Published online 26 May 2021 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* ***W67–W71***
https://doi.org/10.1093/nar/gkab400

BRIO: a web server for RNA sequence and structure motif scan

Andrea Guarracino [1,†], Gerardo Pepe [1,†], Francesco Ballesio [1], Marta Adinolfi [1], Marco Pietrosanto [1], Elisa Sangiovanni[1], Ilio Vitale [2,3], Gabriele Ausiello[1] and Manuela Helmer-Citterich [1,*]

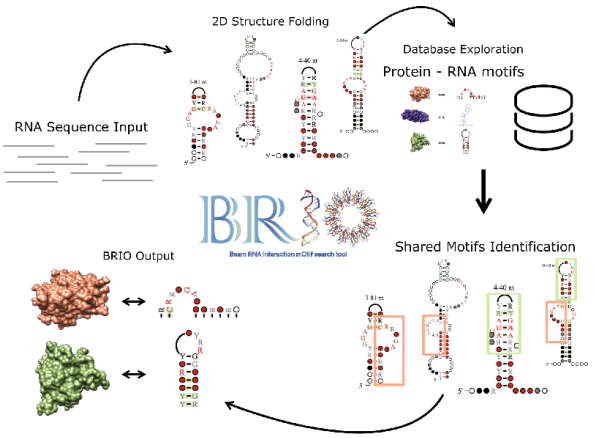
[1]Department of Biology, University of Rome "Tor Vergata", Rome, Italy, [2]IIGM - Italian Institute for Genomic Medicine, c/o IRCSS Candiolo, Italy and [3]Candiolo Cancer Institute, FPO - IRCCS, Candiolo, Italy

Received February 21, 2021; Revised April 27, 2021; Editorial Decision April 28, 2021; Accepted May 22, 2021

ABSTRACT

The interaction between RNA and RNA-binding proteins (RBPs) has a key role in the regulation of gene expression, in RNA stability, and in many other biological processes. RBPs accomplish these functions by binding target RNA molecules through specific sequence and structure motifs. The identification of these binding motifs is therefore fundamental to improve our knowledge of the cellular processes and how they are regulated. Here, we present BRIO (BEAM RNA Interaction mOtifs), a new web server designed for the identification of sequence and structure RNA-binding motifs in one or more RNA molecules of interest. BRIO enables the user to scan over 2508 sequence motifs and 2296 secondary structure motifs identified in *Homo sapiens* and *Mus musculus*, in three different types of experiments (PAR-CLIP, eCLIP, HITS). The motifs are associated with the binding of 186 RBPs and 69 protein domains. The web server is freely available at http://brio.bio.uniroma2.it.

GRAPHICAL ABSTRACT



INTRODUCTION

Molecular interactions are crucial for most biological processes in the cell. The landscape of all possible molecular interactions depends on the actors involved in the interplay. These actors include RNA-binding proteins (RBPs), which play a central role in RNA metabolism, regulating the transcripts throughout their life cycle, and in particular modulating mRNA localization, splicing, stability, and translation (1). Moreover, RBPs are also involved in the regulation of non-coding RNAs. The human proteome includes over 2000 RBPs (2), each one having specific functions and target RNAs. This occurs through the interaction of particular RNA binding domains (RBDs) with specific RNA-binding motifs (3). Of note, some RBPs recognize their target molecules via nucleotide patterns, while others favour specific RNA structural motifs (4). The understanding of RBP function and the identification of the binding motifs are required to get insights into the regulatory mechanisms in which these proteins are involved.

*To whom correspondence should be addressed. Tel: +39 0672594324; Fax: +39 062023500; Email: manuela.helmer.citterich@uniroma2.it
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Present address: Marta Adinolfi, Department of Experimental Oncology, IEO, European Institute of Oncology IRCCS, Milan, Italy.

