Published in final edited form as:

Mol Biotechnol. 2009 February ; 41(2): 180–193. doi:10.1007/s12033-008-9114-z.

Informatic Resources for Identifying and Annotating Structural RNA Motifs

Ajish D. George and

Gen*NY*Sis Center for Excellence in Cancer Genomics, University at Albany-SUNY, Department of Biomedical Sciences, School of Public Health, 1 Discovery Drive, Room 220, Rensselaer, NY 12144

Scott A. Tenenbaum, PhD.

Gen*NY*Sis Center for Excellence in Cancer Genomics, University at Albany-SUNY, Department of Biomedical Sciences, School of Public Health, 1 Discovery Drive, Room 220, Rensselaer, NY 12144, Phone (518) 591-7157; FAX (518) 591-7201

Ajish D. George; ; Scott A. Tenenbaum: stenenbaum@albany.edu

Abstract

Post-transcriptional regulation of genes and transcripts is a vital aspect of cellular processes, and unlike transcriptional regulation, remains a largely unexplored domain. One of the most obvious and most important questions to explore is the discovery of functional RNA elements. Many RNA elements have been characterized to date ranging from cis-regulatory motifs within mRNAs to large families of non-coding RNAs. Like protein coding genes, the functional motifs of these RNA elements are highly conserved, but unlike protein coding genes, it is most often structure and not sequence that conserved. Proper characterization of these structural RNA motifs is both the key and the limiting step to understanding the post-transcriptional aspects of the genomic world. Here we focus on the task of structural motif discovery and provide a survey of the informatics resources geared towards this task.

Introduction

Post-transcriptional regulation of genes and transcripts is a vital aspect of cellular processes and unlike transcriptional regulation, remains a largely unexplored domain. One of the most obvious and most important questions to explore is the discovery of functional RNA elements. Many RNA elements have been characterized to date ranging from cis-regulatory motifs within mRNAs to large families of non-coding RNA such as pre-miRNAs, snRNAs, snoRNAs, gRNAs, tRNAs, rRNAs, and assorted ribozymes. Like protein coding genes, the functional motifs of these RNA elements are highly conserved, but unlike protein coding genes, it is most often structure and not sequence that conserved. Proper characterization of these structural RNA motifs is both the key and the limiting step to understanding the post-transcriptional aspects of the genomic world.

Here we focus on the task of structural motif discovery and the informatics resources and tools geared towards this task. We present first the existing databases of RNA structures and their known instances (Table 1). These range from databases of directly imaged 3D structures to ones where consensus structures have been compiled either manually from literature or by using a computational approach. They also include databases that catalog the result of genome-

wide searches for conserved structures. Complementing these structure databases is a collection of tools for searching out instances of known structures in new sequences (Table 2).

We move on then to tools focusing on the discovery of new structural motifs from a set of related sequences. These are divided into two main families – ones that rely on pre-aligning the sequences (Table 3), and those that can work with unaligned sequences (Table 4). The first group includes notable covariance model based approaches as well as a smattering of classifier driven, Bayesian, thermodynamic, and aggregate approaches while the latter contains many improvements on the Sankoff algorithm for simultaneous sequence/structure alignment along with novel approaches such as shape-abstraction, suffix-arrays, genetic programming, and formal grammars. To aid in the comparison and benchmarking of these motif prediction algorithms, we also provide the two known attempts at compiling standardized datasets of motif-containing sequences (Table 5). The newer of these, TUTR, also contains matched control sets to help properly estimate sensitivity and specificity parameters for each algorithm.

We have not included here the numerous tools for predicting structures in individual sequences, for predicting interactions between RNA structures, or those for folding sequences in a specific association context or with specific thermodynamic constraints. These go well beyond the task of motif prediction as it relates to *families* of functionally related mRNAs and ncRNAs.

Using the listed tools, it should be possible to survey the known space of functional RNA motifs, to search for known motifs in new sequences, and to discover new structure families in related sets of aligned and unaligned sequences. This should provide a good starting point for studies of post-transcriptional regulatory elements and non-coding RNAs.

Acknowledgments

We wish to thank the members of the Tenenbaum Lab for helpful suggestions and discussion, especially Chris Zaleski and Frank Doyle. This work was supported in part by NIH grant U01HG004571 to SAT from the NHGRI.

Bibliography

- Abreu-Goodger C, Merino E. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Research* 2005;33(Web Server issue):W690–2. [PubMed: 15980564]
- Abreu-Goodger C, et al. Conserved regulatory motifs in bacteria: riboswitches and beyond. *Trends in Genetics: TIG* 2004;20(10):475–9. [PubMed: 15363900]
- Anwar M, Nguyen T, Turcotte M. Identification of consensus RNA secondary structures using suffix arrays. *BMC Bioinformatics* 2006;7:244. [PubMed: 16677380]
- Bafna, V.; Zhang, S. FastR: fast database search tool for non-coding RNA. *Proceedings/IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference*; 2004. p. 52-61.
- Bauer M, Klau GW, Reinert K. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics* 2007;8:271. [PubMed: 17662141]
- Berman HM, et al. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal* 1992;63(3):751–9. [PubMed: 1384741]
- di Bernardo D, Down T, Hubbard T. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics (Oxford, England)* 2003;19(13):1606–11.
- Bindewald E, et al. RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Research* 2008;36(Database issue):D392–7. [PubMed: 17947325]
- Bindewald E, Shapiro BA. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA (New York, NY)* 2006;12(3):342–52.

