BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 22 2012, pages 3006–3008
doi:10.1093/bioinformatics/bts554

Databases and ontologies
Advance Access publication September 12, 2012

An RNA Mapping DataBase for curating RNA structure mapping experiments

Pablo Cordero[1], Julius B. Lucks[2] and Rhiju Das[1,3,4,*]

[1]Department of Biochemistry and Biomedical Informatics Program, Stanford University, Stanford, CA 94305, [2]School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853, [3]Department of Biochemistry and [4]Department of Physics, Stanford University, Stanford CA 94305, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: We have established an RNA mapping database (RMDB) to enable structural, thermodynamic and kinetic comparisons across single-nucleotide-resolution RNA structure mapping experiments. The volume of structure mapping data has greatly increased since the development of high-throughput sequencing techniques, accelerated software pipelines and large-scale mutagenesis. For scientists wishing to infer relationships between RNA sequence/structure and these mapping data, there is a need for a database that is curated, tagged with error estimates and interfaced with tools for sharing, visualization, search and meta-analysis. Through its on-line front-end, the RMDB allows users to explore single-nucleotide-resolution mapping data in heat-map, bar-graph and colored secondary structure graphics; to leverage these data to generate secondary structure hypotheses; and to download the data in standardized and computer-friendly files, including the RDAT and community-consensus SNRNASM formats. At the time of writing, the database houses 53 entries, describing more than 2848 experiments of 1098 RNA constructs in several solution conditions and is growing rapidly.

Availability: Freely available on the web at http://rmdb.stanford.edu

Contact: rhiju@stanford.edu

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

Received on January 11, 2012; revised on August 17, 2012; accepted on September 5, 2012

1 INTRODUCTION

Understanding the secondary and tertiary structures of RNAs is critical for dissecting their diverse biological functions, ranging from catalysis in ribosomal RNAs to gene regulation in metabolite-sensing riboswitches and protein-binding elements in RNA messages (Eickbush and Eickbush, 2007; Kudla *et al*., 2009; Nudler and Mironov, 2004; Spahn *et al*., 2001; Yanofsky, 2004). RNA structure has therefore been intensely studied with a variety of biophysical and biochemical technologies (Adilakshmi *et al*., 2006; Getz *et al*., 2007; Waldsich, 2008; Varani and Tinoco, 1991; Wilkinson *et al*., 2006). Among these tools, a facile, information-rich and widely used technique is structure mapping (also called structure probing or footprinting), in which the chemical modification, enzymatic cleavage or degradation rate of an RNA nucleotide correlates with the exposure, flexibility or other structural features of the site. Modern methods often reverse transcribe probed RNA molecules into DNA fragments whose lengths can be subsequently analyzed to infer the locations of probe events. These methods permit the single-nucleotide resolution readout of structural data for RNAs as large as ribosomes (Deigan *et al*., 2009; Culver *et al*., 1999), and in recent years, investigators have developed high-throughput technologies, such as 96-well capillary electrophoresis (Mitra and Shcherbakova, 2008; Mortimer and Weeks, 2007) and deep sequencing (Lucks *et al*., 2011) to perform this step. Furthermore, several bioinformatic pipelines have been implemented to rapidly quantify, map and analyze the resulting data (Aviran *et al*., 2011; Deigan *et al*., 2009; Low and Weeks 2010; Vasa *et al*., 2008; Yoon *et al*., 2011). RNA mapping experiments are now routinely used to improve automated secondary structure modeling (Deigan *et al*., 2009), probe entire viral genomes (Watts and Dang, 2009), simultaneously map arbitrary RNA mixtures through deep sequencing (Kertesz *et al*., 2010; Lucks *et al*., 2011; Underwood *et al*., 2010; Zheng *et al*., 2010) and infer an RNA's 'contact map' by coupling to exhaustive single-nucleotide mutagenesis (Kladwang and Das, 2010; Kladwang *et al*., 2011).

