

Introducción a la Bioinformática

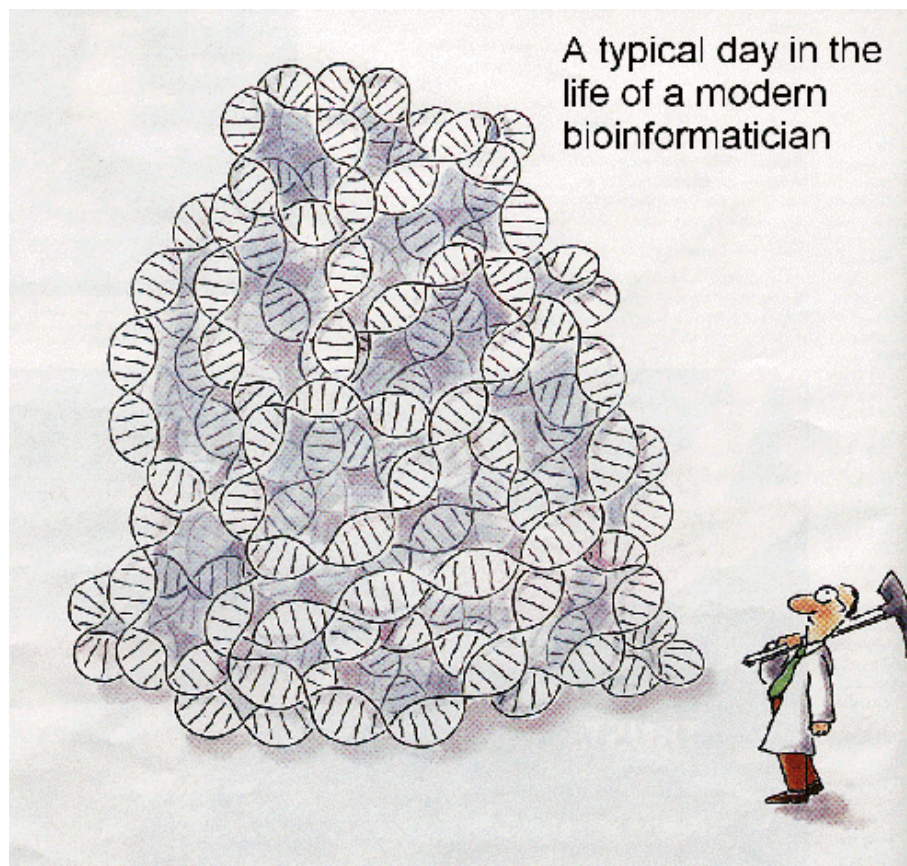
Información biológica en formato electrónico

Bases de datos

Fernán Agüero
Instituto de Investigaciones Biotecnológicas
UNSAM

Slot 1

INFORMACIÓN BIOLÓGICA EN FORMATO ELECTRÓNICO



- **Automatización y robotización** del laboratorio
 - Experimentos de **alta escala** (*high-throughput*)
 - Grandes **volúmenes de datos**
 - Conjuntos de **datos complejos**
 - Datos **interconectados**
-
- Genómica → Secuencias
 - Proteómica → Mass spectra → Secuencias
 - Metabolómica → Mass spectra → fármacos, drogas
 - Spatial Omics → Secuencias + Localizaciones
 - High-content microscopy → Imágenes
 - High-content screenings → datos complejos

Qué tipos de datos manejamos en biotecnología?

- Secuencias (de ADN, de ARN, de proteínas, polímeros)
- Estructuras (de proteínas, de complejos macromoleculares)
- Taxonomía (organismos, especies, relaciones filogenéticas)
- Vías metabólicas celulares (reacciones químicas)
- Compuestos químicos (metabolitos, fármacos)
- Bibliografía (papers, artículos científicos)

Qué tipos de datos manejamos y como los representamos?

- Secuencias biológicas y formatos

>U54469.1

```
CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTTGCCCTGTTCAACAATCGATA
GCTGCCTTTGGCCACCAAAATCCCAAACCTTAATTAAAGAATTAAATAATTCGAATAATAATTAAGCCCAG
TAACCTACGCAGCTTGAGTGCCTAACCGATATCTAGTATACATTTTCGATACATCGAAATCATGGTAGTGT
TGGAGACGGAGAAGGTAAGACGATGATAGACGGCGAGCCGCATGGGTTCGATTTGCGCTGAGCCGTGGCA
GGGAACAACAAAAACAGGGTTGTTGCACAAGAGGGGAGGCGATAGTCGAGCGGAAAAGAGTGCAGTTGGC
```

>ENA|U54469|U54469.1 Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene, complete cds, alternatively spliced.

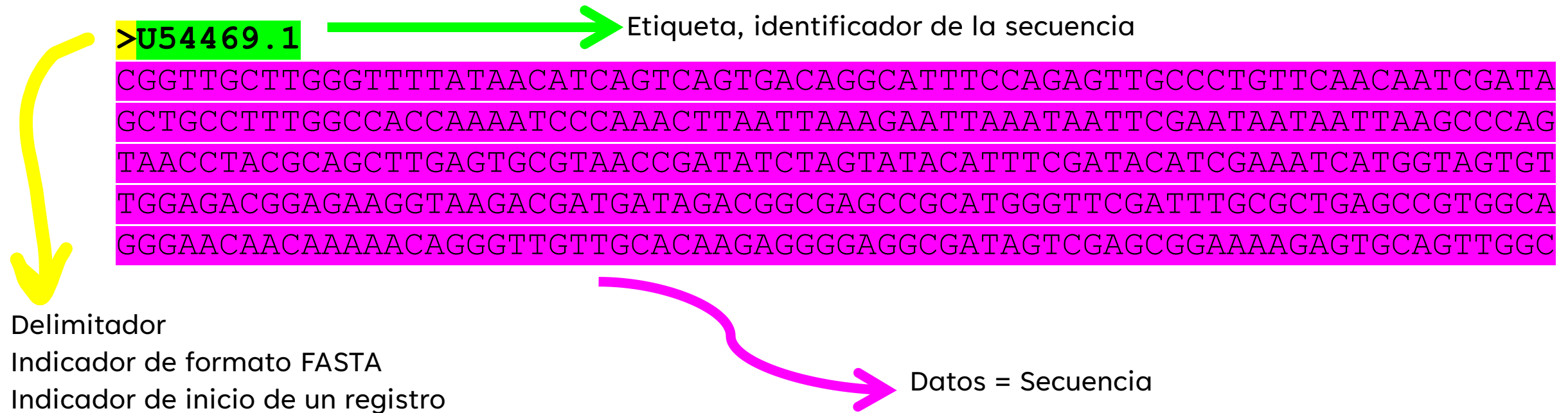
```
CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTTGCCCTGTTCAACAATCGATA
GCTGCCTTTGGCCACCAAAATCCCAAACCTTAATTAAAGAATTAAATAATTCGAATAATAATTAAGCCCAG
TAACCTACGCAGCTTGAGTGCCTAACCGATATCTAGTATACATTTTCGATACATCGAAATCATGGTAGTGT
TGGAGACGGAGAAGGTAAGACGATGATAGACGGCGAGCCGCATGGGTTCGATTTGCGCTGAGCCGTGGCA
GGGAACAACAAAAACAGGGTTGTTGCACAAGAGGGGAGGCGATAGTCGAGCGGAAAAGAGTGCAGTTGGC
```

Qué podemos decir acerca de los datos? Qué tipos de datos vemos? Y cuántos distintos tipos de datos identifican?

Y del formato?

Formato FASTA

- FASTA: el formato más común y popular para trabajar con secuencias



Formato FASTA

- FASTA: el formato más común y popular para trabajar con secuencias

```
>ENA|U54469|U54469.1 Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene, complete cds, alternatively spliced.  
CGGTTGCTTGGGTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTTGCCCTGTTCAACAATCGATA  
GCTGCCTTTGGCCACCAAATCCCAAACCTAATTAAAGAATTAAATAATTCGAATAATAATTAAGCCCAG  
TAACCTACGCAGCTTGAGTTCGTAAACCGATATCTAGTATACATTTTCGATACATCGAAATCATGGTAGTGT  
TGGAGACGGAGAAGGTAAGACGATGATAGACGGCGAGCCGCATGGGTTTCGATTTGCGCTGAGCCGTGGCA  
GGGAACAACAAAAACAGGGTTGTTGCACAAGAGGGGAGGCGATAGTCGAGCGGAAAAGAGTGCAGTTGGC
```



1. Header
2. Feature Table
3. Sequence

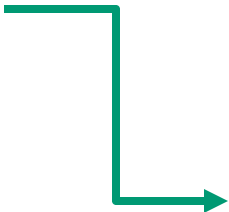
Ejemplo:

<https://www.ebi.ac.uk/ena/browser/api/embl/U54469.1>

ID U54469; SV 1; linear; genomic DNA; STD; INV; 2881 BP.
XX
AC U54469;
XX
DT 19-MAY-1996 (Rel. 47, Created)
DT 23-JUN-2017 (Rel. 133, Last updated, Version 5)
XX
DE Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene,
DE complete cds, alternatively spliced.
XX
KW .
XX
OS Drosophila melanogaster (fruit fly)
OC Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
Pterygota;
OC Neoptera; Holometabola; Diptera; Brachycera; Muscomorpha; Ephydroidea;
OC Drosophilidae; Drosophila; Sophophora.
XX
RN [1]
RP 1-2881
RX DOI; .1074/jbc.271.27.16393.
RX PUBMED; 8663200.
RA Lavoie C.A., Lachance P.E., Sonenberg N., Lasko P.;
RT "Alternatively spliced transcripts from the Drosophila eIF4E gene
produce
RT two different Cap-binding proteins";
RL J Biol Chem 271(27):16393-16398(1996).
XX
RN [2]
RP 1-2881
RA Lasko P.F.;
RT ;
RL Submitted (09-APR-1996) to the INSDC.
RL Paul F. Lasko, Biology, McGill University, 1205 Avenue Docteur
Penfield,
RL Montreal, QC H3A 1B1, Canada
XX
DR MD5; 303680f06f3441eb47a0de3a028a8d06.
DR FLYBASE; FBgn0015218; eIF-4E.

1

- 1. Header
- 2. Feature Table
- 3. Sequence



FH	Key	Location/Qualifiers
FH		
FT	source	1..2881
FT		/organism="Drosophila melanogaster"
FT		/chromosome="3"
FT		/map="67A8-B2"
FT		/mol_type="genomic DNA"
FT		/db_xref="taxon:7227"
FT	gene	80..2881
FT		/gene="eIF4E"
FT	mRNA	join(80..224,892..1458,1550..1920,1986..2085,2317..2404,2466..2881)
FT		/gene="eIF4E"
FT		/product="eukaryotic initiation factor 4E-I"
FT	mRNA	join(80..224,1550..1920,1986..2085,2317..2404,2466..2881)
FT		/gene="eIF4E"
FT		/product="eukaryotic initiation factor 4E-II"
FT	CDS	join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
FT		/codon_start=1
FT		/gene="eIF4E"
FT		/product="eukaryotic initiation factor 4E-II"
FT		/note="Method: conceptual translation with partial peptide sequencing"
FT		/db_xref="GOA:P48598"
FT		/db_xref="InterPro:IPR001040"
FT		/db_xref="PDB:5T47"
FT		/db_xref="PDB:5T48"
FT		/db_xref="UniProtKB/Swiss-Prot:P48598"
FT		/protein_id="AAC03524.1"
FT		/translation="MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGE
FT		PAGNTATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQNEITSFDTVED
FT		FWSLYNHIKPPSEIKLGSDYSLFKKNIRPMWEDAANKQGGRWVITLNKSSKTDLDNLWL
FT		DVLLCLIGEAFDHSDQICGAVINIRGKSNKISIWTDAGNNEEAALIGHKLRDALRLGR
FT		NNSLQYQLHKDTMVKQGSNVKSIYTL"

Más sobre tablas de features

Se usan para describir y mapear características sobre las secuencias

Permiten asignar anotaciones a **posiciones** en la secuencia

Vienen en varias formas

Genbank / EMBL

GFF

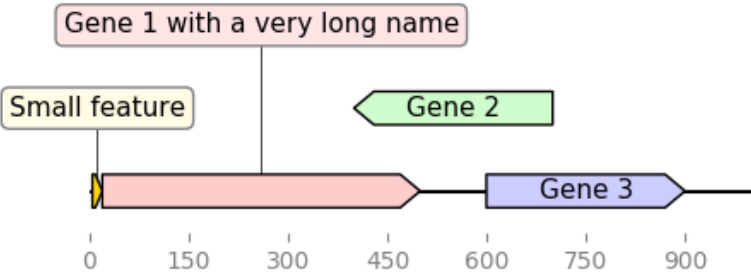
BED

No es importante saberlas de memoria

Es importante el concepto. Capturan información en forma estructurada.

Table 1.1 Indicating locations within the feature table.

345	Single position within the sequence
345..500	A continuous range of positions bounded by and including the indicated positions
<345..500	A continuous range of positions, where the exact lower boundary is not known; the feature begins somewhere prior to position 345 but ends at position 500
345..>500	A continuous range of positions, where the exact upper boundary is not known; the feature begins at position 345 but ends somewhere after position 500
<1..888	The feature starts before the first sequenced base and continues to position 888
(102..110)	Indicates that the exact location is unknown, but that it is one of the positions between 102 and 110, inclusive
123 ^ 124	Points to a site <i>between</i> positions 123 and 124
123 ^ 177	Points to a site <i>between</i> two adjacent nucleotides or amino acids anywhere between positions 123 and 177
join(12..78,134..202)	Regions 12-78 and 134-202 are joined to form one contiguous sequence
complement(4918..5126)	The sequence complementary to that found from 4918 to 5126 in the sequence record
J00194:100..202	Positions 100-202, inclusive, in the entry in this database having accession number J00194

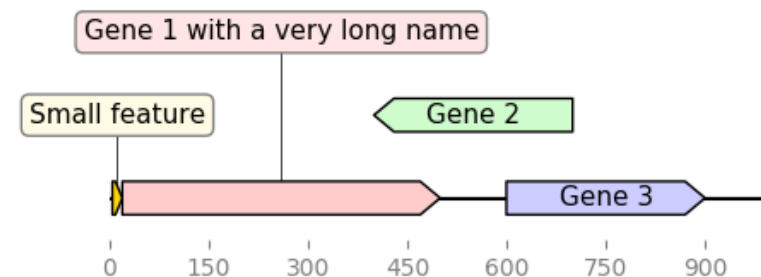


Ejemplo visual de una tabla de features con 4 features

Más sobre tablas de features

EMBL

```
FT      gene           100..500
FT                      /gene="GeneA"
FT                      /locus_tag="GENE01"
FT      mRNA          join(100..200,400..500)
FT                      /gene="GeneA"
FT                      /product="GeneA protein product"
```



Ejemplo visual de una tabla de features con 4 features

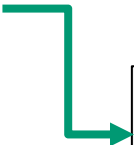
GFF (General Feature Format)

```
##gff-version 3
chr1    RefSeq  gene     100 500 .   +   .   ID=gene01;Name=GeneA
chr1    RefSeq  mRNA     100 500 .   +   .   ID=mrna01;Parent=gene01;Name=GeneA-mRNA
chr1    RefSeq  exon     100 200 .   +   .   ID=exon01;Parent=mrna01
chr1    RefSeq  exon     400 500 .   +   .   ID=exon02;Parent=mrna01
```

BED (Browser Extensible Data)

```
Chr1    99      500    gene01
Chr1    99      500    mrna01
Chr1    99      200    exon01
Chr1    399     500    exon02
```

