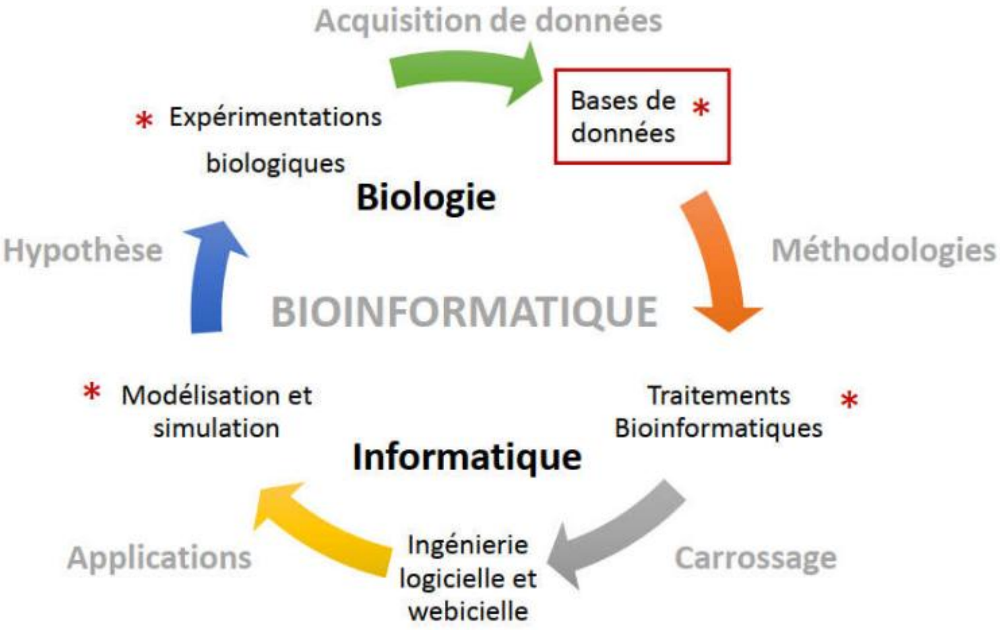
Course 03: Storage of biological bioinformation (databases and data banks)

1. Introduction and definitions:

Recent molecular biology techniques generate a massive amount of data that cannot be managed by traditional publishing methods. In this context, repositories and databases are now a major source of information for the scientific community.


Acquisition de données
* Expérimentations biologiques
Bases de données *
Biologie
Hypothèse
Méthodologies
BIOINFORMATIQUE
* Modélisation et simulation
Traitements Bioinformatiques *
Informatique
Applications
Ingénierie logicielle et webicielle
Carrossage

Often the terms "bank" and "base" are used interchangeably. However, there is a difference not only for the user but also for their IT implementation:

- Data banks:

A dataset relating to a defined area of knowledge and organized to be made available for user consultation.

- Database:

A set of data organized for use by programs corresponding to distinct applications and in such a way as to facilitate the independent evolution of the data and programs.

There are numerous databases of biological interest. This introduction will be limited to a presentation of the main public databases, based on the primary structure of sequences. We will distinguish between two types: generalist and specialized.

These databases can contain information such as: (DNA, proteins, genes and genomes, taxonomy, etc.). They also include a bibliography and biological expertise directly related to the sequences processed.

-Difference between data banks and databases

It should be noted that a data banks is a database (because it is a structured table) but which contains heterogeneous biological information (viruses, bacteria, fungi, plants, animals) whereas a database is more specialized (a database specific to E. coli, to Bacillus, etc.) .

-Role of banks/databases

-Collect information (sequences, physical mapping, genetics, structural and relational data, etc., from : biologists, literature, other databases)

