Nucleic Acids Research Advance Access published November 11, 2010

Nucleic Acids Research, 2010, ***1–4***
doi:10.1093/nar/gkq990

AREsite: a database for the comprehensive investigation of AU-rich elements

Andreas R. Gruber[1,*], Jörg Fallmann[1], Franz Kratochvill[2], Pavel Kovarik[2] and Ivo L. Hofacker[1]

[1]Institute for Theoretical Chemistry and [2]Department of Microbiology and Immunobiology, Max F. Perutz Laboratories, University of Vienna, Vienna, Austria

Received August 15, 2010; Revised October 3, 2010; Accepted October 5, 2010

ABSTRACT

AREsite is an online resource for the detailed investigation of AU-rich elements (ARE) in vertebrate mRNA 3′-untranslated regions (UTRs). AREs are one of the most prominent *cis*-acting regulatory elements found in 3′-UTRs of mRNAs. Various ARE-binding proteins that possess RNA stabilizing or destabilizing functions are recruited by sequence-specific motifs. Recent findings suggest an essential role of the structural mRNA context in which these sequence motifs are embedded. AREsite is the first database that allows to quantify the structuredness of ARE motif sites in terms of opening energies and accessibility probabilities. Moreover, we also provide a detailed phylogenetic analysis of ARE motifs and incorporate information about experimentally validated targets of the ARE-binding proteins TTP, HuR and Auf1. The database is publicly available at: http://rna.tbi.univie.ac.at/AREsite.

INTRODUCTION

AU-rich elements (AREs) are distinct sequence elements in the 3′-untranslated region (UTR) of mRNAs often consisting of one or several AUUUA pentamers located in an adenosine and uridine rich region (1). Numerous proteins directly interact with AREs, thereby modulating mRNA stability or translational efficiency. The importance of these sequence motifs has been highlighted recently by a multitude of studies pointing out that the loss of ARE-mediated mRNA control leads to severe pathologies as AREs affect gene expression on a global scale (2–7).

AREs have been studied bioinformatically early on (8) and today's estimate is that ~7% of the human protein-coding genes contain AREs (9). However, the presence of an ARE consensus motif alone is not enough to qualify a gene as a true *in vivo* target of ARE-binding proteins. Recent computational and experimental evidence (10–13) and the fact that ARE-targeting proteins bind to RNA in single-stranded conformation (14) emphasize the need to analyze the structural context these motifs are embedded in. Furthermore, the mounting comparative genomics data available can be harnessed to identify evolutionarily conserved motif sites. AREsite is the first database that combines sequence annotation of AREs with the prediction of the accessibility and evolutionary conservation of the motif site. In addition to these features, we incorporated information from extensive expert literature search and list experimentally validated targets of the ARE-binding proteins TTP, HuR and Auf1.

DATABASE GENERATION AND CONTENT

In its current version AREsite uses Ensembl release 56 as data basis. For human and mouse, any protein-coding gene that has at least one transcript with a 3′-UTR sequence has been added to the collection. To account for the various definitions of AREs found in literature we decided not to restrict the database to a single motif, but offer the user the possibility to screen for a total of eight different consensus motifs, starting with the plain AUUUA pentamer to the WWWWAUUUAWWWW 13-mer, which resembles the core motif embedded in a stretch of A/U residues. By default, only the representative transcript of the selected gene, which we define as the transcript with the most AUUUA counts in its 3′-UTR sequence, is analyzed in detail. For each transcript we list sequence statistics and calculate the fold enrichment based on an order-0 and an order-1 Markov model for each motif. Beside plain sequence annotation of ARE motifs in transcripts AREsite also offers the researcher to study sequence conservation of motifs on both transcript and genomic level. For each motif site we provide annotated alignments with highlighted conserved motifs and

*To whom correspondence should be addressed. Tel: +43 1 4277 52731; Fax: +43 1 4277 52793; Email: agruber@tbi.univie.ac.at sequence logos (15). Finally, an overview figure in form of a phylogenetic tree depicts the conservation pattern of all detected motif sites. Motif site accessibility in terms of opening energies and probabilities of being unpaired are calculated using RNAplfold (16,17). For each motif we present accessibility values for the core AUUUA pentamer. Furthermore, results are visualized in an interactive SVG plot that allows the user to explore different parameter settings (Figure 1).