1. Header
2. Feature Table
3. Sequence



```
XX
SQ   Sequence 2881 BP; 849 A; 699 C; 585 G; 748 T; 0 other;
      cggttgcttg ggttttataa catcagtcag tgacaggcat ttccagagtt gccctgttca      60
      acaatcgata gctgcctttg gccaccaaaa tcccaaactt aattaaagaa ttaaataatt      120
      cgaataataa ttaagcccag taacctacgc agcttgagtg cgtaaccgat atctagtata      180
      catttcgata catcgaaatc atggtagtgt tggagacgga gaaggtaaga cgatgataga      240
      cggcgagccg catgggttcg atttgcgctg agccgtggca gggaacaaca aaaacagggt      300
      tgttgcacaa gaggggaggc gatagtcgag cggaaaagag tgcagttggc gtggctacat      360
      [...]
      gtcactgtca ataatggatt agacggaaaa gtatattaat aaaaacctac attaaaaccg      2880
      g                                                    2881
//
```



Más sobre los formatos: todos son FLAT FILES

FLAT FILES = archivos de texto plano = .txt

Son archivos de tipo TXT (*.txt)

- Solo contienen caracteres
- Nada más

En un programa de computadora? (Python por ejemplo)

secuencia.py

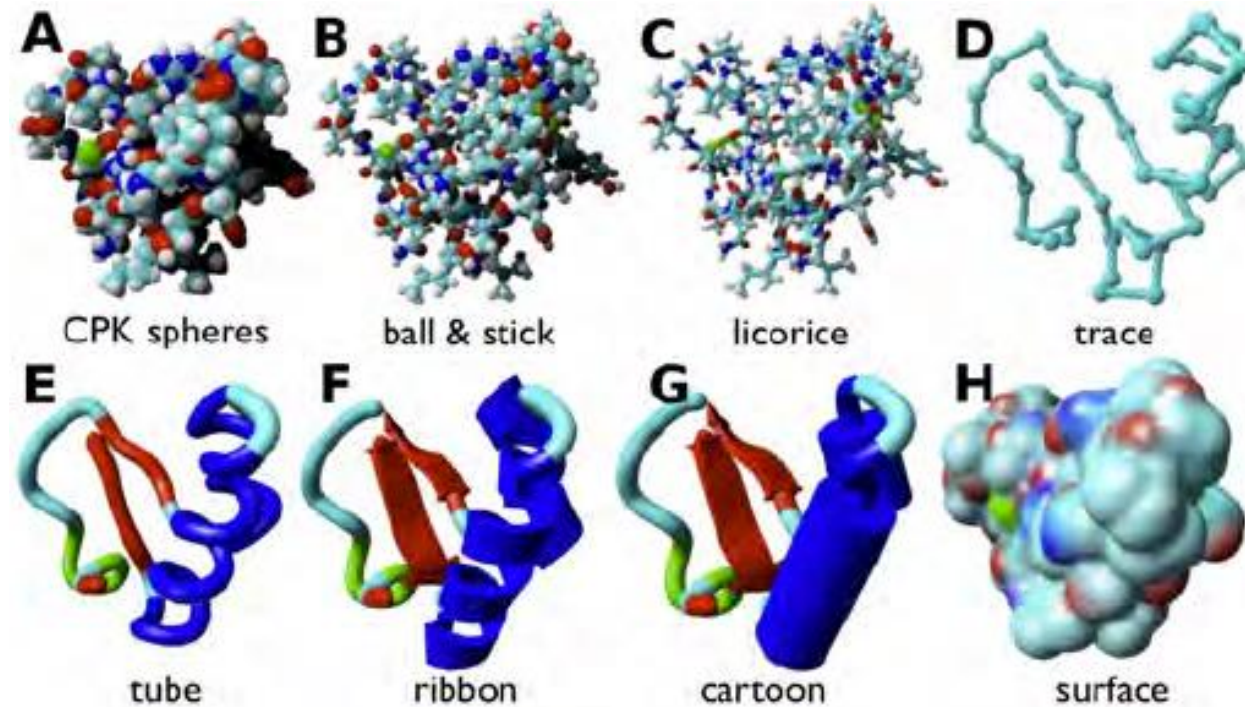
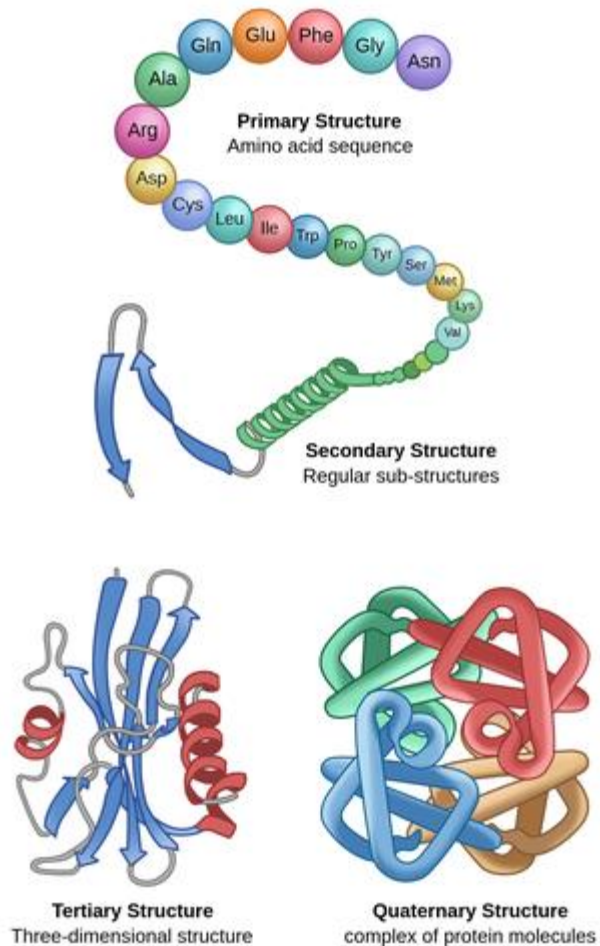
```
1  # defino una variable y guardo la secuencia
2  sequence = "CGGTTGCTTGGGTTTATAACATCAGTCAGTGACAGGCATTTCAGAGTTGCCCTGTTCAACAATCGATA"
3
```

Qué tipos de datos manejamos en biotecnología?

- Secuencias (de ADN, de ARN, de proteínas, polímeros)
- **Estructuras (de proteínas, de complejos macromoleculares)**
- Taxonomía (organismos, especies, relaciones filogenéticas)
- Vías metabólicas celulares (reacciones químicas)
- Compuestos químicos (metabolitos, fármacos)
- Bibliografía (papers, artículos científicos)

Qué tipos de datos manejamos y como los representamos?

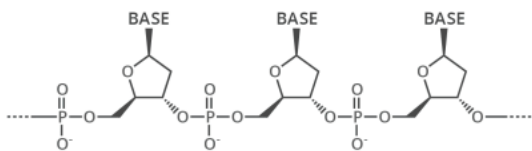
Estructuras



What makes up the chemical structure of DNA?

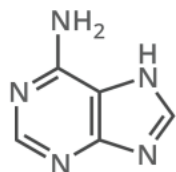


The sugar-phosphate backbone

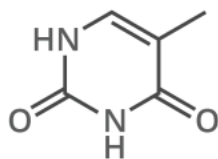


DNA is a polymer made up of units called nucleotides. The nucleotides are made of three different components: a sugar group, a phosphate group, and a base. There are four different bases: adenine, thymine, guanine and cytosine.

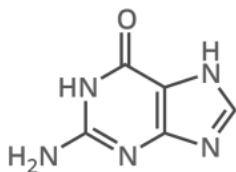
A Adenine



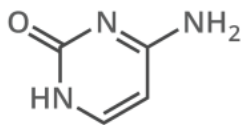
T Thymine



G Guanine

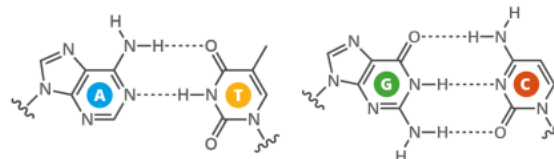


C Cytosine



What holds DNA strands together?

DNA strands are held together by hydrogen bonds between bases on adjacent strands. Adenine (A) always pairs with thymine (T), while guanine (G) always pairs with cytosine (C). Adenine pairs with uracil (U) in RNA.



From DNA to proteins

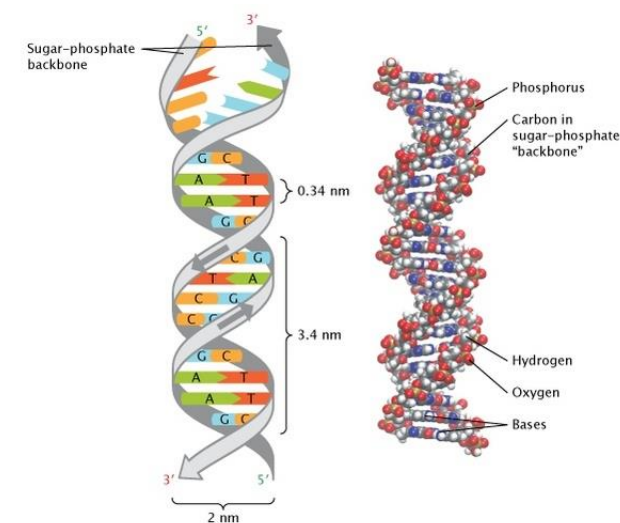
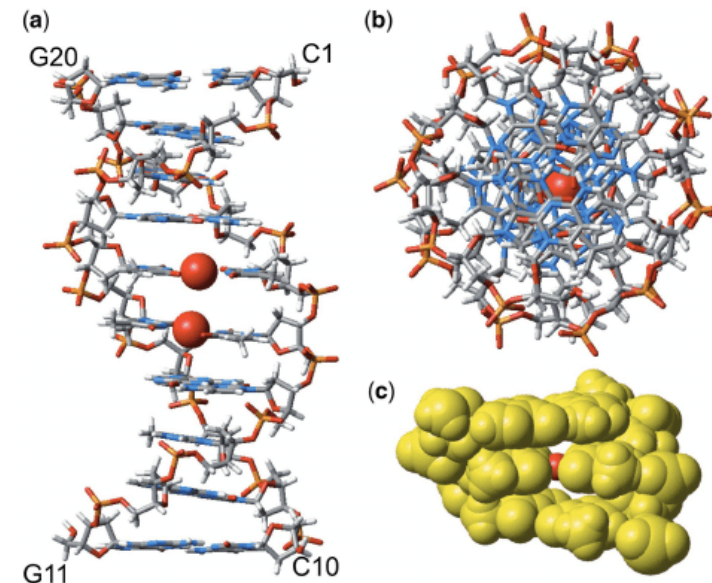
The bases on a single strand of DNA act as a code. The letters form three letter codons, which code for amino acids - the building blocks of proteins.



An enzyme, RNA polymerase, transcribes DNA into mRNA (messenger ribonucleic acid). It splits apart the two strands that form the double helix, then reads a strand and copies the sequence of nucleotides. The only difference between the RNA and the original DNA is that in the place of thymine (T), another base with a similar structure is used: uracil (U).

DNA sequence	T	T	C	C	T	G	A	A	C	C	C	G	T	T	A
mRNA sequence	U	U	C	C	U	G	A	A	C	C	C	G	U	U	A
Amino acid	Phenylalanine	Leucine	Leucine	Leucine	Valine	Asparagine	Asparagine	Asparagine	Proline	Proline	Proline	Glycine	Valine	Valine	Phenylalanine

In multicellular organisms, the mRNA carries genetic code out of the cell nucleus, to the cytoplasm. Here, protein synthesis takes place. 'Translation' is the process of turning the mRNA's 'code' into proteins. Molecules called ribosomes carry out this process, building up proteins from the amino acids coded for.



Formato PDB para estructuras

1. Header
2. Remarks / Comments
3. Primary Structure (sequence)
4. Compuestos químicos
5. Secondary Structure
6. Atomic coordinates

```
HEADER      ELECTRON TRANSPORT      19-MAR-90   2TRX      2TRXA   1
COMPND      THIOREDOXIN                                     2TRXA   2
SOURCE      (ESCHERICHIA $COLI)                               2TRX   4
AUTHOR      S.K.KATTI,D.M.LE*MASTER,H.EKLUND                 2TRX   5
JRNL        AUTH  S.K.KATTI,D.M.LE*MASTER,H.EKLUND           2TRX   7
JRNL        TITL  CRYSTAL STRUCTURE OF THIOREDOXIN FROM ESCHERICHIA 2TRX   8
JRNL        TITL 2 $COLI AT 1.68 ANGSTROMS RESOLUTION          2TRX   9
JRNL        REF   J.MOL.BIOL.                                V. 212   167 1990 2TRX  10
JRNL        REPN  ASTM JMOBAK  UK ISSN 0022-2836              070 2TRX  11
REMARK      2                                                 2TRX  31
REMARK      2 RESOLUTION. 1.68 ANGSTROMS.                     2TRX  32
REMARK      3                                                 2TRX  33
REMARK      3 REFINEMENT. BY THE RESTRAINED LEAST-SQUARES PROCEDURE OF J. 2TRX  34
REMARK      3 KONNERT AND W. HENDRICKSON AS MODIFIED BY B. FINZEL 2TRX  35
REMARK      3 (PROGRAM *PROFFT*). THE R VALUE IS 0.165 FOR 25969 2TRX  36
REMARK      3 REFLECTIONS IN THE RESOLUTION RANGE 8.0 TO 1.68 ANGSTROMS 2TRX  37
REMARK      3 WITH FOBS .GT. 3.0*SIGMA(FOBS)                  2TRX  38
REMARK      3                                                 2TRX  39
SEQRES      1 A 108 SER ASP LYS ILE ILE HIS LEU THR ASP ASP SER PHE ASP 2TRX  74
SEQRES      2 A 108 THR ASP VAL LEU LYS ALA ASP GLY ALA ILE LEU VAL ASP 2TRX  75
SEQRES      3 A 108 PHE TRP ALA GLU TRP CYS GLY PRO CYS LYS MET ILE ALA 2TRX  76
SEQRES      4 A 108 PRO ILE LEU ASP GLU ILE ALA ASP GLU TYR GLN GLY LYS 2TRX  77
SEQRES      5 A 108 LEU THR VAL ALA LYS LEU ASN ILE ASP GLN ASN PRO GLY 2TRX  78
SEQRES      6 A 108 THR ALA PRO LYS TYR GLY ILE ARG GLY ILE PRO THR LEU 2TRX  79
SEQRES      7 A 108 LEU LEU PHE LYS ASN GLY GLU VAL ALA ALA THR LYS VAL 2TRX  80
SEQRES      8 A 108 GLY ALA LEU SER LYS GLY GLN LEU LYS GLU PHE LEU ASP 2TRX  81
SEQRES      9 A 108 ALA ASN LEU ALA                           2TRX  82
HET         MPD 606      8 2-METHYL-2,4-PENTANEDIOL          2TRX 107
HET         MPD 607      8 2-METHYL-2,4-PENTANEDIOL          2TRX 108
HET         MPD 608      8 2-METHYL-2,4-PENTANEDIOL          2TRX 109
FORMUL      3  CU      2(CU1 ++))                             2TRX 110
FORMUL      4  MPD      8(C6 H14 O2)                          2TRX 111
FORMUL      5  HOH     *140(H2 O1)                           2TRX 112
HELIX       1 A1A SER A 11 LEU A 17 1 DISORDERED IN MOLECULE B 2TRX 113
HELIX       2 A2A CYS A 32 TYR A 49 1 BENT BY 30 DEGREES AT RES 39 2TRX 114
HELIX       3 A3A ASN A 59 ASN A 63 1                         2TRX 115
HELIX       4 31A THR A 66 TYR A 70 5 DISTORTED H-BONDING C-TERMS 2TRX 116
HELIX       5 A4A SER A 95 LEU A 107 1                        2TRX 117
HELIX       6 A1B SER B 11 LEU B 17 1 DISORDERED IN MOLECULE B 2TRX 118
SSBOND      1 CYS A 32 CYS A 35                               2TRX 143
ATOM        1 N SER A 1 21.389 25.406 -4.628 1.00 23.22       2TRX 152
ATOM        2 CA SER A 1 21.628 26.691 -3.983 1.00 24.42       2TRX 153
ATOM        3 C SER A 1 20.937 26.944 -2.679 1.00 24.21       2TRX 154
ATOM        4 O SER A 1 21.072 28.079 -2.093 1.00 24.97       2TRX 155
ATOM        5 CB SER A 1 21.117 27.770 -5.002 1.00 28.27       2TRX 156
ATOM        6 OG SER A 1 22.276 27.925 -5.861 1.00 32.61       2TRX 157
ATOM        7 N ASP A 2 20.173 26.028 -2.163 1.00 21.39       2TRX 158
ATOM        8 CA ASP A 2 19.395 26.125 -0.949 1.00 21.57       2TRX 159
ATOM        9 C ASP A 2 20.264 26.214 0.297 1.00 20.89       2TRX 160
ATOM       10 O ASP A 2 19.760 26.575 1.371 1.00 21.49       2TRX 161
ATOM       11 CB ASP A 2 18.439 24.914 -0.856 1.00 22.14       2TRX 162
ATOM       22 CE LYS A 3 21.620 21.104 2.844 1.00 25.84       2TRX 173
ATOM       23 NZ LYS A 3 20.830 20.757 1.615 1.00 25.55       2TRX 174
```

