Biotite: a unifying open source computational biology framework in Python

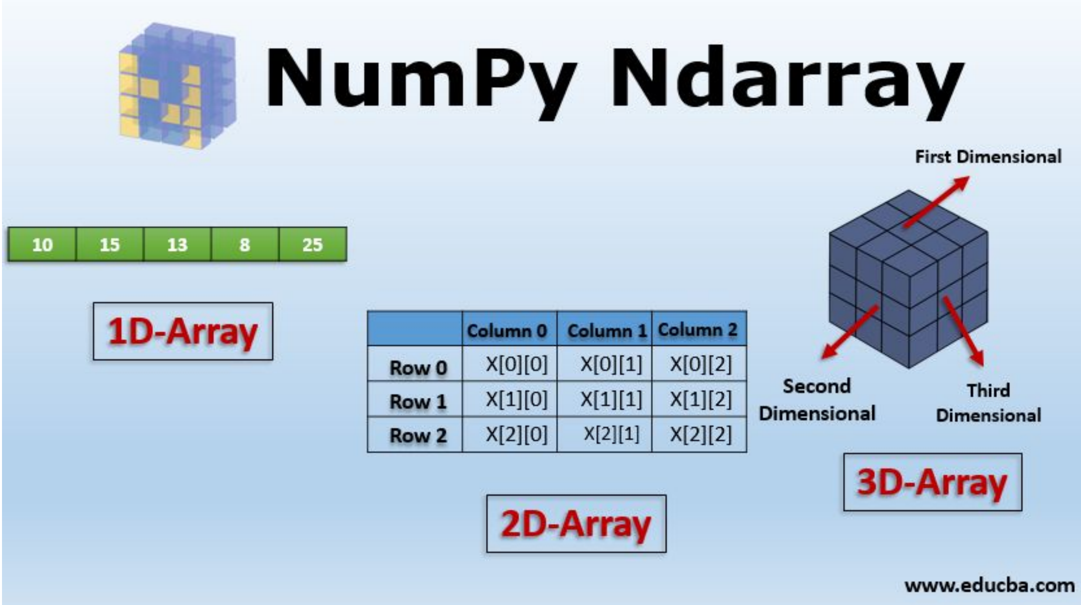
Isaac Chung, Tanmay Thakral

Kunzmann, P., Hamacher, K. Biotite: a unifying open source computational biology framework in Python. *BMC Bioinformatics* 19, 346 (2018). https://doi.org/10.1186/s12859-018-2367-z

What is Biotite?

- open-source computational biology framework in Python
- bundles sequence analysis and structural bioinformatics
- high and fast performance
- fetches relevant reading and writing files
- efficient and intuitive analysis and manipulation of their data

N-Dimensional Array (Ndarray)



10	15	13	8	25

	Column 0	Column 1	Column 2
Row 0	X[0][0]	X[0][1]	X[0][2]
Row 1	X[1][0]	X[1][1]	X[1][2]
Row 2	X[2][0]	X[2][1]	X[2][2]

- multidimensional container of items of the same type and size
- number of dimensions and shape defined by tuple of N non-negative integers