- Busch A, Backofen R. INFO-RNA--a fast approach to inverse RNA folding. *Bioinformatics* (Oxford, England) 2006;22(15):1823–31.
- Chang T, et al. RNAMST: efficient and flexible approach for identifying RNA structural homologs. *Nucleic Acids Research* 2006;34(Web Server issue):W423–W428. [PubMed: 16845040]
- Coventry A, Kleitman DJ, Berger B. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* 2004;101(33):12102–7. [PubMed: 15304649]
- Dalli D, et al. STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics* (Oxford, England) 2006;22(13):1593–9.
- Do CB, Foo C, Batzoglou S. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics* (Oxford, England) 2008;24(13):i68–76.
- Doyle F, et al. Bioinformatic tools for studying post-transcriptional gene regulation: The UAlbany TUTR collection and other informatic resources. *Methods in Molecular Biology* (Clifton, NJ) 2008;419:39–52.
- Dsouza M, Larsen N, Overbeek R. Searching for patterns in genomic data. *Trends in Genetics: TIG* 1997;13(12):497–8. [PubMed: 9433140]
- Eddy SR. Computational analysis of RNAs. *Cold Spring Harbor Symposia on Quantitative Biology* 2006;71:117–28.
- Gardner P, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 2004;5(1):140. [PubMed: 15458580]
- Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of Molecular Biology* 2001;313(5):1003–11. [PubMed: 11700055]
- Griffiths-Jones S, et al. miRBase: microRNA sequences, targets and gene nomenclature. *Nucl Acids Res* 2006;34(suppl1):D140–144. [PubMed: 16381832]
- Griffiths-Jones S, et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucl Acids Res* 2005;33(suppl1):D121–124. [PubMed: 15608160]
- Hamada M, et al. Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics* (Oxford, England) 2006;22(20):2480–7.
- Hofacker IL. RNA consensus structure prediction with RNAalifold. *Methods in Molecular Biology* (Clifton, NJ) 2007;395:527–44.
- Hofacker IL. RNA secondary structure analysis using the Vienna RNA package. *Current Protocols in Bioinformatics/Editorial Board, Andreas D Baxevanis ... [et Al]* 2004;Chapter 12(Unit 122)
- Hofacker IL. Vienna RNA secondary structure server. *Nucleic Acids Research* 2003;31(13):3429–31. [PubMed: 12824340]
- Hofacker IL, Bernhart SHF, Stadler PF. Alignment of RNA base pairing probability matrices. *Bioinformatics* (Oxford, England) 2004;20(14):2222–7.
- Holmes I. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics* 2005;6:73.
- Horesh Y, et al. RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules. *BMC Bioinformatics* 2007;8:366. [PubMed: 17908318]
- Huang H, et al. RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. *Nucleic Acids Research* 2006;34(Web Server issue):W429–34. [PubMed: 16845041]
- Hu Y. GPRM: A genetic programming approach to finding common RNA secondary structure elements. *Nucleic Acids Research* 2003;31(13):3446–9. [PubMed: 12824343]
- Jacobs GH, et al. Transterm--extended search facilities and improved integration with other databases. *Nucleic Acids Research* 2006;34(Database issue):D37–40. [PubMed: 16381889]
- Ji Y, Xu X, Stormo GD. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics* (Oxford, England) 2004;20(10):1591–602.
- Katoh K, Toh H. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* 2008;9:212. [PubMed: 18439255]

- Kin T, Tsuda K, Asai K. Marginalized kernels for RNA sequence data analysis. *Genome Informatics International Conference on Genome Informatics* 2002;13:112–22. [PubMed: 14571380]
- Kiryu H, Kin T, Asai K. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics (Oxford, England)* 2007;23(4):434–41.
- Kiryu H, et al. Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics (Oxford, England)* 2007;23(13):1588–98.
- Klein RJ, Eddy SR. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics* 2003;4:44. [PubMed: 14499004]
- Knight R, Birmingham A, Yarus M. BayesFold: rational 2 degrees folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. *RNA (New York, NY)* 2004;10(9):1323–36.
- Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research* 2003;31(13):3423–8. [PubMed: 12824339]
- Lambert A, et al. Computing expectation values for RNA motifs using discrete convolutions. *BMC Bioinformatics* 2005;6:118. [PubMed: 15892887]
- Le, S.; Maizel, JV.; Zhang, K. An algorithm for detecting homologues of known structured RNAs in genomes. *Proceedings/IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference*; 2004. p. 300–10.
- Lestrade L, Weber MJ. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research* 2006;34(Database issue):D158–62. [PubMed: 16381836]
- Le SY, Zhang K, Maizel JV. A method for predicting common structures of homologous RNAs. *Computers and Biomedical Research, an International Journal* 1995;28(1):53–66.
- Lindgreen S, Gardner PP, Krogh A. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics (Oxford, England)* 2007;23(24):3304–11.
- Liu J, et al. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics* 2005;6:89. [PubMed: 15817128]
- Macke TJ, et al. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research* 2001;29(22):4724–35. [PubMed: 11713323]
- Matsui, H.; Sato, K.; Sakakibara, Y. Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures. *Proceedings/IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference*; 2004. p. 290–9.
- Meyer IM, Miklós I. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PloS Computational Biology* 2007;3(8):e149. [PubMed: 17696604]
- Mignone F, et al. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucl Acids Res* 2005;33(suppl1):D141–146. [PubMed: 15608165]
- Moretti S, et al. R-Coffee: a web server for accurately aligning noncoding RNA sequences. *Nucleic Acids Research* 2007;36(Web Server issue):W10–3. [PubMed: 18483080]
- Pavesi G, et al. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Research* 2004;32(10):3258–69. [PubMed: 15199174]
- Pedersen JS, et al. Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Computational Biology* 2006;2(4):e33. [PubMed: 16628248]
- Pesole G, Liuni S. Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. *Trends in Genetics: TIG* 1999;15(9):378. [PubMed: 10461207]
- Reeder J, Reeder J, Giegerich R. Locomotif: from graphical motif description to RNA motif search. *Bioinformatics (Oxford, England)* 2007;23(13):i392–400.
- Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2001;2:8. [PubMed: 11801179]
- Rocheleau L, Pelchat M. The Subviral RNA Database: a toolbox for viroids, the hepatitis delta virus and satellite RNAs research. *BMC Microbiology* 2006;6:24. [PubMed: 16519798]
- Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics (Oxford, England)* 2004;20(1):58–66.

- Sakakibara Y. Pair hidden Markov models on tree structures. *Bioinformatics* (Oxford, England) 2003;19 (Suppl 1):i232–40.
- Sakakibara Y, et al. Stem kernels for RNA sequence analyses. *Journal of Bioinformatics and Computational Biology* 2007;5(5):1103–22. [PubMed: 17933013]
- Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* (Oxford, England) 2005;21(16):3352–9.
- Steffen P, et al. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* (Oxford, England) 2006;22(4):500–3.
- Tabai Y, et al. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics* 2007;9:33. [PubMed: 18215258]
- Thébaud P, et al. Searching RNA motifs and their intermolecular contacts with constraint networks. *Bioinformatics* (Oxford, England) 2006;22(17):2074–80.
- Touzet H. Comparative analysis of RNA genes: the caRNAc software. *Methods in Molecular Biology* (Clifton, NJ) 2007;395:465–74.
- Veksler-Lublinsky I, et al. A structure-based flexible search method for motifs in RNA. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 2007;14(7):908–26. [PubMed: 17803370]
- Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America* 2005;102(7):2454–9. [PubMed: 15665081]
- Will S, et al. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology* 2007;3(4):e65. [PubMed: 17432929]
- Wilm A, Higgins DG, Notredame C. R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Research* 2007;36(9):e52. [PubMed: 18420654]
- Wilm A, Linnenbrink K, Steger G. ConStruct: Improved construction of RNA consensus structures. *BMC Bioinformatics* 2007;9:219. [PubMed: 18442401]
- Xie J, et al. Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs. *Nucleic Acids Research* 2007;35(Database issue):D183–7. [PubMed: 17099227]
- Xue C, Liu G. RScan: fast searching structural similarities for structured RNAs in large databases. *BMC Genomics* 2007;8:257. [PubMed: 17663795]
- Xu X, Ji Y, Stormo GD. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics* (Oxford, England) 2007;23(15):1883–91.
- Yao Z, Weinberg Z, Ruzzo WL. CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* (Oxford, England) 2006;22(4):445–52.
- Zhang S, et al. Searching genomes for noncoding RNA using FastR. *IEEE/ACM Transactions on Computational Biology and Bioinformatics/IEEE, ACM* 2005;2(4):366–79.
- Zhou Y, et al. GISSD: Group I Intron Sequence and Structure Database. *Nucleic Acids Research* 2007;36 (Database issue):D31–7. [PubMed: 17942415]

