Bases de Datos (BD) Biológicas

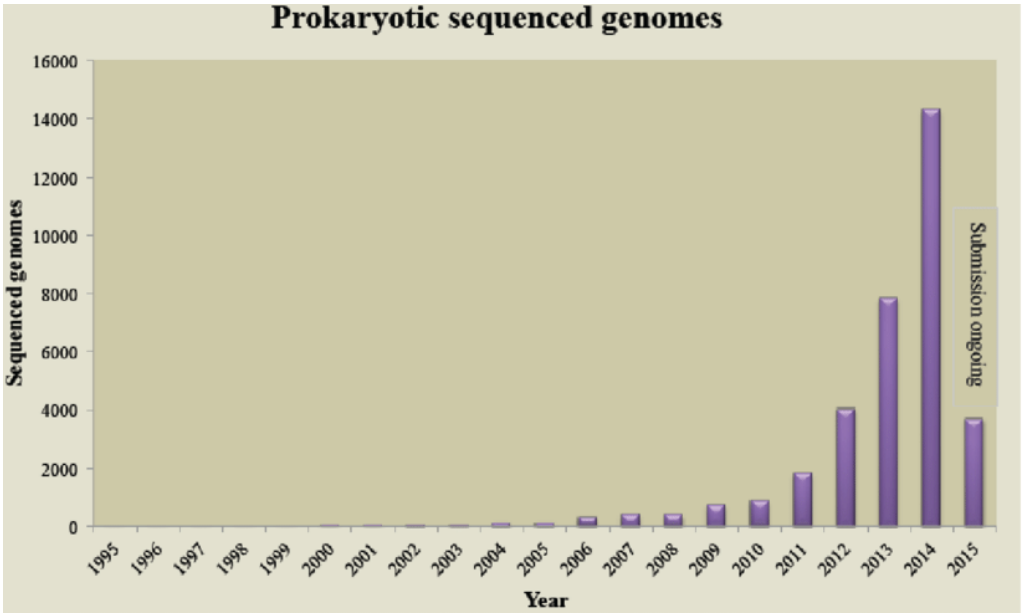
Luis Garreta

Electiva de Bioinformática
MAESTRÍA EN INFORMÁTICA BIOMÉDICA
Universidad el Bosque
Bogotá-Colombia

