

# **Introduccion a la Bioinformática**

## **Información biológica en formato electrónico**

### **Bases de datos**

**Fernán Agüero**

**Instituto de Investigaciones Biotecnológicas**

**UNSAM**

# Bases de datos: introducción: conceptos básicos

Qué es una base de datos?

Una colección de datos

Cómo colecciono los datos?

Decisión del usuario. Diseño de la base de datos.

Puedo usar:

Procesador de texto? (Word)

Si. Permite sólo búsqueda y ordenamiento simples.

Planilla de Cálculo? (Excel)

También. Como los datos están en columnas independientes, se puede ordenar en formas más complejas. Las búsquedas siguen siendo simples.

# Bases de datos: introducción: conceptos básicos: registros

- Una colección de registros (records).
  - Cada registro tiene varios campos.
  - Cada campo contiene información específica.
  - Cada campo contiene datos de un tipo determinado.
    - Ej: dinero, texto, números enteros, fechas, direcciones
  - Cada registro tiene una **clave primaria**. Un identificador **único** que define al registro sin ambigüedad.

## Planilla

*Versión simple de una base de datos*

gi	Accession	version	date	Genbank Division	taxid	organims	Number of Chromosomes
6226959	NM_000014	3	01/06/2000	PRI	9606	homo sapiens	22 diploid + X+Y
6226762	NM_000014	2	12/10/1999	PRI	9606	homo sapiens	22 diploid + X+Y
4557224	NM_000014	1	04/02/1999	PRI	9606	homo sapiens	22 diploid + X+Y
41	X63129	1	06/06/1996	MAM	9913	bos taurus	29+X+Y

# Tipos de datos

- Cada campo de una base de datos contiene un tipo particular de datos
  - 211203
    - Es un numero?
    - Es texto?
    - Es una fecha?
- Ejemplo de una búsqueda: buscar todos los registros en donde el valor almacenado sea mayor que 211203
  - Es obvio que para poder comparar los valores almacenados tenemos que saber qe tipo de valores estamos comparando.
  - Si es una fecha: 21 12 03 < 2 12 04
  - Si es un numero: 211 203 > 21 204
  - Si es texto: 211203  $\neq$  21204, las comparaciones < y > pueden dar distintos resultados (evaluan orden o longitud)

# Tipos de datos

- Numericos (enteros, decimales)
- Texto
- Fechas (DD/MM/YYYY, HH:MM:SS)
- Logicos (boolean) = verdadero / falso
- Geometricos (punto, linea, circulo, poligonos, etc.)

# Bases de datos: conceptos básicos: clave primaria

gi	Accession	version	date	Genbank Division	taxid	organims	Number of Chromosomes
6226959	NM_000014	3	01/06/2000	PRI	9606	homo sapiens	22 diploid + X+Y
6226762	NM_000014	2	12/10/1999	PRI	9606	homo sapiens	22 diploid + X+Y
4557224	NM_000014	1	04/02/1999	PRI	9606	homo sapiens	22 diploid + X+Y
41	X63129	1	06/06/1996	MAM	9913	bos taurus	29+X+Y

**gi = Genbank Identifier: Clave única : Clave primaria**

Cambia con cada actualización del registro correspondiente a la secuencia

**Accession Number: Clave secundaria**

Refiere al mismo locus y secuencia, a pesar de los cambios en la secuencia.

Accession + Version es equivalente al **gi** (representa un identificador único)

**Ejemplo: AF405321.2      Accession: AF405321      Version: 2**

# Bases de datos: bases de datos relacionales

gi	Accession	version	date	Genbank Division	taxid	organims	Number of Chromosomes
6226959	NM_000014	3	01/06/2000	PRI	9606	homo sapiens	22 diploid + X+Y
6226762	NM_000014	2	12/10/1999	PRI	9606	homo sapiens	22 diploid + X+Y
4557224	NM_000014	1	04/02/1999	PRI	9606	homo sapiens	22 diploid + X+Y
41	X63129	1	06/06/1996	MAM	9913	bos taurus	29+X+Y

## Base de datos relacional:


*Normalizar* una base de datos: repartir sub-elementos repetidos en varias tablas, relacionadas a través de un identificador único (clave primaria).

gi	Accession	version	date	Genbank Division	taxid
6226959	NM_000014	3	01/06/2000	PRI	9606
6226762	NM_000014	2	12/10/1999	PRI	9606
4557224	NM_000014	1	04/02/1999	PRI	9606
41	X63129	1	06/06/1996	MAM	9913
taxid	organims	Number of Chromosomes			
9606	homo sapiens	22 diploid + X+Y			
9913	bos taurus	29+X+Y			

# Bases de datos: distribucion de la informacion

gi	annotation
5693	Trypanosoma cruzi chromosome 3, ORF 1234, similar to gi 12345 AF934567 caseine kinase (Candida albicans)
5694	Candida albicans hypothetical protein in region 21922..24568
5695	Sarcocystis cruzi 16SRNA gene
5696	Lutzomyia cruzi cytochrome b; best similarity to gi 1234568

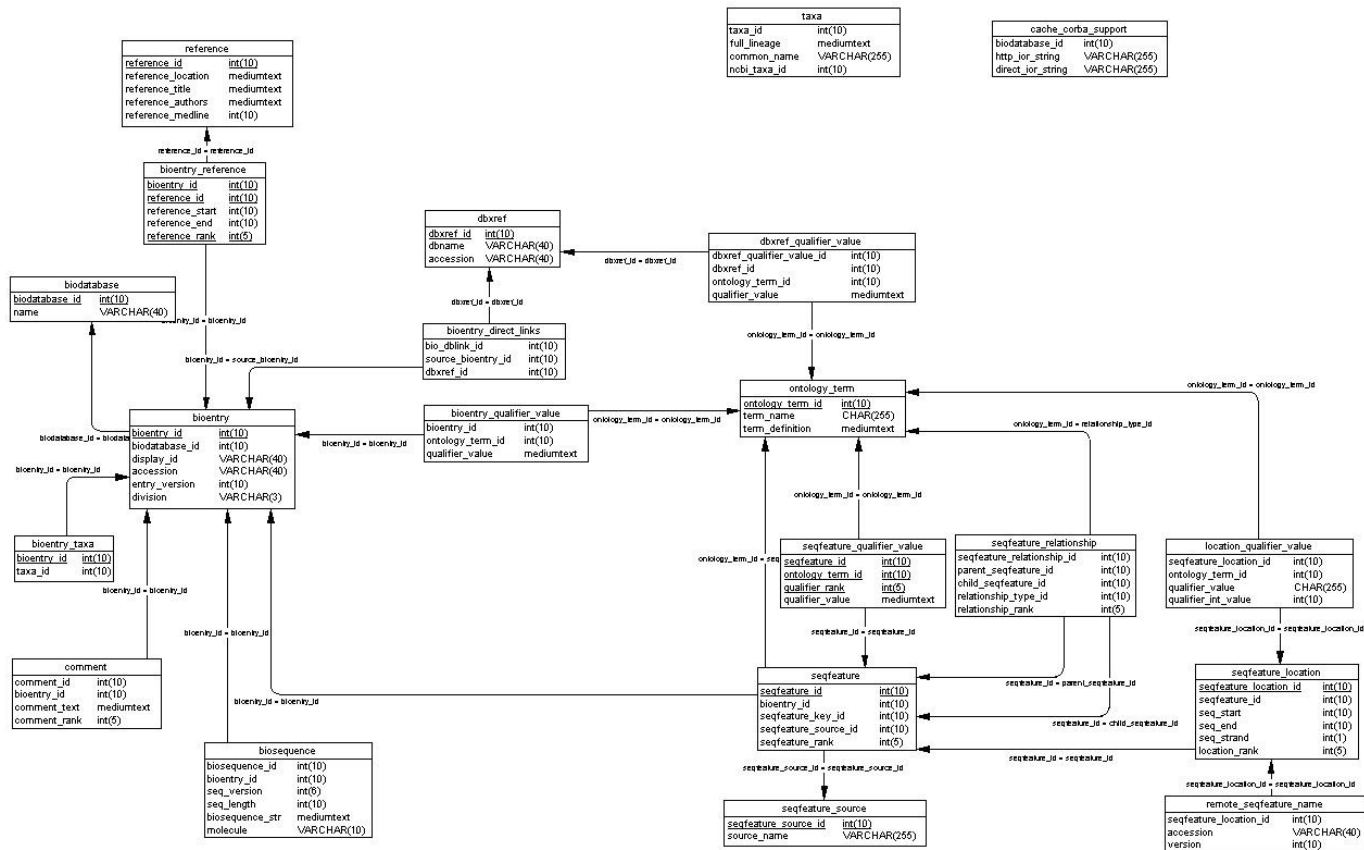
gi	Organism	Annotation	similar to
5693	Trypanosoma cruzi	Chromosome 3, ORF 1234	12345
5694	Candida albicans	Hypothetical protein in region 21922..24568	
5695	Sarcocystis cruzi	16S RNA gene	786512
5696	Lutzomyia cruzi	Cytochrome b	1234568



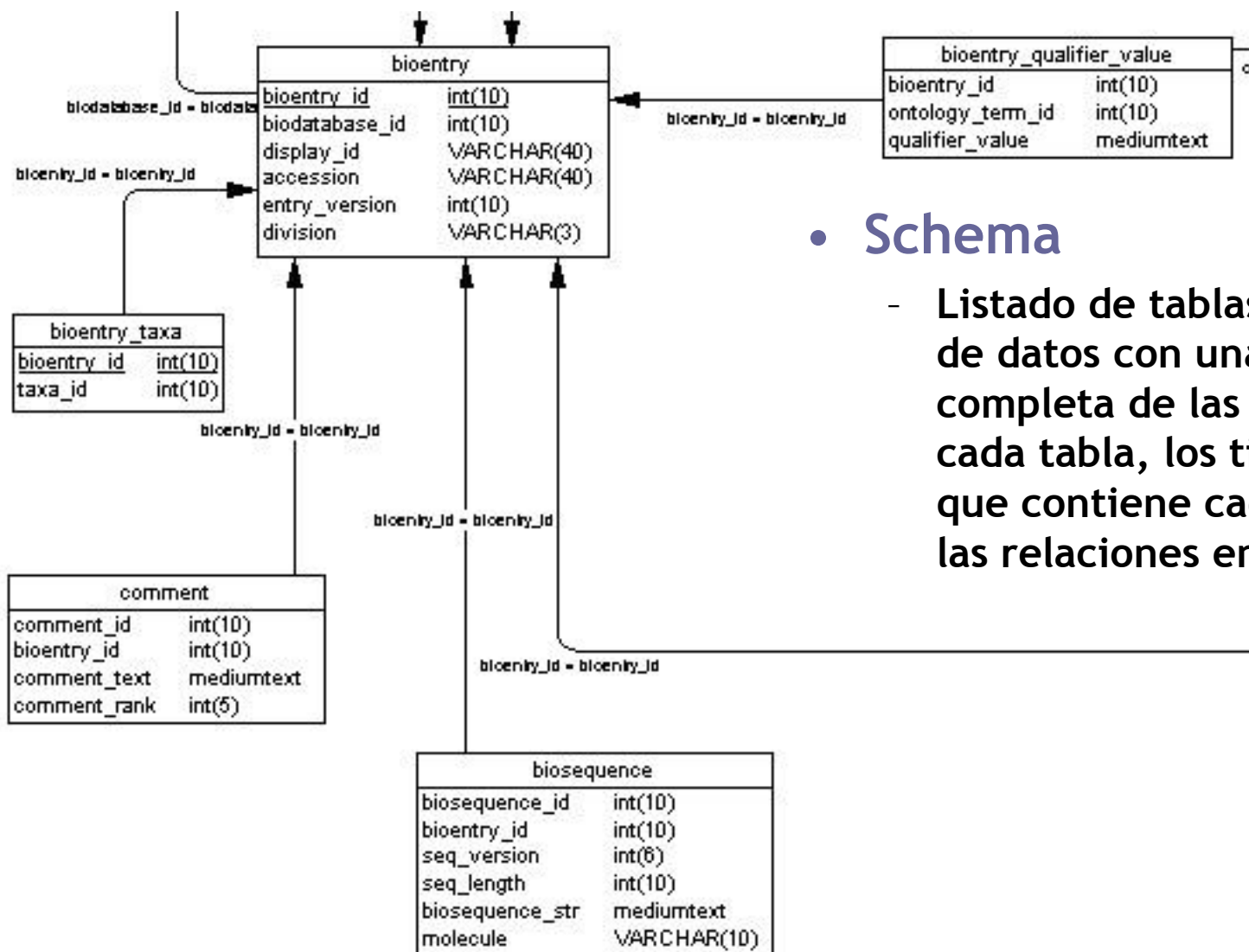


# Schemas

- La distribución de los datos en campos dentro de una tabla y de las relaciones entre tablas y sus campos es lo que se llama el diseño o **schema**



# Schemas (cont)



- Schema

- Listado de tablas de una base de datos con una descripción completa de las columnas de cada tabla, los tipos de datos que contiene cada columna y las relaciones entre tablas.

