"I hope you've got a lot of disk space, Ted.
I think I accidentally just faxed you
the entire Internet."

Introducción a las Bases de datos biológicas

Febrero 2010

Porque las bases de datos ?

- Crecimiento exponencial de los datos biológicos
- Datos (secuencias, 3D estructuras, análisis gel 2D, MS análisis….) no son publicados en revistas, pero si en bases de datos
- Son usadas en investigación biológica, como lo eran la revistas científicas !
- Biólogos dependen de los computadores para almacenar, organizar, buscar, manipular, y recuperar los datos
- Libre Acceso es clave
- Base de todas las herramientas bioinformáticas

Que es una base de datos ?

- Una colección
  - estructurada
  - De fácil búsqueda (indexada) -> tabla de contenido
  - Actualizada periódicamente (release) -> Nuevas dediciones
  - Referencias cruzadas (hipervínculos) -> vínculos con otras DB
- Incluye la herramientas (software) para acceso, actualización, inserción, borrado…. en la DB
- Manejo del almacenamiento de los datos: Texto plano (flat files), tablas vinculadas (bases de datos relacionales)

DB: Texto plano « flat file »

Base de datos de estudiantes:
(texto plano, 3 entradas)

```
Código: 183023
Nombre: Julián
Apellido : Pulecio
Cursos: 19003-01, 21001-01
Email: jpul@ibun.unal.edu.co
//
//
Código: 183024
Nombre: Sonia
Apellido : Cuartas
Cursos : 19003-01, 17001-01
Email: soniacol@hotmail.com
//
Código: 183025
Nombre: Jaime
Apellido : Moreno
Cursos : 19003-01
Email: pm186111@ibun.unal.edu.co
//
```

- Facil de manejar: todas las entradas de pueden ver al tiempo !

Bases de datos « relacionales »

Curso	Nom.Curso
19003-01	Bioinformática
17001-01	Bioquímica Avanzada
21001-01	Análisis Molecular

Alumno	Código
Gutiérrez	183023
Cuartas	183024
Moreno	182425

Curso	Código
19003-03	183023
19003-03	183024
19003-03	182425
17001-01	183024
21001-01	183023

Fácil: manejo y selección de la salida

Algunas estadísticas

- Más de 2000 bases de datos diferentes
- Generalmente accesibles a través de WEB (
  - Google: http://www.google.com/
  - Biohunt: http://www.expasy.org/BioHunt/
  - Amos' links: www.expasy.ch/alinks.html
- Tamaño variable: <100Kb a >150Gb
  - DNA: > 180 Gb
  - Proteinas: 7 Gb
  - Estructuras 3D : 15 Gb
  - Otras: Pequeñas

Clases de DB para las ciencias de la vida

- Secuencias (DNA, proteínas) -> DB primarias
- Genómicas
- Dominios/familias proteícos -> DB secundarias
- Mutación/polimorfismo
- Proteómica (2D gel, MS)
- 3D estructura -> DB de estructuras
- Metabolismo
- Bibliografía
- Otras

Contenido Ideal mínimo de una DB de « secuencias »

- Número de acceso (AC)
- Secuencias !!
- Referencias
- Datos taxonómicos
- ANOTACIONES/CURADURIA
- Palabras claves
- Referencias cruzadas
- Documentación

Ejemplo…Formato de entrada para secuencias

…entrada SWISS-PROT, en formato fasta :

```
>sp|P01588|EPO_HUMAN ERYTHROPOIETIN PRECURSOR - Homo sapiens (Human).
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPD
AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```

Formatos empleados (texto plano):
- fasta
- GCG
- NBRF/PIR
- MSF….
- Formatos estandarizados ?

Bases de datos anotadas - curadas?

- contienen la secuencia, comentarios, referencias de la literatura, notas sobre experimentos
- Derivadas de la integracion de las herramientas de cómputo y conocimiento biológico
  - por ejemplo, genes conocidos y predichos

- Registros añadidos solo despues de verificar su precision y las anotaciones
- Ejemplo :
  SWISS-PROT, OMIM, RefSeq, LocusLink

Ejemplo: Base de datos de secuencia

SWISS-PROT Flat file

Taxonomía

Referencia

```
ID   EPO_HUMAN      STANDARD;      PRT;   193 AA.
AC   P01588; Q9UHA0; Q9UEZ5; Q9UDZ0;
DT   21-JUL-1986 (Rel. 01, Created)
DT   21-JUL-1986 (Rel. 01, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   Erythropoietin precursor.
GN   EPO.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=85137899; PubMed=3838366;
RA   Jacobs K., Shoemaker C., Rudersdorf R., Neill S.D., Kaufman R.J.,
RA   Mufson A., Seehra J., Jones S.S., Hewick R., Fritsch E.F.,
RA   Kawakita M., Shimizu T., Miyake T.;
RT   "Isolation and characterization of genomic and cDNA clones of human
RT   erythropoietin.";
RL   Nature 313:806-810(1985).
```

Ejemplo: Base de datos de secuencia

SWISS-PROT Flat file

Anotaciones

Referencias Cruzadas

Palabras claves

```
...
CC   -!- FUNCTION: ERYTHROPOIETIN IS THE PRINCIPAL HORMONE INVOLVED IN THE
CC       REGULATION OF ERYTHROCYTE DIFFERENTIATION AND THE MAINTENANCE OF A
CC       PHYSIOLOGICAL LEVEL OF CIRCULATING ERYTHROCYTE MASS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY KIDNEY OR LIVER OF ADULT MAMMALS
CC       AND BY LIVER OF FETAL OR NEONATAL MAMMALS.
CC   -!- PHARMACEUTICAL: Available under the names Epogen (Amgen) and
CC       Procrit (Ortho Biotech).
...
DR   EMBL; X02158; CAA26095.1; -.
DR   EMBL; X02157; CAA26094.1; -.
DR   EMBL; M11319; AAA52400.1; -.
DR   EMBL; AF053356; AAC78791.1; -.
DR   EMBL; AF202308; AAF23132.1; -.
DR   EMBL; AF202306; AAF23132.1; JOINED.
...
KW   Erythrocyte maturation; Glycoprotein; Hormone; Signal; Pharmaceutical.
```

Ejemplo: Base de datos de secuencia

Anotación

Secuencia

