Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W423–W428***
doi:10.1093/nar/gkl231

RNAMST: efficient and flexible approach for identifying RNA structural homologs

Tzu-Hao Chang[1], Hsien-Da Huang[2,3,4,*], Tzu-Neng Chuang[1,5], Dray-Ming Shien[1,6] and Jorng-Tzong Horng[1,7,*]

[1]Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan, [2]Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan, [3]Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan, [4]Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan, [5]Department of Electronic Engineering, Ching-Yun University, Chung-Li 320, Taiwan, [6]Department of Electronic Engineering, Chin Min Institute of Technology, Miao-Li, Taiwan and [7]Department of Life Science, National Central University, Chung-Li 320, Taiwan

Received February 14, 2006; Revised March 4, 2006; Accepted March 27, 2006

ABSTRACT

RNA molecules fold into characteristic secondary structures for their diverse functional activities such as post-translational regulation of gene expression. Searching homologs of a pre-defined RNA structural motif, which may be a known functional element or a putative RNA structural motif, can provide useful information for deciphering RNA regulatory mechanisms. Since searching for the RNA structural homologs among the numerous RNA sequences is extremely time-consuming, this work develops a data preprocessing strategy to enhance the search efficiency and presents RNAMST, which is an efficient and flexible web server for rapidly identifying homologs of a pre-defined RNA structural motif among numerous RNA sequences. Intuitive user interface are provided on the web server to facilitate the predictive analysis. By comparing the proposed web server to other tools developed previously, RNAMST performs remarkably more efficiently and provides more effective and flexible functions. RNAMST is now available on the web at http://bioinfo.csie.ncu.edu.tw/~rnamst/.

INTRODUCTION

RNA molecules are involved in numerous biological processes, ranging from gene regulation to protein synthesis. Some regulatory mechanisms involve binding of proteins or metabolites to specific sites of RNA molecules (1–7). For instance, riboswitches are metabolite-binding domain within a specific mRNA, and can regulate both transcription and translation by binding their corresponding targets (4–7).

The function of RNA molecules frequently depends on the motifs conserved in both secondary structure and sequence. Numerous biologically relevant RNA motifs, including signal recognition particle, RNase P, *cis*-acting RNA regulatory elements (eleRNAs) and microRNAs, are found in the untranslated regions (UTRs) of mRNA and are involved in the regulation of gene expression. Most of these elements are conserved better in structures than in sequences (8–10). Searching these highly conserved RNA structural motifs can yield useful information for deciphering RNA regulatory mechanisms.

Several publicly available motif search tools, e.g. tRNAscan-SE (11) and Riboswitch finder (12), were developed to identify transfer RNA and riboswitch RNA against RNA sequence, respectively. In addition to these tools for achieving special types of regulatory RNAs, numerous tools such as Palingol (13), RNAMotif (14), PatSearch (15) and RNABOB (http://selab.wustl.edu/cgi-bin/selab.pl?mode=software#rnabob) were developed so that users could define RNA motifs according to their requirements. Palingol (13) uses a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. RNAMotif (14) devises the definition of RNA structural descriptor to describe an RNA secondary structure and scans sequences for structural homologs. PatSearch (15) is a website for detecting patterns and structural motifs in nucleotide sequences. RNABOB is a program searching for RNA motifs in sequence databases.

Although these tools can successfully define the RNA structural motifs and identify putative motifs among the limited

*To whom correspondence should be addressed. Tel: +886 3 5712121 (ext. 56952); Fax: +886 3 5729288; Email: bryan@mail.nctu.edu.tw
*Correspondence may also be addressed to Jorng-Tzong Horng. Fax: +886 3 4222681; Email: horng@db.csie.ncu.edu.tw

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

range of RNA sequences, the search time can become a critical concern when searching in very large quantities of RNA sequences. In order to implement a web server for the convenient usage of biologists, more efficient tools for identify homologs of RNA structural motifs are crucial. This work develops an efficient and flexible tool, namely RNAMST, for rapidly searching for homologs of a pre-defined RNA structural motif, including hairpin, internal loop as well as multi-branch loop, among numerous RNA sequences. Additionally, intuitive web interface are provided on the web server to facilitate the predictive analysis.