Table 1. Number of sequence and structure motifs identified in *Homo sapiens* and *Mus musculus*

Motif type	*Homo sapiens*	*Mus musculus*	Total
Sequence	2112	184	2296
Structure	2319	189	2508

Recent advances in high-throughput methods to assess targets of RBPs *in vitro* and *in vivo* (5,6) have led to the development of several resources to identify RBPs and their binding motifs. Some tools find sequence motifs from inputs of target RNAs, while others focus on RNA secondary structure (7). Here, we describe BRIO (BEAM RNA Interaction mOtifs), a new web server allowing users to easily search for sequence and structure RNA-binding motifs in one or more RNA molecules of interest. The dataset contains 2508 sequence and 2296 structure motifs associated with the binding of 186 unique proteins and 69 unique protein domains. The motifs were previously identified in *Homo sapiens* and *Mus musculus*, in three different types of experiments (PAR-CLIP, eCLIP, HITS) analyzed in Adinolfi *et al.* (8). The structure motifs are encoded using BEAR (Brand nEw Alphabet for RNA), a powerful context-aware structural string encoding we previously developed and applied in our research on RNA structural characterization and conservation (9,10). This encoding not only stores information about the 'paired' or 'unpaired' status of a nucleotide, but also takes into account the type and length of the secondary structure element (SSE) to which the nucleotide belongs. This means that the BEAR encoding describes RNA secondary structure in a more comprehensive way, but with low algorithmic complexity given the simple string representation.

MATERIALS AND METHODS

Workflow

The overall workflow of BRIO is illustrated in Figure 1. After the pre-processing step (i.e. the calculation of the secondary structure, when not provided by the user), each input RNA molecule is scanned for the identification of sequence and structure motifs. The user can decide to scan subsets of motifs present in our database, selecting the species and the type of experiments in which the motifs were identified.

Here, we provide a description of the datasets, the algorithm, its input and the results provided by the web server.

Datasets

BRIO scans the input RNA molecules using sequence and structure motifs (Table 1). The motifs were previously identified in 186 RBPs and 69 protein domains analyzing 228 PAR-CLIP, eCLIP, HITS-CLIP experiments in *H. sapiens* and *M. musculus* (8,11).

The CLIP experiments are both from human (hg19; 74 experiments performed in 13 different cell lines, principally HEK293 cells) and mouse (mm9; 30 experiments performed in 7 different cell lines, mostly brain and embryonic stem cell), while eCLIP data comes from studies performed in human chronic myelogenous leukemia (K562) and hepatocellular carcinoma (HepG2) cells. The dataset contains 186 RBPs for both human and mouse, and a total of 69 unique protein domains, 12 of which are shared between the two species (51 protein domains unique for human and 6 protein domains unique for mouse). The secondary structure motifs are represented using the BEAR encoding.

The whole dataset of sequence and structure motifs is available for download at the BRIO website.

Input

Users can either input only the RNA primary sequence(s) of interest, or the sequence(s) and the corresponding secondary structure(s) in dot-bracket notation, all in FASTA or multiFASTA format. Alternatively, the input can be uploaded as a text file. The user can choose the preferred type of structural representation, e.g. Minimum Free Energy (MFE) or centroid. Sequences submitted without the secondary structure are automatically folded using RNAfold (in this case, the MFE structure is computed by default) (12). Finally, the dot-bracket notation is translated into a BEAR string. The input RNA molecules are required to be at least three nucleotides long, and shorter than 3000. To search for structure motifs, sequences in input are requested to be at least 50 nucleotides long. At most 100 sequences can be submitted at a time.

Users can choose to compare their RNA molecules to the whole dataset of motifs, or to only *H. sapiens* or *M. musculus* motifs. It is also possible to select a subset of the experiments analyzed (PAR-CLIP, eCLIP or HITS), considering that the eCLIP datasets were obtained from experiments performed only in *H. sapiens*.

Procedure

To identify the motifs, BRIO relies on substitution matrices: the Matrix of BEAR encoded RNA (MBRs) for secondary structure elements (9) and a classic substitution matrix for nucleotides (with score 3 for nucleotide matching and −2 otherwise). For each motif, the algorithm scans the corresponding Position Frequency Matrix in any single input RNA using a sliding-window ungapped alignment. Next, the score of the best match is compared to the score threshold associated with the motif (for more information, see (8)). Finally, the one-sided Fisher's Test is applied to determine if a motif is enriched in the input RNA molecules with respect to a set of background RNAs. This test evaluates whether the motif is identified with a significantly greater proportion in the RNA over the background. The background set of RNA molecules can be specified by the user. By default, all 85640 sequences from Rfam 14.3 are considered (13).