```
FT   SIGNAL       1     27
FT   CHAIN       28    193       ERYTHROPOIETIN.
FT   PROPEP     190    193       MAY BE REMOVED IN PROCESSED PROTEIN.
FT   DISULFID    34    188
FT   DISULFID    56     60
FT   CARBOHYD    51     51       N-LINKED (GLCNAC...).
FT   CARBOHYD    65     65       N-LINKED (GLCNAC...).
FT   CARBOHYD   110    110       N-LINKED (GLCNAC...).
FT   CARBOHYD   153    153       O-LINKED (GALNAC...).
FT   VARIANT    131    132       SL -> NF (IN AN HEPATOCELLULAR
FT                               CARCINOMA).
FT                               /FTId=VAR_009870.
FT   VARIANT    149    149       P -> Q (IN AN HEPATOCELLULAR CARCINOMA).
FT                               /FTId=VAR_009871.
FT   CONFLICT    40     40       E -> Q (IN REF. 1; CAA26095).
FT   CONFLICT    85     85       Q -> QQ (IN REF. 5).
FT   CONFLICT   140    140       G -> R (IN REF. 1; CAA26095).
**
**   ################    INTERNAL SECTION    ################
**CL 7q22;
SQ   SEQUENCE   193 AA;  21306 MW;  C91F0E4C26A52033 CRC64;
     MGVHECPAWL WLLSLLSLP LGLPVLGAPP RLICDSRVLE RYLLEAKEAE NITTGCAEHC
     SLNENITVPD TKVNFYAWKR MEVGQQAVEV WQGLALLSEA VLRGQALLVN SSQPWEPLQL
     HVDKAVSGLR SLTTLLRALG AQKEAISPPD AASAAPLRTI TADTFRKLFR VYSNFLRGKL
     KLYTGEACRT GDR
//
```

Bases de Datos de Secuencias de nucleótidos

EMBL/GenBank/DDJB

- Contienen principalmente la misma información en el plazo de 2-3 días (pocas diferencias en el formato y sintaxis)
- Contienen las secuencias (genes únicos, ESTs, genomas completos, etc.) derivadas de:
  - Los proyectos genoma y centro de secuenciación
  - Científicos Individuales
  - Oficinas de patentes (es decir oficina de patentes europea, EPO)

Datos No-confidenciales son intercambiados diariamente

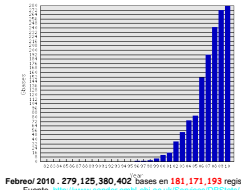
Actualmente: 20 secuencias $x10^6$, sobre 30 $x10^9$ bp;

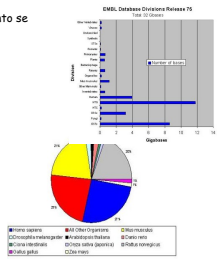
Estadisticas: http://www3.ebi.ac.uk/Services/DBStats/

secuencias > 73´000 especies diferentes;

Impresionante incremento en secuencias de nucleótidos

- Datos en EMBL... su gran incremento se debió al surgimiento de PCR...





Febrero/ 2010 . 279,125,380,402 bases en 181,171,193 registros.
Fuente http://www.ebi.ac.uk/Services/DBStats/

1980: 80 genes totalmente secuenciados!

EMBL/GenBank/DDBJ

- Longitud Heterogenea de secuencias: genomas fragmentos…
- Tamaños de Secuencia:
  - max 300'000 pb / entrada (! secuencias genómicas, sobrelapadas)
  - min 10 bp / entrada
- **Archivo**: nada sale!!!! -> muy redundante!
- lleno de errores: en las secuencias, en anotaciones, en atribución de CDS…
- Ninguna consistencia en las anotaciones; la mayoría de las anotaciones son hechas por los investigadores (submitters); heterogeneidad en la calidad y en la realización y actualización de la información

EMBL/GenBank/DDJB

- Se puede encontrar información inesperada en estas bd:

```
FT   source          1..124
FT                   /db_xref="taxon:4097"
FT                   /organelle="plastid:chloroplast"
FT                   /organism="Nicotiana tabacum"
FT                   /isolate="Cuban cahibo cigar, gift from President Fidel
FT                   Castro"
```

- O:

```
FT   source          1..17084
FT                   /chromosome="complete mitochondrial genome"
FT                   /db_xref="taxon:9267"
FT                   /organelle="mitochondrion"
FT                   /organism="Didelphis virginiana"
FT                   /dev_stage="adult"
FT                   /isolate="fresh road killed individual"
FT                   /tissue_type="liver"
```

Entrada de EMBL : ejemplo

```
ID   LISOD      standard; DNA; PRO; 756 BP.
XX
AC   X64011; S78972;
XX
SV   X64011.1
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   30-JUN-1993 (Rel. 36, Last updated, Version 6)
XX
DE   Listeria  ivanovii sod gene for superoxide dismutase
XX
KW   sod gene; superoxide dismutase.
XX
OS   Listeria ivanovii
OC   Bacteria; Firmicutes; Bacillus/Clostridium group;
OC   Bacillus/Staphylococcus group; Listeria.
XX
RN   [1]
RX   MEDLINE; 92140371.
RA   Haas A., Goebel W.;
RT   "Cloning of a superoxide dismutase gene from Listeria ivanovii by
RT   functional complementation in Escherichia coli and characterization of the
RT   gene product.";
RL   Mol. Gen. Genet. 231:313-322(1992).
XX
RN   [2]
RP   1-756
RA   Kreft J.;
RT   ;
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases.
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL   Hubland, 8700 Wuerzburg, FRG
XX
DR   SWISS-PROT; P28763; SODM_LISIV.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..756
FT                   /db_xref="taxon:1638"
FT                   /organism="Listeria ivanovii"
FT                   /strain="ATCC 19119"
FT   RBS             95..100
FT                   /gene="sod"
FT   terminator      723..746
FT                   /gene="sod"
FT   CDS             109..717
FT                   /db_xref="SWISS-PROT:P28763"
FT                   /transl_table=11
FT                   /gene="sod"
FT                   /EC_number="1.15.1.1"
```

Palabras claves

Taxonomía

Referencias

Referencias cruzadas

Entrada de EMBL : (Cont.)

