Nucleic Acids Research, 2007, Vol. 35, Web Server issue **W335–W338**
doi:10.1093/nar/gkm222

The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures

Andreas R. Gruber, Richard Neuböck, Ivo L. Hofacker and Stefan Washietl*

Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria

Received January 29, 2007; Revised March 20, 2007; Accepted March 28, 2007

ABSTRACT

Many non-coding RNA genes and *cis*-acting regulatory elements of mRNAs contain RNA secondary structures that are critical for their function. Such functional RNAs can be predicted on the basis of thermodynamic stability and evolutionary conservation. We present a web server that uses the RNAz algorithm to detect functional RNA structures in multiple alignments of nucleotide sequences. The server provides access to a complete and fully automatic analysis pipeline that allows not only to analyze single alignments in a variety of formats, but also to conduct complex screens of large genomic regions. Results are presented on a website that is illustrated by various structure representations and can be downloaded for local view. The web server is available at: rna.tbi.univie.ac.at/RNAz.

INTRODUCTION

Functional RNA structures play an important role in a variety of cellular processes. They can be found in independent non-coding RNAs (ncRNAs or 'RNA genes') as well as in untranslated regions of mRNAs (1,2). Defined RNA secondary structures in ncRNAs may be required for specific interactions with proteins as part of a ribonucleoprotein complex (e.g. the signal recognition particle), or for protein interaction during the maturation process of the RNA (e.g. microRNAs interacting with DICER). Also direct catalytic activity of RNAs is possible (e.g. type I and II self-splicing introns). RNA structures in untranslated regions of mRNAs often serve regulatory roles and form interaction partners for proteins (e.g. iron responsive element interacting with IRP1) or small ligands (e.g. riboswitches in bacteria).

Such functional RNAs have moved into focus of interest during the past years and computational techniques for their prediction have become more and more important (3). Methods for prediction of secondary structure models from primary sequence have a long tradition and are readily available (4–6). However, any nucleotide sequence can be folded into a secondary structure using these programs. The challenge is to discriminate real functional RNA structures from random structures. There is general consensus in the community, that the structure information contained in a single sequence is not sufficient to yield reasonable prediction accuracies (7,8). The observation that many functional RNA structures are conserved in evolution, and the massive availability of comparative sequence data, has lead the efforts to predict functional RNAs into a clear direction: a series of programs have been developed recently that try to detect 'structurally conserved' RNAs (9–15).

Although the problem still remains challenging especially when scanning large genomes (3,13,16,17), such programs represent an important addition to today's arsenal of sequence analytic methods. In this article, we describe a web server that scans multiple sequence alignments for functional RNAs using the RNAz algorithm [12].

THE RNAz ALGORITHM

RNAz predicts functional RNA structures based on two criteria: (i) structural conservation, (ii) thermodynamic stability. It first predicts a consensus secondary structure using the RNAalifold algorithm (18). This is essentially an extension of standard minimum free energy folding algorithms with the constraint that all sequences have to fold into a common structure. Compensatory/consistent mutations, i.e. mutations that preserve a secondary structure, are incorporated as 'bonus' into the energy model, while inconsistent mutations are penalized. RNAz measures structural conservation by calculating the ratio of this consensus folding energy to the unconstrained folding energies of the single sequences.

In addition, RNAz calculates a stability score for the sequences in the alignment because functional RNAs are known to be thermodynamically more stable than random

*To whom correspondence should be addressed. Tel: +43-1-4277-52744; Fax: +43-1-4277-52793; Email: wash@tbi.univie.ac.at sequences (8,19). Stability is measured as normalized *z*-score of folding energies. It indicates how many SDs a given sequence is more/less stable than expected for random sequences of the same length and base composition.

Finally, an alignment is classified as 'functional RNA' or 'other' based on these two characteristic measures. A support vector machine learning technique which calculates an optimal combination of both scores is used for this purpose. Details of the algorithm can be found in reference (12).

