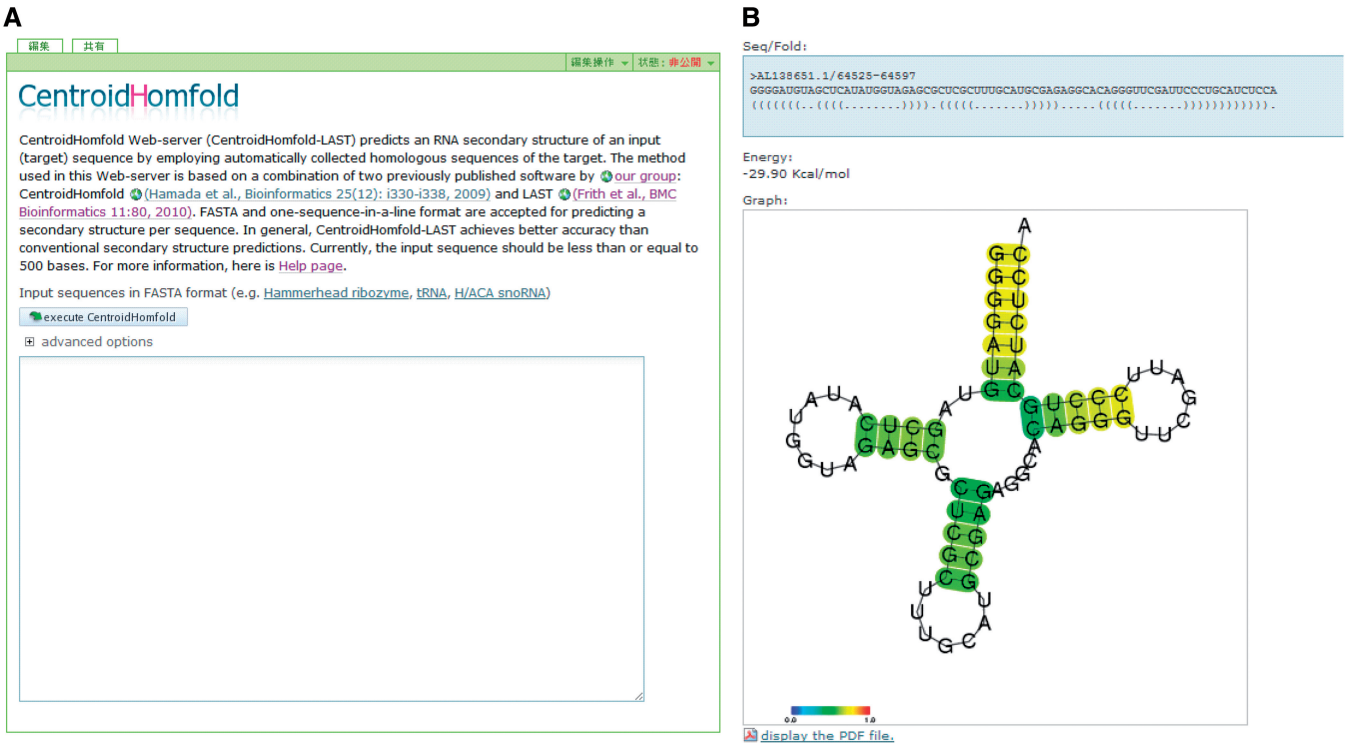

Figure 3. (**A**) A screen-shot of the Web interface of a CentroidHomfold-LAST pipeline (http://www.ncrna.org/centroidhomfold), which is similar to the CentroidFold Web application (26). Users can paste RNA sequence in the text area and push the 'Execute CentroidHomfold' button. Then, the secondary structure of the input is returned immediately (**B**). By expanding the 'Advanced options' control, the parameters in Table 1 are adjustable. (B) An example of output from the Web application. The colors of base pairs in the graphical representation indicate pseudo base-paring probabilities for each base pair (See 'Materials and Methods' section for details). Users can download the PDF file of secondary structure and homologous sequences used in CentroidHomfold.

As performance measures, we used the sensitivity (SEN) and PPV: SEN = TP/(TP + FN) and PPV = TP/(TP + FP) where TP, FP and FN are the numbers of true positive base pairs, false positive base pairs and false negative base pairs, respectively.