These developments could enable novel methods in RNA structural biology, especially if predictive relationships between RNA sequence/structure and these data can be established. However, unlike existing structural biology fields like nuclear magnetic resonance and crystallography, there is no equivalent of the Biological Magnetic Resonance Bank (Ulrich *et al*., 2007) or the Protein Data Bank (Bernstein *et al*., 1977) that stores curated datasets. Structure mapping data are available in the supporting material of papers or self-reported in SNRNASM format (Rocca-Serra *et al*., 2011), but these formats do not typically include error estimates; are not always normalized or background-subtracted with standardized protocols; are not linked to RNA structures and are not straightforward to visualize, which would enable consistency checks during further analysis. We have therefore created an RNA mapping database (RMDB) and are populating it with curated structure mapping measurements in human and machine-readable formats amenable to inferring relationships between sequence/structure and structure mapping data. Data contained in the RMDB are freely available and can be easily integrated with future repositories such as RNAcentral (Bateman *et al*., 2011).

[*]To whom correspondence should be addressed.

2 DATABASE CONTENT AND STRUCTURE

For a specific RNA in defined solution conditions, each structure mapping experiment can be conceptualized as $M \times N$ matrices, where M is the number of measurements made on the RNA, e.g. normalized peak areas calculated by HiTRACE (Yoon *et al.*, 2011), CAFA (Mitra *et al.*, 2008) and ShapeFinder (Vasa *et al.*, 2008); or maximum likelihood parameters (Aviran *et al.*, 2011) and N is the number of nucleotides in the RNA. Entries in the RMDB house these data matrices and are enriched with annotations and free text to describe associated content.

The database currently includes experiments using base methylation by dimethyl sulfate, base adduct formation by 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-*p*-toluene sulfonate, selective 2′ hydroxyl acylation with primer extension (SHAPE) with either *N*-methylisatoic anhydride or 1-methyl-7-nitroisatoic anhydride and hydroxyl radical footprinting. The RMDB is further capable of storing data from enzymatic, in-line and other structure mapping experiments. The RNAs probed include riboswitches, tRNAs, ribozyme and ribosomal domains featured in several published studies as well as human-designed sequences accruing in the internet-scale RNA engineering project EteRNA (http://eterna.stanford.edu) (see Supplementary Table S2). For many RMDB entries, the complementary DNA fragment separation and analysis steps were carried out by 96-well format capillary electrophoresis and the HiTRACE pipeline (Yoon *et al.*, 2011), respectively; entries describing data from (Lucks *et al.*, 2011) were read out and processed using the SHAPE-Seq protocol (Aviran *et al.*, 2011).

New experimental data can be uploaded to the RMDB as a spreadsheet in the SNRNASM Isa-Tab format (Rocca-Serra *et al.*, 2011) or in RDAT file format (a simple text file format; detailed in the Supplement Information and at http://rmdb.stanford.edu/repository/specs/). However, like the well-curated Protein DataBank, public release then requires passing review by the RMDB team; the data must include error estimates or replicates, information on estimated or known structure (at least at the level of secondary structure), associated publications or preprints and descriptions of how the data were processed.

3 FEATURES AND EXAMPLE USE CASES

3.1 For experimentalists

The RMDB is a resource for RNA biochemists and molecular biologists interested in using existing data to guide biological hypotheses, interpret new data or to share their own experimental results. First, users interested in a particular RNA system can quickly find relevant data in the RMDB by using the full-text search field in the upper-right corner of the site. The user can then inspect each entry using the data visualization tools (Fig. 1a and b; Supplementary Material). Second, the integration with the VARNA applet allows for quickly comparing mapping data against structural models. The data can be downloaded in either RDAT or SNRNASM format or exported directly from the VARNA visualization applet for further inspection with other tools. Third, the RMDB also includes a secondary structure prediction server (located at http://rmdb.stanford.edu/structureserver); the server can use structure mapping data to generate sensible secondary structure hypotheses (see Supplementary Material). Finally, experimentalists who wish to submit their data to the RMDB can do so after registering to the site.

