NCBI Minute:  APIs @NCBI

• Closed captioning: www.captionedtext.com and enter 3341486
• The recording, will be on our YouTube channel in the Webinars playlist: <youtube>/user/NCBINLM/playlists
• Use the Questions Pod to ask questions when you think of them. Don’t wait until the end.
• Answers to all questions will be available after the webinar and linked to our Webinars page: <ncbi>/home/coursesandwebinars.shtml
• Materials / Q&A: https://go.usa.gov/xRQDf

Materials / Q&A:
- <youtube> = www.youtube.com

9/06/2017

NCBI Minute:  APIs @NCBI

This is an introduction to NCBI APIs for people new to APIs and for everyone, including the already API-savvy, to give you an idea of the diversity and breadth of APIs available at NCBI.... ....and where to look for more information.

9/06/2017
This webinar is just an introduction....

If you have specific questions about how to use a particular API or problems with a script you have already developed, please send a message to the HelpDesk (info@ncbi.nlm.nih.gov) and we’ll work with you directly to help you.

After this webinar, we’ll be sending you a Survey to ask you for some feedback:

1) Which specific API topics would you like to learn more about for future webinars?
2) What functionality or access to specific types of data would you like to have added to our existing repertoire of APIs?
3) What additional suggestions do you have to help us improve our APIs?

We’d love to hear your thoughts, so that we can improve our services for you!

Using APIs (Application Program Interfaces)

What is an API? What can it do for you?

- URLs are web addresses and are composed of two main components:
  a protocol identifier (ex: HTTP/HTTPS or FTP/SFTP) and a resource name.

- APIs are URLs (with specific syntax) designed for developers to use within their scripts/applications/programs to interact with servers across the internet.

- NCBI's APIs have been developed for 3rd-party developers to access, process and retrieve data from our backend servers and are now also used within our own internal systems, for example: PubChem, EndNote, BioGPS, Molecules, and more...

Think of APIs as an action-based sentence - with specific grammatical parts.
I’ll show you how this works in just a minute.....
Today...

- The NCBI site-wide API: Entrez Programming Utilities (EUtilities or EUtils)
- PubMed Central (PMC)’s ID Converter
- BLAST URL API
- PubChem’s PUG REST
- Entrez Direct (Edirect)

- A couple of others and where to find more information...

The NCBI site-wide API: Entrez Programming Utilities (EUtilities or EUtils)

A set of structured URL-based calls to access the NCBI Entrez search system and databases. These store information covering a diverse array of data including medical and scientific literature, nucleotide and protein sequences, gene records, gene expression, genotype and biological assay studies with data, chemical information, and three-dimensional molecular structures.

Examples of things you can do:
- Search PubMed with a term and retrieve PMIDs.
- Search MedGen with a SnoMedCT ID, then link over to GTR, and retrieve their Summary info.
- Post a list of Gene IDs and fetch their full records.
- Post a PubChem Compound CID and retrieve the SDF-formatted record (to show the 3D coordinates in the Molecules App).

Where to learn more?
https://www.ncbi.nlm.nih.gov/books/NBK25501/

YouTube Videos:
2014 (1 hour): https://www.youtube.com/watch?v=ICFVVeSp30o

EUtils Base URL:
Entrez Programming Utilities Help

https://www.ncbi.nlm.nih.gov/books/NBK25501/

Introduction to the E-utils

* Visit E-utils Introduction
* Please see the Release Notes for details and changes.

The Entrez Programming Utilities (E-utils) are a set of eight server-side programs that provide a stable interface to the Entrez query and database systems at the National Center for Biotechnology Information (NCBI). The E-utils are a fixed URL system that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and return the requested data. The E-utils are therefore the structured interface to the Entrez system, which currently includes 30 databases covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature.

>40 Databases, >1 billion records:

- Biomedical literature (PubMed, PMC...)
- Health-related (MedGen, GTR, ClinVar...)
- Sequences (Nucleotide, SRA, Protein...)
- Genomic data (Genome, dbGaP, dbSNP...)
- Gene information (Gene, HomoloGene, BioSystems...)
- Gene expression (GEO, UniGene...)
- 3D Biomolecule Structure/function (Structure, CDD...)
- Chemicals & Biological Assays (PubChem)
- PubChem Substances (related substance and chemical information)
- Entrez Gene (gene structure, information and links to genes, proteins and genes)
**VERBS (do what?) AND NOUN (with what?)**