# Representación relacional de la información

- Qué criterio usamos para diseñar el schema?
- Cómo distribuimos los datos en tablas/columnas?
- Distintas cosas a tener en cuenta:
  - Eficiencia (economía) al almacenar datos: normalización
  - Consultas que planeamos hacer sobre nuestra base de datos y en el tipo de datos.

# Relaciones entre los datos

- Ejemplos de relaciones
  - Proteins ↔ Bibliographic references

## Proteins

Accession	Description	MW	pI
AF1234	Malate dehydrogenase	36000	6.4
AM44432	Cysteine proteinase	45000	4.5

## Linking table

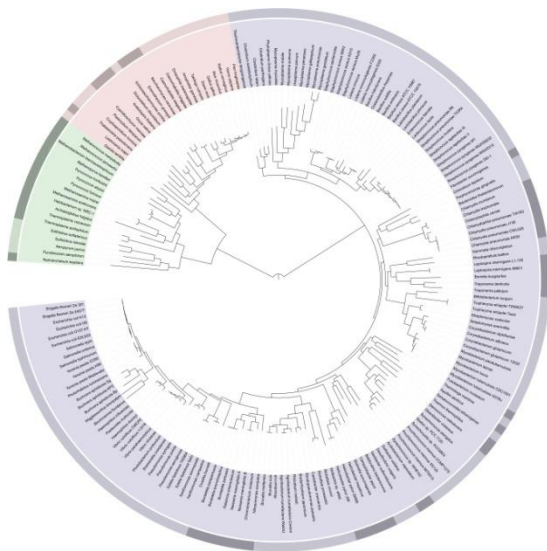
Accession	PubMed ID
AF1234	1234556
AF1234	23445

## Bibliographic References

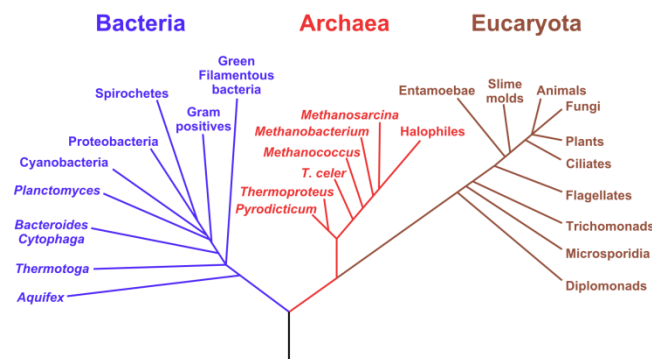
PubMed ID	Journal	Year	Title	Vol
1234556	J Biol Chem	1978	The malate dehydrogenase ...	5
23445	Biochem J	1982	A malate dehydrogenase from ...	13

# Representación de árboles y grafos

- Ejemplo: representación en forma relacional de árboles y grafos
  - Información estructurada jerárquicamente
  - Taxonomy (NCBI), SCOP (Structural Classification of Proteins)



Phylogenetic Tree of Life



**Relational modeling of biological data: trees and graphs.** Aaron J. Mackey. <http://www.oreillynet.com/pub/a/network/2002/11/27/bioconf.html>

# Ejemplo: adjacency list

	Campo	Tipo de dato
PK	Taxon_id	Entero
FK	Parent_id	Entero (ref a PK)
	Nombre	texto

Este tipo de representación se conoce como 'adjacency list':

Cada relación jerárquica 'padre-hijo' está definida en forma explícita.

Taxon_id	Parent_id	nombre
1	-	raíz
2	1	Bacteria
2157	1	Archaea
2759	1	Eukaryota
1224	2	Proteobacteria
...	...	...
543	1236	Enterobacteriaceae
561	543	Escherichia
562	561	Escherichia coli
83333	562	Escherichia coli K12

# Adjacency list: consultas

- Qué consultas podemos hacer sobre los datos organizados en forma de 'adjacency list'?
  - Podemos encontrar el taxón inmediatamente superior de cualquier elemento taxonómico.
  - Podemos encontrar taxones terminales sin 'hijos'
  - Podemos encontrar un taxón (o taxones) buscándolos por nombre
- Y cuáles son difíciles de hacer con esta representación de los datos?
  - Podemos encontrar todos los taxones 'hijos' de un determinado taxón?
    - Ejemplos típicos de este tipo de consultas: buscar todos los mamíferos, todos los vertebrados, o todos los miembros del orden Apicomplexa.
    - Cómo harían esta consulta? Es posible responder estas preguntas con una única consulta sobre la base de datos? Cuántas consultas deberían hacer?

Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostome; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Primates; Catarrhini; Hominidae; Homo/Pan/Gorilla Group; Homo; Homo sapiens

# Representación relacional de árboles: nested set

	Campo	Tipo
PK	Taxon_id	entero
FK	Parent_id	entero
	Left_id	entero
	Right_id	entero
	Nombre	texto

Los valores **left** y **right** son números arbitrarios, pero deben cumplir con la siguiente propiedad:

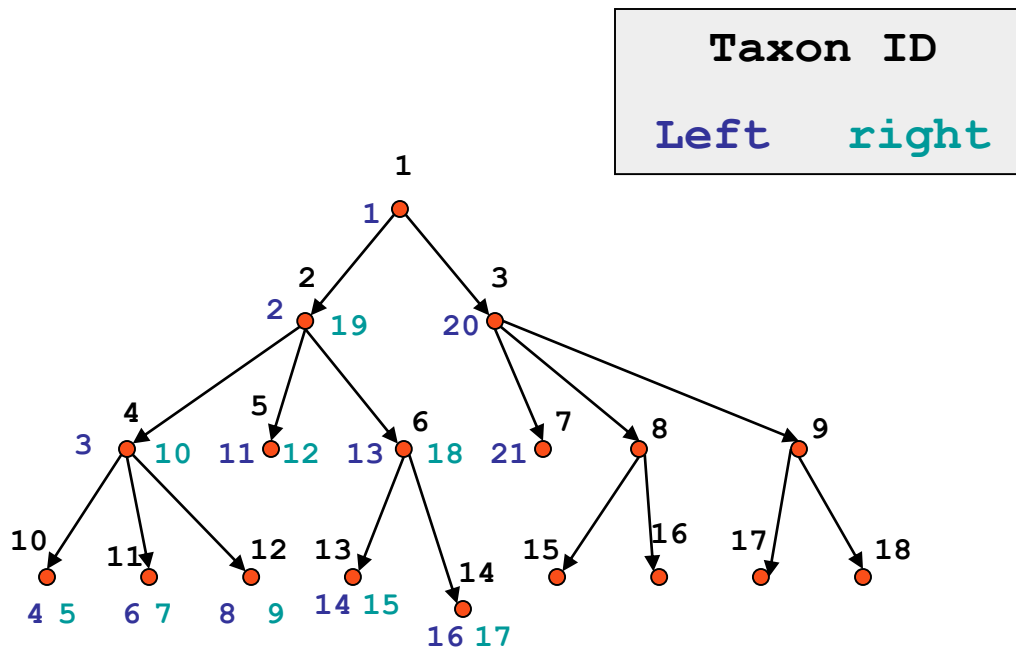
Para cada par 'padre-hijo' los valores del hijo tienen que estar dentro de los valores del padre.

Taxon	Nombre	Parent	Left	Right
1	Root	NULL	1	323458
2	Bacteria	1	21703	87862
3	Archaea	1	87863	92266
4	Eukaryota	1	92267	323456
1224	Proteobacteria	2	23982	49591
...	...	...	...	...
543	Enterobacteriaceae	1236	26681	27938
561	Escherichia	543	26852	26891
562	Escherichia coli	561	26853	26868
83333	Escherichia coli K12	562	26856	26857



# Nested set representation: como calcular left/right?

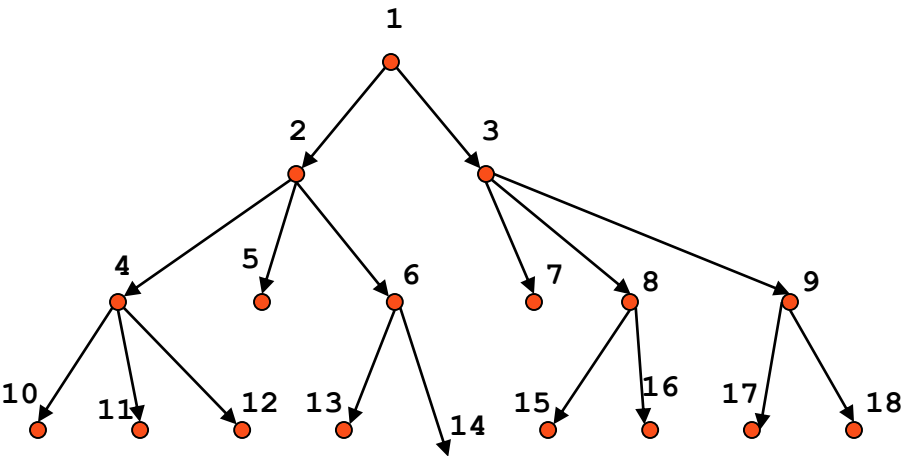
- Cómo se generan los valores para **left** y **right**?
  - Hay que recorrer el árbol asignando estos valores



- Arboles / Grafos
  - Hay distintas maneras de recorrerlos
  - Depth-first
  - Breadth-first

# Materialized Paths

	Campo	Tipo
PK	Taxon_id	entero
	Name	Texto
	Path	Texto



Taxon	Nombre	Path
1	Root	1
2	Bacteria	1.2
3	Archaea	1.3
4	Proteobacteria	1.2.4
5	Cyanobacteria	1.2.5
6	Actinobacteria	1.2.6
7	Crenarchaeota	1.3.7
8	Euryarchaeota	1.3.8
9	Thaumarchaeota	1.3.9
10	Alfa-Proteobacteria	1.2.4.10
11	Gamma-proteobacteria	1.2.4.11
12	Delta-proteobacteria	1.2.4.12
13	Coriobacteridae	1.2.6.13
14	Actinobacteria	1.2.6.14
15	Methanobacteria	1.3.8.15
16	Thermococci	1.3.8.16
17	Cenarchaeales	1.3.9.17
18	Nitrosopumiales	1.3.9.18

En este diseño, el **camino (path)** hacia cada nodo del árbol, está incluido en forma explícita en la información asociada a cada nodo.