For the three best studied ARE-binding proteins TTP, HuR and Auf1, literature was screened for putative or confirmed mRNA targets. We classified the type of evidence for an mRNA being targeted by one of the three proteins by five criteria: (i) direct binding of the protein to the mRNA or its 3′-UTR (e.g. using RNA immunoprecipitation or electrophoretic mobility shift assays); (ii) an independent reporter assay confirming the functionality of the putative binding site; (iii) the loss or overexpression of the ARE-binding protein affects mRNA and/or (iv) the protein level of the target mRNA; (v) the stability of the target mRNA is affected by the lack or excess of the ARE-binding protein as shown by actinomycin D chase experiments or cell-free decay assays. New references will be added on a regular basis.

Figure 2 shows a typical output of an AREsite query. If the user aims for permanent storage of the search results, annotated Genbank files can be downloaded for each analyzed transcript.

Generation of alignments from transcripts

Alignments of orthologous transcripts were generated using data from the Ensembl gene orthology pipeline. For each gene database entry we first collected all orthologous genes from other species that have a strict one to one relation. Next we screened for transcripts that have an annotated 3′-UTR and among those we selected the one that showed the best coverage (at least 75%) of the reference species 3′-UTR. Multiple species whole transcript alignments were then generated with CLUSTAL W. To investigate the sequence conservation of the motif site we finally extract the region containing the motif site plus five flanking nucleotides on each side from the alignments. Each alignment sequence is then searched with the corresponding consensus ARE motif. Finally, detected motifs are used as sequence anchors and sequences are realigned using DIALIGN (18). The same procedure was also applied to the processed and filtered genomic alignments.

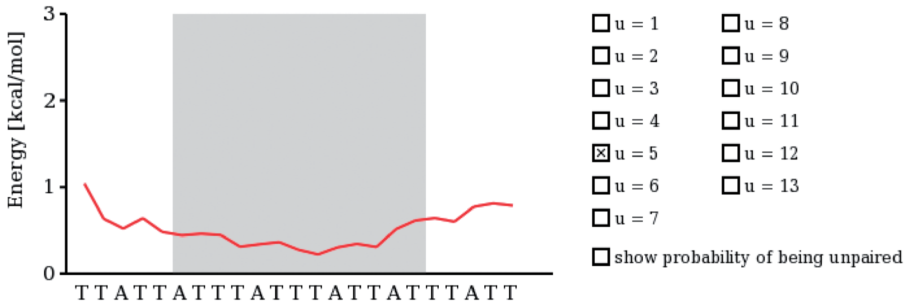
Generation of genomic alignments

Since comparative data at the level of transcripts is still limited, we decided to also incorporate data from genome-wide alignments to get a more refined picture of the conservation pattern of motifs. Interpretation of these data though has to be done with caution since there is no guarantee that the aligned sequences from other species really belong to the gene of interest. We apply, however, filtering strategies that ensure that aligned sequences are homologous over a longer stretch of nucleotides than simply the motif site.

Genomic alignments in MAF format were obtained for each UTR sequence from multiz generated alignments available at the UCSC genome browser (19). For human, corresponding alignments were extracted from 46 species multiple alignments based on the human genome assembly hg19, and for mouse from 30 species multiple alignments based on the assembly mm9. The obtained alignment blocks were often too short for any practical use and so we developed a MAF processing and filtering pipeline, that first merges adjacent MAF blocks to longer ones and then returns alignment windows of 120 nt and a step size of 30 nt. Finally, these windowed alignments were realigned with CLUSTAL W and were filtered to contain only sequences that have a length of at least 50% of the sequence length of the reference species.