Output

BRIO returns a collection of protein binding motifs identified in the input RNA molecules. The results are shown as tables, in two different views: 'Enriched Motifs' and 'Sequences' (Figure 2). The 'Enriched Motifs' tab shows the motifs identified in the input molecules as a whole, while

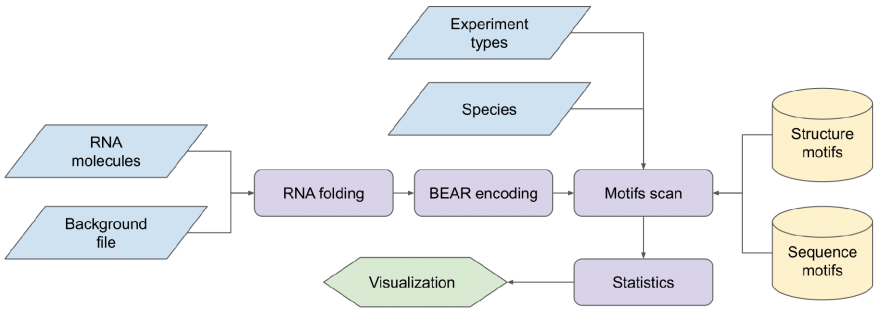
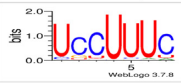
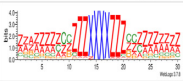
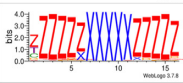
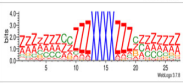
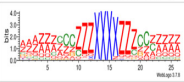


Figure 1. The Workflow used by BRIO to search for motifs in the input RNA sequences. The input RNA molecules are folded, encoded using the BEAR alphabet (9), and scanned for sequence and structure motifs. Filters on species and type of experiment can also be applied. A background file can be used as a comparison with the input sequences. The presence of enriched motifs is determined using Fisher's exact test.

A

Enriched Motifs | Sequences

Logo	Type	Region	Coverage	p-value	Experiment	Protein	Domains	Cell line	Experiment Info	Organism	Download
	Sequence	CDS	1	0.013	PARCLIP	WTAP	-	-	Link	*Homo sapiens*	Download
	Sequence	UTR	1	0.016	PARCLIP	ELAVL1MNASE	RRMx1; RRMx2; RRMx3;	-	Link	*Homo sapiens*	Download
	Structure	UTR	1	0.017	PARCLIP	FMR1	-	-	Link	*Homo sapiens*	Download
	Structure	CDS	1	0.02	PARCLIP	FMR1	-	-	Link	*Homo sapiens*	Download
	Structure	CDS	1	0.024	PARCLIP	HuRMNase	RRMx1; RRMx2; RRMx3;	-	Link	*Homo sapiens*	Download
	Sequence	CDS	1	0.031	PARCLIP	FMR1	-	-	Link	*Homo sapiens*	Download
	Structure	CDS	1	0.033	PARCLIP	CAPRIN1	G3BP1-binding;	-	Link	*Homo sapiens*	Download

B

Enriched Motifs | Sequences

	Name	# Sequence motifs	# Structure motifs	Length
+	chr1:149783661-149783992(-)	3	7	332

Start	End	Enriched Motif	Type	Protein	Experiment
90	96	UCUUUUC	Sequence	ELAVL1MNASE	PARCLIP
98	127	bb"""dddd[cccmmmmcccdddd333bb]	Structure	FMR1	PARCLIP
99	124	b"""dddd[cccmmmmcccdddd333	Structure	CAPRIN1	PARCLIP
99	125	b"""dddd[cccmmmmcccdddd333b	Structure	HuRMNase	PARCLIP
138	144	CCUGCUC	Sequence	WTAP	PARCLIP
184	190	AGAAGGA	Sequence	FMR1	PARCLIP
280	295	:eeeeeoooooeeee	Structure	QKI	PARCLIP
281	296	eeeeeoooooeeeee	Structure	FMR1	PARCLIP
281	297	eeeeeoooooeeeee:	Structure	AGO2	PARCLIP
282	295	eeeeoooooeeee	Structure	ALKBH5	PARCLIP

Figure 2. The output is composed of two views: (**A**) the 'Enriched Motifs' table, showing the enriched motifs found in the input RNA molecules, and (**B**) the 'Sequence' table, showing the results for each input sequence provided. The column description is displayed in both tables when the pointer hovers over the column name (not shown in the figure).

the 'Sequences' tab shows the results for the different input RNA molecules. The content of each table can be sorted in ascending or descending order with respect to each column.