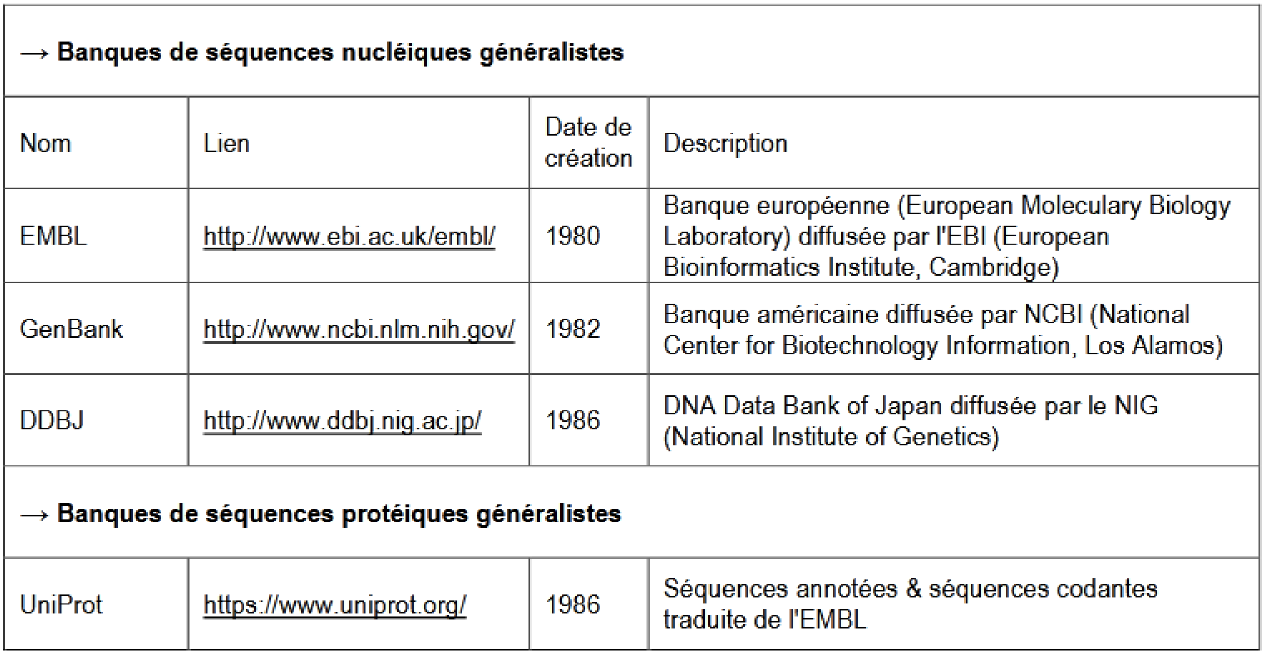
-Store and organize

-Distribute information

-Facilitate exploitation

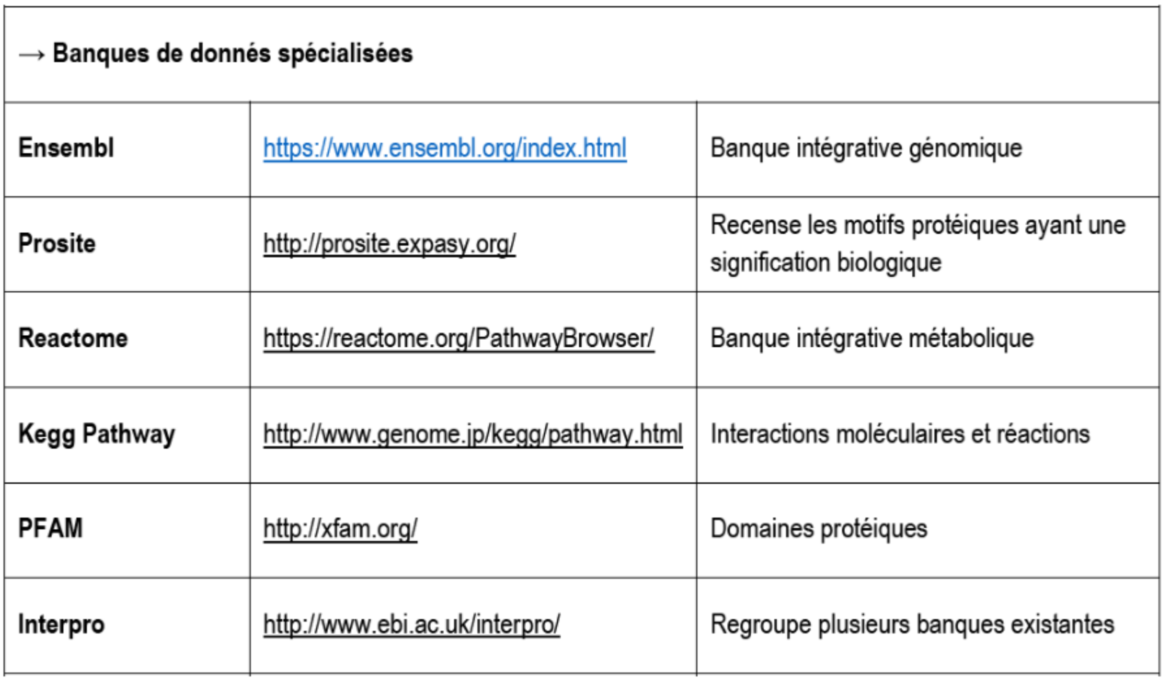
2. Types of data banks

We will distinguish two types of banks, those which correspond to the most exhaustive possible collection of data and which ultimately offer a rather heterogeneous set of information (**generalist data banks**) and those which correspond to more homogeneous data established around a theme (**specialized data banks**) and which offer added value from a particular technique or an interest aroused by a group of scientists.



→ Banques de séquences nucléiques généralistes			
Nom	Lien	Date de création	Description
EMBL	http://www.ebi.ac.uk/embl/	1980	Banque européenne (European Moleculary Biology Laboratory) diffusée par l'EBI (European Bioinformatics Institute, Cambridge)
GenBank	http://www.ncbi.nlm.nih.gov/	1982	Banque américaine diffusée par NCBI (National Center for Biotechnology Information, Los Alamos)
DDBJ	http://www.ddbj.nig.ac.jp/	1986	DNA Data Bank of Japan diffusée par le NIG (National Institute of Genetics)
→ Banques de séquences protéiques généralistes			
UniProt	https://www.uniprot.org/	1986	Séquences annotées & séquences codantes traduite de l'EMBL

Table presenting some general databases



→ **Banques de donnés spécialisées**		
Ensembl	https://www.ensembl.org/index.html	Banque intégrative génomique
Prosite	http://prosite.expasy.org/	Recense les motifs protéiques ayant une signification biologique
Reactome	https://reactome.org/PathwayBrowser/	Banque intégrative métabolique
Kegg Pathway	http://www.genome.jp/kegg/pathway.html	Interactions moléculaires et réactions
PFAM	http://xfam.org/	Domaines protéiques
Interpro	http://www.ebi.ac.uk/interpro/	Regroupe plusieurs banques existantes

Table showing some specialized databases

- **General data banks**