# Materialized Paths: consultas

- **Buscar un nodo y todos sus parentales**
  - Ej: buscar todos los parentales del nodo 'Thermococci'
  - Buscar todos los registros cuyo Path **esté contenido** dentro del nodo de interés
  - Path del nodo 'thermococci' = 1.3.8.16
  - Lista de Paths que cumplen la condición,
    - 1.3.8 (Euryarchaeota), 1.3 (Archaea), 1 (root)
- **Buscar un nodo y todos sus descendientes (directos o indirectos)**
  - Ej: buscar todos los descendientes del nodo 'Bacteria'
  - Buscar todos los registros cuyo Path **contenga** al del nodo de interés
  - Path del nodo de 'Bacteria' = 1.2
  - Lista de Paths que cumplen con la condición,
    - 1.2.4 (Proteobacteria), 1.2.5 (cyanobacteria), 1.2.6 (Actinobacteria), 1.2.4.10 (alpha-proteobacteria), 1.2.4.11 (gamma-proteobacteria), 1.2.4.12 (delta-proteobacteria), etc.

# Entity-Attribute-Value

- También: Object-Attribute-Value
- Usado en casos en donde el número de **atributos** (propiedades, parámetros) utilizados para describir **algo** (un objeto o entidad) es muy grande pero el número de atributos que realmente se utilizan es variable y pequeño.
- El caso más común es el de *historias clínicas* de pacientes
  - Cientos de miles de atributos que se pueden medir, diagnosticar, o evaluar
  - En la consulta el médico pregunta de acuerdo a los síntomas que describe el paciente (filtra atributos) y finalmente se almacena en la base de datos aquellos que son **relevantes**.

# Entity-Attribute-Value

- **Modelar estos datos de la manera tradicional**
  - Una tabla con miles de columnas (una por cada posible atributo)
  - El seguimiento en el tiempo de un paciente implica agregar una fila por cada consulta.
  - En cada fila hay sólo unos pocos hallazgos (positivos), el resto de las columnas están vacías (NULL).
- **Modelar estos datos usando el modelo Entity-Attribute-Value**
  - Una única tabla con tres columnas:

**Tabla de objetos (entidades)**

ID	Nombre	Apellido	...
Paciente 1	Tito	Chocola	

**Tabla de datos**

Entity	Attribute	Value
Paciente 1	1	33
Paciente 1	15	230
Paciente 1	56	

**Tabla de atributos**

Attr ID	Name	Description	Data type	Units	Input validation
1	Edad	Edad ...	Integer	Años	\d+
15	Colesterol en sangre	Descripcion ...	Float	Mg/ml	\d+\.\d*

# Structured Query Language

- **SQL - Structured Query Language**

- Es un lenguaje utilizado por todos los sistemas de manejo de bases de datos relacionales
  - Oracle, Sybase, PostgreSQL, MySQL, SQLite, etc.
- Permite definir tablas, relaciones (DDL)
- Y hacer consultas (DML)

- **DDL - Data Definition Language**

- Subset de SQL utilizado para crear bases de datos, tablas, definir campos, etc.
- CREATE DATABASE, CREATE TABLE
- DROP DATABASE, DROP TABLE,
- ALTER TABLE,

- **DML - Data Manipulation Language**

- Subset de SQL utilizado para hacer consultas, insertar y actualizar datos, etc.
- SELECT FROM TABLE, INSERT INTO TABLE
- UPDATE TABLE
- DELETE FROM TABLE

# SQL - Un ejemplo de consulta

## Proteins

Accession	Description	MW	pI
AF1234	Malate dehydrogenase	36000	6.4
AM44432	Cysteine proteinase	45000	4.5

## Linking table

Accession	PubMed ID
AF1234	1234556
AF1234	23445

## Bibliographic References

PubMed ID	Journal	Year	Title	Vol
1234556	J Biol Chem	1978	The malate dehydrogenase ...	5
23445	Biochem J	1982	A malate dehydrogenase from ...	13

```
SELECT accession, description, journal, year, vol, pages, ...  
FROM proteins, bibliographic_references, linking_table  
WHERE linking_table.accession = proteins.accession  
AND linking_table.pubmed_id = bibliographic_references.pubmed_id  
AND proteins.mw <= 36000;
```

# SQL - Un ejemplo de manipulación de datos

## Proteins

Accession	Description	MW	pI
AF1234	Malate dehydrogenase	36000	6.4
AM44432	Cysteine proteinase	45000	4.5

## Linking table

Accession	PubMed ID
AF1234	1234556
AF1234	23445

## Bibliographic References

PubMed ID	Journal	Year	Title	Vol
1234556	J Biol Chem	1978	The malate dehydrogenase ...	5
23445	Biochem J	1982	A malate dehydrogenase from ...	13

**INSERT INTO** proteins (accession, description, mw, ...)  
**VALUES** ('AF1234', 'Malate dehydrogenase', '36000', ...);

**UPDATE** proteins **SET** mw = 45000 **WHERE** accession = AM44432;

**DELETE** FROM proteins **WHERE** accession = AF1234;

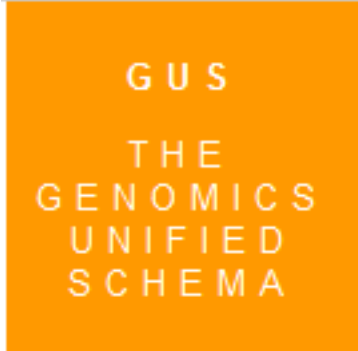


- Reinventar la rueda

- Cuántas maneras hay de organizar información biológica en forma de tablas en una base de datos relacional?
- secuencias + anotación?
- Secuencias + features (propiedades de la secuencia localizables)
- ...

- Después de haber reinventado muchas ruedas ...

- GUS, Genomics Unified Schema
  - PlasmoDB, ToxoDB, CryptoDB (ApiDB), TcruziDB, Allgenes.org,
- Chado, The GMOD Database Schema
  - Wormbase, FlyBase, TaiR, Gramene, SGD, DictyBase



- **Qué es?**
  - Extensive relational database schema
  - Associated application framework
- **Para que se usa?**
  - Para almacenar, integrar, analizar y presentar datos genómicos
- **Modular schema:**
  - Core: tablas conteniendo información de GUS (housekeeping)
  - DOTS: tablas para almacenar información sobre secuencias, genes,
  - SRES: resource tables (to store external resources, controlled vocabularies)
  - RAD: microarray data
  - TESS: transcription binding, transcription factors
  - PROT: proteomics

# CHADO: The GMOD schema



- GMOD = Generic Model Organism Database
- CHADO = DB Schema that underlies many GMOD installations
- Capable of representing many of the general classes of data frequently encountered in modern biology

- **Modular schema**

- Companalysis, for data derived from computational analysis
- Contact, for people, groups, organizations
- Controlled vocabularies
- Expression
- General (for accession numbers and identifiers)
- Genetic
- MAGE, microarray data
- Phenotype,
- Organism, for taxonomic data)
- Publication, for publication references
- Sequence, for sequence, annotation, and features
- Stock, for specimens and biological collections

- **Relational Database Management Systems**
  - **Comerciales**
    - Oracle, Sybase
  - **Open source, gratuitos**
    - PostgreSQL, MySQL
- **Todos usan SQL (standard query language) para**
  - **crear tablas, índices, etc.**
    - CREATE TABLE **taxon** ( **taxon\_id** integer, **name** text, PRIMARY KEY(**taxon\_id**) )
    - ALTER TABLE **taxon** INDEX (name)
  - **ingresar datos**
    - INSERT INTO **taxon** (**taxon\_id**, **name**) VALUES (1, root);
    - UPDATE **taxon** SET **name** = "Trypanosoma cruzi" WHERE **name** = "Schizotrypanum cruzi"
  - **consultar**
    - SELECT **name** FROM **taxon** WHERE **taxon\_id** = 1;
    - SELECT **taxon\_id**, **name** FROM **taxon** WHERE **taxon\_id** IN ('12', '15', '345', '1823')

# Búsquedas en una base de datos: índices

- Para facilitar las búsquedas en una base de datos, se construyen índices.
- Un índice es una lista de claves primarias asociadas a un determinado campo (o grupo de campos)

gi	Accession	version	date	Genbank Division	taxid	organims	Number of Chromosomes
6226959	NM_000014	3	01/06/2000	PRI	9606	homo sapiens	22 diploid + X+Y
6226762	NM_000014	2	12/10/1999	PRI	9606	homo sapiens	22 diploid + X+Y
4557224	NM_000014	1	04/02/1999	PRI	9606	homo sapiens	22 diploid + X+Y
41	X63129	1	06/06/1996	MAM	9913	bos taurus	29+X+Y

gi                      Accession  
6226959 NM\_000014  
6226762 NM\_000014  
4557224 NM\_000014  
41 X63129

## Indices (cont)

- Un ejemplo más complejo: buscar todos los records que contengan la palabra 'kinase' en la descripción de la secuencia

gi	acc	def
214734077	20	Xenopus laevis rhodopsin r
123456743	567	Mus musculus casein kinase

### •Indexar la columna 'def'

word	list of GIs
casein	1234, 3245, 43678, 123456 ...
kinase	432, 5678, 32456, 123456 ...
laevis	36314, 214734, ...
mus	23467, 98732, 123456, 312456, 567
muscu	123467, 98732, 123456, 567983 ...
rhodopsin	214734, 223466, 873212, 23587, 29
xenopus	28462, 36314, 98476, 214734 ...

# Indexar es costoso

- El proceso de indexación es costoso en términos computacionales, pero se realiza una única vez (en realidad cada vez que se actualizan los datos)
- Desde el punto de vista de la base de datos, los índices no son otra cosa que nuevas tablas relacionadas con la tabla que contiene el campo indexado
- Ejemplo más obvio: buscadores de páginas de internet (Google, Altavista). Visitan páginas e indexan los términos que encuentran
  - keyword: url1, url2, url3, url4, etc.

- Son estructuras de datos utilizadas para acelerar la búsqueda de relaciones (tuples) que cumplan alguna determinada condición
  - Igualdad: encontrar Discos donde Banda = Tipitos
  - Otras condiciones son posibles: rangos
    - Encontrar Discos donde Año de Lanzamiento (AL) sea
    - $AL < 1990$  y  $AL > 1980$
- Hay muchos tipos de Indices
  - Convencionales
  - B-Trees
  - Hashing indexes
- Se evalúan de acuerdo a
  - Tiempo de acceso
  - Tiempo que lleva insertar un dato
  - Tiempo que lleva borrar un dato
  - Espacio en disco que ocupan



# Indices convencionales

- **Similares al índice de un libro**
  - El índice contiene una entrada, con un puntero (número de página) al lugar donde están los datos
- **Sparse vs Dense (más o menos densos)**
  - Dense: hay una entrada para cada clave asociada a un objeto
  - Sparse: hay una entrada para algunas claves solamente

Indice (genes)	Genes	Tiempo	Expresión
ABC1	ABC1	1	0.2
BRC2	BRC2	1	0.8
CAM3	BRC2	2	0.3
DHFR	CAM3	2	0.25
EGF-1	DHFR	1	0.1
	DHFR	2	0.3
	DHFR	2	0.4
	EGF-1	1	0.3

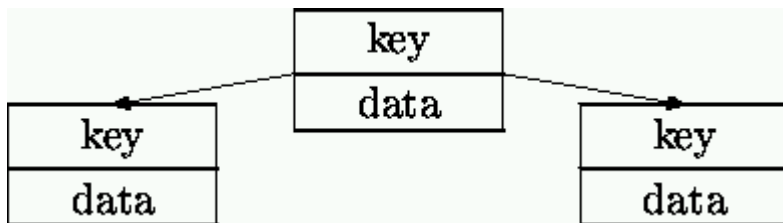
- **Ventajas / Desventajas?**
  - Los índices densos son más rápidos
  - Los índices dispersos ocupan menos espacio
  - **Cuál es el límite a partir del cuál un índice *disperso* se vuelve *denso* ?**
  - Los índices dispersos pueden ajustarse
    - Cuantas claves nos salteamos?
    - Densidad de claves
    - Evaluar # total de filas, # de entradas por cada clave

# Multi-level indexes

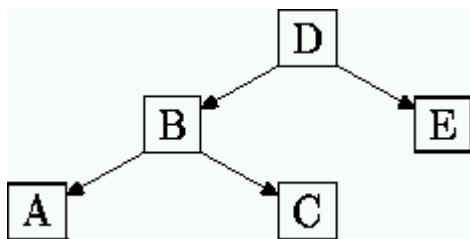
- Los índices convencionales pueden ser muy grandes
- La idea de estos índices es que reduzcan el acceso a disco
- La unidad mínima de I/O en una computadora es mover un bloque de datos del disco a memoria
- Ejemplo
  - Un archivo con 100,000 registros, con 10 datos x gen
  - Un índice disperso, con una entrada x gen: tendríamos 10,000 filas
  - Si asumimos que en un bloque de I/O entran 100 filas, necesitamos acceder a 100 bloques.
- Es deseable mantener los índices en memoria RAM
- Los índices se vuelven costosos cuando crecen los datos
  - Que pasaría en el caso de tener millones de registros?

