Bioinformatics, 31(16), 2015, 2668–2675
doi: 10.1093/bioinformatics/btv213
Advance Access Publication Date: 16 April 2015
Original Paper

OXFORD

Structural bioinformatics

StructureFold: genome-wide RNA secondary structure mapping and reconstruction *in vivo*

Yin Tang[1,2,3,*], Emil Bouvier[4,5], Chun Kit Kwok[2,6,†], Yiliang Ding[1,2,6,‡], Anton Nekrutenko[3,4,5], Philip C. Bevilacqua[2,6,7] and Sarah M. Assmann[1,2,3,7,*]

[1]Department of Biology, [2]Center for RNA Molecular Biology, [3]Bioinformatics and Genomics Graduate Program, [4]Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA 16802, USA, [5]Galaxyproject.org, University Park, PA 16802, USA and Baltimore, MD 21218, USA, [6]Department of Chemistry and [7]Plant Biology Graduate Program, Pennsylvania State University, University Park, Pennsylvania 16802, USA

*To whom correspondence should be addressed.
[†]Present address: Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
[‡]Present address: Department of Cell and Developmental Biology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK
Associate Editor: Anna Tramontano

Received on December 23, 2014; revised on March 30, 2015; accepted on April 12, 2015

Abstract

Motivation: RNAs fold into complex structures that are integral to the diverse mechanisms underlying RNA regulation of gene expression. Recent development of transcriptome-wide RNA structure profiling through the application of structure-probing enzymes or chemicals combined with high-throughput sequencing has opened a new field that greatly expands the amount of *in vitro* and *in vivo* RNA structural information available. The resultant datasets provide the opportunity to investigate RNA structural information on a global scale. However, the analysis of high-throughput RNA structure profiling data requires considerable computational effort and expertise.

Results: We present a new platform, StructureFold, that provides an integrated computational solution designed specifically for large-scale RNA structure mapping and reconstruction across any transcriptome. StructureFold automates the processing and analysis of raw high-throughput RNA structure profiling data, allowing the seamless incorporation of wet-bench structural information from chemical probes and/or ribonucleases to restrain RNA secondary structure prediction via the RNAstructure and ViennaRNA package algorithms. StructureFold performs reads mapping and alignment, normalization and reactivity derivation, and RNA structure prediction in a single user-friendly web interface or via local installation. The variation in transcript abundance and length that prevails in living cells and consequently causes variation in the counts of structure-probing events between transcripts is accounted for. Accordingly, StructureFold is applicable to RNA structural profiling data obtained *in vivo* as well as to *in vitro* or *in silico* datasets. StructureFold is deployed via the Galaxy platform.

Availability and Implementation: StructureFold is freely available as a component of Galaxy available at: https://usegalaxy.org/.

Contact: yxt148@psu.edu or sma3@psu.edu

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

1 Introduction

RNA plays vital roles in myriad cellular functions, including regulation of transcription, RNA processing and stability, and translation (Mortimer *et al.*, 2014; Sharp, 2009). The secondary and tertiary structures of RNAs are integral to their biological functions; e.g. as ligand-binding sensors (riboswitches), enzymes (ribozymes) or temperature sensors (RNA thermometers). Therefore, knowledge of RNA structure, especially *in vivo*, is essential for a mechanistic understanding of RNA function.

Understanding of RNA structure has been greatly advanced through the development of computational approaches for RNA secondary and tertiary structure prediction (Schroeder, 2009). Resultant software packages (Table 1) include RNAstructure (Reuter and Mathews, 2010), ViennaRNA package (Lorenz *et al.*, 2011), MC-Fold/MC-Sym (Parisien and Major, 2008) and SeqFold (Ouyang *et al.*, 2013). Each method has its advantages (cf. Table 1), including pseudoknot predictions, local folding, G-quadruplex folding and prediction of 3D structure. Nonetheless, any purely computational RNA structure prediction method has severe limitations due to incomplete knowledge of free energy parameters, RNA interaction with other molecules ranging from ligands to proteins, and kinetically controlled RNA folding. To overcome these limitations, RNA-modifying agents can be used to provide experimental restraints regarding the single- and double-stranded regions of the RNA. For example, critical RNA structural information can be provided using nucleases (Kertesz *et al.*, 2010; Knapp, 1989; Underwood *et al.*, 2010; Zheng *et al.*, 2010) such as RNase V1, which cleaves double-stranded RNA sites, and RNase S1, which cleaves single-stranded RNA sites. However, a limitation of such nuclease-based structure probing is that it can only be applied *in vitro*. RNA structures *in vivo* can be very different due to protein binding, RNA–RNA interactions, ionic and pH conditions, and other aspects of the cellular milieu such as molecular crowding and the influence of metabolites that are impossible to either completely model *in silico* or reconstitute *in vitro* (Kwok *et al.*, 2013; Zaug and Cech, 1995).

As an alternative to nucleases, several RNA modifying chemicals, including dimethyl sulfate (DMS; Zaug and Cech, 1995) and Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) reagents (Wilkinson *et al.*, 2006) can be employed to report the presence of single-stranded nucleotides in RNA and thus provide experimental information on RNA structure. DMS methylates the Watson–Crick N1 position of adenine and N3 position of cytosine when these positions are not engaged in Watson–Crick base pairing or other interactions. SHAPE reagents acylate the 2′-hydroxyl group on the sugar of flexible nucleotides of all four nucleotide types. An advantage of chemical probing methods is that many can be applied *in vivo* as well as *in vitro* (Kwok *et al.*, 2013; Spitale *et al.*, 2013; Wells *et al.*, 2000; Zaug and Cech, 1995).

