Dinopy (DNA Input and Output in Python) is a Python package that aims to simplify the development of bioinformatics applications by providing efficient facilities for DNA input and output. At the time of writing, there is no library for I/O of DNA specific files available which makes full use of the potential of Cython [1]. Dinopy exports Cython level API bindings which can be used by other Cython applications for increased speedup.

The Power of Cython

Cython aims to combine the performance of C and C++ with Python's convenient development process. This is achieved by translating Cython code to C and compiling it.

Advantages:
- fewer type checks at runtime
- fast numeric operations
- faster execution
- easy integration of efficient C / C++ code
- used in many big projects (SciPy [3], pandas [4], scikit-learn [6], ...)

Downsides:
- manual typing required
- not all Python features supported (yet)

Goals:
- early typing → avoid Python overhead
- use efficient C level functions wherever possible

Core Goals of dinopy

Goal 1 - performance:
- data type (dtype) system, similar to numpy [8]
- efficient (C level) functions to modify sequences (tailored to dtypes)
- generators → small memory footprint
- 2bit and 4bit encoding of sequences
- use faidx index files → random access on FASTA files
- provide Cython level API bindings

Goal 2 - usability:
- different views on sequences
- modular processors to modify views and sequences
- exclusion of information (e.g. read names, quality values) when not needed
- automatic handling of gzipped files

Availability

The dinopy package is available for Python 3.2+ under the MIT License. It has so far only been tested on Linux operating systems. The source code can be downloaded from Bitbucket (https://bitbucket.org/HenningTimm/dinopy) or scan the QR code on the right) and installed using python setup.py install.

References