# Binary trees

- Arboles: nodos conectados con vértices (grafos)



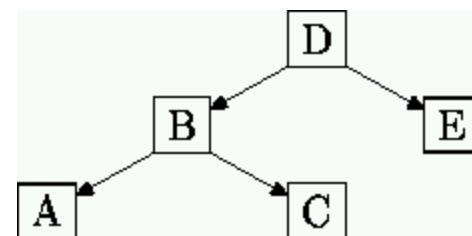
- Para hacer búsquedas, los datos se ignoran.
- Es como si el árbol solo tuviera 'claves'



# Binary trees

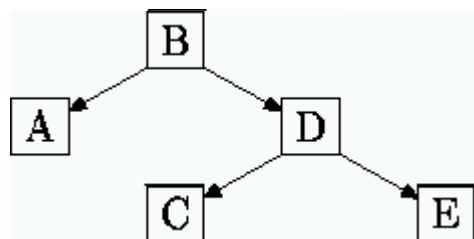
## Reglas para moverse en el árbol

- Ejemplo, buscar A
- Empezar en la raíz, si encontramos A listo!
- Si no, nos movemos, de esta forma:
  - Si la clave del nodo es  $< A$ , a la izquierda
  - Si la clave del nodo es  $> A$ , a la derecha



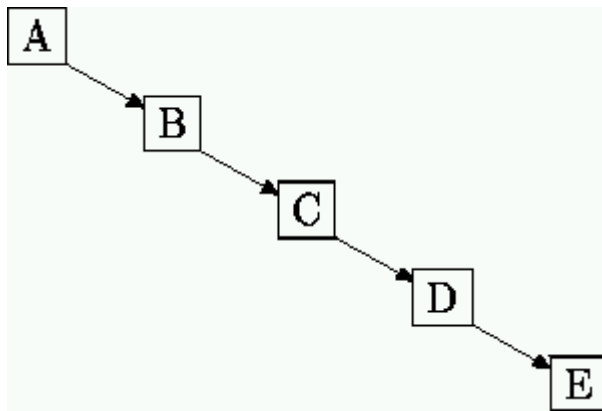
- **Armando un árbol: BDCAE**

- Empezamos por la raíz
- Agregamos nodos siguiendo las mismas reglas de movimiento



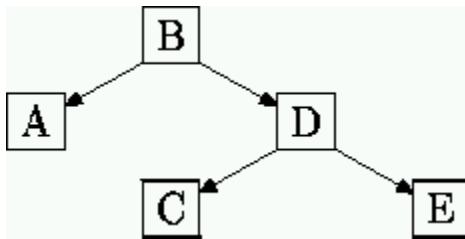
Profundidad promedio de este árbol: **1.5**  
 **$(1 + 1 + 2 + 2) / 4$**

- **Armando un árbol: ABCDE**
  - El orden de los datos afecta el balance del árbol  
(la distribución de las ramas)



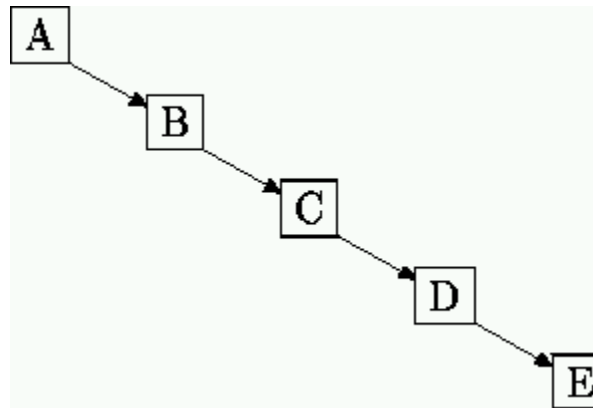
# Más sobre 'binary trees'

- **Arboles balanceados vs no balanceados**
  - Profundidad es inversamente proporcional a la velocidad de las búsquedas



2

1.5



4

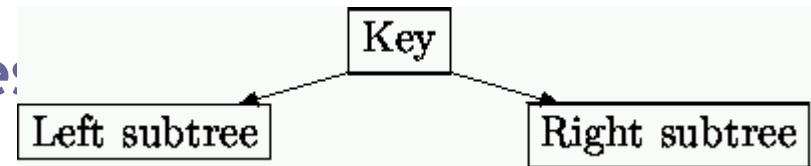
2.5

Profundidad máxima

Profundidad promedio

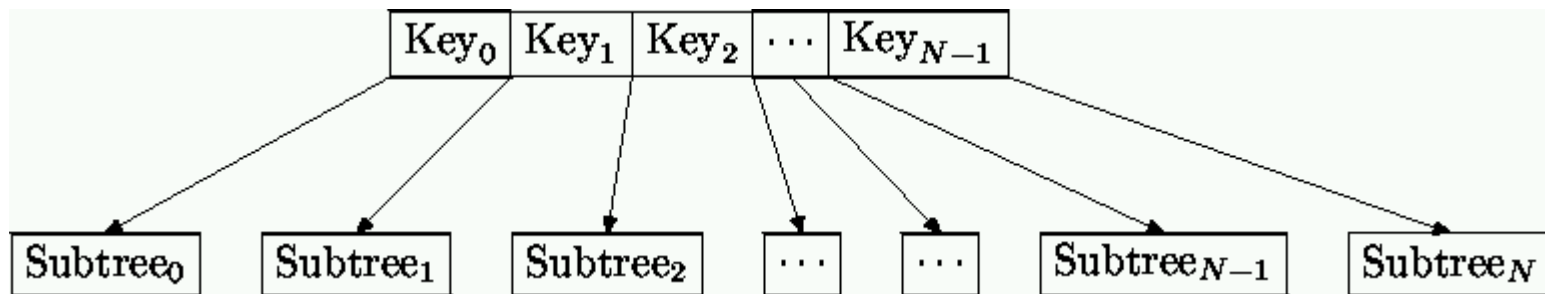
- B-Trees are not 'binary trees'

- Bushy Trees, B



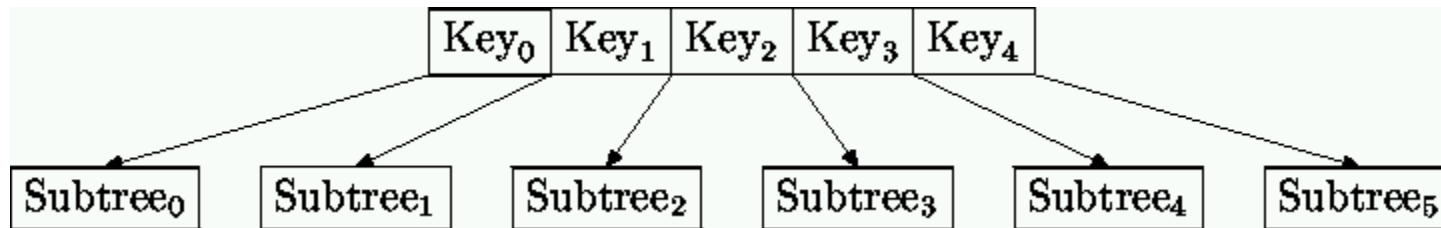
- Son una generalización de los árboles binarios

- Los nodos pueden contener más de una clave
  - Las claves dentro de un nodo están **ordenadas**
  - B-Tree de orden 2 => cada nodo contiene a lo sumo 3 claves
  - B-Tree de orden 3 => cada nodo contiene a lo sumo 4 claves
  - B-Tree de orden ***n*** => a lo sumo ***n-1*** claves

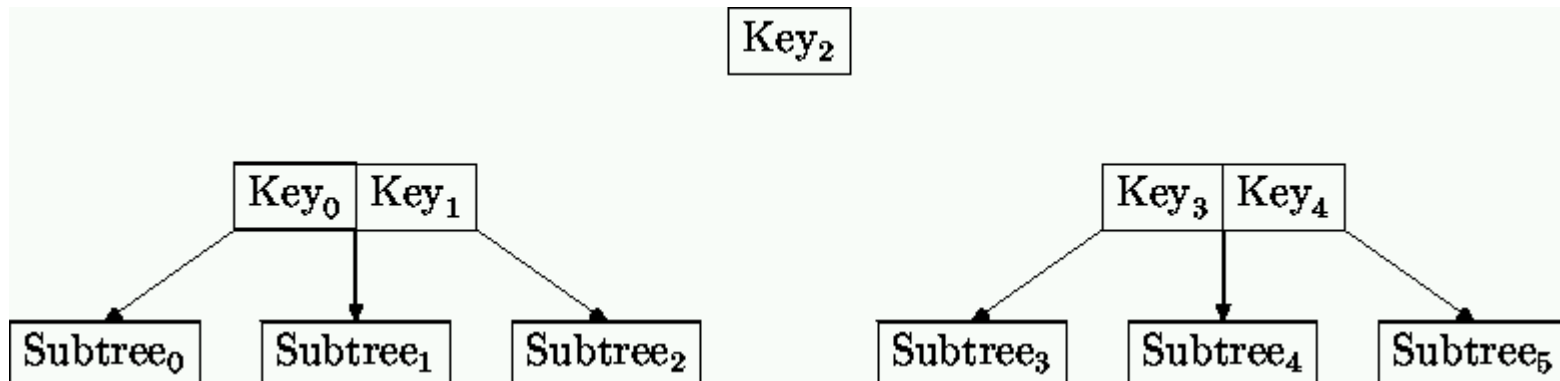




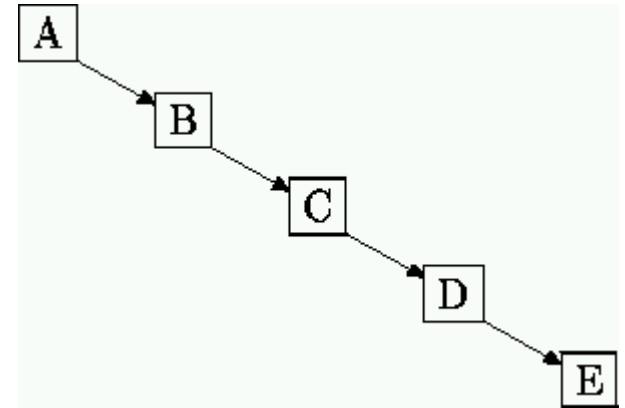
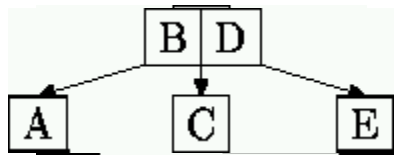
- **B-Tree de orden 5**
  - A lo sumo 4 claves por nodo
  - Agrego una nueva clave (un nuevo objeto/elemento)