https://www.educba.com/numpy-ndarray/

Implementation

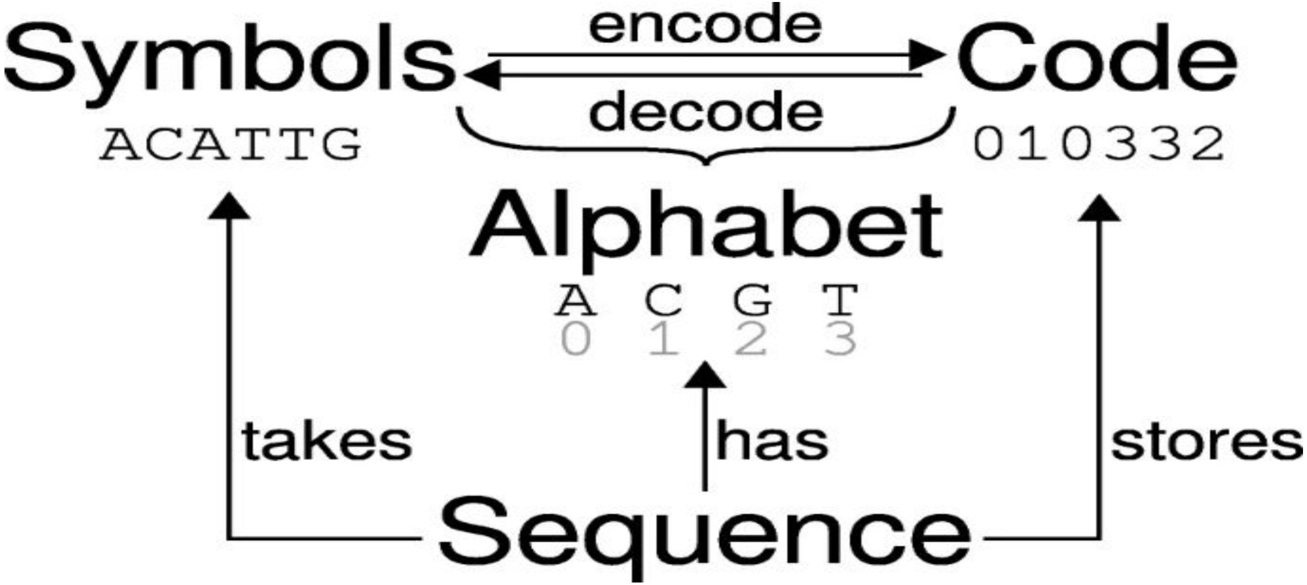
Biotite can be divided into 4 subpackages:

1. Sequence
2. Structure
3. Database
4. Application

Sequences Subpackage

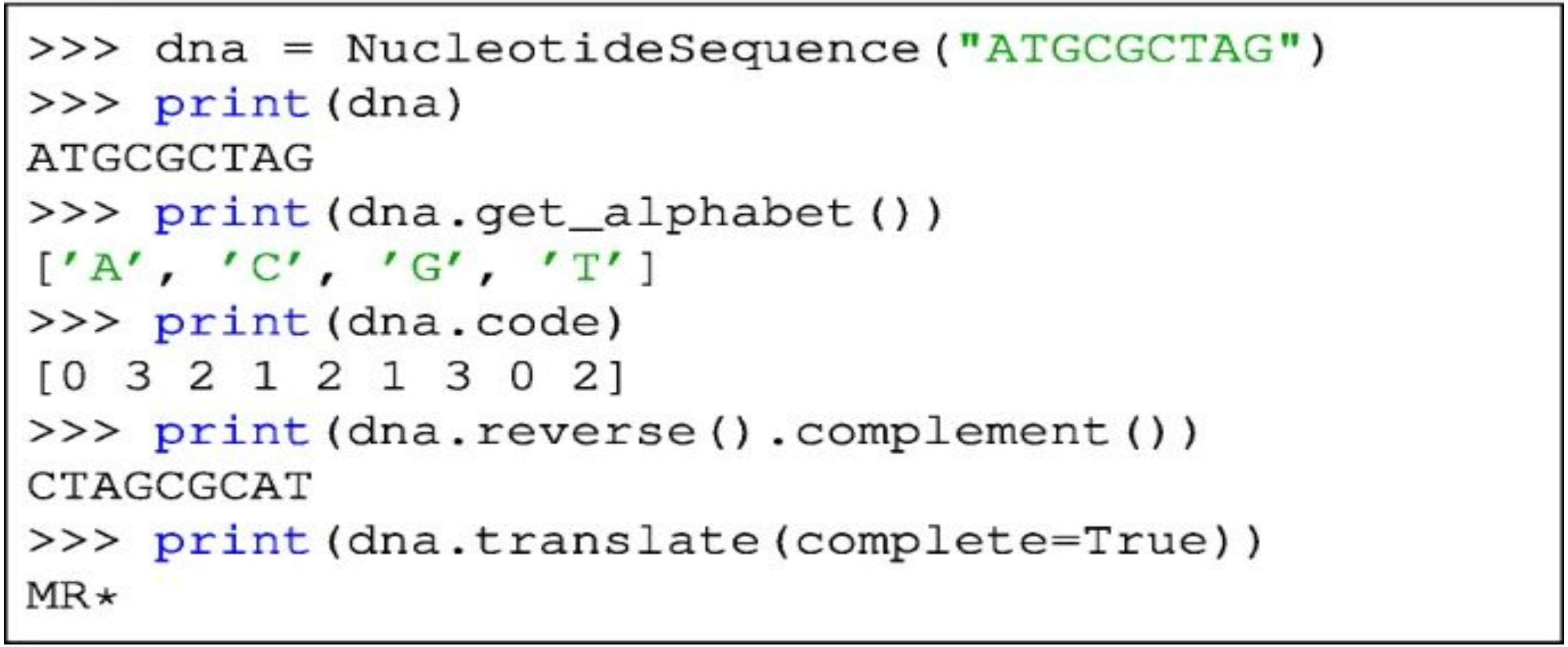
- Tools for handling classical sequences (nucleotide and protein) and also protein structures or pharmacophores
- Contains :
    - Nucleotide and protein sequences
    - Alignments
    - Visualizations