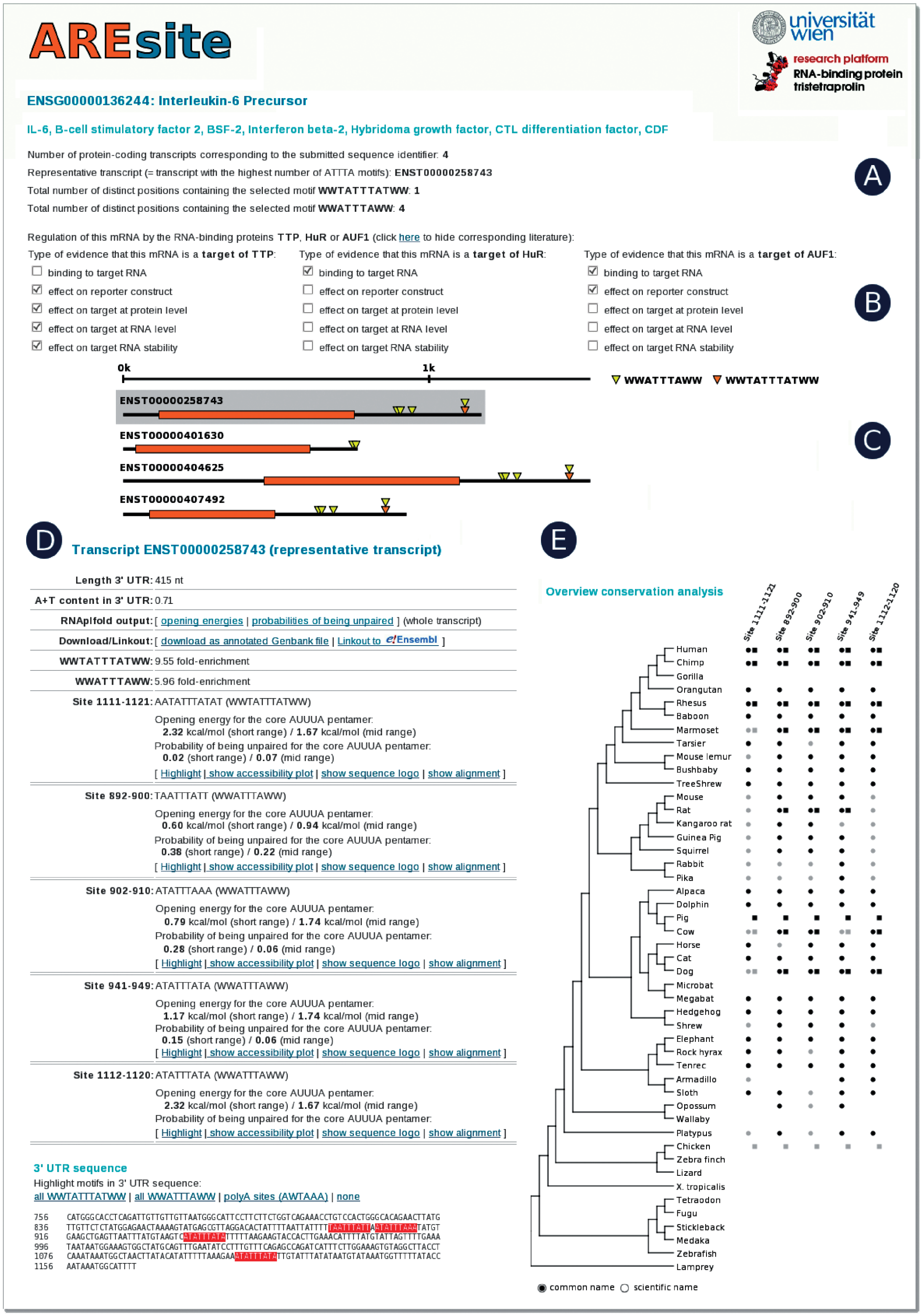
Quantifying motif site accessibility

For the calculation of the motif site accessibility in terms of opening energies and probabilities of being unpaired we used RNAplfold (16) with different parameter settings. RNAplfold is a thermodynamic RNA folding program that calculates local base-pairing probabilities, as well as the probability that a stretch of u consecutive nucleotides is unpaired (17). These probabilities are directly related to the energy needed to open all secondary structures in the respective stretch of nucleotides. The parameter set



□ u = 1
□ u = 2
□ u = 3
□ u = 4
☒ u = 5
□ u = 6
□ u = 7
□ u = 8
□ u = 9
□ u = 10
□ u = 11
□ u = 12
□ u = 13
□ show probability of being unpaired

Figure 1. Screenshot of the interactive SVG plot showing an ARE motif site of the human TNF-alpha gene. TNF-alpha is one of the best characterized ARE-containing genes. Its ARE target site consists of several consecutive ATTTA (AUUUA) motifs which favors the site's accessibility. When using a SVG ready web browser the user can explore the target site and flanking nucleotides with different parameter settings. With default settings ($u = 5$), the plot shows for each nucleotide i the energy that is needed to open local secondary structure for a stretch of five nucleotides (5′–3′) ending at position i.



