Methods xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

StructureFold2: Bringing chemical probing data into the computational fold of RNA structural analysis

David C. Tack [a,b], Yin Tang [c], Laura E. Ritchey [b,d], Sarah M. Assmann [a,d,*], Philip C. Bevilacqua [b,d,e,*]

[a] Department of Biology, Pennsylvania State University, University Park, PA 16802, USA
[b] Department of Chemistry, Pennsylvania State University, University Park, PA 16802, USA
[c] Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA
[d] Center for RNA Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA
[e] Department of Biochemistry & Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 1 November 2017
Received in revised form 12 January 2018
Accepted 29 January 2018
Available online xxxx

Keywords:
DMS
Glyoxal
RNA structure
SHAPE
Structure-seq
Structurome

ABSTRACT

The secondary structure of an RNA is often implicit to its function. Recently, various high-throughput RNA structure probing techniques have been developed to elucidate important RNA structure–function relationships genome-wide. These techniques produce unwieldy experimental data sets that require evaluation with unique computational pipelines. Herein, we present StructureFold2, a user-friendly set of analysis tools that makes precise data processing and detailed downstream analyses of such data sets both available and practical. StructureFold2 processes high-throughput reads sequenced from libraries prepared after experimental probing for reverse transcription (RT) stops generated by chemical modification of RNA at solvent accessible residues. This pipeline is able to analyze reads generated from a variety of structure-probing chemicals (e.g. DMS, glyoxal, SHAPE). Notably, StructureFold2 offers a new fully featured suite of utilities and tools to guide a user through multiple types of analyses. A particular emphasis is placed on analyzing the reactivity patterns of transcripts, complementing their use as folding restraints for predicting RNA secondary structure. StructureFold2 is hosted as a Github repository and is available at (https://github.com/StructureFold2/StructureFold2).

1. Introduction

The roles that RNAs are known to play in the cell have expanded greatly in the past few decades. RNA was first recognized as a rather humble messenger in the classical central dogma of molecular biology, yet today it has been demonstrated to be both a regulator and regulated, catalyst and catalyzed, capable of assuming protean roles including perhaps the original vitalizing rudiment of life itself. Classical targeted RNA structural studies have made use of techniques such as NMR, crystallography, and gel-based structural probing through nucleases or chemical reagents to identify structures of individual RNAs [1], but have not been able to determine the structural contribution of the entire transcriptome. More recently, multiple approaches have combined structural-probing techniques with next-generation sequencing (NGS), thus producing structural information on a genome-wide scale [2]. Structure-seq [3–5], one of these methods developed by our labs, is able to probe the structure of RNA transcriptome-wide *in vivo* by using chemicals that penetrate living cells and covalently modify RNA at sites that are single-stranded and unprotected. Dimethyl sulfate (DMS) methylates the Watson-Crick face of unprotected adenine and cytosine residues. Recent improvements in use of glyoxal as a modifying agent for the Watson-Crick face of guanine [6] have opened the prospect of using glyoxal and its derivatives as supplemental probing reagents in Structure-seq. SHAPE reagents can modify the backbone of every base [7]. When subsequently performing reverse transcription on extracted RNA using a random hexamer in Structure-seq, these modifications prevent reverse transcription read-through and thus result in truncated cDNA fragments. An adapter is then ligated onto the 3′ end of these transcripts to allow amplification and Illumina sequencing. The first nucleotide sequenced is immediately downstream of the solvent accessible RNA nucleotide *in vivo*. Higher chemical reactivity corresponds to a higher probability of base single-strandedness. Other NGS structure probing studies use similar methodologies [2].

The emergence of these techniques necessitates the development of a unified set of computational tools to extract and analyze

* Corresponding authors at: Pennsylvania State University, University Park, PA 16802, USA.
E-mail addresses: sma3@psu.edu (S.M. Assmann), pcb5@psu.edu (P.C. Bevilacqua).

https://doi.org/10.1016/j.ymeth.2018.01.018
1046-2023/© 2018 Elsevier Inc. All rights reserved.

the unique data they generate, resolving differential RNA reactivity and structure, in contrast to the widely-available canonical tools for standard RNA-seq analysis, which resolve differential RNA abundance. Properly calibrated reactivity scores require information from combining two individual samples (Fig. 1): libraries produced without a probing reagent must be subtracted from libraries produced with chemical treatment before any comparison between experimental conditions. The minus reagent libraries account for innate reverse transcription stops due in part to natural RNA modifications or strong *in vitro* structure, allowing for a more accurate appraisal of the true degree of chemical modification *in vivo* and thus single-strandedness of individual bases. Chemical reactivity calculations must also take into account transcript nucleotide composition and overall transcript abundance, while comparing the RT-induced stops of every individual base between untreated and treated libraries (Fig. 1). As there are a wide variety of questions that can be asked concerning RNA structure, unified analysis packages that can enable a non-specialist to both accurately prepare their data and perform a wide array of downstream analyses, as provided here with StructureFold2, should greatly contribute to the continuing advancement of this field (see StructureFold2 Manual).