METHODS

Figure 1 shows the system flow of RNAMST, which comprises RNA structural description generation, data preprocessing and searching structural homologs. Each part of the RNAMST is described in detail below.

RNA structural descriptor generation

The RNAMST web server accepts four input formats, i.e. the RNA structural descriptor, bracket notation, ct format (16) and RNA sequence, to allow users to easily describe an RNA structural motif. An RNA structural descriptor defined previously by RNAMotif is used to describe a specific structural motif, and can be recognized through the step of searching structure homologs. The bracket notation and ct format are transformed to structural descriptor. In addition, the input of RNA sequence is folded by mfold (16), stored as ct format file and then transformed into structural descriptor.

Data preprocessing

The RNAMST designs a preprocessing strategy to prepare all the possible hairpins of the search sequences in advance. For each search sequence, RNAMST calculates the possible helices based on local sequence alignment. For example, Supplementary Figure S1 contains a hairpin and a non-external helix. RNAMST detects all possible hairpins and collects them into the indexed RNA sequence database. Additionally, this work exploits the multilevel index technique to establish the index file based on the order of stem length, loop length and the number of mispairs. Through effectively using these indexed files and adopting binary search algorithm, RNAMST can retrieve the requiring hairpins rapidly and efficiently.

Furthermore, since tandem repeat may produce excessive noise when determining helical regions, the tandem repeats within the search sequences are detected by the Tandem Repeat Finder (17), and then are filtered out from the indexed RNA sequence database. Following the preprocessing, RNAMST can retrieve all the requiring hairpin information in search sequences to fit the input RNA structural motifs extremely quickly.