- **Esearch.fcgi**
  - set of IDs
- **Epost.fcgi**
  - temporary storage of a set of IDs in “Entrez history”
- **Esummary.fcgi**
  - Record summaries
- **Efetch.fcgi**
  - Full data records
- **Elink.fcgi**
  - set of IDs in dbB

---

**EUtils API Examples**

Getting ESummary information for Gene with a Gene ID of 3269 (Human Histamine Receptor H1)


**BASE URL CALL**

Perform this **action** on this **database** with this **term**

*(return the information in this **format**)*
Getting ESummary information for Gene with a Gene ID of 3269 (Human Histamine Receptor H1)


Linking to Find PubMed Record IDs (PMIDs) for Gene ID 3269

/elink.fcgi?dbfrom=gene&id=3269&dbto=pubmed

EUtils API Examples
Linking to Find PubMed Record IDs (PMIDs) for Gene ID 3269

/elink.fcgi?dbfrom=gene&id=3269&dbto=pubmed

The XML file does not appear to have any style information associated with it. The document tree is shown below:

```
<eLinkResult>
  <LinkList>
    <IdList>
      <Id>3269</Id>
    </IdList>
  </LinkList>
</eLinkResult>
```

Getting ESummary information for Gene with a Gene ID of 3269 (Human Histamine Receptor H1)


Linking to Find PubMed Record IDs (PMIDs) for Gene ID 3269

/elink.fcgi?dbfrom=gene&id=3269&dbto=pubmed

Getting ESummary or Efetch information for PMID 24236486 (which was linked to Gene ID 3269)

/esummary.fcgi?db=pubmed&id=24236486 (retype=xml is the default)
/efetch.fcgi?db=pubmed&id=24236486&rettype=medline
/efetch.fcgi?db=pubmed&id=24236486&rettype=xml
Getting ESummary or Efetch information for PMID 24236486

GET ESummary:
/summary.fcgi?db=pubmed&id=24236486 (which was linked to Gene ID 3269)
/efetch.fcgi?db=pubmed&id=24236486&rettype=medline
/efetch.fcgi?db=pubmed&id=24236486&rettype=xml

What they are REALLY useful for: EUtil Pipelines

"Hereditary Hemochromatosis"
35400008 [SNOMED CT CUI]

Information about GTR-listed tests, including who to contact

3 E-UTILITY URL CALLS
EUtils API USAGE GUIDELINES

Limit URL requests to less than 3 per second.
Limit large jobs to either weekends or weekdays - between 9PM and 5AM EST

Failure to comply with this policy may result in an IP address being blocked from accessing NCBI.


Currently, you can register EUtil API calls that you are using by sending an Email to eutilities@ncbi.nlm.nih.gov with the name of your tool and an Email address. If there are any problems with your script, we will contact you using the registered Email before we block you.

Anticipated in the Spring of 2018, we will be initiating an EUtilities API Registration system to enable you to sign up to get an API Key to incorporate in your URL calls. This will enable us to better manage our server access so that we can accommodate the large and still increasing amounts of scripted access.

BONUS: Including the API Key will enable YOU to increase your EUtil URL call rates over the current standards!

A PMC Tool API: ID Converter

This service enables you to find the relevant pairs of these IDs for record listed in PubMed Central (PMC):

- **PMCID** (PMC ID)
- **PMID** (PubMed ID)
- **Manuscript ID** (available only for articles that came in through a manuscript submission system, e.g., NIHMS, Europe PMC, PMC Canada)
- **DOI** (Digital Object Identifier, if the article has one)

CAVEAT: Works for all records that are listed in PMC..... not all in PubMed, though.

Examples of things you can do:

- Finding the PMCID or PMID for a NIHMS ID
- Correlating PMIDs with PMCID(s) (free full text articles!)
- Finding a DOI for a PMCID or PMID.