Nuclease-based cleavage truncates the RNA template available for reverse transcriptase (RT), while the above chemical modifications of the RNA halt the reverse transcriptase one nucleotide before the modification (Ehresmann *et al.*, 1987). Both of these effects can be read out as 'stops' in reverse transcription using conventional methods based on gel (PAGE) or capillary electrophoresis (Fig 1). Such electrophoresis-based methods, however, only read out this structural information for relatively short segments of individual RNAs and so are low-throughput and incapable of providing transcriptome-wide data.

Recently, genome-wide methods employing high-throughput sequencing techniques to identify the RT stops resulting from cleavages or chemical modifications have been developed (e.g. Fig 1). These methods make it possible to obtain structural information on thousands of RNAs in a single experiment. To date, such studies have been conducted *in vitro* in yeast (Kertesz *et al.*, 2010), mouse (Incarnato *et al.*, 2014; Underwood *et al.*, 2010), *Drosophila*, *C. elegans* (Li *et al.*, 2012a), *Arabidopsis* (Zheng *et al.*, 2010) and human (Wan *et al.*, 2014), and *in vivo* in *Arabidopsis* (Ding *et al.*, 2014), yeast (Rouskin *et al.*, 2014; Talkish *et al.*, 2014) and human cells (Rouskin *et al.*, 2014). Another method, SHAPE-MaP, that relies on mutagenesis rather than RT stops, has been applied *in vitro* to the HIV-1 RNA genome (Siegfried *et al.*, 2014). We recently reviewed many of these approaches (Kwok *et al.*, 2015).

While the combination of RNA structure probing with high-throughput sequencing is a major advance in the RNA field, it

Table 1. Comparison of commonly used software for RNA structure prediction

Software	Restraint	Algorithm	Special attributes
RNAstructure	Structural reactivity/ Hard restraint	Thermodynamics	Pseudoknot prediction
ViennaRNA Package	Hard restraint	Thermodynamics	Local folding, G-quadruplex prediction
MC-Fold/MC-Sym	Semi-hard restraint	Nucleotide cyclic motif	3D RNA structure prediction
SeqFold	Structural reactivity	Sample and select	–

RNAstructure (Reuter and Mathews, 2010); ViennaRNA package (Lorenz *et al.*, 2011); MC-Fold/MC-Sym (Parisien and Major, 2008); SeqFold (Ouyang *et al.*, 2013)

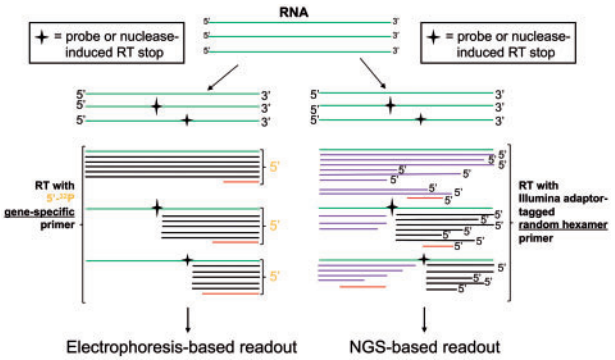

Fig. 1. Electrophoresis-based versus NGS-based RNA structure probing. The left side diagrams the standard electrophoresis-based method, while the right side diagrams the next-generation-sequencing (NGS)-based approach, as exemplified by Structure-Seq. Chemical probe treatment can be *in vivo* or *in vitro*; nuclease treatment is *in vitro*. Chemical probe treatments usually result in a covalent modification of the RNA that terminates reverse transcription (RT). Nuclease treatment results in cleavage of the RNA (not explicitly illustrated). Green lines indicate several copies of a transcript. Gold indicates ^{32}P radiolabel. Black, red, and purple lines indicate cDNA. Red lines indicate RT products resulting from imperfect RT processivity; such RT dropoff products will on average be accounted for by comparison to control data produced without nuclease or chemical probe treatment. In the NGS-based readout, purple lines indicate 5′-runoff RT products; although they do not provide information on RT structure stops, inclusion of these reads (as well as red and black reads) in the Structure-Seq method is critical as it allows proper normalization for transcript abundance

requires intensive computational efforts to derive the explicit predicted reactivities and secondary structures transcriptome-wide from high-throughput RNA structure profiling data. Some efforts have been made to develop algorithms and implement software tools and webservers to facilitate the data analysis process for high-throughput RNA structure profiling (Table 2), which includes reads mapping, normalization, reactivity derivation and RNA structure prediction (Aviran *et al.*, 2011; Li *et al.*, 2012b; Loughrey *et al.*, 2014; Siegfried *et al.*, 2014; Talkish *et al.*, 2014). Here we provide a brief overview of these methods and then indicate the advances provided by StructureFold.

Spats (Aviran *et al.*, 2011; Loughrey *et al.*, 2014) is a method to derive reactivities from Shape-seq raw data (Lucks *et al.*, 2011) based on maximum likelihood, primarily to solve the problem arising from the signal decay that occurs at the 3′ end of each cDNA read when using a single fixed primer that initiates reverse transcription at the 3′ end of the RNA. Mod-seeker (Talkish *et al.*, 2014) is a method that identifies significantly enriched sites of DMS modification by Cochran–Mantel–Haenszel tests on each transcript and defines structural reactivity as the fold enrichment of RT stops on each nucleotide in the DMS treated library relative to a control library with no DMS treatment. While Spats and Mod-seeker provide methods to derive structural reactivities from high-throughput RNA structure profiling data, Spats focus on reactivity information for individual RNAs, while the structural reactivities provided by Mod-seeker are not directly comparable between different RNAs in a transcriptome. In addition, while these programs calculate reactivities, they do not include the essential step of actual prediction of RNA structures.

