Briefings in Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Review

Secondary structure prediction of long non-coding RNA: review and experimental comparison of existing approaches

L.A. Bugnon[1]*, A. Edera[1], S. Prochetto[1,2], M. Gerard[1], J. Raad[1], E. Fenoy[1], M. Rubiolo[1], U. Chorostecki[3], T. Gabaldón[3,4,5], F. Ariel[2], L. Di Persia[1], D.H. Milone[1] and G. Stegmayer[1]

[1]Research Institute for Signals, Systems and Computational Intelligence sinc(i) (CONICET-UNL), Ciudad Universitaria, Santa Fe, Argentina.
[2] IAL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.
[3]Barcelona Supercomputing Center (BSC-CNS), Institute of Research in Biomedicine (IRB), Spain.
[4] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.
[5] Centro de InvestigaciÃ³n BiomÃ©dica En Red de Enfermedades Infecciosas (CIBERINFEC), Barcelona, Spain.

Abstract

Motivation: In contrast to messenger RNAs, the function of the wide range of existing long non-coding RNAs (lncRNAs) largely depends on their structure, which determines interactions with partner molecules. Thus, the determination or prediction of the secondary structure of lncRNAs is critical to uncover their function. Classical approaches for predicting RNA secondary structure have been based on dynamic programming and thermodynamic calculations. In the last 4 years, a growing number of machine learning (ML)-based models, including deep learning (DL), have achieved breakthrough performance in structure prediction of biomolecules such as proteins and have outperformed classical methods in short transcripts folding. Nevertheless, the accurate prediction for lncRNA still remains far from being effectively solved. Notably, the myriad of new proposals has not been systematically and experimentally.
Results: In this work we compare the performance of the classical methods as well as the most recently proposed approaches for secondary structure prediction of RNA sequences using a unified and consistent experimental setup. We use the publicly available structural profiles for 3,023 yeast RNA sequences, and a novel benchmark of well-characterized lncRNA structures from different species. Moreover, we propose a novel metric to assess the predictive performance of methods, exclusively based on the chemical probing data commonly used for profiling RNA structures, avoiding any potential bias incorporated by computational predictions when using dot-bracket references. Our results provide a comprehensive comparative assessment of existing methodologies, and a novel and public benchmark resource to aid in the development and comparison of future approaches.
Availability: Full source code and benchmark datasets are available at: https://github.com/sinc-lab/lncRNA-folding
Contact: lbugnon@sinc.unl.edu.ar
Supplementary information: Supplementary data are available at *Briefings in Bioinformatics* online.

1 Introduction

For decades, the sole function assigned to RNAs was to act as an information messenger between DNA and proteins. Originally, in the central dogma of biology, RNA was considered to play a secondary part in expressing inherited information as proteins. However, our recent ability to sequence entire genomes and transcriptomes served to uncover that the majority of the human genome is transcribed, although nearly 98% of it comprises non-coding regions [33, 63]. Growing evidence has linked non-coding RNAs (ncRNAs) to virtually every step of gene expression regulation, including epigenetics and spatial organization of genetic information in the cell nucleus, transcripts processing and stability, messengers translation, post-transcriptional protein modification and degradation [42]. Non-coding RNAs can be classified according to their size into two large classes [30]. On the one hand active small

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

This is a pre-copyedited, author-produced version of an article accepted for publication in Briefings in Bioinformatics following peer review. The version of record L A Bugnon, A A Edera, S Prochetto, M Gerard, J Raad, E Fenoy, M Rubiolo, U Chorostecki, T Gabaldón, F Ariel, L E Di Persia, D H Milone, G Stegmayer, Secondary structure prediction of long noncoding RNA: review and experimental comparison of existing approaches, Briefings in Bioinformatics, Volume 23, Issue 4, July 2022, bbac205, https://doi.org/10.1093/bib/bbac205

RNAs, less than 50 nt in length, include microRNAs (miRNAs), small interfering RNAs (siRNAs), heterochromatic siRNAs (hetsiRNAs), Piwi-associated RNAs (piRNAs), small nuclear and nucleolar RNAs (snRNAs and snoRNAs, respectively), among others. On the other hand, long non-coding RNA (lncRNAs), larger than 200 nt, may exert their functions as long transcripts without being processed into small RNAs [67]. In particular, several studies shed light on the role of lncRNAs in diverse cellular processes, such as cell-cycle regulation in health and disease [19, 34], including cancer and diabetes [2, 18, 46, 72, 79], among others. In plants, lncRNAs have been associated with development and the dynamic response to the environment [5, 38]. At the molecular level, lncRNAs have been linked to virtually every step of gene expression regulation, including chromosome inactivation, genomic imprinting, chromatin dynamics, protein modifications and stability [45, 47]. Similar to proteins, RNA function is mainly related to its structure [22]. In particular, it should be noted that ncRNAs perform their functions through interaction with other molecules (DNA, RNA, proteins and lipids). In this sense, the lncRNA secondary structure is decisive to determine its interactome and the related functional output [60].

The RNA molecule is an ordered sequence composed of four nucleotides (nts) or bases: adenine (A), cytosine (C), guanine (G) and uracil (U), arranged in the 5âŁ™ to 3âŁ™ direction. The pairing of these four bases within a RNA molecule gives rise to its secondary structure. Canonical base pairs include the WatsonâŁ"Crick base pairs (AâŁ"U and GâŁ"C) and wobble base pairs (GâŁ"U), which provide higher energetic stability to the molecule. These base pairs often result in the formation of a nested structure, where several pairs are stacked and one or more unpaired bases form a loop. RNA secondary structure is a 2-D representation of this self-folding. Figure 1 (top) illustrates an example of the basic motifs: (i) the double stranded regions named *stems*, obtained by the stacking of two or more consecutive base-pairs; (ii) the hairpin *loop*, a single-stranded region at the end of a stem; (iii) a *bulge*, a single stranded region which interrupts a stem on one side; (iv) an *internal loop* stops a stem on either side; and (v) a single stranded region where several stems meet called a *multi-branched loop* [1]. The RNA secondary structure is usually represented in dot-bracket notation, with matching parentheses for paired bases and dots for unpaired bases (Figure 1, bottom). Additional structures called pseudo-knots can be formed when unpaired bases match with distant ones, commonly annotated with other symbols such as angle- and curly-brackets.

Even though there is currently a wide variety of publicly available ncRNA sequences, and their numbers keep growing at an ever-increasing rate [61], most of their structures remain unknown. Therefore, the efficient determination of their secondary structure is of high interest, which can be carried out by physico-chemical methods. For example, from atomic coordinates obtained from X-ray crystallography or nuclear magnetic resonance (NMR) [21, 31, 59]. However, such methods have low throughput and it is very challenging to apply them on lncRNAs not only because of the high experimental costs and resolution limits [48], but also due to their length, low abundance in in-vivo systems and the large diversity of stable structures that they can adopt. To date, the majority of structural elements found in lncRNA sequences has been determined as patterns of base pairings using a combination of chemical or enzymatic probing [74] such as PARS [32], nextPARS [52], SHAPE-seq [37] or DMS-seq [15]. Yet, the structures of only a tiny fraction of RNAs have been experimentally determined, limiting the understanding of this key feature upon functional outputs.