29 de marzo de 2023

Avances en biología molecular y tecnología de secuenciación

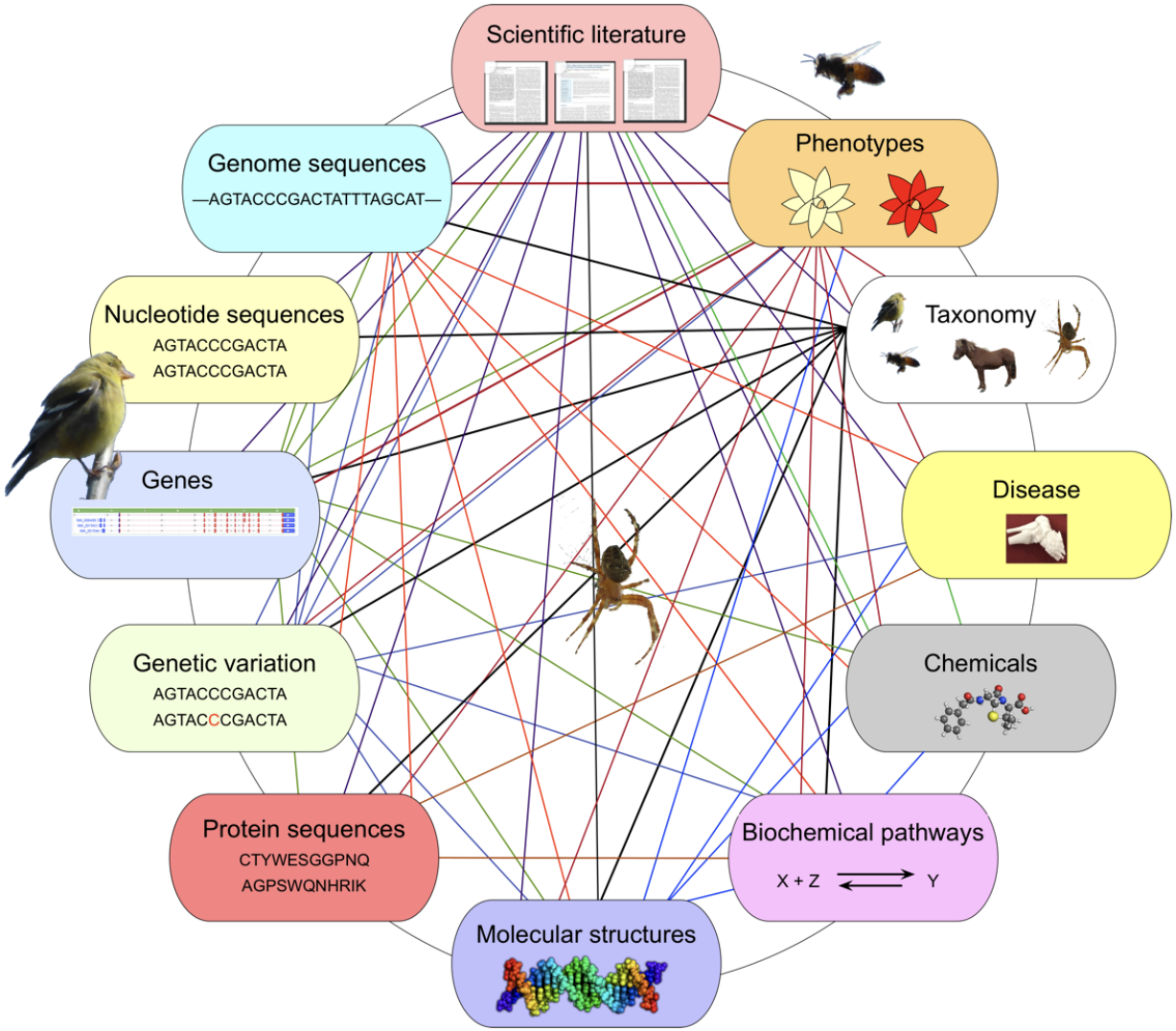
Actualmente, se están descubriendo enormes cantidades de información biológica, como conjuntos de datos de secuenciación sin procesar, proteomas, etc., a un ritmo muy rápido. .



Todo esto debido al rápido avance de la biología molecular, la proteómica y las tecnologías de secuenciación del genoma de bajo costo y alto rendimiento,

Bases de datos biológicas

Los conjuntos de datos relevantes para las ciencias biológicas, como la biología molecular y la bioinformática, se denominan bases de datos biológicas.



¿Por qué son importantes los BD biológicos??

Las BD biológicas tienen un propósito crítico en la recopilación y organización de datos relacionados con los sistemas biológicos.

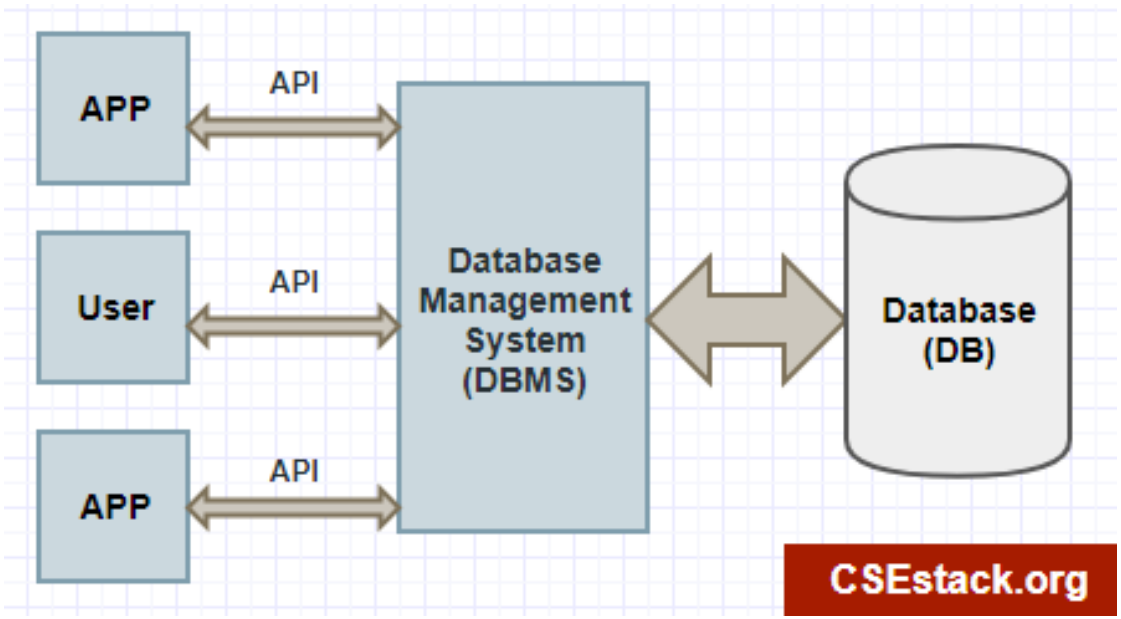
Brindan soporte computacional y una interfaz fácil de usar para análizar los datos biológicos (e.g.secuencias de genes y proteínas, estructuras moleculares, etc).