SAVoR takes RNA sequences and mapped high-throughput RNA structure profiling data as inputs, and outputs individual RNA structures from the RNAfold program in the ViennaRNA package (Lorenz *et al.*, 2011), either including experimental restraints as derived from mapped high-throughput RNA structure profiling data or without any restraints. The online SAVoR platform is easy to use but it cannot be downloaded and installed locally, and it does not provide an option to output transcriptome-wide RNA structural reactivities or structures, which is crucial for deriving global principles of RNA folding and function.

ShapeMapper and the SM_folding_pipeline software developed in the SHAPE-MaP approach (Siegfried *et al.*, 2014) rely on mutagenesis rather than stops during RT. The programs output structural reactivities as well as predicted RNA structures.

Table 2. Comparison of platforms (software) performing analysis on high-throughput RNA structure profiling data (raw sequence)

Platform	Structural reactivity (# transcripts)	RNA structure prediction	Local install	Webserver
Spats	Multiple	No	Yes	No
Mod-seeker	Transcriptome-wide	No	Yes	No
SAVoR	Individual	Yes	No	Yes
ShapeMapper/ SM_folding _pipeline	Multiple	Yes	Yes	No
StructureFold	Transcriptome-wide	Yes	Yes	Yes

Spats (Loughrey *et al.*, 2014); Mod-seeker (Talkish *et al.*, 2014); SaVoR (Li *et al.*, 2012b); ShapeMapper/SM_folding_pipeline (Siegfried *et al.*, 2014); StructureFold (this study).

Here we describe our development of StructureFold, a series of software packages that integrates reads mapping, RT stop count calculation, reactivity derivation, and structure prediction in a web-based interface. The two inputs to StructureFold are transcriptome sequence and (optional) high-throughput RNA structure profiling data; the outputs are structural reactivities on each nucleotide transcriptome-wide and the corresponding RNA structures predicted with or without the experimental restraints. By taking into account the differential abundances and lengths of RNAs in a transcriptome, StructureFold provides values of *in vivo* structural reactivities of nucleotides on different RNAs that are directly comparable. StructureFold is provided as a component of the Galaxy platform (http://www.galaxyproject.org) and takes advantage of the computational tools available in Galaxy (Goecks *et al.*, 2010), such as Bowtie (Langmead *et al.*, 2009) and SAMtools (Li *et al.*, 2009). StructureFold consists of a series of intuitive modules that are easy to use, either online or after local installation.

2 Methods/software implementation

We developed StructureFold as four modules that perform two major functions. The first function employs the first three modules to derive structural reactivities on each nucleotide in a transcriptome based on input of raw sequencing reads. The second function, encompassed in the fourth module, predicts RNA secondary structures transcriptome-wide. StructureFold takes in high-throughput sequencing data from control datasets and from treated datasets consisting of nuclease-based or chemical-based (e.g. DMS or SHAPE) structure probing data and outputs predicted RNA secondary structures in just a few simple steps. Figure 2 provides an overview of how StructureFold works.

StructureFold is implemented as follows:

1. The first module in StructureFold is the **Iterative Mapping** module. This module maps the raw reads from the sequencer to the reference transcriptome library using Bowtie (Langmead *et al.*, 2009), allowing up to the maximum number of mismatches specified by the user. The two inputs are raw sequencing reads and the corresponding reference library (i.e. the transcriptome or RNA genome of interest), provided as Fasta/Fastq and Fasta files, respectively. Bowtie is used to iteratively map the reads to the reference library. Users can use the −5/−3 flag in Bowtie to

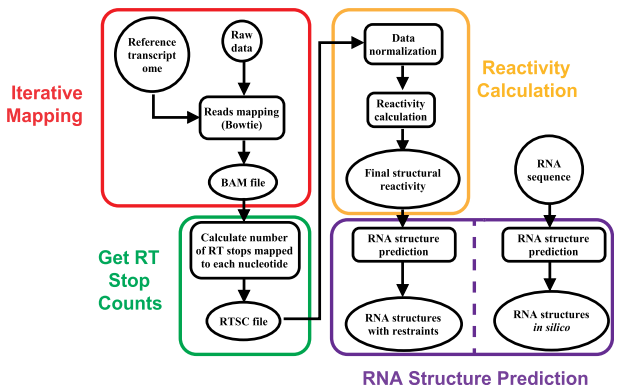

Fig. 2. StructureFold calculates reactivities from experimental RNA structure profiling data and predicts RNA structures with or without these experimental restraints. StructureFold consists of 4 modules: Iterative Mapping, Get RT Stop Counts, Reactivity Calculation, and RNA Structure Prediction. Circles represent inputs, boxes represent processes, and ovals represent outputs. Outputs of modules that are also inputs for the next module are left as ovals

remove a user-specified number of nucleotides at the 5′ and/or 3′ end of each read. This process is useful for removing adapter sequence from the read. Users can also use the –v option to set the flag for the number of mismatches allowed while mapping. If a read cannot be mapped, it may be because of internal mismatches between the read and the reference library sequence, or because adapter sequence is included in the read. Accordingly, if a read cannot be mapped, it is trimmed by a user-specified number of nucleotides at the 3′ or 5′ end, and mapping to the reference library is again attempted. The process is iterated until the read either properly maps or is less than the minimum length requirement specified by the user, which typically corresponds to the minimum length required for unique mapping of a read to the reference library. Each successfully mapped read with its mapping location is collected and the information is stored as a *.bam* file.