In the last decades, many methods have been developed for the computational prediction of RNA secondary structure. The first proposals, dating 15 years ago [41], were based on dynamic programming and thermodynamics calculations [53, 66], identifying a structure with minimum free energy (MFE) according to the principle that RNA molecules exist in energetically stable states, like proteins [4]. Until

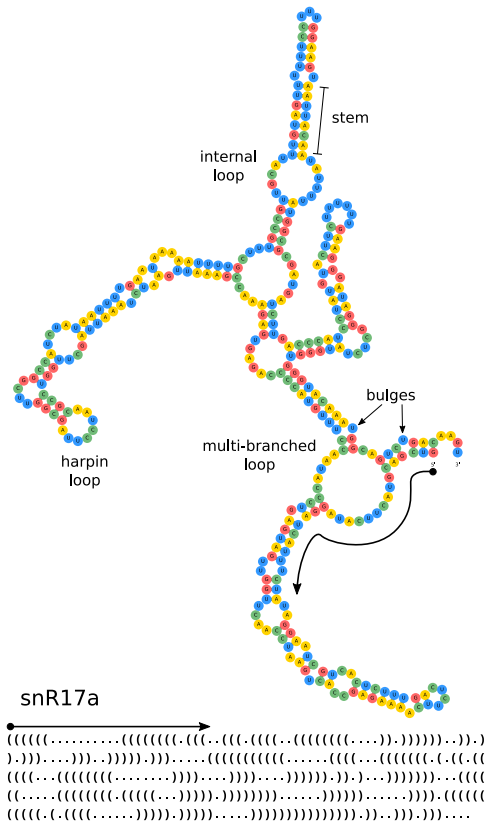

Fig. 1. Example of typical motifs of RNA secondary structure in the transcript snR17a of S. cerevisiae: (top) hairpin loop, internal loop, stem, bulge and multi-branched-loop; (down) corresponding dot-bracket notation.

the irruption of machine learning (ML)-based methods in the field approximately 4 years ago, prediction accuracy has remained almost unaltered. Recent works have shown that newer methods based on Deep Learning (DL) can outperform existing mainstream methods on small datasets, in terms of accuracy and applicability [77].

DL techniques first emerged as an alternative approach to structure prediction problems in proteins with AlphaFold [28] and in RNA secondary structure prediction as well [55, 75]. Compared to classical approaches, DL methods make much weaker assumptions about the thermodynamic mechanic driving RNA folding, which is based on labor-intensive experimental melting data [66]. Thus, they are more suited to detect more complex foldings [20], such as non-canonical base pairing or previously unrecognized base pairing constraints. There are many different DL proposals, which differ in their architectural design, model input-output, training data and optimization algorithms used to adjust their parameters. In general, methods treat the input RNA as a sequence of characters defined by the bases, which can be processed by different well-known computational models for text processing, such as Long Short-Term Memory (LSTM) [23] or Transformer encoders [70] that are well suited for capturing long-range interactions between nucleotides. Other approaches integrate DL techniques with thermodynamic methods to further improve predictions [51]. From a practical point of view, it is difficult to choose among the myriad of new methods available, in particular because existing DL approaches still face several challenges. For example, a huge number of model parameters need to be adjusted, resulting in high computational cost. Another major point is training data because representative and high quality training data is critical to achieve reliable models and high performance in terms of generalization capability, i.e. the ability to make

good structural predictions for RNA sequences that differ from those used for model training.

Several reviews have been published on the topic of RNA secondary structure prediction [17, 54, 77]. Yet, they still do not offer systematic evaluations of ML and DL techniques for comparatively assessing their performance on representative and large benchmark datasets. Notably, previous attempts at creating RNA secondary structure databases for methods comparison [3, 12, 43, 69] are no longer available, have not been updated or lncRNA sequences are not included. Additionally, the secondary structures provided are mostly limited to computational predictions. In this review, instead, we have included only RNA secondary structures experimentally determined by biochemical approaches. We provide a comprehensive overview of the most recent ML- and DL-based methods for lncRNA secondary structure prediction, in comparison to classical methods. Furthermore, particularly for lncRNAs, we have compiled a novel benchmark dataset of manually curated and experimentally validated secondary structures of lncRNAs. In this work we also introduce a novel metric for assessing the performance of predictions using the probing scores obtained from enzymatic approaches as the reference of the true RNA structure. This metric named Mean Average Similarity (MAS) makes comparison between methods less biased than classical metrics based on reference dot-bracket structures obtained by specific folding methods. Our work not only provides a comprehensive comparative assessment of existing methodologies for lncRNA secondary structure prediction in the last 15 years, but also contributes to the bioinformatics community for the development and comparison of future approaches.

2 RNA secondary structure prediction methods

RNA structure prediction has been classically formulated as an optimization problem, where a score is defined for every possible folding of the given RNA sequence, and the predicted folding is the one that maximizes it. The most popular approach was based on thermodynamic models [53], such as Turner's nearest neighbor model [66]. The free energy of each nearest neighbor is calculated by summing up its free energy parameters. The free energy of an entire RNA secondary structure will be the sum of the free energy of each nearest-neighbor loop. An optimal secondary structure with MFE is calculated using dynamic programming, such as the Zuker algorithm [80].

The scores of each local element were obtained from wet-lab experiments reflecting the thermodynamics free energy theory [53]. However, the increasing availability of known RNA structures for training made it possible to successfully drift towards hybrid approaches, which are based on dynamic programming but make a fine-tuned parameter estimation based on ML [24]. Nowadays, methods for RNA structure prediction are shifting towards full ML and DL-based approaches, due to the fact that there is now larger data available for training (mostly of short sequences). A comprehensive summary of the methods that have appeared in the last 15 years in literature is presented in Table 1 and described in detail in Supplementary Material 1. The methods included in the experimental results of this study are only those available, at the time of writing this manuscript, as open access tools (already trained prediction models) in a repository or as public web servers.

3 Data and performance measures

3.1 Data

Saccharomyces cerevisiae **(sce)**. A large benchmark dataset was obtained from [32] where PARS was performed to characterize the secondary structure of the messenger RNAs (mRNAs) and ncRNAs of the budding yeast Saccharomyces cerevisiae. A total of 3,199 transcripts ranging from 71 to 8,145 bases in length were profiled. From this data, we defined 3 distinctive sub-datasets: sce3k, sce188 and sce18. The sce3k subset has 3,023 unique sequences with length larger than 200 nt. The subset sce188 was obtained from the previous one by identifying non-coding transcripts. To this end, we used the coding potential calculator 2 (CPC2) [29], which is species-neutral and has high accuracy for long non-coding transcripts. The sce18 subset was defined by taking only those sequences from sce188 that were not previously classified as mRNA in [32]. Authors used the Vienna package [36] to fold transcripts, calculate the partition function of the structure ensemble and base pairing probabilities, without probing scores as constraints[1].

Curated set of lncRNAs. As indicated in [40], an informative benchmark requires a variety of high-quality structures. We have generated a novel dataset including curated structures of lncRNAs that were determined by experimental analyses, for which the structures have been carefully studied and validated through a variety of biochemical approaches. We have compiled those available from literature and those mentioned in a very recent review on lncRNAs [48]. The curated lncRNAs included in our study are shown in Table 2, where the name, species, length, year of discovery and probing methodology are indicated in the columns, for each test lncRNA in the rows. A brief description of these lncRNAs can be found in Supplementary Material 2. For each lncRNA, its reference structure is represented in two different ways. In the first place, the classical dot-bracket representation, in which the nested base-pairs 2D conformation was obtained with a classical method, using the probing scores as constraints. The other reference representation consists of the probing score per base, which indicates the pairing probability of each nucleotide in the sequence. Since each experimental technique provides a different distribution of score values, a normalization[2] step was performed to place all the scores within the [0, 1] interval, where 0 is unpaired, 1 is paired, and values around 0.5 are uncertain.

3.2 Performance measures

Classical measures. In this case, the focus of performance measures is on the accuracy of predicted base pairs in comparison to a reference structure [40]. Pairs that are both in the prediction and in the reference structure are true positives (TP), while pairs predicted but not in the true structure are false positives (FP). Similarly, a pair in the reference structure that is not predicted is a false negative (FN), and a pair that is neither predicted nor in the true structure is a true negative (TN). To fully characterize the successes and failures of structure prediction, the F_1 score is defined as

$$F_1 = \frac{2\,TP}{2\,TP + FP + FN}.$$

These classical measures were calculated with the scorer program in RNAstructure package[3].

