Sequence analysis

CROSSalive: a web server for predicting the in vivo structure of RNA molecules

Riccardo Delli Ponti\textsuperscript{1,2}, Alexandros Armaos\textsuperscript{1,2}, Andrea Vandelli\textsuperscript{1,2} and Gian Gaetano Tartaglia\textsuperscript{1,2,3,4,}\textsuperscript{*}

\textsuperscript{1}Gene Function and Evolution, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain, \textsuperscript{2}Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain, \textsuperscript{3}Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain and \textsuperscript{4}Department of Biology ‘Charles Darwin’, Sapienza University of Rome, Rome 00185, Italy

\textsuperscript{*}To whom correspondence should be addressed.

Received on May 24, 2019; revised on May 24, 2019; editorial decision on August 16, 2019; accepted on August 26, 2019

Abstract

Motivation: RNA structure is difficult to predict in vivo due to interactions with enzymes and other molecules. Here we introduce CROSSalive, an algorithm to predict the single- and double-stranded regions of RNAs in vivo using predictions of protein interactions.

Results: Trained on icSHAPE data in presence (m6a\textsuperscript{+}) and absence of N6 methyladenosine modification (m6a\textsuperscript{-}), CROSSalive achieves cross-validation accuracies between 0.70 and 0.88 in identifying high-confidence single- and double-stranded regions. The algorithm was applied to the long non-coding RNA Xist (17 900 nt, not present in the training) and shows an Area under the ROC curve of 0.83 in predicting structured regions.

Availability and implementation: CROSSalive webserver is freely accessible at http://service.tartaglialab.com/new_submission/crossalive

Contact: gian.tartaglia@crg.es

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The in vitro structure of an RNA differs from that in vivo for the action of molecules such as RNA-binding proteins (Livi et al., 2015). The complex mechanisms contributing to the formation of structure in vivo are poorly characterized and previous analysis suggests a prevalence of single-stranded regions for all RNA types (Rouskin et al., 2014), although conservation of double-stranded regions has been observed for specific non-coding RNAs (Spitale et al., 2015). In the cellular environment RNA undergoes a number of modifications such as methylation that influence both stability and turnover of the whole transcriptome (Liu and Jia, 2014). Mettl3 is a key component of the complex that methylates adenosine residues at the N\textsubscript{6} (m\textsubscript{6}a) and plays a central role in determining RNA structure in vivo. Indeed, a method of probing RNA structure using the chemical probe NAI-N3 (icSHAPE) indicated that m\textsubscript{6}a promotes transition from double- to single-stranded regions (Spitale et al., 2015). Through analysis of icSHAPE data we developed the CROSSalive method for the prediction of RNA secondary structure in vivo. One important part of our approach is the use of catRAPID predictions of protein interactions to classify single- and double-stranded regions of RNA molecules (Bellucci et al., 2011). catRAPID estimates the binding through van der Waals, hydrogen bonding and secondary structure properties of both protein and RNA sequences.

2 Workflow and implementation

CROSSalive profiles a RNA sequence computing the corresponding secondary structure in vivo with (m\textsubscript{6}a\textsuperscript{+}) and without (m\textsubscript{6}a\textsuperscript{-}) methylation, which is significantly different from that in vitro (Supplementary Fig. S1). The algorithm uses predictions of protein interactions to identify single- and double-stranded regions (Spitale et al., 2015):