Table 1

Structural Motif Catalogs

1. Rfam	http://rfam.janelia.org/	2007
A comprehensive collection of non-coding RNA (ncRNA) families, represented by multiple sequence alignments and profile stochastic context-free grammars (compiled with INFERNA) that aims to facilitate the identification and classification of new members of known sequence families, and distributes annotation of ncRNAs in over 400 complete genome sequences. Allows querying against motif instances by EMBL ID or by de-novo search of upto 2kbp of sequence. References: (Griffiths-Jones et al. 2005)		
2. UTRSite	http://bigshot.ba.itb.cnr.it/UTR/	2008
A database of approximately 60 structural and sequential cis-regulatory functional motifs in RNA. Patterns are annotated with a modified version of the PatScan pattern definition syntax and is directly parseable only by UTRScan. References: (Mignone et al. 2005)		
3. UTRdb	http://www.ba.itb.cnr.it/UTR/	2006
A curated database of 5' and 3' untranslated sequences of eukaryotic mRNAs, derived from several sources of primary data that allows selection and extraction of UTR subsets based on their genomic coordinates and/or features of the protein encoded by the relevant mRNA (e.g. GO term, PFAM domain, etc.). Experimentally validated and predicted instances of UTRsite patterns are annotated and cross-linked. References: (Mignone et al. 2005)		
4. RegRNA	http://regna.mbc.nctu.edu.tw/	2006
RegRNA is an integrated web server for identifying the homologs of Regulatory RNA motifs and elements against an input mRNA sequence. Both sequence homologs or structural homologs of regulatory RNA motifs can be identified. References: (Hsi-Yuan Huang et al. 2006)		
5. TransTerm	http://uther.otago.ac.nz/Transterm.html	2007
An interactive database providing access to mRNA sequences and associated regulatory elements. The mRNA sequences are derived from all gene sequence data in Genbank, including complete genomes, divided into putative 5' UTRs and 3' UTRs, initiation and termination regions and the full CDS sequences. This data can be searched for defined regulatory elements. References: (Jacobs et al. 2006)		
6. RibEx	http://www.ibt.unam.mx/biocomputo/ribex.htm	2005
A web server capable of searching any sequence for known riboswitches as well as other predicted, but highly conserved, bacterial regulatory elements. It allows the visual inspection of the identified motifs in relation to attenuators and open reading frames (ORFs). Any of the ORF's or regulatory elements' sequence can be obtained with a click and submitted to NCBI's BLAST. Alternatively, the genome context of all other genes regulated by the same element can be explored with our genome context tool. Availability: web service, basic documentation, examples References: (Abreu-Goodger et al. 2004) (Abreu-Goodger & Merino 2005)		
7. EvoFold	http://www.cbse.ucsc.edu/~jsp/EvoFold/	2006
Phylogenetic stochastic context-free grammars for identifying functional RNAs were used to survey an eight-way genome-wide alignment of the human, chimpanzee, mouse, rat, dog, chicken, zebra-fish, and puffer-fish genomes for deeply conserved functional RNAs. The result was a large set of candidate RNA structures including many known functional RNAs such as miRNAs, histone 3'UTR stem loops, and various types of known genetic recoding elements. The new predictions include heretofore unknown members of known classes such as novel miRNAs and SECIS elements. References: (Pedersen et al. 2006)		
8. RNAz ncRNAs	http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/ncRNA/	2006
A comparative screen (using RNAz) of vertebrate genomes for structural noncoding RNAs, which evaluates conserved genomic DNA sequences for signatures of structural conservation of base-pairing patterns and exceptional thermodynamic stability predicted thousands of highly-conserved structured RNA elements. Only a small fraction of these sequences has been described previously but more than 40% of the predicted structured RNAs overlap with experimentally detected sites of transcription. References: (Washietl, Hofacker & Stadler 2005)		
9. NDB	http://ndbserver.rutgers.edu/	2008
The goal of the Nucleic Acid Database Project (NDB) is to assemble and distribute structural information about nucleic acids. The NDB processes data for the crystal structures of nucleic acids. Uses PDB formats. References: (Berman et al. 1992)		
10. MiRBase	http://microrna.sanger.ac.uk/	2008
miRBase Sequences is the primary online repository for miRNA sequence data and annotation. miRBase Targets is a comprehensive new database of predicted miRNA target genes. References: (Griffiths-Jones et al. 2006)		
11. RNAJunction	http://majunction.abcc.ncifcrf.gov/	2007
More than 12,000 extracted three-dimensional junction and kissing loop structures as well as detailed annotations for each. If you are interested in RNA as a building block for nano-scale design or if you are analyzing the properties of specific RNA motifs you should find utility in this site. The junctions in this database were extracted using a junction scanning algorithm from a number of structures from the Protein Data Bank. References: (Bindewald et al. 2008)		
12. SnoRNA-LBME-db	http://www-snoRNA.biotoul.fr/	2007
Dedicated database containing human C/D box and H/ACA box small nucleolar RNAs (snoRNAs), and small Cajal body-specific RNAs (scaRNAs) as well as the target sites of their predicted action. References: (Lestrade & Weber 2006)		
13. Sno/scaRNAbase	http://gene.fudan.sh.cn/snoRNAbase.nsf	2007
Provides an easy-to-use gateway to important sno/scaRNA features such as sequence motifs, possible functions, homologues, secondary structures, genomics organization, sno/scaRNA gene's chromosome location, and more. References: (Xie et al. 2007)		
14. SubViral Motifs	http://subviral.med.uottawa.ca/cgi-bin/motifs.cgi	2006
Provides secondary structures and sequences of ribozymes and other ncRNAs from viral genomes. References: (Rocheleau & Pelchat 2006)		
15. GISSD	http://www.rna.whu.edu.cn/gissd/	2008
Group I Intron Sequence and Structure Database (GISSD) is a specialized and comprehensive database for group I introns, including sequences, secondary structures, 3D structures, and internal CDSes where available for known and predicted members of the 14 Group I intron subgroups. References: (Y. Zhou et al. 2008)		