THE RNAz WEB SERVER

The design of the web server was guided by three main goals: (i) minimizing the burden of manual pre-processing and formatting of the input data, (ii) providing a reasonable analysis pipeline that, on the one hand, works 'out of the box' but, on the other hand, can also be customized by the user and (iii) providing reasonable output for humans (e.g. graphical visualization, overview tables) and computers (e.g. annotation files, raw RNAz text output).

The web server operates in two different modes: in 'Standard Analysis' mode, usually one single alignment is analyzed. In this mode, it is also possible to analyze more than one alignment in one session, but the alignments are treated to be independent from each other. In 'Genomic screen' mode, a large number of alignments covering a genomic region can be screened and the results from all alignments are integrated in the end.

In the following, we describe general features of the server that apply to both modes of operation. Special requirements and features of the 'Genomic screen' mode is described in Section 'Conducting genomic screens'. Detailed instructions how to use the server is available as online help.

Uploading sequence alignments

Multiple sequence alignments can be provided by cut-and-paste or uploaded as file (Figure 1A). The server currently can read the following alignment formats: CLUSTALW, FASTA, PHYLIP, NEXUS, MAF and XMFA. Alignments can be generated by any sequence-based alignment program (see (20) and (21) for comparison of different programs on structural RNAs). However, one should not use 'structurally enhanced' alignments generated by programs that consider RNA structures. Although this appears counterintuitive, one has to keep in mind that RNAz was trained on pure sequence alignments and structural alignments could result in artifactually high scores even for alignments without conserved RNA structure.

File uploads are currently limited to 20 MB. This allows, e.g. to screen roughly 2 megabases of 6-way alignments in MAF format.

Local scanning and pre-processing of alignments

The RNAz algorithm works 'globally', i.e. the given alignment is scored as a whole. For long alignments (e.g. alignment of a whole chromosome), this is neither computationally tractable nor biologically meaningful. Therefore, long alignments are scanned in overlapping windows. The window and step size can be set by the user. By default, a window size of 120 and a step size of 40 is used. This window size appears large enough to detect local secondary structures within long ncRNAs and, on the other hand, small enough to find short secondary structures without loosing the signal in a much too long window.

In addition to this step, alignments are filtered in various ways before they are analyzed with RNAz. In particular, automatically generated genomic alignments are full of gap-rich regions, dubious aligned fragments or low-complexity regions. Such alignments are unlikely to contain true conserved structures and, in some cases, can cause artifactual predictions. Sequences that contain, e.g. too many gaps or too many repeat-masked letters are therefore filtered out. Details of the filtering process can be set by the user (Figure 1A).

