Published online 16 November 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D131–D136***
doi:10.1093/nar/gkp981

IRESite—a tool for the examination of viral and cellular internal ribosome entry sites

Martin Mokrejš[1,*], Tomáš Mašek[1], Václav Vopálenský[1], Petr Hlubuček[1], Philippe Delbos[1,2] and Martin Pospíšek[1,3,*]

[1]Department of Genetics and Microbiology, Faculty of Science, Charles University in Prague, Viničná 5, 128 44, Prague 2, Czech Republic, [2]University of Montpellier 2, CNRS, UMR 5235, Place Eugene Bataillon, F-34095 Montpellier, France and [3]Centre of Biomedical Informatics, Institute of Computer Science AS CR, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic

Received September 5, 2009; Revised October 13, 2009; Accepted October 15, 2009

ABSTRACT

The IRESite (http://www.iresite.org) presents carefully curated experimental evidence of many eukaryotic viral and cellular internal ribosome entry site (IRES) regions. At the time of submission, IRESite stored >600 records. The IRESite gradually evolved into a robust tool providing (i) biologically meaningful information regarding the IRESs and their experimental background (including annotation of IRES secondary structures and IRES *trans*-acting factors) as well as (ii) thorough concluding remarks to stored database entries and regularly updated evaluation of the reported IRES function. A substantial portion of the IRESite data results purely from in-house bioinformatic analyses of currently available sequences, *in silico* attempts to repeat published cloning experiments, DNA sequencing and restriction endonuclease verification of received plasmid DNA. We also present a newly implemented tool for displaying RNA secondary structures and for searching through the structures currently stored in the database. The supplementary material contains an updated list of reported IRESs.

INTRODUCTION

Eukaryotic translation initiation employs the 7-methylguanosine cap moiety bound to the 5′-end of mRNA to recruit transcripts to an assembling ribosome and initiate protein synthesis. Certain RNA viruses infecting eukaryotic cells lack the 5′-cap structures and, instead, utilize internal ribosome entry sites (IRESs) for binding the ribosome directly onto their (sub)genomic RNAs and thus for the so-called cap-independent initiation of viral protein synthesis. IRES-dependent translation initiation provides an important advantage to these specific viruses, as it allows them to transcribe their genome independently of nuclear transcription and mRNA modification machineries. Further, it allows them to synthesize their proteins in conditions when overall cellular cap-dependent protein synthesis is greatly reduced due to either cellular antiviral defense and/or the action of viral proteins. Shortly after simultaneous discoveries of the first viral IRESs in poliovirus and encephalomyocarditis virus by Pelletier & Sonenberg and Jang *et al.*, respectively, in 1988 (1,2), Macejak and Sarnow (3) reported the first cellular IRES within mRNA coding for human immunoglobulin heavy-chain binding protein (BiP). Since that time, the number of both viral and cellular IRESs reported has grown substantially.

The rationale for searching for new cellular IRES-containing mRNAs comes from observations that some mRNAs remain translated sufficiently even when the overall cellular protein synthesis is reduced under certain conditions (including stress, apoptosis, viral infection, starvation and mitosis). Another reason for searching for new cellular IRESs is provided by the obvious fact that we now know several functionally and structurally different epitomes of viral IRESs that can be handled by cellular protein synthesis apparatus efficiently. Thus, one can easily hypothesize that binding of the 5′-cap need not to be the only possible way for cellular mRNA to be translated in every condition.

However, it appears that some published results, especially those concerning putative cellular IRESs, can be explained not by IRES function *per se* but rather by the existence of aberrant, shorter mRNAs produced by cryptic transcription, cryptic splicing or the presence of

*To whom correspondence should be addressed. Tel: +420 22195 1719; Fax: +420 22195 1724; Email: martin@natur.cuni.cz
Correspondence may also be addressed to Martin Mokrejš. Tel: +420 22195 1716; Fax: +420 22195 1724; Email: mmokrejs@iresite.org

breakage hot spots within the examined RNA molecule. It has become more and more evident that reliable proof of IRES function, especially when putative IRES activity is very low, requires the application of sophisticated new approaches. Introduction of such approaches to verify function of the previously reported IRESs is one of the reasons for the still growing list of false and/or insufficiently experimentally supported IRESs. At the same time, it can be expected that other cellular and viral IRESs are awaiting discovery. For the most recent review on cap-dependent translation, see Sonenberg and Hinnebusch (4); for recent reviews and research papers containing extensive results on more viral and cellular IRESs, see (5–11). For comprehensive reviews on translational control, see the CSHL Press book series (12–14).