RESULTS AND DISCUSSION

CentroidHomfold-LAST Web application

Figure 1 shows the architecture of the CentroidHomfold Web application (CentroidHomfold–LAST pipeline) with default parameters (see the 'Default' column in Table 1). We used several databases to provide known RNA sequences (see Supplementary Table S1 and the 'Materials and Methods' section for details; users can select a database using the 'Database' option in Table 1). It should be emphasized that each database contains only information about RNA sequences and does not give any common secondary structures or secondary structures of particular RNA sequences. We prepared a compiled database of each data set, using `lastdb` with default parameters. For a given input sequence, homologous sequences with a given *E*-value were found using LAST. The *E*-value, *e*, is the expected number of chance alignments and a score threshold for significant alignment is automatically computed from *e* for each input RNA sequence (see 'Materials and Methods' section). (The parameter *e* corresponds to '*E*-value' in Table 1.) When the number of homologous sequences collected by LAST is large, a subset of those sequences are chosen by selecting the top-*n* sequences. (The parameter *n* corresponds to 'No. of homologous' in Table 1.) We then run CentroidHomfold with those automatically collected homologous sequences. Finally, CentroidHomfold can output 'non-canonical' base pairs in a predicted secondary structure, non-canonical base pairs were removed from the set of base pairs in the predicted secondary structure.

We implemented the above pipeline in a Web application (http://www.ncrna.org/centroidhomfold/), which has an interface that is quite similar to the CentroidFold Web application (26) (Figure 3). The prediction results are shown in two ways: the input sequence with standard base pair notation (Figure 3B, 'Seq/Fold'), and a popular secondary structure graph (Figure 3B, 'Graph'). A PDF version of the graph and a FASTA file of homologous sequences (used in CentroidHomfold) can be downloaded by clicking on a link. The colors of base pairs in the graphical representation indicate the pseudo base-pairing probabilities, which are a measure of the reliability of the predicted base pairs (see 'Materials and Methods' section). Libraries in the Vienna RNA package are employed for the drawing graphical representation of secondary structures. By expanding the 'Advanced options' control, users are able to adjust all of the parameters in Table 1.

CentroidHomfold–LAST substantially outperforms conventional secondary structure predictors

Figure 4 indicates that CentroidHomfold–LAST has substantially improved accuracy compared with the conventional secondary structure predictors CentroidFold (3), Sfold (18), Simfold (17) and RNAfold (6). Figure 5 shows an example of the predictions of the CentroidHomfold–LAST (this work) and the CentroidFold Web application (26) (http://www.ncrna.org/centroidfold/) for a typical tRNA sequence. In this example, CentroidHomfold–LAST successfully predicted a 'clover-leaf' secondary structure while CentroidFold failed.

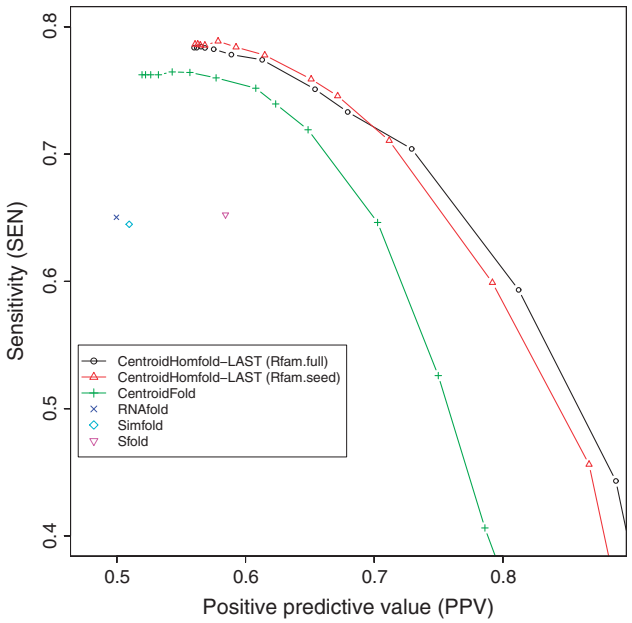

Figure 4. Performance of CentroidHomfold–LAST with default settings ($n = 30$ and $e = 0.01$), compared to the conventional secondary structure predictors CentroidFold, RNAfold, Sfold and SimFold. The axes are the PPV and Sensitivity with respect to base pairs in secondary structures. We used 17 values of the γ parameter ($\gamma \in \{2^k : -5 \leq k \leq 10,\ k \in \mathbb{Z}\} \cup \{6\}$) in order to draw the sensitivity–PPV curves for CentroidHomfold–LAST.