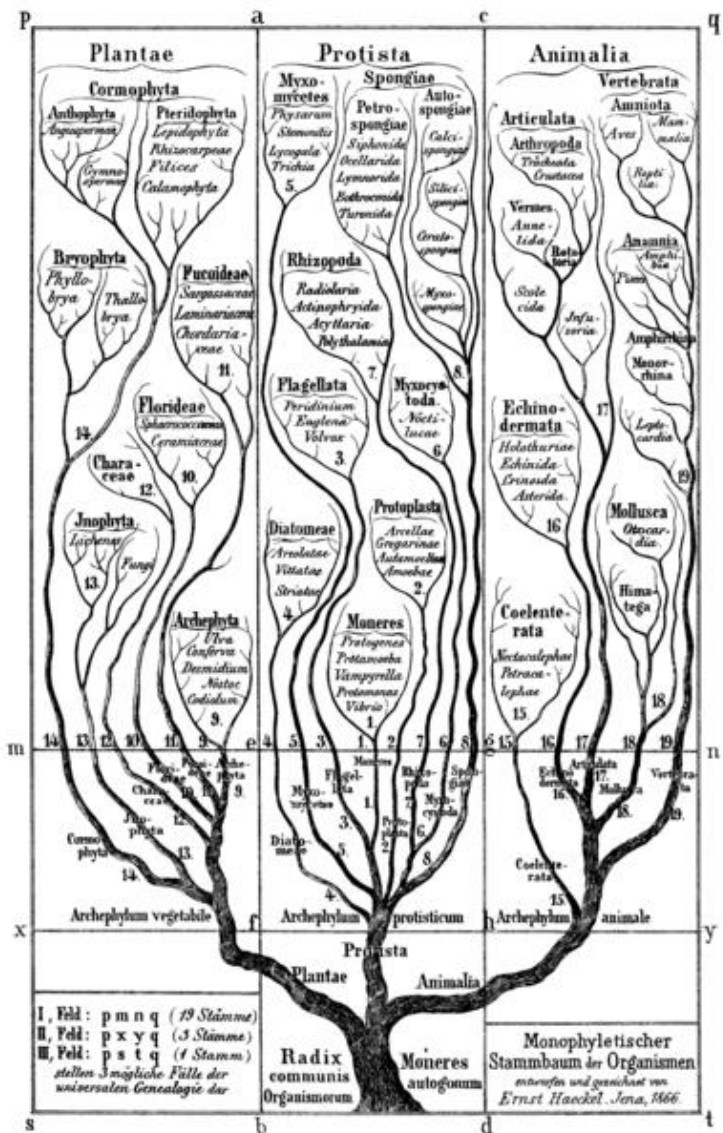
Source: Bioinformatics (4th Ed), Baxevanis, Bader, Wishart.

Figure 12.6 An example of a Protein Data Bank formatted file showing the first ~50 lines of the *Escherichia coli* thioredoxin entry (Protein Data Bank database identifier: 2TRX).

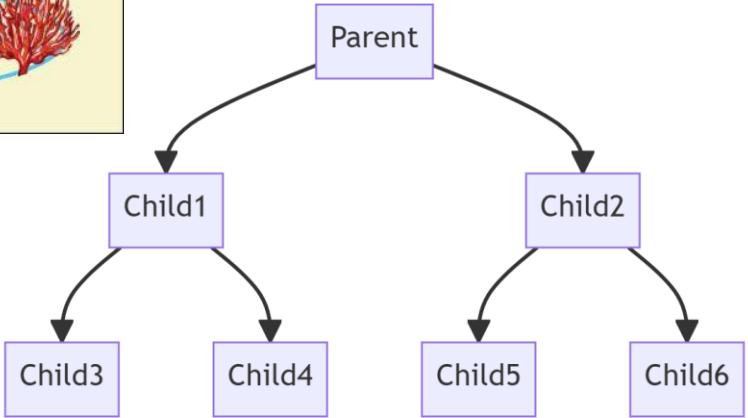
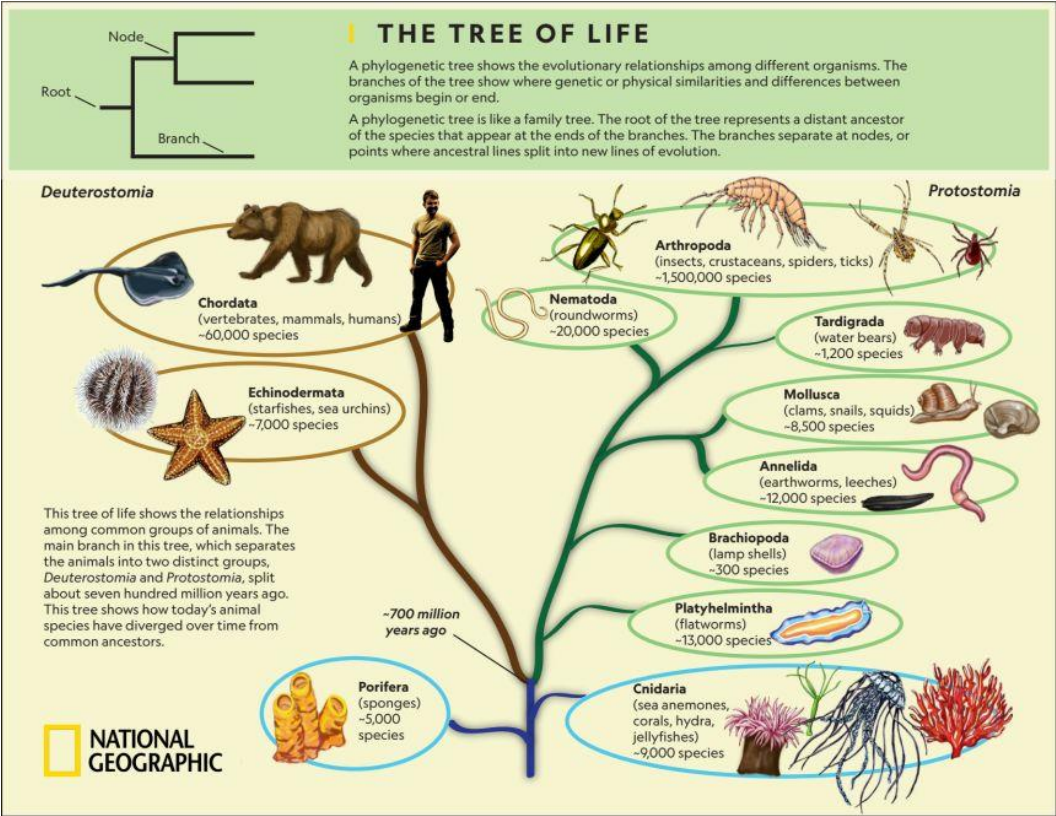
Qué tipos de datos manejamos en biotecnología?

- Secuencias (de ADN, de ARN, de proteínas, polímeros)
- Estructuras (de proteínas, de complejos macromoleculares)
- **Taxonomía (organismos, especies, relaciones filogenéticas)**
- Vías metabólicas celulares (reacciones químicas)
- Compuestos químicos (metabolitos, fármacos)
- Bibliografía (papers, artículos científicos)

Taxonomía, árboles filogenéticos, datos jerárquicos



Ernst Haeckel. General Morphology of Organisms (1866)



Esquema conceptual de datos jerárquicos

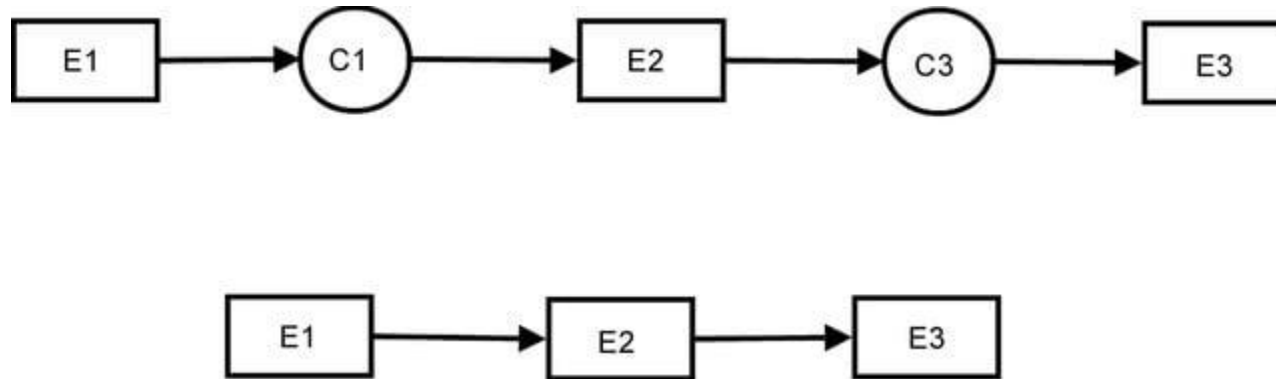
Qué tipos de datos manejamos en biotecnología?

- Secuencias (de ADN, de ARN, de proteínas, polímeros)
- Estructuras (de proteínas, de complejos macromoleculares)
- Taxonomía (organismos, especies, relaciones filogenéticas)
- **Vías metabólicas celulares (reacciones químicas)**
- Compuestos químicos (metabolitos, fármacos)
- Bibliografía (papers, artículos científicos)

Viás metabólicas celulares

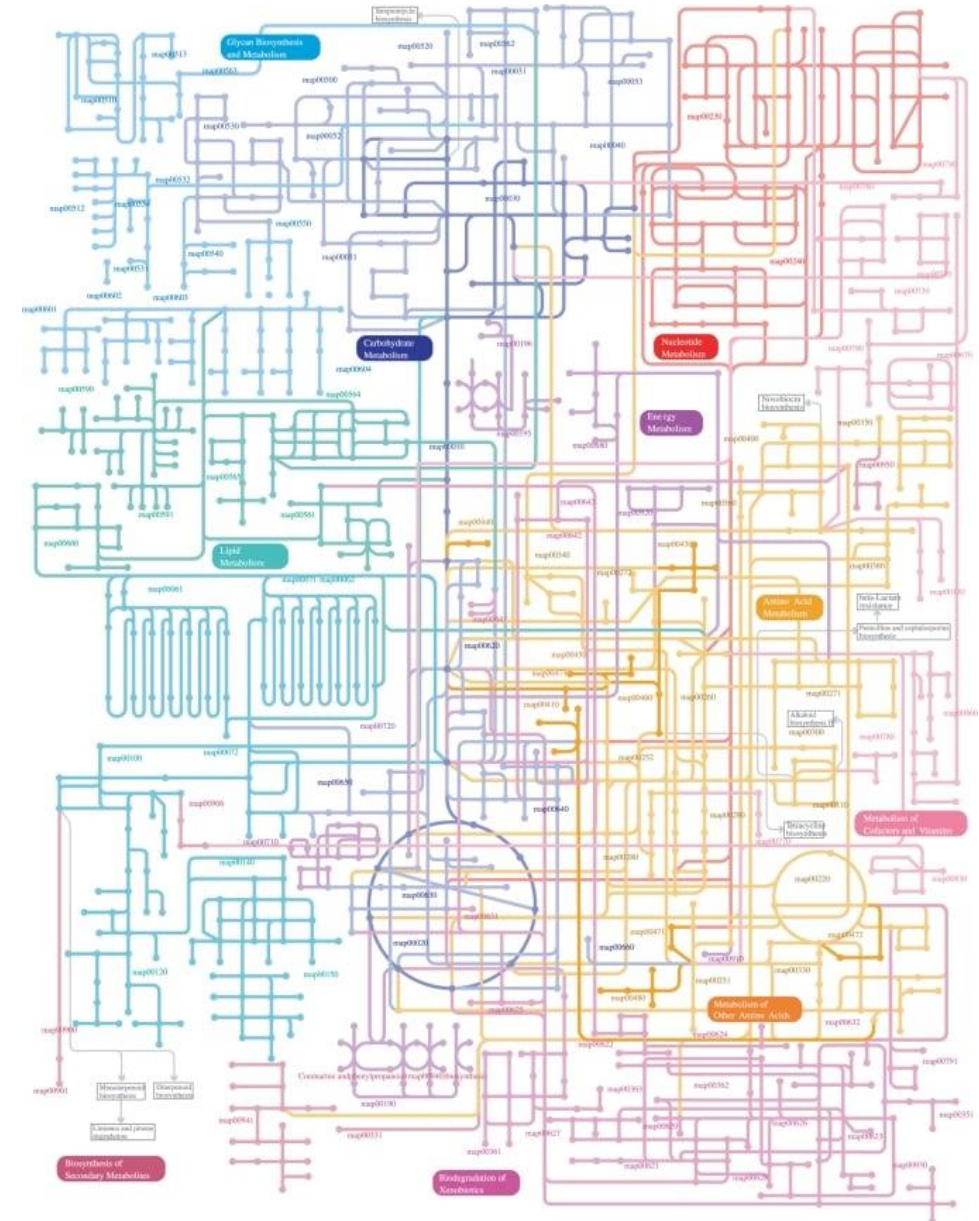
Qué información capturan?