The importance of an *in vivo* RNA structure is often only realized through the accompanying loss or gain of function after a conformational change, for which different conditions may be required. This puts a priority on the ability to rapidly scale the analysis to any amount of data, while at the same time allowing accurate resolution of specific questions. StructureFold2 builds on the strengths of the initial StructureFold suite, which is available at Galaxy [8], while providing high utility and versatility. Through improvements in both the underlying scientific methodology and by reformatting to a new user-friendly implementation that includes a wider variety of tools, we present StructureFold2 as an essential suite of data preparation and analysis tools for working with RNA structure probing data. Computational packages analyzing experimental RNA structure probing data by mutational profiling are also available [9–12]. Due to the modularity of StructureFold2, future iterations should allow such mutational profiling data to be imported into the downstream analysis modules. StructureFold2 allows precise and facile exploration of high-throughput RNA structure data generated by chemical probing followed by next generation sequencing, offering enough simplicity to put basic analyses within the hands of the novice computational biologist, yet enough modularity to enable complex customization at every step for the advanced user (see StructureFold2 Manual).

2. Material and methods

2.1. Improved, more flexible analysis platform

StructureFold2 has shifted from the Galaxy platform, instead emphasizing the flexibility offered through direct distribution of Python [13] scripts via Github. The ability to perform data analysis on a locally controlled system opens up more power and possibilities for StructureFold. Experimenters are free to scale the size and scope of their experiment, and are not tied to a particular remote server. Local control of the scripts allows analyses to be quickly adapted or integrated with other programs, or to be updated when new functionalities become available. The vast majority of StructureFold2 scripts have had a batch processing option added, enabling users to enact one entire analysis step on all of their data at once with simple commands. Thus, processing even large or elaborately designed experiments becomes straightforward and orderly. Output file names are generated automatically, providing streamlining, and preventing a common pitfall in data tracking, especially as more conditions and samples are added.

2.2. Manuals and menus

StructureFold2 aims to put advanced structural analysis within the reach of a researcher with minimal computational background. We have thus added both a detailed standalone manual and a detailed help menu to each individual module. The StructureFold2 Manual contains all of the essential information to get started, common lexicon, information on planning analyses, and tips and tricks to accommodate particular quirks of each study's transcriptome(s). Flowcharts (Supplemental Figs. 1 and 2) illustrate the flow of data through a typical analysis and clarify the use and purpose of each analysis tool. Each module's help menu explains each of the options that can be modified or invoked when executed, adding ease of customizability. However, most of the preset defaults should require few changes for a typical analysis and thus are recommended settings. StructureFold2 requires the use of a read trimming program and a short read aligner (typically cutadapt [14] and Bowtie 2 [15], respectively), and we include scripts to batch run these programs with the recommended settings for a StructureFold2 analysis, further streamlining the process. These scripts can automatically log run information and simplify the analysis by providing consistent intermediate file nomenclature and organization throughout the experiment.

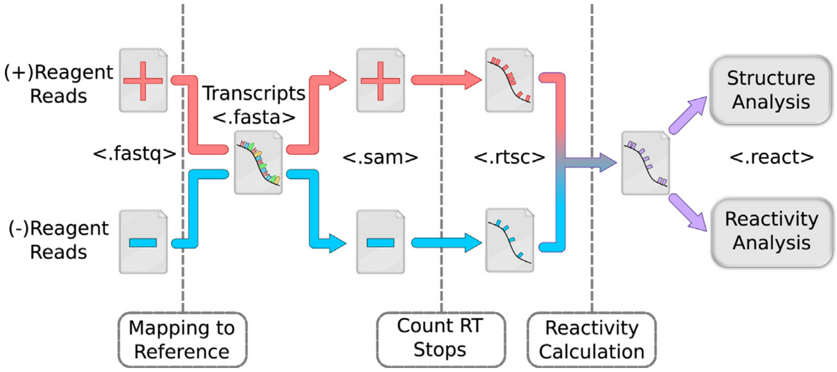

Fig. 1. A synopsis of calculating reactivity values using StructureFold2. Reads from both untreated and probing-reagent-treated libraries are mapped to the reference transcriptome. These mappings are interpreted and summed as per-base reverse transcriptase (RT) stop counts. RT stops from the untreated library are then subtracted from the RT stops from the treated library during the reactivity calculation, yielding an accurate assessment of the *in vivo* chemical reactivity of each base. These reactivity values may be directly analyzed or compared with the values from another condition, inferring structural change as a result of changes in reactivity. Complementary to this approach, the derived reactivity values may guide RNA folding software.