Nucleotide and protein sequences



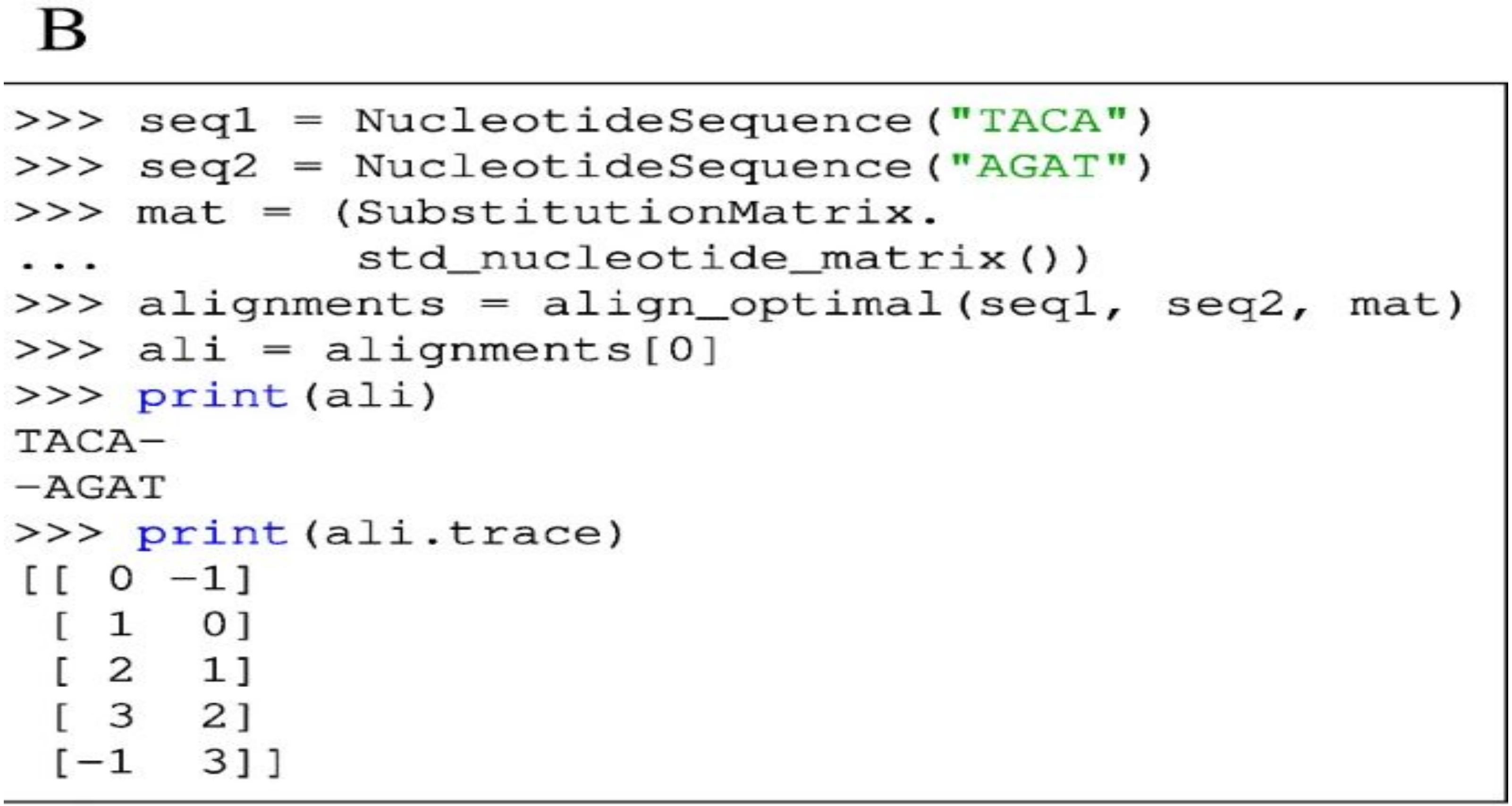
Nucleotide and protein sequences

A



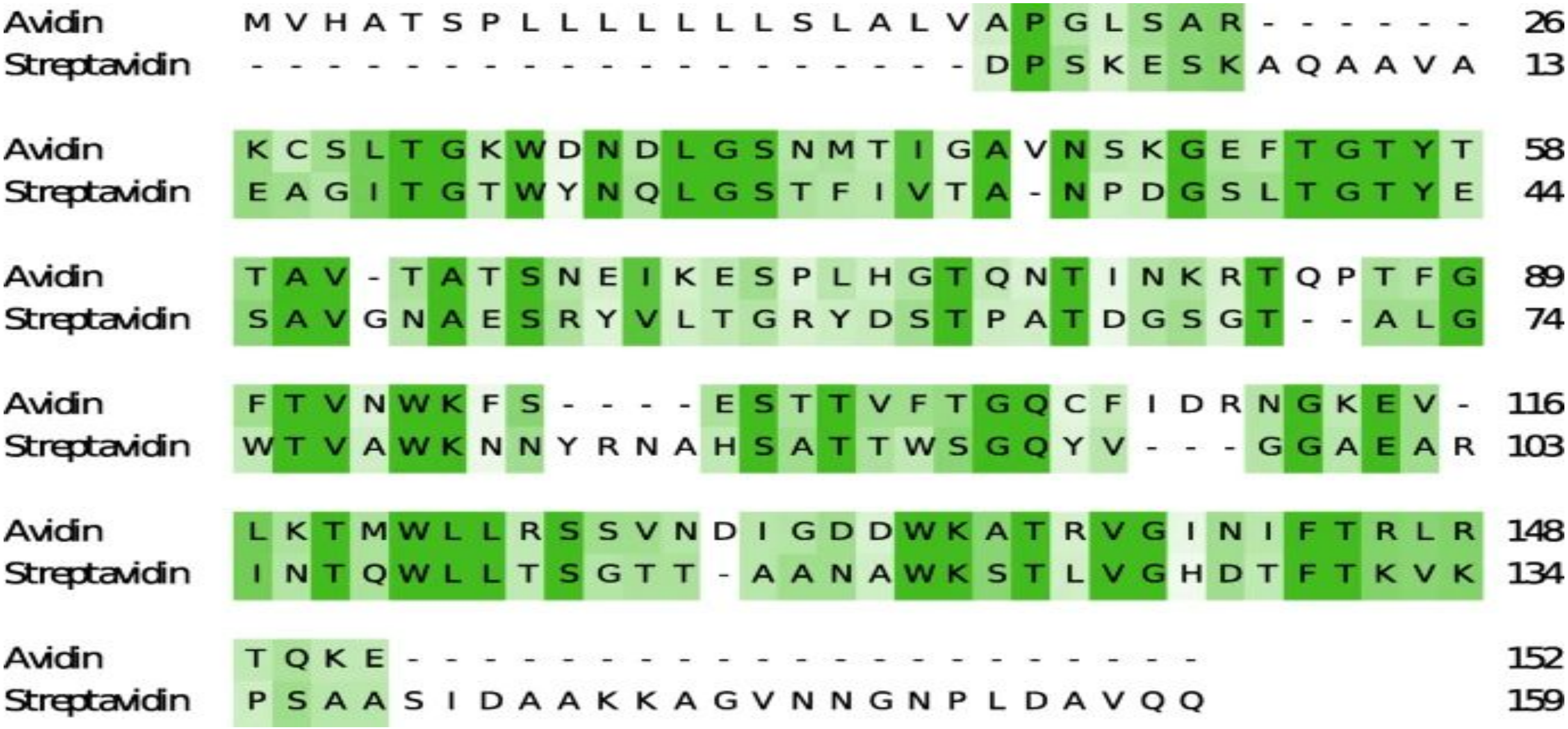
```
>>> dna = NucleotideSequence("ATGCGCTAG")
>>> print(dna)
ATGCGCTAG
>>> print(dna.get_alphabet())
[’A’, ’C’, ’G’, ’T’]
>>> print(dna.code)
[0 3 2 1 2 1 3 0 2]
>>> print(dna.reverse().complement())
CTAGCGCAT
>>> print(dna.translate(complete=True))
MR*
```