Reacciones enzimáticas que convierten unos metabolitos en otros (ej glucosa en etanol)

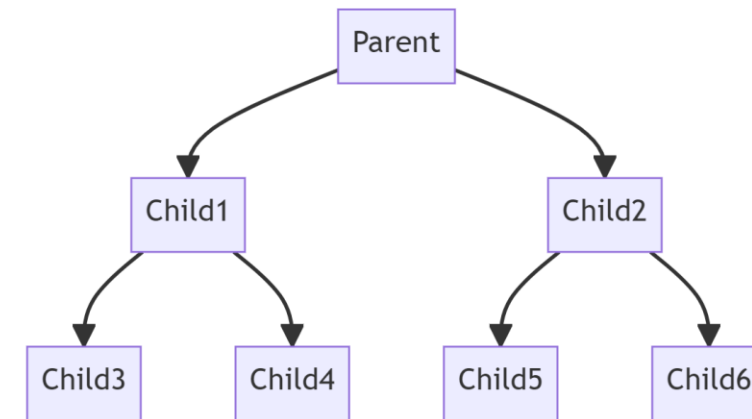
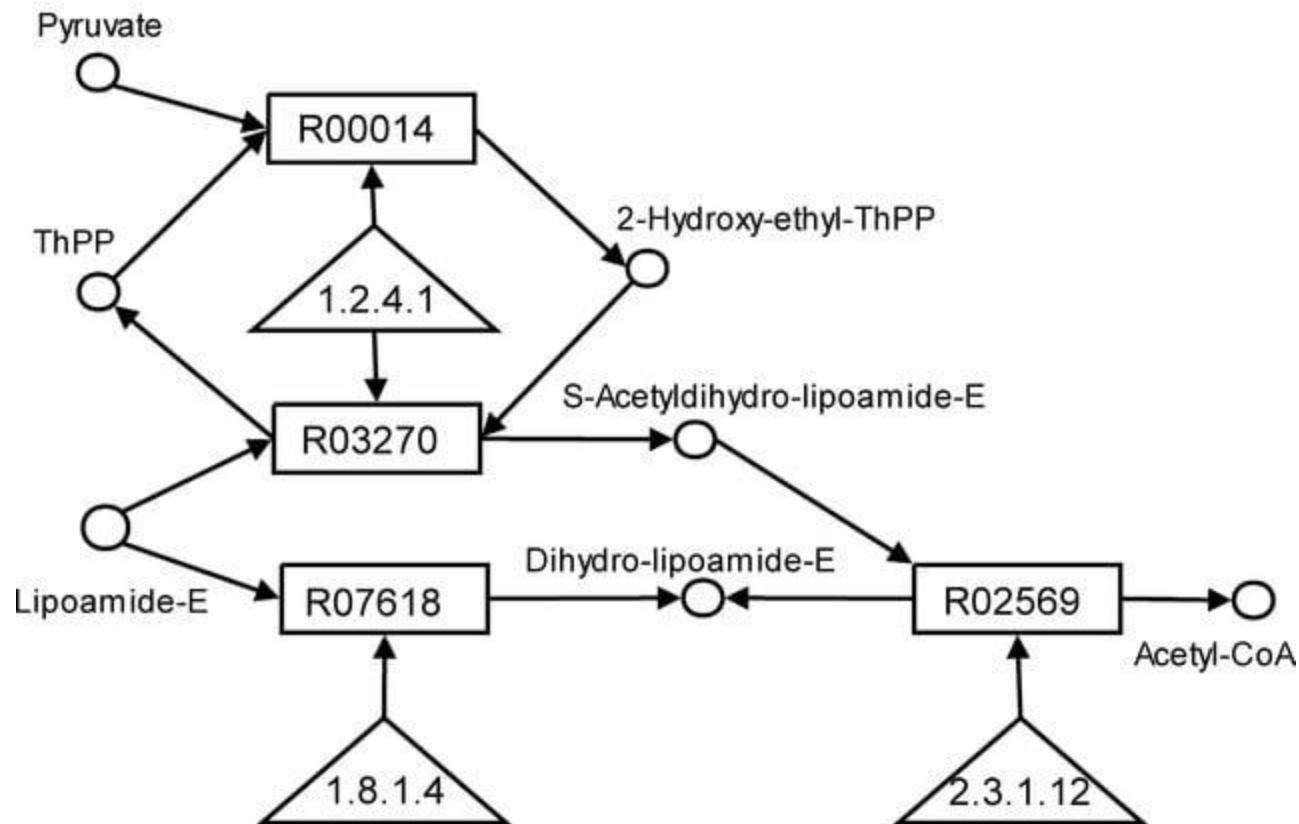


Esquema de una vía metabólica genérica.

E = enzima; C = compuesto químico (metabolito)



Más sobre vías metabólicas: reacciones



Esquema conceptual de datos jerárquicos

Qué tipo de datos son estos? Y cómo los representamos electrónicamente en la computadora?

Nuevo tipo de datos: Grafos

Los **grafos** son objetos de la **matemática discreta**.

Un grafo es una estructura compuesta por conjuntos de **nodos** (o puntos) y **aristas** (o vértices o links).

Son parte de la Teoría de Grafos:

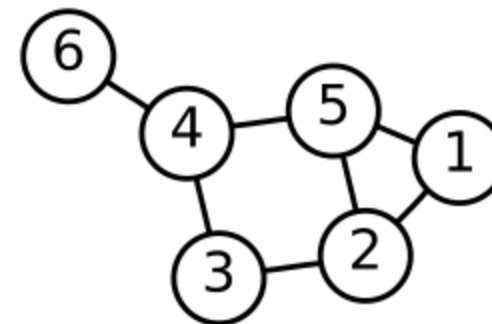
- permiten abstraerse de la **naturaleza** de los objetos
- estudian y resuelven problemas de **conectividad**

Los grafos se pueden describir en FLAT FILES (.txt)

Grafo:

nodos = {1, 2, 3, 4, 5, 6}

aristas = { {1,2}, {2,3}, {2,5}, {3,4}, {4,5}, {1,5}, {4,6} }



A graph with six vertices and seven edges

[https://en.wikipedia.org/wiki/Graph_\(discrete_mathematics\)](https://en.wikipedia.org/wiki/Graph_(discrete_mathematics))

https://en.wikipedia.org/wiki/Graph_theory

Nuevo tipo de datos: Grafos

Los **grafos** son objetos de la **matemática discreta**.

Un grafo es una estructura compuesta por conjuntos de **nod**os (o puntos) y **aristas** (o vértices o links).

Son parte de la Teoría de Grafos:

- permiten abstraerse de la **naturaleza** de los objetos
- estudian y resuelven problemas de **conectividad**

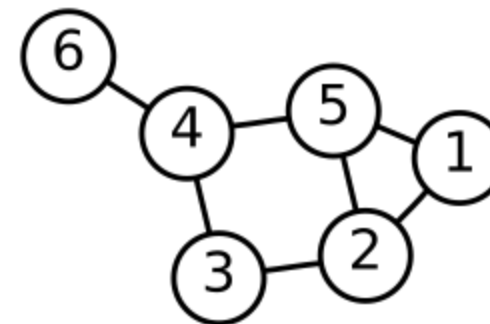
Los grafos se pueden describir en FLAT FILES (.txt)

Grafo:

nod

os = {a, b, c, d, e, f}

aristas = { {a,b}, {b,c}, {b,e}, {c,d}, {d,e}, {a,e}, {d,f} }



A graph with six vertices and seven edges

[https://en.wikipedia.org/wiki/Graph_\(discrete_mathematics\)](https://en.wikipedia.org/wiki/Graph_(discrete_mathematics))

https://en.wikipedia.org/wiki/Graph_theory

Grafos: qué son? Propiedades y operaciones

Propiedades de los grafos:

Grado de conectividad de los nodos (degree)

Direccionalidad de las aristas

Intensidad (sentido vectorial) de cada arista

Las aristas pueden tener asociado un valor numérico (peso, largo, costo)

Posibilidad de identificar los nodos

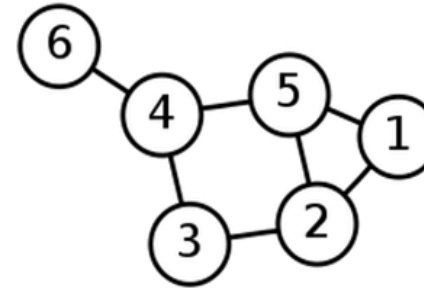
Son elementos de un conjunto

Grafos etiquetados (labeled)

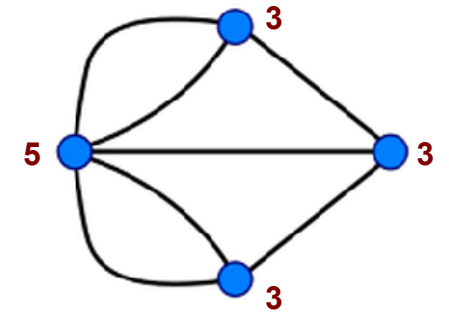
vs no-etiquetados (unlabeled)

Operaciones con grafos (algunos ejemplos):

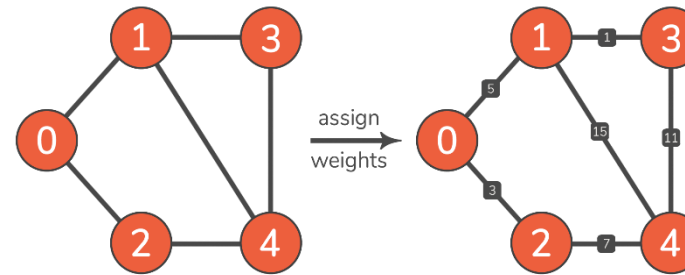
Complementación, Unión, Suma, Intersección, Diferencia, ...



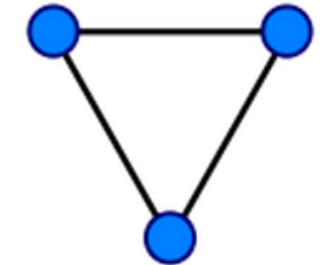
Un grafo etiquetado



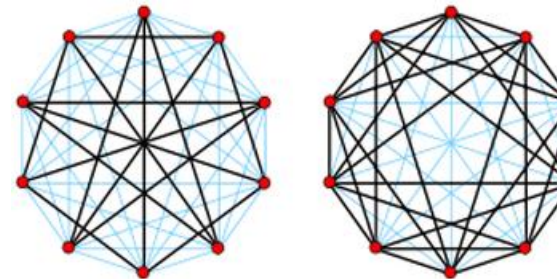
Grados de los nodos



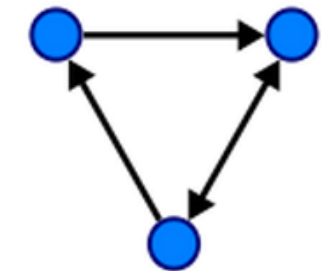
Un grafo pesado



Grafo simple o regular

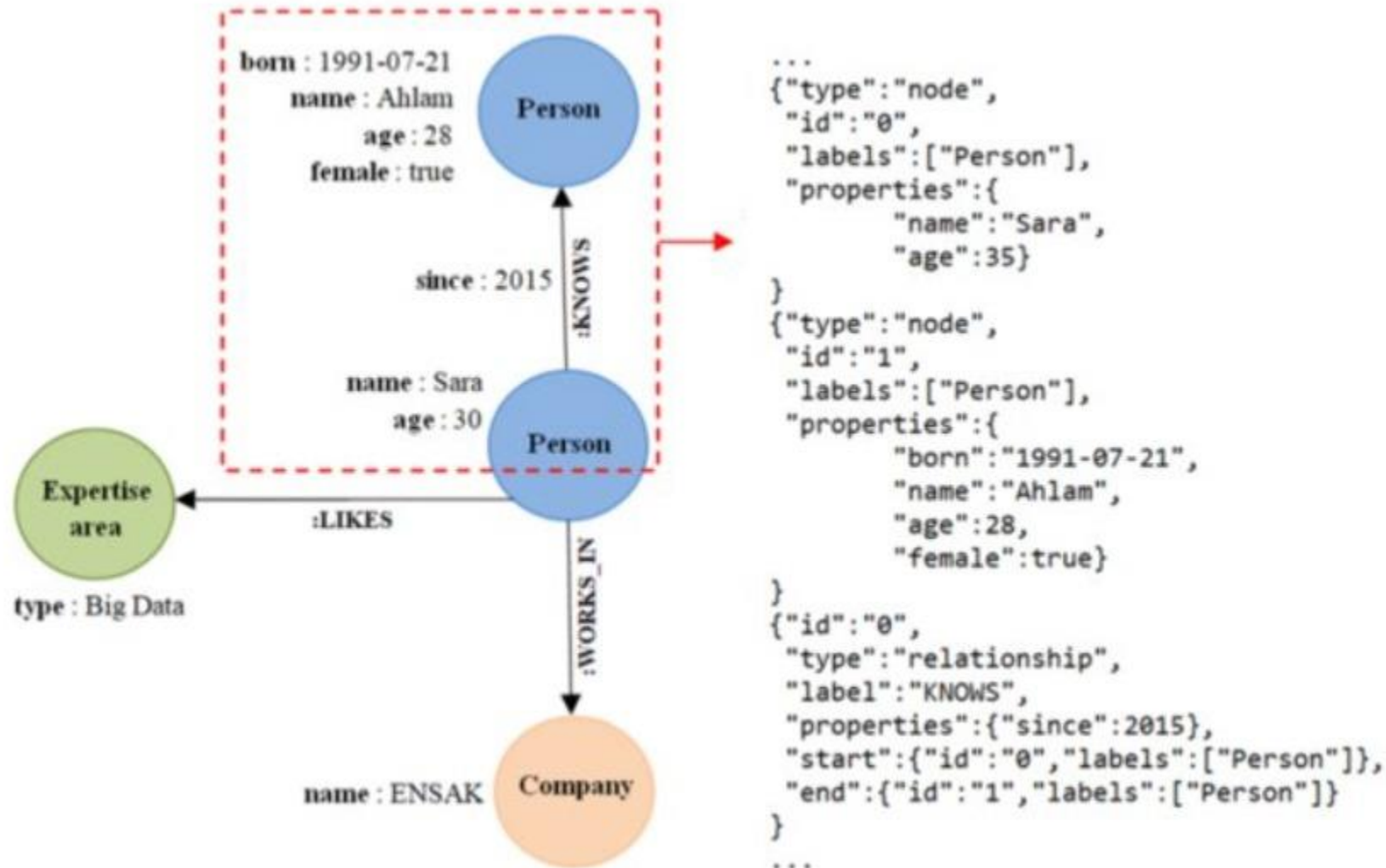


Un grafo y su complemento



Grafo dirigido (red)