Otros propósitos:

- Ayudan a comprender el mecanismo molecular de las enfermedades (humanos, plantas, o animales)
- Ayuda al desarrollo de la medicina personalizada para recetar el fármaco más adecuado.
- Ayuda en el diseño y desarrollo de fármacos, mejora de cultivos y mejora de la calidad nutricional.

Que es una base de datos (BD)?

Una colección organizada de datos o información que se almacena electrónicamente y es accesible desde un sistema informático.



La naturaleza organizada de la BD facilita el acceso, la gestión, la actualización periódica y la búsqueda rápida de los datos/información necesarios desde un sistema informático adecuado.

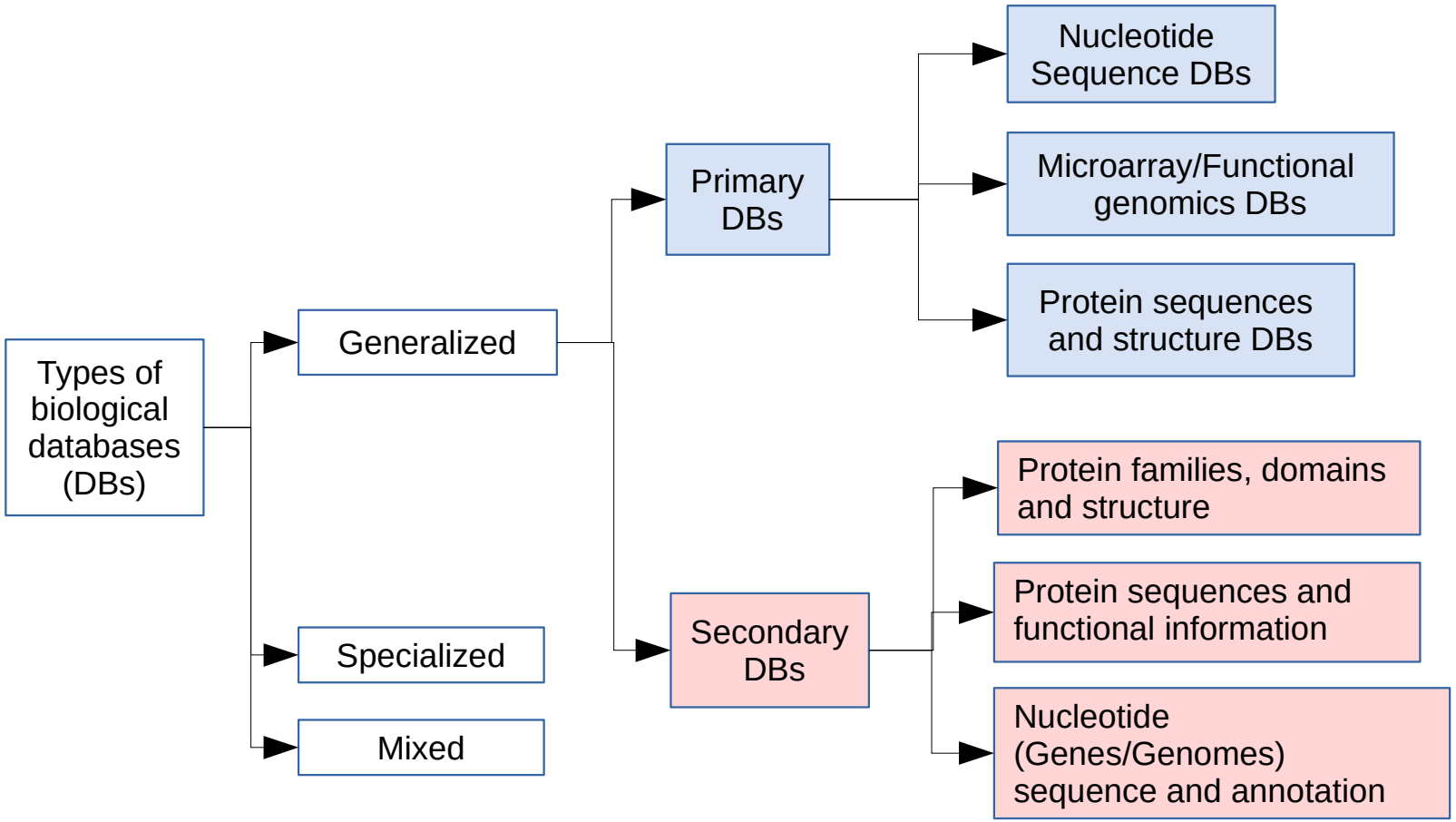
Elementos generales de una base de datos