Alignments



```
>>> seq1 = NucleotideSequence("TACA")
>>> seq2 = NucleotideSequence("AGAT")
>>> mat = (SubstitutionMatrix.
...           std_nucleotide_matrix())
>>> alignments = align_optimal(seq1, seq2, mat)
>>> ali = alignments[0]
>>> print(ali)
TACA-
-AGAT
>>> print(ali.trace)
[[ 0 -1]
 [ 1  0]
 [ 2  1]
 [ 3  2]
 [-1  3]]
```

Visualiser



```
Avidin         M V H A T S P L L L L L L L S L A L V A P G L S A R - - - - - -    26
Streptavidin   - - - - - - - - - - - - - - - - - - - D P S K E S K A Q A A V A    13

Avidin         K C S L T G K W D N D L G S N M T I G A V N S K G E F T G T Y T    58
Streptavidin   E A G I T G T W Y N Q L G S T F I V T A - N P D G S L T G T Y E    44

Avidin         T A V - T A T S N E I K E S P L H G T Q N T I N K R T Q P T F G    89
Streptavidin   S A V G N A E S R Y V L T G R Y D S T P A T D G S G T - - A L G    74

Avidin         F T V N W K F S - - - - E S T T V F T G Q C F I D R N G K E V -   116
Streptavidin   W T V A W K N N Y R N A H S A T T W S G Q Y V - - - G G A E A R   103

Avidin         L K T M W L L R S S V N D I G D D W K A T R V G I N I F T R L R   148
Streptavidin   I N T Q W L L T S G T T - A A N A W K S T L V G H D T F T K V K   134

Avidin         T Q K E - - - - - - - - - - - - - - - - - - - - -                 152
Streptavidin   P S A A S I D A A K K A G V N N G N P L D A V Q Q                 159
```