The presence of contradictory results from reported IRESs and difficulties in designing suitable IRES-proving experiments spawned our idea to record the key features of published IRESs together with their corresponding experimental data in a computer database. This database could then be used to evaluate which IRESs are well supported and characterized by what methods in what reporter gene setups, etc. Our aims are to provide the data accumulated in the IRESite database (http://www.iresite.org) (15) to the scientific community and to determine which of the many claimed IRESs published so far are sufficiently experimentally supported and meet current criteria. The first comparative analysis of data accumulated in the IRESite we published recently together with a discussion on experimental pitfalls associated with attempts to provide evidence of IRES existence and function (16). Here, we provide a description of the current stage of the IRESite, including explanation of newly introduced tools as well as a brief discussion of the approach used to create database entries. We also provide an up-to-date list of viral and cellular IRESs reported in the supplement.

CURRENT STATUS OF THE DATABASE

The IRESite aims to classify IRESs based on experimental evidence supporting their existence and therefore has to deal not only with annotation of sequences, structural data and interacting protein factors but also with a still growing list of experimental methods (used to confirm IRESs by their function) and the corresponding measured values. The IRESite contains two distinct groups of entries: annotated 'natural' transcripts derived from genomes of viral and cellular origins and a much larger set of data consisting of annotated 'engineered' transcripts produced from various DNA vectors. Internally, the database comprises a very complex network of 30 tables consisting of 338 columns covering various types of information (see Supplementary data for the SQL schema). The actual experimental results are represented by 36 columns, available experimental methods are reflected by 32 columns, our own annotations are stored in 76 columns and internal keys used to cross-reference the data and reuse existing data structures are represented by 194 columns. These numbers illustrate the degree of interrelationships between the data or, in other words, how many pieces of information can be both extracted from a single publication and added by curators as a bonus in remarks and evaluations.

Table 1. A portion of the IRESite statistics available at the web site under the 'record counts' reference showing the total numbers of entries classified into different categories (record counts as of August 2009)

	Counts
Viral RNAs containing reported IRES	43
Cellular mRNAs containing reported IRES	70
DNA vectors containing IRES or putative IRES	417
DNA vectors without IRES—negative controls	50
Promoter-less DNA vectors—negative controls	19
Pending unfinished records	3
Total record counts	602
Experimentally determined IRES secondary structures	43
IRES *trans*-acting factors reported	24
In vitro/*in vivo* translation measurements	575

Currently, the IRESite covers to some extent 113 IRESs. This roughly corresponds to 65% of all reported IRES elements (see the list of known IRESs in the Supplementary data). Users can efficiently search for data scattered in >300 scientific publications. A total of 486 entries are based on plasmid sequences. IRESite curators verified 60 plasmids either by restriction endonuclease mapping, PCR or sequencing. Two new plasmid sequences were deposited into GenBank under the accession numbers EU919738 and EU919739. Artificially synthesized IRESs, including defective variants, are covered by nine records. Experimentally determined secondary structures (43 entries) were recently adopted by the RFAM database (17). In total, the IRESite contains 602 records as of August 2009 (Table 1).

The IRESite web site is accessed from 1000 to 2000 individual IP addresses per month (excluding web crawlers and our own addresses). In addition to the web interface, the IRESite also provides an email discussion list in which curators give support to novice scientists aiming to identify and characterize new IRES.

Although the IRESite is a database of experimentally studied IRESs, we included for the user's convenience a number of insect dicistroviral IRESs that are supported only by phylogenetic analyses as well as sequence and structural similarity to several well proven IRESs of this large family. To distinguish these predicted (although quite reliable) IRESs, we placed the term 'pred' at the end of their name and emphasized this in their annotation.

