

Use cases of veriNA3d

Diego Gallego Perez^{1,2}

¹University of Barcelona - Dept. Biochemistry and Molecular Biomedicine

²Institute for Research in Biomedicine - (IRB Barcelona), Barcelona Institute of Science and Technology

2018-11-26

Contents

1	Introduction	2
2	Parsing mmCIF files and the R S4 Object "CIF".	2
2.1	Origin and standardization of mmCIF files	2
2.2	The CIF object	3
3	Main use case: Manage Nucleic Acid datasets.	3
4	Use case: Querying the EMBL-EBI REST API	3
5	Use case: eRMSD	3
6	Use case: Generate substructures	3
	References	3

1 Introduction

The R language provides an excellent interface for statistical analysis, which is also interesting from the point of view of structural data. This gap was filled in 2006 by the R package [bio3d](#) (Grant et al. 2006). It was presented as a suite of tools to handle PDB formatted structures, and trajectories. It integrates a variety of functions to analyse these data, from sequence to 3D structure (RMSD, NMA, PCA... see their [documentation](#) for details). As far as we know, [bio3d](#) represented the only structural package for R until now.

The R package presented in here, [veriNA3d](#), does not replace [bio3d](#) at all. Rather, it was developed on top of it to cover additional necessities. The only common tool integrated in both packages is a parser for mmCIF files (see below). [veriNA3d](#) is mainly intended (but not limited) to the analysis of Nucleic Acids. It integrates a higher level of abstraction than [bio3d](#) since it also allows the analysis of datasets, in addition to analysis of single structures. The functions in the package could be divided in the following blocks (complete list in README file):

- Dataset level: Functions to get and analyse lists of pdb IDs. This includes access to the [representative lists of RNA](#) by (Leontis and Zirbel 2012) and other analytical functions.
- Structure level: Functions to get data, parse mmCIF files and analyse these data.
- Plots: examples to show the results of the previous analysis.

The complete list of functions can be found in the README.md file within the package, also accessible on the gitlab [main page](#).

2 Parsing mmCIF files and the R S4 Object “CIF”.

2.1 Origin and standardization of mmCIF files

Atomic structural data of macromolecules has long been distributed in the PDB file format. However, one of its main limitations is the column size for the coordinates data, which didn't allowed to save molecules with more than 99999 atoms, more than 62 chains or more than 9999 residues (in a chain).

Given that the Protein Data Bank is continuously growing and accepting bigger structures (e.g. a whole *E.coli* ribosome has over 140000 atoms - [pdbID 4V4S](#)), an alternative file format became the standard: the mmCIF file format.

The mmCIFs are an evolution of the Crystallographic Information File (CIF), originally used for small molecule structures. It stands for **macromolecular CIF** file, and it has actually coexisted with the PDB format since the 1997. However, since the PDB is easier to parse and such big structures didn't populate the database at the time, most software has been developed for the PDB format.

The PDB format was definitely frozen in 2014. However, it will still coexist with the standard mmCIF format as long as “old” software evolves to accept mmCIFs. Following this trend, the [bio3d](#) R package integrated a `read.cif` function in their version 2.3. At that time, we had already started the development of our own `cifParser` function. Given that the mmCIF format

is constantly evolving and that both functions take slightly different approaches, we decided to offer our own version of it, which might provide an useful and fast alternative for users working with mmCIF files.

2.2 The CIF object

The R language supports different kinds of [objects](#), being the S3 and S4 the most common ones. The S3 objects Later on, i also integrated a parser for mmCIF files.

3 Main use case: Manage Nucleic Acid datasets

Get Leontis list, change representative structures and analyse them with one of the pipelines

4 Use case: Querying the EMBL-EBI REST API

Query EBI API and how to construct new queries!

5 Use case: eRMSD

For two NMR models of RNA/DNA, compute the eRMSD

6 Use case: Generate substructures

For a given structure (CIF or PDB), generate a smaller PDB with the region of interest and surroundings

References

- Grant, B.J., A.P.C. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S.D. Caves. 2006. "Bio3d: An R Package for the Comparative Analysis of Protein Structures." *Bioinformatics* 22 (21): 2695–6.
- Leontis, N.B., and C.L. Zirbel. 2012. "Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking." In *RNA 3D Structure Analysis and Prediction*, edited by N. Leontis and E. Westhof, 27:281–98. Springer Berlin Heidelberg.