2. The second module in StructureFold is the **Get RT Stop Counts** module. This module calculates the number of RT stops that map to each nucleotide of each transcript. The inputs are a mapped dataset (in *.bam* format), typically from the preceding Iterative Mapping module, and the reference library that was used to map the reads, and the output is an RT Stop Counts (RTSC) text file, which contains RT stop counts mapped to each of the four nucleotides (A, C, G, U) in each transcript.

To obtain the number (count) of RT stops for each nucleotide, the mapping location of each RT stop is first considered. For chemical modification, mapping allows identification of the nucleotide immediately 5′ to the RT stop on the RNA, which is the nucleotide that was modified and so receives the stop count. When nucleases are used for structure mapping, the nucleotide that will receive the count according to this procedure is the nucleotide on the 5′ side of the cleavage site on the RNA. The counts of the RT stops of all the nucleotides of each of the RNAs in the experimentally queried transcriptome are combined into a RT termination (stop) count .txt file for each type of library: control and reagent treated. The control library is used to account for background RT stops arising from reverse transcriptase drop off at non-exogenously modified or non-nuclease cleaved sites, which can result from a number of factors including imperfect processivity of the reverse transcriptase, the presence of endogenous RNA modifications, or degradation of the RNA template.

3. The third module of StructureFold is the **Reactivity Calculation** module. This module calculates reactivity based on the RTSC file generated from the preceding Get RT Stop Counts module. The Reactivity Calculation module begins with a two-step normalization of the RT stop counts:

Step 1: The RT stop counts on each nucleotide of each mapped transcript in the plus and minus reagent conditions are incremented by 1 and the natural log (ln) is taken of the resultant value. This provides the numerator in Equations (1) and (2). These raw data are typically skewed, and taking the ln results in a distribution much closer to Gaussian (Fig 3). Incrementing by 1 avoids an undefined value of the natural logarithm if there are zero RT stop counts.

Step 2: The ln of RT stop counts on each nucleotide of each mapped transcript in the plus and minus reagent conditions are normalized by the transcript's abundance and length. For a transcript, suppose $P_r(i)$ and $M_r(i)$ are the raw '*r*' numbers of RT stops mapped to nucleotide *i* (all four nucleotides are included) on the transcript in the plus (P) and minus (M) reagent libraries, respectively, and *l* is the length of the transcript. $P_r(0)$ and $M_r(0)$ are the raw numbers of 5′-runoff RT reads (purple-coded reads in Fig 1) on the transcript in the plus and minus reagent libraries, respectively. The denominators in Equations 1 and 2 thus represent the abundance normalized by transcript length.

Equation (1) provides the normalized RT stop count for nucleotide *i* in the plus reagent library.

$$P(i) = \frac{ln[P_r(i)+1]}{\left(\sum_{i=0}^{l} ln[P_r(i)+1]\right) \Big/ l} \quad (1)$$

Equation (2) provides the normalized RT stop count for nucleotide i in the minus reagent library.

$$M(i) = \frac{ln[M_r(i)+1]}{\left(\sum_{i=0}^{l} ln[M_r(i)]+1\right) \Big/ l} \quad (2)$$

'Final structural reactivities' (FSRs) are then calculated from the normalized RT stop counts in equations 1 and 2 in three steps. First, values of a term denoted 'raw structural reactivity' (RSR) are defined. Raw structural reactivities (RSRs) are calculated by subtracting the normalized RT stop count on each nucleotide in the minus reagent library from the normalized RT stop count on each corresponding nucleotide in the plus reagent library. If the value is negative, it is set to zero. Second, 2–8% normalization is applied to the RSRs. In 2–8% normalization, the top 2% of the data are initially removed and the average value of the remaining top 8% of the data is then calculated (Low and Weeks, 2010). Then all the data (here, all the RSRs) are divided by this average value to obtain the 'normalized structural reactivities' (NSRs). The purpose of this normalization is to convert RSRs into a range that can be used as restraints for RNA structure prediction in RNAstructure and ViennaRNA package. As with SHAPE data (Deigan *et al.*, 2009), this normalization places most reactivities between 0 and 1, with the few reactivities larger than 1 representing highly reactive nucleotides. Third, the NSRs are capped by a user-specified threshold. Capping avoids extremely high structural reactivities, which could introduce bias when investigating meta properties of RNA structural patterns (Kertesz *et al.*, 2010). The resultant values are the 'final structural reactivities' (FSRs) for each nucleotide in the transcriptome.

The output from this module (.txt file), which is indicative of the tendency of a nucleotide to be single-stranded, can be used as input for the next module or can be used directly (i.e. without the next module) to carry out meta analyses. For instance, we used FSRs to evaluate periodicity in mRNA reactivity and relationships between

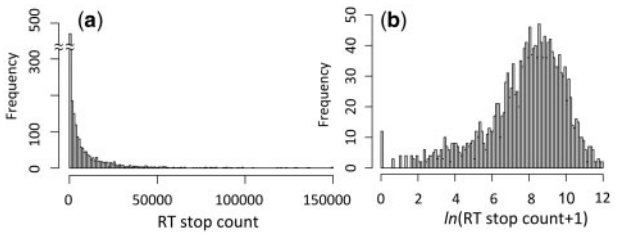

Fig. 3. RT stop count distribution on 18S rRNA in the (+) DMS library from Ding *et al.* (2014). a. Before *ln* transformation. b. After *ln* transformation. The RT stop count distribution is quite skewed before *ln* transformation and is much closer to a Gaussian distribution after *ln* transformation

FSRs and alternative polyadenylation or alternative splicing (Ding *et al.*, 2014).