- For the training and testing we selected RNA fragments carrying the central nucleotide with the highest (single-stranded; 10\textsuperscript{5} non-redundant sequences) and lowest icSHAPE reactivities (double-stranded; 10\textsuperscript{3} non-redundant sequences), following the analysis carried out for CROSS in vitro (Delli Ponti et al., 2017). Each RNA fragment contains a total of 51 nucleotides to allow calculations with catRAPID (Bellucci et al., 2011). The nucleotides are represented as A = (1, 0, 0, 0), C = (0, 1, 0, 0), G = (0, 0, 1, 0) and U = (0, 0, 0, 1).
- The catRAPID approach uses a phenomenological potential that exploits several physico-chemical predictors including RNAfold for the RNA structure (Bellucci et al., 2011). 7797 regions from a library of 640 canonical RNA-binding proteins (Agostini et al., 2013) were analyzed to identify those able to discriminate nucleotides in single- and double-stranded states with accuracies >0.6 (m\textsubscript{6}a\textsuperscript{+}: 228 regions; m\textsubscript{6}a\textsuperscript{-}: 206 regions; Supplementary Figs S2 and S3).
HC scores). The m6a-double-stranded regions with a 0.88 accuracy (10-fold CV on 1%
the quality of the training set by altering the stability and turnover
10-fold CV on 1% HC scores) mainly because m6a removal affects
an independent
in vivo
We used the
information only (Delli Ponti
2016) on the long non-coding
3 Performances
CROSSalive scores were ranked by their absolute value and equal
groups of positives and negatives were selected to assess the per-
formances of the algorithm. From low (50%) to high-confidence (HC) scores (1%, Fig. 1A) the accuracy of the models increases
monotonically reaching a maximum of 0.86 for the m6a+ model when protein interactions are used (10-fold cross-validation, CV). In
comparison, the in vitro icSHAPE model based on RNA sequence information only (Delli Ponti et al., 2017) discriminates single- and
double-stranded regions with a 0.88 accuracy (10-fold CV on 1% HC scores). The m6a-in vivo model shows lower accuracy (0.74 in
10-fold CV on 1% HC scores) mainly because m6a removal affects the quality of the training set by altering the stability and turnover
of the transcriptome (Liu and Jia, 2014). We applied CROSSalive to an independent in vivo SHAPE-Map experiment (Smola et al.,
2016) on the long non-coding Xist (17 900 nt; not in the training).
We used the in vivo m6a+ model because Mettl3 is poorly abundant in
the trophoblasts (Thul et al., 2017) employed in SHAPE-Map and
only few nucleotides are methylated at the 5’ and 3’ of Xist (Patil et al.,
2016). The algorithm achieves an Area under the ROC curve (AUC) of 0.83 on the 15% HC single- and double-stranded regions
ranked by SHAPE reactivity (Fig. 1B). Moreover, CROSSalive profile shows a correlation of 0.45 with the SHAPE-Map one (Fig. 1B).
The m6a+ model trained on RNA sequence information only achieves an AUC of 0.53 (~0 correlation).

4 Conclusions
By using sequence-based information, CROSSalive profiles the RNA secondary structure in vitro. The use of different models (in vivo/ in vitro, m6a+/-m6a-) will help to identify structural regions to
investigate experimentally. As previously done with CROSS (Delli Ponti et al., 2017), CROSSalive can be integrated as a constrain in
thermodynamics-based approaches such as RNAfold, which will allow study structural differences of RNAs in vivo and in vitro
(Lorenz et al., 2016).

Acknowledgements
The authors thank Andrea Cerese and Alessio Colantoni.

Funding
The research leading to these results has received funding from European Research Council RIBOMYLOMEX_309545, European Union’s Horizon 2020 IASIS_727638 and INFORS_825070, as well as Spanish Ministry of Economy and Competitiveness BFU2017-86970-P.

Conflict of Interest: none declared.

References
Agostini,F. et al. (2013) catRAPID omics: a web server for large-scale prediction of protein–RNA interactions. Bioinformatics, 29, 2928–2930.
Bellucci,M. et al. (2011) Predicting protein associations with long noncoding RNAs. Nat. Methods, 8, 444–445.
Delli Ponti,R. et al. (2017) A high-throughput approach to profile RNA structure. Nucleic Acids Res., 45, e35.
Klus,P. et al. (2015) Protein aggregation, structural disorder and RNA-binding ability: a new approach for physico-chemical and gene ontology classification of multiple datasets. BMC Genomics, 16, 1071.
Liu,J. and Jia,G. (2014) Methylation modifications in eukaryotic messenger RNA. J. Genet. Genomics, 41, 21–33.
Livi,C.M. et al. (2015) catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. Bioinformatics, 773–775.
Lorenz,R. et al. (2016) SHAPE directed RNA folding. Bioinformatics, 32, 145–147.
Patil,D.P. et al. (2016) m6A RNA methylation promotes XIST-mediated transcriptional repression. Nature, 537, 369–373.
Rouskin,S. et al. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature, 505, 701–705.
Smola,M.J. et al. (2016) SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. Proc. Natl. Acad. Sci. USA, 113, 10322–10327.
Spitale,R.C. et al. (2015) Structural imprints in vivo decode RNA regulatory mechanisms. Nature, 519, 486–490.
Thul,P.J. et al. (2017) A subcellular map of the human proteome. Science, 356, 6340–6352.