Bases de datos genomicas

Guia de trabajos practicos

Maestría en bioinformática y biología de sistemas

1. Conceptos generales

1.1. Formatos de archivo

Las bases de datos de secuencias biológicas permiten recuperar los datos que tienen almacenadas en archivos de diferentes formatos. En general, estos formatos son de texto para que puedan ser fácilmente procesados por el usuario. A continuación se detallan algunos de los formatos de archivos más comunes.

1.1.1. Fasta

Es, probablemente, el formato más común para almacenar secuencias biológicas. Tiene la ventaja de ser muy sencillo de manipular. La principal desventaja que presenta es que no almacena anotaciones sobre la secuencia. Ejemplo:

```
>gi|845635415|ref|NR_132280.1| Mus musculus polymerase (DNA-directed), delta 4
AGGGCAAAGGCCTGGATGTGCCTGAGGAGCCACCACATGTTATTGCAGGGGAAGACACCCAGTCCCTCAG
CCAGGAGGAAACAGAGCTGGAGCTGCTGAGGCAGTTT ...
```

1.1.2. Texto plano

Muchas bases de datos tienen ofrecen un formato propio para almacenar la información de los registros. Estos contiene además de la secuencia en si, anotaciones sobre ella. El formato propio de cada base de datos se verá más adelante.

La estructura de este formato está diseñada para que pueda ser fácilmente leido por el per humano, pero al mismo tiempo que pueda ser manipulado computacionalmente.

1.1.3. Tabular

Algunos servicios en las bases de datos dan como resultado datos sobre las secuencias en un formato de tabla de texto. Cada fila de la tabla corresponde con una línea de texto y las columnas se separan ususalmente con el carácter *TAB* (ASCII 09).

```
Entry   Entry name   Length  Gene names  Organism
P26313  GLYC_JUNIN   485     GPC GP-C    Junin arenavirus (JUNV)
P14239  NCAP_JUNIN   564     N           Junin arenavirus (JUNV)
Q6XQI4  L_JUNIN      2210    L           Junin arenavirus (JUNV)
Q6IVU5  Z_JUNIN      94      Z           Junin arenavirus (JUNV)
```

1.1.4. XML y similares

El formato XML (*eXtensible Markup Language*) es un formato de texto diseñado para almacenar y comunicar datos a través de Internet. Es fácilmente manipulable computacionalmente, y permite ser leido por el humano (aunque puede no resultar menos sencillo que otros formatos). La información está organizada como un árbol jerárquico. Un documento XML contiene un nodo raíz, que contiene otros nodos (en un nivel jerárquico inferior), estos nodos puede contiene otros a su vez, etc. Los nodos tiene nombres que los identifican (puede haber nodos con el mismo nombre), un contenido y atributos. Ejemplo:

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<padron_electoral>
  <provincia nombre="Buenos Aires">
    <persona>
      <nombre>Juan</nombre>
      <apellido>Pérez</apellido>
      <dni>98.765.432</dni>
    </persona>
    <persona>
      <nombre>Robero</nombre>
      <apellido>Pérez</apellido>
```

```
      <dni>89.567.234</dni>
    </persona>
  </provincia>
  <provincia nombre="La Pampa">
  ...
  </provincia>
</padron_electoral>
```

En el ejemplo, la primer línea es un encabezado que indica que comienza un documento XML. El nodo raíz es 'padron_electoral', este contiene nodos 'provincia'. Los nodos 'provincia' tiene un atributos que se llama 'nombre' y contiene nodos 'persona'. Es importante notar que los nodos se encierran entre los símbolos mayor y menor, y que para indicar el final del contenido de un nodo se agrega un símbolo de barra de división inmediatamente antes del nombre.

Las bases de datos biológicas que ofrecen obtener su información en formato XML, pero en ocasiones es posible hacerlo en otros formatos, que cumplen una función más o menos similar que el XML. Por ejemplo, JSON (*JavaScript Object Notation*) y RDF (*Resource Description Framework*).

1.2. Acceso programático

Todas las bases de datos biológicas ofrecen una acceso su información a través de internet por medio de un navegador y usualmente los resultados son mostrados en páginas interactivas con gran cantidad de recursos y enlaces a los que es posible acceder. Esto permite a un usuario poder acceder a una gran variedad de información de forma intuitiva. Esta interfaz *web* es muy útil para hacer análisis exploratorios pero no lo es sí se quiere automatizar un trabajo, o analizar un gran número de datos.

Las bases de datos más importantes suelen ofrencer otras formas de acceder a su contenido, que permiten la automatización y el trabajo con grandes volúmenes de datos. La opción más sencilla es el acceso directo a los datos crudos por *FTP* (*File transfer protocol*). Otra opción es proveer una interfaz REST (*Representational State Transfer*), para interactuar con la base de datos.

1.2.1. Acceso REST

El acceso REST es un estilo de diseño de *software* para sistemas distribuidos, como lo es internet. No es el objetivo de esta guía entrar en los detalles técnicos, sin embargo es importante saber que usando esta tecnología las bases de datos pueden ofrecer al usuarios capacidades similares de trabajo a las que tendrían si estuvieran usando un navegador, pero además, permite la automatización de tareas.

Para hacer una consulta REST a una base de datos, es necesario saber la URL base del servicio (que puede ser la misma dirección que la página web de la base de datos o no) y poder enviar datos al servidor. Para hacer esto último, hay dos métodos: GET y POST. Para algunos servicios se deberá usar el primero, en otros el segundo y depende de la decisión de quien haya diseñado el servicio.

Los datos que se transfieren al servidor son del tipo clave/valor. Los datos que se le pasan al servidor (valores) están identificados con una palabra clave.

Para hacer una consulta REST necesitamos un programa que nos permita hacer esto o escribir un pequeño programa en algún lenguage de programación que ofrezca esto.

1.2.2. Programas

Hay muchos programas que permiten hacer consultas REST.

Navegadores *web* Todos los navegadores pueden hacer consultas REST, sin embargo se pierde la capacidad de automatización.

wget Es un programa que funciona por línea de comandos. No permite usar el método POST para transferir datos al servidor. Está disponible para **Linux** y **Windows**.
Ejemplo:
`wget "http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=nuccore&term=coli" -O result.xml`
Donde, 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi' es la URL base. 'db=nuccore&term=coli' son los datos que se le pasan al servidor, 'db' es una clave y está asociada al valor 'nuccore', 'term' es otra clave y está asociada al valor 'coli'. El resulado de la consulta que devuelve el servidor se guarda en el archivo 'result.xml'.

curl Es un programa que funciona por línea de comandos. Permite hacer tanto consultas tipo GET como POST. Está disponible para Linux y **Windows**. Ejemplos:

```
curl "http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=nuccore&term=coli" -o result.xml
curl -d db=nuccore -d term=coli "http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi" -o result.xml
```

Donde, 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi' es la URL base. 'db=nuccore&term=coli' son los datos que se le pasan al servidor, 'db' es una clave y está asociada al valor 'nuccore', 'term' es otra clave y está asociada al valor 'coli'. El resulado de la consulta que devuelve el servidor se guarda en el archivo 'result.xml'.

1.2.3. Lenguages de programación

Es posible hacer consultas REST con prácticmente cualquier lenguage de programación moderno. Sin embargo, los más comunmente usados en el ámbito de la bioinformática son Perl, Python, Ruby y Java. Varias bases de datos ofrecen ejemplos particulares para estos lenguajes.

2. Bases de datos primarias

2.1. National Center of Biotechnology Information - NCBI

El NCBI ofrece un conjunto de herramientas para el acceso programático a sus bases de datos que se llama NCBI E-utilities y se basa en una interface REST. Se puede acceder a ellas con el sistema Entrez (*Global Query Cross-Database Search System*).

Cada una de estas herramientas, tiene una URL base particular y está destindada a ofrecer diferentes funciones. Es importante mencionar que Entrez, no solo ofrece acceso a base de datos de secuencias, sino de otro tipo, como por ejemplo bibliográfico, a través de *Pubmed*. Los servicios de Entrez que se discutirán son EInfo, ESearch, ESummary y EFetch. Estos no son los únicos servicios disponibles, pero son los más relevantes.

2.1.1. Formato del registro

El formato de texto plano de NCBI tiene tres secciones principales. El encabezado (Desde *LOCUS* hasta *JOURNAL*), una tabla de características anotadas en la secuencia (*FEATURES*) y la secuencia propiamente (*ORIGIN*). A continuación se muestra, como ejemplo, un registro en este formato.

