Bioinformatics, 2019, 1–3
doi: 10.1093/bioinformatics/btz553
Advance Access Publication Date: 9 July 2019
Applications Note

Structural bioinformatics

VeriNA3d: an R package for nucleic acids data mining

Diego Gallego[1,2], Leonardo Darré[1,3], Pablo D. Dans[1,*] and Modesto Orozco[1,2,*]

[1]Computational Biology Node, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, [2]Department of Biochemistry and Biomedicine, Faculty of Biology, University of Barcelona, Barcelona, Spain and [3]Functional Genomics Laboratory and Biomolecular Simulations Laboratory, Institute Pasteur of Montevideo, Montevideo, Uruguay

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on April 1, 2019; revised on June 20, 2019; editorial decision on July 4, 2019; accepted on July 6, 2019

Abstract

Summary: veriNA3d is an R package for the analysis of nucleic acids structural data, with an emphasis in complex RNA structures. In addition to single-structure analyses, veriNA3d also implements functions to handle whole datasets of mmCIF/PDB structures that could be retrieved from public/local repositories. Our package aims to fill a gap in the data mining of nucleic acids structures to produce flexible and high throughput analysis of structural databases.
Availability and implementation: http://mmb.irbbarcelona.org/gitlab/dgallego/veriNA3d.
Contact: pablo.dans@irbbarcelona.org or modesto.orozco@irbbarcelona.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The Protein Data Bank (PDB; Berman *et al.*, 2000) is one of the most widely used biological databases, and the central repository for structural data coming from NMR, X-Ray or more recently electron microscopy experiments (Fernandez-Leiro and Scheres, 2016). However, initiatives such as NDB (Coimbatore Narayanan *et al.*, 2014) and more recently specific repositories for hand-curated nucleic acid complexes and RNA motifs (Leontis and Zirbel, 2012; Sagendorf *et al.*, 2017) demonstrate the need to treat nucleic acids in a different way than protein, as nucleic acids require particular interpretative tools and specific analyses accounting for the complexity of the backbone polymorphisms (Balaceanu *et al.*, 2017), the helical nature of DNA/RNA molecules, the uniqueness of secondary and tertiary motifs (Petrov *et al.*, 2013), and the distinctive nature of the interactions between nucleic acids and proteins or cations (Hud, 2008; Rice and Correll, 2008). Unfortunately, despite the structural richness in nucleic-acid databases, few tools are available to automatically extract, analyze and integrate data into coherent pipelines (Bottaro *et al.*, 2019; Grant *et al.*, 2006). Here we present veriNA3d, a new R package build on top of bio3d (Grant *et al.*, 2006) that take advantage of bio3d functions to process, organize and explore available structural data, but extending its capabilities to take into account the singularities of nucleic acids structures. Our package incorporates nucleic-acids specific functions and provides the user with the ability to query large nucleic acids datasets, using APIs simple commands.

2 VeriNA3d functions

The functions in veriNA3d can be classified into three main blocks: (i) Dataset analysis; (ii) Single-structure analysis and (iii) Exploratory data analyses. Its power resides in its high degree of abstraction and the simplicity to perform complex tasks with few lines of R code. VeriNA3d offers integration with third-party utilities such as the non-redundant lists of RNA structures (Leontis and Zirbel, 2012), the εRMSD suggested to compare RNA structures (Bottaro *et al.*, 2014), a wrapper to the DSSR (Dissecting the Spatial Structure of RNA) software (Lu *et al.*, 2015) and query functions to access the PDBe REST API (Velankar *et al.*, 2016). In addition, it provides our own parser to read structures in mmCIF/PDB format and functions to process structural data and produce graphical outputs.

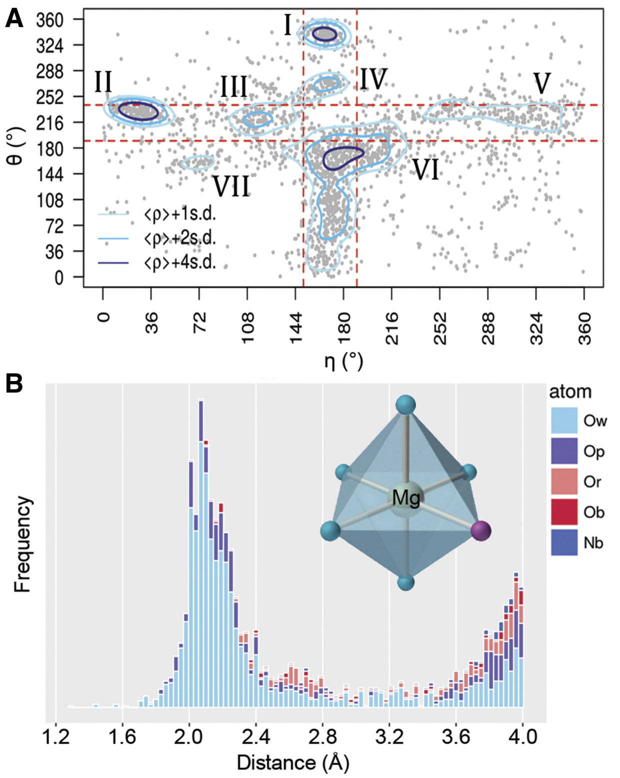

Fig. 1. Structural analysis using veriNA3d. (**A**) η-θ RNA map of non-helical nucleotides exhibiting sugars in North. (**B**) Distribution of distances between Mg^{2+} and its first shell. Ow/Op/Or/Ob: oxygen atoms of water, phosphate, sugar or base. Nb: base nitrogen

3 Comparative analysis and data mining of nucleic acids structures

We exemplify the usefulness of veriNA3d by presenting four relevant examples dealing with public repositories and cutting-edge analyses of DNA/RNA complexes (2 examples in Supporting Information). Simpler examples of veriNA3d usage and functions' description are provided via GitLab.

