Bioinformatics, 34(6), 2018, 1058–1060
doi: 10.1093/bioinformatics/btx704
Advance Access Publication Date: 31 October 2017
Applications Note

OXFORD

Structural bioinformatics

BEAM web server: a tool for structural RNA motif discovery

Marco Pietrosanto[1,*], Marta Adinolfi[1], Riccardo Casula[1], Gabriele Ausiello[1], Fabrizio Ferrè[2] and Manuela Helmer-Citterich[1]

[1]Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, 00133 Rome, Italy and [2]Department of Pharmacy and Biotechnology, University of Bologna Alma Mater, 40126 Bologna, Italy

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 22, 2017; revised on September 15, 2017; editorial decision on October 26, 2017; accepted on October 30, 2017

Abstract

Motivation: RNA structural motif finding is a relevant problem that becomes computationally hard when working on high-throughput data (e.g. eCLIP, PAR-CLIP), often represented by thousands of RNA molecules. Currently, the BEAM server is the only web tool capable to handle tens of thousands of RNA in input with a motif discovery procedure that is only limited by the current secondary structure prediction accuracies.

Results: The recently developed method BEAM (BEAr Motifs finder) can analyze tens of thousands of RNA molecules and identify RNA secondary structure motifs associated to a measure of their statistical significance. BEAM is extremely fast thanks to the BEAR encoding that transforms each RNA secondary structure in a string of characters. BEAM also exploits the evolutionary knowledge contained in a substitution matrix of secondary structure elements, extracted from the RFAM database of families of homologous RNAs. The BEAM web server has been designed to streamline data pre-processing by automatically handling folding and encoding of RNA sequences, giving users a choice for the preferred folding program. The server provides an intuitive and informative results page with the list of secondary structure motifs identified, the logo of each motif, its significance, graphic representation and information about its position in the RNA molecules sharing it.

Availability and implementation: The web server is freely available at http://beam.uniroma2.it/ and it is implemented in NodeJS and Python with all major browsers supported.

Contact: marco.pietrosanto@uniroma2.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Structural motif finding in RNA is a growing branch in the field of computational biology, especially given the rise of new experimental techniques capable of probing structural contexts at single nucleotide resolution (Lu and Chang, 2016; Wan *et al.*, 2014), which in turn allows for more accurate secondary structure predictions (Lorenz *et al.*, 2016). The question that is usually addressed by looking for structural motifs revolves around finding structural determinants associated to specific functions (e.g. protein interaction specificity, main actor of certain interactions or behaviours), for example the determinant of Staufen-RNA specificity (LeGendre *et al.*, 2013). In this sense, data coming from high-throughput *in vivo* experiments such as HITS-CLIP, PAR-CLIP, iCLIP or eCLIP provide a perfect playground, for they are often composed by a large number of molecules (up to 50k RNAs, or even more) with a shared binding ability. Current structural motif finders cannot work over low input size thresholds (i.e. 1000 molecules is a hard limit for most) and, to our knowledge, only our method BEAM (Pietrosanto *et al.*, 2016), and the most recent SMARTIV (Polishchuk *et al.*, 2017), can tackle these large inputs. Ours is, however, the only web server that can both discover motifs in large windows (e.g. downstream or

upstream a binding site or along a 500 nt RNA) and with tens of thousands of molecules.

2 Materials and methods

We extended BEAM, for which the user must provide pre-computed RNA secondary structures converted in BEAR notation through a separate encoding software (Mattei *et al.*, 2014), by letting users upload a standard FASTA format file containing only the RNA sequences. In this case users can choose one out of two possible structural prediction methods: RNAfold from the Vienna Package (Gruber *et al.*, 2008) or MaxExpect from RNAstructure (Reuter and Mathews, 2010). Then, from the dot-bracket notation, RNA structures will be automatically converted into the BEAR encoding. Users can also directly upload a file containing RNA sequences in FASTA format containing the corresponding secondary structure prediction in dot-bracket or in BEAR notation. The same data can be also pasted in a text-area. The users can also upload a background dataset for computing the motif significance; alternatively, the server provides automatic background generation by using RNA sequences from Rfam seed data with a filter that guarantees similar length and amount of structural content with respect to the input (Mattei *et al.*, 2015).

Another available feature is the possibility to upload a BED file, which is the most common output format for CLIP-Seq analysis tools: the webserver will manage all the needed processing steps (namely: extension of the intervals, intersection with a feature file to extract only specific genomic regions, sequence retrieval, secondary structure prediction and motif discovery).

In the output page, a table is provided containing all the RNA structure motifs identified. This table shows the following information, for each identified motif: a WebLogo (Crooks *et al.*, 2004) picture in qBEAR alphabet (Pietrosanto *et al.*, 2016), statistic values (such as *P*-value, coverage, BEAM score etc.), a histogram of the motif position distribution with respect to the 5′ of each RNA, and the motif model structure picture obtained using VARNA (Darty *et al.*, 2009). It is also possible to expand the motif results by listing all sequences with a graphic illustration of the motif position relative to the sequence length, along with the dot-bracket and sequence alignments. This representation of a structural motif provides researchers with an overview of how sub-structures could be involved in the function shared by all, or a subset, of the input RNAs, such as protein–RNA or RNA–RNA interactions.