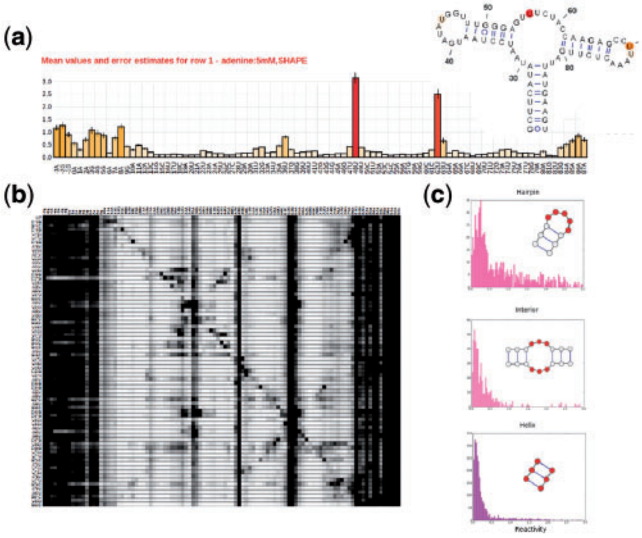

Fig. 1. Different visualization tools for entries in the RNA mapping database. (a) Classic bar plot of 2′-OH acylation (SHAPE) rates across the nucleotides of the adenine-sensing domain of the *add* riboswitch from *Vibrio vulnificus*. The data are from a 'standard state' study averaging 19 replicates across multiple experiments and estimating the resulting errors (shown as error bars); the RNA's crystallographic secondary structure, colored by the SHAPE data, is shown in the inset. (b) Through mutate-and-map data, the RMDB also allows exploring the contact map of the same riboswitch. SHAPE data are shown for constructs with single mutations at each RNA position. (c) Histograms for reactivities found in interior loops, hairpin loops and helices. Nucleotides in the motif for which reactivity data were collected are marked in red. Hairpin loops have higher average reactivities than interior loops, bulges and non-helical elements

3.2 For structural bioinformaticists

The RMDB can extract general properties of mapping data including histograms of reactivities for different secondary structure elements (Fig. 1c). Analyses of structure mapping data are facilitated by the Python/MATLAB RDATkit package (http://rdatkit.simtk.org, see Supplementary Material) for RDAT/SNRNASM-IsaTAB parsing. To demonstrate the utility of the RMDB in extracting new information from multiple datasets, we tested whether the SHAPE method can discriminate between interior and hairpin loops. Using the advanced search feature of the database, we downloaded SHAPE data for each secondary structure element (internal loops, hairpins, helices and bulges) collected in standard state experiments for all non-coding RNAs (ncRNAs) with known structure in the database. Interior and hairpin loops of ncRNA have distinct reactivity distributions, suggesting that SHAPE-directed modeling can be made more accurate by taking this effect into account (Fig. 1c).

3.3 For web-app developers

Data stored in the RMDB are exposed through a RESTful API (described in https://sites.google.com/site/rmdbwiki/web-api) in JSON format, simplifying the creation of web applications that use the data contained in the repository. The RMDB also provides RSS feeds that are automatically updated with new entries (see https://sites.google.com/site/rmdbwiki/rss for details). These tools have allowed integration of the entries in the RMDB into the SNRNASM repository (http://snrnasm.bio.unc.edu/browse.html).

4 DISCUSSION

The throughput of structure mapping experiments has taken significant leaps with multiplexed capillary electrophoresis and next-generation sequencing that allow probing of thousands of RNAs at once. These data should, in principle, permit the development of confident structural biology tools that couple structure mapping measurements to secondary and tertiary structure modeling. However, until recently, researchers have had few resources that enable curation and sharing of high-throughput quantified RNA mapping data. It is our hope that the RMDB will make such projects possible.