1. Entidad: una entidad se refiere a lo que queremos almacenar en una base de datos. P.ej. Secuencias de ADN, Genes, Referencias bibliográficas, etc.
2. Registros: un registro típico se refiere a una combinación de todos los campos de una entidad determinada. Por ej. Registro del gen BRCA1 en GenBank.
3. Identificador: el nombre único que identifica un registro. En el caso de una base de datos simple, un solo archivo contiene múltiples registros. Entre estos
4. Campos: las propiedades de una entidad se denominan campos. P. ej. Nombre del gen, secuencia del gen, mutación (si la hay), etc.

Ejemplo de una base de datos tabular

DNI	Nombre	Apellidos	Sexo	Edad
309854	Pepe	Solano	V	32
205487	Lola	Sanchez	F	60
602548	Rosa	Perez	F	21
652314	Antonio	Cardenas	V	18
506981	Carlos	Lozano	V	51
559850	Fermin	Martin	V	54
203261	Laura	Mendez	F	69
363636	Carlos	Sanchez	V	22
696969	Felisa	Carmona	F	88

Tipos de bases de datos biológicas



Bases de datos primarias

Contienen datos derivados experimentalmente como:

- Secuencias de nucleótidos
- Secuencias de proteínas
- Información estructural de macromoléculas.

Esta información básica puede ir acompañada de:

- Anotación funcional,
- Bibliografías, y
- Enlaces a otras bases de datos.

Los datos son enviados directamente por los investigadores.

A cada dato envíado se les asigna un número de acceso

- Este número es permanente y pasa a formar parte del registro científico.

Bases de datos secundarias

Almacenan la información derivada del análisis de conjuntos de datos primarios

Contienen información altamente seleccionada derivada del análisis de recursos primarios y literatura, usando:

- Análisis computacional complejo
- Análisis manual

Estas bases de datos suelen almacenar información sobre :

- Estructura/secuencias de dominio conservadas
- Secuencias de señales
- Residuos del sitios activo

Aspectos esenciales de las bases de datos primarias y secundarias

	Base de datos principal	Base de datos secundaria
Sinónimos	BD de archivo	BD curada; base de conocimientos
Fuente de datos	Envío directo de datos derivados experimentalmente de los investigadores	Resultados de análisis, investigación e interpretación de la literatura, a menudo de datos en las base de datos primarias
Ejemplos	ENA, GenBank and DDBJ (secuencia de nucleótidos) ArrayExpress and GEO (datos de genómica funcional) Protein Data Bank (PDB; ccoordenadas de estructuras macromoleculares tridimensionales)	InterPro (familias de proteínas, motivos y dominios) UniProt Knowledgebase (secuencia e información funcional sobre proteínas) Ensembl (variación, función, regulación y más en capas en secuencias del genoma completo)