4. The fourth module of StructureFold is the **RNA Structure Prediction** module. This module predicts RNA structures using the RNAstructure program (V5.6; http://rna.urmc.rochester.edu/RNAstructure.html) and ViennaRNA package (V2.1.9). The module can predict RNA secondary structures from sequence data alone, or with inclusion of restraints from structural reactivities as provided by the first three modules. Lastly, the user has the option of creating bar plots to illustrate the distribution of the FSRs of the nucleotides in any transcript specified by the user (Fig 4a).

The RNA Structure Prediction module requires one or more transcript IDs from the reference library as input; these are provided as a .txt file with each ID on one line. StructureFold will find the RNA sequence in the reference library corresponding to the ID provided. If the FSR on each nucleotide of the RNA is provided, the module can output bar plots that illustrate the distribution of the FSRs (generated from the third module) of the nucleotides in any transcript specified by the user (Fig 4a). When the FSRs are provided, the module outputs the predicted structures of the RNA with these restraints (Fig 4b). Otherwise the module outputs the predicted structures without restraints, i.e. an *in silico* prediction that does not incorporate any experimental data (Fig 4c).

Explicit inputs and outputs for all four modules in StructureFold can be found in the Supplementary Note.

2.1 ROC curve generation

ROC curves were generated on regions of 18S and 25S rRNAs for both gel-based and NGS-based readouts. These curves are based on phylogenetic secondary structures, derived by the Gutell lab based on comparative analyses of RNA sequences and their secondary structures (Cannone *et al.*, 2002; http://www.rna.icmb.utexas.edu). We define a single-stranded nucleotide in the phylogenetic structure as condition positive. We vary a reactivity threshold from 0 to 7 evenly with 15 000 divisions. For a given threshold, each nucleotide with FSR (or normalized gel intensity) that is above the threshold is designated as test outcome positive. We calculate sensitivity (true positive rate) and 1—specificity (false positive rate) at each threshold and use this information to generate the ROC curves.

2.2 Gel-based structure probing

In vivo DMS modification was performed on etiolated *Arabidopsis* seedlings, and the structure probing results were analyzed by gel electrophoresis (Fig 1) as described previously (Ding *et al.*, 2014; Kwok *et al.*, 2013). Briefly, the DMS-treated and DMS-untreated (control) RNA were used as the inputs for rRNA structure probing. Approximately 1 μg of total RNA was used for the (+) and the (−) DMS conditions and total RNA was used for dideoxy sequencing. ^{32}P radiolabelled 18S or 25S rRNA-specific primers were used for the reverse transcription: for 18S region one (5′-AACTGATTTAATGAGCCATTCGCAG-3′), for 18S region two (5′-GAGCCCGCGTCGACCTTTTATC-3′), for 18S region three (5′-GGTAATTTGCGCGCCTGCT-3′), for 25S region one (5′-AAGCGCCATCCATTTTCGG-3′). The labeled cDNA fragments were size fractionated by PAGE and detected using a Typhoon phosphorimager. Gel intensity was quantified using ImageQuant 5.2. Gel-based DMS reactivity was normalized (Kwok *et al.*, 2013) by 2–8% normalization based on Low and Weeks (2010).

3 Results

3.1 Comparison between structural reactivities and gel-based assay intensities

We used ROC curves to compare the similarity of structural reactivities generated by the NGS-based Structure-Seq method (Ding *et al.*, 2014) analyzed via StructureFold to DMS reactivities derived from the classical gel-based method of RNA structure-probing (Fig 5). The Structure-Seq FSRs and normalized *in vivo* gel-based assay intensities have similar performance on all of the rRNA regions tested, which strongly validates the DMS reactivities generated from StructureFold. The observation that the ROC curves are above the diagonal line of no discrimination indicates that the methods perform much better than random prediction, and the observation that the curves track each other on each of the plots indicates that the two methods have similar performance.

3.2 Comparison between structural reactivities derived from StructureFold and Mod-seeker

To compare our method of deriving structural reactivities with the previously published method, Mod-seeker, we applied the Mod-seeker approach to our high-throughput structural profiling data on rRNA from Ding *et al.* (2014). For Arabidopsis 18S and 25S rRNAs, we observed a high correlation between the structural reactivities derived using StructureFold and those from Mod-seeker [Pearson correlation coefficient (PCC) = 0.91 and PCC = 0.93, respectively]. In addition, as shown in Table 3, the structural reactivities derived from both programs are consistent with the phylogenetic structures of 18S and 25S rRNA.