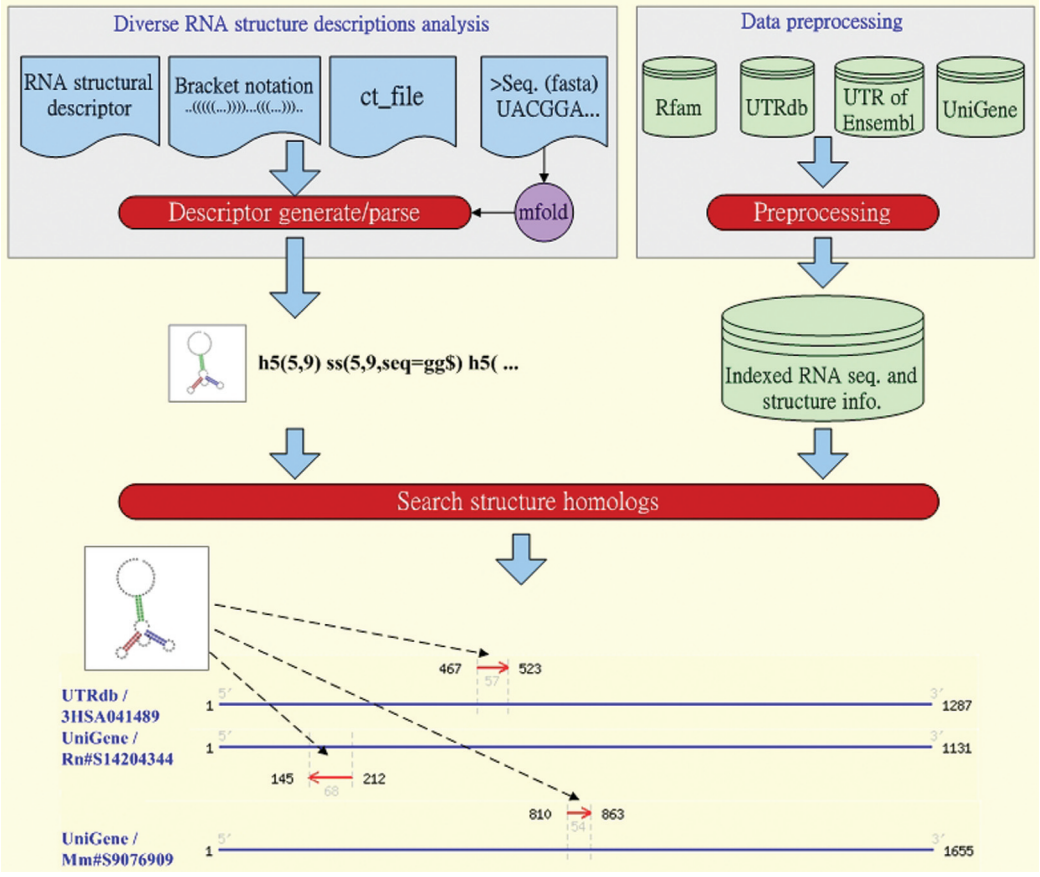

Figure 1. Search process of RNAMST. The figure shows how RNAMST searches a specified RNA structural motif in the given RNA sequences.

	Rfam (11Mbps)	UniGene Oryza ativa (55Mbps)	UniGene Mus musculus (88Mbps)	UniGene Homo sapiens (106Mbps)
RNABOB	9.4sec	95sec	197sec	231sec
RNAMotif	2.3sec	11.3sec	22sec	25sec
RNAMST	1.5sec	7.2sec	11.4sec	13.3sec

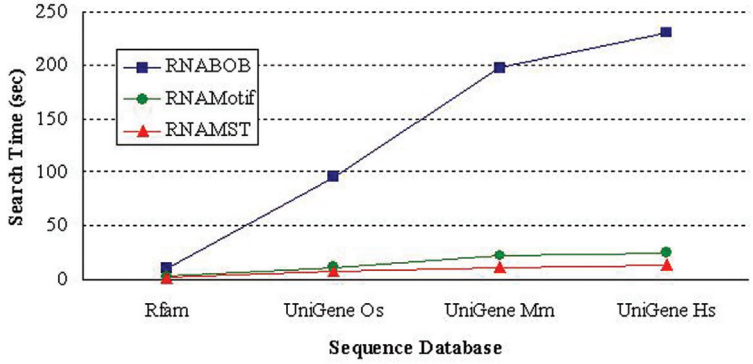

Figure 2. Performance comparison for RNABOB, RNAMotif and RNAMST. Search time of RNABOB, RNAMotif and RNAMST against the different search sequence databases with the IRE motif.

Searching structural homologs

The RNAMST search algorithm involves two steps: the combination of external hairpins and the combination of non-external helices. Initially, RNAMST seeks all possible hairpins satisfying the constraints, which are described in RNA structural descriptor, from the indexed RNA sequence database. Subsequently, RNAMST then seeks the non-external helices using local alignment. If non-external helices are found, RNAMST integrates the external hairpins and non-external helices to fit the RNA structural motif, thus generating results of RNA structural homologs.

The RNAMST program is implemented using C++ programming language and runs under the Linux operating system on a PC server. As to the preprocessing of the search sequence database, Supplementary Table S1 gives the indexing time and required storage.

PERFORMANCE EVALUATION

This section describes the efficiency and flexibility of RNAMST. Several well-known RNA motifs in Rfam (18) are used to search for structural homologs against the sequences with different sizes, and the search time are compared with those experiments done using RNAMotif (14).

To evaluate the performance of searching for structural homologs among large numbers of sequences, RNAMotif (14) is compared with RNAMST developed here in searching for two kind of RNA motifs, iron response element (IRE) (19) and purine riboswitch (5), with different sized databases.