```
LOCUS       AJ131281                348 bp    RNA     linear   VRL 10-DEC-1999
DEFINITION  Lymphocytic choriomeningitis virus z gene, partial, strain MX.
ACCESSION   AJ131281
VERSION     AJ131281.1  GI:3970751
KEYWORDS    RING finger protein; z gene.
SOURCE      Lymphocytic choriomeningitis mammarenavirus (LCMV)
  ORGANISM  Lymphocytic choriomeningitis mammarenavirus
            Viruses; ssRNA viruses; ssRNA negative-strand viruses;
            Arenaviridae; Mammarenavirus.
REFERENCE   1
  AUTHORS   Gibadulinova,A., Zelnik,V., Reiserova,L., Zavodska,E.,
            Zatovicova,M., Ciampor,F., Pastorekova,S. and Pastorek,J.
  TITLE     Sequence and characterisation of the Z gene encoding ring finger
            protein of the lymphocytic choriomeningitis virus MX strain
  JOURNAL   Acta Virol. 42 (6), 369-374 (1998)
   PUBMED   10358742
REFERENCE   2  (bases 1 to 348)
  AUTHORS   Pastorek,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-DEC-1998) Pastorek J., Molecular Biology, Institute
            of Virology, Slovak Academy of Sciences, Dubravska cesta 9, 842 46,
            SLOVAK REPUBLIC
FEATURES             Location/Qualifiers
     source          1..348
                     /organism="Lymphocytic choriomeningitis mammarenavirus"
                     /mol_type="genomic RNA"
                     /strain="MX"
                     /db_xref="taxon:11623"
                     /lab_host="HeLa cells"
     gene            <70..343
                     /gene="z"
     CDS             <70..343
                     /gene="z"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /codon_start=2
                     /product="ring finger protein"
                     /protein_id="CAA10342.1"
                     /db_xref="UniProtKB/TrEMBL:Q9YJW6"
                     /translation="MGQGKSKEKKDTNTGDRAEILPDTTYLGPLNCKSCWQKFDSLVR
                     CHDHYLCRHCLNLLLSVSDRCPLCKCPLPTKLKISTAPSPPPPYEE"
     mat_peptide     <70..338
                     /gene="z"
```

```
                     /product="ring finger protein"
ORIGIN
        1 cgtttagttg cgctgtttgg ttgaacagcc ttttcctgtg agagtacaga gacaaaccta ...
//
```

2.1.2. EInfo

La URL base de este servicio es **http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi**. Su función es simplemente dar los nombres y estadísticas acerca de las bases de datos disponibles. Si se consulta este servicio sin ningún parámetro se devuelve un listado de los nombres de las bases de datos disponibles en el sistema.

El parámetro *db* permite especificar una base de datos particular y obtener información sobre ella. El formato de respuesta es por omisión XML, pero puede ser modificado con el parámetro *retmode*. Este parámetro tiene dos valores posibles, 'xml' y 'json'. Adicionalmente, se puede usar el parámetro *version* que solo acepta un valor '2.0', para especificar que los datos se recuperen en la versión 2.0 Einfo XML. La diferencia entre la versión por omisión y la 2.0 es el agregado de dos términos, 'IsTruncatable' y 'IsRangeable', en los la descripción de los campos de los registros de la base de datos.

Ejemplo:

- **http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=protein&version=2.0**
  Recupera información sobre la base de datos *protein* en el formato XML versión 2.0.

2.1.3. ESearch

La URL base de este servicio es **http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi**. Permite la búsqueda por texto en las bases de datos, como resultado devuelve una lista de números de identificación de cada registro.

Este servicio requiere especificar la base de datos en la cual buscar, para ello se utiliza el parámetro *db*.

Además, se requiere del parámetro *term* para detallar el texto que se desea buscar. La búsqueda por texto recupera registros que coinciden con el texto ingresado en cualquier campo. Sin embargo, es posible ser más específico, y restringir la búsqueda a un campo en particular. Para ello, luego del término de búsqueda se adiciona el nombre del campo encerrado entre corchetes. Además, es posible construir búsquedas más complejas usando los operadores 'AND' , 'OR' y 'NOT'. Los campos de búsqueda disponibles varian con cada base de datos y pueden obtenerse con el servicio EInfo. En la tabla 1 se muestran algunos de los campos de búsqueda para la base de datos *nuccore*.

Ejemplos:

- **esearch.fcgi?db=nuccore&term="complete+genome"[title]+AND+"Escherichia+coli"[Organism]**
  Recupera los registros que corresponden al microrganismo *Escherichia coli* y que contengan la frase "complete genome" en el título. Notar que los espacios son reemplazados por un símbolo '+'.

- **esearch.fcgi?db=nuccore&term="complete+genome"[title]+"Escherichia"[Organism]+NOT+-"Escherichia+coli"[Organism]**
  Recupera los registros de la base de datos *nuccore* que corresponden a genomas completos de microrganismos del género *Escherichia*, pero que no sean de *Escherichia coli*.

- **esearch.fcgi?db=nuccore&term="Escherichia"[Organism]+OR+"Bacillus"[Organism]**
  Recupera los registros de la base de datos *nuccore* que corresponden a los microrganismos del género *Escherichia* o del género *Bacillus*.

Los parámetros *retstart* y *retmax* especifican las opciones de paginación de los resultados. El primero establece el índice del primer registro a recuperar y el segundo la cantidad de registros a recuperar. El valor por omisión de *retmax* es 20.

Ejemplos:

- **esearch.fcgi?db=nuccore&term="Escherichia+coli"[Organism]&retstart=0&retmax=100**
  Recupera los cien primeros registros que corresponden al microrganismo *Escherichia coli*.

- **esearch.fcgi?db=nuccore&term="Escherichia+coli"[Organism]&retstart=100&retmax=100**
  Recupera los registros del 101 al 200, que corresponden al microrganismo *Escherichia coli*.

Field Name	Nombre completo	Descripción
accn	accession	Número de acceso de la secuencia
auth	author	Autores de la publicación
div	division	División de la base de datos
fkey	feature key	*Feature* anotado sobre la secuencia
gene	gene name	Nombre del gene asociado a la secuencia
iss	issue	Número de ejemplar (*issue*) de la publicación
jour	journal	Abreviacíón de la nombre de la publicación
kywd	keyword	Palabras clave dadas por el publicante
mdat	modification date	Fecha de la última modificación
orgn	organism	Nombres científicos y comunes de los organismos, y de los niveles superiores de la clasificación taxonómica
pdat	publication date	Fecha de publicación original
prop	properties	Clasificación por calificadores (*qualifiers*) y tipo de molécula
slen	sequence length	Longitud de la secuencia
titl	title	Contenido del título (*definition*)
uid	uid	Número único asignado a cada secuencia

Tabla 1: Campos de búsqueda de la base de datos *nuccore* de NCBI

El parámetro *rettype* define que información se va a devolver de la búsqueda. Solo acepta dos valores posibles, 'uilist' y 'count'. Con el primero, la opción por omisión, se devuelve una lista con los identificadores de los registros. La segunda devuelve el recuento de los registros.

Ejemplos:

- **esearch.fcgi?db=nuccore&term="Escherichia+coli"[Organism]&rettype=uilist**
  Recupera la lista de los identificadores de registros.

- **esearch.fcgi?db=nuccore&term="Escherichia+coli"[Organism]&rettype=count**
  Recupera el número de resgistro coincidentes con la búsqueda.

El parámetro *retmode* especifíca el formato de respuesta: 'xml' o 'json'.

