



The Kalmanovitz Library and
The Center for Knowledge Management

An Introduction to NCBI E-Utilities

Gilberto da Gente
Bioinformatics Specialist
May 20th 2008

University of California, San Francisco



UCSF

Entrez Query System at NCBI

NCBI
National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search for

SITE MAP
Alphabetical
Resource
About NCBI
An introduction
NCBI
GenBank
Sequence
submission
and software
Literature
databases
PubMed,
Books, and
Central

What does NCBI do?
founded in 1988 as a national resource for
biology information, NCBI creates
bases, conducts research in
nal biology, develops software tools
g genome data, and disseminates
information - all for the better
ing of molecular processes
man health and disease. [More...](#)

Hot Spots

- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools

Genome Association
Whole Genome Association (WGA) resource
searchers with access to genotype and
phenotype information that will help
the link between genes and disease. For

Entrez Functions

- Search one or all databases.
- Generate brief “document summaries” for a list of records.
- Link from one list of records to another.
- Perform Boolean operations on lists of records.
- Format records for display and download.

Entrez Transactions

- Each record in an Entrez database is assigned an integer called a **UID**, or “**unique identifier**” .
- Entrez transactions are performed on lists of UIDs.
- Transactions include Boolean operations and the tracking of links within and between database records.

DocSums

- Brief summaries of database records are generated quickly on frontend servers.
- Full records are retrieved from backend machines.

Search History

- Separate search history is maintained for each database.
- Previous searches can be recalled and combined using a query key and a cookie, called a "WebEnv".
- Available on the Web under the 'History Tab'

Eutilities

- A set of eight server-side programs.
- Support a uniform URL syntax.
- Translate a standard set of URL-encoded input parameters for the array of programs comprising the Entrez system.

Entrez Functions and EUtils

- Global Query: `egquery.fcgi`
- Searches: `esearch.fcgi`
- DocSums: `esummary.fcgi`
- Information: `einfo.fcgi`
- Downloads: `efetch.fcgi`
- Links: `elink.fcgi`
- Uploads: `epost.fcgi`

Spelling: `espell.fcgi` (only used for pubmed)

URL Syntax

Location on the web where e-utility web services are located

Base URL

`http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eutil.fcgi?`

Eutility

Specific e-utility web service request

URL Parameters

BASE/

esearch.fcgi?

db=nucleotide&term=mouse[orgn]

E-Utility

Parameters

esearch.fcgi?
einfo.fcgi?
efetch.fcgi?
esummary.fcgi?
egquery.fcgi?
epost.fcgi?
elink.fcgi?

db = nucleotide
term = mouse[orgn]

"&" symbol used to separate parameters

- We need to know the following for each eutility call:
1. What parameters are available
 2. What values they accept

The Entrez System

Entrez
Core
Engine

- Displays Document Summaries for each matching UID

- Finds Unique Identifiers (UIDs) that match a text query

The Entrez Core

EGQuery

- Finds Unique Identifiers (UIDs) that match a text query

ESearch

ESummary

- Displays Document Summaries for each matching UID

Text Query



EGQuery



Count of UIDs

Text Query



ESearch



Set of UIDs

Set of UIDs



ESummary



DocSums

These work for ALL DATABASES

EGQuery

Performs an Entrez search across all databases

Why use it?

- To find the number of records matching a text query

BASE/

egquery.fcgi?

&term=mouse [orgn]

INPUT

term

Entrez text query

OUTPUT

XML

Number of records matching the query in each database

EGQuery Output

```
<ResultItem>
  <DbName>nucleotide</DbName>
  <MenuName>Nucleotide</MenuName>
  <Count>6305267</Count>
  <Status>Ok</Status>
</ResultItem>

<ResultItem>
  <DbName>protein</DbName>
  <MenuName>Protein</MenuName>
  <Count>129743</Count>
  <Status>Ok</Status>
</ResultItem>

<ResultItem>
  <DbName>genome</DbName>
  <MenuName>Genome</MenuName>
  <Count>1</Count>
  <Status>Ok</Status>
</ResultItem>

<ResultItem>
  <DbName>structure</DbName>
  <MenuName>Structure</MenuName>
  <Count>1144</Count>
  <Status>Ok</Status>
</ResultItem>
```

```
<Result>

  <Term>mouse[orgn] </Term>

  <eGQueryResult>

    <ResultItem>
      <DbName>pubmed</DbName>
      <MenuName>PubMed</MenuName>
      <Count>0</Count>
      <Status>Term or Database is not found</Status>
    </ResultItem>

    <ResultItem>
      <DbName>pmc</DbName>
      <MenuName>PMC</MenuName>
      <Count>870</Count>
      <Status>Ok</Status>
    </ResultItem>

    <ResultItem>
      <DbName>journals</DbName>
      <MenuName>Journals</MenuName>
      <Count>0</Count>
      <Status>Term or Database is not found</Status>
    </ResultItem>

    <ResultItem>
      <DbName>mesh</DbName>
      <MenuName>MeSH</MenuName>
      <Count>0</Count>
      <Status>Term or Database is not found</Status>
    </ResultItem>
```

