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.

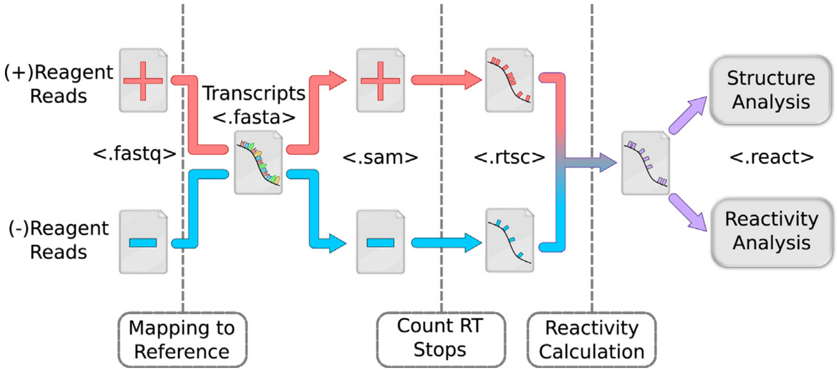

Fig. 1. A synopsis of calculating reactivity values using StructureFold2. Reads from both untreated and probing-reagent-treated libraries are mapped to the reference transcriptome. These mappings are interpreted and summed as per-base reverse transcriptase (RT) stop counts. RT stops from the untreated 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.

2.3. Advances in scientific methodology

In the original StructureFold pipeline [16], we used iterative mapping. This involves mapping the reads to the reference and progressively trimming down reads that did not map, remapping until the read either maps successfully or the read length becomes less than 20 bp. This strategy maximizes the amount of information obtained from each sequencing run. However, with new wet-bench advances such as Structure-seq2 [5], the quality of structure-probing libraries is much improved. Therefore, we now use a more directed and typical mapping paradigm that is much faster. First, an automated Python script guides the user in the removal of specified 5′ and 3′ adapters, low quality bases, and reads that become too short via driving the cutadapt Python package; other read trimming programs may be substituted at this step. Precisely trimmed reads, like those offered by cutadapt and related programs, offer more reliable mappings and take less time to map than mapping via an iterative process, which can be subject to heuristic artifacts. After mapping via an automated script provided to drive Bowtie 2, the mapped reads are subject to an additional quality check step. Reads with a nucleotide mismatch on the most 5′ mapped base, reads with over three mismatches or indels with respect to the reference transcriptome, as well as reads that align anti-sense relative to the annotated transcript are all discarded. This offers enhanced precision by eliminating more artifacts than in previous pipelines, removing both sequencing errors and ambiguous modification signals. These computational improvements build on the molecular improvements found in the Structure-Seq2 protocol [5].

The core equations used to calculate reactivity in StructureFold2 are unchanged from the original StructureFold [16]. However, we have added several options: forgoing 2–8% normalization [17], not taking the natural log when generating reactivity scores, and using the same normalization scale across the reactivity calculation of multiple samples to more precisely measure true changes in reactivity via a uniform normalization scale. When the intention is to compare directly measured chemical reactivity patterns between samples, rather than to compare processed reactivity patterns or folded RNAs generated with reactivities as restraints, applying the 2–8% normalization scale or use of the natural log may overprocess the data. In such cases, disabling these options may offer a more precise look at the raw modification signal. We also have incorporated a module to combine multiple sets of restraints from different probing reagents into one streamlined restraint file.

2.4. Improved analysis toolkit

StructureFold included all the tools necessary to generate per base pair reactivity and pipe these restraints into either RNAStructure [18] or Vienna Package [19] to generate predicted structures. StructureFold2 expands on this foundation, providing a full-fledged suite of tools to allow a more complete and nuanced analysis. First, modules to extract the free energy and single-strandedness of predicted RNA structures have been added to quickly collect and write this information to a convenient .csv file so that the data can be analyzed in a systematic fashion by typical statistics software, such as R [20]. A similar module generates the comparative PPV metric between two directories of folded structures by driving the RNAstructure scorer module, hence consolidating the pairwise structural differences of potentially thousands of RNAs between two conditions into a single file. While none of these metrics based on predicted structures are singularly authoritative, they can provide focus and direction to any study.