```
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases.
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL   Hubland, 8700 Wuerzburg, FRG
XX
DR   SWISS-PROT; P28763; SODM_LISIV.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..756
FT                   /db_xref="taxon:1638"
FT                   /organism="Listeria ivanovii"
FT                   /strain="ATCC 19119"
FT   RBS             95..100
FT                   /gene="sod"
FT   terminator      723..746
FT                   /gene="sod"
FT   CDS             109..717
FT                   /db_xref="SWISS-PROT:P28763"
FT                   /transl_table=11
FT                   /gene="sod"
FT                   /EC_number="1.15.1.1"
FT                   /product="superoxide dismutase"
FT                   /protein_id="CAA45406.1"
FT                   /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVSG
FT                   HAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLKAA
FT                   IESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPVLGL
FT                   DVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
XX
SQ   Sequence 756 BP; 247 A; 136 C; 151 G; 222 T; 0 other;
     cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat        60
     gtaatttctt ttcacataaa taataaacaa tccgaggagg aatttttaat gacttacgaa       120
     ttaccaaaat taccttatac ttatgatgct ttggagccga attttgataa agaaacaatg       180
     gaaattcact atacaaagca ccacaatatt tatgtaacaa aactaaatga agcagtctca       240
     ggacacgcag aacttgcaag taaacctggg gaagaattag ttgctaatct agatagcgtt       300
     cctgaagaaa ttcgtggcgc agtacgtaac cacggtggtg gacatgctaa ccatacttta       360
     ttctggtcta gtcttagccc aaatggtggt ggtgctccaa ctggtaactt aaaagcagca       420
     atcgaaagcg aattcggcac atttgatgaa ttcaaagaaa aattcaatgc ggcagctgcg       480
     gctcgttttg gttcaggatg ggcatggcta gtagtgaaca atggtaaact agaaattgtt       540
     tccactgcta accaagattc tccacttagc gaaggtaaaa ctccagttct tggcttagat       600
     gtttgggaac atgcttatta tcttaaattc caaaaccgtc gtcctgaata cattgacaca       660
     ttttggaatg taattaactg ggatgaacga aataaacgct ttgacgcagc aaaataatta       720
     tcgaaaggct cacttaggtg ggtcttttta tttcta                                 756
//
```

Anotación

Secuencia

Entrada de GenBank : ejemplo

☐ **1:** AJ416340. Escherichia coli ...[gi:16304811]

```
LOCUS       ECO416340               1470 bp    DNA     linear   BCT 07-JUL-2002
DEFINITION  Escherichia coli partial tnpA gene for transposase and blaCTX-M-1
            gene for CTX-M-1 beta-lactamase.
ACCESSION   AJ416340
VERSION     AJ416340.1  GI:16304811
KEYWORDS    blaCTX-M-1 gene; CTX-M-1 beta-lactamase; tnpA gene; Transposase.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1
  AUTHORS   Saladin,M., Cao,V.T., Lambert,T., Donay,J.L., Herrmann,J.L.,
            Ould-Hocine,Z., Verdet,C., Delisle,F., Philippon,A. and Arlet,G.
  TITLE     Diversity of CTX-M beta-lactamases and their promoter regions from
            Enterobacteriaceae isolated in three Parisian hospitals
  JOURNAL   FEMS Microbiol. Lett. 209 (2), 161-168 (2002)
  MEDLINE   22003747
   PUBMED   12007800
REFERENCE   2  (bases 1 to 1470)
  AUTHORS   Arlet,G.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-OCT-2001) Arlet G.J., Bacteriology, UFR
            Saint-Antoine, 27, rue de Chaligny, Paris 75571 cedex 12, FRANCE
```

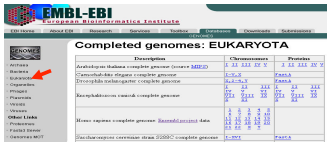
Entrada de GenBank : ejemplo