Database Subpackage

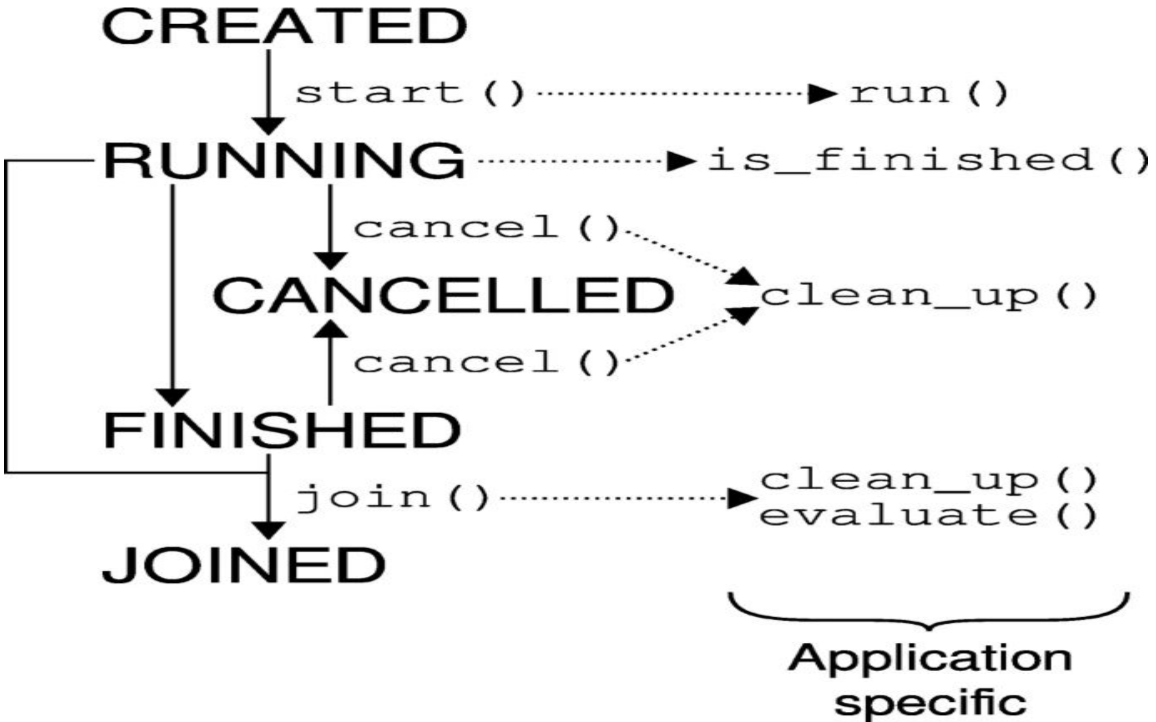
- Download files from RCSB Protein Database (PDB) and NCBI Entrez server (Global Query Cross-Database Search System)

Structural Subpackage

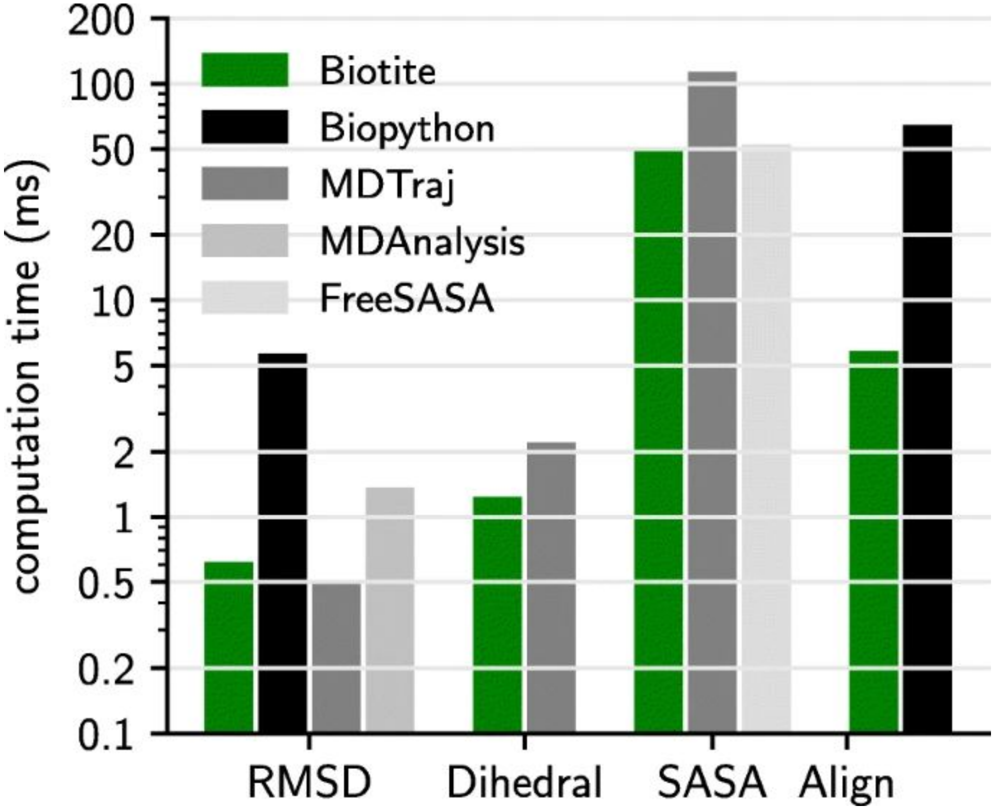
- Biomolecular structure representation
- Atom class: most basic unit of the representation of a biomolecular structure
  - Contains info about atom coordinates with length three ndarray
  - Info about annotations (eg. chain ID, residue ID, atom name)
- AtomArray represents entire structure consisting of multiple atoms
- AtomArrayStack represents multi-model structures