Formatos de archivo: grafos



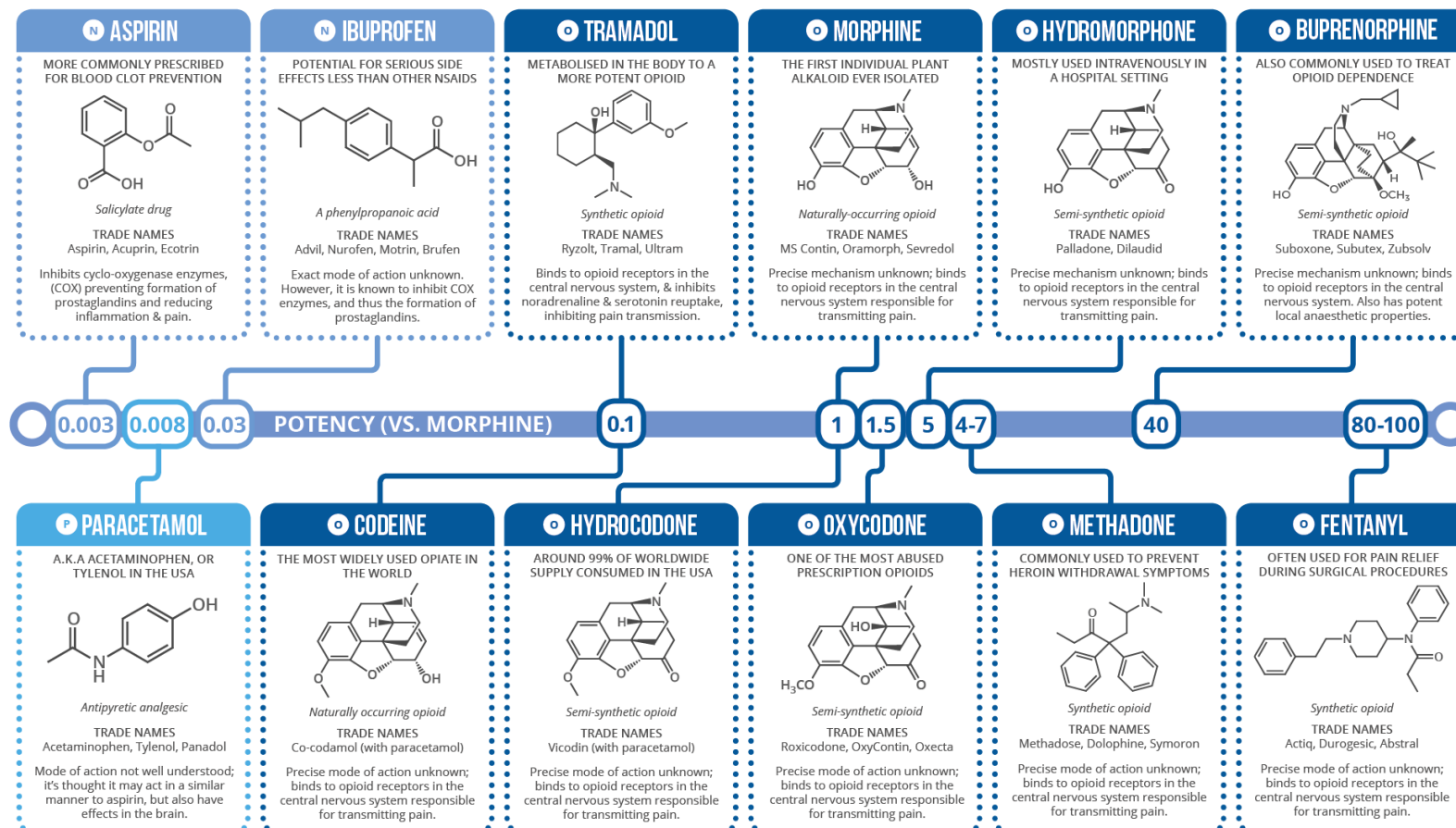
Qué tipos de datos manejamos en biotecnología?

- Secuencias (de ADN, de ARN, de proteínas, polímeros)
- Estructuras (de proteínas, de complejos macromoleculares)
- Taxonomía (organismos, especies, relaciones filogenéticas)
- Vías metabólicas celulares (reacciones químicas)
- **Compuestos químicos (metabolitos, fármacos)**
- Bibliografía (papers, artículos científicos)

Fármacos, drogas, compuestos químicos

A BRIEF GUIDE TO SELECTED COMMON PAINKILLERS

THERE ARE TWO MAIN CLASSES OF PAINKILLERS - PARACETAMOL IS AN EXCEPTION. **Key:** **N** NON-STEROIDAL ANTI-INFLAMMATORY DRUGS **P** PARACETAMOL **O** OPIOID ANALGESICS



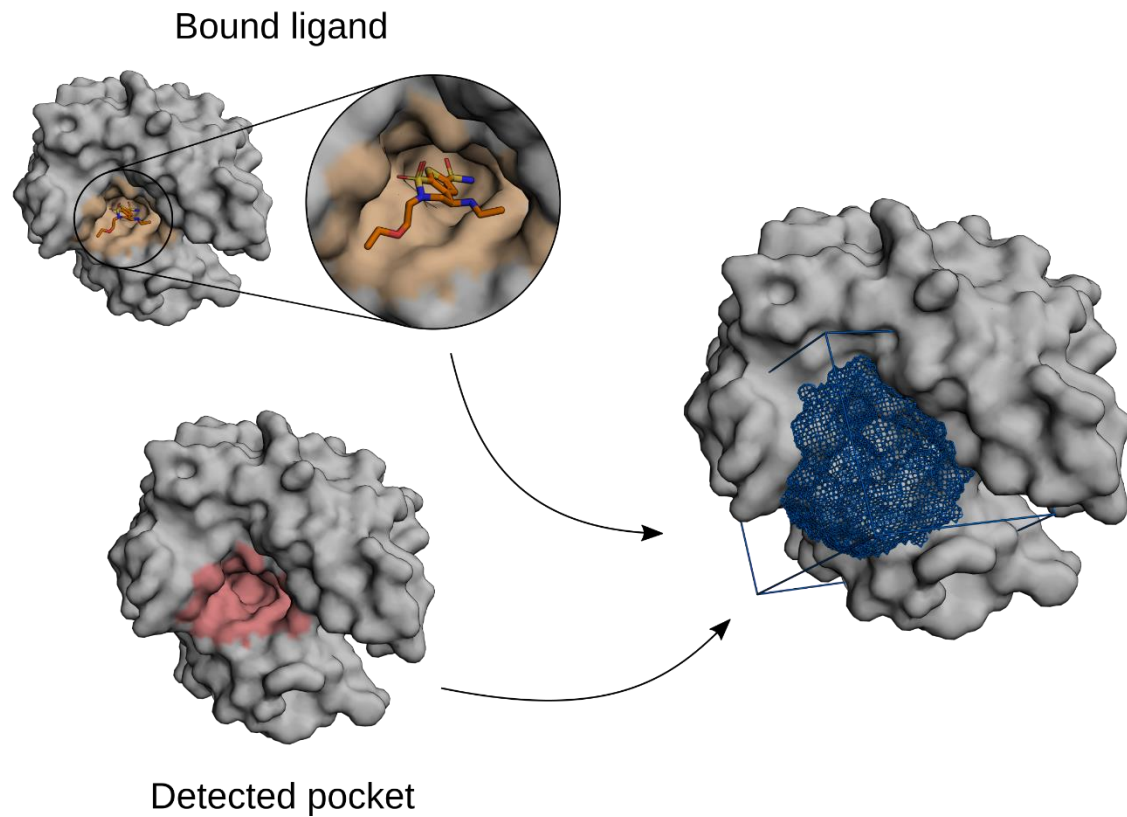
Note: Potency values are for oral administration. Numeric measures of potency are variable; the figures given are merely general approximations, and can be affected by a number of factors.

Fármacos contra el dolor



Interacciones entre fármacos y macromoléculas celulares

Ejemplo: fármaco (droga, ligando)
interaccionando con un sitio funcional
(bolsillo) de una proteína.



Qué tipos de datos manejamos en biotecnología?

- Secuencias (de ADN, de ARN, de proteínas, polímeros)
- Estructuras (de proteínas, de complejos macromoleculares)
- Taxonomía (organismos, especies, relaciones filogenéticas)
- Vías metabólicas celulares (reacciones químicas)
- Compuestos químicos (metabolitos, fármacos)
- **Bibliografía (papers, artículos científicos)**

PubMed®

Advanced

Search

User Guide

Search results

Save

Email

Send to

Display options ⚙

Structure. 2026 Mar 5;34(3):471-486.e7. doi: 10.1016/j.str.2025.12.015. Epub 2026 Jan 30.

Discovery of tankyrase scaffolding inhibitor specifically targeting the ARC4 peptide binding domain

Chiara Bosetti¹, Albert Galera-Prat¹, Sven T Sowa¹, Alexandra Gade², Cláudia Braga³, Shoshy A Brinch⁴, Faranak Nami³, Johan Pääkkönen¹, Veeti Pulju¹, Maureen T Meling⁴, Maria Candamo-Lourido⁴, Jo Waaler⁴, Mads H Clausen³, Lari Lehtiö⁵

Affiliations + expand

PMID: 41619733 DOI: 10.1016/j.str.2025.12.015

Free article

Abstract

Tankyrases are poly-ADP-ribosyltransferases that orchestrate numerous biological processes involved in disease. Their established regulatory roles, particularly within the WNT/ β -catenin pathway, have driven notable drug discovery efforts aimed at inhibiting their catalytic activity. Targeting tankyrases' interaction with proteins through their ARC domains represents an alternative strategy to be explored as a therapeutic approach against specific protein-protein interactions. In this article, we employed a pre-established FRET-based assay to screen the EU-OPENSOURCE libraries for identification of ARC4 inhibitors. We discovered a series of pyrrolone-based compounds, and we synthesized compound S8 (ARCher-142), which binds selectively to ARC4 with a potency of 8 μ M. NMR analysis and X-ray crystallography allowed us to identify the binding site and to rationalize the observed selectivity. Despite binding exclusively to ARC4, the inhibitor can attenuate the WNT/ β -catenin signaling pathway in cells. Our work demonstrates that targeting single ARC domains is possible, offering an inhibition approach tailored to tankyrase ARC4.

Keywords: ARC; TNKS; WNT signaling pathway; ankryrin repeat cluster; drug discovery; inhibitor; protein-protein interactions; pyrrolone; tankyrase; β -catenin.

Copyright © 2026 The Authors. Published by Elsevier Inc. All rights reserved.

PubMed Disclaimer

FULL TEXT LINKS



ACTIONS

- Cite
- Collections
- Permalink

PAGE NAVIGATION

Title & authors

- Abstract
- Conflict of interest statement
- Similar articles
- MeSH terms
- Substances
- Related information
- LinkOut - more resources

PMID- 41619733

OWN - NLM

STAT- MEDLINE

DCOM- 20260706

LR - 20260706

IS - 1878-4186 (Electronic)

IS - 0969-2126 (Linking)

VI - 34

IP - 3

DP - 2026 Mar 5

TI - Discovery of tankyrase scaffolding inhibitor specifically targeting the ARC4 peptide binding domain.

PG - 471-486.e7

LID - S0969-2126(25)00496-4 [pii]

LID - 10.1016/j.str.2025.12.015 [doi]

AB - Tankyrases are poly-ADP-ribosyltransferases that orchestrate numerous biological ...

FAU - Bosetti, Chiara

AU - Bosetti C

AD - Faculty of Biochemistry and Molecular Medicine & Biocenter Oulu, University of Oulu, 90220 Oulu, Finland.

AU - Galera-Prat A

LA - eng

PT - Journal Article [...]

https://pubmed.ncbi.nlm.nih.gov/41619733/?format=pubmed

Qué tipos de datos manejamos en biotecnología?

- **Secuencias (de ADN, de ARN, de proteínas, polímeros)**
 - Cadenas de caracteres (strings), 1D
- **Estructuras (de proteínas, de complejos macromoleculares)**
 - Grafos, 3D
- **Taxonomía (organismos, especies, relaciones filogenéticas)**
 - Grafos, datos jerárquicos
- **Vías metabólicas celulares (reacciones químicas)**
 - Grafos,
- **Compuestos químicos (metabolitos, fármacos)**
 - Cadenas de caracteres (representaciones 2D)
 - Grafos, 3D
- **Bibliografía (papers, artículos científicos)**
 - Cadenas de caracteres (texto)

I need coffee...
Lots and lots and lots
of coffee, with a cake
and possibly some
cookies.

someecards
user card



INTERVALO

Fin del Slot 1

Slot 2

BASES DE DATOS

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, Google Docs)

Si. Permite sólo búsqueda y ordenamiento simples.

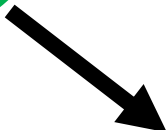
Planilla de Cálculo?
(Excel, Google Calc)

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

- 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**. O sea, un **identificador único** que define al registro sin ambigüedad.

Planilla

Versión simple de
una base de datos



Accession	Version	Date	Division	Tax ID	Organism	No of chromosomes
NM_000014	3	01/06/2000	PRI	9606	Homo sapiens	22 diploid + X + Y
NM_000014	2	12/10/1999	PRI	9606	Homo sapiens	22 diploid + X + Y
NM_000014	1	04/02/1999	PRI	9606	Homo sapiens	22 diploid + X + Y
X63129	1	06/06/1996	MAM	9913	Bos Taurus	29 diploid + 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 busqueda: 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)**
 - 1, 85, 23, 553.45, 1243.32, 3.14159
- **Texto**
 - Hola, “Sos un 10”, GAATTC, “23/Jun/2026”, “3.14159”
- **Tiempo: fechas y timestamps (DD/MM/YYYY, HH:MM:SS)**
 - 23/Jun/2026, 23/06/2026, "2026-08-01 12:34:56"
- **Logicos (boolean)**
 - verdadero / falso
 - 1, 0
- **Geometricos (punto, linea, circulo, poligonos, etc.)**
 - POINT(x,y), POINT(x,y,z), POLYGON(x1, y1, x2, y2)
 - POINT(-34.5669884854995, -58.528324629181434)
- **Secuencias (ADN, Proteinas)**
 - GAATTC, CATATGTTCC, MQSDFHRMKNFANPKSMF

Bases de datos: conceptos básicos: clave primaria

gi	Accession	Version	Date	Division	Tax ID	Organism	No 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 diploid + 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 relacionales

Accession	Version	Date	Division	Tax ID	Organism	No of chromosomes
NM_000014	3	01/06/2000	PRI	9606	Homo sapiens	22 diploid + X + Y
NM_000014	2	12/10/1999	PRI	9606	Homo sapiens	22 diploid + X + Y
NM_000014	1	04/02/1999	PRI	9606	Homo sapiens	22 diploid + X + Y
X63129	1	06/06/1996	MAM	9913	Bos Taurus	29 diploid + X + Y

Accession	Version	Date	Division	Tax ID
NM_000014	3	01/06/2000	PRI	9606
NM_000014	2	12/10/1999	PRI	9606
NM_000014	1	04/02/1999	PRI	9606
X63129	1	06/06/1996	MAM	9913

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).

Tax ID	Organism	No of Chromosomes
9606	Homo sapiens	22 diploid + X + Y
9913	Bos Taurus	29 diploid + 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



- 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

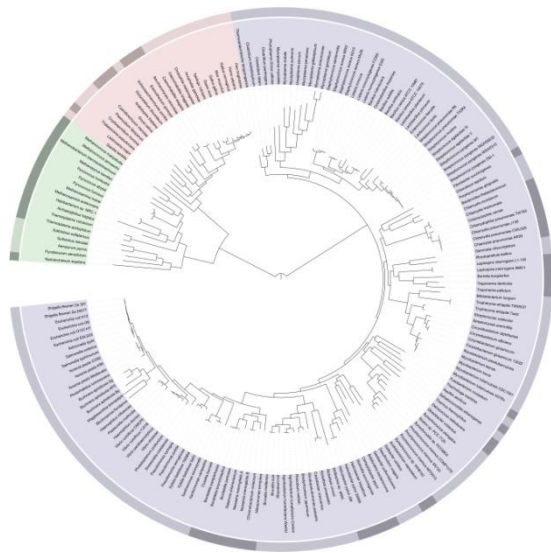
Bibliographic References

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

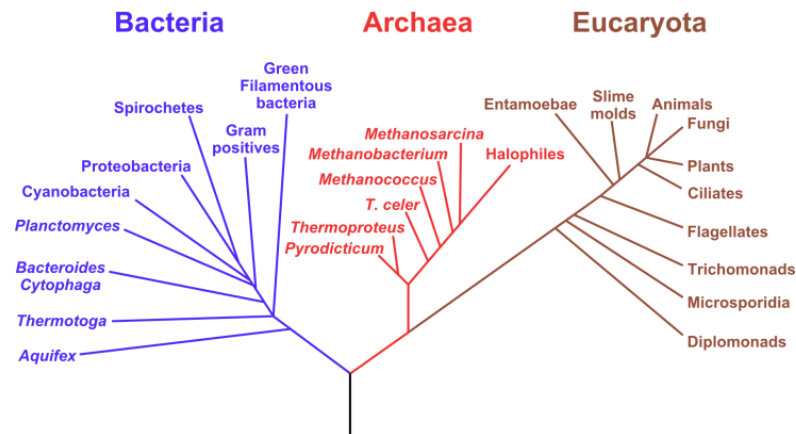
Las claves primarias (o secundarias) **permiten generar asociaciones** (links) entre *datos de distintos tipos*.

