Nucleic Acids Research Advance Access published May 14, 2015

doi: 10.1093/nar/gkv489

Web-Beagle: a web server for the alignment of RNA secondary structures

Eugenio Mattei[1], Marco Pietrosanto[1], Fabrizio Ferrè[2,*] and Manuela Helmer-Citterich[1]

[1]Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy and [2]Department of Pharmacy and Biotechnology (FaBiT), University of Bologna Alma Mater, Via Belmeloro 6, 40126 Bologna, Italy

Received February 14, 2015; Revised April 30, 2015; Accepted May 02, 2015

ABSTRACT

Web-Beagle (http://beagle.bio.uniroma2.it) is a web server for the pairwise global or local alignment of RNA secondary structures. The server exploits a new encoding for RNA secondary structure and a substitution matrix of RNA structural elements to perform RNA structural alignments. The web server allows the user to compute up to 10 000 alignments in a single run, taking as input sets of RNA sequences and structures or primary sequences alone. In the latter case, the server computes the secondary structure prediction for the RNAs on-the-fly using RNAfold (free energy minimization). The user can also compare a set of input RNAs to one of five pre-compiled RNA datasets including lncRNAs and 3′ UTRs. All types of comparison produce in output the pairwise alignments along with structural similarity and statistical significance measures for each resulting alignment. A graphical color-coded representation of the alignments allows the user to easily identify structural similarities between RNAs. Web-Beagle can be used for finding structurally related regions in two or more RNAs, for the identification of homologous regions or for functional annotation. Benchmark tests show that Web-Beagle has lower computational complexity, running time and better performances than other available methods.

INTRODUCTION

When annotating functional RNAs, including secondary structure information can improve the accuracy of the alignments. Moreover, using secondary structure information becomes crucial when aligning two RNA sequences with sequence identity <60% (1). The dot-bracket notation represents the most common encoding for the RNA secondary structure. This notation uses a three characters alphabet encoding for unpaired base '.', an open '(' and a closed ')' base pair. Considering its simplicity, the dot-bracket notation is not very informative. Indeed, direct information about the structural context of the nucleotide is not stored into the encoding and *ad hoc* post-processing procedures are required to extract it. Due to the low informative power of dot-bracket notation, the most generally used approaches rely on shifting from the dot-bracket notation to tree-based encodings (2,3), motif description (4,5) or covariance models (6). Unfortunately, using these more informative data structures leads to higher computational complexity, making these approaches computationally expensive when aligning a large number of RNAs. Recently, we introduced a novel encoding, called BEAR, which allows storing secondary structure information into a string of characters (7). Each character of the encoding stores the information about the type and length of the secondary structure elements the nucleotide belongs to (e.g. a set of characters is used to describe stems, and stems of different length are assigned different characters). Moreover, exploiting this powerful yet simple encoding, we computed a substitution matrix of secondary structure elements called MBR (Matrix of Bear-encoded RNAs) that captures transition rates between secondary structures of functionally related RNAs (7). Given a string encoding of the secondary structure and a substitution matrix for the string characters, it becomes natural to apply classic string alignment algorithms to solve the problem of fast and accurate pairwise RNA secondary structure global and local comparison. In our previous work (7), we showed as a proof of concept that a simple variant of a global sequence alignment algorithm able to take as input the BEAR encoding of two RNAs, and using the MBR to guide the alignment, was able to provide good alignment accuracy in a very short computational time.

Here, we present a web server able to compute RNA sequence and structure alignments using dynamic programming, with lower computational complexity than any other state-of-the-art structural aligner, with equal or better accuracy. Compared to the prototype described in (7), the webserver utilizes an improved implementation where parameters where optimized, the algorithm was extended also to lo-