• These databases contain heterogeneous data:

- Collection as comprehensive as possible

-Enormous wealth of sequences in a single set;

-Great diversity of organisms;

-A lot of information accompanies the sequences

- Nucleus sequence databases

- Protein sequence databases

• **Advantage:** everything can be viewed at once

• **Disadvantages:** difficult to maintain, difficult to query

- **Specialized data banks**

• These databases contain homogeneous data

• Collection organized around a specific theme

•Their purpose is:

- to identify families of sequences around specific biological characteristics such as regulatory signals, gene promoters, peptide signatures or identical genes from different species.

- to group specific classes of sequences such as cloning vectors, restriction enzymes, or all sequences of the same genome.

• For specific needs related to the activity of a group of people, or for bibliographic compilations

- **Advantages:** ease of updating data, verifying its integrity, offering a user-friendly interface,…
- **Disadvantages:** doesn't always target what you want; not all possible banks exist

2.1 General-purpose sequence banks:

2.1.1 Nucleotide Sequence data banks

The data stored in these types of data banks comes from DNA and RNA sequencing. Three well-known nucleic acid databases share information and therefore contain nearly identical sets of sequences. These three data banks have systematically exchanged their contents since 1987 and have adopted a common system of conventions: "DDBJ/EMBL/ GenBank" »:

- **The bank EMBL:** created in 1980 and funded by EMBO (European Molecularly Biology Organization), developed within the European Laboratory of Molecular Biology located at Heidelberg (Germany), it is now distributed by the EBI: http://www.ebi.ac.uk/embl/. As of February 24, 2014, the bank contains 369.5 million sequences.