Representación de árboles y grafos

- Ejemplo: representación en forma tabular 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

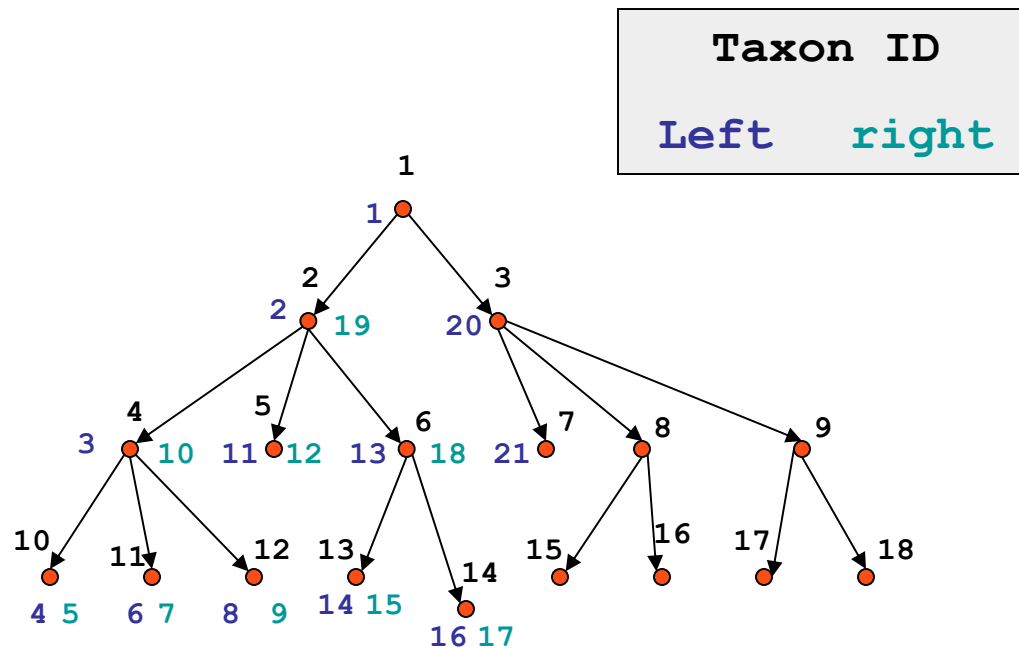
Los valores **left** y **right** son números arbitrarios, pero deben cumplir con la siguiente propiedad:

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?

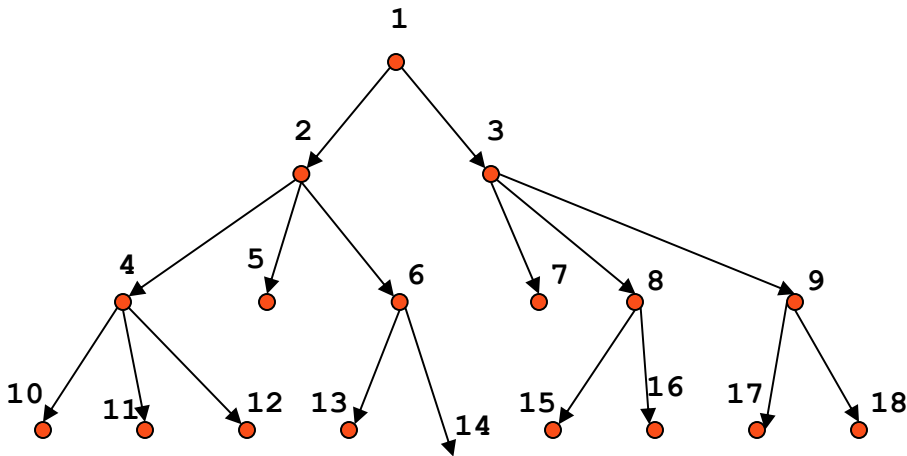
- Cómo se generan los valores para **left** y **right**?
 - Hay que recorrer el árbol asignando estos valores



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

Materialized Paths

	Campo	Tipo
PK	Taxon_id	entero
	Name	Texto
	Path	Texto

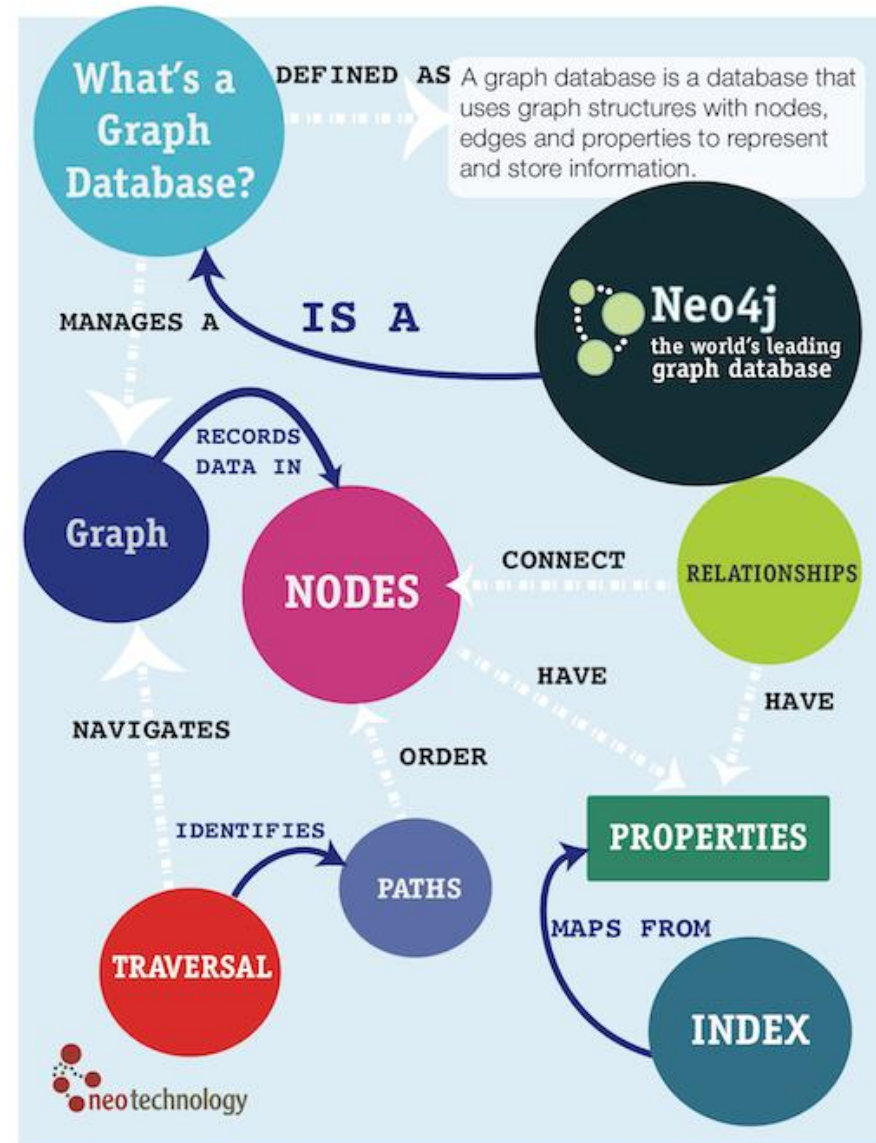
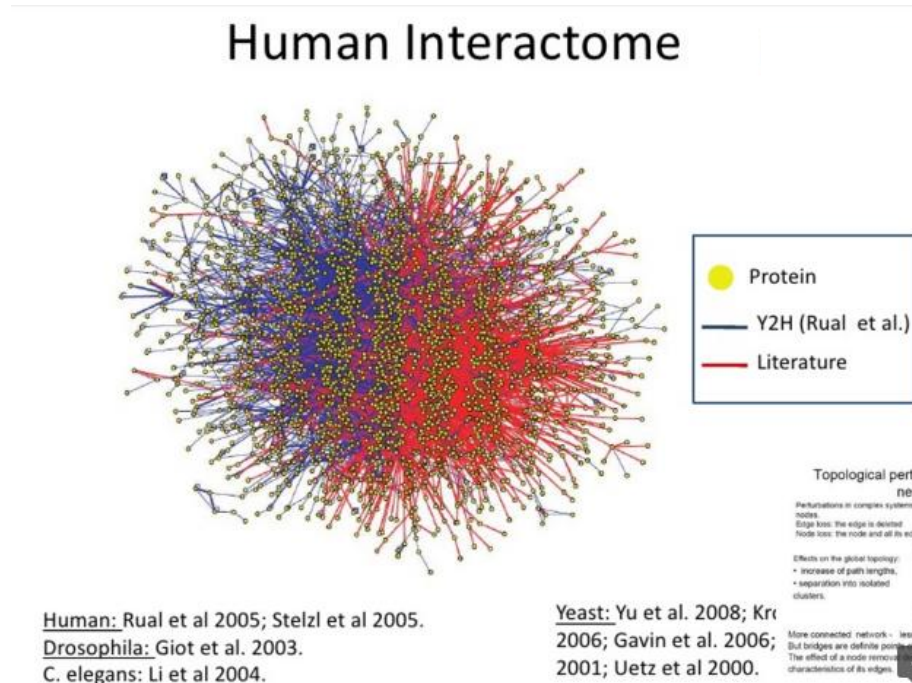


Taxon	Nombre	Path
1	Root	1
2	Bacteria	1.2
3	Archaea	1.3
4	Proteobacteria	1.2.4
5	Cyanobacteria	1.2.5
6	Actinobacteria	1.2.6
7	Crenarchaeota	1.3.7
8	Euryarchaeota	1.3.8
9	Thaumarchaeota	1.3.9
10	Alfa-Proteobacteria	1.2.4.10
11	Gamma-proteobacteria	1.2.4.11
12	Delta-proteobacteria	1.2.4.12
13	Coriobacteridae	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

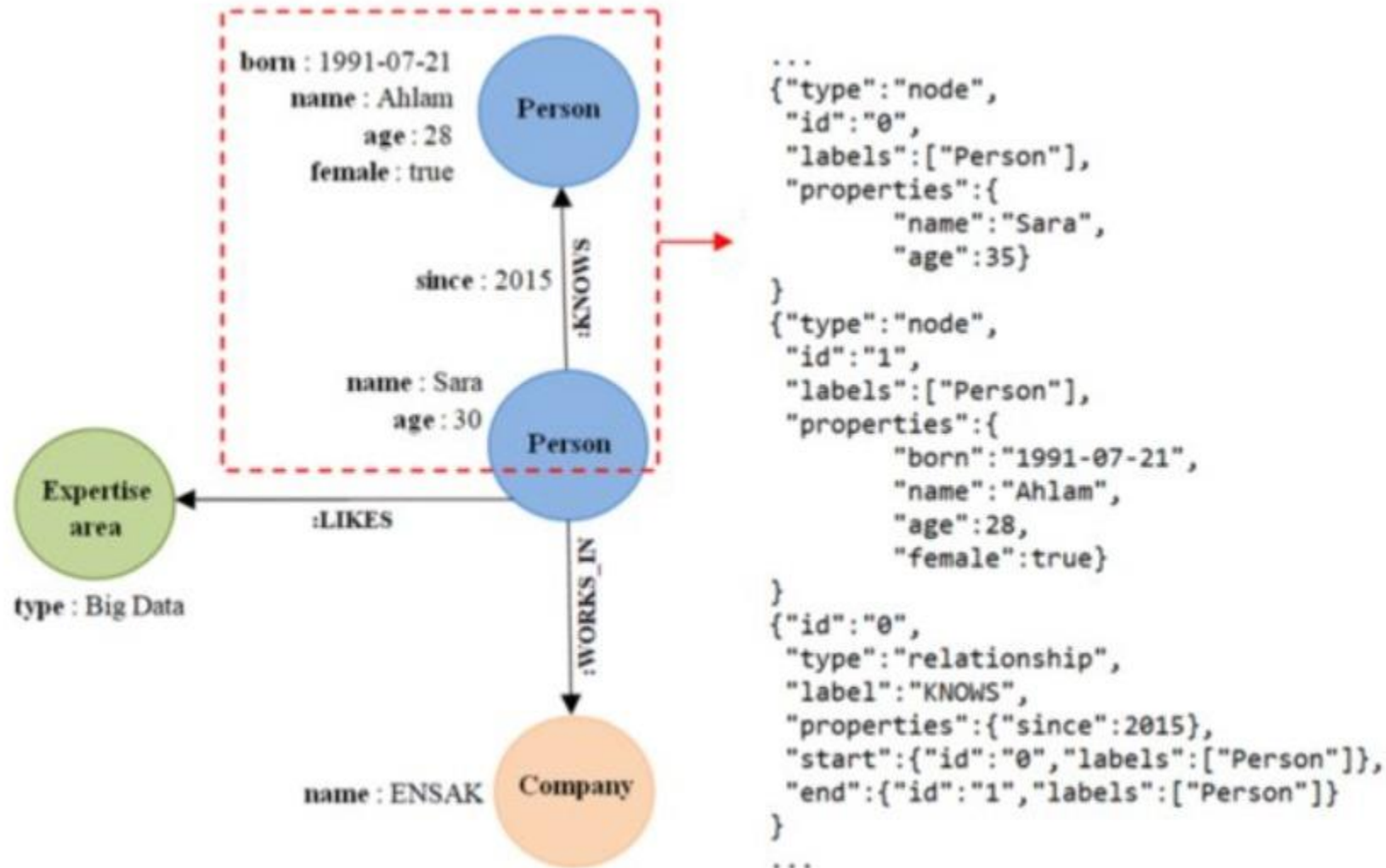
En este diseño, el **camino (path)** hacia cada nodo del árbol, está incluido en forma explícita en la información asociada a cada nodo.

- **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.