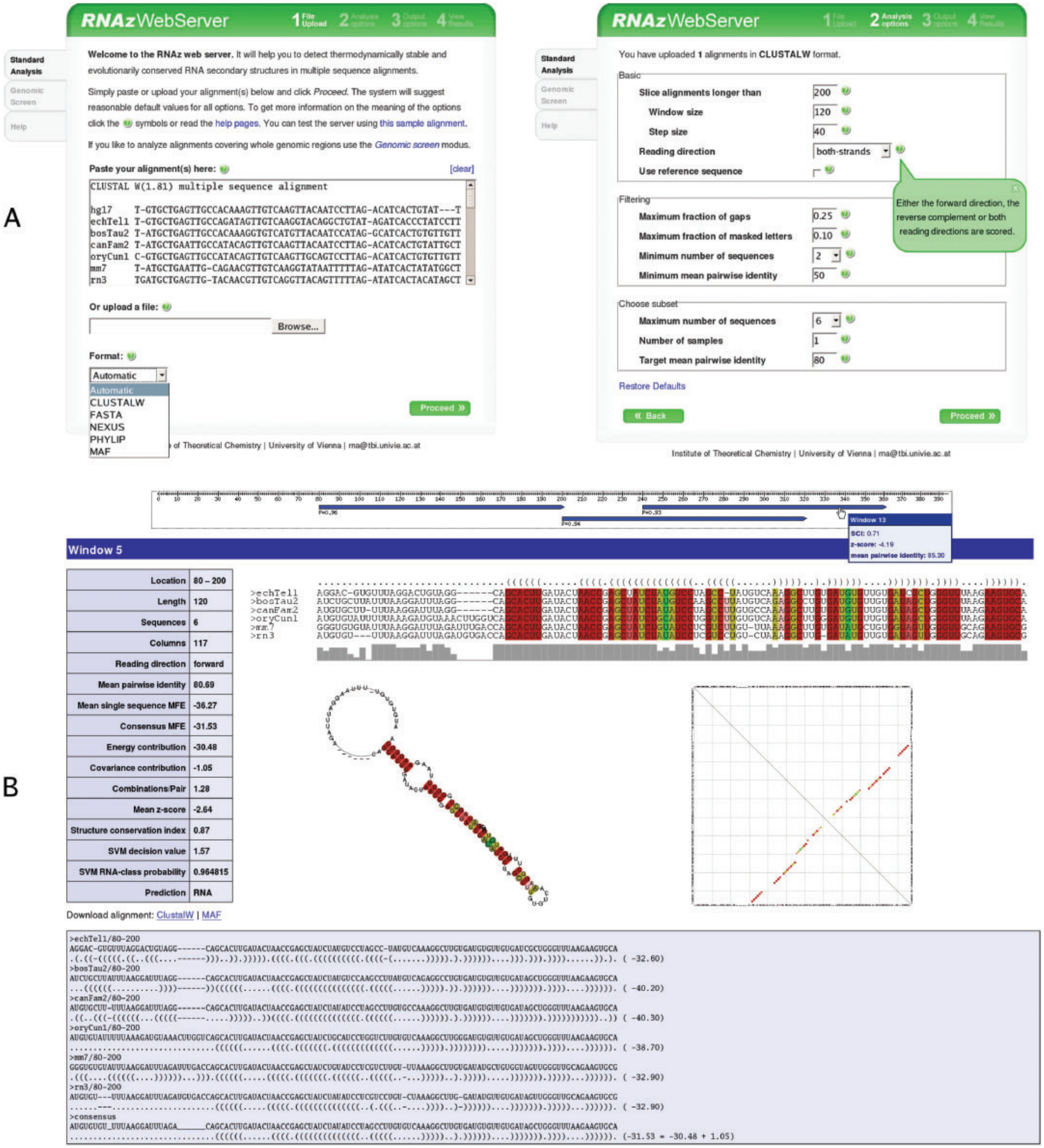
The RNAz program in its current implementation can only analyze alignments with up to six sequences. Six sequences usually hold enough information to allow reasonable predictions. If there are more sequences in the given alignment, the server selects an optimal subset of sequences. A greedy algorithm is used that gradually selects sequences to optimize for a given target diversity in the alignment. By default, a subset of six sequences is chosen which is optimized for a mean pairwise sequence identity of 80%.

The output

Sample output of the server is shown in Figure 1B. In 'Standard Analysis' mode, an overview of each uploaded alignment is shown. Windows containing predicted secondary structures are highlighted and detailed information (z-score, structure conservation index, RNAz *P*-value, etc.) is shown in a table. These results are supplemented by different visualizations of the predicted consensus secondary structure. A typical secondary structure drawing, a 'dot-plot' representing the base-pairing probability matrix, and a structure-annotated alignment are generated. All three visualizations are color coded which makes it easy to identify compensatory/consistent mutations that support a predicted structure. In addition, the raw RNAz output can be viewed as text file. In 'Genomic screen' modus also annotation files in the standard formats BED and GFF are generated if desired. All result files are stored for 30 days on the server and can be downloaded as a single compressed archive file for local viewing.

Conducting genomic screens

For screening genomic regions, the 'Genomic screen' option must be chosen on the first page of the server. In general, the analysis pipeline and the generated output are the same as described above. However, only alignments in MAF and XMFA formats are read. These alignment need to fulfill some requirements: The identifier of the first sequence in the first alignment is used as 'reference'.



Welcome to the RNAz web server. It will help you to detect thermodynamically stable and evolutionarily conserved RNA secondary structures in multiple sequence alignments.