Supplementary Figure S1 shows the RNA structural motif of IRE, which is a short and conserved stem–loop found in UTRs of various mRNA. The IRE regulates mRNA translation by interacting with iron regulatory proteins (19). The RNA structural descriptor of IRE is given in Supplementary Figure S1. Figure 2 gives the comparison of the search time of RNABOB, RNAMotif and RNAMST with different sized databases. For example, RNABOB and RNAMotif spent 231 and 25 s in detecting homologs of IRE motifs in UniGene human sequence database, respectively, while our RNAMST only needed 13.3 s to accomplish the same search. Since IRE is a very small motif and is highly conserved, the constraint of the IRE structural descriptor is quite strict, making it easy to be detected. Despite this, RNAMST is still ~9 and 1.9 times faster than RNABOB and RNAMotif, respectively, and the difference increases as the structure becomes more complex and the constraints loosen.

Supplementary Figure S2 gives the RNA structural descriptor of purine riboswitch, which is a metabolite-binding domain within certain mRNAs and is involved in modulating gene expression (5). Similarly, Figure 3 illustrates the comparison of the search time among RNABOB, RNAMotif and RNAMST for different sized databases. For instance, RNABOB and RNAMotif spend 16 676 and 2048 s in detecting homologs of purine riboswitch, respectively, while our RNAMST spent only 17 s. In this case, owing to the more complex structure and looser constraints of purine riboswitch compared to IRE motif, RNAMST performs ~980 and 120 times faster than RNAMotif and RNABOB, respectively. After the above overall evaluation, our RNAMST has a lot of improvement in search performance.

Flexible structural search

Besides the efficient RNA structural homolog search, RNAMST also consider flexible structures with asymmetric mispairs and bulges to increase its applicability and practicality. For example, to search for homologs allowing asymmetric mispairs and bulges of a long hairpin such as the example of RNA structural motif given in Supplementary Figure S3-A,

	Rfam (11Mbps)	UniGene Oryza ativa (55Mbps)	UniGene Mus musculus (88Mbps)	UniGene Homo sapiens (106Mbps)
RNABOB	1572sec	8627sec	13382sec	16676sec
RNAMotif	138sec	989sec	1664sec	2048sec
RNAMST	3sec	9sec	14sec	17sec

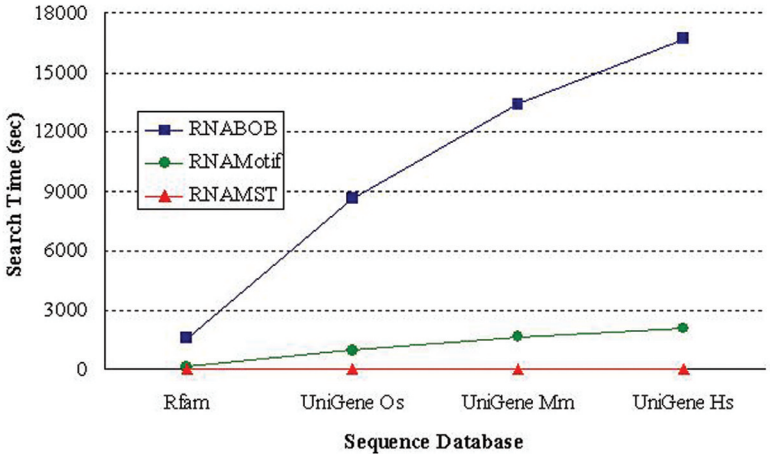

Figure 3. Performance comparison for RNABOB, RNAMotif and RNAMST. Search time of RNABOB, RNAMotif and RNAMST for the different search sequence databases with purine riboswitch motifs.

RNAMotif only yields results in structures with symmetrical mispair (Supplementary Figure S3-B) but fails to find any results in cases involving asymmetrical mispair (Supplementary Figure S3-D) and bulge (Supplementary Figure S3-C, E), despite two types of structures are similar. Because searching for motifs with asymmetrical mispair or bulge is extremely time-consuming, none of the RNA structural motif search tools developed previously support this consideration, despite these structures commonly appearing in RNA secondary structures. RNAMST overcomes this problem through preprocessing so as to support more flexible search for RNA structural homologs.