*To whom correspondence should be addressed. Tel: +39 0512088809; Fax: +39 0512099741; Email: fabrizio.ferre@unibo.it

cal alignments, and the statistical significance of each alignment was computed over background distributions. Moreover, the graphical depiction of the output alignments can help the user in assessing the structurally similar regions. As such, the presented web-server provides a useful tool filling a gap for the structural analysis of RNAs. Other web-server methods are based on simultaneous alignment and folding algorithms (8–10). These methods take two RNA primary sequences as input and give as output a *consensus* secondary structure for the two RNAs, which is in other terms their alignment. This approach is not suitable when looking for a specific query structure in a set of target RNAs because the methods will compute their own *consensus* structure that might differ from the original input structure. Nevertheless, the folding and aligning approach is particularly useful for classification purposes. There are also web servers based on tree-based approaches, such as RNAStrAT (2) that uses a tree representation to encode the two input sequences and to compute the alignments. To assess the quality of our alignments, we used four datasets of curated alignments and secondary structures (described in (7)) showing comparable or better performances than the other available methods but reducing the computational time considerably.

METHODS

The Web-Beagle server is based on the algorithm for RNA structural alignment presented in our previous work (7). In particular, a new context-aware representation for RNA secondary structure, called BEAR, is used to encode structural information within a string of characters as long as the primary sequence. The new encoding allows shifting from the dot-bracket encoding, where each nucleotide could only be paired or unpaired, to a more informative representation where each nucleotide is encoded with a character representing the type and length of the structural element it belongs to (loop, internal loop, stem and bulge). Since the encoding can be handled using strings, it is possible to apply dynamic programming algorithms, the same used for amino acids and nucleotides sequence alignments, to perform structural alignments. The alignment procedure is guided by the MBR substitution matrix in conjunction with affine gap costs scoring system. Transition rates in MBR were computed on a dataset of structural alignments of functionally related RNAs as described in (7). Along with structural information, Web-Beagle takes into account also sequence information during the computation, in the form of a numeric *bonus*, to improve the accuracy of the resulting alignment, especially in unstructured regions. In particular, while filling in the dynamic programming matrix, this sequence *bonus* is added to the MBR score to favor the alignment of identical nucleotides. The user is presented with suggested values for gap opening, extension and sequence *bonus* (the procedure used to compute these parameters is described below). Compared to other approaches to align RNA secondary structure, Web-Beagle shows lower computational complexity ($O(n^2)$) and higher accuracy (see next section).

Usage of Web-Beagle

The server offers three comparison options: (i) Two sets comparison; (ii) One set comparison; (iii) Search. With option (i) the server takes as input two sets of RNAs, namely the query and the target sets and compares the RNAs from the query set with the RNAs from the target. The user can choose between the 'All versus All' or 'pairwise' mode. When the 'All versus All' comparison mode is selected, each RNA in the query set will be aligned with all the RNAs in the target set. By contrast, in the 'pairwise' mode the first RNA in the target set will be aligned only to the first RNA in the target set, the second with the second and so on. This option requires the cardinalities of query and target to be equal. In both cases the maximum number of alignments allowed are 10^4. The 'One set comparison' (ii) allows the user to input a single set of RNAs (maximum 300 RNAs) and obtain all the possible comparisons among those RNAs. Finally, using the 'Search' (iii) option, a single input RNA will be compared with one of the precompiled datasets included in the server, namely Human 3′ UTR, Mouse 3′ UTR, Human lncRNAs, Mouse lncRNAs, Structured Rfam (described in Section 5 of Supplementary Materials). All types of comparison require the input to be formatted using FASTA notation or optionally a modified FASTA including the secondary structure in dot-bracket notation or additionally the BEAR notation. If the structure is not provided, the server will compute the secondary structures on-the-fly using the RNAfold algorithm (11). The default parameters of RNAfold were used producing the minimum free energy structure for each RNA. The user can specify different parameters: (1) the gap opening value; (2) the gap extension value and (3) the sequence *bonus*. If the global alignment is selected, the Needleman-Wunsch algorithm (12), modified accordingly to deal with BEAR encoding and sequence *bonus*, would be used to perform the alignments; the Smith-Waterman algorithm (13), if local. After clicking on the submit button, the server gives in output a loading page reporting the link to the results page and information on the program *status*. Once all the alignments are completed, the user will be redirected to the results page. Alternatively, the user can bookmark the link and access the results at her/his own pace. Each results page reports a table containing all the computed pairwise alignments. Each row of the table contains the two input RNA ids and alignment statistics such as sequence and structural identity and structural similarity percentages. The structural measures were computed using the BEAR alphabet as the fraction of the aligned identical BEAR characters over the total alignment pairs (i.e. structural identity), or the fraction of aligned BEAR character encoding for the same structural element, e.g. two characters belonging to stem alphabet (structural similarity). In addition to alignment statistics (such as alignment score and sequence and structural identity), the output reports for each alignment the associated z-score and p-value. These measures can be useful to assess the statistical significance of the alignment. In particular, we suggest that the z-score should be used as the reference measure. Specifically, alignments with a *z*-score >3 should be considered significant (see Supplementary Materials, Section 4). The results can also be sorted ac-