Table 2

Search Tools for Known Structural Motifs

1. INFERNAL	http://infernal.janelia.org/	2007
Infernal is an implementation of “covariance models” (CMs), which are statistical models of RNA secondary structure and sequence consensus. It is the primary tool used for the Rfam project. Give Infernal a multiple sequence alignment of a conserved structural RNA family, annotated with the consensus secondary structure. The “cmbuild” program builds a statistical profile of your alignment. That CM can be used as a query in a database search to find more homologs of your RNAs (the “cmsearch” program). The latest version also includes the QDB optimization algorithm. Availability: source, extensive documentation, examples References: (S R Eddy 2006)		
2. RSEARCH	http://selab.janelia.org/software.html#rsearch	2003
RSEARCH aligns an RNA query to target sequences, using SCFG algorithms to score both secondary structure and primary sequence alignment simultaneously. It's slow, but somewhat more capable of finding significant remote RNA structure homologies than sequence alignment methods like BLAST. Availability: source, extensive documentation References: (Klein & Sean R Eddy 2003)		
3. RNABOB	http://selab.janelia.org/software.html#rnabob	1996
Fast Pattern searching for RNA secondary structures. RNABOB is an implementation of D. Gautheret's RNAMOT, but with a different underlying algorithm using a nondeterministic finite state machine with node rewriting rules. An RNABOB motif is a consensus pattern a la PROSITE patterns, but with base-pairing. Availability: source, limited documentation References: none		
4. UTRScan	http://www.pesolelab.it/	1999
UTRScan is a web service for finding matches to all secondary structure patterns from the UTRSite database in a set of FASTA sequences. It is backed by the PatSearch program which is also provided as a web service for searching custom patterns against Fasta sequences or collected UTR sequences. Availability: web-services, basic documentation References: (G Pesole & S Liuni 1999)		
5. PatScan	http://www-unix.mcs.anl.gov/compbio/PatScan/	1997
PatScan is a pattern matcher which searches protein or nucleotide (DNA, RNA, tRNA etc.) sequence archives for instances of a pattern which you input. Pattern definition rules are provided. Availability: web-service, source, basic documentation References: (Dsouza, Larsen & Overbeek 1997)		
6. RegRNA	http://regrna.mbc.nctu.edu.tw/	2006
RegRNA is an integrated web server for identifying the homologs of Regulatory RNA motifs and elements against an input mRNA sequence. Both sequence homologs or structural homologs of regulatory RNA motifs can be identified. Availability: web service, basic documentation, examples References: (Hsi-Yuan Huang et al. 2006)		
7. TransTerm	http://uther.otago.ac.nz/Transterm.html	2007
An interactive database providing access to mRNA sequences and associated regulatory elements. The mRNA sequences are derived from all gene sequence data in Genbank, including complete genomes, divided into putative 5' UTRs and 3' UTRs, initiation and termination regions and the full CDS sequences. This data can be searched for defined regulatory elements. Availability: web service, extensive documentation, examples References: (Jacobs et al. 2006)		
8. RibEx	http://www.ibt.unam.mx/biocomputo/ribex.htm	2005
A computational approach that identifies regulatory elements conserved across phylogenetically distant organisms. Intergenic regulatory regions were clustered by orthology of the adjacent genes, and an iterative process was applied to search for significant motifs, enabling new elements of the putative regulon to be added in each cycle. With this approach, we identified highly conserved riboswitches and the Gram positive T-box. Interestingly, we identified many other regulatory systems that appear to depend on conserved RNA structures. Availability: web service, basic documentation, examples References: (Abreu-Goodger et al. 2004) (Abreu-Goodger & Merino 2005)		
9. Locomotif	http://bibiserv.techfak.uni-bielefeld.de/locomotif/	2007
Locomotif: Localization of RNA motifs with generated thermodynamic matchers. Its GUI-based program that allows for the visual design of RNA motifs. The graphical structures are then translated into executable programs to be used for searching a motif in a sequence (plain text or FASTA format). Availability: JavaWS GUI, extensive documentation References: (Janina Reeder, Jens Reeder & Giegerich 2007)		
10. PHMMTS	http://phmmts.dna.bio.keio.ac.jp/	2004
PHMMTS provides a unifying framework and an automata-theoretic model for alignments of trees, structural alignments and pair stochastic context-free grammars. By structural alignment, we mean a pairwise alignment to align “an unfolded RNA sequence” into “an RNA sequence of known secondary structure”. PHMMTS takes a folded RNA sequence and searches for a structural alignment to it in an unfolded sequence. Availability: web service, sources, basic documentation References: (Sakakibara 2003)		
11. Stem Kernel	http://stem-kernel.dna.bio.keio.ac.jp/	2004
Several computational methods based on stochastic context-free grammars have been A novel kernel function, stem kernel, for the discrimination and detection of functional RNA sequences using support vector machines (SVMs) is proposed. The stem kernel is tailored to measure the similarity of two RNA sequences from the viewpoint of secondary structures. The stem kernels are then applied to discriminate members of an RNA family from nonmembers using SVMs. The study indicates that the discrimination ability of the stem kernel is strong compared with conventional methods. Availability: sources, limited documentation Reference: (Sakakibara et al. 2007)		
12. PSTAG	http://pstag.dna.bio.keio.ac.jp/	2004
This software provides an implementation of the “pair stochastic tree adjoining grammars (PSTAGs)” for modeling “pseudoknot” RNA structures, which is an extension of the “pair hidden Markov models on tree structures (PHMMTs)”. Used to align and predict RNA secondary structures including pseudoknots in unfolded sequences using a folded sequence. Availability: web service, sources, basic documentation Reference: (Matsui, Sato & Sakakibara 2004)		
13. RNAinverse	http://rna.tbi.univie.ac.at/	2000

RNAinverse searches for sequences folding into a predefined structure, thereby inverting the folding algorithm. Target structures (in bracket notation) and starting sequences for the search are read alternately from stdin. For each search the best sequence found and its Hamming distance to the start sequence are printed to stdout.

Availability: web service, sources, full documentation, examples

Reference: (Hofacker 2003) (Hofacker 2004)

14. **HomoSTRscan** <http://protein3d.ncifcrf.gov/shuyun/homostrscan.html> 2004

Homologous Structural RNA Scan: A program for discovering homologous RNAs in complete genomes by taking a single RNA sequence with its secondary structure. It takes account of information of both the primary sequence and the secondary structural constraints of the query RNA in detail including each base-pairs in the duplexes and each nucleotide in the single strand. The homologous RNA structures are strictly inferred from a robust statistical distribution of a quantitative measure, maximal similarity score of RNA structures.

Availability: sources, basic documentation, examples

References: (Shu-Yun Le, Jacob V Maizel & Kaizhong Zhang 2004)

15. **RNAmotif** <http://www.scripps.edu/mb/casegr-sh-3.5.html> 2008

The rnamotif program searches a database for RNA sequences that match a "motif" describing secondary structure interactions. A match means that the given sequence is capable of adopting the given secondary structure, but is not intended to be predictive. Matches can be ranked by applying scoring rules that may provide finer distinctions than just matching to a profile. It is an extension of RNAMOT and RNABOB.