New features of the IRESite

The database has grown in terms of both the number of records it contains as well as the experimental evidence being tracked and breadth of the annotation provided.

```
Score = 158
Identities with small structure = 61/61 (100%), Gaps = 0/61 (0%)
Identities with large structure = 54/61 (88%), Gaps = 7/61 (11%)
Strand = Plus / Plus (Minus strand search not supported)

Your query modified by gaps (GappedLength = 61, ActualLength = 61):
>HBV encapsidation signal
1 uguacauguccca cuguuc aagccuccaagcugugccuuggguggcuuuggggcauggaca 61
1 (((.(((((((((.....(((((((((((.....)))))).)))))))))))))).))) 61

The target hit adjusted by gaps (GappedLength = 61, ActualLength = 54):
>IRESite Id:481 2D structure of transcript of plasmid pRKMI1F with artificial IRES
3 auucccuccuccu------ccuccuccuccgaauucggagga-ggaggaggaggagggaau 63
3 ((((((((((((------(((((((((((.....)))))))-))))))))))))))))) 63
```

Figure 1. An example of colored text output from a structure-based search through the secondary structures contained in the IRESite. These are recorded as a series of left and right brackets (paired bases) and dots (unpaired bases). The web site allows users to search by structural motif, optionally accompanied by the primary sequence. Our wrapper around the RNAforester program computes simple statistics for each result and displays the location of the hit within the target structure (numbers to the left/right of the alignment). Please note that position numbers are relative to the experimentally mapped secondary structure region, not to the mRNA sequence coordinates. Here, the hepatitis B virus encapsidation signal was used as a sample query that hit part of the synthetic KMI1 IRES structure with the best (highest) score. Adjustments to both the query and target made by RNAforester to yield the alignment are shown in red.

Structures of many IRESs are well known and/or hypothesized to be important for their function in translation initiation. Therefore, the IRESite now provides tools enabling graphical display of RNA secondary structures as well as searching by secondary structure within RNA structures already stored in the database. Experimentally determined secondary structures of nucleic acids stored in IRESite can be conveniently displayed in our Java-based VARNA-align applet. We improved the original VARNA applet to allow users to save graphical representations of the structure into JPEG and SVG formats and to export plain sequences/structures into various other formats (including dot-bracketed notation). Some of our changes have already been backported into current version of VARNA by its authors (18). We also developed a wrapper around the RNAforester program to enable searching of stored secondary structures by a secondary structure motif represented in dot-bracketed notation (optionally accompanied by a corresponding primary sequence) (19). Resulting combined alignments of the query and target sequences and structures can either be displayed as colored HTML-formatted text or rendered graphically by the above-mentioned VARNA-align applet (Figures 1 and 2 and Supplementary Figures). We are still working on the secondary structure search tool and will improve it even further in the near future.

Local installation of the NCBI BLAST program on our web server enables searching through several subsets of IRESite sequences, and results are rendered as an HTML output with hyperlinks pointing to corresponding IRESite records. The low-complexity filter had to be turned off to facilitate searches for repetitive sequences which sometimes occur within viral and cellular IRESs (20).

The IRESite provides annotated plasmid sequences in GenBank flatfile format (http://iresite.org/IRESite_web.php?page=plasmids) in addition to annotations stored in the database. The annotation in these flatfiles is focused on annotation of promoters, known/putative transcription start sites, reporter protein-coding regions IRES regions and the poly(A) signals. In near future, we will provide a graphical overview of each plasmid map. We added annotation of plasmid name aliases; scientists sometimes rename plasmids in successive publications, but users obviously perform database searches by either name.

The parts of the database dedicated to tracking promoter and potential aberrant splicing issues were improved along with annotation capabilities, which can now record IRES *trans*-acting factors (ITAF) and other RNA-binding proteins.

The search engine was extended to allow more complex queries, including searches by taxonomic groups and aliases (e.g. 'fungi', 'mammals', 'baker's yeast') using the NCBI Taxonomy database (21) and searches for all records with a known secondary structure (http://www.iresite.org/IRESite_web.php?page=search&search_type=extended&form_search.HAS--se__ss_experiment_id=1).

The annotation process

The IRESite is fully manually curated and uses a combined bioinformatic and wet-bench approach to obtain sequence data, which are subjected to annotation in the database. The need for careful assembly of annotated sequences and re-examination of some reported IRES-containing mRNA sequences is strikingly evident from recently published results (e.g. Baranick *et al.* (22) suggested—by employing both detailed bioinformatic as well as sophisticated wet-bench approaches—to revise the opinion about the function of at least four out of six cellular IRESs tested).