Mean absolute similarity (MAS) score. We propose a new score that can reflect the similarity between the nextPARS/PARS/SHAPE probing data and the predicted dot-bracket structures. Given an RNA sequence, this score takes into account the binary paired-unpaired state of each nucleotide n in the predicted structure $b(n) \in 0, 1$, and the corresponding normalized probing score $p(n) \in [0, 1]$. The paired, unpaired and average similarities

[1] https://genie.weizmann.ac.il/pubs/PARS10/pars10_catalogs.html

[2] This is detailed in the source code, results section

[3] http://rna.urmc.rochester.edu/Releases/

Table 1. Methods for RNA secondary structure prediction included in this review.

Method	Year	Type	Repository	Web server
CONTRAFold [16]	2006	Statistical learning	http://contra.stanford.edu/contrafold/download.html	http://contra.stanford.edu/contrafold/server.html
CentroidFold [49]	2009	Statistical decision theory	https://github.com/satoken/centroid-rna-package	http://rtools.cbrc.jp/centroidfold/
ShapeKnots [13]	2010	Dynamic programming	http://rna.urmc.rochester.edu/RNAstructure.html	
ProbKnot [6]	2010	Assembling structures from base-pair probabilities	http://rna.urmc.rochester.edu/RNAstructure.html	https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/ProbKnot/ProbKnot.html
RNAstructure [44]	2010	Thermodynamics	http://rna.urmc.rochester.edu/RNAstructure.html	http://rna.urmc.rochester.edu/RNAstructureWeb/
RNAfold [36]	2011	Dynamic programming	https://github.com/ViennaRNA/ViennaRNA	http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi
IPknot [50]	2011	Integer programming	https://github.com/satoken/ipknot	http://rtips.dna.bio.keio.ac.jp/ipknot/
contextFold [73]	2011	Structured-prediction learning	https://www.cs.bgu.ac.il/∼negevcb/contextfold/ContextFold_1_00.zip	https://www.cs.bgu.ac.il/∼negevcb/contextfold/
RNAshapes [26]	2014	Abstract shape analysis	https://bibiserv.cebitec.uni-bielefeld.de/rnashapes	https://bibiserv.cebitec.uni-bielefeld.de/rnashapes?id=rnashapes_view_webservice
pKiss [26]	2014	Abstract shape analysis	https://bibiserv.cebitec.uni-bielefeld.de/pkiss	https://bibiserv.cebitec.uni-bielefeld.de/pkiss?id=pkiss_view_webservive
SPOT-RNA [55]	2019	ResNet + biLSTM	https://github.com/jaswindersingh2/SPOT-RNA	https://sparks-lab.org/server/spot-rna
LinearFold [24]	2019	Dynamic programming + statistical learning	https://github.com/LinearFold/LinearFold	http://linearfold.org
LinearPartition [76]	2020	Dynamic programming + base pairing probabilities	http://github.com/LinearFold/LinearPartition	http://linearfold.org/partition
E2Efold [8]	2020	Transformer	https://github.com/ml4bio/e2efold	
rna-state-inf [71]	2020	Bi-LSTM	https://github.com/dwillmott/rna-state-inf	
SPOT-RNA2 [56]	2021	Ensemble of deep learning models	https://github.com/jaswindersingh2/SPOT-RNA2	https://sparks-lab.org/server/spot-rna2
UFold [20]	2021	U-net	https://github.com/uci-cbcl/UFold	https://ufold.ics.uci.edu
MXfold2 [51]	2021	Deep learning + thermodynamic parameters	https://github.com/keio-bioinformatics/mxfold2	http://www.dna.bio.keio.ac.jp/mxfold2/
2dRNA-Fold [39]	2021	Deep learning + reinforcement learning	https://github.com/Urinx/2dRNA-Fold	http://biophy.hust.edu.cn/new/2dRNA

Table 2. Dataset of curated and validated lncRNAs.

Name	Species	Length	Year	Measurement technology
NORAD#1 [11]	*H. sapiens*	1,903	2021	nextPARS
NORAD#2 [11]	*H. sapiens*	1,862	2021	nextPARS
NORAD#3 [11]	*H. sapiens*	1,614	2021	nextPARS
CYRANO[27]	*H. sapiens*	4,419	2020	SHAPE
MEG3 [68]	*H. sapiens*	1,595	2019	SHAPE
RepA [35]	*M. musculus*	1,630	2017	SHAPE + chemical probing
PAN [62]	*Human gammaherpes virus 8*	1,077	2017	SHAPE-MaP
XIST [57]	*M. musculus*	17,779	2016	SHAPE-MaP
lincRNAp21sense [9]	*H. sapiens*	311	2016	SHAPE
lincRNAp21antisense [9]	*H. sapiens*	303	2016	SHAPE
HOTAIR [58]	*H. sapiens*	2,154	2015	SHAPE + chemical probing
MALAT1 [7]	*H. sapiens*	8,415	2014	SHAPE
ROX2 [25]	*D. melanogaster*	573	2013	SHAPE+PARS

are defined as

$$\begin{aligned} s^+(b,p) &= 1 - \frac{1}{|M^+|}\sum_{n\in M^+}|b(n)-p(n)|, \\ s^-(b,p) &= 1 - \frac{1}{|M^-|}\sum_{n\in M^-}|b(n)-p(n)|, \\ s(b,p) &= 1 - \frac{1}{|M|}\sum_{n\in M}|b(n)-p(n)|, \end{aligned}$$

where the sets of paired and unpaired nucleotides are

$$\begin{aligned} M^+ &= \{n\colon p(n) \geq 1/2+\epsilon/2\}, \\ M^- &= \{n\colon p(n) < 1/2-\epsilon/2\}, \end{aligned}$$

and $M = M^+ \cup M^-$. In this definition, ϵ is the uncertainty level around the undetermined/no-measure score (1/2). That is, all the nucleotides with scores in the range $[1/2-\epsilon/2, 1/2+\epsilon/2]$ are ignored in the similarity measure.

As the normalized probing scores are distributed in the [0,1] range, the maximum/minimum value of the average similarities will generally not be 1/0. This is because predicted structures are binary, preventing perfect matchings with the real-valued probing scores. Therefore, a re-scaling step must be done to take into account the best and the worst binary structures,

$$\begin{aligned} b_b(n) &= \begin{cases} 1 & \text{if } p(n) \geq 1/2+\epsilon/2 \\ 0 & \text{otherwise} \end{cases} \\ b_w(n) &= 1 - b_b(n). \end{aligned}$$

Using the average similarities for these extreme cases, the normalized MAS score is defined as

$$MAS(b,p) = \frac{s(b,p) - s(b_w,p)}{s(b_b,p) - s(b_w,p)}$$

This score was tested in several cases, which are provided as an interactive notebook in the source code repository, section results.

4 Results

4.1 Large experimental comparison on RNA secondary structure prediction

Figure 2 reports the results for the prediction methods evaluated on the largest dataset (sce3k). The figure shows violin plots with the median on the y-axis (blue diamond) of F_1 (top) and MAS score (bottom) for all the comparative methods in the x-axis. Methods with F_1 lower than 0.10 were not included in the analysis (E2Efold mostly predicted pseudoknots and 2dRNA-fold took a very large time to run for each sequence and could predict only a very small number of sequences). The figure also shows the corresponding coverage (gray bar in the background) for each method, i.e. the percentage of sequences from the full dataset that was effectively predicted by each method. It can be seen that most methods provided predictions for all 3,023 sequences. The best methods achieved a median F_1 score of around 0.50: ProbKnot, RNAshapes, RNAfold, ShapeKnots and RNAstructure. This result is expected considering that the reference dot-bracket for this dataset was obtained with RNAfold and the other methods with good performance share the same approach (thermodynamic modeling and dynamic programming), or are even directly based on RNAfold. The least performing method according to this metric achieved a F_1 score of around 0.20. With respect to the MAS score, when looking at the similarity between experimental probing data and the predictions, most methods achieved more stable predictive performances (low variance) with median values around 0.60. According to this measure, which is not biased towards any particular computational method used to derive the reference dot-bracket, all methods have a median performance higher than 0.50. Also, it can be seen that the MAS scores distribution for each method is less dispersed than the corresponding F_1 scores distribution, showing a more consistent measurement across the different sequences in the dataset.