- Como me excedo del límite, parto el nodo a la mitad



- **Ejemplo: ABCDE**
  - Esta cadena de texto, en un árbol binario, da un árbol desbalanceado
  - Probemos usando un B-Tree de orden 3



- **B-trees son estructuras de datos especializadas**
  - Uso en discos (lento)
  - Almacenamiento de grandes volúmenes de datos
- **Permiten realizar búsquedas extremadamente rápidas**
  - No se recorren todos y cada uno de los nodos para obtener una respuesta

## Recorriendo árboles

- **Depth-first**
  - Recorrido en profundidad primero
  - Se visita cada nodo 3 veces
    - Al visitarlo por primera vez (desde el nodo parental)
    - Al visitarlo por segunda vez desde el nodo hijo izquierdo
    - Al visitarlo nuevamente (desde el nodo hijo derecho)
- **Breadth-first (level order)**
  - Recorrido exhaustivo de cada nivel de profundidad del árbol (hacia lo ancho)
- **Ejemplos interactivos:**
  - <http://nova.umuc.edu/~jarc/idsv/lesson1.html>

**Importante:** no se busca en el total de los datos disponibles, sino sobre un subset pre-computado.

- Buscadores de páginas en internet
- PubMed / Entrez / SRS
- BLAST

# Motores de búsqueda: búsquedas simples

- Los motores de búsqueda ofrecen búsquedas simples
- No imponen restricciones
- El usuario tipea palabras libremente
- Usan estrategias para intentar “adivinar” la intención del usuario (sobre qué campo de la base de datos buscar)

## Ejemplo: term mapping - Entrez (PubMed)

- Entrez busca en una serie de listas para ver si la palabra que ingresaron se encuentra en alguna
- **MeSH (Medical Subject Headings):** vocabulario controlado utilizado para indexar artículos en PubMed.
- **Journals:** nombre completo del journal, abreviaturas usadas en MEDLINE y números ISSN.
- **Lista de frases:** cientos de miles de frases generadas a partir de MeSH y otros vocabularios controlados similares.
- **Indice de autores:** apellido e iniciales.
- **Stopwords:** palabras comunes, presentes en casi todos los registros de la base de datos (a, an, by, of, the ... )

# Búsquedas simples: pros / cons

- **Ventajas**

- rápidas de formular
- no hay que leer el manual
- ni hacer un curso 😊

- **Desventajas**

- poco selectivas



# Búsquedas avanzadas

- Presuponen un cierto conocimiento sobre la organización subyacente de los datos
- Hay que especificar sobre qué campos buscar:  
⇒ hay que conocer los campos
- **Entrez:** se especifican entre corchetes
- Tags predefinidos (hay que conocerlos)
  - `Escherichia coli[organism]`
  - `review[publication type]`
  - `attenuator[feature key]`
- **SRS:** formulario avanzado (no hay que conocer términos o tags)

- **Entrez provee además**

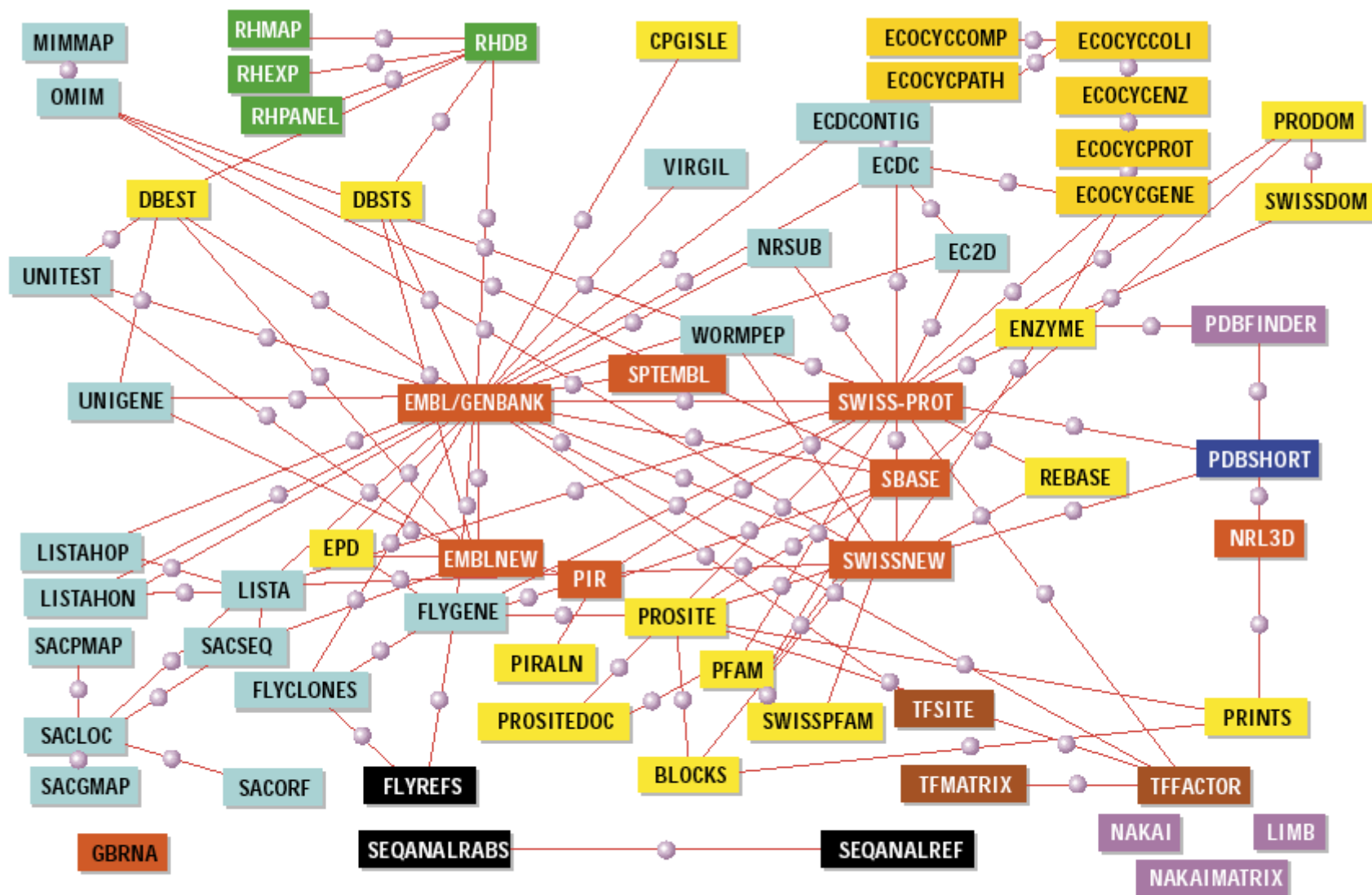
- **Límites:** especie de formulario avanzado que les permite limitar la búsqueda a un campo determinado, sin tener que conocer los tags)
- **History:** una historia de las búsquedas que van realizando. En cualquier momento pueden combinar búsquedas o volver sobre alguna de ellas
- **Preview/Index:** les permite probar una búsqueda (preview) y ver el número de registros que selecciona o ver los índices y el número de registros asociados a cada uno de ellos
- **Details:** permite analizar la traducción que realizó Entrez de la búsqueda que realizamos (uso de sinónimos, límites, etc)

# Operadores lógicos

- En búsquedas simples o avanzadas siempre tienen a disposición operadores lógicos para encadenar términos
- **AND (intersección)**
  - human AND genome
  - +human +genome
  - human && genome
- **OR (unión)**
  - human OR genome
  - human | | genome
- **NOT (subconjunto)**
  - human NOT genome

# Orden de los términos en un query

- El orden de los términos es importante
- Un query se evalúa de izquierda a derecha
  - **human NOT genome** no es lo mismo que **genome NOT human**
- Si el query tiene muchos términos pueden forzar el orden de evaluación usando paréntesis
  - **human AND cancer AND (cell OR science OR nature)**
  - **casein kinase NOT (human OR mouse)**



■ Sequence 
 ■ Protein Structure 
 ■ SeqRelated 
 ■ Genome 
 ■ Metabolic 
 ■ Literature 
 ■ Others 
 ■ Transfac 
 ■ Mapping

# The NAR Molecular Biology Database Collection

El numero de Enero de cada año está dedicado a bases de datos biológicas

OXFORD JOURNALS CONTACT US

## Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals > Life Sciences > Nucleic Acids Research > Volume 40, Issue D1

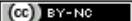
### Database issue

Volume 40 Issue D1 January 2012

[Clear](#) [Get All Checked Abstracts](#)

#### Articles

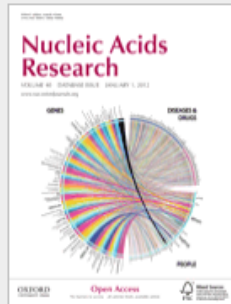
☐ Michael Y. Galperin and Xosé M. Fernández-Suárez  
**The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection**  
Nucl. Acids Res. (2012) 40(D1): D1-D8 doi:10.1093/nar/gkr1196  
» [Abstract](#) » [FREE Full Text \(HTML\)](#) » [Full Text \(PDF\)](#) » [Screen PDF](#) » [Database Summaries](#)  
OPEN ACCESS 

☐ Robert D. Finn, Paul P. Gardner, and Alex Bateman  
**Making your database available through Wikipedia: the pros and cons**  
Nucl. Acids Res. (2012) 40(D1): D9-D12 doi:10.1093/nar/gkr1195  
» [Abstract](#) » [FREE Full Text \(HTML\)](#) » [Full Text \(PDF\)](#) » [Screen PDF](#)  
OPEN ACCESS 

☐ Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Evan Bolton, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Ian M. Fingerman, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, Sergey Krasnov, David Landsman, David J. Lipman, Zhiyong Lu, Thomas L. Madden, Tam Madsen, Donna R. Mackeith, Anna Marchler-Bauer, Nadine Miller, Heng

» [Previous](#) | [Next Issue](#) »

**This Issue**  
January 2012 40 (D1)



» [Index By Author](#)  
» [Cover Image](#)  
» [Table of Contents \(PDF\)](#)

» [Articles](#)  
» [Front-Matter/Back-Matter](#)

2009

## **TcSNP:** a database of genetic variation in *Trypanosoma cruzi*

Alejandro A. Ackermann, Santiago J. Carmona and Fernán Agüero\*

☐ Author Affiliations

Instituto de Investigaciones Biotecnológicas, Universidad de San Martín – CONICET, San Martín, 1650, Argentina

\*To whom correspondence should be addressed. Tel: +54 11 4580 7255; Fax: +54 11 4752 9639; Email: [fernan@unsam.edu.ar](mailto:fernan@unsam.edu.ar)

Received August 15, 2008.  
Revision received September 24, 2008.  
Accepted October 18, 2008.

2011

## **PCDB:** a database of protein conformational diversity



Ezequiel I. Juritz, Sebastian Fernandez Alberti and Gustavo D. Parisi\*

☐ Author Affiliations

Universidad Nacional de Quilmes, Centro de Estudios e Investigaciones, Roqu Peña 352, Bernal, Argentina

\*To whom correspondence should be addressed. Tel: +54(011)43657100; Fax: +54(011)43657182; Email: [gusparisi@gmail.com](mailto:gusparisi@gmail.com)

Received Au  
Revision received Nov  
Accepted Nov

## **TDR Targets:** a chemogenomics resource for neglected diseases



María P. Magariños<sup>1</sup>, Santiago J. Carmona<sup>1</sup>, Gregory J. Crowther<sup>2</sup>, Stuart A. Ralph<sup>3</sup>, David S. Roos<sup>4</sup>, Dhanasekaran Shanmugam<sup>4</sup>, Wesley C. Van Voorhis<sup>2</sup> and Fernán Agüero<sup>1</sup>\*

☐ Author Affiliations

<sup>1</sup>Instituto de Investigaciones Biotecnológicas, Universidad de San Martín, San Martín, Buenos Aires, Argentina, <sup>2</sup>Department of Medicine, University of Washington, Seattle, WA, USA, <sup>3</sup>Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Victoria, Australia and <sup>4</sup>Department of Biology and Penn Genomics Institute, University of Pennsylvania, Philadelphia, PA, USA

☐ \*To whom correspondence should be addressed. Tel: +54 11 4580 7255 (Ext. 310); Fax: +54 11 4752 9639; Email: [fernan@unsam.edu.ar](mailto:fernan@unsam.edu.ar), [fernan.aguero@gmail.com](mailto:fernan.aguero@gmail.com)

Received September 15, 2011.  
Revision received October 24, 2011.  
Accepted October 25, 2011.