El parámetro *reldate* permite limitar la búsqueda a los 'n' días anteriores. Este parámetro requiere el uso del parámetro *datetype*, que especifica que tipo de fecha se está buscando (por ejemplo, fecha de publicación o fecha de última actualización). Los tipos de fechas disponibles es dependiente de cada base de datos. Se pueden consultar con el sistema EInfo.

Los parámetros *mindata* y *maxdate*, permiten la búsqueda de registros en un rango determinado de fechas. También requiere el uso del parámetro *datetype*.

Ejemplos:

- **esearch.fcgi?db=protein&term=egfr&reldate=60&datetype=pdat&retmax=100**
  Recupera los identificadores de 100 registros de la base de datos *protein* que contienen el término 'egfr' (de *epidermal growth factor receptor*) en alguno de los campos y hayan sido publicadas en los últimos 60 días.

- **esearch.fcgi?db=protein&term=egfr&datetype=mdat&mindate=2014&maxdate=2014**
  Recupera los identificadores de todos los registros de la base de datos *protein* que contienen el término 'efgr' y cuya última actualización haya sido durante el año 2014.

2.1.4. ESummary

La URL base es http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi. La función de este servicio es la de proporcionar la información esencial de un registro a modo de resumen. Esto evita tener que descargar el registro completo. Usualmente, se lo utiliza como un paso secundario luego de usar el servicio ESearch.

Este servicio requiere de los parámetros *db*, para especificar una base de datos y *id* para indicar de que registros se desea obtener información. Se pueden especificar hasta 100 registros separando los identificadores con una coma.

Ejemplos:

- **esummary.fcgi?db=protein&id=894216091**
  Recupera el resumen del registro con el identificador 894216091 en la base de datos *protein*.

- **esummary.fcgi?db=protein&id=894216091,906849281,906849280**
  Recupera el resumen de los registros con los identificadores 894216091, 906849281 y 906849280, de la base de datos *protein*.

ESummary acepta los parámetros *retstart*, *retmax* y *retmode* de la misma forma que ESearch.

2.1.5. EFetch

La URL base es **http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi**. La función de este servicio es la de descargar la información completa de un registro de una base de datos a partir de su número de identificación. Suele usarse, luego de usar el servicio ESearch.

Al igual que ESummary requiere los parámetros *db* y *id* para especificar que registros se quiere obtener y de que base de datos.

Ejemplos:

- **efetch.fcgi?db=protein&id=894216091**
  Recupera el resumen del registro con el identificador 894216091 en la base de datos *protein*.

- **efetch.fcgi?db=protein&id=894216091,906849281,906849280**
  Recupera el resumen de los registros con los identificadores 894216091, 906849281 y 906849280, de la base de datos *protein*.

Los parámetros *rettype* y *retmode* controlan el tipo de información delregistro se quiere descargar y en que formato respectivamente. Los tipos y modos disponibles dependen de la base de datos en particular. En la tabla 2 se muestran los tipos y modos para algunas de las bases de datos.

Es posible utilizar los parámetros *retstart* y *retmax*, para especificar una fracción de los resultados, al igual que en ESearch y ESummary.

El parámetro *strand* permite recuperar, para registros que corresponden a secuencias de moléculas de DNA, una cadena específica. Hay dos valores permitidos: ‘1’ para la cadena positiva (o sentido) y ‘2’ para la cadena negativa (o antisentido).

Los parámetros *seq_start* y *seq_stop* define la región de la secuencia que se desea recuperar. La primer base de la secuencia es la número ‘1’.

El parámetro *complexity* permite recuperar los registros con diferentes grados de complejidad. Los valores permitidos son números del ‘0’ al ‘4’, que corresponden de mayor complejidad de los datos (‘0’) a menor complejidad (‘4’). Esta opción no está disponible para todas las bases de datos.

2.1.6. Acceso FTP

La dirección base para el acceso por *FTP* es **ftp://ftp.ncbi.nlm.nih.gov/**. En cuanto a secuencias de nucleótidos, las dos colecciones más importantes están en **ftp://ftp.ncbi.nlm.nih.gov/refseq/release/**, para las secuencias de referencia y **ftp://ftp.ncbi.nlm.nih.gov/genbank/** para todas las secuencias. Debido a su gran tamaño, los datos en ‘GenBank’ están organizados en divisiones de acuerdo al origen de la secuencia. A su vez, dentro de cada división hay muchos archivos, todos ellos siguen una nomenclatura racional para que sea sencillo identificarlos. El nombre de cada archivo dentro de una división está formado por un prefijo que identifica a la división, un número y una extensión de archivo. En la tabla 3 se muestran las divisiones existentes.

Tipo de registro	rettype	retmode
All Databases		
Document summary	docsum	xml, default
List of UIDs in XML	uilist	xml
List of UIDs in plain text	uilist	text
Base de datos *gene*		
XML	null	xml
Gene table	gene_table	text
Bases de datos *nuccore, nucest, nucgss, protein y popset*		
Full record in XML	native	xml
FASTA	fasta	text
TinySeq XML	fasta	xml
Opciones adicionales para *nuccore, nucest, nucgss y popset*		
GenBank flat file	gb	text
GBSeq XML	gb	xml
Opciones adicionales para *nuccore y protein*		
Feature table	ft	text
Opciones adicionales para *nuccore*		
GenBank flat file with full sequence (contigs)	gbwithparts	text
CDS nucleotide FASTA	fasta_cds_na	text
CDS protein FASTA	fasta_cds_aa	text
Opciones adicionales para *protein*		
GenPept flat file	gp	text
GBSeq XML	gp	xml
Base de datos *pubmed*		
XML	null	xml
Abstract	abstract	text

Tabla 2: Opciones para los parámetros *rettype* y *retmode* para el servicio EFetch en NCBI. La lista que se muestra corresponde a un subconjunto seleccionado de las bases de datos disponibles. Para obtener la referencia completa se puede visitar **http://www.ncbi.nlm.nih.gov/books/NBK25499/table/chapter4.T._valid_values_of__retmode_and/**

Categoría	Prefijo	Descripción
PRI	gbpri	Primates
ROD	gbrod	Roedores
MAM	gbmam	Mamiferos (no primates, ni roedores)
VRT	gbvrt	Vertebrados (no mamíferos)
INV	gbinv	Invertebrados
PLN	gbpln	Plantas, hongos, plastidos y otros eukariotas
BCT	gbbct	Bacterias y archaeas
VRL	gbvrl	Viruses
PHG	gbphg	Bacteriofagos
ENV	gbenv	Secuencias de muestras ambientales
SYN	gbsyn	Secuencias de cosntrucciones sintéticas
EST	gbest	*Expressed sequence tags*. Secuencias cortas de cDNA
GSS	gbgss	*Genome survey sequences*. Secuencias genómicas cortas
TSA	gbtsa	Ensamblados de transcriptómica (*Transcriptome shhotgun assemblies*)
HTC	gbhtc	Secuencias de cDNA por métodos de alto rendimiento (*high throughput* . Esta categoría incluye secuencias de cDNA no terminadas. Cuando la secuenciación se completa se mueve a la categoría taxonómica que le corresponde
HTG	gbhtg	Secuencias genómicas por métodos de alto rendimiento (*high throughput*)
STS	gbsts	*Sequence tagged sites*
PAT	gbpat	Secuencias relacionadas con aplicaciones patentadas
UNA	gbuna	Secuencias no anotadas.
CON	gbcon	Rearmados en proyectos de secuenciación (*Contigs*)

Tabla 3: Categorías de las bases de datos de GenBank

2.2. European Nucleotide Archive - ENA

La base de datos ENA ofrece varios servicios para la búsqueda y descarga de registros programaticamente usando una interface REST. Además da acceso FTP para descargar sus datos. De todos los servicios que ofrece, se discutirán el servicio de recuperación de datos y el de búsqueda por texto.

2.2.1. Formato del registro

El formato del registro en texto plano se muestra debajo. Su estructura es muy similar al formato de texto plano de NCBI.