In the 'Enriched Motifs' table (Figure 2A), by default, the motifs identified in the input RNA molecules are sorted according to the one-sided Fisher's Test *P*-value. For each motif, the table reports:

- the logo of the motif in the qBEAR alphabet for structure motifs or in IUPAC nucleic acid notation for sequence motifs;
- the type of the motif: sequence or structure;
- the type of mapping regions from the GENCODE annotation (8,14) of the RNAs datasets where the motif was originally found. The annotation includes UTR, CDS, and transcript for those RBPs known to act in the nucleus on unspliced RNAs;
- the coverage, which represents the number of input RNA molecules in which the motif is identified divided by the total number of query molecules;
- the one-sided Fisher's Test *P*-value;
- the type of experiment;
- the protein associated with the RNA sequence or secondary structure motif in the CLIP experiment reported in Adinolfi *et al*. (8);
- the protein domain associated with the RNA secondary structure motif (note that this information is not always available);
- the cell line used in the eCLIP experiments;
- the link to the experiment page (for eCLIP data), or to the corresponding article (for PAR-CLIP and HITS data);
- the organism in which the experiment was performed (*H. sapiens* or *M. musculus*).

The last column reports the link for the download of the information on the motif described in each row.

In the 'Sequences' table (Figure 2B), by default, the entries are sorted according to the start position of the motif in the sequence. For each entry, the table reports:

- the start and the end position of the motif in the selected sequence;
- the representation of the motif in BEAR alphabet for structure motifs or in IUPAC nucleic acid notation for sequence motifs;
- the type of the motif: sequence or structure;
- the protein associated with the RNA sequence or secondary structure motif in the experiments analyzed by Adinolfi *et al*. (8);
- the type of the experiment (PAR-CLIP, eCLIP, HITS).

RESULTS

We describe, here, two examples of the use of BRIO web server to search for putative binding proteins. As input RNA molecules, we selected RNA sequences belonging to Rfam families whose interactors are already described in the literature, and checked the server capability to identify known motifs. These input datasets are also available in the BRIO web server.

In the first example, we analyzed sequences of the *Herpesvirus saimiri* U RNAs (HSUR) family (RF01802). HSURs are viral small RNAs that seem to be involved in the regulation of the stability of host mRNAs. They have been shown to associate, *in vivo*, with two proteins: heterogeneous nuclear ribonucleoprotein D (HNRNPD) and ELAV like RNA binding protein 1 (ELAVL1, best known as HuR) (15). In particular, HuR is known to contribute to the stability of mRNAs and to recognize short sequence motifs of low statistical significance. The HSUR family consists of four sequences, two of which share 100% sequence identity. In our analysis, three different HSUR sequences were given as input to BRIO and launched against the PAR-CLIP dataset of binding motifs. The web server identified several sequence motifs with high coverage (= 1) and low *P*-values (in the $[10^{-6}–10^{-5}]$ range). After sorting the solutions according to ascending *P*-values, we found in the top positions sequence motifs binding the PAZ Piwi domain of AGO, and the HuR and ELAVL1MNASE RNA binding domains. With a significant score, we also detected PUM2, a post-transcriptional repressor interacting with the 3′ UTRs of its target mRNAs. Of note, the structure motifs identified in this run did not show full coverage or a very low *P*-value, although the HuR, AGO and ELAVL1MNASE protein binding motifs were listed with *P*-values in the [0.024–0.028] range and found in two out of the three input sequences. The identification of structure motifs is generally dependent on the number of input sequences (the more, the better). The best hits can be identified also by inspecting the 'Sequences' table in which the motifs identified in the single sequences are shown. In particular, this view allows the user to see the motifs identified according to the position in the sequence, but also sorted by protein name. Using this last view, it is easy to see if single proteins or domains are repeated, and if the motifs identified are sequence and/or structure motifs.

In the second example, we used BRIO to search for protein binding motifs in ten U2 spliceosomal RNA sequences (the first listed in the RF00004 Rfam family) launched against the eCLIP dataset of binding motifs. In the 'Enriched Motifs' Table, a structural motif associated with the SF3B1 protein showed a very low *P*-value. SF3B1 is known to interact with U2 small nuclear ribonucleoprotein (snRNP), which is composed of U2 snRNAs and their associated polypeptide. Of note, several additional sequence and structure motifs are listed in the BRIO output table, some of which are cytoplasmic and therefore the interaction is not possible, while others can be used as suggestions for or to confirm an experimental test.