2012



# Beca de Iniciación (Doctoral)

Proyecto PICT-2013, Búsqueda de marcadores para diagnóstico molecular de la Enfermedad de Chagas.

Beca por 3 años

Contactar a: [fernan@unsam.edu.ar](mailto:fernan@unsam.edu.ar)

Análisis bioinformático de genomas, diseño de ensayos moleculares (ej PCR, microarrays de péptidos) para validación de marcadores



**UNSAM**  
UNIVERSIDAD  
NACIONAL DE  
SAN MARTÍN







- Nucleotide databases:

- Genbank: International Collaboration
  - NCBI (USA), EMBL (Europe), DDBJ (Japan and Asia)
  - European Nucleotide Archive (ENA) - Europe
  - Sequence Read Archive (SRA) - USA
- Organism specific databases
  - FlyBase
  - ChickBASE
  - pigbase
  - SGD (Saccharomyces Genome Database)

- Protein Databases:

- NCBI:
  - Genpept: Translated Proteins from Genbank Submissions
- EMBL
  - TrEMBL: Translated Proteins from EMBL Database
- SwissProt:
  - recibe secuencias peptídicas
  - cura y anota secuencias provenientes de TrEMBL

*(Gratuita para uso académico. Restricciones sobre los descubrimientos hechos utilizando la base de datos. La versión de 1998 es gratuita y libre de todas las restricciones.)*

- <http://www.expasy.ch> (última versión no-gratuita)
- NCBI tiene la última versión gratuita.

- **Structure databases:**

- PDB: Protein structure database.
  - <http://www.rscb.org/pdb/>
- MMDB: NCBI's version of PDB with entrez links.
  - <http://www.ncbi.nlm.nih.gov>
- SCOP: structural classification of proteins
  - family, superfamily, fold
- CATH: structural classification of proteins
  - class, architecture, topology, homology
- FSSP: fold classification based on structure-structure alignment

- **Genome Mapping Information:**

- <http://www.il-st-acad-sci.org/health/genebase.html>
- NCBI(Human)
- Genome Centers:
  - Stanford, Washington University, UCSC
- Research Centers and Universities

- **Literature databases:**

- NCBI: Pubmed: All biomedical literature.
  - [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)
  - Abstracts and links to publisher sites for
    - full text retrieval/ordering
    - journal browsing.
- Publisher web sites.

- **Pathways Database:**

- KEGG: Kyoto Encyclopedia of Genes and Genomes:  
[www.genome.ad.jp/kegg/kegg/html](http://www.genome.ad.jp/kegg/kegg/html)
- **BioCyc: Pathway/Genome Databases and Pathway Tools**
- [www.biocyc.org](http://www.biocyc.org)

- Es un Banco: no se intenta unificar datos.
  - No se pueden modificar las secuencias sin el consentimiento del autor (submitter).
  - No se intenta unificar (puede haber más de una secuencia para un locus/gen).
  - Puede haber registros de diversas calidades de secuencia y diferentes fuentes ==> Se separan en varias divisiones de acuerdo a:
    - Secuencias de alta calidad en divisiones taxonómicas.
      - PRI -> Primates
      - MAM -> Mamíferos
      - INV -> Invertebrados
    - Secuencias de baja calidad en divisiones uso-específicas.
      - GSS -> Genome Sequence Survey
      - EST -> Expressed Sequence Tags
      - HTG -> High Troughput Sequencing (unfinished contigs, BACs, cosmids, chromosomes).

- Redundante
- Con errores
- Difícil de actualizar
- Para poder corregir, mejorar y mantener actualizada la anotación de los registros, el NCBI creó RefSeq (colección curada de registros de GenBank)
  - toma records de GenBank y los actualiza/corrije
  - unifica para reducir redundancia
  - Accession numbers del tipo XX\_123456

# Bases de datos primarias

- Una base de datos primaria es un repositorio de datos derivados de un experimento o de conocimiento científico.
  - Genbank (Repositorio de secuencias nucleotídicas)
  - Protein DB, Swissprot
  - PDB
  - Pubmed (literatura)
  - Genome Mapping
  - Kegg (Kyoto Encyclopedia of Genes and Genomes, base de datos de vías metabólicas)

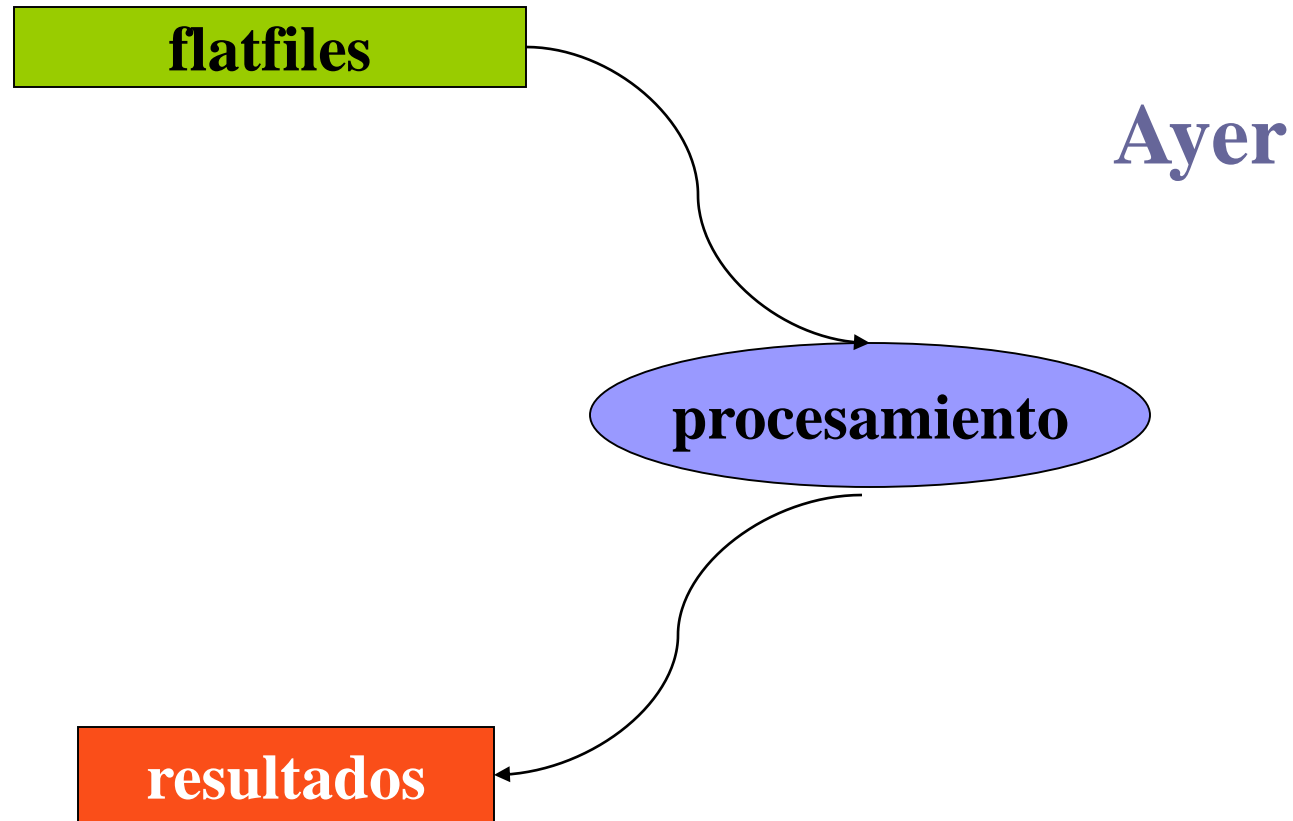


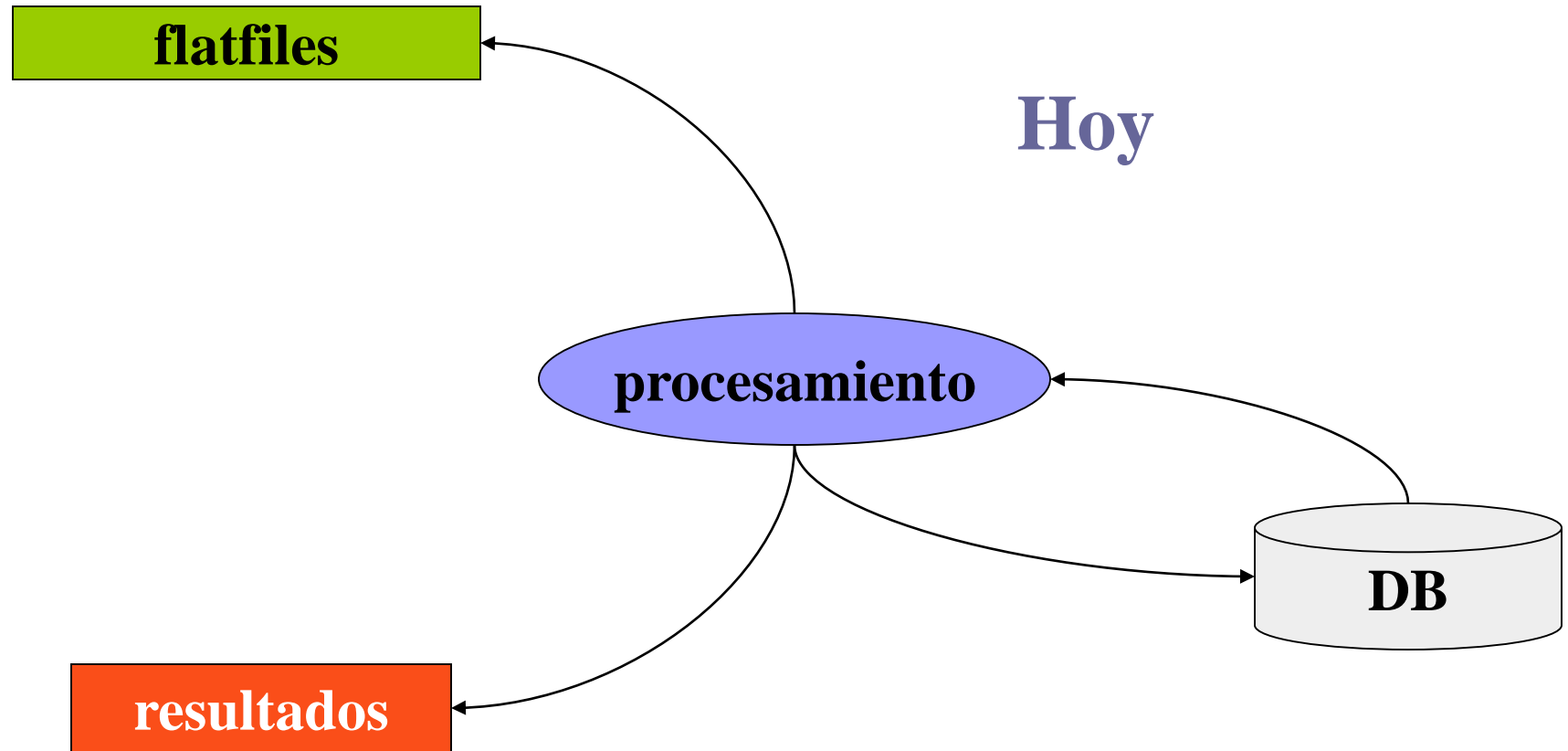
# Bases de datos secundarias

- Una base de datos secundaria contiene información derivada de otras fuentes (primarias, entre otras).
  - Refseq (Colección curada de GenBank en NCBI)
  - Unigene (Clustering de ESTs en NCBI)
- Las bases de datos organismo específicas son en general una mezcla entre primaria y secundaria.