ESearch

Performs an Entrez search in one specified database

BASE/

esearch.fcgi?

db=nuccore&term=mouse [orgn]

INPUT

db

Entrez database to search

term

Entrez text query

OUTPUT

XML

Total number of records matching the query

Partial list of matching UIDs

Term translations

ESearch Output

Total number of records found

retmax

retstart

first record = retstart

Matching UIDs

quantity = retmax

```
<eSearchResult>  
<Count>6305267</Count>  
<RetMax>20</RetMax>  
<RetStart>0</RetStart>  
<IdList>  
<Id>49619226</Id>  
<Id>49615287</Id>  
<Id>49615286</Id>  
<Id>49615285</Id>  
<Id>49615284</Id>  
<Id>49615283</Id>  
<Id>49615282</Id>  
<Id>49615281</Id>  
<Id>49615280</Id>  
<Id>49615279</Id>  
<Id>49615278</Id>  
<Id>49615277</Id>  
<Id>49615276</Id>  
<Id>49615275</Id>  
<Id>49615274</Id>  
<Id>49615273</Id>  
<Id>49615272</Id>  
<Id>49615271</Id>  
<Id>49615270</Id>  
<Id>49615269</Id>  
</IdList>
```

Retrieval Parameters

These work for ESearch, ESummary, and EFetch

retstart

First record to retrieve from UID set
(default = 0)

retmax

Number of records to retrieve from
UID set

(84, 23, 19, 55, 20, 96, 73)

&retmax=4



(84, 23, 19, 55)

&retstart=2&retmax=4



(19, 55, 20, 96)

ESearch Output

Term translations: Equivalent to the Details link on the web

```
<TranslationSet>
  <Translation>
    <From>mouse%5Borgn%5D</From>
    <To>%22Mus+musculus%22%5BOrganism%5D</To>
  </Translation>
</TranslationSet>
<TranslationStack>
  <TermSet>
    <Term>"Mus musculus"[Organism]</Term>
    <Field>Organism</Field>
    <Count>6305267</Count>
    <Explode>Y</Explode>
  </TermSet>
</TranslationStack>
</eSearchResult>
```

mouse[orgn]

"Mus musculus" [Organism]

ESummary

Retrieves Document Summaries matching a set of UIDs

Why use it?

- It's fast and can download large sets with one URL
- If EFetch does not support your database

BASE/

esummary.fcgi?

db=nucleotide&id=49619226,49615287

id

db

INPUT

Set of UIDs

Entrez database to search

OUTPUT

XML

DocSums, often with more data than web Entrez provides

ESummary Output

```
<eSummaryResult>
<DocSum>
  <Id>49619226</Id>
  <Item Name="Caption" Type="String">NM_008496</Item>
  <Item Name="Title" Type="String">Mus musculus lectin, galactose binding,
  <Item Name="Extra" Type="String">gi|49619226|ref|NM_008496.4|</Item>
  <Item Name="Gi" Type="Integer">49619226</Item>
  <Item Name="CreateDate" Type="String">2000/01/04</Item>
  <Item Name="UpdateDate" Type="String">2004/07/02</Item>
  <Item Name="Flags" Type="Integer">0</Item>
  <Item Name="TaxId" Type="Integer">10090</Item>
</DocSum>
</eSummaryResult>
```

❑ 1: [NM_008496](#)

Mus musculus lectin, galactose binding, soluble 7 (Lgals7), mRNA
gi|49619226|ref|NM_008496.4|[49619226]

The Entrez System

- Provides formatted data records indexed by UID to the Core Engine
- Provides Entrez links for each UID

Entrez
Data
Bases

Entrez Databases

EInfo

- Provides general information about a specific database

EFetch

- Provides formatted data records indexed by UID

ELink

- Provides Entrez links for each UID

The function of these depends on the chosen Entrez database

Entrez Database

EInfo

Database
Statistics

set of UIDs

EFetch

Formatted Data

set of UIDs in db A

ELink

set of UIDs in db B

EInfo

Provides indexing fields and overall statistics for an Entrez database

Why use it?

- To get a list of all indexing fields in a database
- To find all possible links to and from a database

BASE/

einfo.fcgi?

db=nucleotide

INPUT

db

Entrez database to search

OUTPUT

XML

- General indexing statistics
- List of available Entrez links for the database
- List of indexing fields and counts of records

