Use cases of veriNA3d

Diego Gallego Perez [1,2]

[1]University of Barcelona - Dept. Biochemistry and Molecular Biomedicine
[2]Institute for Research in Biomedicine - (IRB Barcelona), Barcelona Institute of Science and Technology

2018-12-10

Contents

1 Introduction: Structural Bioinformatics in R

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 formated structures, and trajectories. It integrates a variety of functions to analyse from sequence to 3D structure data (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:

- 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

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 continously 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 **m**acro**m**olecular **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 definetely frozen in 2014. However, it will still coexist with the standard mmCIF format as long as all softwares evolve 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

Parsing a particular file format often involves creating a new class of object. In R, the principal objects are called S3, S4 and RC. Our container for mmCIF data is an S4 object called **CIF**, in contrast with the S3 object (called **pdb**) in bio3d - it is worth noting that this difference does **not** affect the compatibility between the two packages (see below for details).

Since different mmCIF files usually have different sections of data (in addition to the coordinates), we carried out an analisis that checked which ones are always present in all mmCIF files (this included all mmCIF files in the Protein Data Bank in March 2018), and reached a list of 14 items:

- Atom_site
- Atom_sites
- Atom_type
- Audit_author
- Audit_conform
- Chem_comp
- Database_2
- Entity
- Entry
- Exptl
- Pdbx_database_status
- Struct
- Struct_asym
- Struct_keywords

The detailed description of each data sections can be found in the mmCIF main site.

The CIF object is created by the `cifParser` function and contains these 14 sections of data, which can be accessed with the CIF accessors. To see the accessor functions run:

```
library(veriNA3d)
?cif_accessors
```

To read a mmCIF file and access the coordinates data, use:

```
## To parse a local mmCIF file:
# cif <- cifParser("your-file.cif")
## To download from PDB directly:
cif <- cifParser("1bau")
cif
#>
#> -- mmCIF with ID: 1BAU ------------------------------------------------------
#>
#> Author description:  NMR STRUCTURE OF THE DIMER INITIATION COMPLEX OF HIV-1 GENOMIC RNA, MINIMIZED AVERAGE
#> mmCIF version:       5.279
#>
#> To extract coordinates and other data use accessor functions
#> (type ?cif_accessors for details)
## To see the coordinates:
coords <- cifAtom_site(cif)
head(coords)
#>   group_PDB id type_symbol label_atom_id label_alt_id label_comp_id
```

```
#> 1      ATOM  1           O           O5'            .             G
#> 2      ATOM  2           C           C5'            .             G
#> 3      ATOM  3           C           C4'            .             G
#> 4      ATOM  4           O           O4'            .             G
#> 5      ATOM  5           C           C3'            .             G
#> 6      ATOM  6           O           O3'            .             G
#>   label_asym_id label_entity_id label_seq_id pdbx_PDB_ins_code Cartn_x
#> 1             A               1            1                 ?  23.989
#> 2             A               1            1                 ?  24.965
#> 3             A               1            1                 ?  25.937
#> 4             A               1            1                 ?  26.741
#> 5             A               1            1                 ?  25.259
#> 6             A               1            1                 ?  24.868
#>   Cartn_y Cartn_z occupancy B_iso_or_equiv pdbx_formal_charge auth_seq_id
#> 1   8.289 -15.135         1              0                  ?           1
#> 2   9.100 -14.503         1              0                  ?           1
#> 3   9.725 -15.512         1              0                  ?           1
#> 4   8.744 -16.162         1              0                  ?           1
#> 5  10.527 -16.627         1              0                  ?           1
#> 6  11.838 -16.252         1              0                  ?           1
#>   auth_comp_id auth_asym_id auth_atom_id pdbx_PDB_model_num
#> 1            G            A          O5'                  1
#> 2            G            A          C5'                  1
#> 3            G            A          C4'                  1
#> 4            G            A          O4'                  1
#> 5            G            A          C3'                  1
#> 6            G            A          O3'                  1
```

2.3 Bidirectional compatibility with bio3d

Using the `cifPArser` will often be the first step to analyse a structure. However, if the analisys requires any of the bio3d functions, then a conversion should be done with the `cifAsPDB` function.

```
pdb <- cifAsPDB(cif)
pdb
#>
#>  Call:  "1BAU"
#>
#>    Total Models#: 1
#>      Total Atoms#: 1486,  XYZs#: 4458  Chains#: 2  (values: A B)
#>
#>      Protein Atoms#: 0  (residues/Calpha atoms#: 0)
#>      Nucleic acid Atoms#: 1486  (residues/phosphate atoms#: 46)
#>
#>      Non-protein/nucleic Atoms#: 0  (residues: 0)
#>      Non-protein/nucleic resid values: [ none ]
#>
#>    Nucleic acid sequence:
```

```
#>       GGCAAUGAAGCGCGCACGUUGCCGGCAAUGAAGCGCGCACGUUGCC
#>
#> + attr: atom, xyz, calpha, model, flag, call
```

It takes a CIF object and generates an equivalent pdb object (as used by all bio3d functions). In addition, all veriNA3d functions are prepared to accept as input either the CIF or pdb objects. Therefore, the compatibility between the two packages is bidirectional.

