Bioinformatics, 33(3), 2017, 441–443
doi: 10.1093/bioinformatics/btw627
Advance Access Publication Date: 30 September 2016
Applications Note

OXFORD

Structural bioinformatics

SEQualyzer: interactive tool for quality control and exploratory analysis of high-throughput RNA structural profiling data

Krishna Choudhary, Luyao Ruan, Fei Deng, Nathan Shih and Sharon Aviran*

Department of Biomedical Engineering and Genome Center, University of California at Davis, Davis, CA 95616, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 1, 2016; revised on September 25, 2016; accepted on September 26, 2016

Abstract

Summary: To serve numerous functional roles, RNA must fold into specific structures. Determining these structures is thus of paramount importance. The recent advent of high-throughput sequencing-based structure profiling experiments has provided important insights into RNA structure and widened the scope of RNA studies. However, as a broad range of approaches continues to emerge, a universal framework is needed to quantitatively ensure consistent and high-quality data. We present SEQualyzer, a visual and interactive application that makes it easy and efficient to gauge data quality, screen for transcripts with high-quality information and identify discordant replicates in structure profiling experiments. Our methods rely on features common to a wide range of protocols and can serve as standards for quality control and analyses.

Availability and Implementation: SEQualyzer is written in R, is platform-independent, and is freely available at http://bme.ucdavis.edu/aviranlab/SEQualyzer.

Contact: saviran@ucdavis.edu

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

1. Introduction

RNA serves many important functions in cells, which require folding into specific structures. This necessitates identification of RNA structures. Data from structure profiling experiments have yielded important insights into RNA structure (Mortimer *et al.*, 2014) and helped improve computational structure-prediction accuracy (Deigan *et al.*, 2009; Lorenz *et al.*, 2015). Experiments use reagents that modify RNA residues in a structure-dependent manner (Weeks, 2010). Modifications are detected as terminations or mutations via reverse transcription and subsequent sequencing (Kutchko and Laederach, 2016). Noise in detection is measured using a control assay. These measurements are then combined to yield final chemical reactivity scores, which bear structural information for each residue. A wide range of recent and emerging techniques harness high-throughput sequencing to probe structure in a massively parallel fashion, driving the field towards transcriptome-wide (TW) studies (Lu and Chang, 2016). Whereas quality control of structural data has been traditionally addressed with visual inspection or simple tests, the scale and complexity of recent datasets preclude such approaches. This poses a need for automated quality control and experiment design (Aviran and Pachter, 2014; Choudhary *et al.*, 2016). Here, we present SEQualyzer (**S**tructure-profiling **E**xperiment **Qua**lity Ana**lyzer**), which serves this need by providing easy-to-interpret quality metrics and visualizations thereof in an interactive R Shiny application. SEQualyzer rapidly and quantitatively evaluates data reproducibility to identify transcripts of high quality, making use of novel metrics developed in previous work (Choudhary *et al.*, 2016). It further provides data summaries, allowing complete flexibility to optimize reactivity scoring schemes. SEQualyzer is applicable to a range of protocols notwithstanding differences in reagents, modification detection methods, priming strategies, sequencing choices (single-end or paired-end), or scoring

schemes (see Manual for diverse examples). It thus has the potential to serve as the go-to tool for quality control and analyses.

2. Methods

2.1 Data input

SEQualyzer takes read counts information (single/paired-end) for single or multiple transcripts and replicates thereof. Inputs are tallied stop counts at each residue in a format that is easy to conform to (see SEQualyzer Manual). Users can either process raw reads into such format independently or use the StructureFold platform for this purpose (Tang *et al.*, 2015). Optional inputs are local coverages at each residue (see Supplementary Information). Sequence information is input in FASTA format.

2.2 Reactivity scoring strategies

Users have flexibility to optimize reactivity scores, choosing between difference (Aviran *et al.*, 2011; Tang *et al.*, 2015) or ratio (Talkish *et al.*, 2014) of detection frequencies in modified and control assays. One can also choose to subsequently apply standard normalization routines and/or log-transform read counts or reactivities (Sloma and Mathews, 2015).

2.3 Quality measures

We calculate quality estimates, namely signal-to-noise ratio (SNR) and Coverage Quality Index (CQI), as previously described (Choudhary *et al.*, 2016). SNR is calculated for each residue as the ratio of mean to standard deviation of a reactivity score, obtained from experimental or simulated replicates. It can be adapted for different purposes:

1. To evaluate replicates, per-residue SNR is calculated from simulated replicates, obtained via bootstrap (as in Fig. 1). Results are plotted and summarized as mean of SNR values across residues. Optionally, a formula to estimate SNR (see Supplementary Information) can rapidly evaluate large-scale datasets.
2. To gauge replicate agreement, SNR is applied to real replicate data.
3. To facilitate identification of discordant replicates, pairwise comparisons are illustrated as a correlation matrix and pairwise mean SNR.
4. To identify regions with poor-quality data, rolling mean SNR is plotted for a user-defined window size. For example, in Figure 1, rolling mean SNR indicates that data quality decreases towards 5' end and is poor for almost half the length of the *hepC* IRES domain.
5. To obtain high-level summaries, SNR histograms are generated.
6. To sift through TW data, transcripts are scored and ranked by their mean SNR values.

