Software.ncrna.org: web servers for analyses of RNA sequences

Kiyoshi Asai[1,2,*], Hisanori Kiryu[2], Michiaki Hamada[2,3], Yasuo Tabei[1,2], Kengo Sato[2,4,5], Hiroshi Matsui[5], Yasubumi Sakakibara[5], Goro Terai[2,4,6] and Toutai Mituyama[2]

[1]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwa-no-ha, Chiba 277-8561, [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, [3]Mizuho Information & Research Institute, Inc., 2-3, Kanda-Nishikicho, Chiyoda-ku, Tokyo 101-8443, [4]Japan Biological Informatics Consortium, 10F TIME24 Building, 2-45 Aomi, Koto-ku, Tokyo 135-8073, [5]Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522 and [6]INTEC Systems Institute, Inc., 1-3-3 Shinsuna, Koto-ku, Tokyo 136-0075, Japan

ABSTRACT

We present web servers for analysis of non-coding RNA sequences on the basis of their secondary structures. Software tools for structural multiple sequence alignments, structural pairwise sequence alignments and structural motif findings are available from the integrated web server and the individual stand-alone web servers. The servers are located at http://software.ncrna.org, along with the information for the evaluation and downloading. This website is freely available to all users and there is no login requirement.

INTRODUCTION

Comparisons, alignments and motif identification are essential procedures for extracting valuable information from biological sequences. Many effective software tools for these purposes are available for use with amino acid and DNA sequences, but their efficiency for RNA sequences is limited because they do not accommodate analysis of possible secondary structures. Practical analyses of multiple RNA sequences in light of their secondary structures have been difficult because of their extremely high computational costs, but several algorithms have been proposed and there are a few websites of software tools that support structure-based analyses of RNA sequences, e.g. Vienna RNA Package (http://www.tbi.univie.ac.at/~ivo/RNA/), Sfold (http://sfold.wadsworth.org) and BiBiServ (http://bibiserv.techfak.uni-bielefeld.de/).

Recent progress in RNA sequence analysis has created a demand for rapid and accurate structure-based analyses of multiple RNA sequences. To this end, we have developed several software tools for comparison (1,2), alignment (3–8) and motif identification (9) of multiple RNA sequences; searches for conserved miRNAs (10); prediction of common secondary structures from multiple sequence alignments (11) and calculation of base-pairing probabilities for long sequences (12). Using these software tools, we have developed an integrated web server and stand-alone web servers (software.ncrna.org) that support multiple alignment, pairwise alignment and extraction of structural motifs of RNA sequences.

METHODS

The integrated web server and the stand-alone web servers we developed offer three types of RNA sequence analyses based on common potential secondary structures: pairwise alignment, multiple alignment and structural motif extraction. SCARNA, PHMMTS (pair hidden Markov models on tree structures), PSTAG (pair stochastic tree adjoining grammar), Murlet and MXSCARNA can be used on their stand-alone web servers as well as on our integrated server. In addition, the source codes for PHMMTS, PSTAG, Murlet and MXSCARNA are available for download. The brief introductions to those tools follow, while the detailed evaluations are described in the refs (3–9) and their summaries given on web (http://software.ncrna.org).

Pairwise alignment

SCARNA (3) is a rapid structural pairwise alignment tool for RNA sequences of unknown secondary structure.

*To whom correspondence should be addressed. Tel: +81 33599 8089; Fax: +81 33599 8081; Email: asai@k.u-tokyo.ac.jp

This program separately aligns the 5′ and 3′ regions of stem candidates, which are extracted from each RNA sequence in light of base-pairing probabilities (12,13), by use of an engineered DP algorithm that incorporates rough consideration of consistency. We compared SCARNA with several other alignment tools by using Gardner's benchmark dataset (14) and a dataset comprising 5S ribosomal RNA, 5.8S ribosomal RNA and Hammerhead ribozyme from the Rfam database (15). The alignment accuracies of SCARNA for sequences with low similarities were not as high as those of programs that evaluate secondary structures more strictly, e.g. Foldalign (16), Dynalign (17) and PMcomp (18). However, the computational speed of SCARNA was approximately one order of magnitude faster (i.e. <1 min for 1000 bases) and allowed alignment of sequences longer than 1000 bases.