Nodes and **Edges** are used to represent and store information



Formatos de archivo: grafos

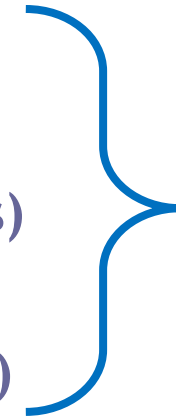


Navegando un mar de datos

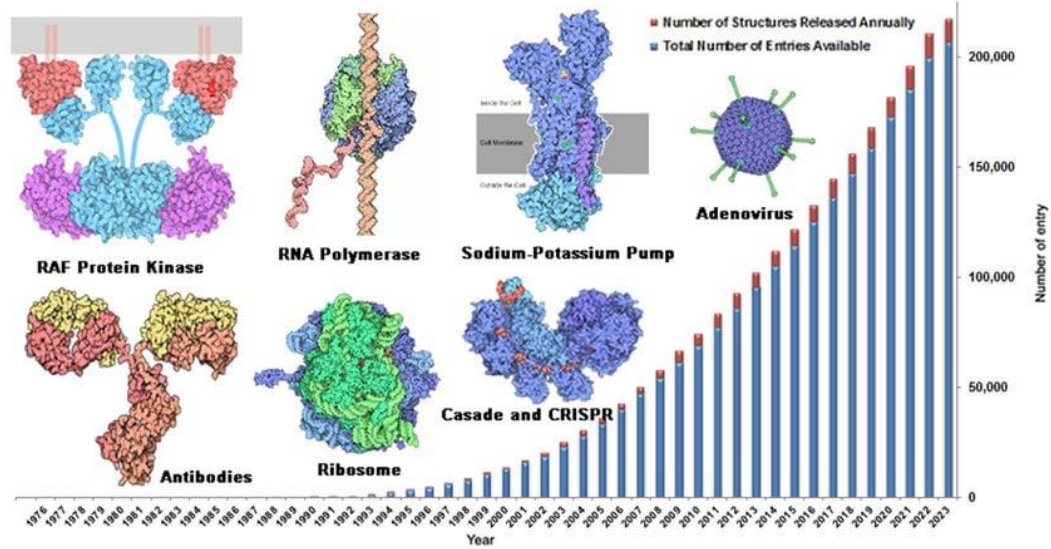
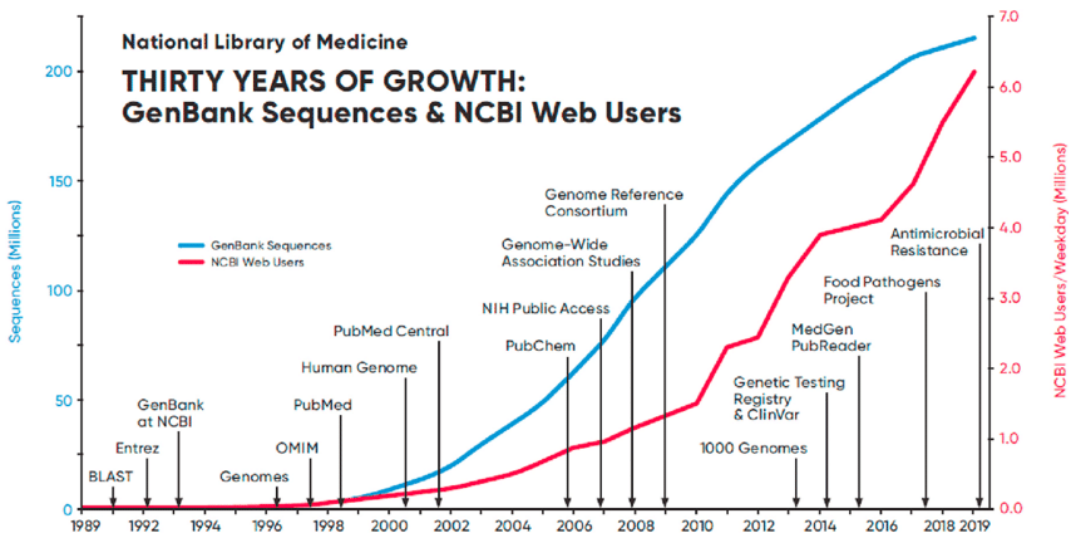
BASES DE DATOS BIOLÓGICAS

Qué tipos de datos manejamos en biotecnología?

- Secuencias (de ADN, de ARN, de proteínas, polímeros)
- Estructuras (de proteínas, de complejos macromoleculares)
- Taxonomía (organismos, especies, relaciones filogenéticas)
- Vías metabólicas celulares (reacciones químicas)
- Compuestos químicos (metabolitos, fármacos)
- Bibliografía (papers, artículos científicos)



Crecimiento y volumen de bases de datos biológicas





**257,571**
Structures from the
PDB archive

**1,062,058**
Computed Structure
Models (CSM)

Data Storage Growth

PDB Core Archive and Related Holdings (single copy)

Year	Size of PDB legacy archive snapshot	Size of PDB versioned archive	Size of EMDB Core Archive
2025	1583 GB for 246,905 PDB structures	298 GB	29 TB for 52,944 EMDB entries
2024	1437 GB for 229,564 PDB structures	269 GB	21 TB for 41,282 EMDB entries
2023	1242 GB for 214,121 PDB structures	231 GB	13 TB for 32,033 EMDB entries
2022	1086 GB for 199,755 PDB structures	218 GB	7.5 TB for 25,319 EMDB entries
2021	923 GB for 185,541 PDB structures	174 GB	4.5 TB for 18,059 EMDB entries
2020	822 GB for 173,005 PDB structures	148 GB	2.9 TB for 13,731 EMDB entries
2019	577 GB for 159,140 PDB structures	117 GB	1.7 TB for 10,370 EMDB entries
2018	441 GB for 136,472 PDB structures	85 GB	592 GB for 5,753 EMDB entries

Announcements

July 10, 2026
RefSeq Release 236 is available for FTP

This release includes:

Proteins: 482,864,455
Transcripts: 83,806,434
Organisms: 182,465

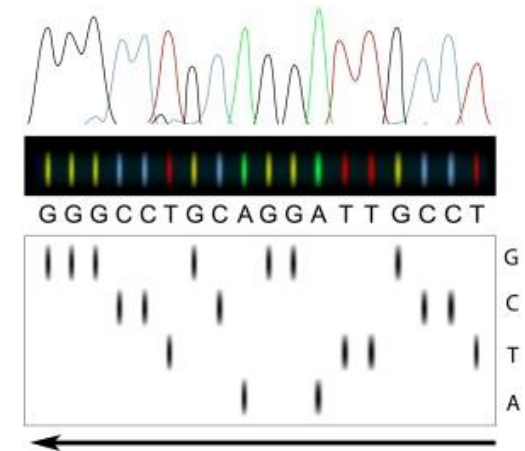
Available at: <https://ftp.ncbi.nlm.nih.gov/refseq/release/>
Documentation: [Release Notes](#)

See [previous announcements](#) or subscribe to [NCBI's refseq-announce mail list](#) to receive announcements.

Bases de datos primarias

- Una base de datos **primaria** es un repositorio de datos derivados de un **experimento** o de **conocimiento científico validado**.

- GenBank, ENA (secuencias nucleotídicas)
- dbSNP (mutaciones, SNPs)
- GenPept, Swissprot, TrEMBL (secuencias de proteínas)
- PDB (estructuras de proteínas, y macromoléculas)
- PubMed (literatura)
- KEGG (base de datos de vías metabólicas)



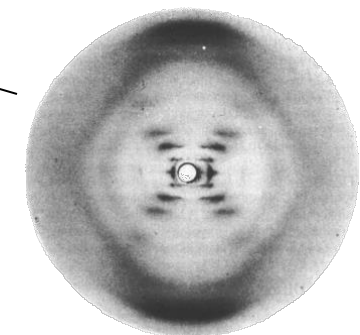
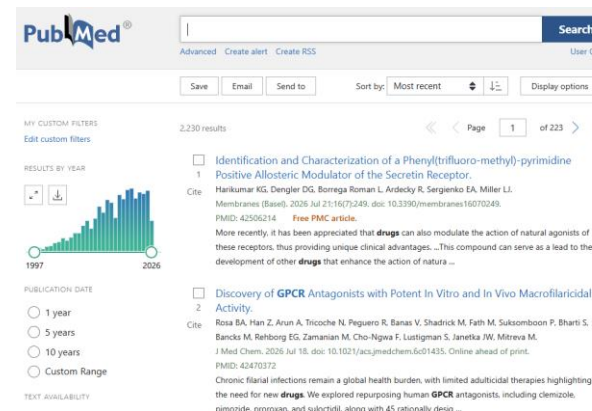
KEGG Home
Release notes
Current statistics
KEGG Database
KEGG overview
KEGG mapping
Color codes
KEGG identifier
KEGG WebLinks
Entry format

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of biological systems, such as cells, organisms, ecosystems and the biosphere from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See Release notes (July 1, 2026) for new and updated features.

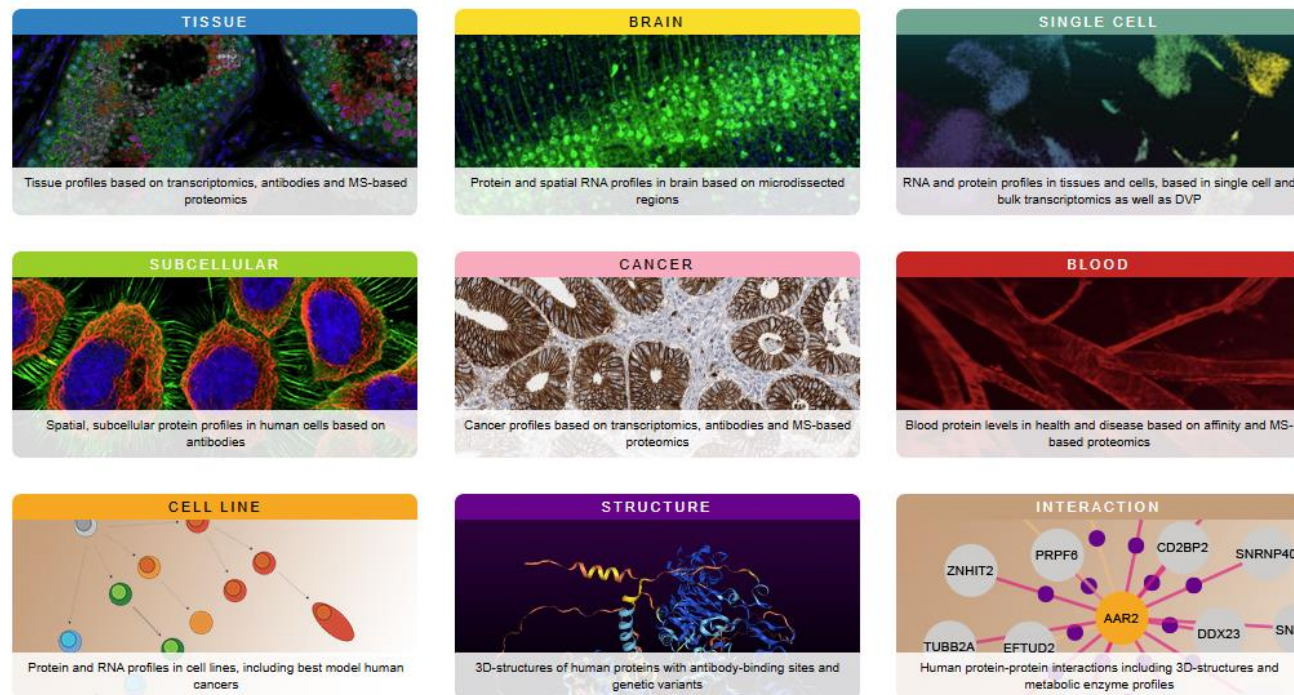
Announcement: KEGG reorganized into 23 types of data objects

New article: KEGG Syntax for comparison of organisms, organism groups and viruses by conserved gene repertoires



- 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)
 - TrEMBL (secuencias de proteínas traducidas a partir de secuencias de DNA)
- Las bases de datos organismo específicas son en general una mezcla entre primaria y secundaria.
- Hoy en día muchas otras bases de datos son híbridos
 - Integran diversos tipos de datos
 - son **primarias** para algunos tipos de información
 - y son a la vez **secundarias** para otros tipos

The Human Protein Atlas



Core Data Types and Sub-Atlases

Tissue Atlas: Protein expression profiles across normal human organs using antibody-based microarrays and RNA-seq.

Single Cell Type Atlas: Single-cell and single-nucleus RNA sequencing datasets mapping cell-type-specific expression in healthy tissues.

Cell Atlas: Subcellular localization data and immunofluorescence microscopy images of cell lines.

Pathology Atlas: Protein expression data linked to patient Survival statistics across major cancer types.

Brain Atlas: Detailed regional mapping and distribution of proteins in human, mouse, and pig brains.

Blood Protein Atlas: Profiles of circulating blood proteins and immune cell-type restrictions for biomarker discovery.

Metabolic Atlas: Maps of human metabolic enzymes integrated with 3D structural contexts.

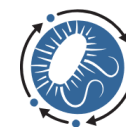
Cell Line Atlas: Transcriptomic and antibody-based protein expression data from standard human cell lines.

Protein Structure/Interaction Resource: Predictive 3D structures and protein-protein interaction networks powered by AI models like AlphaFold

<https://www.proteinatlas.org/>

Nucleotide databases:

- Genbank: International Collaboration
 - NCBI (USA), EMBL (Europe), DDBJ (Japan and Asia)
 - European Nucleotide Archive (ENA) - Europe
 - Sequence Read Archive (SRA)
- Organism specific databases
 - Human Genome Browser (UCSC)
 - Mouse Genome Informatics (MGI)
 - Drosophila Genetics Database (FlyBase)
 - Escherichia coli (EcoCYC)
 - Saccharomyces Genome Database (SGD)
 - The Arabidopsis Information Resource (TAIR)



EcoCYC



Saccharomyces
GENOME DATABASE



- 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 divisiones.



Divisiones de la base de datos GenBank

Secuencias de **alta calidad** en divisiones taxonómicas:

PRI → Primates

MAM → Mamíferos

INV → Invertebrados

Secuencias de **menor calidad** en divisiones
uso-específicas:

GSS → Genome Sequence Survey

EST → Expressed Sequence Tags

HTG → High Throughput Genomic Sequencing (unfinished
contigs, BACs, cosmids, chromosomes).

Organismal Divisions

Database	Division	BLAST	Example
BCT	Bacterial sequences	nr, month	
PRI	Primate sequences	nr, month	Human Phase 3
ROD	Rodent sequences	nr, month	
MAM	Other mammalian sequences	nr, month	
VRT	Other vertebrate sequences	nr, month	
INV	Invertebrate sequences	nr, month	Drosophila, C. elegans Phase 3
PLN	Plant and Fungal sequences	nr, month	Arabidopsis Phase 3
VRL	Viral sequences	nr, month	
PHG	Phage sequences	nr, month	
RNA	Structural RNA sequences	nr, month	
SYN	Synthetic and chimeric sequences	nr, month	
UNA	Unannotated sequences	nr, month	

Functional Divisions

Database	Division	BLAST	Example
EST	Expressed Sequence Tags	dbest, month	
STS	Sequence Tagged Sites	dbsts, month	
GSS	Genome Survey Sequences	dbgss, month	
HTG	High Throughput Genomic sequences	htgs, month	All Organisms: Phase 0, 1, and 2

- **Phase 0** sequences are single-pass reads of a single clone (not contigs usually).
- **Phase 1** sequences are unfinished, unordered, and contain gaps.
- **Phase 2** sequences are unfinished, ordered, and can contain one or more gaps.
- **Phase 3** sequences are high quality finished sequences that do not contain gaps.