In order to provide a statistical analysis of results, we used Friedman test and critical difference (CD) with post-hoc Nemenyi test ($\alpha = 0.05$) [14]. The Friedman test showed that the differences in the F_1 and MAS scores distributions are statistically significant ($p < 1\text{E-}15$). The CD diagram is shown in Figure 3. Here, the corresponding CD diagram for F_1

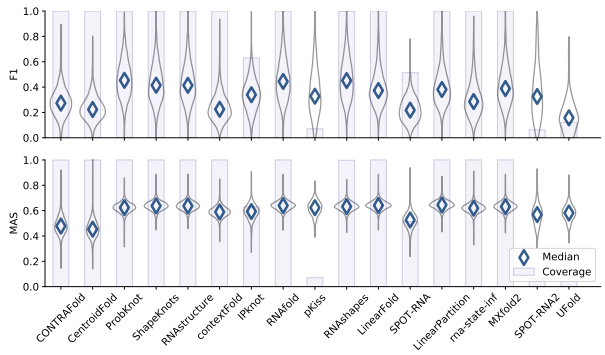

Fig. 2. Comparison of RNA prediction methods on the sce3k dataset. F_1 (top), MAS score (bottom).

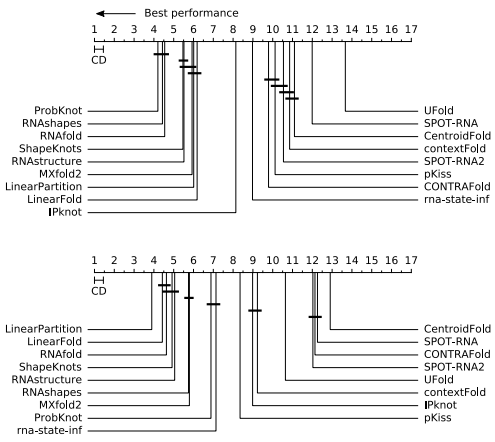

Fig. 3. Critical difference diagram for the RNA prediction methods on the sce3k dataset for F_1 (top) and MAS score (bottom).

indicates that ProbKnot, RNAshapes and RNAfold are the best methods for this dataset. Then there is a second group of statistically similar methods, one of which is based on ML (MXfold2). Very interestingly, the rest of the methods are well separated (different) among them. Regarding MAS, the best method is LinearPartition, well-separated and followed by classical methods and MXfold2, which are not statistically different among them.

4.2 Experimental comparison focused on lncRNAs

Figure 4 shows the comparison of prediction methods on the sce188 dataset, composed of 188 transcripts with lengths between 200 and 1,301 bases. The figure reports violin plots of the F_1 (top) and MAS scores (bottom) for all the comparative methods (x-axis). Methods with F_1 lower than 0.10 were not included. The figure also shows the corresponding coverage (gray bar) for each method. In this case, almost all methods could predict 100% of the sequences, except for UFold whose predictions are restricted to a maximum of 600 nt per sequence. The best methods according to F_1 are, again, those sharing the same classical approach as the method used in the reference structures. The rest of the methods have F_1 between 0.17 and 0.40, not that far from the best ones ($F_1 \approx 0.50$).

The MAS score shows that most methods are very close in performance, between 0.50 and 0.60. Unlike the results obtained from the sce3k dataset, the slightly wider dispersion observed here in the plots indicates that there is more heterogeneity in the predictions of sce188, which only includes transcripts predicted as lncRNAs. The Friedman test indicates that differences in the F_1 and MAS score are statistically significant ($p < 1\text{E-}15$). However, given the smaller number of sequences,

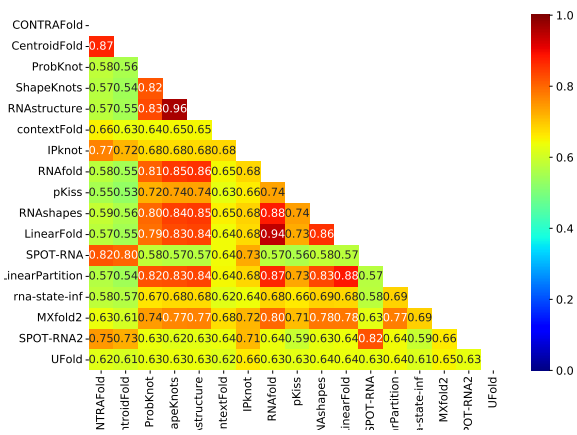

Fig. 5. Paired/unpaired prediction rate between methods on the dataset sce188. Each value of the heatmap shows how similar the predictions of two methods are on average.

the statistical power of the CD method is lower than that of the sce3k dataset (see Supplementary Material 3, Figure S1).

Given the close results achieved by the methods in this dataset, the similarity of each prediction was analyzed in more detail. To this end, for each sequence the structures predicted by a pair of methods were compared in terms of which nucleotides were predicted as paired or unpaired. The average rate over the predictions obtained for each pair of methods is shown with a pair-wise heatmap in Figure 5. These comparisons show that the most similar methods are RNAstructure and ShapeKnots, which is very reasonable since the last one is based on the same thermodynamic model. This high similarity is in agreement with the previous results of F_1 and MAS score, as well as with the critical differences where both methods are similarly ranked. The second most similar methods are LinearFold and RNAfold, which is expected since the first method was proposed precisely to improve the computational performance (from cubic to linear time) of RNAfold without modifying the results. Interestingly, it can be seen that the structures predicted by CONTRAFold, CentroidFold, IPknot, SPOT-RNA and SPOT-RNA2 very frequently display high similarity between them, and are different (<0.60) in comparison to the other methods.

Figure 6 shows the comparative results for the sce18 dataset. Surprisingly here, all methods achieved a noticeably better performance high. In sce18, most methods have very high median F_1 (around 0.80), with RNAfold and LinearFold showing a median $F_1 = 0.86$, very closely followed by LinearPartition and ProbKnot ($F_1 = 0.83$), RNAstructure

Fig. 4. Comparison of performance for the RNA secondary structure prediction methods on the sce188 dataset. F_1 (top), MAS score (bottom)

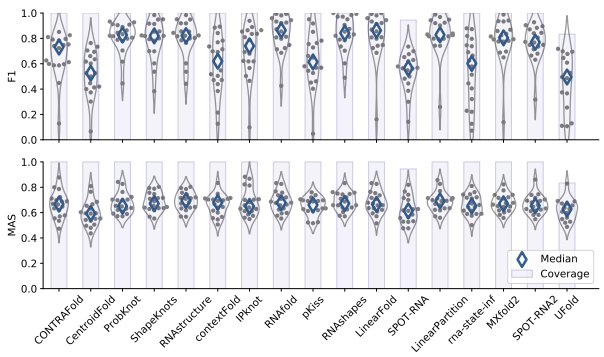

Fig. 6. Comparison of RNA prediction methods on the sce18 dataset. F_1 (top), MAS score (bottom).

and ShapeKnot ($F_1 = 0.82$), MXfold2 ($F_1 = 0.80$) and SPOT-RNA2 ($F_1 = 0.77$). Importantly, these last two methods are ML-based, suggesting that they are able to achieve good predictive performance for lncRNAs. However, there is a large dispersion in the F_1 plots, hinting at a high heterogeneity in the predictions for the sce18 sequences. The other methods achieved lower scores, which are between $F_1 = 0.50$ (UFold) and $F_1 = 0.74$ (SPOT-RNA). Regarding the MAS score, all methods achieved median scores between 0.60 and 0.70, with more compact predictions and less dispersion than F_1. This indicates that all methods, no matter the approach, can reach dot-bracket structures with high correlation to the probing data. For this dataset, these results indicate that ML-based methods achieved performances very close to the classical approaches, and which are also considerably high for lncRNAs. This is an impressive result for ML-based models in comparison to very well-known and established methods for RNA structure prediction, making them very competitive for this task (Supplementary Material 3, Figure S2).