3 Results

For large datasets the application was tested, along with about a hundred unique datasets, with CLIP-Seq data for SLBP (Zhang *et al.*, 2012) (stem-loop binding protein)-interacting RNAs (GSE62154), LIN28A (Cho *et al.*, 2012; Zeng *et al.*, 2016) targets and the other DoRiNA (Blin *et al.*, 2015) datasets, for which all the significant motifs retrieved were presented in the original work (Pietrosanto *et al.*, 2016). In some datasets, we analysed more than 35K RNA sequences in a single run. In particular SLBP has been known to interact with dsRNA (Brooks *et al.*, 2015; Li *et al.*, 2010; Zhang *et al.*, 2012), and the accurate structural context (Fig. 1) can be retrieved with little effort. The server has been tested on datasets of up to 100k RNAs and up to 5 motifs per dataset, and the computational time is similar to that of the BEAM standalone version, as the post analyses take negligible time to compute. The only consistent time added is the time taken by the secondary structure

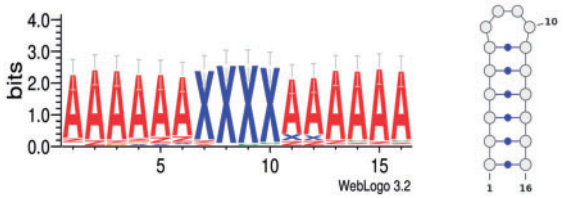

Fig. 1. SLBP putative interaction motif. On the left, a logo describing the identified structural motif is shown in qBEAR notation, in which A stands for medium size stem of a hairpin, and X for a short size terminal loop. On the right, an instance of the motif secondary structure is shown

prediction and eventually the genomic interval pre-processing by means of BEDtools (Quinlan, 2014), if used. Current limitations and associated graphs are reported in the Supplementary Material and in the online documentation.

4 Conclusion

The BEAM web server is a web application that allows the analyses of RNA datasets in search of secondary structure motifs. It can work with tens of thousands of molecules (see Supplementary Material for more information) with a length up to 2000 nt (if folding predictors are used, different limits are applied, see Supplementary Material).

Therefore, this is the only tool that can tackle the task of structural motif discovery of big datasets (such as CLIP-Seq) along their full length.

Moreover, our framework enables researchers to access the tool without additional scripting thanks to the automation provided by the web server. For advanced users, this resource is a fast test ground for BEAM and a precious time saver for downstream analysis.

Acknowledgements

We acknowledge ELIXIR-IIB (elixir-italy.org), the Italian Node of the European ELIXIR infrastructure (elixir-europe.org) and CINECA for supporting the development of ‘BEAM: a new method for discovering RNA secondary structure motifs – Validation’ through the ELIXIR-IIB HPC@CINECA call.

Funding

This work was supported by the EPIGEN Flagship Project MIUR-CNR to M.H.C.

Conflict of Interest: none declared.

References

Blin,K. *et al.* (2015) DoRiNA 2.0-upgrading the dorina database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, **43**, D160–D167.

Brooks,L. *et al.* (2015) A multiprotein occupancy map of the mRNP on the 3′ end of histone mRNAs. *RNA*, **21**, 1943–1965.

Cho,J. *et al.* (2012) LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell*, **151**, 765–777.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Darty,K. *et al.* (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Gruber,A.R. *et al.* (2008) The Vienna RNA websuite. *Nucleic Acids Res.*, **36**, W70–W74.

LeGendre,J.B. *et al*. (2013) RNA targets and specificity of staufen, a double-stranded RNA-binding protein in *Caenorhabditis elegans*. *J. Biol. Chem*., **288**, 2532–2545.

Li,X. *et al*. (2010) Predicting *in vivo* binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, **16**, 1096–1107.

Lorenz,R. *et al*. (2016) RNA folding with hard and soft constraints. *Algorithms Mol. Biol*., **11**, 8.

Lu,Z. and Chang,H.Y. (2016) Decoding the RNA structurome. *Curr. Opin. Struct. Biol*., **36**, 142–148.

Mattei,E. *et al*. (2014) A novel approach to represent and compare RNA secondary structures. *Nucleic Acids Res*., **42**, 6146–6157.

Mattei,E. *et al*. (2015) Web-Beagle: a web server for the alignment of RNA secondary structures. *Nucleic Acids Res*., **43**, W493–W497.

Pietrosanto,M. *et al*. (2016) A novel method for the identification of conserved structural patterns in RNA: From small scale to high-throughput applications. *Nucleic Acids Res*., **44**, 8600–8609.

Polishchuk,M. *et al*. (2017) A combined sequence and structure based method for discovering enriched motifs in RNA from in vivo binding data. *Methods*, **118–119**, 73–81.

Quinlan,A.R. (2014) BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinf*., **47**, 11.12.1–11.12.34.

Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.

Wan,Y. *et al*. (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.

Zeng,Y. *et al*. (2016) Lin28A binds active promoters and recruits Tet1 to regulate gene expression. *Mol. Cell*, **61**, 153–160.

Zhang,M. *et al*. (2012) Interaction of histone mRNA hairpin with stem-loop binding protein and regulation of the SLBP-RNA complex by phosphorylation and proline isomerization. *Biochemistry*, **51**, 3215–3231.