```
ID   AJ131281; SV 1; linear; genomic RNA; STD; VRL; 348 BP.
XX
AC   AJ131281;
XX
DT   03-DEC-1998 (Rel. 57, Created)
DT   10-DEC-1999 (Rel. 62, Last updated, Version 2)
XX
DE   Lymphocytic choriomeningitis virus z gene, partial, strain MX
XX
KW   RING finger protein; z gene.
XX
OS   Lymphocytic choriomeningitis mammarenavirus
OC   Viruses; ssRNA viruses; ssRNA negative-strand viruses; Arenaviridae;
OC   Mammarenavirus.
XX
RN   [1]
RP   1-348
RA   Pastorek J.;
RT   ;
RL   Submitted (02-DEC-1998) to the INSDC.
RL   Pastorek J., Molecular Biology, Institute of Virology, Slovak Academy of
RL   Sciences, Dubravska cesta 9, 842 46, SLOVAK REPUBLIC.
...
XX
DR   MD5; 520554560ea4a63dc1fe3f175b37a319.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..348
FT                   /organism="Lymphocytic choriomeningitis mammarenavirus"
FT                   /lab_host="HeLa cells"
FT                   /strain="MX"
FT                   /mol_type="genomic RNA"
FT                   /db_xref="taxon:11623"
FT   CDS             <70..343
FT                   /codon_start=2
FT                   /gene="z"
FT                   /product="ring finger protein"
...
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
FT                   /protein_id="CAA10342.1"
FT                   /translation="MGQGKSKEKKDTNTGDRAEILPDTTYLGPLNCKSCWQKFDSLVRC
FT                   HDHYLCRHCLNLLLSVSDRCPLCKCPLPTKLKISTAPSPPPPYEE"
FT   mat_peptide     <70..338
FT                   /gene="z"
FT                   /product="ring finger protein"
XX
SQ   Sequence 348 BP; 95 A; 100 C; 77 G; 76 T; 0 other;
     cgtttagttg cgctgtttgg ttgaacagcc ttttcctgtg agagtacaga gacaaaccta        60
...
```

//

2.2.2. Recuperación de datos

La URL base de este servicio es **http://www.ebi.ac.uk/ena/data/view/**'. El uso más sencillo es la búsqueda por el código de identificación del registro que se quiere recuperar. Es posible, la búsqueda por rango separando dos códigos de identificación con el símbolo '-'.
Ejemplo:

- **http://www.ebi.ac.uk/ena/data/view/KP718916**
  Recupera el registro 'KP718916' que corresponde a una secuencia de HIV-1.

Otra opción, es la búsqueda por nombre taxonómico. Este tipo de búsqueda acepta nombres comunes, nombres científicos y números de identificación de taxa. En este tipo de búsqueda la URL base es **http://www.ebi.ac.uk/ena/data/view/Taxon:** Por ejemplo, la búsquedas que se muestra debajo recuperan los registros asociados a una especie de laucha.

- **http://www.ebi.ac.uk/ena/data/view/Taxon:calomys laucha**
- **http://www.ebi.ac.uk/ena/data/view/Taxon:small vesper mouse**
- **http://www.ebi.ac.uk/ena/data/view/Taxon:56211**

Además, es posible obtener información acerca de proyectos de secuenciación. Para ello, se necesita el número de identificación del proyecto (*project_id*) o el código de acceso del estudio (*study_accession*). La URL para estas búsquedas se muestran debajo.

- **http://www.ebi.ac.uk/ena/data/view/study_accession**
  Ejemplo: **http://www.ebi.ac.uk/ena/data/view/PRJDB1793**
- **http://www.ebi.ac.uk/ena/data/view/Project:project_id**
  Ejemplo: **http://www.ebi.ac.uk/ena/data/view/Project:254551**

El formato de descarga de los datos es por omisión HTML, sin embargo, esta no es la mejor opción si se desea hacer consulta programáticas. ENA ofrece otros formatos para descargar los datos. Para ello, se utiliza la opción *display* en el URL de búsqueda. En la tabla 4 se muestran los formatos aceptados.

Opción	Descripción	Ejemplo
html	Valor por omisión. Muestra los resultados en formato html. Disponible para todos los tipos de datos	**view/KP718916&display=html**
xml	Muestra los resultados en formato Xml. Disponible para todos los tipos de datos	**view/KP718916&display=xml**
text	Muestra los datos en formato de texto plano. Disponible solo para secuencias con anotación y ensambladas	**view/KP718916&display=text**
fasta	Muestra los datos en formato fasta. Disponible solo para secuencias con anotación, ensambladas y *trace*	**view/KP718916&display=fasta**
fastq	Muestra los datos en formato fastq. Solo disponible para *trace*	**view/TI1&display=fastq**

Tabla 4: Formatos disponibles en ENA

Además de poder especificar el formato del contenido, se puede definir si se desea que la descarga sea un archivo de texto plano o un archivo comprimido. Para ello, se usa el parámetro *download*. Los valores permitidos son *txt* y *gzip*.

Este sistema de descarga limita la cantidad de descargas a 100 000 registros por consulta. Este límite es para evitar la descarga no intencional de un gran cantidad de datos. Si, es necesario descargar más registros se utiliza la opción *lentgh*, cuyo valor es el nuevo número de límite que se desea. Otro parámetro importante es *offset* que establece cual es el primer registro que se descargará. Esto es útil para hacer descargar fragmentadas.

2.2.3. Búsqueda por texto

La búsqueda por texto en ENA usa la URL base **http://www.ebi.ac.uk/ena/data/warehouse/search**. La búsqueda se realiza en particiones de las bases de datos denominadas *domain* y *result*. La partición *domain* es de mayor nivel de abstracción y divide el contenido según su naturaleza. La partición *result* es

Domain	Result	Descripción
assembly	assembly	Genomas ensamblados.
sequence	sequence_release	Secuencias de nucleótidos (Release).
sequence	sequence_update	Secuencias de nucleótidos (Update).
coding	coding_release	Secuencias codificantes (Release).
coding	coding_update	Secuencias codificantes (Update).
noncoding	noncoding_release	Secuencias no codificantes (Release).
noncoding	noncoding_update	Secuencias no codificantes (Update).
read	read_experiment	Lecturas crudas agrupadas por experimento.
read	read_run	Lecturas crudas.
read	read_study	Lecturas crudas agrupadas agrupadas por estudio.
analysis	analysis	Análisis de secuencias de nucleótidos a partir de lecturas.
analysis	analysis_study	Análisis de secuencias de nucleótidos a partir de lecturas, agrupados por estudio.
trace	read_trace	Trazas de capilares.
study	study	Estudios.
taxon	taxon	Clasificación taxonómica.
sample	sample	Muestras.
environmental	environmental	Muestras ambientales.

Tabla 5: Particiones *domain* y *result* disponibles en ENA.

un refinamiento de *domain* que divide la información con algún criterio adiocional. Por ejemplo, el *domain* 'sequence' agrupa los registros que contiene información de secuencias nucleotídicas. Para este *domain* hay dos *result* disponibles: 'sequence_release' y 'sequence_update'. El primero contiene todas las secuencias de la última versión publicada de la base de datos y el segundo tiene las secuencias que se agregaron o actualizaron desde esa última publicación. Este último conjunto suele ser mucho más pequeño. En la tabla 5 se muestran todos los *domain* y *result* disponibles.

Las búsquedas programáticas por texto, además de los parámetros *domain* y *result*, requieren del parámetro URL *query*. El contenido del parámetro *query* no es libre, es decir no es posible hacer una búsqueda arbitraria de texto. El contenido del parámetro se construye por opciones de filtrado unidos por operadores lógicos (*AND*, *OR*, *NOT*). Si se tienen varias opciones de filtrado, se pueden agrupar entre paréntesis.

Las opciones de filtrado se aplican a los distintos campos que posee cada registro. Los campos disponibles varian en cada *domain* y *result*. También, cada campo de búsqueda tienen valores de un tipo particular: *Booleano*, *Vocabulario restringido*, *Fecha*, *Númerico*, *Textual*, *Geo-espacial* y *Taxonómico*. En la tabla 7 se muestran algunos de los campos de búsqueda más comunes para el *domain=sequence* y *result=sequence_release*. Estos campos se comparan con operadores de igualdad ('='), diferencia ('!='), mayor y menor ('>', '>=', '<=' y '<'), excepto los campos de ubicación geográfica y taxonomía. La búsqueda en estos campos requiere el uso de funciones particulares. En la tabla 6 se muestran algunas funciones que provee ENA para este tipo de búsqueda.

Función	Descripción	Parámetros	Ejemplo
geo_box1	La ubicación se encuentra dentro de un cuadrado definido por la esquina superior derecha y la esquina inferior izquierda.	latitud suroeste, longitud suroeste, latitud noreste y longitud noreste.	geo_box1(-20,10,20,50)
geo_circ	La unicación se encuentra dentro de un circulo definido por el centro y el radio.	Latitud, longitud y radio (km).	geo_circ(44,00,00)
tax_eq	Registros que coinciden con un identificador de taxonomía del NCBI	Identificador de taxonomía de NCBI(numérico).	tax_eq(9606)
tax_tree	Registros que coinciden con un identificador de taxonomía del NCBI y todos sus descendientes.	Identificador de taxonomía de NCBI(numérico).	tax_tree(2759)

Tabla 6: Funciones para la búsqueda por geolocalización y por taxonomía

La búsqueda por texto ofrece las mismas opciones para manipular el formato de los datos a recibir que la búsqueda por identificador que se discutió en la primer sección. Se usan los parámetros *display*, *download*, *length* y *offset* ya vistos (ver la tabla 4). Este tipo de búsqueda ofrece una opción más para el

accession	Texto	Número de acceso.
base_count	Número	Longitud de la secuencias.
cell_line	Texto	línea celular de donde se obtuvo la muestra.
country	Texto	Ubicación de la muestra: País, región y localidad.
description	Texto	Breve descripción de la secuencia.
environmental_sample	Booleano	Especifíca si el material pertenece a muestra ambiental.
keywords	Texto	Palabras clave asociadas a la secuencia.
last_updated	Fecha	Fecha de la última actualización.
mol_type	Vocabulario restringido	Tipo molecular de la molécula *in vivo*.

Tabla 7: Lista de algunos de los campos de búsqueda ofrecidos por ENA.

parámetro *display*, el valor 'report'. Esta opción permite descargar algunos campos de los registros y no el registro completo. Es útil para tener en un resumen de los registros que se desea descargar. Cuando se usa esta opción se requiere el uso del parámetro *fields* para indicar los campos que se desea obtener. Debajo se muestran algunos de los campo más usuales para el *domain* 'sequence' y *result* 'sequence_release'.

- accession
- base_count
- cell_line
- description
- environmental_sample
- host
- keywords
- last_updated
- location
- mol_type
- tax_division
- tax_id
- scientific_name

La referencia completa de los campos de búsqueda y los campos disponibles para los reportes se puede consultar en **http://www.ebi.ac.uk/ena/data/warehouse/usage**.

2.3. DNA Data Bank of Japan - DDBJ

La base de datos DDBJ ofrece, además de la interfaz web, el acceso por FTP y una API web llamada WABI (Web API for Biology). Estos dos son los recursos que permiten el acceso programático a esta base de datos.

2.3.1. Formato del registro

El formato del registro en texto plano se muestra debajo. Su estructura es casi idéntica al registro de texto plano de NCBI.