# Formas de representar la información

- En una base de datos, la información está representada en forma compleja
- El usuario sin embargo tiene acceso a formas más simples de representación de los datos: flatfiles
- Ejemplos de archivos simples (flatfiles): FASTA, GenBank/EMBL
- En general son archivos de texto (o HTML en el caso de páginas web) conteniendo todos los datos de un registro, organizados de alguna forma particular.
- Ejemplos:
  - GenBank/EMBL, FASTA, Swissprot





# Ejemplo de formato: GenBank

LOCUS          XELRHODOP                  1684 bp      mRNA      linear      VRT 15-FEB-1996  
DEFINITION    Xenopus laevis rhodopsin mRNA, complete cds.  
ACCESSION     L07770  
VERSION       L07770.1   GI:214734  
KEYWORDS      G protein-coupled receptor; phototransduction protein; retinal  
              protein; rhodopsin; transmembrane protein.  
SOURCE        Xenopus laevis (African clawed frog)  
  ORGANISM    Xenopus laevis  
              Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
              Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;  
              Xenopodinae; Xenopus.  
REFERENCE     1   (bases 1 to 1684)  
  AUTHORS     Knox,B.E., Scalzetti,L.C., Batni,S. and Wang,J.Q.  
  TITLE       Molecular cloning of the abundant rhodopsin and transducin from  
              Xenopus laevis  
  JOURNAL     Unpublished (1992)  
REFERENCE     2   (bases 1 to 1684)  
  AUTHORS     Batni,S., Scalzetti,L., Moody,S.A. and Knox,B.E.  
  TITLE       Characterization of the Xenopus rhodopsin gene  
  JOURNAL     J. Biol. Chem. 271 (6), 3179-3186 (1996)  
  MEDLINE     96216396  
  PUBMED      8621718  
COMMENT       Original source text: Xenopus laevis (tissue library: lambda-ZAPII)  
              adult retina cDNA to mRNA.

# Ejemplo de formato: GenBank (cont)

```
FEATURES                     Location/Qualifiers
    source                    1..1684
                              /organism="Xenopus laevis"
                              /db_xref="taxon:8355"
                              /tissue_type="retina"
                              /dev_stage="adult"
                              /tissue_lib="lambda-ZAPII"
    CDS                       110..1174
                              /note="gene accession number U23808"
                              /codon_start=1
                              /product="rhodopsin"
                              /protein_id="AAC42232.1"
                              /db_xref="GI:214735"
                              /translation="MNGTEGPNFYVPMSNKTGVVRSPFDYPQYYLAEPWQYSALAAAYM
FLLILLGLPINFMTLFVTIQHKKLRTPLNYILLNLVFNHFMVLCGFTVTMYTSMHGY
FIFGQTGCYIEGFFATLGGEVALWSLVLAVERYMVVCKPMANFRFGENHAIMGVAFT
WIMALSCAAPPLFGWSRYIPEGMQCSCGVDYYTLKPEVNNESFVIYMFIVHFTIPLIV
IFFCYGRLLCTVKEAAAQQQESATTQKAEKEVTRMVVIMVVFFLICWVPYAYVAFYIF
THQGSNFGPVMFMTVPAFFAKSSAIYNPVIYIVLNKQFRNCLITTLCCGKNPFGDEDGS
SAATSKTEASSVSSSQVSPA"
    misc_feature              189..1684
                              /note="sequenced from clone pXOP71"
    variation                  1224
                              /note="clone pXOP5 contained deletion from bp 1224-1534"
```

# Bases de datos: formatos: EMBL

```
ID   XLRHODOP   standard; RNA; VRT; 1684 BP.
XX
AC   L07770;
XX
SV   L07770.1
XX
DT   12-DEC-1992 (Rel. 34, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 7)
XX
DE   Xenopus laevis rhodopsin mRNA, complete cds.
XX
KW   G protein-coupled receptor; phototransduction protein; retinal protein;
KW   rhodopsin; transmembrane protein.
XX
OS   Xenopus laevis (African clawed frog)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Amphibia;
OC   Batrachia; Anura; Mesobatrachia; Pipoidae; Pipidae; Xenopodinae; Xenopus.
XX
RN   [1]
RP   1-1684
RA   Knox B.E., Scalzetti L.C., Batni S., Wang J.Q.;
RT   "Molecular cloning of the abundant rhodopsin and transducin from Xenopus
RT   laevis";
RL   Unpublished.
XX
RN   [2]
RP   1-1684
RX   MEDLINE; 96216396.
RA   Batni S., Scalzetti L., Moody S.A., Knox B.E.;
RT   "Characterization of the Xenopus rhodopsin gene";
RL   J. Biol. Chem. 271(6):3179-3186(1996).
XX
79  DR   SWISS-PROT; P29403; OPSD_XENLA.
```

# Bases de datos: formatos: EMBL (cont)

FH	Key	Location/Qualifiers
FH		
FT	source	1..1684
FT		/db_xref="taxon:8355"
FT		/organism="Xenopus laevis"
FT		/dev_stage="adult"
FT		/tissue_type="retina"
FT		/tissue_lib="lambda-ZAPII"
FT	CDS	110..1174
FT		/codon_start=1
FT		/db_xref="SWISS-PROT:P29403"
FT		/note="gene accession number U23808"
FT		/product="rhodopsin"
FT		/protein_id="AAC42232.1"
FT		/translation="MNGTEGPNFYVPMSNKTGVVRSPFDYPQYYLAEPWQYSALAAAYMF
FT		LLILLGLPINFMTLFTVIQHKKLRTPLNYILLNLVFNHFMVLCGFTVTMYTSMHGYFI
FT		FGQTGCYIEGFFATLGGEVALWSLVLAVERYMVVCKPMANFRFGENHAIMGVAFTWIM
FT		ALSCAAPPLFGWSRYIPEGMQCSCGVDYYTLKPEVNNESFVIYMFIVHFTIPLIVIFFC
FT		YGRLLCTVKEAAQQQESATTQKAEKEVTRMVVIMVVFFLICWVPYAYVAFYIFTHQGS
FT		NFGPVMFTVPAFFAKSSAIYNPVIYIVLNKQFRNCLITTLCCGKNPFGDEDGSSAATSK
FT		TEASSVSSSQVSPA"
FT	misc_feature	189..1684
FT		/note="sequenced from clone pXOP71"
FT	variation	1224
FT		/note="clone pXOP5 contained deletion from bp 1224-1534"



# Feature tables

- Una de las regiones más importantes (en cuanto a cantidad de información)
- El espectro de ‘features’ que se pueden representar es amplio e incluye regiones de una secuencia que pueden:
  - contar con una función biológica
  - afectar o ser el resultado de la expresión de una función biológica
  - interaccionar con otras moléculas
  - afectar la replicación de una secuencia
  - afectar o ser el resultado de recombinación de diferentes secuencias
  - ser reconocidas como una unidad repetitiva
  - tener estructura secundaria o terciaria
  - mostrar variación
  - haber sido corregidas o revisadas

- **Feature key [fkey]**

- una palabra clave que indica un grupo funcional
- **Ejemplos:** source, CDS, RBS, repeat\_region

- **Location**

- instrucciones para localizar el feature
- **Ejemplos:** 1..1000, 23..400, join(544..589,688..1032)

- **Qualifiers**

- información adicional acerca del feature

# Feature keys

Key	Description
-----	-----
<b>attenuator</b>	Sequence related to transcription termination
<b>C_region</b>	Constant region of immunoglobulin light and heavy chain, and T-cell receptor alpha, beta and gamma chains
<b>CAAT_signal</b>	'CAAT box' in eukaryotic promoters
<b>CDS</b>	Sequence coding for amino acids in protein (includes stop codon)
<b>conflict</b>	Independent determinations differ
<b>D-loop</b>	Displacement loop
<b>D-segment</b>	Diversity segment of immunoglobulin heavy chain and T-cell receptor beta-chain
<b>enhancer</b>	Cis-acting enhancer of promoter function
<b>exon</b>	Region that codes for part of spliced mRNA
<b>GC_signal</b>	'GC box' in eukaryotic promoters
<b>iDNA</b>	Intervening DNA eliminated by recombination
<b>intron</b>	Transcribed region excised by mRNA splicing
<b>J_segment</b>	Joining segment of immunoglobulin light and heavy chains, And T-cell receptor alpha, beta and gamma-chains
<b>LTR</b>	Long terminal repeat
<b>mat_peptide</b>	Mature peptide coding region (does not include stop codon)
<b>misc_binding</b>	Miscellaneous binding site
<b>misc_difference</b>	Miscellaneous difference feature also used to describe variability that arises as a result of genetic manipulation (e.g. site directed mutagenesis).
...	

# Feature keys [fkey]

- Constituyen un vocabulario controlado, organizado en forma jerárquica

## gene

- \* **misc\_signal**
- \* **promoter**
- \* CAAT\_signal
- \* TATA\_signal
- \* -35\_signal
- \* -10\_signal
- \* GC\_signal
- \* **RBS**
- \* **polyA\_signal**
- \* **enhancer**
- \* **attenuator**
- \* **terminator**
- \* **rep\_origin**

## misc\_RNA

- \* **prim\_transcript**
- \* **precursor\_RNA**
- \* **mRNA**
- \* 5'clip
- \* 3'clip
- \* 5'UTR
- \* 3'UTR
- \* exon
- \* **CDS**
- \* sig\_peptide
- \* transit\_peptide
- \* mat\_peptide
- \* **intron**
- \* **polyA\_site**
- \* **rRNA**
- \* **tRNA**
- \* **scRNA**
- \* **snRNA**
- \* **snoRNA**

- A location can be one of the following:
  - A single base
  - A contiguous span of bases (1..1009)
  - A site between two bases (23^24)
  - A single base chosen from a range of bases (23.79)
  - A single base chosen from among two or more specified bases
  - A joining of sequence spans (join(1..1009,2130..5401))
  - A reference to an entry other than the one to which the feature belongs i.e. a remote entry), followed by a location referring the remote sequence.

# Qualifiers

- `/qualifier_name=value`
  - Free text
  - Controlled vocabulary or enumerated values
  - Citations or reference numbers
  - Sequences
  - Feature labels

Qualifier	Description
-----	-----
<code>/allele</code>	Name of the allele for given gene.
<code>/anticodon</code>	Location of the anticodon of tRNA and the amino acid for which it codes
<code>/bound_moiety</code>	Moiety bound
<code>/cell_line</code>	Cell line from which the sequence was obtained
<code>/cell_type</code>	Cell type from which the sequence was obtained
<code>/chromosome</code>	Chromosome from which the sequence was obtained
<code>/citation</code>	Reference to a citation providing the claim of or evidence for a feature
<code>/clone</code>	Clone from which the sequence was obtained
<code>/clone_lib</code>	clone library from which the sequence was obtained
...	