```
FEATURES             Location/Qualifiers
     source          1..1470
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /strain="SCI"
                     /db_xref="taxon:562"
                     /country="France:Paris"
     repeat_region   1..514
                     /insertion_seq="ISEcp1"
     gene            1..307
                     /gene="tnpA"
     CDS             <1..307
                     /gene="tnpA"
                     /function="transposition"
                     /codon_start=2
                     /transl_table=11
                     /product="transposase"
                     /protein_id="CAC95167.1"
                     /db_xref="GI:16304812"
                     /db_xref="SPTREMBL:Q932Z8"
                     /translation="EKRGNAENYIKEAKYDMAVGHLLLKSFWANEAVFQMMMLSYNLF
BASE COUNT      429 a    300 c    376 g    365 t
ORIGIN
        1 tgaaaagcgt ggtaatgctg aaaactatat caaagaagcc aaatacgaca tggcggtggg
       61 tcatctcttg ctaaagtcat tttgggcgaa tgaagccgtg tttcaaatga tgatgctttc
      121 atataaccta tttttgttgt tcaagtttga ttccttggac tcttcagaat acagacagca
```

Divisiones EMBL

- EMBL ha sido divida en sub-bases de datos para permitir el fácil mantenimiento y busqueda de datos
  - fun, hum, inv, mam, org, phg, pln, pro, rod, syn, unc, vrl, vrt
  - est, gss, htg, htc, sts, patent

http://www.ebi.ac.uk/genomes/



DB genómicas

- Contienen información sobre genes, su localización (mapping), nomenclatura y links a DB de secuencia; **usualmente no contiene secuencias**;
- Existen para los organismo más importantes en investigación de las ciencias de la vida;
- Ejemplos: MIM, GDB (human), MGD (mouse), FlyBase (Drosophila), SGD (yeast), MaizeDB (maize), SubtiList (B.subtilis), etc.;
- Formato: generalmente relacional (Oracle, SyBase or AceDb).

MIM

- OMIM™: Online Mendelian Inheritance in Man
- Catalogo de genes humanos y desordenes genéticos
- Contiene un sumario de literatura, fotos, y referencias. Contiene numerosos links a información de secuencia y articulos.

MIM:

*133170 ERYTHROPOIETIN; EPO

Alternative titles; symbols

EP

TABLE OF CONTENTS

TEXT
REFERENCES
SEE ALSO
CONTRIBUTORS
CREATION DATE
EDIT HISTORY

Database Links

Gene Map Locus: 7q21

Note: pressing the symbol will find the citations in MEDLINE whose text most closely matches the text of the preceding OMIM paragraph, using the Entrez MEDLINE neighboring function.

TEXT

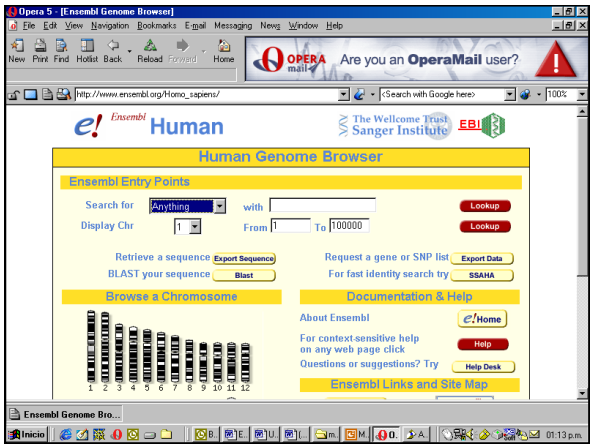
Human erythropoietin is an acidic glycoprotein hormone with molecular weight 34,000. As the prime regulator of red cell production, its major functions are to promote erythroid differentiation and to initiate hemoglobin synthesis. Sherwood and Shouval (1986) described a human renal carcinoma cell line that continuously produces erythropoietin. Eschbach et al. (1987) demonstrated the effectiveness of recombinant human erythropoietin in treating the anemia of end-stage renal disease. Lee-Huang (1984) cloned human erythropoietin cDNA in E. coli. McDonald et al. (1986) and Shoemaker and Mitsock (1986) cloned the mouse gene and the latter workers showed that coding DNA and amino acid sequence are about 80% conserved between man and mouse. This is a much higher order of conservation than for various interferons, interleukin-2, and GM-CSF.

Proyectos DB de Secuencia y genómica

- Ensembl
- TIGR

Ensembl: automatic annotation of eukaryotic genomes

- Contiene todas las secuencias de DNA del genoma humano normalmente disponibles para el dominio público.
- Anotación automatizada: usando diferentes herramientas de software, identificando dentro de las secuencias de DNA:
  - Genes (conocidos o predecidos)
  - Polimorfismos de un solo nucleótido (SNPs)
  - Repeticiones
  - Homologías
- Creado y mantenido por el EBI y el Sanger Center (UK)



Ensembl: www.ensembl.org

Con Ensembl usted puede ...

- Buscar secuencias de DNA en el genoma humano
- Visualizar los mapas cromosómicos
- Encontrar genes, SNPs y sitios comunes con el genoma de ratón
- Mirar proteínas y familias de proteínas

•Ensemble provee:

- Identificación del 90% de los genes humanos conocidos
- Predicción de 10,000 genes adicionales, todos con evidencia soportada

BD de secuencias de proteína

- **SWISS-PROT**: creada en 1986 (A.Bairoch)

  http://www.expasy.org/sprot/

- **TrEMBL**: creada en 1996; complementa a SWISS-PROT; derivada de las traducciones de CDS EMBL automatizadas (versión «proteómica» de EMBL)

- **PIR-PSD**: Protein Information Resources

  http://pir.georgetown.edu/

BD de secuencias de proteína

- **PRF**: Protein Research Foundation (Japan): Peptide/Protein Sequence Database (PRF/SEQDB)

  http://www.prf.or.jp/en/index.html

- **GenPept:** producida a partir del analisis del release de GenBank correspondiente para regiones codificantes traducidas.

- Muchas bases de datos de proteina especializada para familias especificas o grupos de proteínas.

  - Ejemplos: **YPD** (proteínas de levadura), **AMSDb** (peptidos antibacterianos), **GPCRDB** (receptores 7 TM), **IMGT** (sistema immune) etc.

Swissprot

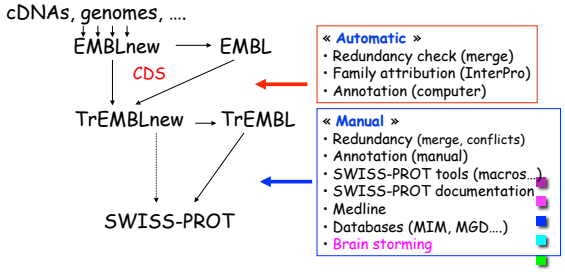
http://www.expasy.ch/sprot/

SWISS-PROT es una base de datos de secuencias de proteína,

- Colaboración entre el SIB (CH) y EMBL/EBI, 1986
- Curada y Anotada (descripción de la función, estructura de los dominios, modificaciones post- traduccionales, variantes)
- Mínima redundancia y alta Integración con otras BD.
- Release semanal; disponible en cerca de 50 servidores a través del mundo, la fuente principal que es ExPASy

SWISS-PROT
Protein knowledgebase
TrEMBL
Computer-annotated supplement to SWISS-PROT

Historia de una entrada en SWISS-PROT



Once in SWISS-PROT, the entry is no more in TrEMBL, but still in EMBL (archive)

CDS: proposed and submitted at EMBL by authors or by genome projects (can be experimentally proved or derived from gene prediction programs). TrEMBL does not translate DNA sequences nor use gene prediction programs: only take CDS already annotated in the EMBL entry.

TrEMBL: ejemplo

TrEMBL: Q9UDZ0