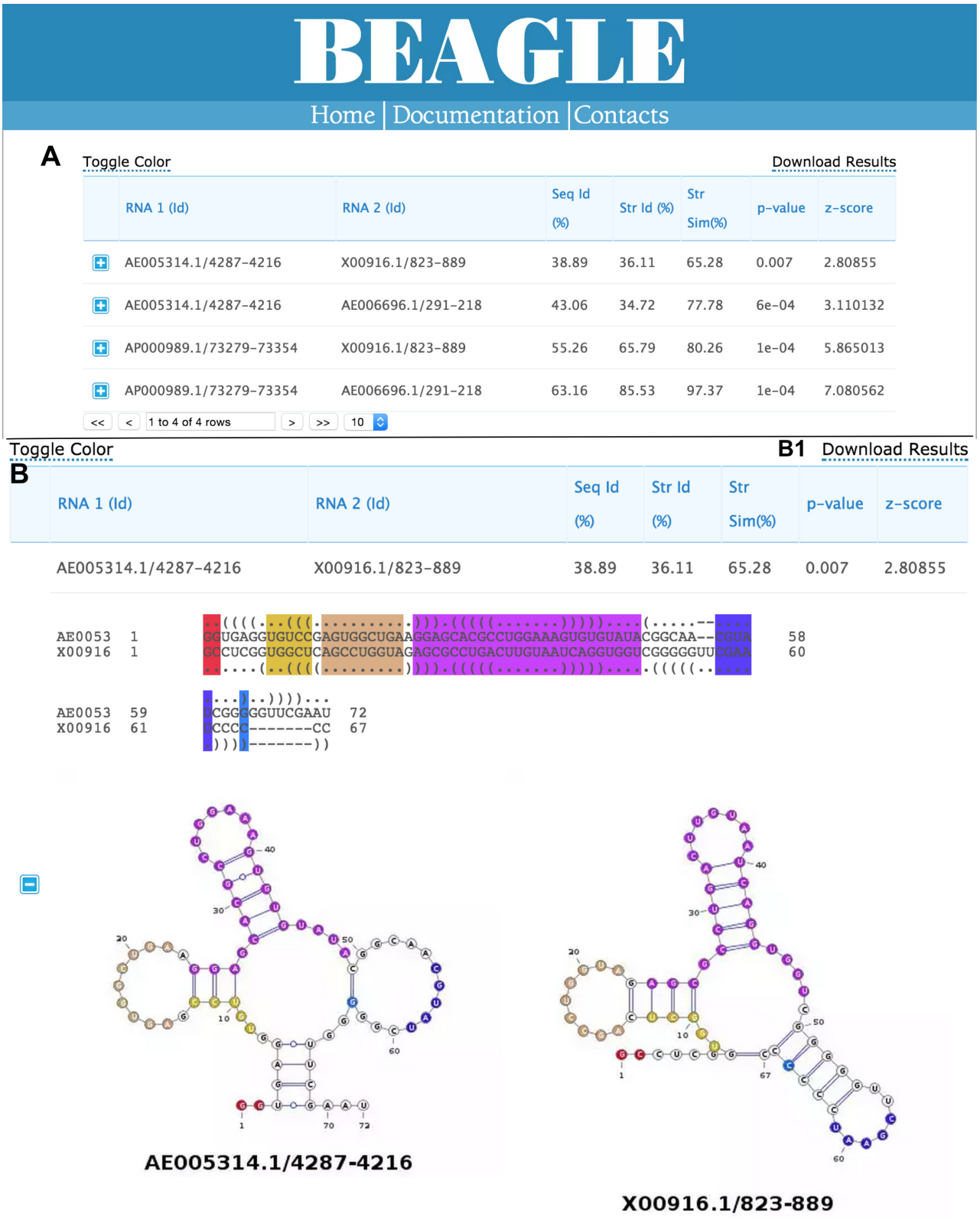

Figure 1. An example of the Web-Beagle output page. (**A**) Results page reporting a table containing all the computed pairwise alignments. Each row of the table contains the two input RNA ids and alignment statistics such as sequence and structural identity and similarity percentages. Moreover, *z*-score and p-value for the statistical evaluation of the alignment are displayed. Clickable column headings allow the user to sort the table by one of the parameters. By clicking on a row, the specific alignment for the selected pair of RNAs will be displayed (**B**). A color-coded representation for the alignment will help the user to identify aligned structural regions between the two RNAs. Along with the textual also a color-coded graphical representation will be displayed. Results can be exported as a flat file using the 'download results' button (b1).

cording to any of the previous parameters. By clicking on a row, the sequence and structure color-coded alignment for that specific pair of RNAs appears along with a color-coded graphical representation of the RNA secondary structures produced using VARNA (14). The color code helps the user to identify aligned structural ungapped regions between the two RNAs (Figure 1). Results can also be exported as static text files. The results page remains available and accessible for two weeks. The website also includes documentation pages with a complete description of the alphabet, the substitution matrix, information on input and output formats, and results of the performance evaluation and the parameters selection.

Experimental results

The method performances have been tested on the same four datasets of curated pairwise structural alignments of RNAs presented in (7). In brief, the four datasets (namely BRAliBase, RNAspa, RNA STRAND and RRS) were obtained by merging together structural and alignment information from different curated databases (1,15–18). The different level of sequence identity in the pairwise alignments and the many classes of non-coding RNA represented in the datasets assure a good level of variability reducing the possibility of biased results. Moreover, the secondary structures are not derived from *consensus* secondary structures, therefore there is not perfect agreement among the aligned structures. In (7) we showed that these reference datasets contain alignments of pairs of