Availability: sources, extensive documentation

References: (Macke et al. 2001)

16. **ERPIN** <http://tagc.univ-mrs.fr/erpin/> 2005

Easy RNA Profile IdentificationN is an RNA motif search program. Unlike most RNA pattern matching programs, ERPIN does not require users to write complex descriptors before starting a search. Instead ERPIN reads a sequence alignment and secondary structure, and automatically infers a statistical "secondary structure profile" (SSP). An original Dynamic Programming algorithm then matches this SSP onto any target database, finding solutions and their associated scores. Web service allows search with precompiled RNA motifs.

Availability: web service, sources, binaries, full documentation, examples

References: (D Gautheret & A Lambert 2001) (Andre Lambert et al. 2005)

17. **RSmatch/RADAR** <http://datalab.njit.edu/biodata/rna/RSmatch/server.htm> 2007

RADAR can align structure-annotated RNA sequences so that both sequence and structure information are taken into consideration. This server is capable of performing database search, multiple structure alignment, and pair-wise structure comparison. In addition, RADAR provides two salient features: (1) constrained alignment of RNA secondary structures, and (2) prediction of the consensus structure for a set of RNA sequences.

Availability: web service, binaries, full documentation, examples

References: (J. Liu et al. 2005)

18. **Rscan** <http://bioinfo.au.tsinghua.edu.cn/member/cxue/rscan/RScan.htm> 2007

RScan is designed to quickly find structural similarities for a query sequence with known or predicted secondary structure in a genomic database. The input format of a structured query is the output of Vienna's RNAfold.

Availability: sources, basic documentation, examples

References: (Xue & G. Liu 2007)

19. **INFO-RNA** <http://www.bioinf.uni-freiburg.de/Software/INFO-RNA/start.html> 2007

We consider the inverse RNA folding problem, which is the design of RNA sequences that fold into a desired structure. Given a set of base pairs, we aim at finding an RNA sequence that is going to adopt these pairs. Additionally, restrictions on the sequence level can be specified by constraints given in IUPAC symbols. The resulting sequences could be used in a BLAST or related sequence search.

Availability: web service, full documentation

References: (Busch & Backofen 2006)

20. **FastR/PfastR** <http://ribozyme.ucsd.edu/fastr/> 2007

Given an RNA sequence with a known secondary structure, efficiently compute all structural homologs (computed as a function of sequence and structural similarity) in a genomic database. Structural filters that eliminate a large portion of the database allow us to search a typical bacterial database in minutes on a standard PC with high sensitivity and specificity. The web service allows querying of input sequences against a number of RNA structure profiles.

Availability: web service, limited documentation

References: (Bafna & Shaojie Zhang 2004) (Shaojie Zhang et al. 2005)

21. **MilPat** <http://cat.toulouse.inra.fr/~rnaworld/MilPat/MilPat.pl> 2006

Searching for RNA structure profiles in target sequences according to a constraint network model. Also allows searching for target sequences in interaction with RNA structures.

Availability: web service, basic documentation

References: (Thebault et al. 2006)

22. **STRMS** <http://www.cs.bgu.ac.il/~vaksler/STRMS.htm> 2006

STRMS is an RNA motif search tool. Prefolds the target RNA sequences using an mfold based approach and converts them into structure trees. Then takes a query structure, builds a tree-structure from it and runs tree-alignment of the query tree against the target database.

Availability: sources, basic documentation, examples

References: (Veksler-Lubinsky et al. 2007)

23. **RNAMST** <http://bioinfo.csie.ncu.edu.tw/~rnamst/> 2006

RNAMST is an efficient and flexible RNA Motif Search Tool for RNA structural homologs. RNAMST web server accepts four different kinds of input formats to facilitate user to describe a RNA structure easily. Beside, several databases are provided and have been processed by our algorithm. Therefore, the user can easily and quickly search the RNA structural homologs against the huge amount of sequences. In addition, RNAMST is able to search structure with asymmetric mismatch and bulges that makes the search more comprehensive and practical.

Availability: web service, full documentation

References: (Chang et al. 2006)