Where to learn more?

https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/

PMC ID Converter Base-URL:

https://www.ncbi.nlm.nih.gov/pmc/utils/idconv/v1.0/?
PMC Developer Resources & the ID Converter API

https://www.ncbi.nlm.nih.gov/pmc/tools/developers/

PMC ID Converter API Example

Find the other IDs for the NIHMS Submission ID 311352

https://www.ncbi.nlm.nih.gov/pmc/utils/idconv/v1.0/?ids=NIHMS311352

BASE URL CALL  Perform this action on this database with this term
(return the information in this format!)
Find the other IDs for the NIHMS Submission ID 311352

https://www.ncbi.nlm.nih.gov/pmc/utils/idconv/v1.0/?ids=NIHMS311352

You can retrieve the results in several formats xml (default), html, json, or "csv".

https://www.ncbi.nlm.nih.gov/pmc/utils/idconv/v1.0/?ids=NIHMS311352&format=csv

**BASE URL CALL**  Perform this **action** on this **database** with this term
(return the information in **this format**)!
You can retrieve the results in several formats: xml (default), html, json, or "csv".

https://www.ncbi.nlm.nih.gov/pmc/utils/idconv/v1.0/?ids=NIHMS311352
https://www.ncbi.nlm.nih.gov/pmc/utils/idconv/v1.0/?ids=NIHMS311352&format=csv

PMC ID Converter API **USAGE GUIDELINES**

Each lookup request can contain ≥ 1 identifiers, but **do not mix the types of IDs in a single request.**

≤200 IDs can be searched in a single request.

For a larger number of IDs, split your list up!

*(If you have a really large set of IDs you’d like to match, consider downloading this file: [ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/PMC-ids.csv.gz]*)
BLAST’s URL API

The NCBI-BLAST Common URL API allows you to run sequence similarity searches remotely on the NCBI web servers (Web BLAST).

This is a 3 step process:
- **Submission** of the job to the servers,
- **Checking** to see if the job is done,
- **Retrieval** of the results once the job is done.

A set of parameters is available to incorporate in the calls to tailor your BLAST search and specify the format for retrieval of the search results. These are described on the NCBI BLAST GitHub page.

Examples of things you can do:
- Search for an identical or near identical sequence to find out more about it.
- Search for similar sequences in other organisms to identify potential homologs.
- Search for where the sequence may be match a region on a genome or chromosome to enable annotation for where it is encoded.
- and more...

Where to learn more?
- [https://ncbi.github.io/blast-cloud/dev/api.html](https://ncbi.github.io/blast-cloud/dev/api.html)

BLAST URL API Base:

---

**BLAST URL API**

3-step set for BLAST URL API Calls

1) **Submit** your query sequences/IDs for comparison/calculation & get back an RID

2) **Check** on the progress of your BLAST search with the RID

3) **When the Results are ready** – **retrieve** them in your choice of format.

---

1) Compare the sequence of human acetylcholinesterase with those of known structures in the PDB database.

   Submit an ID to a particular BLASTp calculation & get back an RID (Request ID).


   • **Look for the RID value for use in step 2!**
1) Compare the sequence of human acetylcholinesterase with those of known structures in the PDB database. Submit an ID to a particular BLASTp calculation & get back an RID (Request ID).


• Look for the RID value for use in step 2!

2) Check on the progress of your BLAST search with the RID

/Blast.cgi?CMD=Get&RID=UY11SZ3W016

• Look for a “READY” status to indicate the BLAST search has completed and results are available for access.
3-step set for BLAST URL API Calls

1) Compare the sequence of human acetylcholinesterase with those of known structures in the PDB database.
   Submit an ID to a particular BLASTp calculation & get back an RID (Request ID).
   • Look for the RID value for use in step 2!

2) Check on the progress of your BLAST search with the RID
   /Blast.cgi?CMD=Get&RID=UY11SZ3W016
   • Look for a “READY” status to indicate the BLAST search has completed and results are available for access

3) When the Results are ready – retrieve them in your choice of format.
   /Blast.cgi?CMD=Get&VIEW_RESULTS=FromRes&RID=UY11SZ3W016&FORMAT_TYPE=TEXT
   /Blast.cgi?CMD=Get&VIEW_RESULTS=FromRes&RID=UY11SZ3W016&FORMAT_TYPE=XML2_S
BLAST URL API USAGE GUIDELINES

1) Submit your query sequences/IDs for comparison/calculation & get back an RID

Please submit at least 3 seconds apart.

2) Check on the progress of your BLAST search with the RID

You need to let the BLAST Search do its work!
So delay the initial check for at least 1 minute after submission
And at least 30 seconds for each subsequent check.

3) When the Results are ready – retrieve them in your choice of format.

Finally, please limit your Web BLAST searches to less than 100 per day.
You need to do more than that? Consider using the “BLAST in the Cloud”.

PubChem’s Specialty API: PUG REST

A set of structured URL-based calls to access and retrieve information from the PubChem set of databases (Compound, Substance & BioAssay).