RNAs often different in terms of sequence identity and length (Supplementary Table S1). We used these alignments as reference to assess the performances of our method using the SPS (Sum of PairS) score (1). SPS is defined as the fraction of aligned nucleotides in the reference alignment that are correctly aligned (Supplementary Materials, Section 3). Firstly, we computed the best parameters for gap opening, gap extension and sequence *bonus* by running an exhaustive search on the datasets. In particular, we run the algorithm several times on each dataset using different sets of parameters. The goal was to find a common set of parameters that can be used independently from the characteristics of the input RNAs. By comparing all the results together we identified a generic set of parameters maximising the accuracy across the datasets. Note that each dataset was used independently (i.e. the best set of parameters computed on one dataset was then tested on all the other ones), and that the datasets are non-redundant, in order to avoid over-fitting (Supplementary Material, Section 2). We compared the performances of our method with those obtained using other state-of-the-art tools, namely LocARNA (19), gardenia (3), RNAStrAT (2), RNAdistance (11), RNAforester (20) and with those from the sequence-only Needleman-Wunsch algorithm (using the *needle* implementation in the Emboss package) (21). Our method shows a better overall accuracy than other state-of-the-art methods especially when the sequence identity between the input RNAs is <65% (Supplementary Table S1). Additionally, Web-Beagle has lower computational complexity and running time than the other methods. As an additional test, we decided to test the performances of the methods on the same datasets but using the predicted secondary structures obtained using RNAfold (11) (Supplementary Table S2). As expected, all methods are less accurate (except for *needle* that does not use structural information), and we found a positive correlation between the accuracy of the predicted secondary structure and the alignment quality. LocARNA offers the best overall performances in this case, since it does not use the provided structures but recomputes them while aligning. Nevertheless, Web-Beagle performances are not far from those of LocARNA and better than any other method and still maintaining lower computational time (Supplementary Table S3).