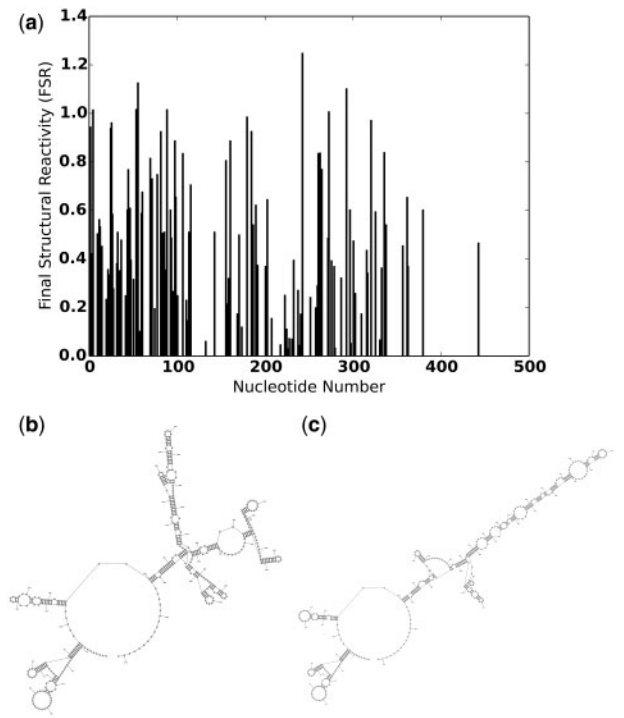

Fig. 4. Sample outputs of StructureFold. a. Final structural reactivity (FSR) distribution of *AT5G53300.3* (Ubiquitin Conjugating Enzyme 10) used for prediction in panel b. Predicted RNA structure of *AT5G53300.3* with *in vivo* restraints derived from Ding *et al.* (2014). c. Predicted RNA structure of *AT5G53300.3* (471 nt) *in silico* (unrestrained)

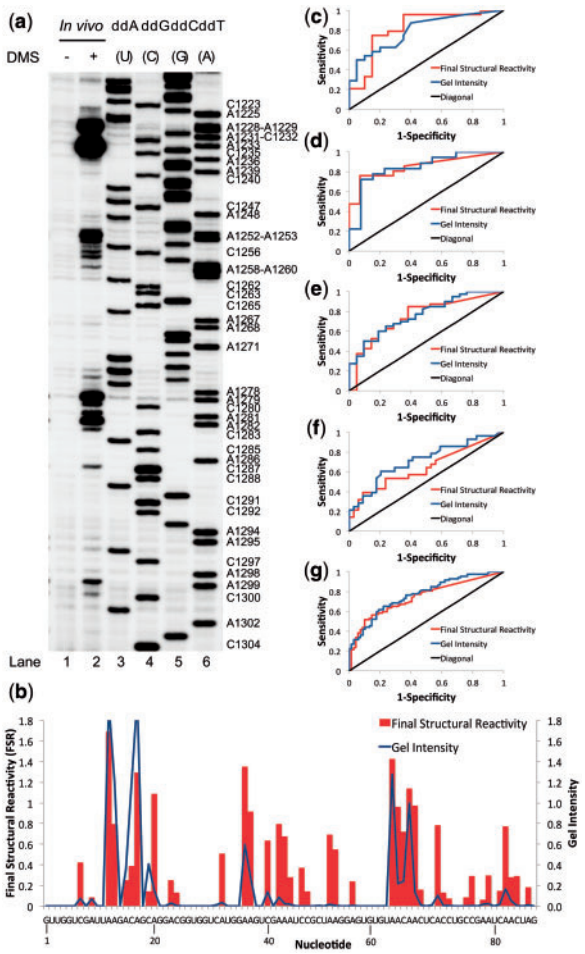

Fig. 5. Comparison between Final Structural Reactivities (FSRs) derived by StructureFold (Ding *et al.*, 2014) versus gel-based assay in which reactivity is read out as band intensity. a. *In vivo* DMS treatment was performed on 1 μg of total RNA and analyzed using ^{32}P radiolabelled 25S rRNA-specific primer (lane 2). A control with no DMS treatment was performed in parallel (lane 1). Dideoxy sequencing was conducted using 1 μg of total RNA (lanes 3–6). The region of 25S rRNA analyzed is nucleotides 1217–1303. b. FSRs derived by StructureFold (red) compared to normalized gel-based assay intensities (blue) on 25S rRNA for nucleotides 1217–1303. $R = 0.69$. c. ROC curve based on FSRs derived by StructureFold versus gel-based reactivities on 25S rRNA (nucleotides 1217–1303). d–g. ROC curves based on FSRs derived by StructureFold versus gel-based reactivities on 18S rRNA: nucleotides 17–86 (d), 87–207 (e), 298–428 (f), panels d–f combined (g)

3.3 Sample outputs from StructureFold

In addition to providing FSRs, StructureFold can provide RNA structures in the following two ways:

3.3.1 Predict RNA structures in silico

The RNA Structure Prediction module can predict RNA structures at a user-chosen temperature without the imposition of any restraints. In this case, StructureFold employs RNAstructure (Reuter and Mathews, 2010) or ViennaRNA package (Lorenz *et al.*, 2011) with default thermodynamic parameters to predict RNA structures directly from the RNA sequence according to thermodynamics. Candidate structures selected are those with minimum free energy.

3.3.2 Predict RNA structures with experimental restraints

StructureFold can take as input experimentally-generated high-throughput RNA structure profiling data to restrain structure prediction. By sequentially employing the preceding Iterative Mapping, Get RT Stop Counts, and Reactivity modules, StructureFold will derive the FSR reactivity for each nucleotide, which is a quantification of the likelihood that the nucleotide is single-stranded.