There are many ways to access PubChem’s servers, PUG REST has two facets...

PUG VIEW REST – enables the downloading of PubChem records sections from the servers
PUG REST – enabling search and retrieval of pieces of information from backend servers

Examples of things you can do:

- Search for chemical synonyms (other things they are called: nicknames, IUPAC, etc.)
- Find some chemical properties for a particular chemical.
- Get whole sections of PubChem Compound records (such as Pharmacology or Toxicology information).
- See the 2D structure for a chemical.

Where to learn more?

PubChem PUG REST Base URLs:
pug_view/data/
PubChem’s PUG REST

PUG REST Search/Data Retrieval Calls

PUG_VIEW – Using a CID, you can retrieve whole sections of a PubChem Compound record.
PUG REST Search/Data Retrieval Calls

PUG_VIEW – Using a CID, you can retrieve whole sections of a PubChem Compound record.


CAVEAT: This method only works with CIDs (similar to the EUtil “Post” function), so how do I find the CIDs?... (you can use the EUtil ESearch or PUG REST)

PUG - Searching with a "name" to retrieving CID(s).


Searching with a "name" to retrieve a chemical property or even several!

/compound/name/benadryl//property/MolecularFormula/TXT
/compound/name/benadryl//property/MolecularFormula,MolecularWeight,InChIKey/XML

Searching with a "name" to retrieving a picture (which happens to be a 2D structure!)

/compound/name/benadryl//PNG
PUG - Searching with a “name” to retrieving CID(s).


Searching with a “name” to retrieve a chemical property or even several!

/compound/name/benadryl/property/MolecularFormula/TXT
/compound/name/benadryl/property/MolecularFormula,MolecularWeight,InChIKey/XML

Searching with a “name” to retrieving a picture (which happens to be a 2D structure!)

/compound/name/benadryl/PNG

PubChem’s PUG REST USAGE GUIDELINES

≤5 requests per second

if you want a large amount of information, we have download files including specialty “Extras” ones:
Entrez Direct (EDirect) – an alternative to Eutils for those who need the data, not the URL....

Entrez Direct (EDirect) provides UNIX command-line access to the data in NCBI's suite of interconnected databases. Multi-step pipelines can be quickly concatenated linking record searching, retrieval, data parsing, and results formatting.

INSTALLATION OF A FILE
CONTAINING ASUITE OF APPLICATIONS IS REQUIRED
UNIX and Macintosh computers with Perl installed can run the installer and quickly start using EDirect.
For Windows/PCs, you’ll need to first install the Cygwin UNIX-emulation environment and a couple of additional applications, and then the installation file.

Examples of things you can do:
• Use line commands to automatically generate EUtils APIs, then parse and reformat the data!
• Download only parts of records
• Extract and concatenate data in specific fields
• Create a custom data table
• and lots more....

Where to learn more?
https://www.ncbi.nlm.nih.gov/books/NBK179288/
EDirect Cookbook:
https://github.com/NCBI-Hackathons/EDirectCookbook

Training Programs:
Webinar & Courses (NLM, NCBI – 4 hour at NIH 9/18)
4-Part 10-15min YouTube Videos (2015):
https://www.youtube.com/watch?v=BLnYW33Mtb0

EDirect Example
Create a table for genes located between position 44.9M & 45.9M on human chromosome 19.
(With this information it’ll be easy to download the sequences!)

> -sort Chromosome | \
> efetch -format docsum | \
> xtract -head ‘ Gene ID\tSymbol\tAccession\tChr. Start\tChr. Stop’ \
> -pattern DocumentSummary -element Id Name -block GenomicInfoType \
> -element ChrAccVer ChrStart ChrStop