This work presents an example of finding microRNA (miRNA) mir-1 motif using flexible search. The miRNAs take the form of a long hairpin structure (Supplementary Figure S4) and trigger a *cis*-acting translational repressive activity by involving Dicer enzyme. The RNA structural descriptor of the mir-1 is given in Supplementary Figure S4. Supplementary Figure S5 lists the sequences in mir-1 family in Rfam that can be identified using mir-1 motif by RNAMST. All seven known sequences in the mir-1 family of Rfam were identified correctly by RNAMST with flexible structure search, but only one of seven sequences, AC002442.1, was identified by RNAMotif.

INTERFACE

Indexed databases are provided on the web server, including UTRdb (20), Rfam (18), UniGene and UTR sequences from Ensembl . Users can input the description of a desired RNA structural motif via four different ways and select the indexed RNA databases through the user web interface. Moreover, the search results can rapidly be displayed on the webpage. Figure 4 shows the web interface of RNAMST. RNAMST involves the following steps in searching for homologs of a specific RNA structural motif.

Firstly, users can input the specific RNA structural motif using an RNA structural descriptor. In addition to RNAMST web server also accepts other inputs, including bracket notation, files in ct format and FASTA format RNA sequences.

Secondly, owing to many of the functional RNA structures being discovered in the UTR, RNAMST web server provides UTR related databases, including UTRdb and UTR sequences from Ensembl. Furthermore, RNAMST also provides RNA sequence databases, including Rfam and UniGene. All of these databases were indexed during the preprocessing phase. Users simply need to select the sequences of interest. Accordingly, users can conveniently and rapidly search these databases for homologs of an input RNA structural motif.

Thirdly, the predictive constraints can be specified by users. Besides the input of the structural descriptor, the other formats can set the search similarity constraint to add the mispairs and bulges for the query structural motif. RNAMST supports two kinds of base pairing to fold helices: Watson–Crick pairing (AT and GC pairing), and both Watson–Crick pairing and G–U pairing.

Finally, the RNAMST generates the output for the discovered RNA structural homologs, as well as provides additional information for further analysis, including description of the search sequence containing the homologs, alignment and RNA secondary structure of the homolog.