ACKNOWLEDGEMENTS

We thank the authors of RNAstructure, VARNA and the Stanford Visualization Team for making their source code freely available, the Laederach laboratory for aid with SNRNASM integration, the Clote laboratory for data presentation comments and suggestions, R. Astorga for design comments and the members of the Das laboratory and C.C. VanLang for help with data curation and manuscript preparation. The server that houses the RMDB was graciously donated by K. Beauchamp.

Funding: The Burroughs-Wellcome Foundation (CASI to R.D.), a Hewlett-Packard Stanford Graduate Fellowship (to P.C.) and the National Institutes of Health (R01 GM102519 to R.D.).

Conflict of Interest: none declared.

REFERENCES

Adilakshmi,T. *et al.* (2006) Hydroxyl radical footprinting *in vivo*: mapping macromolecular structures with synchrotron radiation. *Nucleic Acids Res.*, **34**, e6.

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

Bateman,A. *et al.* (2011) RNAcentral: a vision for an international database of RNA sequences. *RNA*, **17**, 1941–1946.

Bernstein,F.C. *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Culver,G. *et al.* (1999) Identification of an RNA–protein bridge spanning the ribosomal subunit interface. *Science*, **285**, 2133–2135.

Darty,K. *et al.* (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Das,R. *et al.* (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.

Deigan,K.E. *et al.* (2009) Accurate SHAPE directed RNA structure determination. *Proc. Natl Acad. Sci.*, **106**, 97–100.

Eickbush,T.H. and Eickbush,D. (2007) Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics*, **175**, 477–485.

Getz,M. *et al.* (2007) Review NMR studies of RNA dynamics and structural plasticity using NMR residual dipolar couplings. *Biopolymers*, **86**, 384–402.

Kertesz,M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **267**, 103–107.

Kladwang,W. *et al.* (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA*, **17**, 522–534.

Kladwang,W. and Das,R. (2010) A mutate-and-map strategy for inferring base pairs in structured nucleic acids: proof of concept on a DNA/RNA helix. *Biochemistry*, **49**, 7414–7416.

Kladwang,W. *et al.* (2011) Two-dimensional chemical mapping of non-coding RNAs. *Nat. Chem.*, **3**, 954–962.

Kladwang,W. *et al.* (2011) Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry*, **50**, 8049–8056.

Kudla,G. *et al.* (2009) Coding-sequence determinants of gene expression in *Escherichia coli. Science*, **324**, 255–258.

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.*, **108**, 11063–11068.

Mathews,D. and Turner,D. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**, 270–278.

Mitra,S. *et al.* (2008) High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Res*, **36**, e63.

Mortimer,S.A. and Weeks,K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE Chemistry. *J. Am. Chem. Soc.*, **129**, 4144–4145.

Nudler,E. and Mironov,A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, **29**, 11–17.

Rocca-Serra,P. *et al.* (2011) Sharing and archiving nucleic acid structure mapping data. *RNA*, **17**, 1204–1212.

Spahn,C.M.T. *et al.* (2001) Hepatitis C virus IRES RNA-induced changes in the conformation of the 40S ribosomal subunit. *Science*, **291**, 1959–1962.

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

Ulrich,E.L. *et al.* (2007) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.

Varani,G. and Tinoco,I. (1991) RNA structure and NMR spectroscopy. *Quart. Rev. Biophys.*, **24**, 479–532.

Vasa,S.M. *et al.* (2008) ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA*, **14**, 1979–1990.

Waldsich,C. (2008) Dissecting RNA folding by nucleotide analog interference mapping (NAIM). *Nat. Protoc.*, **3**, 811–823.

Watts,J.M. *et al.* (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.

Wilkinson,K.A. *et al.* (2006) Selective 29-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*, **1**, 1610–1616.

Yanofsky,C. (2004) The different roles of tryptophan transfer RNA in regulating trp operon expression in E. coli versus B. subtilis. *Trends Genet.*, **20**, 367–374.

Yoon,S. *et al.* (2011) HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics*, **27**, 1798–805.

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.