```
LOCUS       AJ131281                 348 bp    RNA     linear   VRL 10-DEC-1999
DEFINITION  Lymphocytic choriomeningitis virus z gene, partial, strain MX.
ACCESSION   AJ131281
VERSION     AJ131281.1  GI:3970751
KEYWORDS    RING finger protein; z gene.
SOURCE      Lymphocytic choriomeningitis mammarenavirus (LCMV)
  ORGANISM  Lymphocytic choriomeningitis mammarenavirus
            Viruses; ssRNA viruses; ssRNA negative-strand viruses;
            Arenaviridae; Mammarenavirus.
REFERENCE   1
  AUTHORS   Gibadulinova,A., Zelnik,V., Reiserova,L., Zavodska,E.,
            Zatovicova,M., Ciampor,F., Pastorekova,S. and Pastorek,J.
  TITLE     Sequence and characterisation of the Z gene encoding ring finger
            protein of the lymphocytic choriomeningitis virus MX strain
  JOURNAL   Acta Virol. 42 (6), 369-374 (1998)
   PUBMED   10358742
REFERENCE   2  (bases 1 to 348)
  AUTHORS   Pastorek,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-DEC-1998) Pastorek J., Molecular Biology, Institute
            of Virology, Slovak Academy of Sciences, Dubravska cesta 9, 842 46,
            SLOVAK REPUBLIC
FEATURES             Location/Qualifiers
     source          1..348
                     /organism="Lymphocytic choriomeningitis mammarenavirus"
                     /mol_type="genomic RNA"
                     /strain="MX"
                     /db_xref="taxon:11623"
                     /lab_host="HeLa cells"
     gene            <70..343
                     /gene="z"
     CDS             <70..343
                     /gene="z"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /codon_start=2
                     /product="ring finger protein"
                     /protein_id="CAA10342.1"
                     /db_xref="GI:3970752"
                     /translation="MGQGKSKEKKDTNTGDRAEILPDTTYLGPLNCKSCWQKFDSLVR
                     CHDHYLCRHCLNLLLSVSDRCPLCKCPLPTKLKISTAPSPPPPYEE"
     mat_peptide     <70..338
                     /gene="z"
                     /product="ring finger protein"
ORIGIN
        1 cgtttagttg cgctgtttgg ttgaacagcc ttttcctgtg agagtacaga gacaaaccta
 ...
      301 gatatcaaca gccccaagcc caccacctcc ctacgaagag taacaccg
//
```

Tipo	Base de datos	Descripción
DNA	na	Busca en todas las bases de datos dispoibles. Es la opción usada por defecto.
	ddbj	Base de datos propioa de DDBJ.
	wgs	Base de datos de proyectos de secuenciación de genomas completos (Whole genome sequences).
	mga	Secuenciación masiva para anotación genómica (Mass sequence for Genome Annotation.
Protein	aa	Busca en todas las bases de datos disponibles.
	uniprot	Base de datos Uniprot.
	dad	Secuencias de aminoácidos de DDBJ.
	patent_aa	Secuencias derivadas de patentes.

Tabla 8: Bases de datos disponibles en *gententry*

2.3.2. Acceso por Web API

DDBJ ofrece API para dos servicios diferentes. El servicio 'getentry WebAPI' que permite la recuperar registros de la base de dato por número de acceso y el servicio ARSA WebAPI, que permite la búsqueda de registros por su contenido textual (no de secuencia. los dos servicios usan una arquitectura REST para las consultas.

2.3.3. Getentry WebAPI

El servicio getentry permite recuperar registro de una de las bases de datos de DDBJ por su número de acceso. Las consultas se pueden hacer mediante el método GET o usando una URL semántica. En ambos casos, la URL base es la misma: **http://getentry.ddbj.nig.ac.jp/getentry**.

La síntaxis para especificar una consulta usando el método GET es **http://getentry.ddbj.nig.ac.jp/getentry?database=nombre_de_la_base_de_datos&accession_number=número_de_acceso**.
Ejemplo:

- **http://getentry.ddbj.nig.ac.jp/getentry?database=ddbj&accession_number=KP177965** Recupera el registro con número de identificación 'KP177965'.

La misma consulta, usando una URL semántica es **http://getentry.ddbj.nig.ac.jp/getentry/nombre_de_la_base_de_datos/número_de_acceso/revisión**, donde 'revisión' es opcional. Los registros pueden ser actualizados en el tiempo, cada nueva versión de un registro se carateriza por la fecha y hora de la actualización.
Ejemplo:

- **http://getentry.ddbj.nig.ac.jp/getentry/ddbj/AY646354/2010-02-26+14:07:01**'
  Recupera la revisión 2010-02-26 14:07:01 del registro

Para saber que revisiones están disponibles para un registro particular se puede hacer una consulta a la **http://getentry.ddbj.nig.ac.jp/gethistory**, con un formato similar a la de 'getentry'.

Los parámetros disponibles para las consultas son el *número de acceso*, *base de datos*, *revisión*, *formato*, *tipo de archivo*, *mostrar eliminados* y *limite*.

El parámetro *accession_number* se utiliza para especificar el número. Se puede indicar el número de acceso solo (ej. AY646354), o el número de versión (AY646354.1). Es posible especificar varios número de acceso en un sola consulta, separándolos con una coma. También se puede indicar un rango de números de acceso, conectando dos número de accesos con un símbolo '-'. Este parámetro es el único obligatorio.

El parámetro *database* permite especificar la base de datos sobre la cual se realiza la búsqueda. Las báses de datos disponibles se muestran en la tabla 8.

El parámetro *revision* permite indicar el código de revisión del registro que se quiere obtener. Consiste en la fecha y la hora con el formato **yyyy-MM-dd hh:mm:ss** o **yyyy-MM-dd hh:mm:ss release**.

El parámetro *format* permite especificar el formato de retorno de los datos. En la tabla 9 se muestran los posibles valores que puede tomar.

El parámetro *filetype* permite especificar que tipo de archivo es el que se quiere recuperar los datos. Los valores posibles son *html*, *text* y *gz* (texto plano comprimido).

Es posible recuperar información de registro viejos y que fueron eliminados de las versiones más recientes de las bases de datos. Para ello, se debe usar el parámetro *show_suppressed*. Los valores posibles son *true* y *false*. El valor por defecto es *false*.

Formato	Descripción
flatfile	Formato de texto plano.
xml	Formato INSDSeq-XML.
fasta	Secuencias de DNA o proteína en formato fasta.
trans	Secuencia codificante traducida a proteínas (fasta).
cds	Secuencia codificante en nucleótidos (fasta).

Tabla 9: Formatos de retorno de *getentry*.

Para limitar la cantidad de resultados que se desea recuperar se utiliza el parámetro *limit*. Cuyo valor es el número máximo de registro que se puede recuperar. El valor por defecto es 10. Si se especifica 0, se repueran todos los registros disponibles.

2.3.4. ARSA WebAPI.

La URL base para la búsqueda es **http://ddbj.nig.ac.jp/arsa/search**. Los resultados pueden ser recuperados en idioma inglés o japonés. Para definir el idioma es necesario agregar a la dirección base el el capo *lang* con valor 'en'. De esta forma la URL se convierte en: **http://ddbj.nig.ac.jp/arsa/search?lang=en**. El formato de respuesta es HTML, porque está pensado para ser mostrado en un navegador como una página web. Si se desea, obtener esta información en un formato que pueda ser fácilmente manipulado se tienen dos opciones. Ambas implican usar una URL base diferente.

Una de las opciones es descargar el contenido de todos los registros en formato plano, para ello, hay que cambiar a la URL **http://ddbj.nig.ac.jp/arsa/downloadAll?lang=en**.

La otra opción es recuperar los identificadores de cada registro y luego descargar cada uno de ellos invididualmente con el servicio 'getentry', para ello, se debe cambiar la URL base a **http://ddbj.nig.ac.jp/arsa/searchAllIds?lang=en**.

Hay dos condiciones de búsqueda, la búsqueda rápida (*Quick Search*) y la búsqueda avanzada (*Advanced Search*). La condición de búsqueda se define con el parámetro *cond*, sus posibles valores son: 'quick_search' y 'advanced_search'. Sin embargo, en esta guía se trabajará solamente con la condición de búsqueda simple.