3 Getting data from Application Programming Interfaces (API)

Getting data is the first step of any pipeline, and parsing files is just one of the many ways data can be accessed. Application Programming Interfaces (API) are an intermediate point of access to a remote database. APIs offer the users a series of *endpoints* or *calls*, which are just **links**. Thus, instead of dealing directly with the database with SQL or other language, the user can just use the appropriate *call* to send a query to the API, and it will return the desired data.

In addition to parsing mmCIF files, veriNA3d also offers a series of functions to send queries to different APIs. Since sending queries to remote APIs requires Internet access, the full functionality of veriNA3d might depend on a good connection.

To see all the query functions, run:

```
?queryFunctions
```

IMPORTANT NOTE: The APIs accessed by veriNA3d are free of use with no limit of queries per user. However, this could change if the users of the APIs use them irresponsibly. Servers could eventually fall down if they receive more *calls* than they can actually process. To avoid that, veriNA3d actually saves in memory the result of any query, and any time that you use the same query again, it will take the cached result. To see this effect, run this test:

```
## Run a query for the first time, which will access the API
tech <- queryTechnique("4KQX", verbose=TRUE)
#> [1] "Querying: http://www.ebi.ac.uk/pdbe/api/pdb/entry/summary/4KQX"
#> [1] "Getting expType from API"
#> [1] "Saving expType in RAM"

## Run the same query for the second time, which will get it from memory
tech <- queryTechnique("4KQX", verbose=TRUE)
#> [1] "Querying: http://www.ebi.ac.uk/pdbe/api/pdb/entry/summary/4KQX"
#> [1] "Getting expType from RAM"
```

However, this is only saved across the current session of R, and any script that uses the query functions will send them to the API every time it is run. VeriNA3d does not guarantee the correct service of the APIs and it does not monitor any of your processes. However, the API providers can known which IP address is querying their services at any time. To avoid overloading the servers, please make a responsible use (e.g. save locally the data that you use frequently).

3.1 Querying the EMBL-EBI REST API

An invaluable resource for structural & computational biologists is the PDBe REST API (Velankar et al. 2015). Around this technology, the package includes the following set of functions:

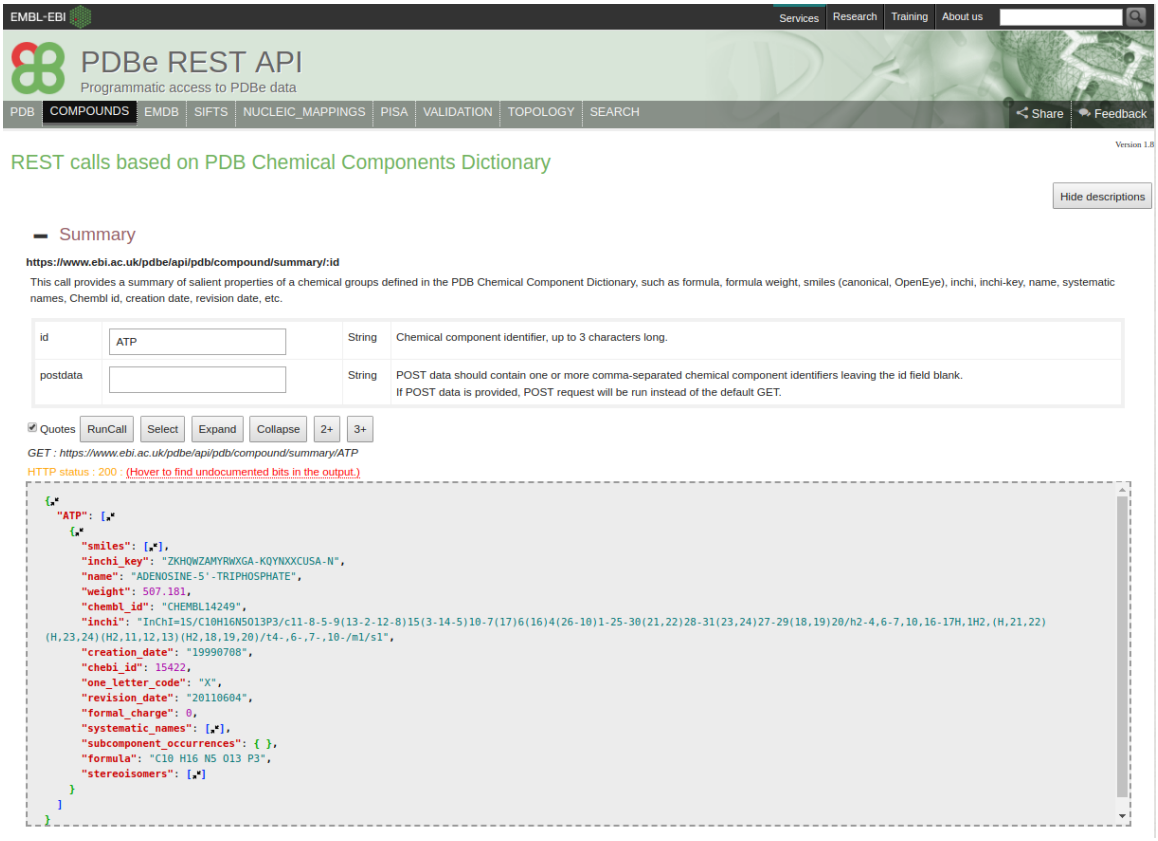
- queryAuthors: List of authors
- queryReldate: Release date
- queryDepdate: Deposition date
- queryRevdate: Revision date
- queryDescription: Author description
- queryEntities: Entitity information
- queryFormats: File formats for the given ID
- queryModres: Modified residues
- queryLigands: Ligands in structure
- queryOrgLigands: Ligands in structure (substracting ions)
- queryResol: Resolution (if applicable)
- queryTechnique: Experimental Technique
- queryStatus: Released/Obsolete and related status information

