

Programming for Biomedical Informatics

Lecture 6

“Data Integration & Summary Analysis”

<https://github.com/biomedical-informatics/pbi>

<https://biomedical-informatics.github.io/pbi-home/>

Ian Simpson

ian.simpson@ed.ac.uk

BioPython

- BioPython - <https://biopython.org/>
- Available via pip, conda, or source download
 - `pip install biopython`
 - `conda install conda-forge::biopython`
 - <http://biopython.org/DIST/biopython-1.84.tar.gz>
- The Biopython package provides a library **Bio.Entrez** that can be used to access the eUtils
- This is the simplest way to begin working with the eUtils using Python
- Once you are comfortable with the way the system (APIs) operate you may choose to simply use urllib (or similar) to code more flexibly with
- Biopython has excellent “cookbooks” which are code recipes to achieve common tasks



Accessing NCBI's Entrez databases

Entrez (<https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html>) is a data retrieval system that provides users access to NCBI's databases such as PubMed, GenBank, GEO, and more. You can access Entrez from a web browser to manually enter queries, or you can use Biopython's `Bio.Entrez` module for programmatic access to Entrez. The latter allows you for example to search PubMed or download GenBank records from within a Python script.

The `Bio.Entrez` module makes use of the Entrez Programming Utilities (also known as eUtils), which consist of eight tools that are described in detail on NCBI's page at <https://www.ncbi.nlm.nih.gov/books/NBK25501/>. Each of these tools corresponds to a specific function in the `Bio.Entrez` module, as described in the sections below. This module ensures that the correct URL is used for the queries, and that NCBI's guidelines for responses are being followed.

The output returned by the Entrez Programming Utilities is typically in XML format. If you want to process this output, you have several options:

1. Use `Bio.Entrez`'s parser to parse the XML output into a Python object;
2. Use one of the XML parsers available in Python's standard library;
3. Read the XML output as raw text, and parse it by string searching and manipulation.

Accessing NCBI's Entrez databases

Entrez Guidelines

EInfo: Obtaining information about the Entrez databases

ESearch: Searching the Entrez databases

EPost: Uploading a list of identifiers

ESummary: Retrieving summaries from primary IDs

EFetch: Downloading full records from Entrez

ELink: Searching for related items in NCBI Entrez

EGQuery: Global Query - counts for search terms

ESpell: Obtaining spelling suggestions

Parsing huge Entrez XML files

HTML escape characters

⊕ Handling errors

⊕ Specialized parsers

Using a proxy

⊖ Examples

PubMed and Medline

Searching, downloading, and parsing Entrez Nucleotide records

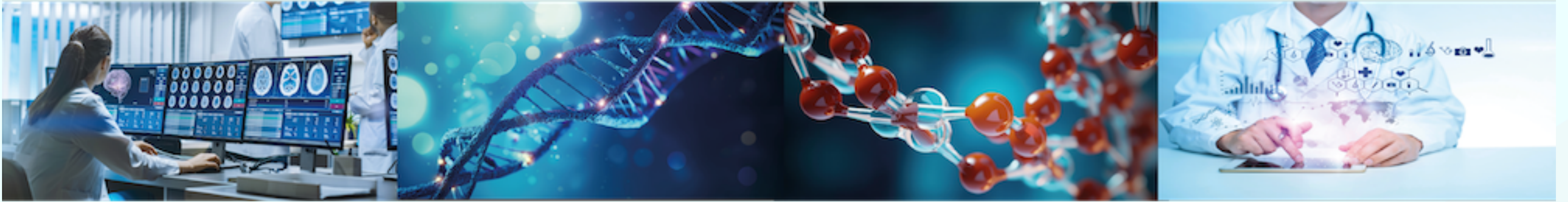
Searching, downloading, and parsing GenBank records

Finding the lineage of an organism

⊕ Using the history and WebEnv

The NCBI-NLM eUtilities

- **EInfo (database statistics)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi
 - Provides the number of records indexed in each field of a given database, the date of the last update of the database, and the available links from the database to other Entrez databases.
- **ESearch (text searches)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi
 - Responds to a text query with the list of matching UIDs in a given database (for later use in ESummary, EFetch or ELink), along with the term translations of the query.
- **EPost (UID uploads)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/epost.fcgi
 - Accepts a list of UIDs from a given database, stores the set on the History Server, and responds with a query key and web environment for the uploaded dataset.
- **ESummary (document summary downloads)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi
 - Responds to a list of UIDs from a given database with the corresponding document summaries.
- **EFetch (data record downloads)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi
 - Responds to a list of UIDs in a given database with the corresponding data records in a specified format.
- **ELink (Entrez links)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi
 - Responds to a list of UIDs in a given database with either a list of related UIDs (and relevancy scores) in the same database or a list of linked UIDs in another Entrez database; checks for the existence of a specified link from a list of one or more UIDs; creates a hyperlink to the primary LinkOut provider for a specific UID and database, or lists LinkOut URLs and attributes for multiple UIDs.
- **EGQuery (global query)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/egquery.fcgi
 - Responds to a text query with the number of records matching the query in each Entrez database.
- **ESpell (spelling suggestions)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/espell.fcgi
 - Retrieves spelling suggestions for a text query in a given database.
- **ECitMatch (batch citation searching in PubMed)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/ecitmatch.cgi
 - Retrieves PubMed IDs (PMIDs) corresponding to a set of input citation strings.



Programming for Biomedical Informatics

Next Lecture
“Biomedical Evidence”

Ask Questions on the Piazza Discussion Board

Background