Simply paste or upload your alignment(s) below and click *Proceed*. The system will suggest reasonable default values for all options. To get more information on the meaning of the options click the symbols or read the help pages. You can test the server using this sample alignment.

If you like to analyze alignments covering whole genomic regions use the *Genomic screen* modus.

Paste your alignment(s) here: [clear]

```
CLUSTAL W(1.81) multiple sequence alignment


hg17     T-GTGCTGAGTTGCCACAAAGTTGTCAAGTTACAATCCTTAG-ACATCACTGTAT---T
echTel1  T-GTGCTGAGTTGCCAGATAGTTGTCAAGGTACAGGCTGTAT-AGATCACCCTATCCTT
bosTau2  T-ATGCTGAGTTGCCACAAAGGTGTCATGTTACAATCCATAG-GCATCACTGTGTTGTT
canFam2  T-ATGCTGAATTGCCATACAGTTGTCAAGTTACAATCCTTAG-ACATCACTGTATTGCT
oryCun1  C-GTGCTGAGTTGCCATACAGTTGTCAAGTTGCAGTCCTTAG-ACATCACTGTGTTGTT
mm7      T-ATGCTGAATTG-CAGAACGTTGTCAAGGTATAATTTTTAG-ATATCACTATATGGCT
rn3      TGATGCTGAGTTG-TACAACGTTGTCAGGTTACAGTTTTTAG-ATATCACTACATAGCT
```

Or upload a file:

Format: Automatic

Institute of Theoretical Chemistry | University of Vienna | rna@tbi.univie.ac.at

You have uploaded **1** alignments in **CLUSTALW** format.

Basic

Slice alignments longer than	200
Window size	120
Step size	40
Reading direction	both-strands
Use reference sequence	□

Either the forward direction, the reverse complement or both reading directions are scored.

Filtering

Maximum fraction of gaps	0.25
Maximum fraction of masked letters	0.10
Minimum number of sequences	2
Minimum mean pairwise identity	50

Choose subset

Maximum number of sequences	6
Number of samples	1
Target mean pairwise identity	80

Institute of Theoretical Chemistry | University of Vienna | rna@tbi.univie.ac.at

Window 13

SCI: 0.71
z-score: -4.19
mean pairwise identity: 85.30

Location	80 – 200
Length	120
Sequences	6
Columns	117
Reading direction	forward
Mean pairwise identity	80.69
Mean single sequence MFE	-36.27
Consensus MFE	-31.53
Energy contribution	-30.48
Covariance contribution	-1.05
Combinations/Pair	1.28
Mean z-score	-2.64
Structure conservation index	0.87
SVM decision value	1.57
SVM RNA-class probability	0.964815
Prediction	RNA

```
         ..............................((((((......((((.((((((((((((((..(((((......)))))..).))))))....))))))).))))....)))))).
>echTel1 AGGAC-GUGUUUAGGACUGUAGG------CAGCACUUGAUACUAACCGAGCUAUCUAUGUCCUAGCC-UAUGUCAAAGGCUUGUGAUGUGUUGUGAUCGCUGGGUUUAAGAAGUGCA
>bosTau2 AUCUGCUUAUUUAAGGAUUUAGG------CAGCACUUGAUACUAACCGAGCUAUCUAUGUCCAAGCCUUAUGUCAGAGGCCUGUGAUGUGUUGUGAUAGCUGGGUUUAAGAAGUGCA
>canFam2 AUGUGCUU-UUUAAGGAUUUAGG------CAGCACUUGAUACUAACCGAGCUAUCUAUAUCCUAGCCUUGUGCCAAAGGCUUGUGAUGUGUUGUGAUAGCUGGGUUUAAGAAGUGCA
>oryCun1 AUGUGUAUUUUUAAAGAUGUAAACUUGGUCAGCACUUGAUACUAACCGAGCUAUCUGCAUCCUGGUCUUGUGUCAAAGGCUUGGGAUGUGUUGUGAUAGCUGGGUUUAAGAAGUGCA
>mm7     GGGUGUGUAUUUAAGGAUUUAGAUUUGACCAGCACUUGAUACUAACCGAGCUAUCUGUAUCCUCGUCUUGU-UUAAAGGCUUGUGAUAUGCUGUGGUAGUUGGGUUGCAGAAGUGCG
>rn3     AUGUGU---UUUAAGGAUUUAGAUGUGACCAGCACUUGAUACUAACCGAGCUAUCUAUAUCCUCGUCCUGU-CUAAAGGCUUG-GAUAUGUUGUGAUAGUUGGGUUGCAGAAGUGCG
```