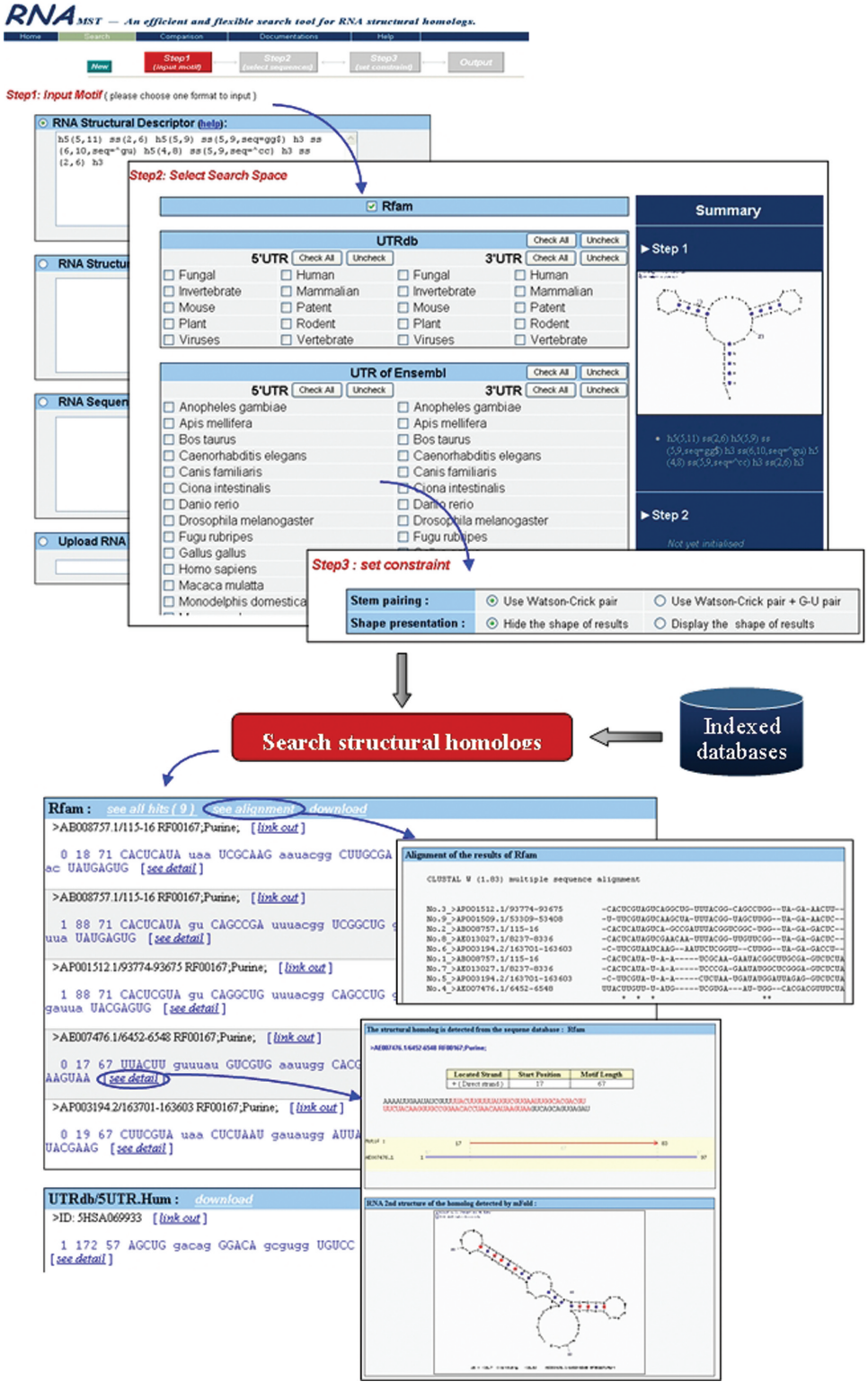

Figure 4. Web interface of RNAMST.

DISCUSSIONS AND CONCLUSION

RNAMST markedly improves the search time compared with other RNA motif search tools. Although other RNA motif tools such as Palingol, RNAMotif, RNABOB and PatSearch can achieve acceptable search times for searching for strict motifs in a small search sequence, they are not efficient enough to be implemented as a web server when searching in a large search sequence. Notably, the performance comparison among

Table 1. Features of RNAMST compared with other tools

Comparing Features	RNAMST	RNAMotif (14)	RNABOB	PatSearch (15)	Palingol (13)	Riboswitch finder (12)	tRNAscan-SE (11)
Diverse RNA structures search	Yes	Yes	Yes	Yes	Yes	—	—
High speed search against large amount of sequences	Yes	—	—	—	—	Yes	Yes
Web interface with real-time response	Yes	—	—	—	—	Yes	Yes
Flexible search with bugle and mispair	Yes	—	—	—	—	—	—
Accept multiple RNA description	Yes	—	—	—	—	—	—

the RNAMotif, RNABOB and our RNAMST is probably unfair due to the RNAMST incorporates a preprocessing strategy to improve the search. In order to provide a convenient web server for identifying RNA structural homologs, the authors suggest that the proposed methodology is a reasonable way to make the web service of this analysis become possible.