```
ID   Q9UDZ0      PRELIMINARY;      PRT;   136 AA.
AC   Q9UDZ0;
DT   01-MAY-2000 (TrEMBLrel. 13, Created)
DT   01-MAY-2000 (TrEMBLrel. 13, Last sequence update)
DT   01-JUN-2000 (TrEMBLrel. 14, Last annotation update)
DE   ERYTHROPOIETIN PROTEIN (FRAGMENT).
GN   ERYTHROPOIETIN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 93384593. [NCBI, ExPASy, Israel, Japan]
RA   Funakoshi A., Kuta H., Baba T., Shimizu S.;
RT   "Gene expression of mutant erythropoietin in hepatocellular
RT   carcinoma.";
RL   Biochem. Biophys. Res. Commun. 195(2):717-722(1993).
DR   EMBL; S65458; AAD15964.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   INTERPRO; IPR001323; -.
DR   INTERPRO; IPR002013; -.
DR   PFAM; PF00758; EPO_TPO; 1.
DR   PRINTS; PR00272; ERYTHROPTN.
DR   PROTOMAP; Q9UDZ0;
DR   PRESAGE; Q9UDZ0.
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
FT   NON_TER       1       1
SQ   SEQUENCE   136 AA;  15173 MW;  BCB9B1F0D819DAB3 CRC64;
     EPCSLHDNIT VPDTKVNFYA WKRMEVEQQA VEVWQGLALL SEAVLRGQAL LVNSSQPWEP
     LQLHVDKAVS GLRSLTTLLR ALGAQKEAIS PPDAASAAPL RTITADTFRK LFRVYSNFLR
     GKLKLYTGEA CRTGDR
//
```

« Old » TrEMBL which does not exist anymore, because it has been integrated into the SWISS-PROT EPO_HUMAN entry: low redundancy

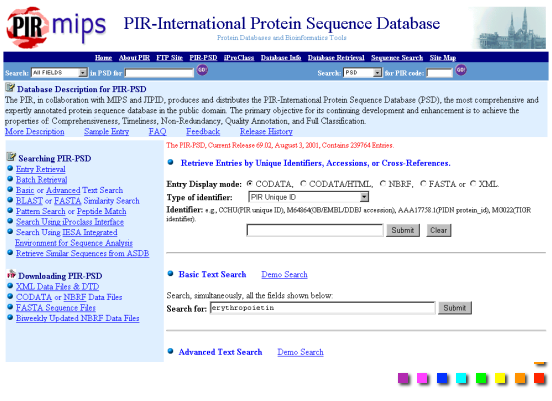
SWISS-PROT / TrEMBL (SPTR) : minima redundancia

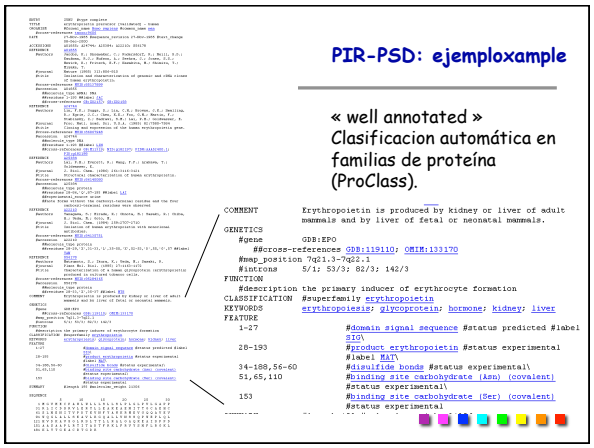
Human EPO: Blastp results

☐	sp	P01588	EPO_HUMAN ERYTHROPOIETIN PRECURSOR [EPO] [Homo sapiens]	302 2e-81
☐	sp	P07865	EPO_MACFA ERYTHROPOIETIN PRECURSOR [EPO] [Macaca fasc...	273 8e-73
☐	sp	Q28533	EPO_MACMU ERYTHROPOIETIN PRECURSOR [EPO] [Macaca mula...	272 3e-72
☐	sp	P33708	EPO_FELCA ERYTHROPOIETIN PRECURSOR [EPO] [Felis silve...	254 6e-67
☐	sp	P29676	EPO_RAT ERYTHROPOIETIN PRECURSOR [EPO] [Rattus norveg...	248 4e-65
☐	sp	P48617	EPO_BOVIN ERYTHROPOIETIN PRECURSOR [EPO] [Bos taurus]	247 6e-65
→ ☐	tn	AAK28825	EPO [Epo] [Mus musculus]	246 2e-64
→ ☐	sp	P07321	EPO_MOUSE ERYTHROPOIETIN PRECURSOR [EPO] [Mus musculus]	246 2e-64
→ ☐	sp	P33709	EPO_SHEEP ERYTHROPOIETIN PRECURSOR [EPO] [Ovis aries]	243 1e-63
→ ☐	tr	Q8NYB8	ERYTHROPOIETIN PRECURSOR [EPO] [Sus scrofa]	241 4e-63
→ ☐	sp	P49157	EPO_PIG ERYTHROPOIETIN PRECURSOR (FRAGMENT) [EPO] [Su...	241 4e-63
☐	tr	Q9GKA2	ERYTHROPOIETIN [Oryctolagus cuniculus]	240 1e-62
☐	tr	Q9GKA3	ERYTHROPOIETIN [Oryctolagus cuniculus]	240 1e-62
☐	sp	P33707	EPO_CANFA ERYTHROPOIETIN PRECURSOR (FRAGMENT) [EPO] [...	226 2e-58

¿cuántas secuencias hay en SWISS-PROT + TrEMBL ?

■100'000 + 540'000 = aprox. 400'000





PIR-PSD: ejemploxample

« well annotated »
Clasificacion automática en familias de proteína (ProClass).

```
COMMENT         Erythropoietin is produced by kidney or liver of adult
                mammals and by liver of fetal or neonatal mammals.
GENETICS
   #gene        GDB:EPO
     #cross-references GDB:119110; OMIM:133170
   #mmp_position 7q21.3-7q22.1
   #introns     5/1; 53/3; 82/3; 142/3
FUNCTION
   #description the primary inducer of erythrocyte formation
CLASSIFICATION  #superfamily erythropoietin
KEYWORDS        erythropoiesis; glycoprotein; hormone; kidney; liver
FEATURE
     1-27               #domain signal sequence #status predicted #label
                        SIG\
     28-193             #product erythropoietin #status experimental
                        #label MAT\
     34-188,56-60       #disulfide bonds #status experimental\
     51,65,110          #binding site carbohydrate (Asn) (covalent)
                        #status experimental\
     153                #binding site carbohydrate (Ser) (covalent)
                        #status experimental
```

BD de familia/dominio de Proteinas

- Bases de datos secundarias: resultantes del análisis de secuencias encontradas en las BD primarias
- Curadas manualmente (PROSITE, Pfam, etc.) o generada automáticamente (ProDom, DOMO)
- Algunos dependen del método usado para detectar si una proteína pertenece a una familia /dominio particular (patrones, perfiles, HMM)

Prosite

- Creada en 1988 (SIB)
- Contiene dominios funcionales anotados completamente, basado en dos métodos: patrones y perfiles
- Las entradas se depositan en PROSITE en dos archivos distintos:
  - Patrones y perfiles con la lista de todos los apareamientos en la versión de SWISS-PROT
  - Documentation

Prosite (pattern): ejemplo

```
ID   EPO_TPO; PATTERN.
AC   PS00817;
DT   OCT-1993 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE   Erythropoietin / thrombopoietin signature.
PA   P-x(4)-C-D-x-R-[LIVM](2)-x-[KR]-x(14)-C.
NR   /RELEASE=38,88008;
NR   /TOTAL=14(14); /POSITIVE=14(14); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0; /PARTIAL=1;
CC   /TAXO-RANGE=??E??; /MAX-REPEAT=1;
CC   /SITE=3,disulfide; /SITE=11,disulfide;
DR   P48617, EPO_BOVIN , T; P33707, EPO_CANFA , T; P33708, EPO_FELCA , T;
DR   P01588, EPO_HUMAN , T; P07865, EPO_MACFA , T; Q28513, EPO_MACMU , T;
DR   P07321, EPO_MOUSE , T; P49157, EPO_PIG   , T; P29676, EPO_RAT   , T;
DR   P33709, EPO_SHEEP , T; P42705, TPO_CANFA , T; P40225, TPO_HUMAN , T;
DR   P40226, TPO_MOUSE , T; P49745, TPO_RAT   , T;
DR   P42706, TPO_PIG   , P;
DO   PDOC00644;
//
```

Diagnostic performance

List of matches

Prosite (profile): ejemplo