Download alignment: ClustalW | MAF

```
>echTel1/80-200
AGGAC-GUGUUUAGGACUGUAGG------CAGCACUUGAUACUAACCGAGCUAUCUAUGUCCUAGCC-UAUGUCAAAGGCUUGUGAUGUGUUGUGAUCGCUGGGUUUAAGAAGUGCA
.(.((-(((((.((..(((....------))).)).))))).((((.(((.((((((((((.((((-(.......))))).).))))))....))).))).))))......)).). ( -32.60)
>bosTau2/80-200
AUCUGCUUAUUUAAGGAUUUAGG------CAGCACUUGAUACUAACCGAGCUAUCUAUGUCCAAGCCUUAUGUCAGAGGCCUGUGAUGUGUUGUGAUAGCUGGGUUUAAGAAGUGCA
...((((((..........))))------))((((((......((((.(((((((((((((((.(((((......))))).)).))))))....))))))).))))....)))))). ( -40.20)
>canFam2/80-200
AUGUGCUU-UUUAAGGAUUUAGG------CAGCACUUGAUACUAACCGAGCUAUCUAUAUCCUAGCCUUGUGCCAAAGGCUUGUGAUGUGUUGUGAUAGCUGGGUUUAAGAAGUGCA
.((..(((-((((((...(((((------.....)))))..))((((.((((((((((((((.((((((......)))))).).))))))....))))))).))))))))))..)) ( -40.30)
>oryCun1/80-200
AUGUGUAUUUUUAAAGAUGUAAACUUGGUCAGCACUUGAUACUAACCGAGCUAUCUGCAUCCUGGUCUUGUGUCAAAGGCUUGGGAUGUGUUGUGAUAGCUGGGUUUAAGAAGUGCA
..............................((((((......((((.(((((((.((((((((((((((......))))).)))))))))).....))))))).))))....)))))). ( -38.70)
>mm7/80-200
GGGUGUGUAUUUAAGGAUUUAGAUUUGACCAGCACUUGAUACUAACCGAGCUAUCUGUAUCCUCGUCUUGU-UUAAAGGCUUGUGAUAUGCUGUGGUAGUUGGGUUGCAGAAGUGCG
.(((....((((((....))))))...))).((((((.....(((((.(((((((.((((((..(((((..-...)))))..).))))).....))))))).)))))...)))))). ( -32.90)
>rn3/80-200
AUGUGU---UUUAAGGAUUUAGAUGUGACCAGCACUUGAUACUAACCGAGCUAUCUAUAUCCUCGUCCUGU-CUAAAGGCUUG-GAUAUGUUGUGAUAGUUGGGUUGCAGAAGUGCG
......---......................((((((.....(((((.((((((((((((((..(.(((..-....))))..)-))))))....))))))).)))))...)))))). ( -32.90)
>consensus
AUGUGUGU_UUUAAGGAUUUAGA______CAGCACUUGAUACUAACCGAGCUAUCUAUAUCCUAGCCUUGUGUCAAAGGCUUGUGAUGUGUUGUGAUAGCUGGGUUUAAGAAGUGCA
..............................((((((......((((.((((((((((((((..(((((......)))))..).))))))....))))))).))))....)))))). (-31.53 = -30.48 + 1.05)
```

Figure 1. (**A**) Screenshots of the RNAz web server. The interface to upload alignments in *Standard Analysis* mode is shown left. The various slicing and filtering options together with context-sensitive online help is shown right. (**B**) Sample output of an alignment of ~400 columns that contains a H/ACA snoRNA in the middle and that was scored in overlapping windows of 120 columns and step size 40. The overview panel shows windows that were predicted to contain a significant RNA structure with RNAz classification probability higher than 90%. Below, detailed results for the window from positions 80 to 200 are shown which contains the first of the two stem-loops that are typical for H/ACA snoRNAs. The output consists of a table summarizing alignment characteristics and RNAz results, graphical representations of the consensus structure (structure annotated alignment, secondary structure drawing, base-pairing probabilities 'dot-plot') and secondary structure models in dot/bracket notation for each single sequence in comparison to the consensus structure.