Several limitations of current implementation of the RNAMST are addressed below. RNAMST pre-calculates and stores all possible hairpin structures of the search sequence databases. Presently, RNAMST accumulates the hairpin structures which the length of hairpin stem ranges from 3 to 30 nt and the length of hairpin loop ranges from 3 nt to two times the length of the corresponding hairpin stem. Thus, the desired RNA structural homologs beyond the limitation of the pre-indexed database cannot be found. Another limitation is the pre-indexed hairpin structures and the resulting RNA structural homologs are computationally predicted by mfold. Users should note that the predicted RNA secondary structures can provide potential candidates for experimental confirmation. Owing to the complicated structures of the pseodonots, the web server does not support the search for pseudonots yet.

This works developed an efficient and flexible web server, RNAMST, capable of rapidly searching homologs of predefined RNA structures against numerous RNA sequences or databases. Table 1 briefly lists the advantages of RNAMST compared with other tools. The web interface enables users to easily implement the various steps of RNAMST. The pre-indexed RNA databases, such as UTRdb, Unigene, UTR of Ensembl and Rfam, are created in advance and supported on the web. The proposed RNAMST web server can provide a fast and convenient analysis for investigators who are interested in regulatory RNA motifs and elements.

AVAILABILITY

The RNAMST web server will be continuously maintained and updated. The web server is now freely available at http://bioinfo.csie.ncu.edu.tw/~rnamst/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 95-3112-E-009-002. Special thanks for the financially supports from National Research Program For Genomic Medicine (NRPGM), Taiwan. This work was also partially supported by MOE ATU. Funding to pay the Open Access publication charges for this article was provided by National Science Council of the Republic of China.

Conflict of interest statement. None declared.

REFERENCES

1. Klaff,P., Riesner,D. and Steger,G. (1996) RNA structure and the regulation of gene expression. *Plant Mol. Biol.*, **32**, 89–106.
2. Gygi,S.P., Rochon,Y., Franza,B.R. and Aebersold,R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, **19**, 1720–1730.
3. Gray,N.K. and Hentze,M.W. (1994) Regulation of protein synthesis by mRNA structure. *Mol. Biol. Rep.*, **19**, 195–200.
4. Barrick,J.E., Corbino,K.A., Winkler,W.C., Nahvi,A., Mandal,M., Collins,J., Lee,M., Roth,A., Sudarsan,N., Jona,I. *et al.* (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl Acad. Sci. USA*, **101**, 6421–6426.
5. Mandal,M., Boese,B., Barrick,J.E., Winkler,W.C. and Breaker,R.R. (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*, **113**, 577–586.
6. Soukup,J.K. and Soukup,G.A. (2004) Riboswitches exert genetic control through metabolite-induced conformational change. *Curr. Opin. Struct. Biol.*, **14**, 344–349.
7. Winkler,W.C. and Breaker,R.R. (2003) Genetic control by metabolite-binding riboswitches. *Chembiochem*, **4**, 1024–1032.
8. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, REVIEWS0004.
9. Eddy,S.R. (2002) Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.
10. Gray,N.K. and Wickens,M. (1998) Control of translation initiation in animals. *Annu. Rev. Cell Dev. Biol.*, **14**, 399–458.
11. Schattner,P., Brooks,A.N. and Lowe,T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.*, **33**, W686–689.
12. Bengert,P. and Dandekar,T. (2004) Riboswitch finder–a tool for identification of riboswitch RNAs. *Nucleic Acids Res.*, **32**, W154–W159.
13. Billoud,B., Kontic,M. and Viari,A. (1996) Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Res.*, **24**, 1395–1403.
14. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
15. Grillo,G., Licciulli,F., Liuni,S., Sbisa,E. and Pesole,G. (2003) PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.*, **31**, 3608–3612.
16. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
17. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
18. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
19. Hentze,M.W. and Kuhn,L.C. (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl Acad. Sci. USA*, **93**, 8175–8182.
20. Mignone,F., Grillo,G., Licciulli,F., Iacono,M., Liuni,S., Kersey,P.J., Duarte,J., Saccone,C. and Pesole,G. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **33**, D141–D146.