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data

- **GenBank (Genetic) Sequence Databank):** created in 1982 by the company IntelliGenetics and now distributed by the NCBI (National Center for Biotechnology Information): http://www.ncbi.nlm.nih.gov/. As of February 2014, the database contained 171,123,749 sequences. GenBank contains a sub-database of proteins, translated from nucleic acid sequences, called GenPept .

GenBank Overview

What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2011 Jan 39(Database issue):D32-7). There are approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional

- **The DDBJ (DNA Databank of Japan):** created in 1986 and distributed by the NIG (National Institute of Genetics , Japan), recorded a total of 81,994,905 DNA sequences in December 2019 (DDBJ 2019).

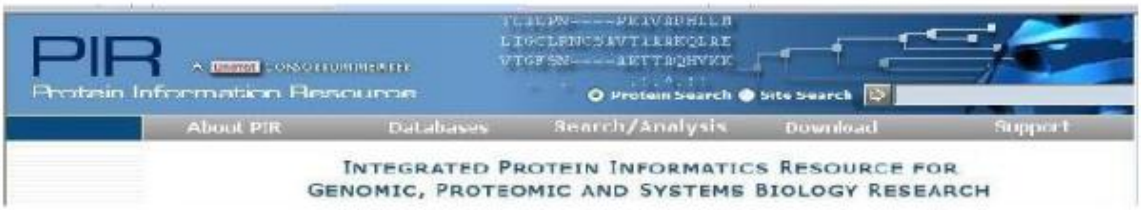
http://www.ddbj.nig.ac.jp/searches-e.html
DDBJ
DNA Data Bank of Japan
Accession DNA Protein Taxonomy Udn-search
Accession numbers
○ DDBJ ○ UniProt ○ PDB ○ DAD ○ PRF ○ Patent >> more
HOME
Submission
How to Use
Search/Analysis
FTP/WebAPI
Report/Statistics
Contact Us
Japanese
HOME > Search and Analysis
Notice
Termination of a part of DDBJ services (2008.8.15)

2.1.2. Protein banks

The data stored in these databases comes from the translation of DNA sequences or from protein sequencing (rare because it is long and expensive):

• The **SwissProt bank :** is a protein bank created in 1986 at the University of Geneva and maintained since 1987 as part of a collaboration between this university (via ExPASy , Expert Protein) and other institutions. Analysis System) and the EBI. It also includes annotated sequences from the PIR-NBRF database as well as translated coding sequences from the EMBL. As of February 2014, the database contained 542,503 sequences compressing 192,888,369 amino acids.

• **PIR-NBRF (Protein Information Resource-National Biomedical Research Foundation)** created in 1984 by the NBRF (National Biomedical Research Foundation). It is now a dataset from MIPS (Martinsried Institute for Protein Sequences , Munich, Germany) and the Japanese bank JIPID (Japan International Protein Information Database);


PIR
Protein Information Resource
A UniProt CONSORTIUM MEMBER
○ Protein Search ● Site Search
About PIR
Databases
Search/Analysis
Download
Support
INTEGRATED PROTEIN INFORMATICS RESOURCE FOR
GENOMIC, PROTEOMIC AND SYSTEMS BIOLOGY RESEARCH

2.2 Specialized Bases

• **The Protein Data Bank (PDB),** created in 1971, is the reference database of protein structures obtained experimentally by X-ray crystallography, NMR spectroscopy, and cryo -electron microscopy (the most recently used technique). The coordinates of the atoms forming a protein's structure, the sequence details, and the crystallization conditions are the main pieces of information available for each structure in the database. Structural homologs are detected from this database.

EPSON Web-To-Page Print Print Preview
RCSB PDB PROTEIN DATA BANK
A MEMBER OF THE PDB
An Information Portal to Biological Macromolecular Structures
PDB Statistics
CONTACT US | HELP | PRINT PAGE
Site Search
Advanced Search
Home Search
News
Complete News
Newsletter
Discussion Forum

Are you missing data updates? The PDB archive has moved to ftp://ftp.wwpdb.org/.
For more information click here.

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

ECD: based on the nucleic acid sequences of Escherichia coli.