EInfo Output: Indexing Fields

```
<eInfoResult>
  <DbInfo>
    <DbName>nucleotide</DbName>
    <MenuName>Nucleotide</MenuName>
    <Description>Nucleotide sequence record</Description>
    <Count>42526878</Count>
    <LastUpdate>2004/07/03 09:13</LastUpdate>
    <FieldList>
      <Field>
        <Name>ALL</Name>
        <Description>All terms from all searchable fields</Description>
        <TermCount>137604309</TermCount>
        <IsDate>N</IsDate>
        <IsNumerical>N</IsNumerical>
        <SingleToken>N</SingleToken>
        <Hierarchy>N</Hierarchy>
      </Field>
      <Field>
        <Name>UID</Name>
        <Description>Unique number assigned to publication</Description>
        <TermCount>0</TermCount>
        <IsDate>N</IsDate>
        <IsNumerical>Y</IsNumerical>
        <SingleToken>Y</SingleToken>
        <Hierarchy>N</Hierarchy>
      </Field>
      <Field>
        <Name>FILT</Name>
        <Description>Limits the records</Description>
        <TermCount>93</TermCount>
        <IsDate>N</IsDate>
        <IsNumerical>N</IsNumerical>
        <SingleToken>Y</SingleToken>
        <Hierarchy>N</Hierarchy>
      </Field>
    </FieldList>
  </DbInfo>
</eInfoResult>
```

Info Output: Links

```
<LinkList>
  <Link>
    <Name>nucleotide_comp_genome</Name>
    <Menu>Components to Genome</Menu>
    <Description>Genome(s) using this record as component</Description>
    <DbTo>genome</DbTo>
  </Link>
  <Link>
    <Name>nucleotide_comp_nucleotide</Name>
    <Menu>Assembly</Menu>
    <Description>Link to master record</Description>
    <DbTo>nucleotide</DbTo>
  </Link>
  <Link>
    <Name>nucleotide_gene</Name>
    <Menu>Gene Links</Menu>
    <Description>Link to related Genes</Description>
    <DbTo>gene</DbTo>
  </Link>
  <Link>
    <Name>nucleotide_genome</Name>
    <Menu>Assembly to Genome</Menu>
    <Description>Genome record containing nucleotide sequence</Description>
    <DbTo>genome</DbTo>
  </Link>
  <Link>
    <Name>nucleotide_geo</Name>
    <Menu>GEO Profile Links</Menu>
    <Description>GEO records associated with nucleotide record</Description>
    <DbTo>geo</DbTo>
  </Link>
</LinkList>
```

EFetch

Retrieves formatted data records matching a set of UIDs

Why use it?

- To download data records

BASE/

efetch.fcgi?

db=nucleotide&id=49619226,49615287

INPUT

db

Entrez database to search

id

Set of UIDs

OUTPUT

Varied

Formatted data records

Databases that Support EFetch

Literature

PubMed
Journals
PubMed Central
OMIM

Sequences

CoreNucleotide
CoreEST
CoreGSS
Protein
Genome
Popset
SNP

Other

Gene
Taxonomy

EFetch Formatting Parameters

rettype

Determines the type of data record returned (flat file, FASTA, EST, accession, etc.)

retmode

Determines the format (mode) of data record returned (text, HTML, XML)

Be warned!

- These settings are very dependent on the database
- These settings interact with one another
- Not all possible combinations are supported

The Entrez System

Entrez
History
Server

- Stores separate search histories for each Entrez database

The Entrez History Server

EPost

ESearch

ELink

Stores UID lists resulting from previous searches

The History Server represents the location of stored UID sets with two parameters:

WebEnv

A string specifying a cookie assigned by the History Server

query_key

An integer equivalent to the History number on the web

EPost

Stores a list of UIDs on the History Server

Why use it?

To upload a large file or set of UIDs

BASE/

epost.fcgi?

db=nucleotide&id=49619226,49615287

INPUT

db

Entrez database containing UIDs

id

List of UIDs

WebEnv

Pre-existing WebEnv to use

OUTPUT

XML

WebEnv

query_key

EPost Output

```
<ePostResult>  
  <QueryKey>25</QueryKey>  
  <WebEnv>0ey1GksCqOpNLoKz5VOZp_d09SRyaDM71cTGftM5vH2aqQKY1Kd0</WebEnv>  
</ePostResult>
```

query_key

WebEnv

UIDs can be passed to EPost in two ways:

1. In the &id parameter in a URL (limited to ~500)
2. In the &id parameter using the POST method (can post thousands)

Using method 2, UIDs can be posted from a file

Using ESearch to Post Results

BASE/ **esearch.fcgi?**

db=nucleotide&term=mouse[orgn]&usehistory=y

```
<eSearchResult>
  <Count>6305267</Count>
  <RetMax>20</RetMax>
  <RetStart>0</RetStart>
  <QueryKey>24</QueryKey>
  <WebEnv>0yGUFNCct4ny9fSm_WKy6HRW646NWQFmy3bIdHIMicaXFh_pEEGo</WebEnv>
  <IdList>
    <Id>49619226</Id>
    <Id>49615287</Id>
    <Id>49615286</Id>
    <Id>49615285</Id>
    <Id>49615284</Id>
    <Id>49615283</Id>
    <Id>49615282</Id>
    <Id>49615281</Id>
    <Id>49615280</Id>
    <Id>49615279</Id>
    <Id>49615278</Id>
    <Id>49615277</Id>
    <Id>49615276</Id>
    <Id>49615275</Id>
```

query_key

WebEnv

The Entrez System

