# UNIVERSITÀ DEGLI STUDI DI TORINO

*Structural bioinformatics*

# RNA Structure Framework: Automated transcriptome-wide reconstruction of RNA secondary structures from high-throughput structure probing data

Danny Incarnato[1,2], Francesco Neri[1], Francesca Anselmi[1,2], and Salvatore Oliviero[1,2]*.

[1] Human Genetics Foundation (HuGeF), via Nizza 52, 10126 Torino, Italy.

[2] Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università di Torino, Via Accademia Albertina, 13 - 10123 Torino.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

**ABSTRACT**

**Summary:** The rapidly increasing number of discovered non-coding RNAs makes the understanding of their structure a key feature toward a deeper comprehension of gene expression regulation. Various enzymatic- and chemically-based approaches have been recently developed to allow whole-genome studies of RNA secondary structures. Several methods have been recently presented that allow high-throughput RNA structure probing (CIRS-seq, Structure-seq, SHAPE-seq, PARS, etc.) and unbiased structural inference of residues within RNAs in their native conformation. We here present an analysis toolkit, named RNA Structure Framework (RSF), which allows fast and fully-automated analysis of high-throughput structure probing data, from data pre-processing to whole-transcriptome RNA structure inference.

**Availability and Implementation:** RSF is written in Perl and is freely available under the GPLv3 license from http://rsf.hugef-research.org.

**Contact:** salvatore.oliviero@hugef-torino.org

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

## 1 INTRODUCTION

The advent of high-throughput method has rapidly led to the annotation of thousands of novel transcripts (ENCODE Project Consortium *et al.*, 2012; Derrien *et al.*, 2012; Djebali *et al.*, 2012), mostly lacking coding capabilities (Bánfai *et al.*, 2012; Guttman *et al.*, 2013). As for large ribonucleoprotein complexes (Krummel *et al.*, 2010; Nagai *et al.*, 2001), these RNAs are thought to regulate gene expression through interactions mediated by their structure (Tsai *et al.*, 2010; Wang and Chang, 2011). In the last years, a variety of methods have been developed to interrogate RNA secondary structures on a genome-wide scale (Underwood *et al.*, 2010; Li *et al.*, 2012; Kertesz *et al.*, 2010; Wan *et al.*, 2012; Lucks *et al.*, 2011; Ding *et al.*, 2014; Rouskin *et al.*, 2014; Wan *et al.*, 2014; Incarnato *et al.*, 2014), but no tool has been developed to enable efficient analysis of the large amount of data generated by these methods, with the exception of the *SeqFold* package, which has been developed to analyze PARS data (Ouyang *et al.*, 2013). Since the analysis of these data and the subsequent inference of RNA structures constitutes a bottle-neck of these protocols, we here provide the RNA Structure Framework (RSF), an open-source framework to analyze high-throughput structure probing data, and to minimize the efforts to get from raw sequencing reads to secondary structures.

## 2 IMPLEMENTATION

RSF is implemented in Perl as a modular package. An outline of a sample RSF data analysis workflow is shown in Figure 1. Detailed information on the core modules is provided in the Supplementary Material, as well in the package archive.

RSF is composed of three core modules, and a number of other utilities. The *reference-builder* module builds the transcriptome reference used in the read-mapping step. This module requires an Internet connection to query the UCSC genome SQL database (http://genome.ucsc.edu) and obtain transcript annotation. Since the *reference-builder* module also requires the genome reference sequence for the species of interest, this can be either provided by the user in multi-FASTA format, or can be automatically downloaded from the UCSC DAS server (http://genome.ucsc.edu/cgi-bin/das/). The *analyzer* module is the core of the framework. It requires a reference index, and a set of at least 2 FastQ files (3 for CIRS-seq), one for each condition (e.g. non treated control and DMS treatment). Reads are pre-processed using the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to clip adapter sequences, while mapping of reads on the reference index is performed using the Bowtie v1 tool (Langmead *et al.*, 2009). Alternatively, the user can use different tools for reads mapping, and then provide to the module the SAM files instead of FastQ files. Following the mapping step, reads are sorted and the number of reads mapping at each position of each transcript are calculated in the provided conditions. Since each read gives information only on the base immediately preceding the start mapping position (Figure S1), the module automatically accounts for eventual trimming of bases from the 5' of the read. Once raw counts have been computed, the *analyzer* module calculates normalized raw reactivity scores using one of two possible scoring schemes. The first scheme (Kertesz *et al.*, 2010; Incarnato *et al.*, 2014) assumes an uniform distribution of read mappings across samples, but it is

*To whom correspondence should be addressed.

more susceptible to cross-sample variations since it uses the total number of mapped reads to normalize libraries for different sequencing depths. The second scheme (Ding *et al.*, 2014) instead, is less affected by sample-to-sample variations in the distribution of read mappings, as it performs per-transcript normalization by considering the average number of RT-stops across the transcript being analyzed.