Table 3. Comparison between the structural reactivities derived on Arabidopsis rRNAs using StructureFold or Mod-seeker

(a) 18S rRNA

	Structural reactivity from StructureFold		
	High	High + medium	Low (%)
Single-stranded	86.7% (96.7%)	75% (92.2%)	48
Base-paired	13.3% (3.3%)	25% (7.8%)	52

	Structural reactivity from Mod-seeker	
	Significant	Non-significant (%)
Single-stranded	76.8% (93.1%)	48.7
Base-paired	23.2% (6.9%)	51.3

(b) 25S rRNA

	Structural reactivity from StructureFold		
	High	High + medium	Low (%)
Single-stranded	78.6% (93.1%)	70.0% (91.9%)	44.2
Base-paired	21.4% (6.9%)	30.0% (8.1%)	55.8

	Structural reactivity from Mod-seeker	
	Significant	Non-significant (%)
Single-stranded	70.0% (91.8%)	45.7
Base-paired	30.0% (8.2%)	54.3

Number in each cell is the percentage of the nucleotides with the specified level of reactivity that are single-stranded or base-paired in the phylogenetic structure. Number in parentheses is the percentage of the nucleotides with the specified level of reactivity that are single-stranded or positioned either at the end of a helix or adjacent to a helical defect in the phylogenetic structure. Here, types of reactivities are defined as follows: **High:** reactivity ≥ 0.6; **Medium:** $0.6 >$ reactivity ≥ 0.3; **Low:** $0.3 >$ reactivity; **Significant:** Reactivity of significantly enriched sites of DMS modification identified by Mod-seeker; **Non-significant:** Reactivity of the sites that are not significantly enriched in DMS modification

The RNA Structure Prediction module in StructureFold then predicts RNA structures using these reactivities as restraints. The method converts reactivities into pseudo-free energy terms, or 'soft restraints' that are used to adjust the purely model oligonucleotide based thermodynamic free energies (Turner and Mathews, 2010; Xia *et al.*, 1998) of the candidate structures. Reactivities are used as restraints in an analogous way as previously described for DMS and SHAPE data (Cordero *et al.*, 2012; Deigan *et al.*, 2009). After our own assessment (Supplementary Table S1), we opted to use the same slope and intercept parameters for StructureFold as those used by Hajdin *et al.* (2013). The corresponding equation is:

$$\Delta G(i) = 1.8ln[\text{Reactivity}(i) + 1] - 0.6 \quad (\text{in kcal mol}^{-1}) \qquad (3)$$

where '*i*' is the *i*th nucleotide in a given transcript.

StructureFold outputs include the RNA structure predicted with and without restraints and a bar plot that illustrates the per nucleotide reactivity of the RNA (Fig 4). It is evident from Figure 4 that this RNA structure predicted with restraints is quite different from that predicted without restraints, an observation that is generally

true (Ding *et al.*, 2014) and highlights the importance of experimentally-derived restraints.

4 Discussion

StructureFold is the first computational platform designed specifically for transcriptome-wide RNA structure mapping and reconstruction. In addition to enabling transcriptome-wide prediction of RNA secondary structures *in silico* (unrestrained), StructureFold can output FSRs that can then be used for transcriptome-wide prediction of experimentally-restrained RNA structures. In the future, we anticipate StructureFold outputs being used to assess and compare the restrained predictions resulting from RNA structure profiling data generated by numerous experimental methods, including nuclease, SHAPE, DMS, and other chemical probing agents (Kwok *et al.*, 2015).

StructureFold has several advantages over other software platforms for analysis of RNA structure (Table 2). Compared to SAVoR (Li *et al.*, 2012b), StructureFold outputs transcriptome-wide structural reactivities on each nucleotide, allowing investigation of global structural patterns. Compared to Spats (Aviran *et al.*, 2011) and Mod-seeker (Talkish *et al.*, 2014), StructureFold predicts RNA structures in addition to deriving structural reactivities. In contrast to Mod-seeker, StructureFold takes the abundance of transcripts into consideration and thus can be used for comparative analyses of *in vivo* as well as *in vitro* high throughput RNA structure data within and between transcriptomes. Compared to ShapeMapper, StructureFold provides a user-friendly online platform where high-throughput data probing RNA structures can be analyzed directly.

StructureFold is straightforward to use online or to install for local use on UNIX systems. Through incorporation into the Galaxy platform, StructureFold takes advantage of the computational infrastructure and tools, as well as the ongoing curation provided by Galaxy (Goecks *et al.*, 2010). For example, outputs from read mapping algorithms in Galaxy such as Bowtie (Langmead *et al.*, 2009) and BWA (Li and Durbin, 2009) directly provide the input to the Get RT Stop Counts module of StructureFold.

In the future, StructureFold could be extended in several ways:

(i) *Methods for reads trimming*

StructureFold employs iterative mapping to remove adapter sequences; this method uses the exact biological sequence to identify the presence of adapter sequence but requires more computational time. The *cutadapt* program (Martin, 2011) is commonly used to remove adapter sequence but has difficulty to identify the adapter sequence if only a small portion of the adapter is present. Iterative mapping and *cutadapt* could possibly be integrated.

(ii) *More options to derive reactivities from RT stop count files*

Currently, three experimental approaches have been applied for *in vivo* transcriptome-wide profiling of RNA structure: Structure-Seq (Ding *et al.*, 2014), Mod-seq (Talkish *et al.*, 2014) and DMS-seq (Rouskin *et al.*, 2014). Each approach has its advantages and limitations. For example, StructureFold takes the transcript abundance into account and avoids skewness of the raw number of RT stops by ln transformation but it is still susceptible to experimental noise to some extent, especially on low-abundance RNAs. RNAs with low abundance have fewer RT stops mapped in total, and so are proportionately more affected by stops resulting from drop off of the RT primer. One way to compensate for this is simply to obtain more reads. Mod-seeker (Talkish *et al.*, 2014) is more resistant to experimental noise than StructureFold because it identifies significantly enriched sites of DMS modification before deriving the final DMS reactivity. Mod-seeker and StructureFold have very similar performance on deriving structural reactivities of individual rRNAs as shown in Table 3; however, DMS reactivities derived by Mod-seeker on different transcripts are not directly comparable because each nucleotide in a library is normalized by the same value, which does not take into account differences in transcript abundance. DMS-Seq (Rouskin *et al.*, 2014) normalizes RT stops mapped to each position proportionally to the most highly reactive base within a given structured window (50–200 nucleotides in length). The method avoids local bias of the RT stops on each RNA, but the length of the window is fairly arbitrary (Aviran and Pachter, 2014). Moreover, local region normalization assumes independence of the different regions of the transcript, which is not always a valid assumption (Behrouzi *et al.*, 2012; Strulson *et al.*, 2014). In the future, StructureFold could be expanded to provide options for users to choose among methods of reactivity calculation.