In order to determine differences in the computational cost of the methods, the running time of each method has been calculated based on the sequences in the sce18 dataset (Supplementary Material 3, Figure S3). For this calculation, all methods were run in the same hardware conditions. These results show how the sequence length clearly impacts on the computing time of each method. Almost all methods increase time for longer transcripts with the same law, showing different exponential behavior, with more cost for pKiss and ShapeKnots.

Finally, in order to understand why all methods had such high performance on the sce18 dataset, we have measured the maximum length of the hairpins present in the reference structures. The average number of hairpins per sequence was 22.32 for sce3k, 7.48 for sce188 and 5.72 for sce18, while the average of the maximum hairpin length per sequence (normalized by the sequence length) was 0.08 for sce3k, 0.16 for sce188 and 0.23 for sce18. Interestingly, sce18 sequences have the fewest but longest hairpins compared to those found in the sce188 and sce3k. We hypothesize that the sce18 secondary structures are more stable and well-formed, and thus are easier to predict by any method. A possible explanation may be that lncRNAs are more stable since their function is more dependent on the structure, compared to mRNAs. This is in line with previous findings showing that mRNAs are poorly structured [32].

4.3 Prediction analyses on a curated list of well-characterized lncRNAs

To further evaluate the methods, we have tested the RNA structure prediction with a set of lncRNAs whose structures have been experimentally validated (details in Table 2). Figure 7-(left) reports violin plots with MAS score, for each method in the x-axis. These results show that most methods provide predictions with high similarity to probing scores, with median values larger than 0.60, and for the least performing methods the median similarities are around 0.50 for CONTRAFold, CentroidFold, SPOT-RNA, SPOT-RNA2 and UFold.

In order to analyze the differences in the prediction of the curated lncRNAs, Figure 7-(right) shows violin plots with the MAS score of the methods for each lncRNA in the x-axis. In this figure, the lncRNAs are ordered according to their lengths, from short to large. The median similarities are around 0.60 for all lncRNAs, no matter the length of the lncRNA to predict. Indeed, for the largest sequence (XIST) the median MAS score is 0.61 and all methods are very close to this median. This indicates that actually all methods and all approaches are capable of predicting the XIST, and also PAN, pairing patterns considerably well as measured by probing methods. It is also interesting to notice that MAS scores are very similar for the three NORAD fragments at the three different temperatures, which is expected given that the fragments resemble each other and have a similar structural function [11]. In comparison, the F_1 violin plots for this dataset (see Supplementary Material 3, Figure S4) show a large dispersion among methods predictions, for example the performance ranges from 0.10 to 0.80 for the same method depending on the lncRNA to predict. This variance makes it hard to distinguish among methods.

4.4 Detailed analysis of a curated lncRNA: NORAD

For an in-detail comparison across methods, we chose NORAD as a leading case since it is dysregulated in various types of cancer and the function is largely mediated by its structure. Figure 8 shows the results for the prediction of the secondary structure of NORAD#1 at 37 °C or all the methods reviewed here. NORAD#1 is a fragment of 1,903 bases in length within NORAD, whose reference structure has been profiled in high detail [11] using nextPARS pipeline [10]. The reference structure was built with RNAstructure and the experimentally determined nextPARS scores as constraints. For this sequence, 2dRNA-Fold could not be run because of not enough RAM, and E2EFold and UFold because of restrictions on sequence length. The figure shows in bars the resulting predictive performances based on F_1 and MAS score. In addition, it also shows the normalized paired MAS, indicated with '+', and the normalized unpaired MAS, indicated with '−'.

The analysis of the F_1 indicates that the best methods are LinearPartition and pKiss, but with low F_1, closely followed by ProbKnot and RNAstructure. It has to be noticed that the reference structure was built with RNAstructure + nextPARS scores as constraints. According to the MAS score, most methods were capable of correctly predicting the 70% of the pairing in the sequence. Here rna-state-inf is the best method, closely followed by ShapeKnots, RNAstructure, LinearPartition, RNAfold, pKiss and MXFold2, which are all equally good performing. The least performing methods reached a 0.45 of MAS score. The paired and unpaired MAS scores are good indicators of the capability of methods of correctly predicting both types of pairing in the sequence. For example, all winning methods here are good predictors of paired nucleotides, while methods with the lowest performance predict simpler structures and mostly identify unpaired nucleotides.

For a more comprehensive comparison among methods, we performed a deeper analysis of the RNA structures predicted by the top-3 best methods and the one with the lowest performance (Figure 9). For this analysis, the package *draw_rna*[4] was used, which plots the structures and colors them in a range from blue to yellow ($5' \rightarrow 3'$, respectively). For visual comparisons, the same color indicates the same part of the sequence. It is very interesting to see that, according to F_1 (top) the 2nd best method is pKiss, but it can be clearly seen that the resulting structure is quite

[4] https://github.com/DasLab/draw_rna

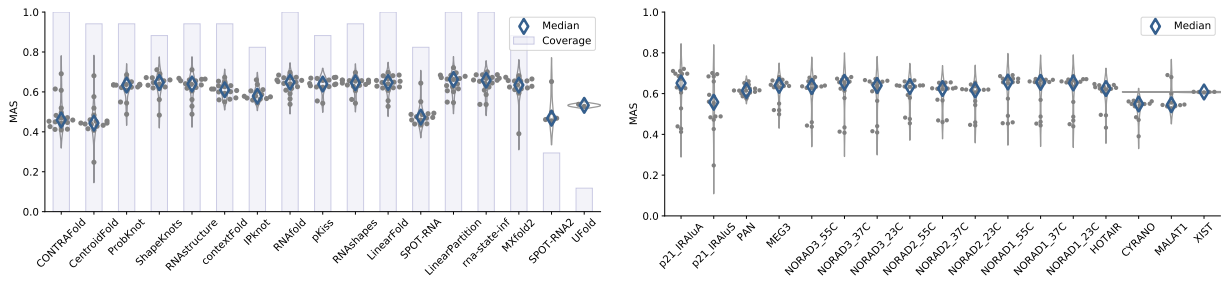

Fig. 7. Comparison of MAS score performance for the RNA secondary structure prediction methods on the curated lncRNAs dataset. Results grouped by prediction method (left); each dot represents a lncRNA. Results grouped by sequence (right); each dot represents a prediction method.

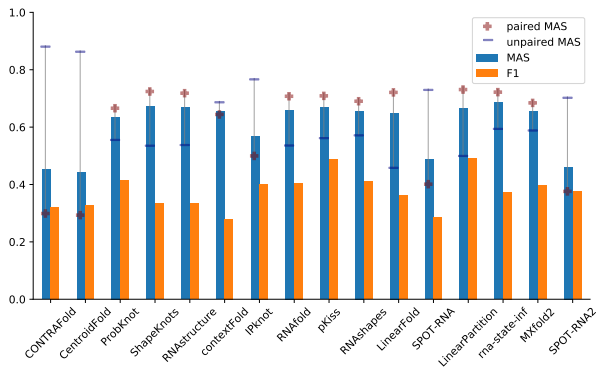

Fig. 8. Comparison of F_1 and MAS scores for the RNA secondary structure prediction methods on NORAD#1 at 37°C.

different from the reference one. Moreover, the method with the lowest F_1 (top, right) seems to be more similar to the reference one than the 2nd best method. Instead, the MAS score (bottom) provides a better ranking: all 3-top methods are visually very similar to the reference one; and the prediction of the method with the lowest performance (bottom, right) is very different to the reference.

For the top-best method in Figure 10 we used the software Circlecompare[5] that compares two RNA structures (reference and prediction) for the same sequence by showing the nucleotides in a circle. It uses the following color scheme: green for pairs present in both predicted and reference structure, red for pairs present in predicted structure only, and black for pairs present in reference structure only. It can be seen that there are more bases in green for LinearPartition, which is reflected by its high F_1 score. It also can be seen that rna-state-inf proposes different base pairs, but most of the paired nucleotides in the reference are paired in the prediction, which explains its high MAS score. Also, both methods predict a high number of pair bases that are not part of the reference structure. Moreover, most of the long-range pairings are different from those from the reference, while local structure is more preserved.