Table 3

Consensus Structures in Aligned Sequences

1. Inferral	http://inferral.janelia.org/	2007
Inferral is an implementation of “covariance models” (CMs), which are statistical models of RNA secondary structure and sequence consensus. It is the primary tool used for the Rfam project. Give Inferral a multiple sequence alignment of a conserved structural RNA family, annotated with the consensus secondary structure. The “cmbuild” program builds a statistical profile of your alignment. That CM can be used as a query in a database search to find more homologs of your RNAs (the “cmsearch” program). The latest version also includes the QDB optimization algorithm. Pre-aligned input. Availability: source, extensive documentation, examples References: (S R Eddy 2006)		
2. RNAz	http://www.tbi.univie.ac.at/~wash/RNAz/	2006
RNAz is a program for predicting structurally conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments. It can be used in genome wide screens to detect functional RNA structures, as found in noncoding RNAs and cis-acting regulatory elements of mRNAs. Pre-aligned input. Availability: sources, windows binary, extensive documentation, examples References: (Washietl, Hofacker & Stadler 2005)		
3. EvoFold	http://www.cbse.ucsc.edu/~jsp/EvoFold/	2004
EvoFold is a comparative method for identifying functional RNA structures in multiple-sequence alignments. It is based on a probabilistic model-construction called a phylo-SCFG and exploits the characteristic differences of the substitution process in stem-pairing and unpaired regions to make its predictions. Each prediction consists of a specific secondary structure and a folding potential score. Pre-aligned input. Availability: linux binary, basic documentation, static results References: (Pedersen et al. 2006)		
4. ddbRNA	http://dibernardo.tigem.it/wiki/index.php/DdbRNA	2003
An algorithm able to detect conserved secondary structures in both pairwise and multiple DNA sequence alignments with computational time proportional to the square of the sequence length. Pre-aligned input. Availability: cross-platform jar, limited documentation References: (di Bernardo, Down & Hubbard 2003)		
5. RNAalifold (Vienna)	http://rna.tbi.univie.ac.at/	2007
RNAalifold reads aligned RNA sequences from stdin or <i>file.aln</i> and calculates their minimum free energy (mfe) structure, partition function (pf) and base pairing probability matrix. Currently, the input alignment has to be in CLUSTAL format. It returns the mfe structure in bracket notation, its energy, the free energy of the thermodynamic ensemble and the frequency of the mfe structure in the ensemble. Pre-aligned input. Availability: web service, sources, full documentation, examples References: (Hofacker 2007)		
6. SCA	http://rna.tbi.univie.ac.at/cgi-bin/SCA.cgi	2008
Allows the use of a variety of methods including distance between individually folded structures, global minimum free energies, and folding space searches. Wraps RNAalifold, RNAdistance, RNApdist and other algorithms using cost, distance, and clustering methods to approximate a conserved structure. Pre-aligned input. Availability: web service, limited documentation, examples References: none		
7. QRNA	http://selab.janelia.org/software.html	2003
A prototype noncoding RNA gene finder, based on comparative genome sequence analysis. QRNA uses comparative genome sequence analysis to detect conserved RNA secondary structures, including both ncRNA genes and cis-regulatory RNA structures. Pre-aligned input. Availability: sources, extensive documentation, examples References: (Rivas & S R Eddy 2001)		
8. McCaskill-MEA	http://www.ncrna.org/papers/McCaskillMEA/	2005
The McCaskill-MEA method first computes the base pairing probability matrices for all the sequences in the alignment and then obtains the base pairing probability matrix of the alignment by averaging over these matrices. The consensus secondary structure is predicted from this matrix such that the expected accuracy of the prediction is maximized. We show that the McCaskill-MEA method performs better than other methods, particularly when the alignment quality is low and when the alignment consists of many sequences. Pre-aligned input. Availability: sources, limited documentation, examples, benchmarks References: (Kiryu, Kin & Asai 2007)		
9. ERPIN	http://tagc.univ-mrs.fr/erpin/	2006
Unlike most RNA pattern matching programs, ERPIN does not require users to write complex descriptors before starting a search. Instead ERPIN reads a sequence alignment and secondary structure, and automatically infers a statistical “secondary structure profile” (SSP). An original Dynamic Programming algorithm then matches this SSP onto any target database, finding solutions and their associated scores. In the latest version (unpublished) Erpin computes E-values for matches. Pre-aligned input. Availability: web service, sources, full documentation, examples References: (D Gautheret & A Lambert 2001) (Andre Lambert et al. 2005)		
10. MSARI	http://groups.csail.mit.edu/cb/MSARi/	2004
A highly accurate method for identifying genes with conserved RNA secondary structure by searching multiple sequence alignments of a large set of candidate orthologs for correlated arrangements of reverse-complementary regions. This approach is growing increasingly feasible as the genomes of ever more organisms are sequenced. A program called msari implements this method and is significantly more accurate than existing methods in the context of automatically generated alignments, making it particularly applicable to high-throughput scans. Subsequently lists RNAz and Pfold as more accurate. Pre-aligned input. Availability: sources, limited documentation, examples References: (Coventry, Kleitman & Berger 2004)		
11. Pfold	http://www.daimi.au.dk/~compbio/rnafold/	2003
A practical way of predicting RNA secondary structure that is especially useful when related sequences can be obtained. The method improves a previous algorithm based on an explicit evolutionary model and a probabilistic model of structures. Pre-aligned input. Availability: web service, limited documentation, examples References: (Knudsen & Hein 2003)		
12. ILM	http://www.cse.wustl.edu/~zhang/projects/rna/ilm/	2003
Iterative loop matching (ILM) is an extended dynamic programming algorithm that is able to predict RNA secondary structures including pseudoknots. ILM can not only predict consensus structures for aligned homologous sequences, using combined thermodynamic and covariance scores, but can also be applied to individual sequences, using thermodynamic information alone. Pre-aligned input.		

Availability: sources, limited documentation, examples
References: (Ruan, Stormo & Weixiong Zhang 2004)

13. **BayesFold** <http://bayes.colorado.edu/Bayes/> 2003

BayesFold is a web application that finds, ranks, and draws the likeliest structures for a sequence alignment. Foldings are based on the predictions of the Bayesian statistical method. BayesFold provides convenient structure comparison and formatting functionality, and produces publication-quality graphics. Pre-aligned input.

Availability: web service (MSIE only), extensive documentation

References: (Knight, Birmingham & Yarus 2004)

14. **KnetFold** <http://knetfold.abcc.ncifcrf.gov/> 2006

KNetFold is a new software for predicting the consensus RNA secondary structure for a given alignment of nucleotide sequences. It uses an innovative classifier system (a hierarchical network of k -nearest neighbor classifiers) to compute for each pair of alignment positions a "base pair" or "no base pair" prediction. We evaluated the accuracy of the KNetFold algorithm with a set of 49 RNA sequence alignments obtained from the RFAM database. In our recent publication, we show that for this test set, the performance of the method is higher compared to the programs PFOLD and RNAalifold. Pre-aligned input

Availability: web service, sources, basic documentation, examples

References: (Bindewald & Shapiro 2006)

15. **ConStruct** <http://www.biophys.uni-duesseldorf.de/construct3/> 2008

ConStruct is an RNA alignment editor and consensus structure prediction tool. It combines multiple sequence alignment, thermodynamic structure prediction and statistics in a semiautomatical fashion. Its sophisticated GUI guides the user through correcting an initial sequence alignment with respect to a consensus structure. Its built-in structure prediction routines allow for optimal secondary structures, suboptimal secondary structures and also tertiary interactions e.g. pseudoknots. Pre-aligned input

Availability: sources, debian package, extensive documentation, examples

References: (Wilm, Linnenbrink & Steger 2008)

16. **SimulFold** <http://www.cs.ubc.ca/~irmtraud/simulfold/> 2007

SimulFold 1.0 is a computer program for co-estimating an RNA structure including pseudo-knots, a multiple-sequence alignment and an evolutionary tree, given a set of evolutionarily related RNA sequences as input. In other words, you give SimulFold an initial alignment of RNA sequences as input and it will predict a consensus RNA structure which may include pseudo-knots while simultaneously estimating the sequence alignment and the evolutionary tree relating the RNA sequences. Pre-aligned input.

Availability: sources, limited documentation, examples

References: (Meyer & Miklos 2007)

17. **STRAL** <http://www.biophys.uni-duesseldorf.de/stral/> 2006

StrAl is an alignment tool designed to provide multiple alignments of non-coding RNAs following a fast progressive strategy. It combines the thermodynamic base pairing information derived from RNAfold calculations in the form of base pairing probability vectors with the information of the primary sequence. Thus the scoring system is composed of two major parts evaluating the given structural and the sequence information, respectively. Pre-aligned input.

Availability: web service, sources, benchmarks

References: (Dalli et al. 2006)

18. **R-Coffee/RM-Coffee** <http://www.tcoffee.org/> 2000

R-Coffee is a multiple RNA alignment package, derived from T-Coffee, designed to align RNA sequences while exploiting secondary structure information. R-Coffee uses an alignment-scoring scheme that incorporates secondary structure information within the alignment. It works particularly well as an alignment improver and can be combined with any existing sequence alignment method. Uses any of a number of sequence aligners alongside pairwise structure aligners incorporating a novel score-improving scheme. Alignment step first.

