Methods xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

The RNA 3D Motif Atlas: Computational methods for extraction, organization and evaluation of RNA motifs

Lorena G. Parlea [a], Blake A. Sweeney [a], Maryam Hosseini-Asanjan [c], Craig L. Zirbel [b], Neocles B. Leontis [c,*]

[a] Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403, USA
[b] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA
[c] Department of Chemistry, Bowling Green State University, Bowling Green, OH 43403, USA

ARTICLE INFO

Article history:
Received 14 March 2016
Received in revised form 21 April 2016
Accepted 22 April 2016
Available online xxxx

Keywords:
Structured RNA molecules
Hairpin loop
Internal loop
Multi-helix junction loop
Non-Watson-Crick basepair
RNA 3D Motif

ABSTRACT

RNA 3D motifs occupy places in structured RNA molecules that correspond to the hairpin, internal and multi-helix junction "loops" of their secondary structure representations. As many as 40% of the nucleotides of an RNA molecule can belong to these structural elements, which are distinct from the regular double helical regions formed by contiguous AU, GC, and GU Watson-Crick basepairs. With the large number of atomic- or near atomic-resolution 3D structures appearing in a steady stream in the PDB/NDB structure databases, the automated identification, extraction, comparison, clustering and visualization of these structural elements presents an opportunity to enhance RNA science. Three broad applications are: (1) identification of modular, autonomous structural units for RNA nanotechnology, nanobiology and synthetic biology applications; (2) bioinformatic analysis to improve RNA 3D structure prediction from sequence; and (3) creation of searchable databases for exploring the binding specificities, structural flexibility, and dynamics of these RNA elements. In this contribution, we review methods developed for computational extraction of hairpin and internal loop motifs from a non-redundant set of high-quality RNA 3D structures. We provide a statistical summary of the extracted hairpin and internal loop motifs in the most recent version of the RNA 3D Motif Atlas. We also explore the reliability and accuracy of the extraction process by examining its performance in clustering recurrent motifs from homologous ribosomal RNA (rRNA) structures. We conclude with a summary of remaining challenges, especially with regard to extraction of multi-helix junction motifs.

Contents

Abbreviations: HL, Hairpin Loop; IL, Internal Loop; 3WJ, 3-way Junction; 4WJ, 4-way Junction; MHJ, Multi-Helix Junction; 2D structure, Secondary structure; PDB, Protein Data Bank; NDB, Nucleic Acid Database; WC, Watson-Crick; non-WC, non-Watson-Crick; BP, basepair; NR, Non-redundant; nt, nucleotide; S/R, Sarcin/Ricin; mRNA, messenger RNA; tRNA, transfer RNA; rRNA, ribosomal RNA; SSU, small ribosomal subunit; LSU, large ribosomal subunit.

* Corresponding author.
E-mail addresses: lorenan@bgsu.edu (L.G. Parlea), bsweene@bgsu.edu (B.A. Sweeney), zirbel@bgsu.edu (C.L. Zirbel), leontis@bgsu.edu (N.B. Leontis).

http://dx.doi.org/10.1016/j.ymeth.2016.04.025
1046-2023/

1. Introduction

This contribution concerns the computational extraction, analysis, and organization of RNA 3D motifs. In this introductory section, we define the different types of 3D motifs we observe in atomic-resolution RNA structures, discuss their properties and functions, and identify those that are amenable to current methods for extraction and clustering. Then we discuss, with reference to the wider goals of RNA bioinformatics, some reasons for systematically analyzing atomic resolution RNA 3D structures to identify, extract, and cluster 3D motifs, including construction of computational tools for RNA structure prediction and analysis. In the Materials and Methods section we discuss the selection of a target set of reliable, non-redundant (NR) RNA 3D structure files for analysis. We also provide computational details of methods currently used to build and maintain the RNA 3D Motif Atlas, see http://rna.bgsu.edu/rna3dhub/motifs [1]. In the Theory section, we provide the conceptual framework used to annotate, classify and cluster RNA motifs into coherent groups intended for downstream bioinformatic analysis. We begin the Results section by reviewing the current content of the RNA 3D Motif Atlas. Then we assess how well the current implementation of the computational pipeline organizes RNA 3D motif instances by tracking the clustering of motif instances from corresponding positions of homologous ribosomal RNA (rRNA) 3D structures from different organisms. We conclude with a summary of outstanding issues in extraction and classification of hairpin loops (HL) and internal loops (IL) and challenges in extending the 3D Motif Atlas to linker regions (defined below) and multi-helix junction (MHJ) loops.

Other workers have developed similar methods to identify, extract, and cluster RNA 3D motifs and websites to make them available in searchable formats [2–6]. This contribution is not meant as a comprehensive comparison of all the available methods, but as an attempt to provide detailed explanation of our own approach, as well as an extensive discussion of its limitations and the opportunities for future work in the field.

1.1. What are "RNA 3D Motifs"?

We use this term to refer to modular arrangements in 3D space of mutually interacting RNA nucleotides localized within the secondary structure, that is, nucleotides delimited by a set of mutually nested AU, GC, or GU *cis* Watson-Crick (WC) basepairs [7]. For our purposes, the secondary structure separates the nucleotides of the linear sequence into two disjoint classes, those that form the secondary structure, per se, and all the rest. The former comprise the WC-paired helices. The latter constitute the so-called "loops" and "linker segments" of RNA chains. These are the nucleotides that may form 3D motifs. Some ambiguity, however, remains regarding those nucleotides that form "isolated" WC pairs that occur within or between 3D motifs and which are not stacked contiguously on other WC pairs, on at least one side. It is not a simple,

Please cite this article in press as: L.G. Parlea et al., Methods (2016), http://dx.doi.org/10.1016/j.ymeth.2016.04.025