The sequence of NORAD contains 12 recognizable and sequence-similar NORAD repeat units (NRUs) that originated by tandem duplication at the rise of mammals and still share sequence homology [65]. The NRUs are ≈300 nt in length. Most NRUs contain one or two binding sites for the two homologs of Pumilio (Pum) in mammals. NORAD-regulated Pum targets are enriched in genes involved in cell division, mitosis and chromosome number instability [64]. The individual NRUs can be studied independently, facilitating isolation of specific interaction partners. Figure 11 shows the comparison of MAS score performance for the RNA secondary structure prediction methods of NRU1, NRU2, NRU3 and NRU4 within NORAD#1 at 37°C. It can be seen that NRU1, NRU2 and NRU4 are hard to predict for all methods, reaching at best a similarity around 0.50. Besides, NRU 3 is predicted by most methods with scores near 0.80. The best one here were LinearFold, RNAfold, RNAshapes and pKiss.

Several conserved elements, including a small and a larger hairpin are peculiarly found in some NRUs and not others[64]. Thus, a possible explanation between performance of the methods for the different NRUs could be that, although most NRUs fold independently, there are occasional long range interactions. In the case of NRU1, it was shown that there are inter-NRU interactions between NRU1-10 [78]. Furthermore, it seems that NRU2 and NRU4 fold mostly independently, but they also have small interaction with regions outside them [11]. Instead the NRU3 folds completely independently, thus it could be locally more structured than the rest of the NRUs and that is why it is better predicted by most methods.

5 Conclusions

In this study we provided a comprehensive review and experimental comparison of classical methods as well as the most recently proposed. We used data sets validated by chemical probing methods, including a novel benchmark with a compilation of well-characterized lncRNAs from different species. Thus, our study also provides a novel public benchmark to aid in the development and comparison of future approaches.

Most available secondary structures published were obtained with computational prediction methods based on dynamic programming and thermodynamics models, which sets upis an unfair basis of comparison for methods based in any different approach. Therefore, we proposed a novel score to assess the performance of methods exclusively based on the chemical-probing data used for profiling RNA structures. The MAS score has shown to be less biased, and thus a better measure for performance comparison of secondary structure prediction methods.

Key Points

- We provide a comprehensive comparative assessment of existing methodologies for lncRNA secondary structure prediction in the last 15 years.
- Each method has been explained and compared experimentally on the same data set of curated lncRNAs.
- A new unbiased measure of performance is provided based only on the chemical-probing data.
- This study will help the bioinformatics community in the development and comparison of future approaches for lncRNA secondary structure prediction.

[5] https://rna.urmc.rochester.edu/Text/CircleCompare.html

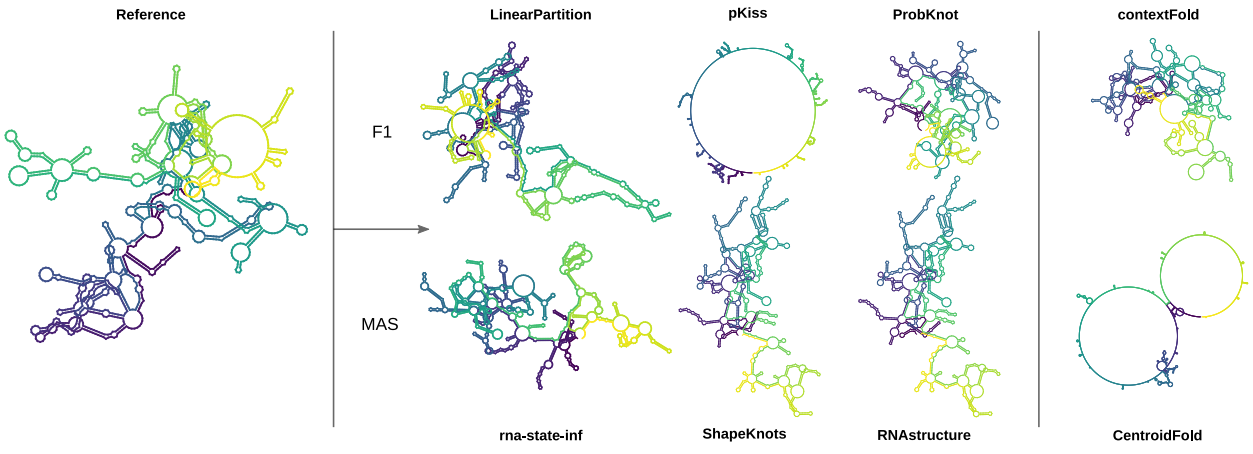

Fig. 9. In-depth analysis of structure predictions for NORAD#1 at 37°C. Plots visualize the reference (left), the predictions of the top-3 methods (middle) and last prediction in the ranking (right) according to F_1 (top) and MAS (bottom). Same color indicates the same part of the sequence from 5′ (blue) to 3′ (yellow).

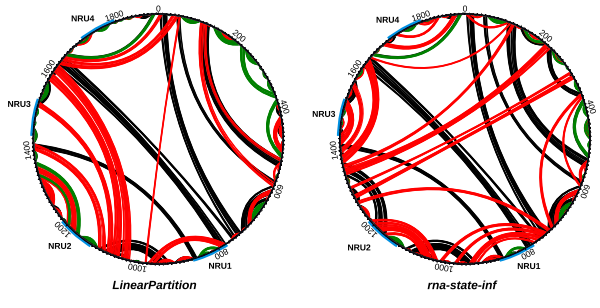

Fig. 10. Circular plot for the best method in predicting NORAD#1 at 37°C according to F_1 (left) and MAS score (right). Pairs present in both predicted and reference structure (green). Pairs present in predicted structure only (red). Pairs present in reference structure only (black).

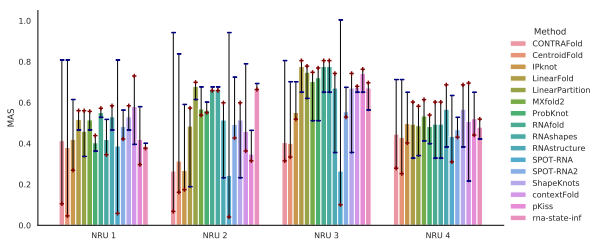

Fig. 11. Comparison of MAS score performance for the RNA secondary structure prediction methods of NRU1, NRU2, NRU3 and NRU4 within NORAD#1 at 37°C.

Acknowledgements

Authors would like to thank Dr. Jessica Brown, Dra. Sztuba-Solinska, Dra Anna Pyle and Rafael Araujo Tavares, Dra. Alisha Jones, Dr. Marco Marcia, Dr. Kevin Weeksand and Dr.Ilik for providing the reference data for the lncRNA experimentally validated.

Funding

This work was supported by ANPCyT (PICT 2018 3384, PICT 2018 2905, PICT 2019 3420) and UNL (CAI+D 2020 115). Researchers from sinc(i) and IAL are collaborating in the framework of the Program Science and Technology against Hunger, supported by the Argentinian Ministry of Science, to study and develop ncRNAs as exogenous bioactive molecules in agriculture. UC was funded by MICINN (IJC2019-039402-I). The work used computational resources from the Pirayu cluster, acquired with funds from the Santa Fe Science, Technology and Innovation Agency (ASACTEI), Project AC-00010-18, Res. No. 117/14. This equipment is part of the National High Performance Computing System of the Ministry of Science and Technology of Argentina. We also acknowledged the support of NVIDIA Corporation for the donation of GPUs used for this research.
Conflict of Interest: none to be declared.

Biographies

L.A. Bugnon is an Assistant Researcher in the Bioinformatics lab at sinc(i) institute, National Scientific and Technical Research Council (CONICET), and Assistant Professor at Universidad Nacional del Litoral (UNL), Santa Fe, Argentina. He works in machine learning research applied to bioinformatics.