Availability: web service, limited documentation

References: (Wilm, Higgins & Notredame 2008) (Moretti et al. 2008)

Table 4

Consensus Structures in Unaligned Sequences

1. RNAShapes	http://bibiserv.techfak.uni-bielefeld.de/rnashapes/	2008
RNA shape abstraction maps structures to a tree-like domain of shapes, retaining adjacency and nesting of structural features, but disregarding helix lengths. Shape abstraction integrates well with dynamic programming algorithms, and hence it can be applied during structure prediction rather than afterwards. This avoids exponential explosion and can still give us a non-heuristic and complete account of properties of the molecule's folding space. RNAShapes offers three powerful RNA analysis tools in one single software package: Computation of a small set of representative structures of different shapes, complete in a well-defined sense; Computation of accumulated shape probabilities; Comparative prediction of consensus structures, as an alternative to the over-expensive Sankoff Algorithm.		
Availability: SOAP service, sources, binaries, full documentation, examples		
References: (Steffen et al. 2006)		
2. RNAmine	http://rnamine.ncrna.org/	2006
Frequent stem pattern miner from unaligned RNA sequences (RNAmine) is a software tool to extract the structural motifs from a set of RNA sequences. The potential secondary structures of the RNA sequences are represented by directed labeled graphs with label taxonomy, and the common secondary structures are extracted by using graph mining technique. RNAmine is used for motif finding, cluster detection and common secondary structure prediction from a set of RNA sequences.		
Availability: web service, limited documentation, benchmarks		
References: (Hamada et al. 2006)		
3. MX-SCARNA	http://mxscarna.ncrna.org/	2008
MXSCARNA (Multiplex Stem Candidate Aligner for RNAs) is a multiple alignment tool for RNA sequences using progressive alignment based on pairwise structural alignment algorithm of SCARNA. This software is fast enough for large scale analyses, while the accuracies of the alignments are better than or comparable with the existing algorithms which are computationally much more expensive in time and memory.		
Availability: web service, sources, full documentation, benchmarks		
References: (Tabei et al. 2008)		
4. SOKOS/CAN	http://www.cbrc.jp/sokos/	2007
An experimental implementation of stochastic or probabilistic context-free grammar (SCFG) for RNA sequence analysis with capability of computing the marginalizes count kernel which is a metric similarity between two RNA sequences. The similarity takes into account of potential RNA secondary structures of the RNA's. SOKOS/CAN can be used for generic RNA sequence analysis including secondary structure prediction and homology search.		
Availability: sources, basic documentation, examples, benchmarks		
References: (Kin, Tsuda & Asai 2002)		
5. RNAGA	http://protein3d.ncifcrf.gov/shuyun/rnaga.html	2003
A program for predicting a secondary structure common to a number of phylogenetically related sequences without the need for pre-aligned RNA sequences. One of the remarkable features of RNAGA is that RNA secondary structures are automatically optimized by not only the free energy of the formation of the structure but also the structural similarity among homologous sequences.		
Availability: web service, sources, full documentation, examples		
References: (S Y Le, K Zhang & J V Maizel 1995)		
6. Carnac	http://bioinfo.lifl.fr/carnac/	2004
Predicts if sequences share a common secondary structure. When this structure exists, Carnac is then able to correctly recover a large amount of the folded stems. The input is a set of single-stranded RNA sequences that need not to be aligned. The folding strategy relies on a thermodynamic model with energy minimization. It combines information coming from locally conserved elements of the primary structure and mutual information between sequences with covariations too.		
Availability: web service, sources, basic documentation, examples		
References: (Touzet 2007)		
7. comRNA	http://ural.wustl.edu/~yji/comRNA/	2004
The algorithm applies graph-theoretical approaches to automatically detect common RNA secondary structure motifs in a group of functionally or evolutionarily related RNA sequences. The advantages of this method are that it: does not require the presence of global sequence similarities (but can take advantage of it); does not require prior structural alignment; and is able to detect pseudoknot structures. It finds sets of stable stems conserved across multiple sequences, and assembles compatible conserved stems to form consensus secondary structure motifs.		
Availability: sources, basic documentation, examples, benchmarks		
References: (Ji, Xu & Stormo 2004)		
8. GPRM	http://bioinfo.life.nctu.edu.tw/tools.php	2003
GPRM is aimed at finding common secondary structure elements, not a global alignment, in a sufficiently large family (e.g. more than 15 members) of unaligned RNA sequences. It is not applicable to finding the possible folding of a single sequence. Besides, owing to the hardware limitation of our current PC server, GPRM is currently limited to finding structure elements with no more than five stems.		
Availability: web service (server unreachable)		
References: (Y. Hu 2003)		
9. CMfinder	http://bio.cs.washington.edu/yzizhen/CMfinder/	2005
CMfinder is a RNA motif prediction tool. It is an expectation maximization algorithm using covariance models for motif description, carefully crafted heuristics for effective motif search, and a novel Bayesian framework for structure prediction combining folding energy and sequence covariation. This tool performs well on unaligned sequences with long extraneous flanking regions, and in cases when the motif is only present in a subset of sequences. CMfinder also integrates directly with genome-scale homology search, and can be used for automatic refinement and expansion of RNA families.		
Availability: web service, sources, basic documentation, examples, benchmarks		
References: (Yao, Weinberg & Ruzzo 2006)		
10. RNAProfile	http://www.pesolelab.it/	2005
We present an algorithm that takes as input a set of unaligned RNA sequences expected to share a common motif, and outputs the regions that are most conserved throughout the sequences, according to a similarity measure that takes into account both the sequence of the regions and the secondary structure they can form according to base-pairing and thermodynamic rules. Only a single parameter is needed as input, which denotes the number of distinct hairpins the motif has to contain. No further constraints on the size, number and position of the single elements comprising the motif are required.		
Availability: sources, basic documentation		
References: (Pavesi et al. 2004)		
11. RNA Sampler	http://ural.wustl.edu/~xingxu/RNASampler	2008
The algorithm applies a probabilistic sampling approach and combines intrasequence base pairing probabilities and intersequence base alignment probabilities to prediction consensus structure on two sequences. It is extended by using a consistency-based method to incorporate pairwise structural information to predict the common structure conserved among multiple sequences.		

Availability: sources, basic documentation, benchmarks

References: (Xu, Ji & Stormo 2007)

12. **RNAspa** <http://faculty.biu.ac.il/~unger/RNAspa/> 2007

We developed the RNAspa program, which comparatively predicts the secondary structure for a set of ncRNA molecules in linear time in the number of molecules. We observed that in a list of several hundred suboptimal minimal free energy (MFE) predictions, as provided by the RNAsubopt program of the Vienna package, it is likely that at least one suggested structure would be similar to the true, correct one. The suboptimal solutions of each molecule are represented as a layer of vertices in a graph. The shortest path in this graph is the basis for structural predictions for the molecule.