SUMMARY

Web-Beagle is a versatile web server for the alignment of RNA secondary structures. It offers three different comparison strategies making it suitable for a wide range of analysis types. The basic usage would be finding the structurally conserved regions in two functionally related RNAs, which could be associated to the RNAs biological roles. Moreover, it can be used for the comparison of the RNAs detected in CLIP experiments (22) looking for common structures implied in the recognition. The web server could also be beneficial for the annotation of new ncRNAs using two different strategies. The ncRNAs can be compared with a precompiled datasets of RNAs looking for common substructures. Alternatively, the ncRNAs could be compared among themselves and then clustered using unsupervised learning methods in order to find new classes of homology. The analysis on four datasets of curated pairwise alignments shows that the server outperforms in computational time and performances the other state-of-art methods. In addition to structural alignments, the server produces a user-friendly representation of the alignments helping users in identifying structural similarities among RNAs. Given the reduced computational time, the server is particularly suitable for annotation analysis and large-scale comparisons. Despite the fact that better performances are obtained using reliable secondary structures, the server still performs well on predicted secondary structures. Moreover, a tool for the accurate comparison of RNA secondary structure will probably become more useful and crucial also thanks to the new experimental techniques for the identification of RNA secondary structure at transcriptome resolution (23), that will make available more (and more reliable) structural data.

AVAILABILITY

Web-Beagle is available at http://beagle.bio.uniroma2.it.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors want to thank Alberto Calderone for his precious help and support.

FUNDING

EPIGEN flagship project MIUR-CNR (to M.H.C.). Funding for open access charge: Programmi di Ricerca di rilevante Interesse Nazionale (PRIN) 2010 (prot. 20108XYHJS_006 to M.H.C.).
Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Guignon,V., Chauve,C. and Hamel,S. (2005) An Edit Distance Between RNA Stem-Loops. In: Consens,M and Navarro,G (eds). *String Processing and Information Retrieval Lecture Notes in Computer Science*, Springer. Berlin Heidelberg, Vol. **3772**, pp. 335–347.
3. Blin,G., Denise,A., Dulucq,S., Herrbach,C. and Touzet,H. (2007) Alignments of RNA structures. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 309–322.
4. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
5. Chang,T.-H., Huang,H.-D., Chuang,T.-N., Shien,D.-M. and Horng,J.-T. (2006) RNAMST: efficient and flexible approach for identifying RNA structural homologs. *Nucleic Acids Res.*, **34**, W423–W428.
6. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
7. Mattei,E., Ausiello,G., Ferrè,F. and Helmer-Citterich,M. (2014) A novel approach to represent and compare RNA secondary structures. *Nucleic Acids Res.*, **42**, 6146–6157.
8. Havgaard,J.H., Lyngsø,R.B. and Gorodkin,J. (2005) The Foldalign web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res.*, **33**, 650–653.
9. Fu,Y., Sharma,G. and Mathews,D.H. (2014) Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. *Nucleic Acids Res.*, **42**, 13939–13948.
10. Smith,C., Heyne,S., Richter,A.S., Will,S. and Backofen,R. (2010) Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. *Nucleic Acids Res.*, **38**, W373–W377.
11. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
12. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
13. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
14. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
15. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
16. Horesh,Y., Doniger,T., Michaeli,S. and Unger,R. (2007) RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules. *BMC Bioinformatics*, **8**, 366.
17. Andronescu,M., Bereg,V., Hoos,H.H. and Condon,A. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.
18. Widmann,J., Stombaugh,J., McDonald,D., Chocholousova,J., Gardner,P., Iyer,M.K., Liu,Z., Lozupone,C.A., Quinn,J., Smit,S. *et al.* (2012) RNASTAR: an RNA STructural Alignment Repository that provides insight into the evolution of natural and artificial RNAs. *RNA*, **18**, 1319–1327.
19. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
20. Höchsmann,M., Töller,T., Giegerich,R. and Kurtz,S. (2003) Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf.*, **2**, 159–168.
21. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
22. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M., Jungkamp,A., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
23. Wan,Y., Qu,K., Ouyang,Z. and Chang,H.Y. (2013) Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nat. Protoc.*, **8**, 849–869.