3.1 The RNA η-θ conformational landscape

The two-dimensional space defined by the RNA backbone pseudo-dihedrals η (eta: C4'$_{i-1}$, P$_i$, C4'$_i$, P$_{i+1}$) and θ (theta: P$_i$, C4'$_i$, P$_{i+1}$, C4'$_{i+1}$) is a simple method to characterize RNA conformational landscape, similar to Ramachandran plots in proteins. The first work analyzing η-θ was based on a small local repository of hand-curated structures (Wadley *et al.*, 2007). Using veriNA3d it is possible now to directly retrieve the data from weekly updated datasets (Leontis and Zirbel, 2012), allowing for an up-to-date version of the η-θ map (Fig. 1A). The script used to generate results for our first example, which filters out the helical nucleotides with sugar in North conformation, is reported in the Supplementary Table S1. Comparing to the original η-θ map (Wadley *et al.*, 2007), the updated version generated using veriNA3d identified a previously undescribed cluster with low density (VII in Fig. 1A) in the η-θ (North) space and reduces the importance of cluster V.

3.2 Mg^{2+}-RNA interaction sites across the PDB

Interactions with cations, especially Mg^{2+}, have been recognized as being crucial not only to stabilize specific RNA motifs (Bergonzo *et al.*, 2016), but also in recognition and catalysis (Lee *et al.*, 2008). With few lines of code, veriNA3d allowed us to scrutinize the atoms in the first coordination sphere of Mg^{2+} present in all RNA structures in the PDB (March 2019; resolution $\leq$ 2.0 Å). In our second example, the generated set of structures was analyzed in the light of two recent studies (Leonarski *et al.*, 2017; Zheng *et al.*, 2015), computing the distribution of distances to the central Mg^{2+} in 5 categories according to the type of atom bound in the first shell (Fig. 1B and Supplementary Table S2 for the code). We found a strong prevalence of Mg^{2+}–O_{water} (Ow) and Mg^{2+}–$O_{phosphate}$ (Op) contacts at short distances. On the contrary, contacts with the nucleobase or sugar atoms are enriched in the second peak (~4 Å) indicative of water-mediated interactions.

4 Conclusions and perspectives

We presented veriNA3d, an R package that greatly simplifies the systematic structural analysis of large nucleic acids datasets, addressing the physicochemical specificities of DNA/RNA and their molecular partners. Our package takes advantage of the extensive graphical and statistical capabilities of the R environment, and can be freely downloaded for local use.

Acknowledgements

M.O. is an ICREA researcher. P.D.D. and L.D. are PEDECIBA and ANII researchers.

Funding

The ERC Council (SimDNA and 2shRNA) and the H2020 project BioExcel (to M.O.). ANII and Institute Pasteur of Montevideo (postdoctoral funding) (to L.D.).

Conflict of Interest: none declared.

References

Balaceanu,A. *et al.* (2017) The role of unconventional hydrogen bonds in determining BII propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**, 21–28.

Bergonzo,C. *et al.* (2016) Divalent ion dependent conformational changes in an RNA stem-loop observed by molecular dynamics. *J. Chem. Theory Comput.*, **12**, 3382–3389.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bottaro,S. *et al.* (2019) Barnaba: software for analysis of nucleic acid structures and trajectories. *RNA*, **25**, 219–231.

Bottaro,S. *et al.* (2014) The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.*, **42**, 13306–13314.

Coimbatore Narayanan,B. *et al.* (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res.*, **42**, D114–D122.

Fernandez-Leiro,R. and Scheres,S.H.W. (2016) Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, **537**, 339–346.

Grant,B.J. *et al.* (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.

Hud,N.V. (ed) (2008) *Nucleic Acid–Metal Ion Interactions*. The Royal Society of Chemistry, Cambridge, UK.

Lee,T.-S. *et al.* (2008) Role of Mg2+ in hammerhead ribozyme catalysis from molecular simulation. *J. Am. Chem. Soc.*, **130**, 3053–3064.

Leonarski,F. *et al.* (2017) Mg^{2+} ions: do they bind to nucleobase nitrogens? *Nucleic Acids Res.*, **45**, 987–1004.

Leontis,N.B. and Zirbel,C.L. (2012) *Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking*. Springer, Berlin, Heidelberg, pp. 281–298.

Lu,X.-J. *et al.* (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142. doi: 10.1093/nar/gkv716.

Petrov,A.I. *et al*. (2013) Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, **19**, 1327–1340.

Rice,P.A. and Correll,C.C. (2008) *Protein–Nucleic Acid Interactions: Structural Biology*. Royal Society of Chemistry, Cambridge, UK.

Sagendorf,J.M. *et al*. (2017) DNAproDB: an interactive tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res*., **45**, W89–W97.

Velankar,S. *et al*. (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res*., **44**, D385–95.

Wadley,L.M. *et al*. (2007) Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. *J. Mol. Biol*., **372**, 942–957.

Zheng,H. *et al*. (2015) Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection. *Nucleic Acids Res*., **43**, 3789–3801.