```
PROSITE: PS50097

ID   BTB; MATRIX.
AC   PS50097;
DT   DEC-1999 (CREATED); DEC-1999 (DATA UPDATE); DEC-1999 (INFO UPDATE).
DE   BTB domain profile.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=67;
MA   /DISJOINT: DEFINITION=PROTECT; NI=6; NJ=62;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=.9751; R2=.02068202; TEXT='-LogE';
MA   /CUT_OFF: LEVEL=0; SCORE=363; N_SCORE=8.5; MODE=1; TEXT='';
MA   /CUT_OFF: LEVEL=-1; SCORE=287; N_SCORE=6.5; MODE=1; TEXT='';
MA   /DEFAULT: D=-20; I=-20; BI=-50; EI=-50; MI=-105; MD=-105; IM=-105; DM=-105; MM=1; MO=-2;
MA   /I: BI=0; BI=-105; BD=-105;
MA   /M: SY='C'; M=-6,-30,28,-14,-9,-15,-20,-14,-19,-15,-17,-14,-8,-19,-14,-15,0,0,-9,-32,-17,-12;
MA   /M: SY='D'; M=-16,41,-28,53,15,-34,-11,-1,-33,0,-27,-25,21,-11,0,-8,2,-6,-26,-38,-19,7;
MA   /M: SY='V'; M=-2,-23,-8,-28,-24,-1,-24,-25,16,-20,7,6,-20,-25,-23,-20,-10,-4,24,-23,-9,-24;
MA   /M: SY='T'; M=-2,-13,-18,-19,-13,-7,-24,-19,6,-8,-2,1,-11,-17,-11,-10,-1,10,10,-24,-6,-13;
MA   /M: SY='L'; M=-11,-30,-22,-33,-24,15,-32,-23,25,-29,35,17,-26,-27,-23,-22,-24,-9,16,-17,3,-24;
MA   /M: SY='V'; M=0,-11,-18,-13,-10,-12,-20,-13,1,-6,-4,2,-10,-19,-6,-7,-4,-2.8,-25,-9,-9;
MA   /M: SY='V'; M=1,-25,-3,-29,-25,-2,-26,-26,17,-22,10,7,-23,-25,-23,-22,-11,-3,24,-27,-10,-25;
MA   /M: SY='D'; M=-6.7,-26.8,7,-25.6,-7,-27.0,-23,-17.8,-13.0,-3.3,-6,-23,-27,-17.3;
MA   /I: I=-5; MI=0; IM=0; DM=-15; MD=-15;
MA   /M: SY='G'; M=-6.8,-27.8,-3,-27.22,-7,-30,-8,-26,-19.10,-14,-8,-9.2,-9,-24,-28,-21,-6;
MA   /M: SY='K'; M=-7,-4,-23,-4.7,-23,-13,-2,-21.10,-18,-9,-3,-12.7,9,-2,-4,-16,-25,-12,6;
MA   /M: SY='E'; M=-8,-6,-21,-8.1,-15,-21,-7,-7,-1,-10,-5,-3,-14.0,-1,-2,-2,-6,-26,-9,-1;
MA   /M: SY='F'; M=-12,-28,-22,-34,-26.31,-31,-21.18,-26.16.9,-22,-27,-27,-21,-20,-9.14,-6,13,-26;
MA   /M: SY='R'; M=13,-9,-24,-10,-3,-11,-21.7,-17.7,-16,-4,-4,8.2.9,-9,-9,-16,-20,-1,-2;
MA   /M: SY='A'; M=21,-15,-8,-22,-17,-10,-10,-23.0,-15,-5,-5,-14,-16,-17,-19.6.6,12,-24,-15,-17;
MA   /M: SY='H'; M=-15.5,-22.2,-1,-20,-16.65,-26,-8,-21,-5.15,-19.6,-2,-2,-11,-26,-32.7,0;
MA   /M: SY='K'; M=-12,-5,-29,-5.5,-25,-18,-8,-26.34,-24,-9,-1,-14.8.34,-8,-8,-17,-20,-10.5;
MA   /M: SY='A'; M=4,-12,-12,-16,-10,-6,-18,-14,-2,-13,-1,-2,-11,-17,-12,-13,-3.1,2,-24,-8,-11;
MA   /M: SY='V'; M=-7,-26,-19,-31,-26.7,-32,-24.27,-23.14.11,-22,-25,-23,-22,-13.0.28,-19.3,-26;
MA   /M: SY='L'; M=-10,-30,-20,-30,-21.9,-30,-20.22,-29.47.20,-29,-29,-20,-20,-29,-10.12,-20,0,-21;
MA   /M: SY='A'; M=18,-6,0,-12,-8,-18,-6,-16,-15,-10,-18,-12,-2,-14,-8,-13.18,11,-5,-32,-19,-8;
...
```

- Tabla de puntaje especifica de la posicion de los residuos.
- Cuales son conservados o degenerados
- Posiciones que toleran inserciones

ProDom

- Consiste en una compilación automatizada del alineamiento de dominios homólogos

PRINTS

- Compendio de fingerprints de motivos proteícos
- La mayoría de familias de proteína poseen varios motivos conservados
- Fingerprint: grupo de motivos (simple o compuesto, por ejemplo multidominios) = característica de los miembros de una familia
- Un miembro de la familia exhibe todos los elementos del fingerprint, mientras que los miembros de la subfamilia pueden poseer solamente una parte

BD de Familia/dominio de Proteinas:BD compuestas

Ejemplo: **InterPro**

- Reune a PROSITE, PRINTS, Pfam, ProDom y SMART en un recurso integrado de familias de proteínas protein families, dominios y sitios funcionales;
- Single set of «documents» linked to the various methods;
- Utilizado para mejorar la anotación funcional en SWISS-PROT (clasificación de proteínas desconocidas)

InterPro: ejemplo

IPR001323

Name
Erythropoietin/thrombopoeitin

Type
Family

Abstract
Erythropoietin, a plasma glycoprotein, is the primary physiological mediator of erythropoiesis [1] . It is involved in the regulation of the level of peripheral erythrocytes by stimulating the differentiation of erythroid progenitor cells, found in the spleen and bone marrow, into mature erythrocytes [2] . It is primarily produced in adult kidneys and foetal liver, acting by attachment to specific binding sites on erythroid progenitor cells, stimulating their differentiation [3] . Severe kidney dysfunction causes reduction in the plasma levels of erythropoietin, resulting in chronic anaemia - injection of purified erythropoietin into the blood stream can help to relieve this type of anaemia. Levels of erythropoietin in plasma fluctuate with varying oxygen tension of the blood, but androgens and prostaglandins also modulate the levels to some extent [3] . Erythropoietin glycoprotein sequences are well conserved, a consequence of which is that the hormones are cross-reactive among mammals, i.e. that from one species, say human, can stimulate erythropoiesis in other species, say mouse or rat [4] .

Thrombopoeitin (TPO), a glycoprotein, is the mammalian hormone which functions as a megakaryocytic lineage specific growth and differentiation factor affecting the proliferation and maturation from their committed progenitor cells acting at a late stage of megakaryocyte development. It acts as a circulating regulator of platelet numbers.

Examplelist
P33708
P33709
P49745
view matches for the examples

Publications
1. Shoemaker C.B., Mitsock L.D. 849-858 (1986)
2. Takeuchi M., Takasaki S., Miyazaki H., Kato T., Hoshi S., Kochibe N., Kobata A. J. Biol. Chem. 263: 3657-3663 (1988)
3. Lin F.K., Lin C.H., Lai P.H., Browne J.K., Egrie J.C., Smalling R., Fox G.M., Chen K.K., Castro M., Suggs S. Gene 44: 201-209 (1986)
4. Nagao M., Suga H., Okano M., Masuda S., Narita H., Ikura K., Sasaki R. Nucleotide sequence of rat erythropoietin. 1171: 99-102 (1992)

Children
IPR003013

Signatures
PROSITE PS00817 EPO_TPO
PFAM PF00758 EPO_TPO

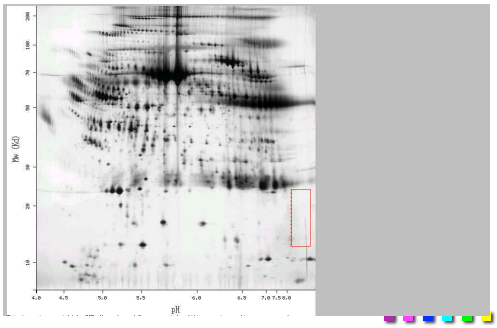
Matches
Table Graphical

BD de Proteomica

- Contain informations obtained by 2D-PAGE: master images of the gels and description of identified proteins
- Examples: SWISS-2DPAGE, ECO2DBASE, Maize-2DPAGE, Sub2D, Cyano2DBase, etc.
- Format: composed of image and text files
- Most 2D-PAGE databases are "federated" and use SWISS-PROT as a master index
- There is currently no protein Mass Spectrometry (MS) database (not for long…)

This protein does not exist in the current release of SWISS-2DPAGE.

EPO_HUMAN (human plasma)



BD 7: estructura 3D

- Contain the spatial coordinates of macromolecules whose 3D structure has been obtained by X-ray or NMR studies
- Proteins represent more than 90% of available structures (others are DNA, RNA, sugars, virus, complex protein/DNA…)
- PDB (Protein Data Bank), SCOP (structural classification of proteins (according to the secondary structures)), BMRB (BioMagResBank; RMN results)
- DSSP: Database of Secondary Structure Assignments.
  HSSP: Homology-derived secondary structure of proteins.
  FSSP: Fold Classification based on Structure-Structure Assignments.
- Future: Homology-derived 3D structure db.

PDB: Protein Data Bank

- Managed by Research Collaboratory for Structural Bioinformatics (RCSB) (USA).
- Contains macromolecular structure data on proteins, nucleic acids, protein-nucleic acid complexes, and viruses.
- Specialized programs allow the vizualisation of the corresponding 3D structure.
- Currently there are ~16'000 structure data for about 4'000 different molecules, but far less protein family (highly redundant) !

PDB: example