Application Subpackage

In this subpackage `Biotite` offers interfaces to external software, such as NCBI BLAST.These interfaces wrap the execution of the respective program on the local machine

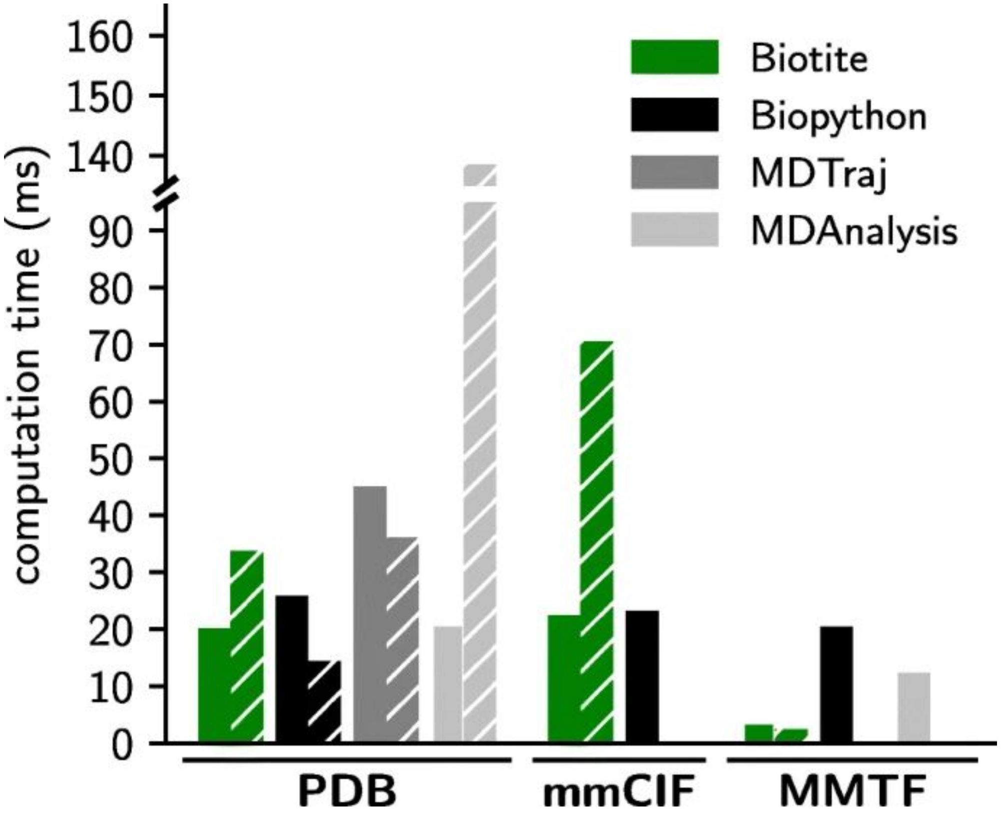


Performance of implemented analysis algorithms



- RMSD: root-mean-square-deviation
- Dihedral: angle between two planes
- SASA: accessible surface area
- Align: protein or nucleotide sequences alignment

Performance of structure file input and output



- PDB: Protein Database
- mmCIF: Crystallographic Information File
- MMTF: Macromolecular Transmission Format