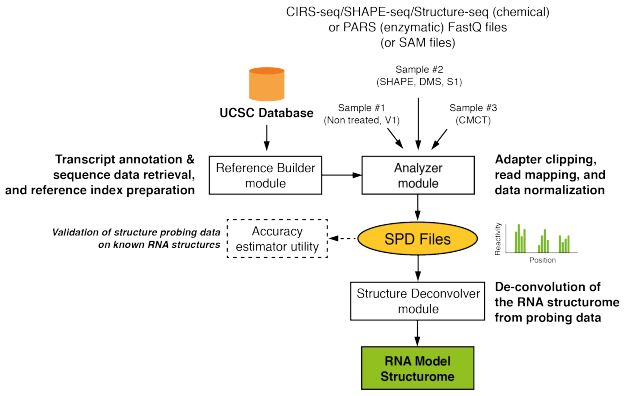


**Fig. 1.** Overview of the RNA Structure Framework pipeline.

Finally, overall reactivities for each position of each analyzed transcript are normalized to values ranging from 0 (less single-stranded) to 1 (more single-stranded). The *analyzer* module actually incorporates three different normalization methods: 2-8% normalization (Lucks *et al.*, 2011), 90% winsorising (Incarnato *et al.*, 2014), and Box-plot normalization (Low and Weeks, 2010). For each transcript being analyzed, the module reports data in a text-based Structure Probing Data (SPD) file (see Supplementary Material). The third, and last, core component of the toolkit is the *structure-deconvolver* module. This module takes a set of SPD files and deconvolutes RNA structures from experimental data, using one of three different approaches (Fig. S2): (1) Hard-constrained structure prediction using the ViennaRNA package (Lorenz *et al.*, 2011) , (2) Soft-constrained structure prediction using the RNAstructure tool (Reuter and Mathews, 2010), or an (3) Iterative cluster-refinement approach. The hard constraint method uses the ViennaRNA package to predict a minimum free energy (MFE) structure by imposing the constraint that RNA positions exceeding a given reactivity cutoff (default: 0.7) must be unpaired. The soft constraint approach uses instead the whole set of reactivity data for an RNA, by first converting it into a SHAPE data file (using the *spd2shape* utility provided with the RSF package), that can be supplied to the RNAstructure software. RNAstructure then uses this data to compute a pseudo-energy term to adjust the free energy of individual nucleotides (Deigan *et al.*, 2009). The third approach (Fig. S3), instead, is a variant of that employed by SeqFold (Ouyang *et al.*, 2013) software. Briefly, the partition folding for the RNA is computed using the ViennaRNA package, then a backtracking through the Boltzmann ensemble of structures is performed, and structures are then clustered using Hamming distance, with a low-stringency cutoff (default: *distance* = 0.5). The base-pair probability profile (BPP) for each cluster is then calculated, and the cluster that better correlates to RNA reactivity data is selected. Following the best-fitting cluster selection, the cluster is iteratively refined by performing a progressively more stringent clustering (default: *distance* = *distance* - 0.01), followed by the subtraction of individual structures that contribute to lower the correlation coefficient. Finally, the higher-correlation cluster of structures (Boltzmann sub-ensemble) is returned, as well as the minimum expected accuracy (MEA) structure for the RNA (defined as the structure in which only the concordant base-pairs in at least the 50% of cluster structures are reported).

Predicted structures can be reported in dot-bracket or connectivity table (CT) notations, as well as in PostScript or SVG graphical formats.

Additional tools and utilities are also shipped with the RSF package, and are described in the Supplementary Information.

## 3 CONCLUSION

The advent of high-througput RNA structure probing methods has provided a large amount of transcriptome-wide scale structural data, although robust tools for the rapid elaboration of this information are currently missing. RNA Structure Framework is a user-friendly toolkit that enables automated inference of RNA secondary structures on a transcriptome-wide scale, in a few steps. It can use data derived from many different structure probing methods, both chemical (CIRS-seq, SHAPE-seq, Structure-seq), and enzymatic (PARS). Due to the rapid evolution of the field, and to the absence of a *golden standard* in the analysis of high-throughput structure probing data, the implementation of several different scoring, normalization and structure prediction methods enables a high degree of analysis flexibility to the user. The use of high-throughput structure probing methodologies coupled to RSF provides an important toolkit for the genome-wide analysis of RNA structures.

*Conflict of interest:* none declared.

## ABBREVIATIONS

CIRS-seq: Chemical Inference of RNA Structures (Incarnato *et al.*, 2014)
PARS: Parallel Analysis of RNA Structure (Kertesz *et al*., 2010)
RSF: RNA Structure Framework
SAM: Sequence Alignment/Map format (http://samtools.sourceforge.net/)
SHAPE-seq: Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (Lucks *et al.*, 2011)
SPD: Structure Probing Data format
SQL: Structured Query Language
SVG: Scalable Vector Graphics format

## REFERENCES

Bánfai,B. *et al.* (2012) Long noncoding RNAs are rarely translated in two human cell lines. *Genome Research,* **22**,1646–1657

Deigan,K.E. *et al.* (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 97–102.

Derrien,T. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research,* **22**, 1775-1789

Ding,Y. *et al.* (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, **505**, 696–700.

Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

ENCODE Project Consortium *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Guttman,M. *et al.* (2013) Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell*, **154**, 240–251.

Incarnato,D. *et al.* (2014) Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biol.*, **15**, 491.

Kertesz,M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.

Krummel,D.A.P. *et al.* (2010) Structure of spliceosomal ribonucleoproteins. *F1000 Biol Rep*, **2**.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,F. *et al.* (2012) Global analysis of RNA secondary structure in two metazoans. *Cell Rep*, **1**, 69–82.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.

Low,J.T. and Weeks,K.M. (2010) SHAPE-directed RNA secondary structure prediction. *Methods*, **52**, 150–158.

Lucks,J.B. *et al.* (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 11063–11068.

Nagai,K. *et al.* (2001) Structure and assembly of the spliceosomal snRNPs. *Biochem. Soc. Trans.*, **29**, 15–26.

Ouyang,Z. *et al.* (2013) SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.*, **23**, 377–387.

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

Rouskin,S. *et al.* (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.

Tsai,M.-C. *et al.* (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.

Underwood,J.G. *et al.* (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods*, **7**, 995–1001.

Wan,Y. *et al.* (2012) Genome-wide measurement of RNA folding energies. *Mol. Cell*, **48**, 169–181.

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

Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, **43**, 904–914.