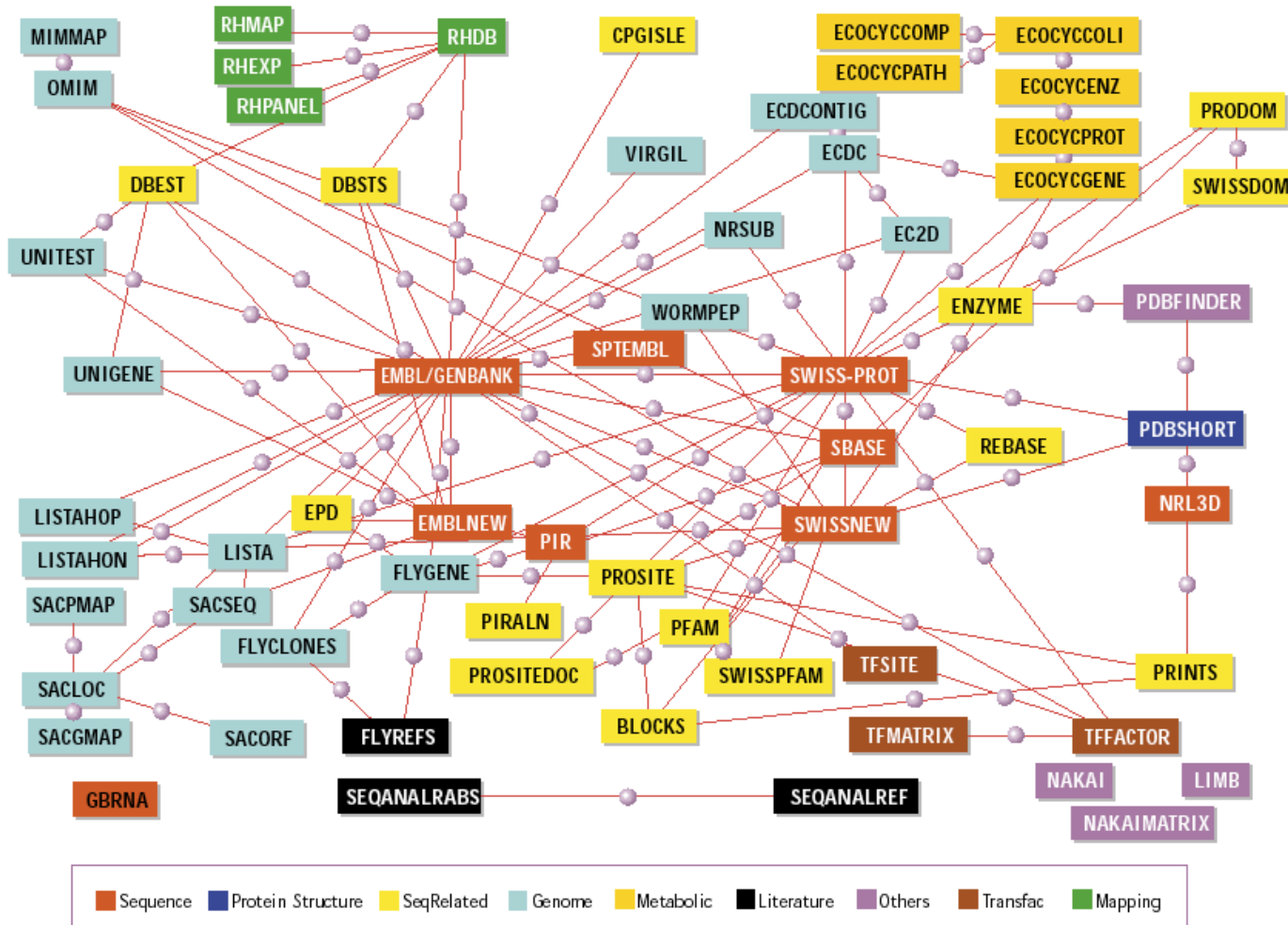
<https://www.ncbi.nlm.nih.gov/genbank/htgs/divisions/>

Concepto de Banco

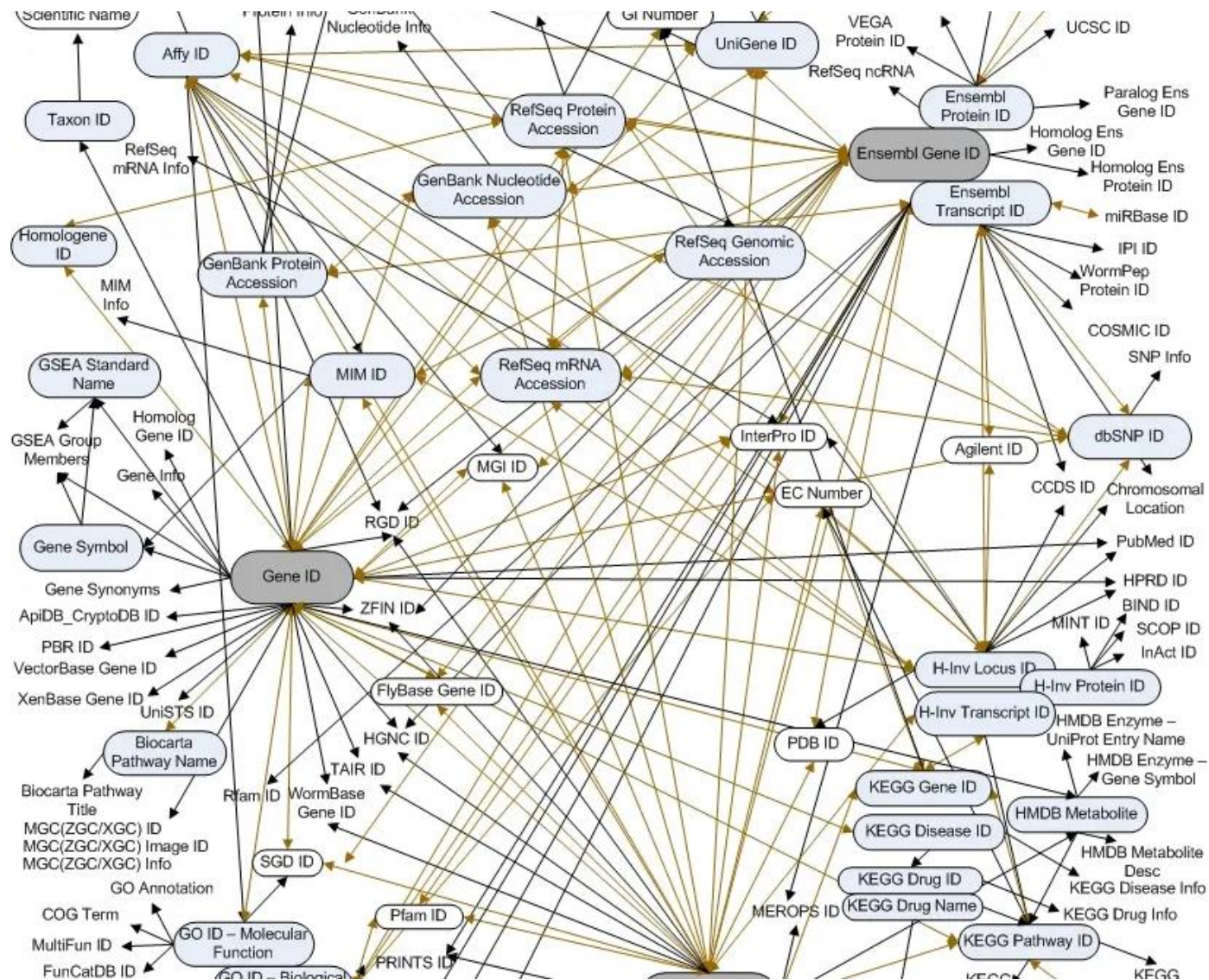
- Los depósitos son del depositante (submitter)
 - Redundante: misma secuencia depositada por varios depositantes
 - Con errores: solo el depositante puede corregirlos
 - Difícil de actualizar
-
- Para poder corregir, mejorar y mantener actualizada la anotación de los registros, el NCBI creó RefSeq (colección curada de registros de GenBank)
 - toma records de GenBank y los actualiza/corrije
 - unifica para reducir redundancia
 - Accession numbers del tipo XX_123456



Interconexiones entre bases de datos



Las claves primarias o secundarias (**accession numbers**) se usan para generar los links o asociaciones entre diversos recursos bioinformáticos.



Las claves primarias o secundarias (**accession numbers**) se usan para generar los links o asociaciones entre diversos recursos bioinformáticos.

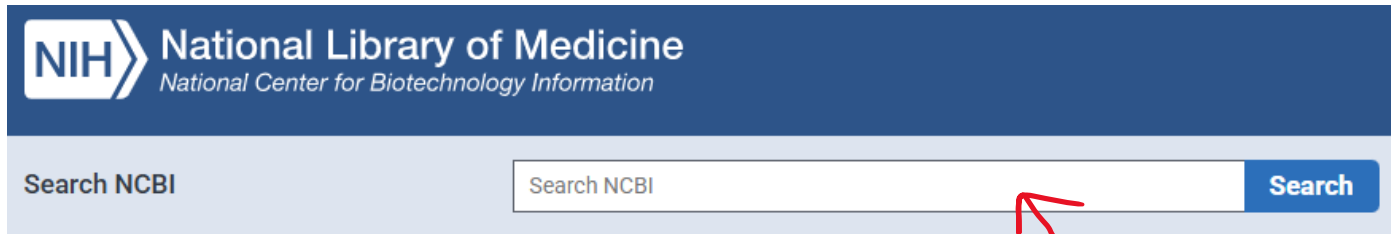
Búsquedas en bases de datos: búsquedas simples



UniProtKB ▾

Advanced | List Search

<https://www.uniprot.org/>



NIH National Library of Medicine
National Center for Biotechnology Information

Search NCBI Search NCBI Search

<https://www.ncbi.nlm.nih.gov/search/>

No imponen restricciones
El usuario tipea palabras libremente
Se usan estrategias para intentar
“adivinar” la intención del usuario



ENA European Nucleotide Archive

Enter text search terms Search

Examples: histone, BN000065

Enter accession View

Examples: Taxon:9606, BN000065, PRJEB402

<https://www.ebi.ac.uk/ena/browser/home>

Ejemplo: automatic term mapping - Entrez (NCBI)

- El usuario tipea libremente un **término** Entrez busca en qué **listas** mapea ese término:
- **MeSH (Medical Subject Headings):** vocabulario controlado utilizado para indexar artículos en PubMed.  Explosión de términos y sinónimos
- **Journals:** nombre completo del journal, abreviaturas usadas en MEDLINE y números ISSN.  Cancer, BMC Cancer, Nat Rev Cancer
- **Lista de frases:** cientos de miles de frases generadas a partir de MeSH y otros vocabularios controlados similares.  “breast cancer”
- **Indice de autores:** apellido e iniciales.  Alice Cancer, Pablo F Cancer, Matko Cancer, ...
- **Stopwords:** palabras comunes, presentes en casi todos los registros de la base de datos (a, an, by, of, the ...)

Ejemplo: automatic term mapping - Entrez (NCBI)

1

NIH National Library of Medicine
National Center for Biotechnology Information

Search NCBI

Search NCBI Search

Term mapping

2

NIH National Library of Medicine
National Center for Biotechnology Information

MeSH

Summary ▾ 20 per page ▾ Send to: ▾

Search results

Items: 1 to 20 of 409

☐ [Neoplasms](#)

1. New abnormal growth of tissue. Malignant **neoplasms** show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign **neoplasms**.
Year introduced: /diagnosis was NEOPLASM DIAGNOSIS 1964-1965

<https://www.ncbi.nlm.nih.gov/mesh/?term=cancer>

3

- [All MeSH Categories](#)
[Diseases Category](#)
Neoplasms
[Cysts](#)
- [Arachnoid Cysts](#)
 - [Bone Cysts +](#)
 - [Branchioma](#)
 - [Breast Cyst](#)
 - [Bronchogenic Cyst](#)
 - [Chalazion](#)
 - [Choledochal Cyst](#)
 - [Colloid Cysts](#)
 - [Dermoid Cyst](#)
 - [Epidermal Cyst](#)
 - [Esophageal Cyst](#)
 - [Follicular Cyst](#)
 - [Ganglion Cysts](#)
 - [Lymphocele](#)
 - [Mediastinal Cyst](#)
 - [Mesenteric Cyst](#)
 - [Mucocoele](#)
 - [Ovarian Cysts +](#)
 - [Pancreatic Cyst +](#)
 - [Parovarian Cyst](#)
 - [Pilonidal Sinus](#)
 - [Ranula](#)
 - [Synovial Cyst +](#)
 - [Tarlov Cysts](#)
 - [Thyroglossal Cyst](#)
 - [Urachal Cyst](#)
- [Hamartoma](#)
- [Hamartoma Syndrome, Multiple +](#)
 - [Pallister-Hall Syndrome](#)
 - [Tuberous Sclerosis](#)
- [Neoplasms by Histologic Type](#)
- [Histiocytic Disorders, Malignant +](#)
 - [Leukemia +](#)
 - [Lymphoma +](#)
 - [Neoplasm, Lymphatic Tissue +](#)
 - [Neoplasms, Complex and Mixed +](#)
 - [Neoplasms, Connective and Soft Tissue +](#)
 - [Neoplasms, Germ Cell and Embryonal +](#)
 - [Neoplasms, Glandular and Epithelial +](#)
 - [Neoplasms, Gonadal Tissue +](#)

Term explosion

Búsquedas simples: pros / cons

- Ventajas

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

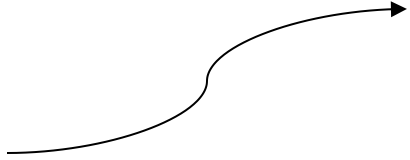
- Desventajas

- 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)
- <https://www.ncbi.nlm.nih.gov/books/NBK3799/table/catalog.Taa/>
 - `Escherichia coli[organism]`
 - `review[publication type], 2026[DP]`
 - `attenuator[feature key]`
- **EMBL-EBI:** formulario avanzado (no hay que conocer términos o tags)



PubMed:
Cancer[Journal]
Cancer[AU]
Cancer[TI]
Cancer[AB]

Búsquedas avanzadas: Uniprot (EMBL-EBI)

Find your protein

UniProtKB ▾

Advanced | List Search

Examples: Insulin, APP, Human, P05067, organism_id:9606

Advanced Search¹

Searching in
UniProtKB ▾

Gene Name [GN] SC5D Remove

AND Taxonomy [OC] Mammalia (mammals) [40674] Remove

AND Keyword [KW] activator Remove

AND All a4_human, P05067, cdc7 human Remove

Add Field

Type * in the search box for the selected field.

- UniProtKB AC
- Entry Name [ID]
- Secondary Accession
- Protein Name [DE]
- Gene Name [GN]
- Organism [OS]
- Taxonomy [OC]
- Virus host
- Protein Existence [PE]
- Function ▶
- Subcellular location ▶

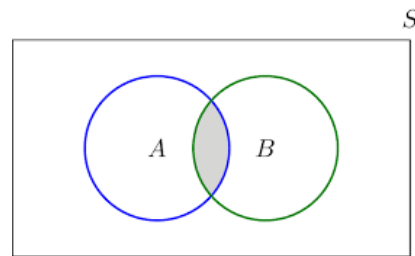
- Las búsquedas avanzadas son más **fáciles de construir**.
- **OK** para usuarios interaccionando **via web**.
- **NOT OK** para interacciones desde otras interfaces:
 - Acceso programático (via APIs)
 - Acceso repetitivo

Operadores lógicos

- En búsquedas simples o avanzadas siempre tienen a disposición operadores lógicos para encadenar términos

- **AND (intersección)**

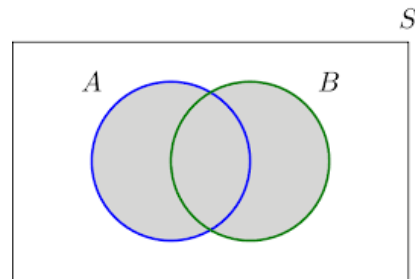
- human AND genome
- +human +genome
- human && genome



Los operadores lógicos funcionan sobre conjuntos. Para pocos conjuntos pueden visualizarse con diagramas de Venn.

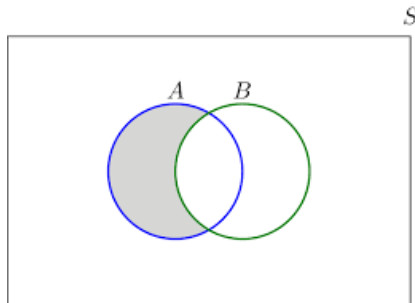
- **OR (unión)**

- human OR genome
- human || genome



- **NOT (subconjunto)**

- human NOT genome



https://es.wikipedia.org/wiki/Teoría_de_conjuntos

Orden de los términos en una búsqueda

- 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)

My favorite part of the work day
is the end of it.



someecards
user card