```
HEADER    LYASE(OXO-ACID)                    01-OCT-91  12CA     12CA  2
COMPND    CARBONIC ANHYDRASE /II (CARBONATE DEHYDRATASE) (/HCA II)   12CA   3
COMPND    2 (E.C.4.2.1.1) MUTANT WITH VAL 121 REPLACED BY ALA (/V121A) 12CA  4
SOURCE    HUMAN (HOMO SAPIENS) RECOMBINANT PROTEIN                   12CA   5
AUTHOR    S.K.NAIR,D.W.CHRISTIANSON                                  12CA   6
REVDAT    1  15-OCT-92 12CA   0                                      12CA   7
JRNL      AUTH   S.K.NAIR,T.L.CALDERONE,D.W.CHRISTIANSON,C.A.FIERKE  12CA   8
JRNL      TITL   ALTERING THE MOUTH OF A HYDROPHOBIC POCKET.         12CA   9
JRNL      TITL 2 STRUCTURE AND KINETICS OF HUMAN CARBONIC ANHYDRASE  12CA  10
JRNL      TITL 3 /II$ MUTANTS AT RESIDUE VAL-121                     12CA  11
JRNL      REF   J.BIOL.CHEM.                V. 266 17320 1991      12CA  12
JRNL      REFN   ASTM JBCHA3  US ISSN 0021-9258                071 12CA  13
REMARK  1                                                        12CA  14
REMARK  2                                                        12CA  15
REMARK  2 RESOLUTION. 2.4  ANGSTROMS.                            12CA  16
REMARK  3                                                        12CA  17
REMARK  3 REFINEMENT.                                            12CA  18
REMARK  3  PROGRAM              PROLSQ                           12CA  19
REMARK  3  AUTHORS              HENDRICKSON,KONNERT              12CA  20
REMARK  3  R VALUE                0.170                          12CA  21
REMARK  3  RMS.D. BOND DISTANCES     0.011  ANGSTROMS           12CA  22
REMARK  3  RMS.D BOND ANGLES         1.3    DEGREES             12CA  23
REMARK  4                                                        12CA  24
REMARK  4 N-TERMINAL RESIDUES SER 2, HIS 3, HIS 4 AND C-TERMINAL    12CA  25
REMARK  4 RESIDUE LYS 260 WERE NOT LOCATED IN THE DENSITY MAPS AND,  12CA  26
REMARK  4 THEREFORE, NO COORDINATES ARE INCLUDED FOR THESE RESIDUES.  12CA  27
------
```

PDB (cont.)