Each provided alignment must contain a sequence with this identifier and at least for this reference sequence correct genomic positions must be provided in the alignment. The MAF and XMFA file formats provide fields to store this information.

Also in this mode, alignments are sliced if necessary and filters are applied. After scoring of filtered alignment windows, RNA predictions in overlapping windows are combined to non-overlapping genomic 'loci'. The genomic location of the predicted loci can be downloaded as BED or GFF annotation file and are presented in an overview table. It is also possible to upload an annotation file with already available annotation. This information will be included in the overview table and allows to compare the predictions with existing annotation. Each prediction shown in the overview table is linked to detailed result pages with illustrations and tables explained earlier.

IMPLEMENTATION

The web server was implemented using Apache, Perl, BioPerl (22), CGI and client-side JavaScript. The analysis pipeline builds upon the programs of the RNAz package version 1.0. As of writing this article, the system makes use of 4 Intel XEON 2.20 GHz CPUs for performing the calculations.

ACKNOWLEDGEMENTS

The authors would like to thank Stephan Bernhart for useful discussions. This work was supported by the Austrian Genome Initiative GEN-AU of the BMWF (projects 'non-coding RNA' and 'Bioinformatics Integration network'). Funding to pay the Open Accesss publication charges for this article was provided by GEN-AU.

Conflict of interest statement. None declared.

REFERENCES

1. Bompfünewerer,A.F., Flamm,C., Fried,C., Fritzsch,G., Hofacker,I.L., Lehmann,J., Missal,K., Mosig,A., Müller,B. *et al.* (2005) Evolutionary patterns of non-coding RNAs. *Th. Biosci.*, **123**, 301–369.
2. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, REVIEWS0004.
3. Athanasius, F. Bompfünewerer Consortium: Backofen, R., Bernhart, S. H., Flamm, C., Fried, C., Fritzsch, G., Hackermüller, J., Hertel, J., Hofacker, I. L. *et al.* (2007) RNAs everywhere: genome-wide annotation of structured RNAs. *J. Exp. Zool. B Mol. Dev. Evol*, **308**,1–25.
4. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
5. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
6. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
7. Rivas,E. and Eddy,S. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
8. Washietl,S. and Hofacker,I.L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, **342**, 19–39.
9. Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
10. diBernardo,D., Down,T. and Hubbard,T. (2003) ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics*, **19**, 1606–1611.
11. Coventry,A., Kleitman,D.J. and Berger,B. (2004) MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, **101**, 12102–12107.
12. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, **102**, 2454–2459.
13. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
14. Torarinsson,E., Sawera,M., Havgaard,J., Fredholm,M. and Gorodkin,J. (2006) Thousands of corresponding human an mouse genomic regions unalignable in primary sequece contain common RNA structure. *Genome Res.*, **6**, 885–889.
15. Uzilov,A.V., Keegan,J.M. and Mathews,D.H. (2006) Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, **7**, 173.
16. Washietl,S., Hofacker,I.L., Lukasser,M., Hüttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.
17. Washietl,S., Pedersen,J.S., Korbel,J.O., Gruber,A.R., Stocsits,C., Hackermüller,J., Hertel,J., Lindemeyer,M., Reiche,K. *et al.* (2007) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.*in press.
18. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
19. Clote,P., Ferre,F., Kranakis,E. and Krizanc,D. (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578–591.
20. Gardner,P.P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
21. Wilm,A., Mainz,I. and Steger,G. (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 19.
22. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I. *et al.* (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.