It is also interesting that using Rfam.full (which contains a large number of predicted RNA sequences) gave slightly better performances than using Rfam.seed (which consists of only manually-curated sequences). This suggests that tentative RNA sequences included in Rfam.full improve the accuracy.

In order to confirm robustness of CentroidHomfold–LAST, we conducted the experiments with various values for the parameters n and e, where n is the maximum number of homologous sequences and e is a threshold E-value for homologous searches (Supplementary Table S2). This figure indicates that CentroidHomfold–LAST is robust to changes in n and e, although a smaller n does yield a slightly worse performance.

We also conducted experiments using various probabilistic models (Supplementary Figure S3): the ProbCons model (23) and the CONTRAlign model (21) for probability distribution of pairwise alignments; the CONTRAfold model (2) and the BL model (22) for probability distributions of secondary structures Supplementary Figure S3 indicates that using the BL model resulted in better performance than using the CONTRAfold model (while using the ProbCons model and the CONTRAlign model lead to similar performances). Moreover, using the BL model is faster than using the CONTRAfold model (Supplementary Table S2). Therefore, the BL model is employed as the default setting in our Web application.

One drawback of CentroidHomfold–LAST is that it is slower than conventional algorithms for secondary structure prediction. It takes 451 (Rfam.full.99) and 267 (Rfam.seed.99) s to compute secondary structures of 151 RNA sequences using CentroidHomfold–LAST (Table 2), while it takes only 30 s using CentroidFold (3). This is because our method computes base-paring probability matrices for all the input sequences (target and

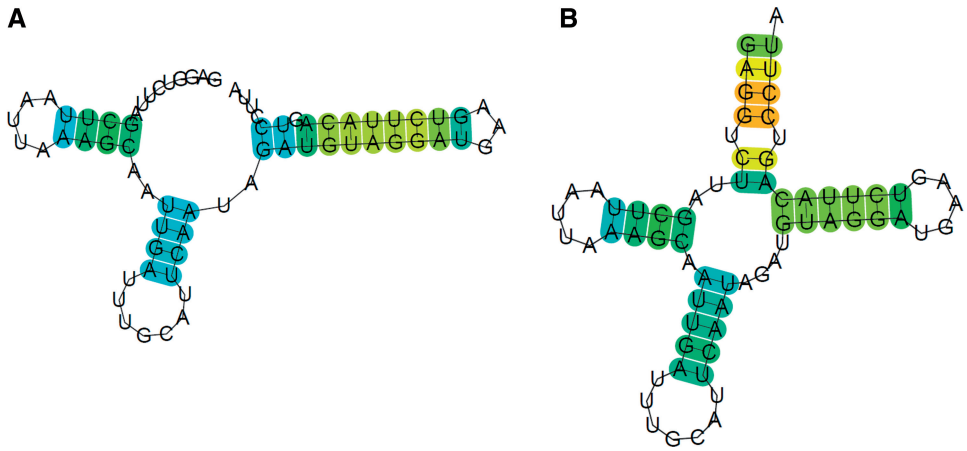

Figure 5. Example of a secondary structure predicted by the CentroidFold Web application (26), http://www.ncrna.org/centroidfold/ (**A**) and the CentroidHomfold-LAST Web application (**B**). A typical tRNA sequence GAGGUCUUAGCUUAAUUAAAGCAAUUGAUUUGCAUUCAAUA GAUGUAGGAUGAAGUCUUACAGUCCUUA (L07095.1/5018-5086) was used.

Table 2. Total computational time in seconds

	Rfam.full.99	Rfam.seed.99
Lastal	62	14
Select homologous from lastal results	54	22
CentroidHomfold	334	229
Total	451	267

Each value indicates the total computational time for predicting secondary structures for 151 RNA sequences (17 values of γ were used in order to draw the performance curves in Figure 4). The row 'Lastal' includes a score threshold calculation (by using lastex) and a lastal search of a given database. The row 'Select homologous from lastal results' includes collection of actual RNA sequences in the database using Lastal results. We used an Intel(R) Xeon(R) CPU X5550 machine with a Linux OS.

homologous sequences) and all the alignment probability matrices between the target sequence and each of homologous sequences.