In StructureFold2, more emphasis has been placed on the reactivity values rather than on the predicted RNA folds of transcripts; thus, tools are included to batch calculate many metrics that were not included in the original StructureFold. A new module, "reactivity statistics", has been included to derive average reactivity, standard deviation of reactivity, Gini of reactivity, and maximum reactivity of any number of transcripts shared between two or more conditions, enabling the user to build an easy-to-analyze .csv file with appropriately named columns containing these metrics. Identifying transcripts that exhibit a change in any of these metrics between conditions can be resolved almost instantly. For a more detailed comparison between two sets of reactivity values of the same transcripts, modules to batch-calculate RMSD/NRMSD of reactivity or the number of shared reactivity maxima between conditions are also included. Perhaps the most interesting improvement is a module to directly calculate and compare windows of reactivity along a transcript in different conditions. This is useful because RNA often folds on a local scale, or has recognition motifs that are local [21]. This reactivity windows module allows a user to query all transcripts for sliding windows of n nucleotide size, with steps of x nucleotides for change in reactivity between conditions. These results can be filtered by the top gain/loss/net change of reactivity within the windows and saved to another easy-to-use .csv file for numerical analysis, or to a .fasta file logging the sequences of these windows. The latter files allow the user to perform a quick MEME [22] analysis to identify any sequence patterns within these windows, or to check for nucleotide composition biases. The combination of both the reactivity statistics and reactivity windows modules offers a streamlined toolkit to rapidly home in on structural changes associated with reactivity changes between conditions. Looking at changes in predicted structure or reactivity on an entire transcript as well as in small windows allows the process to rapidly converge onto sequences and genes of interest in a minimal amount of time once the initial reactivity scores have been calculated. For example, ligands may force changes in the structures of 5′ UTRs [23,24] and these changes could be discerned by the above approach. A module to bin the reactivity values along a single transcript or bin multiple transcripts together into a single distribution adds another way to check for larger positional trends of reactivity changes between two or more conditions. Another module extracts reactivity changes between two conditions from a set number of base pairs starting from either the 5′ or 3′ end of a single transcript or the sum of these changes from multiple transcripts, reporting either the net sum or average per base pair change. Many known key changes in structure occur on or near the ends of transcripts, thus making such a tool useful for exploratory analysis for either single transcripts or for probing patterns of reactivity change among specific functional classes of genes that may share positional structural regulatory changes.

We present StructureFold2 as a compact yet empowering pipeline, capable of rapidly processing next-generation structural information generated by chemical probing into readily accessible RNA structural data. StructureFold2 offers a sophisticated entrée into the untapped reserve of biological novelty and mechanisms awaiting discovery in the RNA structurome. StructureFold2 can be coupled with a variety of experimental RNA structural methods, including the Structure-seq2 pipeline from our own labs. The new modules contained within StructureFold2 facilitate direct analyses of changes in RNA reactivity, while also identifying changes resolved via structure prediction. Building on an intersection of molecular, statistical, and computational disciplines, StructureFold2 offers a facile platform, allowing users with a minimal computational background to complete these analyses.

Acknowledgments

We thank David Mathews, Joseph Waldron, and Elizabeth Jolley for proofreading the manuscript and offering suggestions. This work was supported by the National Science Foundation Plant Genome Research Program [NSF-IOS-1339282].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ymeth.2018.01.018.

References

[1] C. Ehresmann, F. Baudin, M. Mougel, P. Romby, J.P. Ebel, B. Ehresmann, Probing the structure of RNAs in solution, Nucleic Acids Res. 15 (22) (1987) 9109–9128.