The list of functions is **intendedly limited** in comparison with the dozens of *enpoints* of the REST API. Integrating them all would innecessarily increase the total amount of functions of the package. Moreover, the API might offer more and more *endpoints* with the time, and trying to keep them all would make the manteinance of this package more difficult. To allow the users access to their desired endpoints, an alternative method is provided.

The core of all these (and other) query functions is `queryAPI`, which integrates all the error-handling and cache functionalities. With the `queryAPI` function, any user can design their own queries, with a simple process. Herein a couple of examples.

3.1.1 Example 1

This snapshot shows the REST API website and a *call* that is not implemented in veriNA3d:



The link of this *endpoint* is: https://www.ebi.ac.uk/pdbe/api/pdb/compound/summary/ATP

The `queryAPI` function can understand and send this query using the arguments 'ID', 'API', 'string1', and 'string2' properly:

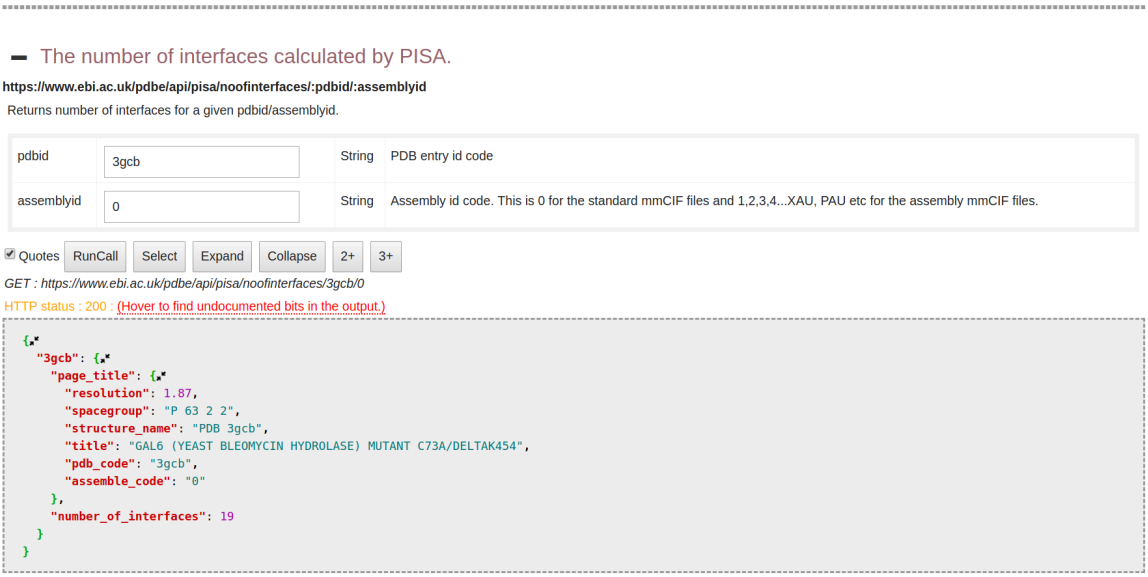
```
atpsummary <- queryAPI(ID="ATP", API="ebi",
                       string1="pdb/compound/summary/", string2="")
str(atpsummary$ATP)
#> 'data.frame':    1 obs. of  15 variables:
#>  $ smiles                 :List of 1
#>   ..$ :'data.frame': 2 obs. of  3 variables:
#>   .. ..$ program: chr  "CACTVS" "OpenEye OEToolkits"
#>   .. ..$ version: chr  "3.341" "1.5.0"
#>   .. ..$ name   : chr  "Nc1ncnc2n(cnc12)[C@@H]3O[C@H](CO[P@](O)(=O)O[P@@](O)(=O)O[P](O)(O)=O)[C@@H](O)[C@H
#>  $ inchi_key              : chr "ZKHQWZAMYRWXGA-KQYNXXCUSA-N"
#>  $ name                   : chr "ADENOSINE-5'-TRIPHOSPHATE"
#>  $ weight                 : num 507
#>  $ chembl_id              : chr "CHEMBL14249"
#>  $ inchi                  : chr "InChI=1S/C10H16N5O13P3/c11-8-5-9(13-2-12-8)15(3-14-5)10-7(17)6(16)4(26-1
#>  $ creation_date          : chr "19990708"
#>  $ chebi_id               : int 15422
#>  $ one_letter_code        : chr "X"
#>  $ revision_date          : chr "20110604"
#>  $ formal_charge          : int 0
#>  $ systematic_names       :List of 1
```

```
#>   ..$ :'data.frame': 2 obs. of  3 variables:
#>   .. ..$ program: chr  "ACDLabs" "OpenEye OEToolkits"
#>   .. ..$ version: chr  "10.04" "1.5.0"
#>   .. ..$ name   : chr  "adenosine 5'-(tetrahydrogen triphosphate)" "[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)