In original publications, IRES regions were often referred to by their position with respect to either the

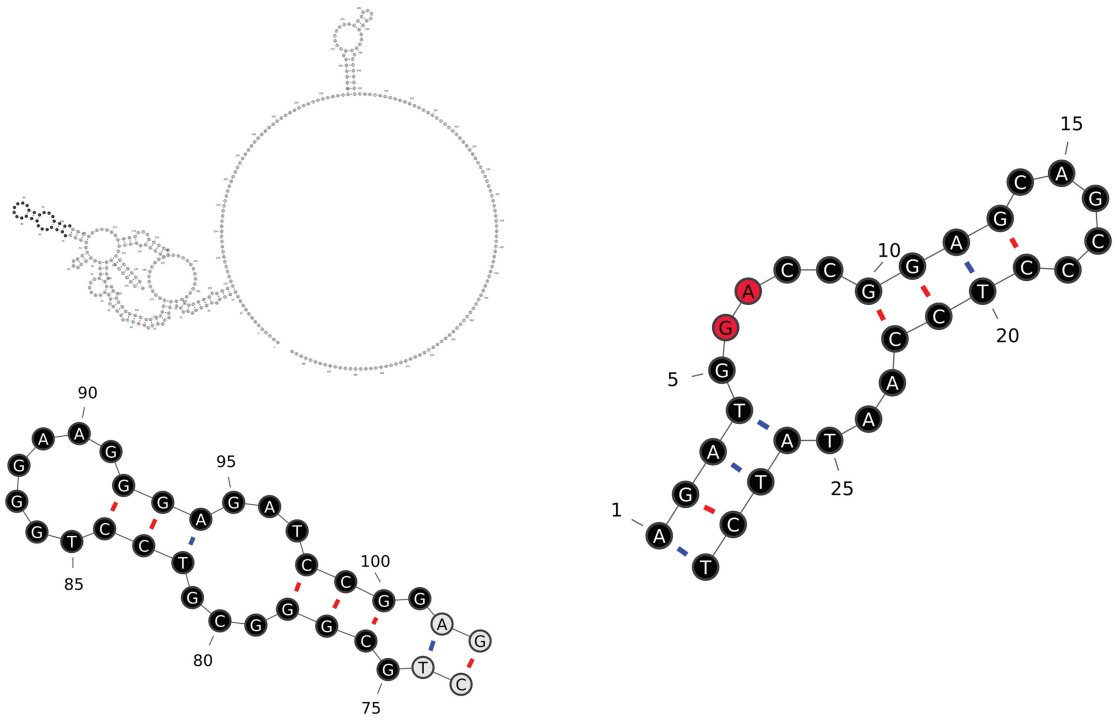

Figure 2. The Java-based VARNA-align program is automatically opened in user's internet browser to render the secondary structure of the IRES being accessed or to render search results graphically as two structures. The user can interactively move parts of the structure, zoom in/out and perform several other actions. In the upper left corner, a complete c-myc IRES structure (IRESite_ID: 35) is shown in gray at low resolution. Its matching part (found by RNAforester) is shown below in the lower left corner, and the adjusted query structure is shown on the right side The RNAforester search mode used for searches through the IRESite is called 'small-in-large'. Here, the query is treated by the RNAforester algorithm as a small structure, whereas the target structure is a large structure. Bases in black circles represent matching positions (primarily reflecting structural attributes and, to a lesser extent, nucleotides). Bases in gray circles are in the positions surrounding (outside) the matching region, and bases in red circles are in positions having no counterpart in the other structure (inside the matching region).

5′-end of the mRNA or the initiator AUG codon. Alternatively, they were delimited by primers or restriction sites used for cloning. We aim to build the IRESite upon full-length sequences of mRNAs and viral RNAs as well as complete 'engineered' transcript sequences actually used in particular experiments. Therefore, database entries are based only on the sequences, which we have (i) received from authors of the published findings, (ii) determined by sequencing or confirmed by restriction endonuclease mapping and PCR analyses of DNA vector samples (which were obtained either from the original authors or elsewhere and match the initially investigated samples) or (iii) reverse engineered based on published cloning procedures (often with the aid of public sequence databases and molecular-cloning computer programs). We prefer to present fewer entries that nonetheless correspond well to details of the original experiments and focus on thorough annotation and verification of transcript models. We discussed some differences between automated and manual approaches to database filling in more detail in our previous review concerning IRESite, UTRdb (http://utrdb.ba.itb.cnr.it) and IRESdb (http://www.rangueil.inserm.fr/IRESdatabase) (16).