While SNR measures variability, CQI assesses the coverage required to ensure user-specified data quality. Users provide an acceptable percentage error in the mean reactivity. A significance level, representing the probability that reactivities can be reproduced within allowed range of error, is also input. From these parameters, we calculate the variance, assuming reactivities would be normally distributed around observed values if the experiment were to be repeated. Next, to estimate desirable local coverage, we use a formula that links it with variance, reactivity and noise at each residue (see Supplementary Information). Taking ratio of desired to observed coverages provides indices for all residues of a transcript. We summarize the indices for three reactivity categories—high, medium and low—as 95th percentile of the indices per category, which we call

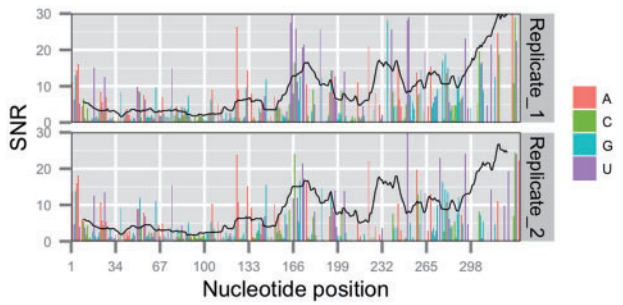

Fig. 1. Sample result: Per-residue (bars) and rolling mean (black line; window size 20) SNR via bootstrap for *hepC* IRES domain replicates (data: Loughrey *et al.*, 2014) (Color version of this figure is available at *Bioinformatics* online.)

CQI. CQIs less/more than 1 indicate sufficient/insufficient coverage. Additionally, SEQualyzer depicts error bars around reactivities and summarizes data as stop or mutation counts, local coverage, reactivity distributions and read count tallies for A, C, G and U. Summaries of the distribution of reads among transcripts are provided as Lorenz curve and histogram of coverages.

2.4 Application features

Quality metrics are implemented for visualization in R Shiny interface. Starting from TW data, users can filter transcripts based on coverage, length, or mean SNR, can easily switch between selected transcripts listed in a menu, and can zoom into regions of transcripts. Thus, SEQualyzer enables users to scope data from residue level to whole transcriptome level. By using the R *parallel* package, we speed up bootstrap and other time-consuming computations. Users can choose to include sequence information in plots as well as save all analyses.

3. Conclusions

SEQualyzer provides insights into the quality of high-throughput RNA structural data and helps identify high-quality components in an interactive and efficient framework. Its outputs permit easy quality evaluation and identification of discordant replicates. Metrics such as CQI can be used as a guideline for experiment design. Since SEQualyzer is applicable to a range of assays, it will help bridge differences in protocols and standardize how quality is assessed. As new data formats and quality control standards emerge, SEQualyzer can be readily extended to be even more comprehensive.

Funding

This work was supported by National Institutes of Health grant [HG006860].

Conflict of Interest: none declared.

References

Aviran,S. *et al.* (2011) Modeling and automation of sequencing-based characterization of RNA structure. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 11069–11074.

Aviran,S. and Pachter,L. (2014) Rational experiment design for sequencing-based RNA structure mapping. *RNA*, **20**, 1864–1877.

Choudhary,K. *et al.* (2016) Metrics for rapid quality control in RNA structure probing experiments. *Bioinformatics*, doi:10.1093/bioinformatics/btw501.

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

Kutchko,K.M. and Laederach,A. (2016) Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *WIREs RNA*, doi:10.1002/wrna.1374.

Lorenz,R. *et al*. (2015) SHAPE directed RNA folding. *Bioinformatics*, **32**, 145–147.

Loughrey,D. *et al*. (2014) SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res*., **42**, e165.

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

Mortimer,S.A. *et al*. (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet*., **15**, 469–479.

Sloma,M.F. and Mathews,D.H. (2015) Improving RNA secondary structure prediction with structure mapping data. *Methods Enzymol*., **553**, 91–114.

Talkish,J. *et al*. (2014) Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA*, **20**, 713–720.

Tang,Y. *et al*. (2015) StructureFold: Genome-wide RNA secondary structure mapping and reconstruction *in vivo*. *Bioinformatics*, **31**, 2668–2675.

Weeks,K.M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol*., **20**, 295–304.