Another drawback of the Web application is that users can only employ database based on the Rfam databases (Supplementary Table S1) and the length of the input sequence is limited to 500. A standalone pipeline system is therefore available from http://www.ncrna.org/software/centroidhomfold/pipeline/, in which there is no restriction for the length of the input RNA sequence and an original (user defined) RNA database can be employed for the candidates of homologous sequences. In our future work, we plan to develop an interface that enables users to upload RNA sequences for the database in the Web application, and provide more databases derived from, for example, the UCSC genome browser.

CONCLUSION

By combining CentroidHomfold (12) and LAST (13), we have developed a novel Web application (CentroidHomfold–LAST pipeline) that predicts the secondary structure of an input RNA sequence using automatically collected homologous sequences of the input sequence. Benchmark experiments with a commonly used data set indicated that CentroidHomfold–LAST substantially outperformed conventional secondary structure predictors (e.g. RNAfold and CentroidFold). The Web application is freely available from http://www.ncrna.org/centroidhomfold/. To the best of our knowledge, there is no Web application that has the same function as ours and it will be useful in a number of studies of RNAs, especially non-coding RNAs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Grant-in-Aid for Scientific Research on Innovative Areas (in parts). Funding for open access charge: Grant-in-Aid for Scientific Research on Innovative Areas.

Conflict of interest statement. None declared.

REFERENCES

1. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet*., **15**, 17–29.
2. Do,C., Woods,D. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
3. Hamada,M., Kiryu,H., Sato,K., Mituyama,T. and Asai,K. (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.
4. Markham,N. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol*., **453**, 3–31.
5. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.
6. Hofacker,I., Fontana,W., Stadler,P., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem*., **125**, 167–188.
7. Mathews,D., Disney,M., Childs,J., Schroeder,S., Zuker,M. and Turner,D. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.
8. Hamada,M., Kiryu,H., Iwasaki,W. and Asai,K. (2011) Generalized centroid estimators in bioinformatics. *PLoS ONE*, **6**, e16450.
9. Bernhart,S., Hofacker,I., Will,S., Gruber,A. and Stadler,P. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
10. Hamada,M., Sato,K. and Asai,K. (2011) Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic Acids Res*., **39**, 393–402.
11. Seemann,S., Gorodkin,J. and Backofen,R. (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res*., **36**, 6355–6362.
12. Hamada,M., Sato,K., Kiryu,H., Mituyama,T. and Asai,K. (2009) Predictions of RNA secondary structure by combining homologous sequence information. *Bioinformatics*, **25**, i330–i338.
13. Frith,M.C., Hamada,M. and Horton,P. (2010) Parameters for accurate genome alignment. *BMC Bioinformatics*, **11**, 80.
14. Kielbasa,S.M., Wan,R., Sato,K., Horton,P. and Frith,M. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res*., **21**, 487–493.
15. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al*. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res*., **37**, D136–D140.
16. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*., **31**, 3429–3431.
17. Andronescu,M., Condon,A., Hoos,H., Mathews,D. and Murphy,K. (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, **23**, 19–28.
18. Ding,Y., Chan,C.Y. and Lawrence,C.E. (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res*., **32**, W135–W141.
19. Sankoff,D. (1985) Simultaneous solution of the RNA folding alignment and protosequence problems. *SIAM J. Appl. Math*, **45**, 810–825.
20. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
21. Do,C., Gross,S. and Batzoglou,S. (2006) Contralign: Discriminative training for protein sequence alignment. In Apostolico,A., Guerra,C., Istrail,S., Pevzner,P.A. and Waterman,M.S. (eds), *RECOMB*, Vol. 3909 of *Lecture Notes in Computer Science*. Springer, pp. 160–174.
22. Andronescu,M., Condon,A., Hoos,H.H., Mathews,D.H. and Murphy,K.P. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.

23. Do,C., Mahabhashyam,M., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
24. Sheetlin,S., Park,Y. and Spouge,J.L. (2005) The Gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment. *Nucleic Acids Res.*, **33**, 4987–4994.
25. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
26. Sato,K., Hamada,M., Asai,K. and Mituyama,T. (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res.*, **37**, W277–W280.