A. Edera is a Postdoctoral Researcher in the Bioinformatics lab at sinc(i) institute. His research includes computational aspects of biological problems with focus on plants.

S. Prochetto is a Postdoctoral Researcher in the Evolution and Development lab at IAL institute, CONICET, and works closely with the Bioinformatics lab at sinc(i). His research is focused on the evolution of photosynthesis using a bioinformatics approach.

M. Gerard is an Assistant Researcher at sinc(i), CONICET, and Teaching Assistant in the Department of Informatics at UNL. His research interests include bioinformatics, machine learning and swarm intelligence.

J. Raad is a Doctoral Researcher in the Bioinformatics lab at sinc(i) institute, CONICET, and Adjunct Professor at UNL. His research interests include bioinformatics and machine learning.

E. Fenoy is a bioinformatician with experience in immunoinformatics and proteomics. He currently holds a postdoctoral position at the Research institute for signals, systems and computational intelligence, sinc(i), CONICET.

M. Rubiolo is an Assistant Researcher at sinc(i), CONICET, and Teaching Assistant at UNL and Adjunct Professor at UTN. His research interests include machine learning, data mining and bioinformatics.

U. Chorostecki is a postdoctoral researcher in the Comparative Genomics group jointly affiliated to the Biomedical Research Institute (IRB) and the Barcelona Supercomputing Centre (BSC), at Barcelona

(Spain). He is interested in the study of the relationship between structure, function, and evolution in lncRNAs.

T. Gabaldón is ICREA Research Professor and Group Leader of the Comparative Genomics group at the Barcelona Supercomputing Centre (BSC) and the Institute for Research in Biomedicine (IRB) at Barcelona (Spain). He is interested in organismal and molecular evolution.

F. Ariel is Professor in the Faculty of Biochemistry at UNL, and Independent Researcher at the IAL Institute, CONICET. He is the leader of the Epigenetics and noncoding RNAs lab at IAL. His current research interest involves plant long noncoding RNAs in the regulation of genome topology dynamics.

L. Di Persia is Professor in the Department of Informatics at UNL, and Independent Researcher at the sinc(i) institute, CONICET. His research interests include biomedical signal processing, machine learning and Bioinformatics

D.H. Milone is Full Professor in the Department of Informatics at UNL and Principal Research Scientist at sinc(i), CONICET. He was founder and first Director of the sinc(i) institute. His research interests include statistical learning, signal processing and neural computing, with applications to biomedical signals and bioinformatics.

G. Stegmayer is Professor in the Department of Informatics at UNL, and Independent Researcher at the sinc(i) institute, CONICET, Argentina. She is the Leader of the Bioinformatics lab at sinc(i). Her current research interest involves machine learning, data mining and pattern recognition in bioinformatics.

References

[1]Achar, A. and Sætrom, P. (2015). RNA motif discovery: a computational overview. *Biology Direct*, **10**(1).

[2]Anastasiadou, E., Jacob, L., and Slack, F. (2018). Non-coding rna networks in cancer. *Nat Rev Cancer*, **18**(1), 5–18.

[3]Andronescu, M., Bereg, V., Hoos, H. H., and Condon, A. (2008). RNA STRAND: The RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**(1).

[4]Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**(4096), 223–230.

[5]Ariel, F., Romero-Barrios, N., Jégu, T., Benhamed, M., and Crespi, M. (2015). Battles and hijacks: noncoding transcription in plants. *Trends in Plant Science*, **20**(6), 362–371.

[6]Bellaousov, S. and Mathews, D. H. (2010). ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *RNA*, **16**(10), 1870–1880.

[7]Brown, J. A., Bulkley, D., Wang, J., Valenstein, M. L., Yario, T. A., Steitz, T. A., and Steitz, J. A. (2014). Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. *Nature Structural & Molecular Biology*, **21**(7), 633–640.

[8]Chen, X., Li, Y., Umarov, R., Gao, X., and Song, L. (2020). Rna secondary structure prediction by learning unrolled algorithms.

[9]Chillón, I. and Pyle, A. M. (2016). Inverted repeatAluelements in the human lincRNA-p21 adopt a conserved secondary structure that regulates RNA function. *Nucleic Acids Research*, page gkw599.

[10]Chorostecki, U., Willis, J. R., Saus, E., and Gabaldon, T. (2021a). Profiling of RNA structure at single-nucleotide resolution using nextPARS. In *Methods in Molecular Biology*, pages 51–62. Springer US.

[11]Chorostecki, U., Saus, E., and Gabaldón, T. (2021b). Structural characterization of NORAD reveals a stabilizing role of spacers and two new repeat units. *Computational and Structural Biotechnology Journal*, **19**, 3245–3254.

[12]Danaee, P., Rouches, M., Wiley, M., Deng, D., Huang, L., and Hendrix, D. (2018). bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research*, **46**(11), 5381–5394.

[13]Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2008). Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, **106**(1), 97–102.

[14]Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, **7**(1), 1–30.

[15]Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014). In vivo genome-wide profiling of rna secondary structure reveals novel regulatory features. *Nature*, **505**(7485), 696–700.

[16]Do, C., Woods, D., and Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**(14), e90–e98.

[17]Fallmann, J., Will, S., Engelhardt, J., Grüning, B., Backofen, R., and Stadler, P. F. (2017). Recent advances in RNA folding. *Journal of Biotechnology*, **261**, 97–104.

[18]Fatica, A. and Bozzoni, I. (2013). Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics*, **15**(1), 7–21.

[19]Flynn, R. A. and Chang, H. Y. (2014). Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell*, **14**(6), 752–761.

[20]Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., and Xie, X. (2020). UFold: Fast and accurate RNA secondary structure prediction with deep learning.

[21]Fürtig, B., Richter, C., Wöhnert, J., and Schwalbe, H. (2003). NMR spectroscopy of RNA. *ChemBioChem*, **4**(10), 936–962.

[22]Graf, J. and Kretz, M. (2020). From structure to function: Route to understanding lncrna mechanism. *BioEssays*, **42**(12), 2000027.

[23]Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**(8), 1735–1780.

[24]Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D. A., and Mathews, D. H. (2019). LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics*, **35**(14), i295–i304.

[25]Ilik, I. A., Quinn, J. J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R. C., Luscombe, N., Backofen, R., Chang, H. Y., and Akhtar, A. (2013). Tandem stem-loops in roX RNAs act together to mediate x chromosome dosage compensation in drosophila. *Molecular Cell*, **51**(2), 156–173.

[26]Janssen, S. and Giegerich, R. (2014). The RNA shapes studio. *Bioinformatics*, **31**(3), 423–425.

[27]Jones, A. N., Pisignano, G., Pavelitz, T., White, J., Kinisu, M., Forino, N., Albin, D., and Varani, G. (2020). An evolutionarily conserved RNA structure in the functional core of the lincRNA cyrano. *RNA*, **26**(9), 1234–1246.

[28]Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**(7873), 583–589.

[29]Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., and Gao, G. (2017). Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research*, **45**(W1), W12–W16.

[30]Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermuller, J., Hofacker, I. L., *et al.* (2007). Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, **316**(5830), 1484–1488.

[31]Keel, A. Y., Rambo, R. P., Batey, R. T., and Kieft, J. S. (2007). A general strategy to solve the phase problem in RNA crystallography. *Structure*, **15**(7), 761–772.

[32]Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**(7311), 103–107.

[33]Lee, H., Zhang, Z., and Krause, H. M. (2019). Long noncoding RNAs and repetitive elements: Junk or intimate evolutionary partners? *Trends in Genetics*, **35**(12), 892–902.

[34]Lee, J. T. and Bartolomei, M. S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell*, **152**(6), 1308–1323.

[35]Liu, F., Somarowthu, S., and Pyle, A. M. (2017). Visualizing the secondary and tertiary architectural domains of lncRNA RepA. *Nature Chemical Biology*, **13**(3), 282–289.

[36]Lorenz, R., Bernhart, S. H., zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, **6**(1).