El parámetro *query* define el texto que se desea buscar. Existen varias opciones que permiten refinar la búsqueda:

Coincidencia por conjunción Búsquedas que contienen todas las palabras clave dadas. Se debe usar el parámetro *operator* con el valor 'AND'.
Ejemplo:
search?lang=en&cond=quick_search&query=brain+phosphatidylethanolamine&operator=AND
Recupera todos los registros que contienen las palabras 'brain' y 'phosphatidylethanolamine', aunque se encuentren en lugares diferentes en el registro. El primer resultado de la búsqueda del ejemplo es un registro que contiene la palabra 'brain' en la definición del registro (título principal) y la palabra 'phosphatidylethanolamine' en el título de una de las referencias bibliográficas que aparecen en ese registro.

Coincidencia por disyunción Búsquedas que contienen al menos una de las palabras clave dadas. Se indican en el parámetro *operator*' y con el valor 'OR'.
Ejemplo:
search?lang=en&cond=quick_search&query=tinea+versicolor&operator=OR
Recupera registros de 'Tinea versicolor', 'Trametes versicolor' y 'Tinea pellionella'.

Negación Búsquedas que contienen una o más palabras clave y que carecen de otras que son dadas explicitamente. Se indican con la palabra clave 'NOT' dentro del valor del parámetro *query*, la palabra clave que se encuentra a su derecha es la que debe estar ausente.
Ejemplo:
search?lang=en&cond=quick_search&query=homo+NOT+sapiens&operator=AND'
Recupera los registros que contiene la palabras 'homo' pero no 'sapiens'.

Coincidencia parcial Búsquedas que contienen parte de una palabra. Se indican dentro del valor del parámetro *query*. Al final del término buscado se agrega el símbolo '*'.
Ejemplo:
search?lang=en&cond=quick_search&query=candid*
Recupera registros de '*Candida dubliniensis*', '*Candidatus Phytoplasma*', 'Junin virus strain candid1', entre otros.

Nombre del campo	Abr.	Description
PrimaryAccessionNumber	pa	Primer número de acceso en ACCESSION
AccessionNumber	an	Números de acceso en ACCESSION
Division	dv	Division en LOCUS (ej. VRL=virus)
SequenceLength	sl	Longitud de secuencia en LOCUS
MolecularType	mt	Tipo de molecula in LOCUS (ej. RNA)
MolecularForm	mf	Estructura molecular (ej. linear)
Date	dt	Última fecha de publicación en LOCUS
Definition	df	Texto en DEFINITION
Comment	cm	Texto en COMMENT
Keyword	kw	Texto en KEYWORDS
Organism	og	Texto en el organismo de ORGANISM
Lineage	ln	Texto en el linaje de ORGANISM
ReferenceAuthor	ra	Texto en AUTHORS de REFERENCE
ReferenceTitle	rt	Texto en TITLE de REFERENCE
ReferenceJournal	rj	Texto en JOURNAL de REFERENCE
ReferencePubmedID	rp	Texto en PUBMED de REFERENCE
Feature	fe	Texto en FEATURES

Tabla 10: Campos de búsqueda disponibles en DDBJ

Coincidencia de frases Búsquedas que contienen varios términos en el orden dado. Se indican encerrando entre comillas los términos.
Ejemplo:
search?lang=en&cond=quick_search&query="mouse+brain"&operator=AND
Recupera los registro que contengan la frase ‘mouse brain’.

Búsquedas en campos Búsquedas que deben coincidir en un campo particular del registro. Se indican dentro del valor del parámetro *query* nombrando el campo en el que se debe hacer la búsqueda antes de la palabra clave buscada y separados por el símbolo ’:’. La tabla 10 muestra los campos de búsqueda disponibles.
Ejemplo:
search?lang=en&cond=quick_search&query=df:"homo+sapiens"&operator=AND
Recupera registros que tienen la frase ‘homo sapiens’ en su definición (título principal).

Búsqueda por expresiones regulares Algunas búsquedas en campos permiten usar expresiones regulares. Deben encerrarse con el símbolo ’/’. En la tabla 11 se muestran las expresiones regulares permitidas.

Búsqueda por rango de valores Algunas búsquedas en campos permiten indicar un rango de valores para la coincidencia. Deben encerrarse entre corchetes. La palabra clave ’TO’ debe estar presente entre el valor mínimo y el máximo del rango. Se puede usar el símbolo ’*’ para indicar un valor de inicio o final del rango desconocido.
Ejemplos:
search?lang=en&cond=quick_search&query=pa:[GL000000+TO+GL000010]&operator=AND
Recupera los registros existentes cuyo número de acceso primario sea ’GL000001’, ’GL000002’, . . . y ’GL000010’. **search?lang=en&cond=quick_search&query=sl:[0+TO+100]&operator=AND**
Recupera todos los registros cuya longitud de secuencia sea 100 o menos.

2.3.5. Acceso FTP

DDBJ permite descargas varias bases de datos. Nos enfocaremos en *DDBJ* (nucleotidos), *DAD* (aminoácidos) y *wgs* (secuenciación de genomas completos). La URL base para ellas son:

ddbj: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/ddbj

ddbj - release notes: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/release_note_archive/ddbj

dad: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dad

dad - release notes: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/release_note_archive/dad

wgs: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/wgs

wgs - Lista de organismos: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/wgs/WGS_ORGANISM_LIST.txt

Significado	Símbolo	Ejemplo
Cualquier caracter individual	.	/ho.o/ coincide con 'homo' y con 'Rhodococcus'.
Cualquier caracter cero o más veces	*	/XX1*/ coincide con 'XX', 'XX1', 'XX11', etc
Cualquier caracter cero o una vez	?	/XY?123/ coincide con 'XY123' and 'X123'
Cualquier caracter una o más veces	+	/XY1*/ coincide con 'XY1', 'XY11',etc, pero no con 'XY'
Cualquier caracter en un grupo	[abc]	/ho[dm]o/ coincide con 'homo' y con 'Rhodococcus', pero no con 'Achoropsyche'
Cualquier caracter excepto algunos	[^abc]	/ho[^d]o/ coincide con 'homo'y con 'Achoropsyche', pero no con 'Rhodococcus'
Cualquier caracter en un rango	[a-z0-9]	/XY[1-9]00/ coincide con 'XY100', 'XY200', etc, 'XYA00', 'XYB00', etc
Cualquier caracter exactamente n vences	{n}	/XY12/ coincide con 'XY11' pero no con 'XY0' o 'XY000'
Cualquier caracter n o más veces	{n,}	/XY12,/ coincide con 'XY11' y 'XY111', pero no con 'XY1'
Cualquier caracter n hasta m veces	{n,m}	/XY12,3/ coincide con 'XY11' y con 'XY111', pero no con 'XY1' o 'XY1111'

Tabla 11: Expresiones regulares permitidas en las búsquedas por texto en ARSA.

La información en las bases de datos *ddbj* y *dad* está organizada en diferentes categorías, según el origen de las secuencias. Los archivos de datos que pueden ser descargados se corresponden con estas categorías y siguen una nomenclatura racional. En la tabla 12 se detallan las categorías. El nombre de cada archivo está formado por un prefijo que corresponde con alguna de las categorías, después un número de orden creciente y finalmente la extensión. El formato de los archivos es texto plano, comprimido. Para las bases de datos *ddbj* y *dad*, posible descargar archivos en formato `fasta` y `XML`. Estos se encuentran en las subcarpetas `fasta/` y `xml/insdxml/v1.4` respectivamente.

Categoría	Prefijo	Descripción
HUM	ddbjhum	Humano
PRI	ddbjpri	Primates (no humano)
ROD	ddbjrod	Roedores
MAM	ddbjmam	Mamiferos (no primates, ni roedores)
VRT	ddbjvrt	Vertebrados (no mamíferos)
INV	ddbjinv	Invertebrados
PLN	ddbjpln	Plantas, hongos, plastidos y otros eukariotas
BCT	ddbjbct	Bacterias y archaeas
VRL	ddbjvrl	Viruses
PHG	ddbjphg	Bacteriofagos
ENV	ddbjenv	Secuencias de muestras ambientales
SYN	ddbjsyn	Secuencias de cosntrucciones sintéticas
EST	ddbjest	*Expressed sequence tags*. Secuencias cortas de cDNA
GSS	ddbjgss	*Genome survey sequences*. Secuencias genómicas cortas
TSA	ddbjtsa	Ensamblados de transcriptómica (*Transcriptome shhotgun assemblies*)
HTC	ddbjhtc	Secuencias de cDNA por métodos de alto rendimiento (*high throughput* . Esta categoría incluye secuencias de cDNA no terminadas. Cuando la secuenciación se completa se mueve a la categoría taxonómica que le corresponde
HTG	ddbjhtg	Secuencias genómicas por métodos de alto rendimiento (*high throughput*)
STS	ddbjsts	*Sequence tagged sites*
PAT	ddbjpat	Secuencias relacionadas con aplicaciones patentadas