ENSG00000136244: Interleukin-6 Precursor

IL-6, B-cell stimulatory factor 2, BSF-2, Interferon beta-2, Hybridoma growth factor, CTL differentiation factor, CDF

Number of protein-coding transcripts corresponding to the submitted sequence identifier: **4**

Representative transcript (= transcript with the highest number of ATTTA motifs): **ENST00000258743**

Total number of distinct positions containing the selected motif **WWTATTTATWW**: **1**

Total number of distinct positions containing the selected motif **WWATTTAWW**: **4**

Regulation of this mRNA by the RNA-binding proteins **TTP**, **HuR** or **AUF1** (click here to hide corresponding literature):

Type of evidence that this mRNA is a **target of TTP**:	Type of evidence that this mRNA is a **target of HuR**:	Type of evidence that this mRNA is a **target of AUF1**:
□ binding to target RNA	☑ binding to target RNA	☑ binding to target RNA
☑ effect on reporter construct	□ effect on reporter construct	☑ effect on reporter construct
☑ effect on target at protein level	□ effect on target at protein level	□ effect on target at protein level
☑ effect on target at RNA level	□ effect on target at RNA level	□ effect on target at RNA level
☑ effect on target RNA stability	□ effect on target RNA stability	□ effect on target RNA stability

Transcript ENST00000258743 (representative transcript)

Length 3' UTR:	415 nt
A+T content in 3' UTR:	0.71
RNAplfold output:	[opening energies | probabilities of being unpaired] (whole transcript)
Download/Linkout:	[download as annotated Genbank file | Linkout to Ensembl]
WWTATTTATWW:	9.55 fold-enrichment
WWATTTAWW:	5.96 fold-enrichment
Site 1111-1121:	AATATTTATAT (WWTATTTATWW) Opening energy for the core AUUUA pentamer: **2.32** kcal/mol (short range) / **1.67** kcal/mol (mid range) Probability of being unpaired for the core AUUUA pentamer: **0.02** (short range) / **0.07** (mid range) [Highlight | show accessibility plot | show sequence logo | show alignment]
Site 892-900:	TAATTTATT (WWATTTAWW) Opening energy for the core AUUUA pentamer: **0.60** kcal/mol (short range) / **0.94** kcal/mol (mid range) Probability of being unpaired for the core AUUUA pentamer: **0.38** (short range) / **0.22** (mid range) [Highlight | show accessibility plot | show sequence logo | show alignment]
Site 902-910:	ATATTTAAA (WWATTTAWW) Opening energy for the core AUUUA pentamer: **0.79** kcal/mol (short range) / **1.74** kcal/mol (mid range) Probability of being unpaired for the core AUUUA pentamer: **0.28** (short range) / **0.06** (mid range) [Highlight | show accessibility plot | show sequence logo | show alignment]
Site 941-949:	ATATTTATA (WWATTTAWW) Opening energy for the core AUUUA pentamer: **1.17** kcal/mol (short range) / **1.74** kcal/mol (mid range) Probability of being unpaired for the core AUUUA pentamer: **0.15** (short range) / **0.06** (mid range) [Highlight | show accessibility plot | show sequence logo | show alignment]
Site 1112-1120:	ATATTTATA (WWATTTAWW) Opening energy for the core AUUUA pentamer: **2.32** kcal/mol (short range) / **1.67** kcal/mol (mid range) Probability of being unpaired for the core AUUUA pentamer: [Highlight | show accessibility plot | show sequence logo | show alignment]