[37]Loughrey, D., Watters, K. E., Settle, A. H., and Lucks, J. B. (2014). Shape-seq 2.0: systematic optimization and extension of high-throughput chemical probing of rna secondary structure with next generation sequencing. *Nucleic acids research*, **42**(21), e165–e165.

[38]Lucero, L., Ferrero, L., Fonouni-Farde, C., and Ariel, F. (2020). Functional classification of plant long noncoding RNAs: a transcript is known by the company it keeps. *New Phytologist*, **229**(3), 1251–1260.

[39]Mao, K. and Xiao, Y. (2021). Learning the fastest RNA folding path based on reinforcement learning and monte carlo tree search. *Molecules*, **26**(15), 4420.

[40]Mathews, D. H. (2019). How to benchmark RNA secondary structure prediction accuracy. *Methods*, **162-163**, 60–67.

[41]Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, **288**(5), 911–940.

[42]Mortimer, S. A., Kidwell, M. A., and Doudna, J. A. (2014). Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, **15**(7), 469–479.

[43]Murthy, V. L. (2003). RNABase: an annotated database of RNA structures. *Nucleic Acids Research*, **31**(1), 502–504.

[44]Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**(1).

[45]Rinn, J. L. and Chang, H. Y. (2020). Long noncoding RNAs: Molecular modalities to organismal functions. *Annual Review of Biochemistry*, **89**(1), 283–308.

[46]Rizki, G. and Boyer, L. A. (2015). Lncding epigenetic control of transcription to cardiovascular development and disease. *Circulation Research*, **117**(2), 192–206.

[47]Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F., and Crespi, M. (2018). Splicing regulation by long noncoding RNAs. *Nucleic Acids Research*, **46**(5), 2169–2184.

[48]Ross, C. J. and Ulitsky, I. (2022). Discovering functional motifs in long noncoding RNAs. *WIREs RNA*.

[49]Sato, K., Hamada, M., Asai, K., and Mituyama, T. (2009). CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Research*, **37**(Web Server), W277–W280.

[50]Sato, K., Kato, Y., Hamada, M., Akutsu, T., and Asai, K. (2011). IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**(13), i85–i93.

[51]Sato, K., Akiyama, M., and Sakakibara, Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications*, **12**(1).

[52]Saus, E., Willis, J. R., Pryszcz, L. P., Hafez, A., Llorens, C., Himmelbauer, H., and Gabaldón, T. (2018). nextPARS: parallel probing of RNA structures in illumina. *RNA*, **24**(4), 609–619.

[53]Schroeder, S. J. and Turner, D. H. (2009). Optical melting measurements of nucleic acid thermodynamics. In *Methods in Enzymology*, pages 371–387. Elsevier.

[54]Seetin, M. G. and Mathews, D. H. (2012). RNA structure prediction: An overview of methods. In *Bacterial Regulatory RNA*, pages 99–122. Humana Press.

[55]Singh, J., Hanson, J., Paliwal, K., and Zhou, Y. (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, **10**(1).

[56]Singh, J., Paliwal, K., Zhang, T., Singh, J., Litfin, T., and Zhou, Y. (2021). Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics*, **37**(17), 2589–2600.

[57]Smola, M. J., Christy, T. W., Inoue, K., Nicholson, C. O., Friedersdorf, M., Keene, J. D., Lee, D. M., Calabrese, J. M., and Weeks, K. M. (2016). Shape reveals transcript-wide interactions, complex structural domains, and protein interactions across the xist lncrna in living cells. *Proceedings of the National Academy of Sciences*, **113**(37), 10322–10327.

[58]Somarowthu, S., Legiewicz, M., Chillón, I., Marcia, M., Liu, F., and Pyle, A. M. (2015). HOTAIR forms an intricate and modular secondary structure. *Molecular Cell*, **58**(2), 353–361.

[59]Spokoini-Stern, R., Stamov, D., Jessel, H., Aharoni, L., Haschke, H., Giron, J., Unger, R., Segal, E., Abu-Horowitz, A., and Bachelet, I. (2020). Visualizing the structure and motion of the long noncoding rna hotair. *RNA*, **26**(5), 629–636.

[60]Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding rnas and its biological functions. *Nature Reviews Molecular Cell Biology*, **22**(2), 96–118.

[61]Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLOS Biology*, **13**(7), e1002195.

[62]Sztuba-Solinska, J., Rausch, J. W., Smith, R., Miller, J. T., Whitby, D., and Grice, S. F. L. (2017). Kaposi's sarcoma-associated herpesvirus polyadenylated nuclear RNA: a structural scaffold for nuclear, cytoplasmic and viral proteins. *Nucleic Acids Research*, **45**(11), 6805–6821.

[63]The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

[64]Tichon, A., Gil, N., Lubelsky, Y., Solomon, T. H., Lemze, D., Itzkovitz, S., Stern-Ginossar, N., and Ulitsky, I. (2016). A conserved abundant cytoplasmic long noncoding RNA modulates repression by pumilio proteins in human cells. *Nature Communications*, **7**(1).

[65]Tichon, A., Perry, R. B.-T., Stojic, L., and Ulitsky, I. (2018). SAM68 is required for regulation of pumilio by the NORAD long noncoding RNA. *Genes & Development*, **32**(1), 70–78.

[66]Turner, D. H. and Mathews, D. H. (2009). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, **38**(suppl_1), D280–D282.

[67]Ulitsky, I. and Bartel, D. P. (2013). lincRNAs: Genomics, evolution, and mechanisms. *Cell*, **154**(1), 26–46.

[68]Uroda, T., Anastasakou, E., Rossi, A., Teulon, J.-M., Pellequer, J.-L., Annibale, P., Pessey, O., Inga, A., Chillón, I., and Marcia, M. (2019). Conserved pseudoknots in lncRNA MEG3 are essential for stimulation of the p53 pathway. *Molecular Cell*, **75**(5), 982–995.e9.

[69]van Batenburg, F. H. D. (2000). PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Research*, **28**(1), 201–204.

[70]Vaswani, A. *et al.* (2017). Attention is all you need. NIPS'17, page 6000âŁ"6010, Red Hook, NY, USA. Curran Associates Inc.

[71]Willmott, D., Murrugarra, D., and Ye, Q. (2020). Improving RNA secondary structure prediction via state inference with deep recurrent neural networks. *Computational and Mathematical Biophysics*, **8**(1), 36–50.

[72]Wu, L. and Belasco, J. G. (2008). Let me count the ways: Mechanisms of gene regulation by miRNAs and siRNAs. *Molecular Cell*, **29**(1), 1–7.

[73]Zakov, S., Goldberg, Y., Elhadad, M., and Ziv-ukelson, M. (2011). Rich parameterization improves RNA structure prediction. *Journal of Computational Biology*, **18**(11), 1525–1542.

[74]Zampetaki, A., Albrecht, A., and Steinhofel, K. (2018). Long non-coding RNA structure and function: Is there a link? *Frontiers in Physiology*, **9**.

[75]Zhang, H., Zhang, C., Li, Z., Li, C., Wei, X., Zhang, B., and Liu, Y. (2019). A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. *Frontiers in Genetics*, **10**.

[76]Zhang, H., Zhang, L., Mathews, D. H., and Huang, L. (2020). LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics*, **36**(Supplement_1), i258–i267.

[77]Zhao, Q., Zhao, Z., Fan, X., Yuan, Z., Mao, Q., and Yao, Y. (2021). Review of machine learning methods for RNA secondary structure prediction. *PLOS Computational Biology*, **17**(8), e1009291.

[78]Ziv, O., Farberov, S., Lau, J. Y., Miska, E., Kudla, G., and Ulitsky, I. (2021). Structural features within the NORAD long noncoding RNA underlie efficient repression of pumilio activity.

[79]Zou, Q., Li, J., Hong, Q., Lin, Z., Wu, Y., Shi, H., and Ju, Y. (2015). Prediction of MicroRNA-disease associations based on social network analysis methods. *BioMed Research International*, **2015**, 1–9.

[80]Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, **9**(1), 133–148.