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Symbol</th>
<th>Accession</th>
<th>Chr. Start</th>
<th>Chr. Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10452</td>
<td>TOMM40</td>
<td>NC_000019.10</td>
<td>4491219</td>
<td>44903688</td>
</tr>
<tr>
<td>348</td>
<td>APOE</td>
<td>NC_000019.10</td>
<td>44905748</td>
<td>44909394</td>
</tr>
<tr>
<td>341</td>
<td>AP0C1</td>
<td>NC_000019.10</td>
<td>44914246</td>
<td>44919348</td>
</tr>
<tr>
<td>346</td>
<td>AP0C4</td>
<td>NC_000019.10</td>
<td>44942237</td>
<td>44945495</td>
</tr>
<tr>
<td>100533990</td>
<td>AP0C4–AP0C2</td>
<td>NC_000019.10</td>
<td>44942237</td>
<td>44949564</td>
</tr>
<tr>
<td>344</td>
<td>AP0C2</td>
<td>NC_000019.10</td>
<td>44945981</td>
<td>44949564</td>
</tr>
<tr>
<td>79090</td>
<td>TRAPPC6A</td>
<td>NC_000019.10</td>
<td>45178301</td>
<td>45162927</td>
</tr>
<tr>
<td>368552</td>
<td>BLOC1S3</td>
<td>NC_000019.10</td>
<td>45178744</td>
<td>45202714</td>
</tr>
<tr>
<td>90332</td>
<td>EXOC3L2</td>
<td>NC_000019.10</td>
<td>45234210</td>
<td>45212620</td>
</tr>
<tr>
<td>100420902</td>
<td>LOC10428902</td>
<td>NC_000019.10</td>
<td>45234808</td>
<td>45234194</td>
</tr>
<tr>
<td>2068</td>
<td>ERC2</td>
<td>NC_000019.10</td>
<td>45370646</td>
<td>45349836</td>
</tr>
<tr>
<td>10849</td>
<td>CD3EAP</td>
<td>NC_000019.10</td>
<td>45406128</td>
<td>45410765</td>
</tr>
</tbody>
</table>
Two examples of the many others available...

Variation Services (Sequence Position Deletion Insertion, SPDI)
https://api.ncbi.nlm.nih.gov/variation/v0/
- Converts genetic variation data formats (HGVS notation, VCF files)
- Searches with your own variant data to match them up with existing variants and retrieves their RefSNP information.

Genome Remapping Service (ReMap)
https://www.ncbi.nlm.nih.gov/genome/tools/remap/docs/api
- Maps genomic annotation data from one coordinate system to another (a.k.a. 'liftover').
  Three variations exist:
  - Assembly-Assembly ReMap - from one genomic assembly to another
  - Clinical ReMap - from an assembly sequence to a RefSeqGene sequence
  - and Alt Loci ReMap - for the mapping of features between the Primary assembly and the Assembly Patches and Alternate Loci available for GRC genomic assemblies

*ReMap provides a Perl script to download for use of the integrated API calls.*
Argh!!! This is so much!
Where can I find the information?

Welcome to NCBI

The National Center for Biotechnology Information (NCBI) provides access to biomedical and genomic data resources.

About the NCBI Mission

NCBI provides a variety of resources that allow researchers, clinicians, and the general public to discover, analyze, and apply biomedical data.

Develop

Use NCBI APIs and code libraries to build applications.

Submit

Deposit data or manuscripts into NCBI databases.

APIs

NCBI provides several public APIs that allow programmatic access to many databases and tools.

Entrez Programming Utilities (E-utils)
The E-utils are the public API to the NCBI Entrez system and allow access to all Entrez databases including PubMed, GenBank, and PubMed Central. The E-utils are a suite of eight commands that accept a single URL syntax for search, list, and retrieval operations. A common package named Entrez Direct consists of several executables that allow the E-utils to be invoked directly from a Unix-style command line.

BLAST URL API
The BLAST API allows developers to submit BLAST searches for processing at NCBI or cloud service provider(s) using HTTPS. The API can then check the status of submitted searches and retrieve results when ready in a single format.

PubMed Central (PMC) APIs
PMC provides several APIs that provide programmatic access to various services that deal with PMC literature content, including the validation tests, Open Access web services, and an E-utils to provide programmatic access to other NCBI databases.

Wrap Up

As you may imagine, with all of the many resources developed at NCBI, there are a lot of diverse APIs that are available. Other NLM groups have APIs too:


At NCBI, we are in the process of cataloging all of the NCBI-specific APIs and will be adding them to the Develop page shortly.

Check in on the NCBI Develop site - YOUR NCBI SCRIPTING HOME!


Webinars & Courses are available to help! (ex: EDirect course at NIH - Sept. 18th)

Hackathons - help scripters use APIs to create specific useful tools!

(ex: upcoming Hackathon in Pittsburgh - Sept. 25-27)

Keep an eye on the NCBI Lean site for more information

- Direct URLs & Places to Learn More -

Entrez Programming Utilities (EUtilities, EUtils)

PMC ID Converter

BLAST URL API

PubChem’s PUG REST

Entrez Direct (EDirect)


Questions?

For help with NCBI resources: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)