[2] P.C. Bevilacqua, L.E. Ritchey, Z. Su, S.M. Assmann, Genome-wide analysis of RNA secondary structure, Annu. Rev. Genet. 50 (2016) 235–266.

[3] Y. Ding, C.K. Kwok, Y. Tang, P.C. Bevilacqua, S.M. Assmann, Genome-wide profiling of *in vivo* RNA structure at single-nucleotide resolution using structure-seq, Nat. Protoc. 10 (7) (2015) 1050–1066.

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

[5] L.E. Ritchey, Z. Su, Y. Tang, D.C. Tack, S.M. Assmann, P.C. Bevilacqua, Structure-seq2: sensitive and accurate genome-wide profiling of RNA structure *in vivo*, Nucleic Acids Res. 45 (14) (2017) e135.

[6] D. Mitchell 3rd, L.E. Ritchey, H. Park, P. Babitzke, S.M. Assmann, P.C. Bevilacqua, Glyoxals as *in vivo* RNA structural probes of guanine base pairing, RNA 24 (1) (2018) 114–124.

[7] R.C. Spitale, P. Crisalli, R.A. Flynn, E.A. Torre, E.T. Kool, H.Y. Chang, RNA SHAPE analysis in living cells, Nat. Chem. Biol. 9 (1) (2013) 18–20.

[8] E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Cech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Gruning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, J. Goecks, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update, Nucleic Acids Res. 44 (W1) (2016) W3–W10.

[9] M. Zubradt, P. Gupta, S. Persad, A.M. Lambowitz, J.S. Weissman, S. Rouskin, DMS-MaPseq for genome-wide or targeted RNA structure probing *in vivo*, Nat. Meth. 14 (1) (2017) 75–82.

[10] A.N. Sexton, P.Y. Wang, M. Rutenberg-Schoenberg, M.D. Simon, Interpreting reverse transcriptase termination and mutation events for greater insight into the chemical probing of RNA, Biochemistry 56 (35) (2017) 4713–4721.

[11] S. Busan, K.M. Weeks, Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2, RNA (2017).

[12] N.A. Siegfried, S. Busan, G.M. Rice, J.A. Nelson, K.M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP), Nat. Methods 11 (9) (2014) 959–965.

[13] Python Software Foundation. Python Language Reference, version 2.7. http://www.python.org.

[14] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet J. 17 (1) (2011) 10–12.

[15] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (4) (2012) 357–359.

[16] Y. Tang, E. Bouvier, C.K. Kwok, Y. Ding, A. Nekrutenko, P.C. Bevilacqua, S.M. Assmann, StructureFold: genome-wide RNA secondary structure mapping and reconstruction *in vivo*, Bioinformatics 31 (16) (2015) 2668–2675.

[17] K.E. Deigan, T.W. Li, D.H. Mathews, K.M. Weeks, Accurate SHAPE-directed RNA structure determination, Proc. Natl. Acad. Sci. USA 106 (1) (2009) 97–102.

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

[19] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster, Fast folding and comparison of RNA secondary structures, Monatshefte für Chemie/Chem. Monthly 125 (2) (1994) 167–188.

[20] R Development Core Team. R: A language and enviornment for statistical computation, 2008. http://www.R-project.org.

[21] I. Tinoco, C. Bustamante, How RNA folds, J. Mol. Biol. 293 (2) (1999) 271–281.

[22] T.L. Bailey, M. Boden, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, J. Ren, W.W. Li, W.S. Noble, MEME SUITE: tools for motif discovery and searching, Nucleic Acids Res. 37 (Web Server issue) (2009) W202–8.

[23] P.J. McCown, K.A. Corbino, S. Stav, M.E. Sherlock, R.R. Breaker, Riboswitch diversity and distribution, RNA 23 (7) (2017) 995–1011.

[24] A.V. Sherwood, T.M. Henkin, Riboswitch-mediated gene regulation: novel RNA architectures dictate gene expression responses, Annu. Rev. Microbiol. 70 (2016) 361–374.