# Feature tables: ejemplos

source	1..1509 /organism="Mus musculus" /strain="CD1"
promoter	<1..9 /gene="ubc42"
mRNA	join(10..567,789..1320) /gene="ubc42"
CDS	join(54..567,789..1254) /gene="ubc42" /product="ubiquitin conjugating enzyme" /function="cell division control" /translation="MVSSFLLAEYKNLIVNPSEHFKISVNEDNLTEGPPDTLY QKIDTVLLSVISLLNEPNPDSPANVDAAKSYRKYLYKEDLESYPMEKSLDECS AEDIEYFKNVPVNVLPVPSDDYEDEEMEDGTYILTYDDEDEEEDEEMDDE"
exon	10..567 /gene="ubc42" /number=1
intron	568..788 /gene="ubc42" /number=1
exon	789..1320 /gene="ubc42" /number=2
polyA_signal	1310..1317 /gene="ubc42"

**Un gen eucariótico**

# Feature tables: ejemplos

```
source          1..9430
                /organism="Lactococcus sp."
                /strain="MG1234"
-35_signal      160..165
                /gene="galA"
                /evidence=EXPERIMENTAL
-10_signal      179..184
                /gene="galA"
                /evidence=EXPERIMENTAL
CDS             405..1934
                /gene="galA"
                /product="galactose permease"
                /function="galactose transporter"
                /evidence=EXPERIMENTAL
CDS            2003..3001
                /gene="galM"
                /product="aldose 1-epimerase"
                /EC_number="5.1.3.3"
                /function="mutarotase"
CDS            3235..4537
                /gene="galk"
                /product="galactokinase"
                /EC_number="2.7.1.6"
                /evidence=EXPERIMENTAL
```

**Un operon bacteriano**



# Feature tables: ejemplos

```
source      1..5300
             /organism="Cloning vector pABC"
             /lab_host="Escherichia coli"
             /focus

source      1..5138
             /organism="Escherichia coli"
             /strain="K12"

source      5139..5247
             /organism="Aequorea victoria"
             /dev_stage="adult"

source      5248..5300
             /organism="Escherichia coli"
             /strain="K12"

CDS          join(complement(<1..799),complement(5080..5120))
             /gene="mob1"
             /product="mobilization protein 1"

CDS          complement(1697..2512)
             /gene="Km"
             /product="kanamycin resistance protein"

CDS          3037..3711
             /gene="rep1"
             /product="replication protein 1"

CDS          complement(4170..4829)
             /gene="Cm"
             /product="chloramphenicol resistance protein"
```

**Un vector de  
clonado (circular)**

# Feature tables: qualifiers (cont)

- Cada feature key tiene asociada una descripción y una serie de calificadores posibles

Feature Key	attenuator
Organism scope	prokaryotes
Molecule scope	DNA
Definition	<ol style="list-style-type: none"><li>1) region of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons;</li><li>2) sequence segment located between the promoter and the first structural gene that causes partial termination of transcription</li></ol>
Optional qualifiers	<pre>/citation=[number] /db_xref="&lt;database&gt;:&lt;identifier&gt;" /evidence=&lt;evidence_value&gt; /gene="text" /label=feature_label /locus_tag="text" (single token) /map="text" /note="text" /phenotype="text" /usedin=accnum:feature_label</pre>

# Formato FASTA

>identificador texto descriptivo ↵

Secuencia de nucleótidos o amino ácidos ↵

en múltiples líneas si es necesario ↵

en múltiples líneas si es necesario ↵

*\n = newline, enter, return*



## Ejemplo:

>gi|41|emb|X63129.1|BTA1AT B.taurus mRNA for alpha-1-anti-trypsin

GACCAGCCCTGACCTAGGACAGTGAATCGATAATGGCACTCTC

CATCACGCGGGGCCTTCTGCTGCTGGC

>gi|214734|L07770|XELRHODOP Xenopus laevis rhodopsin mRNA

ACCGTACGACCGGTGACCTGTGACCAACAACCCGGGTGAAAAC

ACGTCTCGACGACAGTGAGACTG

Hay otros formatos más amigables para la computadora

- Ejemplo: cuando estamos escribiendo *programas* o *scripts* para automatizar una tarea
- **ASN.1 (NCBI)**
  - **Abstract Syntax Notation One**(notación sintáctica abstracta 1, **ASN.1**) es una norma para representar datos independientemente de la máquina que se esté usando y sus formas de representación internas
  - <http://es.wikipedia.org/wiki/ASN.1>
- **XML**
  - XML, siglas en inglés de Extensible Markup Language (lenguaje de marcas extensible), es un metalenguaje extensible de etiquetas
  - [http://es.wikipedia.org/wiki/Extensible\\_Markup\\_Language](http://es.wikipedia.org/wiki/Extensible_Markup_Language)

# Ejemplo ASN.1

## Xenopus laevis rhodopsin mRNA, complete cds

GenBank: L07770.1

```
Seq-entry ::= set { level 1 , class nuc-prot , descr { source { org { taxname "Xenopus laevis" , common "African clawed frog" , db { { db "taxon" , tag id 8355 } } , orname { name binomial { genus "Xenopus" , species "laevis" } , lineage "Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidae; Pipidae; Xenopodinae; Xenopus; Xenopus" , gcode 1 , mgcode 2 , div "VRT" } } , subtype { { subtype tissue-type , name "retina" } , { subtype dev-stage , name "adult" } , { subtype tissue-lib , name "lambda-ZAPII" } } } , pub { pub { gen { serial-number 1 } , gen { cit "Unpublished" , authors { names std { { name name { last "Knox" , initials "B.E." } } , { name name { last "Scalzetti" , initials "L.C." } } , { name name { last "Batni" , initials "S." } } , { name name { last "Wang" , initials "J.Q." } } } , date std { year 1992 } , title "Molecular cloning of the abundant rhodopsin and transducin from Xenopus laevis" } } } , pub { pub { gen { serial-number 2 } , muid 96216396 , article { title { name "Characterization of the Xenopus rhodopsin gene." } , authors { names std { { name name { last "Batni" , initials "S." } } , { name name { last "Scalzetti" , initials "L." } } , { name name { last "Moody" , initials "S.A." } } , { name name { last "Knox" , initials "B.E." } } } , affil str "Department of Biochemistry and Molecular Biology, State University of New York Health Science Center, Syracuse, New York 13210, USA." } , from journal { title { iso-jta "J. Biol. Chem." , ml-jta "J Biol Chem" , issn "0021-9258" , name "The Journal of biological chemistry." } , imp { date std { year 1996 , month 2 , day 9 } , volume "271" , issue "6" , pages "3179-3186" , language "eng" } } , ids { pubmed 8621718 , medline 96216396 } } , pmid 8621718 } } , create-date std { year 1993 , month 4 , day 28 } , update-date std { year 1996 , month 2 , day 15 } } , seq-set { seq { id { genbank { name "XELRHODOP" , accession "L07770" , version 1 } , gi 214734 } , descr { title "Xenopus laevis rhodopsin mRNA, complete cds." , molinfo { biomol mRNA } , genbank { source "Xenopus laevis (tissue library: lambda-ZAPII) adult retina cDNA to mRNA." , keywords { "G protein-coupled receptor" , "phototransduction protein" , "retinal protein" , "rhodopsin" , "transmembrane protein" } , entry-date std { year 1996 , month 2 , day 15 } } } , inst { repr raw , mol rna , length 1684 , strand ss , seq-data ncbi2na 'AC8127D2FA8D129F72A35FEA400201120A4F7F731080A1FCC89E714E068120AD40FFCED54ED41007AAEB18254F63C574B3C7C92253A433D247A79F13BD79D35E7EAF14341F4E17EFEF14D491080748115701C4D79E05EB3F94347D3AD7BBABD1AE10EC45D0E469C7D37FA501EBE713E0A7DFE711FAEB82E95DEB47ACB3E96F808CCEBADE4254E941F58F6A8814E73CEAEC97D13A34E9FEDFB9E75D777DA3AD48C4D522A0E439D39A2C871C447825E2B410E0D7FBCDC4EF4FB51F453D578FB4DF7DE73AD979DE47B408A79254904A0DE71454829E2022B4520EBEF34EB6FF7D78DEFAB95CE5CEE93DC4DF45452A7707FA54B7D385B549FDFE5089DE73710D7B4DC4FB7E0404BD6C1E7E34515EE7BA020D4F6B8E08E9D7925175084827DF7B77D494AEDD793089F452A7B74AB59E5D110F54D1F095DEC7EF9829020F512FF0CFC554F79505FA11EC2278553C79EA0A5427FBE4F78EE35FD24803ABA3D0E0FD1429EC4C10C13CB78291754548839047CFCDDEDFF7E133E39E7DCF4EB470400B54FF10E41E02C3B3FFB0CC304CFD390DD779F3E90ADE332E28C8494855F93C035ECF003F7F 90B'H } , annot { { data ftable { { data imp { key "variation" } , comment "clone pXOP5 contained deletion from bp 1224-1534" , location pnt { point 1223 , id gi 214734 } } } , { data imp { key "misc_feature" } , comment "sequenced from clone pXOP71" , location int { from 188 , to 1683 , id gi 214734 } } } } } , seq { id { genbank { accession "AAC42232" , version 1 } , gi 214735 } , descr { title "rhodopsin [Xenopus laevis]" , molinfo { biomol peptide , tech concept-trans } } , inst { repr raw , mol aa , length 354 , seq-data ncbieaa "MNGTEGPNFYVPM SNKTGVVRSFPDYPQYYLAEPWQYSALAA YMFL LILGLPINFMTL FVTIQHKKLRTP LNYILLNLV FANHFVMLCGFTV TMYTSMHG YFIFGQTGCYIEGFFATLGG E VALWSLVVLAVERYM VVCKPMANFRFGENHAIMGVAFTWMH LSCAAPPTLVFGWSRYIPEGMQCSCGV DYITLKP EVNNE SFVIYMFIVHFTIPLIVIFFCYGRLLCTVKEAAAAQQOESAT TQKAEKEVTRMVVIMVVFLLICWVPYAYVAFYIFTHOGS NFGPVMFTVPAF
```

# Ejemplo ASN.1

## **Xenopus laevis rhodopsin mRNA, complete cds**

GenBank: L07770.1

```
Seq-entry ::= set {
  level 1 ,
  class nuc-prot ,
  descr {
    source {
      org {
        taxname "Xenopus laevis" ,
        common "African clawed frog" ,
        db {
          {
            db "taxon" ,
            tag id 8355 } } ,
        orgname {
          name binomial {
            genus "Xenopus" ,
            species "laevis" } ,
          lineage "Eukaryota; Metazoa; Chordata; Craniata;
Vertebrata; Euteleostomi; Amphibia; Batrachia; Anura; Mesobatrachia;
Pipoidea; Pipidae; Xenopodinae; Xenopus; Xenopus" ,
          gcode 1 ,
          mgcode 2 ,
          div "VRT" } } ,
        subtype {
          {
            subtype tissue-type ,
            name "retina" } ,
          {
            subtype dev-stage ,
            name "adult" } ,
          {
            subtype tissue-lib ,
            name "lambda-ZAPII" } } } ,
        pub {
          pub {
            gen {
              serial-number 1 } ,
            gen {
              cit "Unpublished" ,
```

```
        authors {
          names std {
            { name name {
              last "Knox" ,
              initials "B.E." } } ,
            { name name {
              last "Scalzetti" ,
              initials "L.C." } } ,
            { name name {
              last "Batni" ,
              initials "S." } } ,
            { name name {
              last "Wang" ,
              initials "J.Q." } } } } ,
          date std {
            year 1992 } ,
            title "Molecular cloning of the abundant rhodopsin
and transducin from Xenopus laevis" } } } ,
        pub {
          pub {
            gen {
              serial-number 2 } ,
            muid 96216396 ,
            article {
              title {
                name "Characterization of the Xenopus
rhodopsin gene." } ,
                authors {
                  names std {
                    { name name {
                      last "Batni" ,
                      initials "S." } } ,
                    { name name {
                      last "Scalzetti" ,
                      initials "L." } } ,
                    { name name {
                      last "Moody" ,
                      initials "S.A." } } ,
                    { name name {
                      last "Knox" , initials "B.E." } } } ,
                  affil str "Department of Biochemistry and Molecular
Biology, State University of New York Health Science Center, Syracuse,
New York 13210, USA." } ,
```