Availability: sources, limited documentation, benchmarks

References: (Horesh et al. 2007)

13. **X-INS-I/Q-INS-I** <http://align.bmr.kyushu-u.ac.jp/mafft/software/source65.html> 2008

Part of MAFFT. Methods are suitable for a global alignment of highly diverged ncRNA sequences. Q-INS-i: Applicable to up to <200 sequences, <1,000 nt; Uses the Four-way Consistency objective function (Katoh and Toh, submitted) for incorporating structural information. X-INS-i: Applicable to up to <50 sequences, <1,000 nt; a framework based on the Four-way Consistency objective function to build a multiple structural alignment by combining pairwise structural alignments given by an external program. At present, the external program can be selected from MXSCARNA, LaRA and FOLDALIGN (the local and global options).

Availability: sources, limited documentation, benchmarks

References: (Katoh & Toh 2008)

14. **RAF** <http://contra.stanford.edu/contrafold/> 2008

An efficient algorithm for simultaneous alignment and consensus folding of unaligned RNA sequences. Algorithmically, RAF exploits sparsity in the set of likely pairing and alignment candidates for each nucleotide (as identified by the CONTRAfold or CONTRAlign programs) to achieve an effectively quadratic running time for simultaneous pairwise alignment and folding. RAF's fast sparse dynamic programming, in turn, serves as the inference engine within a discriminative machine learning algorithm for parameter estimation

Availability: yet-to-be-published

References: (Do, Foo & Batzoglou 2008)

15. **StemLoc** <http://biowiki.org/StemLoc> 2005

An SCFG based algorithm using "alignment envelopes" and "fold envelopes" to simultaneously constrain both the alignment and the secondary structures of the sequences being compared and make the Sankoff algorithm tractable. Stemloc moves from all vs all pairwise alignments into global or local focused searches for alignments/structures.

Availability: sources, full documentation, examples

References: (Holmes 2005)

16. **MASTR** <http://servers.binf.ku.dk/mastr/> 2008

Using Markov chain Monte Carlo in a simulated annealing framework, the algorithm MASTR (Multiple Alignment of STructural RNAs) iteratively improves both sequence alignment and structure prediction for a set of RNA sequences. This is done by minimizing a combined cost function that considers sequence conservation, covariation and basepairing probabilities. The results show that the method is very competitive to similar programs available today, both in terms of accuracy and computational efficiency. Unaligned input.

Availability: web service, sources, limited documentation, examples

References: (Lindgreen, Gardner & Krogh 2007)

17. **Murlet** <http://software.ncrna.org/> 2005

A variant of the Sankoff algorithm that uses an efficient scoring system to reduce the time and space requirements considerably without compromising on the alignment quality. First, our algorithm computes the match probability matrix that measures the alignability of each position pair between sequences as well as the base pairing probability matrix for each sequence. These probabilities are then combined to score the alignment using the Sankoff algorithm. By itself, our algorithm does not predict the consensus secondary structure of the alignment but uses external programs for the prediction. Unaligned input.

Availability: web service, sources, basic documentation, benchmarks

References: (Kiryu et al. 2007)

18. **Pmcomp/pmmulti** <http://www.tbi.univie.ac.at/RNA/PMcomp/> 2004

pmcomp and pmmulti are two programs to perform pairwise and progressive multiple alignments of RNA sequences. pmcomp is a variant of Sankoff's algorithm for simultaneous folding and alignment, which takes as input pre-computed base pair probability matrices from McCaskill's algorithm as produced by RNAfold -p. Thus the method can also be viewed as way to compare base pair probability matrices. pmmulti is a simple wrapper program that does progressive multiple alignments by repeatedly calling pmcomp.

Availability: sources, limited documentation, examples

References: (Hofacker, Bernhart & Stadler 2004)

19. **LocARNA** <http://www.bioinf.uni-freiburg.de/Software/LocARNA/> 2007

A structure-based clustering approach that is capable of extracting putative RNA classes from genome-wide surveys for structured RNAs. The LocARNA (local alignment of RNA) tool implements a novel variant of the Sankoff algorithm that is sufficiently fast to deal with several thousand candidate sequences. The method is also robust against false positive predictions, i.e., a contamination of the input data with unstructured or nonconserved sequences.

Availability: web service, sources, examples, benchmarks

References: (Will et al. 2007)

20. **MARNA** <http://biwww2.informatik.uni-freiburg.de/Software/MARNA/> 2005

MARNA is a multiple alignment of RNA s taking into consideration both the primary sequence and the secondary structure. It is based on pairwise comparisons using costs of edit operations. The edit operations can be divided into edit operations on arcs and edit operations on bases. Additionally, MARNA predicts a consensus sequence as well as a consensus structure.

Availability: web service, sources, limited documentation, examples

References: (Siebert & Backofen 2005)

21. **Seed** <http://bio.site.uottawa.ca/software/seed/> 2006

Program takes as input a set of unaligned RNA sequences and produces a set of secondary structure motifs. Suffix arrays are used to enumerate complementary regions, possibly containing interior loops, as well for matching RNA secondary structure expressions.

Availability: sources, limited documentation

References: (Anwar, Nguyen & Turcotte 2006)

22. **lara** <https://www.mi.fu-berlin.de/w/LiSA/Lara> 2008

lara ("lagrangian relaxed structural alignment") is a tool for the sequence-structure alignment of RNA sequences. It employs methods from combinatorial optimization to compute feasible solutions for an integer linear program. lara computes all pairwise sequence-structure alignments of the input sequences and passes this information on to T-Coffee which computes a multiple sequence-structure alignment given the pairwise alignments. This is in contrast to plara where we compute a multiple sequence-structure alignment in a progressive fashion.

Availability: web service (at T-Coffee), sources, full documentation

References: (Bauer, Klau & Reinert 2007)

Table 5

Benchmark Data for Consensus Structure Prediction

1. TUTR	http://ribonomics.albany.edu/	2008
To facilitate the development, evaluation, and training of new software programs that identify RNA motifs, we created the UAlbany training UTR (TUTR) database, which is a collection of validated sets of sequences containing experimentally defined regulatory motifs. Presently, eleven training sets have been generated with associated indexes and “answer sets” provided that identify where the previously characterized RNA motif [the iron responsive element (IRE), AU-rich class-2 element (ARE), selenocysteine insertion sequence (SECIS), etc.] resides in each sequence. For each training set, control sets of corresponding size are also provided that are composed not of random sequences but rather sequences from the concatenated space of genomic UTRs.		
References: (Doyle et al. 2008)		
2. BRAlIBase	http://people.binf.ku.dk/pgardner/bralibase/	2004
A dataset for benchmarking secondary structure prediction algorithms. We have compiled RNA sequence alignments consisting of up to 11 sequences derived from reliable sources. These have been used to test several RNA analysis packages. Each alignment contains at least one reference sequence with (preferably) an experimentally verified secondary structure. Experimental verification of a structure may be from a variety of sources: x-ray crystallography, NMR, enzymatic structure probing or phylogenetic inference. Most new algorithms have already been benchmarked using this set.		
References: (Gardner & Giegerich 2004)		