3' UTR sequence

Highlight motifs in 3' UTR sequence:
all WWTATTTATWW | all WWATTTAWW | polyA sites (AWTAAA) | none

```
756     CATGGGCACCTCAGATTGTTGTTGTTAATGGGCATTCCTTCTTCTGGTCAGAAACCTGTCCACTGGGCACAGAACTTATG
836     TTGTTCTCTATGGAGAACTAAAAGTATGAGCGTTAGGACACTATTTTAATTATTTTTAATTTATTAATATTTAAATATGT
916     GAAGCTGAGTTAATTTATGTAAGTCATATTTATATTTTTAAGAAGTACCACTTGAAACATTTTATGTATTAGTTTTGAAA
996     TAATAATGGAAAGTGGCTATGCAGTTTGAATATCCTTTGTTTCAGAGCCAGATCATTTCTTGGAAAGTGTAGGCTTACCT
1076    CAAATAAATGGCTAACTTATACATATTTTTAAAGAAATATTTATATTGTATTTATATAATGTATAAATGGTTTTTATACC
1156    AATAAATGGCATTTT
```

Overview conservation analysis

Figure 2. Snapshot of a typical AREsite results page (gene: human IL6). (**A**) Basic statistics about the selected gene. (**B**) Experimental evidence collected for this gene. For each of the ARE-binding proteins TTP, HuR and Auf1 we list the type of evidence. The user can choose to see the supporting publications which are directly linked to Pubmed. (**C**) Overview figure that shows all know transcripts of the selected gene and highlights detected ARE motifs in the 3′-UTRs. The representative transcript which is analyzed in detail is shown in a gray box. (**D**) Detailed summary of the analysis results for the representative transcript. For each motif site the user can choose to display accessibility plots, genomic and transcript alignments together with sequence logos. (**E**) Overview figure of the conservation analysis. Black circles (genomic alignments) and boxes (transcript alignments) indicate that the corresponding ARE motif was also detected in the sequence of the corresponding species.

$W = 80$, $L = 40$ models the effects of cotranscriptional folding and has been previously used to predict siRNA binding (20). AREsite features also a different parameter setting ($W = 240$, $L = 120$), which considers longer base pair spans and shows improved results on siRNA binding as well as on RNA–RNA interaction (H. Tafer, personal communication). For each detected motif site we list the accessibility values ($u = 5$) for the core AUUUA pentamer for both parameter settings (short range, mid range).

DISCUSSION

In this contribution we have introduced AREsite, a database for the detailed investigation of ARE motifs in terms of motif site accessibility and evolutionary conservation. In its current state AREsite reports 3275 human protein coding genes which have at least one occurrence of the consensus motif WUAUUUAUW in their $3'$-UTR sequences. This corresponds to $\sim$16% of the human protein coding genes. For 711 of those genes AREsite lists experimental evidence that they are targets of ARE-binding proteins. The requirements which are needed to qualify a gene as an *in vivo* target of ARE-binding proteins are still poorly understood. AREsite with its features of conservation pattern analysis and accessibility prediction can help researchers to unravel the underlying mechanism. Recent studies (11,13) demonstrate the great value of combining computational accessibility prediction and wet-lab data. When interpreting accessibility predictions one has to keep in mind, however, that low accessibility does not necessarily exclude a gene from being an *in vivo* target. mRNA regulation is a complex system and the binding of one factor might lead to structural rearrangements which can make a formerly cryptic site accessible or vice versa (21). In the context of AREs, this concept has been nicely demonstrated by using artificially designed mRNA openers and closers to control mRNA stability (22). The accurate modeling of these combinatorial effects will be among the most challenging issues for future work.

FUNDING

University of Vienna ‘‘Research platform: Structural and Functional Analysis of mRNA Molecules Targeted by the RNA-binding Protein Tristetraprolin’’ (to P.K. and I.L.H.) Funding for open access charge: University of Vienna.

Conflict of interest statement. None declared.