```
SHEET   3  S10 PHE   66  PHE    70 -1  O  ASN    67   N  LEU    60   12CA  68
SHEET   4  S10 TYR   88  TRP    97 -1  O  PHE    93   N  VAL    68   12CA  69
SHEET   5  S10 ALA  116  ASN   124 -1  O  HIS   119   N  HIS    94   12CA  70
SHEET   6  S10 LEU  141  VAL   150 -1  O  LEU   144   N  LEU   120   12CA  71
SHEET   7  S10 VAL  207  LEU   212  1  O  ILE   210   N  GLY   145   12CA  72
SHEET   8  S10 TYR  191  GLY   196 -1  O  TRP   192   N  VAL   211   12CA  73
SHEET   9  S10 LYS  257  ALA   258 -1  O  LYS   257   N  THR   193   12CA  74
SHEET  10  S10 LYS   39  TYR    40  1  O  LYS    39   N  ALA   258   12CA  75
TURN    1  T1 GLN   28  VAL    31     TYPE VIB (CIS-PRO 30)          12CA  76
TURN    2  T2 GLY   81  LEU    84     TYPE IIIPRIME) (GLY 82)        12CA  77
TURN    3  T3 ALA  134  GLN   137     TYPE I (GLN 136)               12CA  78
TURN    4  T4 GLN  137  GLY   140     TYPE I (ASP 139)               12CA  79
TURN    5  T5 THR  200  LEU   203     TYPE VIA (CIS-PRO 202)         12CA  80
TURN    6  T6 GLY  233  GLU   236     TYPE III (GLY 235)             12CA  81
CRYST1  42.700  41.700  73.000 90.00 104.60  90.00 P 21       2  12CA  82
ORIGX1   1.000000  0.000000  0.000000        0.00000                12CA  83
ORIGX2   0.000000  1.000000  0.000000        0.00000                12CA  84
ORIGX3   0.000000  0.000000  1.000000        0.00000                12CA  85
SCALE1   0.023419  0.000000  0.006100        0.00000                12CA  86
SCALE2   0.000000  0.023981  0.000000        0.00000                12CA  87
SCALE3   0.000000  0.000000  0.014156        0.00000                12CA  88
ATOM    1  N   TRP    5      8.559  -0.751  10.738  1.00 13.37     12CA  89
ATOM    2  CA  TRP    5      7.743  -1.668  11.585  1.00 13.42     12CA  90
ATOM    3  C   TRP    5      6.786  -2.502  10.667  1.00 13.47     12CA  91
ATOM    4  O   TRP    5      6.422  -2.085   9.607  1.00 13.57     12CA  92
ATOM    5  CB  TRP    5      6.997  -0.917  12.645  1.00 13.34     12CA  93
ATOM    6  CG  TRP    5      5.784  -0.209  12.221  1.00 13.40     12CA  94
ATOM    7  CD1 TRP    5      5.681   1.084  11.797  1.00 13.29     12CA  95
ATOM    8  CD2 TRP    5      4.417  -0.667  12.221  1.00 13.34     12CA  96
ATOM    9  NE1 TRP    5      4.388   1.418  11.515  1.00 13.30     12CA  97
ATOM   10  CE2 TRP    5      3.588   0.375  11.797  1.00 13.35     12CA  98
ATOM   11  CE3 TRP    5      3.837  -1.877  12.645  1.00 13.39     12CA  99
ATOM   12  CZ2 TRP    5      2.216   0.208  11.656  1.00 13.39     12CA 100
ATOM   13  CZ3 TRP    5      2.465  -2.043  12.504  1.00 13.33     12CA 101
ATOM   14  CH2 TRP    5      1.654  -1.001  12.009  1.00 13.34     12CA 102
-----
```

BD 8: Metabólicas

- Contain informations that describe enzymes, biochemical reactions and metabolic pathways;
- ENZYME and BRENDA: nomenclature databases that store informations on enzyme names and reactions;
- Metabolic databases: EcoCyc (specialized on Escherichia coli), KEGG, EMP/WIT;
  Usualy these databases are tightly coupled with query software that allows the user to visualise reaction schemes.

BD 9: Bibliográfica

- Bibliographic reference databases contain citations and abstract informations of published life science articles;
- Example: Medline
- Other more specialized databases also exist (example: Agricola).

Medline

- MEDLINE covers the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences
- more than 4,000 biomedical journals published in the United States and 70 other countries
- Contains over 10 million citations since 1966 until now
- Contains links to biological db and to some journals
- New records are added to PreMEDLINE daily!
  - Many papers not dealing with human are not in Medline !
  - Before 1970, keeps only the first 10 authors !
  - Not all journals have citations since 1966 !

Medline/Pubmed

- the National Center for Biotechnology Information (NCBI)
- PubMed provides access to bibliographic information such as MEDLINE, PreMEDLINE, HealthSTAR, and to integrated molecular biology databases (composite db)
- PMID: 10923642 (PubMed ID), UI: 20378145 (Medline ID)

Otras DB

- Hay muchas bases de datos que no pueden ser clasificadas;
- Ejemplos: ReBase (enzimas restricción), TRANSFAC (factores de transcripción), CarbBank, GlycoSuiteDB (Azucares ligados), Proteína-proteína interacción db (DIR, ProNet, Interact), Proteasa db (MEROPS), patentes en biotecnología db, etc.;
- Como también otros aspectos de las aspectos de macromoléculas y biología molecular.

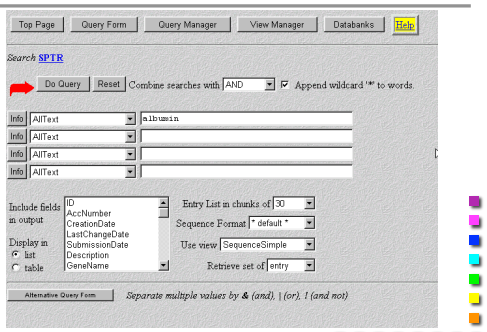
Proliferación de DB

- Cual es la mejor DB para análisis de secuencia?
- Cual tiene la mejor calidad de datos ?
- Cual es la más completa ?
- Cual es la más actualizada ?
- Cual es la menos redundante ?
- Cual es la más indexada (permite búsquedas complejas) ?
- Cual es la que responde más rápido ?
- …….??????

Herramientas de búsqueda en DB

- **Sequence Retrieval System** (SRS, Europe) permite a cualquier db flat-file ser indexada con cualquier otra DB; permite formular consultas a través de un rango amplio de DBs usando una interfase simple, sin importar la estructura de los datos y los lenguajes de consulta…
- **Entrez** (USA): menos flexible que SRS pero explota el concepto de « neighbouring », el cual permite relacionar artículos de diferentes db para ser unidos a otros, sean o no referencias cruzadas especificas.
- **ATLAS**: Especifica para db de secuencias de macromoléculas (i.e. NRL-3D)
- ….

SRS



Entrez-protein

- NCBI: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
- Reune muchos recursos, incluyendo SWISS-PROT, PIR, PRF, PDB, y traducciones de regiones codificantes anotadasen el GenBank (« Genpept ») and RefSeq.
  - PRF: Protein Research Foundation (Japan): DB Peptido/Secuencias de proteínas (PRF/SEQDB)
  - PDB: Protein Data Bank (3D structure)
  - RefSeq: NCBI Reference Sequence project
  - PIR - International Protein Sequence Database
- Proteínas y DNA secuencias