(iii) *More alternatives for predicting RNA structures*

StructureFold employs the free energy minimization program in RNAstructure (Reuter and Mathews, 2010) or the ViennaRNA package (Lorenz *et al.*, 2011) to predict RNA structures. Both RNAstructure and ViennaRNA predict RNA structures according to thermodynamic parameters (Turner Rules) measured on model oligonucleotides and provide the candidate structures with the lowest free energy. RNAstructure is able to use the experimentally-derived reactivities as soft restraints, which improves the accuracy of the prediction. By contrast, the ViennaRNA package incorporates the experimental data only as hard restraints for RNA structure prediction. Hard restraints are digital (0/1) restraints that designate each nucleotide to be base-paired or non-base paired, which may result in less accurate structure prediction than the use of soft restraints. The ViennaRNA package uniquely allows local folding (RNALfold) and G-quadruplex prediction, which are advantageous for some applications.

MC-Fold/MC-Sym (Parisien and Major, 2008), another commonly used RNA structure prediction package, can also use restraints for structure prediction, wherein the user must designate the reactivity for each nucleotide as high, low or medium. MC-Fold uses a knowledge-based scoring function rather than thermodynamics for 2D RNA structure prediction, and MC-Sym is further able to predict the 3D structures of the RNAs based off the 2D MC-Fold structures. In the future, incorporation of MC-Fold and MC-Sym algorithms as choices in StructureFold could provide additional information on individual RNA structures.

We believe that StructureFold is well-suited for use by any scientist interested in generating and analyzing high-throughput RNA structure profiling data. Given the current rapid development of the RNA structure field, the user-friendly nature of the StructureFold platform, and the plentiful computational resources available on the Galaxy platform, StructureFold will facilitate genome-wide RNA structure prediction by both experimentally-focused and computationally-focused groups.

Acknowledgements

We thank the Penn State Galaxy team for advice on software development, Prof. Dave Mathews for advice on the manuscript, Dr. Alex Spasic and Mr. Lin An for testing StructureFold implemented in Galaxy, and Prof. Joel McManus for assistance with Mod-seeker comparisons.

Funding

This work was supported by National Science Foundation-IOS-1339282 with initial support from the Penn State Huck Institutes of the Life Sciences and Human Frontiers in Science Program grant RGP0002/2009-C to PCB and SMA.

Conflict of Interest: none declared.

References

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

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

Behrouzi,R. *et al.* (2012) Cooperative tertiary interaction network guides RNA folding. *Cell*, **149**, 348–357.

Cannone,J.J. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.

Cordero,P. *et al.* (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, **51**, 7037–7039.

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

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

Ehresmann,C. *et al.* (1987) Probing the structure of RNAs in solution. *Nucleic Acids Res.*, **15**, 9109–9128.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Hajdin,C.E. *et al.* (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. USA*, **110**, 5498–5503.

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.

Knapp,G. (1989) Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods Enzymol.*, **180**, 192–212.

Kwok,C.K. *et al.* (2013) Determination of *in vivo* RNA structure in low-abundance transcripts. *Nat. Commun.*, **4**, 2971.

Kwok,C.K. *et al.* (2015) The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem. Sci.*, **40**, 221–232.

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

Li,F. *et al.* (2012b) SAVoR: a server for sequencing annotation and visualization of RNA structures. *Nucleic Acids Res.*, **40**, W59–W64.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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

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.

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. USA*, **108**, 11063–11068.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.

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

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

Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.

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.

Schroeder,S.J. (2009) Advances in RNA structure prediction from sequence: new tools for generating hypotheses about viral RNA structure-function relationships. *J. Virol.*, **83**, 6326–6334.

Sharp,P.A. (2009) The centrality of RNA. *Cell*, **136**, 577–580.

Siegfried,N.A. *et al.* (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*, **11**, 959–965.

Spitale,R.C. *et al.* (2013) RNA SHAPE analysis in living cells. *Nat. Chem. Biol.*, **9**, 18–20.

Strulson,C.A. *et al.* (2014) Molecular crowders and cosolutes promote folding cooperativity of RNA under physiological ionic conditions. *RNA*, **20**, 331–347.

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

Turner,D.H. and Mathews,D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38**, D280–D282.

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

Wells,S.E. *et al.* (2000) Use of dimethyl sulfate to probe RNA structure *in vivo*. *Methods Enzymol.*, **318**, 479–493.

Wilkinson,K.A. *et al.* (2006) Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*, **1**, 1610–1616.

Xia,T. *et al.* (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*, **37**, 14719–14735.

Zaug,A.J. and Cech,T.R. (1995) Analysis of the structure of Tetrahymena nuclear RNAs *in vivo*: telomerase RNA, the self-splicing rRNA intron, and U2 snRNA. *RNA*, **1**, 363–374.

Zheng,Q. *et al.* (2010) Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.*, **6**, e1001141.