Otra clasificación: por contenido / tipo de datos

	Integrales	Especializadas
Fuente de datos	Contienen datos de muchos organismos y muchos tipos diferentes de secuencias	Contienen datos de organismos individuales, categorías/funciones específicas de secuencias o datos generados por tecnologías de secuenciación específicas.
Ejemplos	Nucleotidos: • GenBank (USA), • EMBL (EUROPA), • DDBJ (Japón)	Específico del organismo: • Human Genome Sequencing • GDB: Genome Database (human mapping information) • MGD: Mouse Genome Database • SGD: Saccharomyces Genome Database
	Proteínas: • Entrez Protein, • Swiss-Prot, • UniProt	Categorías de secuencia o funciones: • TRANSFAC: Transcription Factors • Vector Database
	Estructuras de la proteínas: • PDB: Protein Data Bank, • MMDB: Molecular Modeling Database	Datos generados por tecnologías de secuenciación específicas: • EST: Expressed Sequence Tags • GSS: Genome Survey Sequences • STS: Sequence Tagged Sites • HTG: High Throughput Sequences
	Genomas y mapas: • Entrez Genome	

Pregunta 01: Cuál es la secuencia de nucleótidos completa del gen Bcl-2 de humanos?

Se cree ampliamente que Bcl-2 es un gen supresor de la apoptosis. La sobreexpresión de la proteína en las células cancerosas puede bloquear o retrasar el inicio de la apoptosis.

- Acceda al NCBI o GenBank https://www.ncbi.nlm.nih.gov/genbank/
- Seleccionar la BD de nucleótidos
- Buscar el término: "gen bcl-2 homo sapiens"
- Seleccionar la entrada de interés.
- Descargue la secuencia en formato FASTA.
- Descargue la proteína(s) (secuencia de aminoácidos) que expresa este gen.

Pregunta 02: Cuál es la estructura Tridimensional de una proteína de la familia Bcl-2 ?

Un surco hidrofóbico prominente está presente en la superficie de las proteínas antiapoptóticas. Este surco es el sitio de unión de péptidos que imitan la región BH3 de varias proteínas proapoptóticas como Bak y Bad.

- Acceda al Protein Data Bank o PDB: https://www.rcsb.org/
- Realize la búsqueda ingresando la secuencia de aminoácidos de la proteína.
- Seleccione una entrada.
- Descargue el archivo de la estructura.
- Visualizela en un visor de estructuras de proteínas (e.g. VMD o rasmol).

Pregunta 03: Qué literatura científica hay sobre la familia de proteínas Bcl-2 ?

Las proteínas Bcl-2 forman una familia de proteínas relacionadas evolutivamente que son esenciales para mantener el equilibrio entre la muerte celular y la proliferación celular, con diferentes miembros de la familia actuando como moduladores positivos o negativos de la apoptosis.

- Acceda ra la página de inicio del NCBI (http://www.ncbi.nlm.nih.gov/).
- Seleccione la BD de literatura científica Pubmed.
- Busqué el término "bcl-2 protein family"
- Descargue un artículo de interés.

Formato FASTA

https://www.ncbi.nlm.nih.gov/genbank/fastaformat/

El formato FASTA es un formato basado en texto para representar secuencias de nucleótidos o secuencias de péptidos mediante códigos de una sola letra.

Una secuencia en formato FASTA comienza con una descripción de una sola línea, seguida de líneas de datos de secuencia.

La línea de descripción se distingue de los datos de secuencia por un símbolo mayor que (">") en la primera columna. Se recomienda que todas las líneas de texto tengan menos de 80 caracteres de longitud.

Ejemplo Formato FASTA

Nucleótidos:

```
>XR_002086427.1 Candida albicans SC5314 uncharacterized ncRNA (SCR1), ncRNA
TGGCTGTGATGGCTTTTAGCGGAAGCGCGCTGTTCGCGTACCTGCTGTTTGTTGAAAATTTAAGAGCAAA
GTGTCCGGCTCGATCCCTGCGAATTGAATTCTGACGCTAGAGTAATCAGTGTCTTTCAAGTTCTGGTAAT
GTTTAGCATAACCACTGGAGGGAAGCAATTCAGCACAGTAATGCTAATCGTGGTGGAGGCGAATCCGGTG
GCACCTTGTTTGTTGATAAATAGTGCGGTATCTAGTGTTGCAACTCTATTTTT
```

Aminoácidos:

```
>gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase
MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGKNQEEVVTTSYAFQTAKLRQIRA
AHVQGGNSLQVLNFVIFPHLNYDLPFFGADLVTLPGGHLIALDMQPLFRDDSAYQAKYTEPILPIFHAHQ
QHLSWGGDFPEEAQPFFSPAFLWTRPQETAVVETQVFAAFKDYLKAYLDFVEQAEAVTDSQNLVAIKQAQ
LRYLRYRAEKDPARGMFKRFYGAEWTEEYIHGFLFDLERKLTVVK
```

Formato Genbank

https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

- Formato más completo que presenta tres secciones: Definición, Características y Secuencia.
- En definición podemos ver información acerca de la longitud, número de acceso, anotación y referencias bibliográficas donde aparece.
- En características tenemos información sobre la secuencia codificante, secuencia de aminoácidos y otras características (como dominios proteicos, exónes e intrónes).
- En la de secuencia aparecen las coordenadas iniciales de cada línea seguida por la secuencia nucleotídica.
- El formato acaba en //.

