entrezpy: a dedicated Python library to dynamically interact with NCBI Entrez databases

Jan P. Buchmann and Edward C. Holmes *Contact:* jan.buchmann@sydney.edu.au

This work was supported by an ARC Australian Laureate Fellowship [FL170100022] to E.C.H.

NCBI Entrez databases at your fingertips

entrezpy is a dedicated Python library to interact with NCBI Entrez databases [0] via the E-Utilities [1].

entrezpy ***has been designed 'to do one thing and do it well'. It enables the querying and downloading data from the Entrez databases, one of the largest life sciences data repositories, while giving a developer the freedom to easily integrate specific analysis functions.***

entrezpy facilitates the implementation of queries to query or download data from the Entrez databases, e.g. search for specific sequences, publications, or fetch your favorite genome. For more complex queries entrezpy offers the class Conduit to assemble, run query pipelines, or reuse previous queries.

Availability

- **License**: LGPLv3
- **Python** : ≥ 3.6
- **Source:** https://gitlab.com/ncbipy/entrezpy
- **PyPi:** https://pypi.org/project/entrezpy/
- **Documentation, examples, tutorials:** http://entrezpy.readthedocs.io/
- **Publication:** https://doi.org/10.1093/bioinformatics/btz385

Synopsis

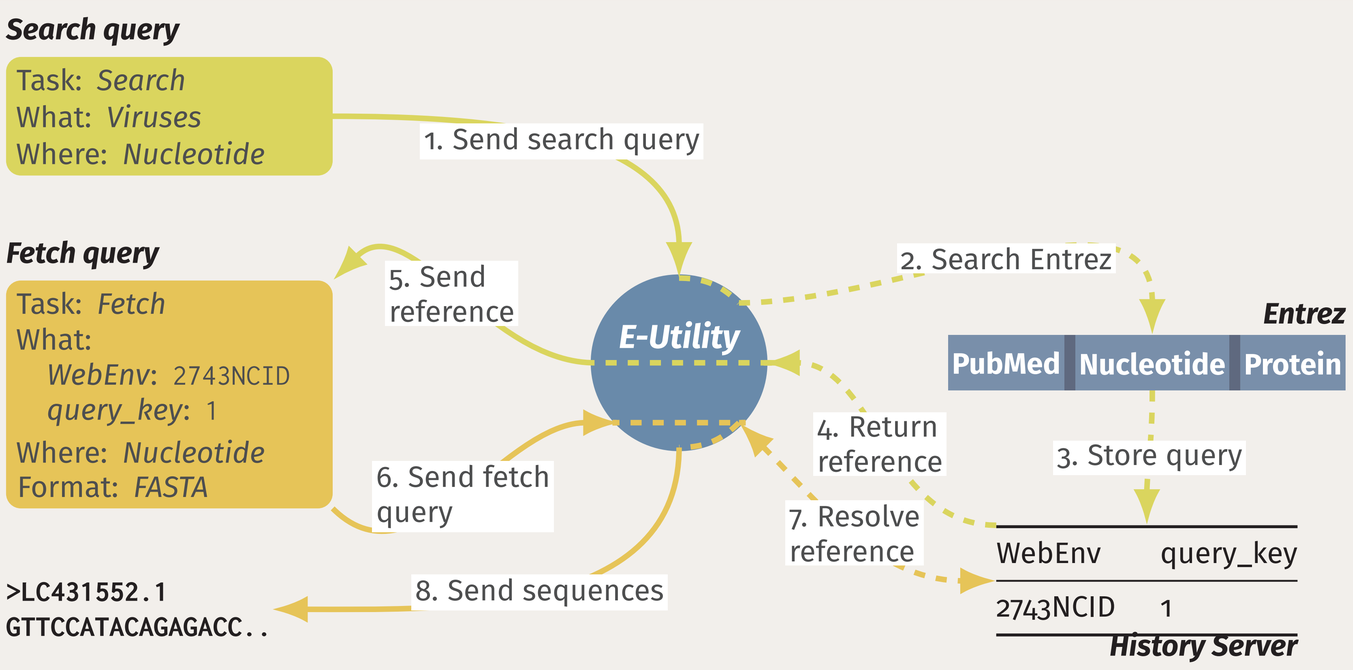
Install **entrezpy** ***and start python interpreter***

```
$ pip install entrezpy --user
$ python3
```

Fetch 10 nucleotide sequences for Influenza H3N2 HA submitted between 2000 and 2019

```
>>> import entrezpy.conduit
>>> c = entrezpy.conduit.Conduit('myemail')
>>> fetch_influenza = c.new_pipeline()
>>> sid = fetch_influenza.add_search({'db':'nucleotide', 'rettype':'count' 'term':'H3N2 [organism] AND HA', 'sort':'Date Released', '
    mindate':2000, 'maxdate':2019, 'datetype':'pdat'})
>>> fetch_influenza.add_fetch({'retmax':10, 'retmode':'text', 'rettype':'fasta'}, dependency=sid)
>>> c.run(fetch_influenza)
```

Basic NCBI Entrez request with E-Direct and entrezpy.Conduit



Query	E-Utility	POST parameter
Search	esearch.fcgi	db='nucleotide'&term='viruses[ORGN]&usehistory=y'
Feacth	efetch.fcgi	WebEnv='2743NCID'&query_key=1&rettype=fasta'

```
esearch -db nucleotide -query "viruses[orgn]" | efetch -format fasta
```

```
import entrezpy.conduit
# Create new Conduit instance
c = entrezpy.conduit.Conduit('email')
# Create empty Conduit pipeline
p = c.new_pipeline()
# Add search query and store its id
sid = p.add_search({'db':'nucleotide', 'term':'Viruses[orgn]'})
p.add_fetch(dependency=sid)
c.run(p)
```

Customizing

The analyzer parameter allows to use custom data analyzers as callbacks based in entrezpy.base.analyzer.EutilsAnalyzer. These need to be implemented by inheriting the base classes and adjusting virtual methods. run returns the analyzer.

```
import entrezpy.conduit
# Create new Conduit instance
c = entrezpy.conduit.Conduit('email')
# Create empty Conduit pipeline
p = c.new_pipeline()
# Add search query and store its id
sid = p.add_search({'db':'protein', 'term
    ':'APY22758.1 OR ABU40994.1'})
# Link search to taxonomy database
lid = p.add_link({'db':'taxonomy'},
    dependency=sid)
# Fetch taxa information in JSON
p.add_fetch({'retmode':'json', 'rettype':
    'docsum'}, dependency=lid, analyzer=
    DocsumAnalzyer())
r = c.run(p).get_result()
```

Features

- Supports NCBI API keys
- Multithreading support
- Pure Python, no external dependencies
- Connection error handling
- Caching and retrieving previous results
- Fully documented code

Why use entrezpy?

Versatile

- The modular design allows to design new and highly specific analyzer for specific datasets without the need to adjust the request process.
- Analyze and process the data as soon as it has been retrieved and configure follow-up queries on-the-fly.

Automated handling of NCBI limits

- entrezpy automatically configures itself to retrieve large datasets according to the implemented E-Utility function and limits enforced by NCBI.

Control every E-Utils parameter

entrezpy allows to configure every E-Utils parameter.

Supported E-Utils

- Esearch
- Efetch
- Elink
- Epost
- Esummary

References

0 : https://doi.org/10.1093/nar/gkw1071
1 : https://www.ncbi.nlm.nih.gov/books/NBK25497