DISCUSSION

The BRIO web server allows researchers to identify sequence and structure protein binding motifs in a set of one or more RNA molecules. The BRIO dataset encompasses 2296 sequence and 2508 structure motifs associated with 186 RNA binding proteins and 69 protein domains from several CLIP experiments. BRIO takes advantage of the BEAR encoding to represent structural motifs. This string encoding allows us to include the structural context of each nucleotide in the secondary structure representation, without increasing algorithm complexity. To the best of our knowledge, no other existing web server offers the possibil-

ity to search for sequence and structure motifs associated with RNA binding proteins. Indeed, few other web servers are available to identify motifs in RNA sequences but, to our knowledge, only the RegRNA 2.0 server (16) allows the user to search also in a database of sequence and structural motifs (such as splicing sites, polyadenylation sites, motifs in 5′ and 3′ UTR, etc) that are not specifically associated with RNA binding proteins.

Together with its friendly interface, BRIO can support scientists in their investigations on groups of RNA molecules of interest, their putative RBPs, and the roles these proteins play in RNA regulation.

DATA AVAILABILITY

The web server is freely available at http://brio.bio.uniroma2.it. The source code and all the data are available at https://github.com/helmercitterich-lab/BRIO.

ACKNOWLEDGEMENTS

We acknowledge ELIXIR-IIB (elixir-italy.org), the Italian Node of the European ELIXIR infrastructure (elixir-europe.org), and CINECA for supporting FB in the development of this work through the ELIXIR-IIB HPC@CINECA call.

FUNDING

Associazione Italiana per la Ricerca sul Cancro (AIRC) [IG 23539 to M.H.C.].
Conflict of interest statement. None declared.

REFERENCES

1. Glisovic,T., Bachorik,J.L., Yong,J. and Dreyfuss,G., (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.
2. Corley,M., Burns,M.C. and Yeo,G.W. (2020) How RNA-binding proteins interact with RNA: molecules and mechanisms. *Mol. Cell*, **78**, 9–29.
3. Lunde,B.M., Moore,C. and Varani,G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.
4. Cook,K.B., Hughes,T.R. and Morris,Q.D. (2015) High-throughput characterization of protein-RNA interactions. *Brief. Funct. Genomics*, **14**, 74–89.
5. Hannigan,M.M., Zagore,L.L. and Licatalosi,D.D. (2018) Mapping transcriptome-wide protein-RNA interactions to elucidate RNA regulatory programs. *Quant Biol.*, **6**, 228–238.
6. Ferrè,F., Colantoni,A. and Helmer-Citterich,M. (2016) Revealing protein-lncRNA interaction. *Brief. Bioinform.*, **17**, 106–116.
7. Sasse,A., Laverty,K.U., Hughes,T.R. and Morris,Q.D. (2018) Motif models for RNA-binding proteins. *Curr. Opin. Struct. Biol.*, **53**, 115–123.
8. Adinolfi,M., Pietrosanto,M., Parca,L., Ausiello,G., Ferrè,F. and Helmer-Citterich,M. (2019) Discovering sequence and structure landscapes in RNA interaction motifs. *Nucleic Acids Res.*, **47**, 4958–4969.
9. 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.
10. Pietrosanto,M., Adinolfi,M., Guarracino,A., Ferrè,F., Ausiello,G., Vitale,I. and Helmer-Citterich,M. (2021) Relative information gain: Shannon entropy-based measure of the relative structural conservation in RNA alignments. *NAR Genom Bioinform*, **3**, lqab007.
11. Blin,K., Dieterich,C., Wurmus,R., Rajewsky,N., Landthaler,M. and Akalin,A. (2015) DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, **43**, D160–D167.
12. 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.
13. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: Wikipedia, clans and the 'decimal' release. *Nucleic Acids Res.*, **39**, D141–D145.
14. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
15. Cook,H.L., Mischo,H.E. and Steitz,J.A. (2004) The Herpesvirus saimiri small nuclear RNAs recruit AU-rich element-binding proteins but do not alter host AU-rich element-containing mRNA levels in virally transformed T cells. *Mol. Cell. Biol.*, **24**, 4522–4533.
16. Chang,T.-H., Huang,H.-Y., Hsu,J.B.-K., Weng,S.-L., Horng,J.-T. and Huang,H.-D. (2013) An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC Bioinformatics*, **14**(Suppl. 2), S4.