UNA	ddbjuna	Secuencias no anotadas.
CON	ddbjcon	Rearmados en proyectos de secuenciación (Contigs)

Tabla 12: Categorías de las bases de datos de DDBJ

3. Bases de datos secundarias

3.1. Uniprot

Uniprot es un conjunto de bases de datos que reune información sobre anotación de proteínas. Toma datos de protéinas de otras bases de datos y luego agrega y cura la anotación. Uniprot está compuesta de tres bases de datos, *UniParc*, *UniProtKB* y *Uniref*. *UniParc* (*Uniprot archive*) es una colección extensiva de practicamente todas las secuencias de proteínas disponibles en las bases de datos públicas, sin importar si están anotadas o no. *UniprotKB* es una base de datos derivada de *UniParc* que contiene solamente las proteínas que contienen anotaciones, manuales o generadas automáticamente. Esta subdividida en otras dos bases de datos, *Swiss-Prot* y *TrEMBL*. *Swiss-Prot* contiene las proteínas con anotaciones curadas de forma manual, mientras que *TrEMBL* contiene aquellas que fueron anotadas automáticamente. Cuando la anotación de un miembros *TrEMBL* se revisa, es pasada a *Swiss-Prot*. El tamaño de *Swiss-Prot* es cerca del 1 % de *UniprotKB*. Por último, *UniRef* contiene conjunto de proteínas (*clusters*) con diferente grado de identidad. *UniRef* ofrece tres subconjuntos:

UniRef100 Contiene secuencias idénticas de tamaño total o menor.

UniRef90 Contiene secuencias con una identidad del 90 % o más y cubriendo 80 % de la longitud de la proteína de mayor tamaño en el conjunto.

UniRef50 Contiene secuencias con una identidad del 50 % o más y cubriendo 80 % de la longitud de la proteína de mayor tamaño en el conjunto.

Uniprot ofrece una interface Rest para acceder a sus bases de datos. Con ella es posible, hacer búsquedas por texto y recuperar registros por código de identificación. Además, *Uniprot* ofrece un servicio que permite mapear identificadores entre las bases de datos propias (por ejemplo, entre *UniprotKB* y *UniParc*), y entre bases de datos de propias y enternas.

3.1.1. Búsqueda por identificador

La URL base para recuperar un registro por identificador es `http://www.uniprot.org/base_de_datos/codigo.formato`. Cada base de datos utiliza códigos de identificación de diferente naturaleza. Una misma proteína, tiene un identificador diferente en *Uniprot* y en *UniParc*. Los formatos aceptados son 'txt', 'fasta', 'tab', 'xml' y 'rdf'. El formato 'tab' recupera una tabla de texto que contiene el código y nombre del registro, el estado de revisión, el nombre de los genes y proteínas, el organismo de origen y su longitud.

Ejemplos:

- **http://www.uniprot.org/uniprot/Q16654.fasta**
- **http://www.uniprot.org/uniprot/Q16654.xml**
- **http://www.uniprot.org/uniprot/Q16654.tab**
- **http://www.uniprot.org/uniparc/UPI000000D984.xml**
- **http://www.uniprot.org/uniref/UniRef50_Q16654.fasta**

Todos corresponden a la enzima piruvato deshidrogenasa.

De esta forma es posible obtener registros de a uno a la vez. Si se quiere obtener varios registros en una sola consulta es necesario utilizar otro servicio llamado *batch*. La URL base de este servicio es **http://www.uniprot.org/batch/**. Se requieren de dos parámetros. Uno de ellos es *format*, que especifica el formato del registro, los valores aceptados son txt', 'fasta' y 'xml'. El segundo parámetro es *query* que acepta una lista de indentificadores cuyos registro se desea obtener. Se pueden obtener registros de *UniProtKB*, *UniRef* y *UniParc*, sin embargo en la lista de identificadores que se pasa todos deben corresponder a la misma base de datos. Este servicio requiere que los datos sean pasados por un pedido tipo POST. Por ello, no es posible probarlo en el navegador *web*. Para hacer este tipo de consulta se puede escribir un pequeño programa en *Perl*, *Python*, *Ruby*, *Java*, etc o usar el programa *cURL*, u otro similar. Aunque este servicio permite descargar muchos registros a la vez, si se tiene un gran número de registros, conviene dividir la carga en varias consultas.

Ejemplos para cURL:

- **curl -L -d format=xml -d "query=P13368 Q16654" http://www.uniprot.org/batch/ -o out.xml**
  Permite recuperar los registros P13368 y Q16654, en formato 'xml' y los guarda en el archivo 'out.xml'.

Formato	Descripción
html	Formato HTML para ver en un navegador.
tab	Muestra información parcial del registro en formato de texto tabular.
xls	Muestra el mismo contenido que la opción 'tab', pero en formato de tabla de MS-Excel.
fasta	Muestra solo la secuencia de la proteína en formato fasta.
gff	Muestra anotaciones de secuencia.
txt	Muestra el registro completo en formato de texto plano.
xml	Muestra el registro completo en formato XML.
rdf	Muestra el registro completo en formato rdf (*Resource Description Framework*).
list	Muestra una lista de identificadores.
rss	Devuelve un RSS Feed (*Really Simple Syndication*.

Tabla 13: Formatos permitidos por *UniProt*

- **curl -L -d format=txt -d "query=P13368 Q16654" http://www.uniprot.org/batch/ -o out.txt** Permite recuperar los registros P13368 y Q16654, en formato de texto plano y los guarda en el archivo 'out.txt'.

3.1.2. Búsqueda por texto

La URL base de este servicio es **http://www.uniprot.org/uniprot/**. Requiere del parámetro *query* que contiene el texto que se quiere buscar. El contenido de *query* esta formado por varios campos unidos por operadores booleanos. Sino se especifica ningún campo, la búsqueda se hace sobre todos los campos. En la tabla 14 se muestran algunos de los campos de búsqueda más usuales.

El parámetro *format* permite especificar el formato que se desea para el registro. En la tabla 13 se muestran los formatos permitidos. Cuando se usa el parámetro *format* con el valor 'tab', se tiene la opción de especificar que datos se quieren obtener en la tabla. Para ello, se debe usar el parámatro *columns* e indicar en su valor las columnas de datos. Para especificar más de un valor, solamente hay que separarlos con una coma. En la tabla 15 se muestran algunos valores posibles. Para obtener una referencia completa, visitar la URL `http://www.uniprot.org/help/uniprotkb_column_names`.

El parámetro *include* permite incluir información adicional para algunos formatos en particular. Cuando el formato es 'fasta', se incluyen secuencias de las isoformas de la proteína. Cuando el formato es 'rfd' se incluyen descripciones de los datos referenciados. Para los demás formatos no tiene efecto.

Los parámetros *limit* y *offset* permiten especificar el tamaño máximo de registros a recuperar y el índice en el cual se comienza. Por omisión, el valor de *limit* no está definido, por lo que se descargan todos los registros.

Los parámetros *sort* y *desc* permiten modificar el orden en que se obtienen los resultados. El parámetro *sort* permite especificar el campo del registro se utiliza como criterio para asignar el orden. El parámetros *desc* permite establecer si el orden es ascendente (*desc=no*) o descendente (*desc=yes*). Por omisición el orden es descendente.

3.1.3. Mapeo de identificadores

Uniprot ofrece un servicio que permite mapear idenficador entre bases de datos internas y externas. La URL base es *http://www.uniprot.org/help/mapping/*. Para usar este sercicio es necesario espeficicar cuatro parámetros: *from*, *to*, *format* y *query*. Los parámetros *from* y *to* permiten especificar las bases de datos de origen y de destino. Para ello, cada base de datos tiene un código de identificación. En la tabla 16 se muestran algunos de ellos. Para obtener una lista completa visitar la URL **http://www.uniprot.org/help/programmatic_access#id_mapping_examples**.

El parámetro *format* permite especificar el formato de los resultados. El único valor aceptado es 'tab', que muestra una tabla de texto de dos columnas, en la primera de ellas se encuentra el identificador de origen y en la segunda el identificador de la base de datos de destino. Es posible que a un mismo identificador en una base de datos le correspondan varios en otra base de datos.