PHMMTS (4,5) and PSTAG (6) are tools for aligning RNA sequences of unknown secondary structure to RNA sequences with known secondary structure. PHMMTS evaluates only pseudoknot-free structures, whereas PSTAG can accept pseudoknotted structures. When compared with ClustalW (19) by using tRNA and Hammerhead ribozyme datasets, PHMMTS was more accurate in regard to correct assignment of secondary structures. In a comparison with PHMMTS and ClustalW by using RNA sequences of HDV_ribozyme, an RNA family in PseudoBase (20) that includes pseudoknotted structures, PSTAG was more accurate in correct assignment of secondary structures.

Multiple alignment

Murlet (7) and MXSCARNA (8) are structural multiple alignment tools for RNA sequences. Murlet is based on pair SCFG (stochastic context-free grammar), has dramatically decreased computational costs, and is applicable to RNA sequences as long as 300 bases. MXSCARNA is an extension of SCARNA that offers progressive alignment and is applicable to RNA sequences as long as 5000 bases though the accuracies for those longer than 500 bases are not confirmed and the lengths are restricted to 1000 in the web server. We validated Murlet and MXSCARNA by using the BRAlibaseII benchmark dataset (14) and the dataset of Kiryu *et al.* (7).

Both tools showed comparative accuracies in SPS (sum-of-pairs score) with ProbCons (21). The accuracies in potential common secondary structures were evaluated by MCC (Mathew's correlation coefficient), and both tools showed comparative accuracy with Stemloc (22).

Motif extraction

RNAmine (9) is a tool for extraction of structural motifs. This program uses a graph-mining technique to identify local sequences with frequent stem patterns from among a set of RNA sequences. RNAmine is currently available only on the integrated web server.

Additional tools

In addition to the six software programs described, the tools SOKOS/CAN (1), Stem Kernel (2), miRRim (10), McCaskill-MEA (11) and Rfold (12) are available for download. Although pairwise alignment is the default method, kernels can be used as similarities in an alternative approach for comparing two biological sequences. SOCOS/CAN and Stem Kernel are tools for sequence comparison, both of which use features of the potential secondary structures to calculate the kernel function. SOKOS/CAN calculates the marginalized kernel on SCFG, and Stem Kernel compares the sequences by the kernel based on all possible stem patterns.

Predicting non-coding RNAs is difficult because general characteristic sequence patterns are not known. For specific families of non-coding RNAs, however, realistic predictions are possible. We developed miRRim (10) as a tool for finding conserved miRNAs.

McCaskill-MEA (11) is a method used to predict consensus secondary structures from given multiple alignments. Rfold (12) is a tool for calculating the local base-pairing probabilities without using sliding windows; it is based on the full energy model of the Vienna RNA Package (23).

DESCRIPTION OF SERVICES

Table 1 shows a list of software tools available at http://software.ncrna.org/.

Table 1. List of software tools at http://software.ncrna.org

Software tool	Function	Pseudo-knot	Download	Integrated web server		Stand-alone web server
				Max. no. of seq.	Max. length	
SCARNA	Pairwise alignment	Yes	N/A	5	1000	Yes
PHMMTS	Pairwise alignment (to known structure)	No	C++ source	5	1000	Yes
PSTAG	Pairwise alignment (to known structure)	Yes	C++ source	5	70	Yes
MXSCARNA	Multiple alignment	No	C++ source	10	1000	Yes
Murlet	Multiple alignment	No	C++ source	5	300	Yes
RNAmine	Motif extraction	No	contact	10	500	No
SOKOS/CAN	Sequence comparison	No	C source	N/A	N/A	No
Stem Kernel	Sequence comparison	Yes	C++ source	N/A	N/A	No
miRRim	miRNA finding	No	Source script	N/A	N/A	No
McCaskill-MEA	Common secondary structure prediction	No	C++ source	N/A	N/A	No
Rfold	Base pairing probabilities	No	C++ source	N/A	N/A	No

The integrated web server and the stand-alone web servers offer web interfaces for use of SCARNA, PHMMTS, PSTAG, Murlet, MXSCARNA and RNAmine. On the integrated web server, users can select one of the service types: multiple alignment, pairwise alignment or structural motif extraction. On the menu for 'multiple alignment', users can select either Murlet or MXSCARNA. Either direct input or uploading of a file of RNA sequences in multi-FASTA format is accepted. The server outputs a multiple alignment with annotations of the predicted common secondary structure, a figure of the structure and a phylogenetic tree of the sequences.