Ejemplo Formato Genbank

https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

```
LOCUS       PSEAVRB                2271 bp    DNA     linear   BCT 26-APR-1993
DEFINITION  P.syringae avirulence protein (avrB) gene, complete cds.
ACCESSION   M21965
VERSION     M21965.1  GI:151050
KEYWORDS    avirulence protein.
SOURCE      Pseudomonas syringae
  ORGANISM  Pseudomonas syringae
            Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales;
            Pseudomonadaceae; Pseudomonas.
REFERENCE   1  (bases 1 to 2271)
  AUTHORS   Tamaki,S., Dahlbeck,D., Staskawicz,B. and Keen,N.T.
  TITLE     Characterization and expression of two avirulence genes cloned from
            Pseudomonas syringae pv. glycinea
  JOURNAL   J. Bacteriol. 170 (10), 4846-4854 (1988)
   PUBMED   3049552
COMMENT     Original source text: Pseudomonas syringae (pv. glycinea, strain
            race 4) DNA, clone pPSG0002.
            Draft entry and computer-readable sequence for [1] kindly provided
            by N.Keen, 12-JAN-1989.
FEATURES             Location/Qualifiers
     source          1..2271
                     /organism="Pseudomonas syringae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:317"
     CDS             477..1442
                     /note="avirulence protein avrB"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="AAA25726.1"
                     /db_xref="GI:151051"
                     /translation="MGCVSSKSTTVLSPQTSFNEASRTSFRALPGPSQRQLEVYDQCL
                     VYRQTAYLHGNVIPLNTFRVATDTEYLRDRVAHLRTELGAKALKQHLQRYNPDRIDHT
                     NASYLPIIKDHLNDLYRQAISSDLSQAELISLIARTHWWAASAMPDQRGSAAKAEFAA
                     RAIASAHGIELPPFRNGNVSDIEAMLSGEEEFVEKYRSLLDSDCF"
ORIGIN      5 bp upstream of PstI site.
        1 ctgcagctgt tgcacaggta tttgacgtgc gggcagctct ggttgccgca ggtgtagtgt
       61 ttgaccgccc aggatcgagg tgccgcaggc cagcattttg gtgacggact ccacttcgat
      121 gtcgcgtaag ccgccctcct caaggctacg cgatcgtaag aacgggcgat ttgcagccgc
      181 taaatgccac acagctcaag caaactacac agcacaacat attagcgttt atgtggtggt
      241 ttaacatact taagtgtgtt ggcatttaat gtacagccaa aacgaggtaa ttattcatgg
      301 gctgcgtctc gtcaaaaagc accacagtgc tttctccaca gacatctttt aatgaagcct
      361 cccgtacgtc tttcagagca ctcccggcc catcgcaaag acaattggag gtctatgatc
      421 acagggcata ttgtcagagc atgtacaact caattcgctc tgctggagat gaaatttcca
```

Header

Features

Sequence
(complete sequence not shown)

Formato PDB Estructuras de Proteínas

https://www.rcsb.org/docs/general-help/structures-without-legacy-pdb-format-files

- El formato PDB consta de líneas de información en un archivo de texto.
- Cada línea de información en el archivo se llama registro.
- Un archivo PDB generalmente contiene varios tipos diferentes de registros, dispuestos en un orden específico para describir una estructura.
  - Los registros HEADER, TITLE y AUTHOR.
  - Los registros REMARK
  - Los registros SEQRES
  - Los registros ATOM
  - Los registros HETATM

Ejemplo de Archivo en Formato PDB

```
HEADER     EXTRACELLULAR MATRIX                         22-JAN-98   1A3I
TITLE      X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA     X-RAY DIFFRACTION
AUTHOR     R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR    2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
```