El parámetro *query* permite indicar el identificador de la base de datos de origen que se quiere mapear. Es posible especificar varios, separándolos con un espacio en blanco.

Ejemplos:

- **curl -L -d from=ACC -d to=ID -d format=tab -d "query=Q6IUF8 Q6IUF7 Q6IUF9 P26578" http://www.uniprot.org/mapping/**
  Recupera el código de identificación de *UniProt* a partir de su número acceso.

Campo	Ejemplo	Descripción
accession	accession:P62988	Registros con número de acceso P62988.
active	active:no	Registros obsoletos.
author	author:ashburner	Registros cuyo uno de los co-autores sea Ashburner.
citation	citation:(author:ashburner journal:nature)	Registros cuyo uno de los co-autores sea Ashburner y haya sido publicado en *Nature*.
cluster	cluster:UniRef90_A5YMT3	Registros que pertenecen a *UniRef90* y cuya secuencia secuencia representativa sea la del registro A5YMT3 de *UniProtKB*.
created	created:[20121001 TO *]	Registros creados después del 01/10/2012.
database	database:(type:pdb 1aut)	Registros que hagan referencia al registro 1AUT de la base de dato externa PDB
domain	domain:3fe-4s	Registros que contengan un dominio 3Fe-4S (formado por trés atomos de hierro y cuatro de azufre).
ec	ec:1.1.1.1	Registros que pertenezcan a enzima con E.C. *number* 1.1.1.1 (Alcohol deshidrogenasa).
family	family:serpin	Registros que pertenezcan a la familia de Serpinas.
fragment	fragment:yes	Lists all entries with an incomplete sequence.
gene	gene:erbb1	Registros de proteínas codificadas por el gen *erbb1*.
go	go:cytoskeleton	Registros asociados con el término GO 'cytoskeleton'
host	host:mouse	Registros de las proteínas de organismos que infectan a ratones
id	id:P00750	Registros cuyo número de acceso primario es P00750.
interactor	interactor:P00520	Registros de proteínas para las que está descripta una interacción con la proteína del registro P00520.
keyword	keyword:toxin	Registros asociados con la palabra clave 'toxin'
length	length:[100 TO 200]	Registros de proteínas cuyo tamaño está entre 100 y 200 residues.
mass	mass:[200000 TO *]	Registros de proteínas cuyo peso molecular es mayor a 200KDa
modified	modified:[20140701 TO 20150701]	Registros que fueron actualizados entre el 01/07/2014 y el 01/07/2015
name	name:"nucleoprotein"	Resgistros cuyo nombre coincida con 'nucleoprotein'
organelle	organelle:Mitochondrion	Registros de proteínas codificadas por el cromosoma mitocondrial.
organism	organism:"mus musculus"	Registros de todas las proteínas codificadas por *Mus musculus*.
reviewed	reviewed:yes	Registros de las proteínas revisadas (*Swiss-Prot*).
taxonomy	taxonomy:arenaviridae	Registros de que pertenezcan al taxón *Arenaviridae*.

Tabla 14: Campos de búsqueda por texto más comunes en Uniprot.

Campo	Descripción
id	Código de identificación.
entry name	Nombre del registro.
comments	Comentarios agregados por el autor al registro.
ec	Número E.C. (Enzyme Catalog).
existence	Evidencia de la existencia de la proteína.
families	Familias a la que pertenece la proteína
features	Tabla de anotaciones.
genes	Genes relacionados con la proteína.
go	Término de Gene-Ontology relacionados con la proteína.
keywords	Palabras clave dadas por el autor.
last-modified	Fecha de la última modificación.
length	Longitud de la secuencia de la proteína.
organism	Nombre del organismo de origen.
organism-id	Número de identificación taxonómica.
reviewed	Indica si la proteína fue revisada (Pertenece a Swiss-Prot)
sequence	Secuencia de aminoácidos de la proteína
3d	Muestra que resultados de estructura tridimensional existen para la proteína

Tabla 15: Lista de algunos campos que se pueden recuperar en el formato de texto tabular en *UniProt*.

Base de datos	Identificador
Bases de datos de *Uniprot*	
UniProtKB AC/ID	ACC+ID
UniProtKB AC	ACC
UniProtKB ID	ID
UniParc	UPARC
UniRef50	NF50
UniRef90	NF90
UniRef100	NF100
Gene name	GENENAME
Otras bases de datos	
EMBL/GenBank/DDBJ	EMBL_ID
EMBL/GenBank/DDBJ CDS	EMBL
GI number	P_GI
RefSeq Protein	P_REFSEQ_AC
RefSeq Nucleotide	REFSEQ_NT_ID
Bases de datos de estructuras 3D	
PDB	PDB_ID

Tabla 16: Identificadores de las bases de datos para el servicio de mapeo de bases de datos de *UniProt*

- **curl -L -d from=ACC -d to=P_GI -d format=tab -d "query=Q6IUF8 Q6IUF7 Q6IUF9 P26578" http://www.uniprot.org/mapping/**
  Recupera el código GI de las bases de datos de *NCBI* a partir de su número acceso.
- **curl -L -d from=ID -d to=PDB_ID -d format=tab -d "query=THIO_ECOLI" http://www. uniprot.org/mapping/**
  Recupera el código de los registros de la bases de datos PDB asociados a la proteína tiorredoxina de *Escherichia coli*.

4. Ejercitación

Ejercicio 1

Utilice el servicio EInfo de Entrez para construir una tabla que contenga el nombre de la base de datos, la descripción, el número de registros, y la fecha de la última modificación para todas las bases de datos disponibles.

Ejercicio 2

Obtenga de la base de datos 'nuccore' de NCBI el número de registros que corresponden con secuencias del virus HIV-1 que fueron publicadas en cada año desde 1980 hasta 2015. Haga un gráfico de barras para mostrar estos resultados. Repita esta operación usando ENA y DDBJ.

Ejercicio 3

- Buscar información sobre la base de datos MESH (NCBI).
- Buscar en esta base de datos el nombre del gen relacionado con la enfermedad Fibrosis Quística (*Cystic fibrosis*).
- Buscar en la base de datos 'gene' de NCBI, los registros de este gen en el ser humano (en formato XML).
- Recuperar la lista de exones. ¿En qué locus está ubicado el gen?.
- Buscar en Swiss-Prot los registros que están asociados a esta proteína.
- Recuperar las secuencias de la proteína.

Ejercicio 4

Recuperar los registros que corresponden a secuencias nucleotídicas de genomas completos de virus de la familia *Flaviviridae*. Generar un histograma de los tamaños de las secuencias. Obtener de cada registro el código del 'TaxID'. Buscar en la base de datos 'Taxonomy' el nombre científico de cada uno y el acrónimo (si corresponde).

Ejercicio 5

Buscar en la ENA, los registros de secuencias de HIV-1 que fueron obtenidas en el hemisferio sur del continente americano en ENA.

Ejercicio 6

Recuperar todos los registros que corresponden a secuencias de 'Amapari virus' en la base de datos 'nuc_core'. Recuperar de ellos los identificadores de las publicaciones relacionadas. Obtener de 'pubmed' el título del artículo, los autores y el *abstract*.

Ejercicio 7

- Buscar en base de datos 'SRA' (*Sequence Read Archive*) de NCBI los registros que corresponden a los proyectos de secuenciación de virus de la familia *Baculoviridae*.
- Recuperar de cada ellos el nómbre del organismo secuenciado y su 'TaxId', el título del proyecto, el número de bases leidas, la longitud de la secuencia obtenida.
- generar histogramas de la calidad de la lectura.
- determinar el porcentage de bases que no fueron determinadas (son 'N' en la secuencia).

Ejercicio 8

Buscar en *UniProtKB* las secuencias de la enzima Hipoxantina-guanina fosforribosiltransferasa (*Hypoxanthine-guanine phosphoribosyltransferase*, HGPRT) de eucariotas.

- Generar un alineamiento.
- Distinguir cuales pertenecen a *Swiss-Prot* y cuales a *TrEMBL.*
- Mostrar para cada una de ellas que evidencia hay de su existencia.
- Buscar si tienen referencia a la base de datos pFam (*protein families*).

Ejercicio 9

Buscar en las bases de datos de KEGG (*Kyoto Encyclopedia of Genes and Genomes*), las enfermedades relacionadas con la vía metabólica de síntesis de Arginina. Buscar para cada una de ellas, la proteína en *Swiss-Prot* que le corresponde y el nombre del gen asociado. Buscar los SNP disponibles en la base de datos 'SNP' de NCBI. Recuparar la ubicación (cromosoma y posición de base) y la secuencia reportada.