REFERENCES

1. Barreau,C., Paillard,L. and Osborne,H.B. (2005) AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res.*, **33**, 7138–7150.
2. Hao,S. and Baltimore,D. (2009) The stability of mRNA influences the temporal order of the induction of genes encoding inflammatory molecules. *Nat. Immunol.*, **10**, 281–288.
3. Lu,J.Y., Sadri,N. and Schneider,R.J. (2006) Endotoxic shock in AUF1 knockout mice mediated by failure to degrade proinflammatory cytokine mRNAs. *Genes Dev.*, **20**, 3174–3184.
4. Ghosh,M., Aguila,H.L., Michaud,J., Ai,Y., Wu,M.T., Hemmes,A., Ristimaki,A., Guo,C., Furneaux,H. and Hla,T. (2009) Essential role of the RNA-binding protein HuR in progenitor cell survival in mice. *J. Clin. Invest.*, **119**, 3530–3543.
5. Katsanou,V., Milatos,S., Yiakouvaki,A., Sgantzis,N., Kotsoni,A., Alexiou,M., Harokopos,V., Aidinis,V., Hemberger,M. and Kontoyiannis,D.L. (2009) The RNA-binding protein Elavl1/HuR is essential for placental branching morphogenesis and embryonic development. *Mol. Cell. Biol.*, **29**, 2762–2776.
6. Taylor,G.A., Carballo,E., Lee,D.M., Lai,W.S., Thompson,M.J., Patel,D.D., Schenkman,D.I., Gilkeson,G.S., Broxmeyer,H.E., Haynes,B.F. *et al.* (1996) A pathogenetic role for TNF alpha in the syndrome of cachexia, arthritis and autoimmunity resulting from tristetraprolin (TTP) deficiency. *Immunity*, **4**, 445–454.
7. Hodson,D.J., Janas,M.L., Galloway,A., Bell,S.E., Andrews,S., Li,C.M., Pannell,R., Siebel,C.W., MacDonald,H.R., De Keersmaecker,K. *et al.* (2010) Deletion of the RNA-binding proteins ZFP36L1 and ZFP36L2 leads to perturbed thymic development and T lymphoblastic leukemia. *Nat. Immunol.*, **11**, 717–724.
8. Bakheet,T., Frevel,M., Williams,B.R., Greer,W. and Khabar,K.S. (2001) ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucleic Acids Res.*, **29**, 246–254.
9. Halees,A.S., El-Badrawi,R. and Khabar,K.S. (2008) ARED organism: expansion of ARED reveals AU-rich element cluster variations between human and mouse. *Nucleic Acids Res.*, **36(Database issue)**, 137–140.
10. Hackermüller,J., Meisner,N.C., Auer,M., Jaritz,M. and Stadler,P.F. (2005) The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: a quantitative model. *Gene*, **345**, 3–12.
11. Li,X., Quon,G., Lipshitz,H.D. and Morris,Q. (2010) Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, **16**, 1096–107.
12. Rabani,M., Kertesz,M. and Segal,E. (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl Acad. Sci. USA*, **105**, 14885–90.
13. Kazan,H., Ray,D., Chan,E.T., Hughes,T.R. and Morris,Q. (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, **6**, e1000832.
14. Hudson,B.P., Martinez-Yamout,M.A., Dyson,H.J. and Wright,P.E. (2004) Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat. Struct. Mol. Biol.*, **11**, 257–264.
15. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
16. Bernhart,S.H., Hofacker,I.L. and Stadler,P.F. (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.
17. Bompfünewerer,A.F., Backofen,R., Bernhart,S.H., Hertel,J., Hofacker,I.L., Stadler,P.F. and Will,S. (2008) Variations on RNA folding and alignment: lessons from Benasque. *J. Math. Biol.*, **56**, 129–44.
18. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
19. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38(Database issue)**, 613–619.
20. Tafer,H., Ameres,S.L., Obernosterer,G., Gebeshuber,C.A., Schroeder,R., Martinez,J. and Hofacker,I.L. (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.
21. Kedde,M., van Kouwenhove,M., Zwart,W., Oude Vrielink,J.A., Elkon,R. and Agami,R. (2010) A Pumilio-induced RNA structure switch in p27-$3'$ UTR controls miR-221 and miR-222 accessibility. *Nat. Cell Biol.*, **12**, 1014–1020.
22. Meisner,N.C., Hackermüller,J., Uhl,V., Aszódi,A., Jaritz,M. and Auer,M. (2004) mRNA openers and closers: modulating AU-rich element-controlled mRNA stability by a molecular switch in mRNA secondary structure. *Chembiochem*, **5**, 1432–1447.