#>  $ subcomponent_occurrences:'data.frame':    1 obs. of  0 variables
#>  $ formula                 : chr "C10 H16 N5 O13 P3"
#>  $ stereoisomers           :List of 1
#>   ..$ :'data.frame': 1 obs. of  2 variables:
#>   .. ..$ name        : chr "9-{5-O-[(S)-hydroxy{[(R)-hydroxy(phosphonooxy)phosphoryl]oxy}phosphoryl]-beta
#>   .. ..$ chem_comp_id: chr "HEJ"
```

- The common root in all the REST API *endpoints* is "https://www.ebi.ac.uk/pdbe/api/", which is internally managed by the function by using *API="ebi"*.
- The *string1="pdb/compound/summary/"* indicates everything that comes after the root and before the ID.
- The *ID="ATP"* obviously represents the desired structure, either a 4 character string for a pdbID or <= 3 character string for compounds.
- The *string2=""* is also a necessary argument that reflects that nothing comes after the ID.

3.1.2 Example 2

This snapshot shows a second *call* not implemented in veriNA3d:



The link of this *endpoint* is: https://www.ebi.ac.uk/pdbe/api/pisa/noofinterfaces/3gcb/0

The proper call with `queryAPI` would be:

```
PISAsummary <- queryAPI(ID="3gcb", API="ebi",
                        string1="pisa/noofinterfaces/", string2="0")
str(PISAsummary$"3gcb")
#> List of 2
```

```
#>  $ page_title          :List of 6
#>   ..$ resolution    : num 1.87
#>   ..$ spacegroup    : chr "P 63 2 2"
#>   ..$ structure_name: chr "PDB 3gcb"
#>   ..$ title         : chr "GAL6 (YEAST BLEOMYCIN HYDROLASE) MUTANT C73A/DELTAK454"
#>   ..$ pdb_code      : chr "3gcb"
#>   ..$ assemble_code : chr "0"
#>  $ number_of_interfaces: int 19
```

This second example shows a case in which the "string2" argument is necessary. If you are unsure about the real link that is actually being constructed, you can always use *verbose=TRUE* to see it printed.

4 Calculate the ϵRMSD

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

5 Generate substructures

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

6 Manage Nucleic Acid datasets

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

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.

Velankar, S., G. Van Ginkel, Y. Alhroub, G.M. Battle, J.M. Berrisford, M.J. Conroy, J.M. Dana, et al. 2015. "PDBe: Improved Accessibility of Macromolecular Structure Data from PDB and EMDB." *Nucleic Acid Research* 44 (D1): D385–395.