For example, we annotated the 'eIF4G IRES' using four different mRNAs (IRESite_IDs: 548, 576, 573, 574) while omitting some other variants of the eIF4G messages initiated from different promoter. However, the latter two entries actually contain no IRES; due to the complicated exon structure of the messages and partial overlap with the reported eIF4G IRES, we included them in the database for clarity and completeness. Similarly, we created three crTMV IRES records (IRESite_IDs 39, 602, 603) to reflect the natural occurrence of the IRES in full-length genomic and two subgenomic RNAs, even though the IRES is functional only in some instances. Bag1 and Apaf-1 transcripts serve as another model example of the annotation issues. In the case of Bag1 (IRESite_ID: 106), the RefSeq NM__004323.4 sequence is 3.8 kb in length; however, we could only confirm the first 1.3 kb using experimentally determined mRNA and expressed sequence tags (ESTs) from GenBank. In the case of Apaf-1 (IRESite_ID: 110), we initially hunted for the longest mRNA sequence and thus picked the seemingly full-length sequence of AF013263. However, we later realized that this is a chimeric cDNA product containing Apaf-1 mRNA in region 1-5244 nts; the remainder (up to position 7042 nts) is a transcript of another gene.

As a result of the methods employed for sequence verification and annotation, many of the sequences deposited

in the IRESite are unique and cannot be found in other public repositories. The data set comprising 'natural' transcripts of viral and cellular origin has no perfect match in GenBank in 26% of the cases, whereas none of the 486 'engineered' plasmid-derived transcripts and almost none of the complete plasmid sequences themselves could be found in GenBank (23).

Examples of 'natural' transcripts unique to the IRESite include extended genomic sequences of ERAV genome (updated, extended sequences were obtained directly from authors of the original work; see IRESite_ID: 286), derived mRNA transcripts from endogenous retrotransposons (idefix in IRESite_ID: 68, gypsy in IRESite_ID: 69) and cloned fragments of nuclear genes (AQP4 in IRESite_ID: 491, Cat-1 in IRESite_ID: 438) amplified from organisms locally available to the original authors.

AVAILABILITY

IRESite data are freely available at http://www.iresite.org, and flatfiles with certain data are available upon request. When any part of a record is changed, its associated version number is automatically increased (note that multiple version numbers exist for each record subsection); a note is also appended to the Changelog. Users can search full-text IRESite data through their favorite public search engines. Several IRESite records are cross-referenced from GenBank and in the future, we plan to make this more common. All users are encouraged to submit any IRES-related experimental data and annotations through the web interface.

Our VARNA-align package is based on the VARNA package (http://varna.lri.fr/). Our VARNA improvements were submitted to the original author and are available at http://www.iresite.org/VARNA.

The original RNAforester software is available from http://bibiserv.techfak.uni-bielefeld.de/rnaforester. Our modifications of the RNAforester and our Python-based wrapper around it are available at http://www.iresite.org/RNAforester.

SUPPLEMENTARY DATA

Supplementary Data containing an up-to-date commented list of the reported IRESs, an SQL schema of the database and some examples of the VARNA-align output accompanied by comments are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank all the following colleagues who provided us with their sequences, plasmid samples, additional data or clarifications: M. Candeias, B. Crabb, J. M. Cuezva, A. de Tomassi, K. Grobe, T. Hinton, M. Holcik, M. Honda, J. Jimenez, R. Kaufmann, T. Komarova, A. Kroeger, A. Lauring, P. W.-L. Li, R. Lloyd, C. Logg, N. S. Magnuson, D. Maier, E. Martinez-Salas, W. Mikulits, J. Mitchell, T. Ohlmann, A. Olbrechts, M. Petz, A.-C. Prats, R. Rijnbrand, R. C. Sanchez, K. Sherril, W. Sossin, K. Spriggs, E. Terry, C. Thoma, C. Traboni, M. Vanhoucke, C. Vaury, A. Willis.