On the menu for 'pairwise alignment', users can select SCARNA, PHMMTS or PSTAG. In SCARNA, either direct input or uploading of a file of RNA sequences in multi-FASTA format is accepted. The server outputs a pairwise alignment with annotations of the predicted common secondary structures and a figure of the structure that includes the two aligned sequences. PHMMTS and PSTAG accept either direct input or uploading of a file of RNA sequences of unknown secondary structures in multi-FASTA format as query sequences, and direct input of an RNA sequence of known secondary structure and its secondary structure in dot-bracket format as the template structure. The server outputs the result of alignments of the query sequences to the template structure with annotations of the secondary structures and the same kind of figures of the structures as SCARNA.

On the 'structural motif extraction' menu, users can select RNAmine, which accepts either direct input or uploading of a file of RNA sequences in multi-FASTA format. The server outputs the extracted motifs as abstract figures of the secondary structures and the list of the members by sequence name and positions. Detailed figures and the structure-annotated sequence are linked to the members.

For each of the software tools described, after the server outputs the results the user can continue to a homology search of the RNA sequences by BLAT for various genomes. Hits of the search are equipped with links to UCSC GenomeBrowser for functional RNAs (24).

FUTURE PLANS

In addition to web servers, web services for sequence analysis tools are desirable. We have already developed soap-based web services for Murlet, MXSCARNA, PHMMTS and PSTAG. The services will start shortly.

CONCLUSION

We have developed web servers for analysis of RNA sequences in light of their secondary structures. The web server offers six software tools for multiple sequence alignment, pairwise alignment and extraction of structural motifs of RNA sequences. These servers provide practical speed of services for the tasks that have been thought to require high computational costs.

REFERENCES

1. Kin,T., Tsuda,K. and Asai,K. (2002) Marginalized kernels for RNA sequence data analysis. *Genome Inform.*, **13**, 112–122.
2. Sakakibara,Y., Popendorf,K., Ogawa,N., Asai,K. and Sato,K. (2007) Stem kernels for RNA sequence analyses. *J. Bioinform. Comput. Biol.*, **5**, 1103–1122.
3. Tabei,Y., Tsuda,K., Kin,T. and Asai,K. (2006) SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *Bioinformatics*, **22**, 1723–1729.
4. Sakakibara,Y. (2003) Pair hidden Markov models on tree structures. *Bioinformatics*, **19**, i232–i240.
5. Sato,K. and Sakakibara,Y. (2005) RNA secondary structural alignment with conditional random fields. *Bioinformatics*, **21**, ii237–ii242.
6. Matsui,H., Sato,K. and Sakakibara,Y. (2005) Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures. *Bioinformatics*, **21**, 2611–2617.
7. Kiryu,H., Tabei,Y., Kin,T. and Asai,K. (2007) Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, **23**, 1588–1598.
8. Tabei,Y., Kiryu,H., Kin,T. and Asai,K. (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinform.*, **9**, 33.
9. Hamada,M., Tsuda,K., Kudo,T., Kin,T. and Asai,K. (2006) Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, **22**, 2480–2487.
10. Terai,G., Komori,T., Asai,K. and Kin,T. (2007) miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity. *RNA*, **13**, 2081–2090.
11. Kiryu,H., Kin,T. and Asai,K. (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics*, **23**, 434–441.
12. Kiryu,H., Kin,T. and Asai,K. (2008) Rfold: an exact algorithm for computing local base pairing probabilities. *Bioinformatics*, **24**, 367–373.
13. McCaskill,J. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
14. Gardner,P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
15. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
16. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.

17. Harmanci,A.O., Sharma,G. and Mathews,D.H. (2007) Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*, **8**, 130.

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

19. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

20. van Batenburg,F.H.D., Gultyaev,A.P., Pleij,W.A., Ng,J. and Oliehoek,J. (2000) PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res.*, **28**, 201–204.

21. Do,C.B., Mahabhashyam,M.S.P., Brudno,M. and Batzoglou,S. (2005) PROBCONS: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, **15**, 330–340.

22. Holmes,I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinform.*, **6**, 73.

23. Hofacker,I. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

24. Kin,T., Yamada,K., Terai,G., Okida,H., Yoshinari,Y., Ono,Y., Kojima,A., Kimura,Y., Komori,T. and Asai,K. (2006) fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.*, **35**, D145–D148.