FUNDING

Ministry of Education Youth and Sports of the Czech Republic (grant numbers LC06066, MSM0021620813, MSM0021620858 and 1M06014); Czech Science Foundation (grant number 301/07/0607). Funding for open access charge: Ministry of Education Youth and Sports of the Czech Republic.

Conflict of interest statement. None declared.

REFERENCES

1. Jang,S.K., Krausslich,H.G., Nicklin,M.J., Duke,G.M., Palmenberg,A.C. and Wimmer,E. (1988) A segment of the 5′ nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during *in vitro* translation. *J. Virol.*, **62**, 2636–2643.
2. Pelletier,J. and Sonenberg,N. (1988) Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature*, **334**, 320–325.
3. Macejak,D.G. and Sarnow,P. (1991) Internal initiation of translation mediated by the 5′ leader of a cellular mRNA. *Nature*, **353**, 90–94.
4. Sonenberg,N. and Hinnebusch,A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**, 731–745.
5. Doudna,J. and Sarnow,P. (2007) Translation initiation by viral internal ribosome entry sites. In Mathews,M.B., Sonenberg,N. and Hershey,J.W.B. (eds), *Translational Control in Biology and Medicine*. Cold Spring Harbor Monograph 48, pp. 129–153.
6. Elroy-Stein,O. and Merrick,W. (2007) Translation initiation via cellular internal ribosome entry sites. In Mathews,M.B., Sonenberg,N. and Hershey,J.W.B. (eds), *Translational Control in Biology and Medicine*. Cold Spring Harbor Monograph 48, pp. 155–172.
7. Martinez-Salas,E. (2008) The impact of RNA structure on picornavirus IRES activity. *Trends Microbiol.*, **16**, 230–237.
8. Niepmann,M. (2009) Internal translation initiation of picornaviruses and hepatitis C virus. *Biochim. Biophys. Acta.*, doi:10.1016/j.bbagrm.2009.05.002.
9. Hellen,C.U. (2009) IRES-induced conformational changes in the ribosome and the mechanism of translation initiation by internal ribosomal entry. *Biochim. Biophys. Acta.*, doi:10.1016/j.bbagrm.2009.06.001.
10. Balvay,L., Rifo,R.S., Ricci,E.P., Decimo,D. and Ohlmann,T. (2009) Structural and functional diversity of viral IRESes. *Biochim. Biophys. Acta.*, doi:10.1016/j.bbagrm.2009.07.005.
11. Baird,S.D., Turcotte,M., Korneluk,R.G. and Holcik,M. (2006) Searching for IRES. *RNA*, **12**, 1755–1785.
12. Hershey,J., Mathews,M. and Sonenberg,N. (1996) *Translational Control*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
13. Sonenberg,N., Hershey,J. and Mathews,M. (2000) *Translational Control of Gene Expression*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
14. Mathews,M., Sonenberg,N. and Hershey,J. (2007) *Translational Control in Biology and Medicine*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

15. Mokrejs,M., Vopalensky,V., Kolenaty,O., Masek,T., Feketova,Z., Sekyrova,P., Skaloudova,B., Kriz,V. and Pospisek,M. (2006) IRESite: the database of experimentally verified IRES structures (www.iresite.org). *Nucleic Acids Res.*, 34, D125–D130.
16. Mokrejs,M., Vopalensky,V., Masek,T. and Pospisek,M. (2007) Bioinformatical approach to the analysis of viral and cellular internal ribosome entry sites. In Lee B,Kwang (ed.), *New Messenger RNA Research Communications*. Nova Science Publishers, Hauppauge, NY, pp. 133–166.
17. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
18. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
19. Hochsmann,M., Voss,B. and Giegerich,R. (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 53–62.
20. Baird,S.D., Lewis,S.M., Turcotte,M. and Holcik,M. (2007) A search for structurally similar cellular internal ribosome entry sites. *Nucleic Acids Res.*, **35**, 4664–4677.
21. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
22. Baranick,B.T., Lemp,N.A., Nagashima,J., Hiraoka,K., Kasahara,N. and Logg,C.R. (2008) Splicing mediates the activity of four putative cellular internal ribosome entry sites. *Proc. Natl Acad. Sci. USA*, **105**, 4733–4738.
23. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.