NRL3D: database of protein sequences whose three-dimensional structure has been determined.

TFD: transcription factor base.

Prosite : databases of protein motifs. It can be considered a dictionary that lists protein motifs with biological significance.

CATH: based on hierarchical (ordered) classifications of protein structures. IMGT: database of immunoglobulin and T-cell receptor sequences.

GENATLAS: database of information derived from the mapping of human genes.

KEGG: basics of metabolic pathways.

The basics of patterns:

We know that certain DNA or protein segments are crucial in sequence analysis because they correspond to specific sites of biological activity, such as gene regulatory elements or peptide signatures. This is why specialized databases have naturally developed around these sequences. The use of specialized databases, such as motif databases, has become an essential tool in sequence analysis for attempting to determine the function of unknown proteins or to identify the family to which a previously uncharacterized sequence belongs.

- **TFD or IMD:** are used for gene promoter sequences
- **Prosite or BLOCKS:** are used for unknown proteins or protein sequences translated from cDNA or genomic sequences.

To detect a feature on a sequence, it is enough to run a program which will try to identify the presence of certain patterns listed in these databases and thus predict whether the tested sequence belongs to a group of sequences having a common signature.

Immunological banks

They specialize in the following information:

- Sequences

- Receptor (T cell, for example)

- Complex MHC (Major Histocompatibility Complex)

- HLA system

2D or 3D Structured Banks

They specialize in the following information:

- 3D coordinates of proteins *

- Secondary structure of proteins

- Structural domains

- Active center of enzymes

- Receptor-ligand complexes

- Atlas of structural topology of proteins

2.3 Bibliographic Databases

Bibliographic databases list all categories of bibliographic objects: books, scientific journals, articles…

For example, PubMed is a bibliographic database in the biological and biomedical sciences, with coverage beginning in 1946 and containing over 30 million references. In addition to articles indexed in MEDLINE, PubMed also contains supplementary references, including open access articles from PubMed Central and books from the NCBI. It was developed by the NCBI and is hosted by the U.S. National Library of Medicine (NLM) of the National Institutes of Health .

PubMed® comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

3. Structuring and organization

Large general-purpose sequence databases such as GenBank or EMBL are international projects that are leaders in the field. They have now become indispensable to the scientific community because they bring together essential data and results, some of which are no longer reproduced in the scientific literature.

3.1. Files and formats

The sequences are generally stored as text files which can be either personal files (present in a personal space), or public files (sequences from banks) accessible by Web tools.

The format corresponds to the set of presentation rules (constraints) to which the sequence(s) in a given file are subject. The format allows:

- Automated formatting

- Homogeneous storage of information

- The subsequent computer processing of the information.

A single piece of information in a database is called an "entry".

To help users find their way around, all this information is made available to the scientific community according to an organization into sections or fields.

3.1.1. The FASTA format

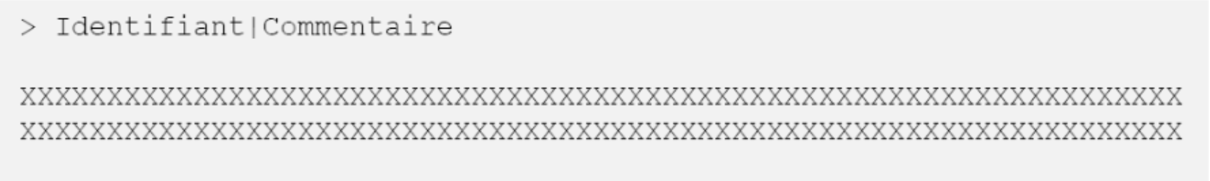
There are several formats, the most common of which is the FASTA format:

Also called format (Pearson) is a text file format used to store biological sequences of a nucleic or protein nature.

The sequence, in the form of lines of up to 80 characters, is preceded by a title line (name, definition, etc.) which must begin with the character ">". Several sequences can be placed in the same file.

The simplicity of the FASTA format makes the manipulation and reading (or syntactic analysis) of sequences easy through the use of text processing tools and programming languages such as C++, Java, Python, R, Matlab or Perl.

Thus, a FASTA file has the following format (the Xs representing nucleic acids or amino acids):

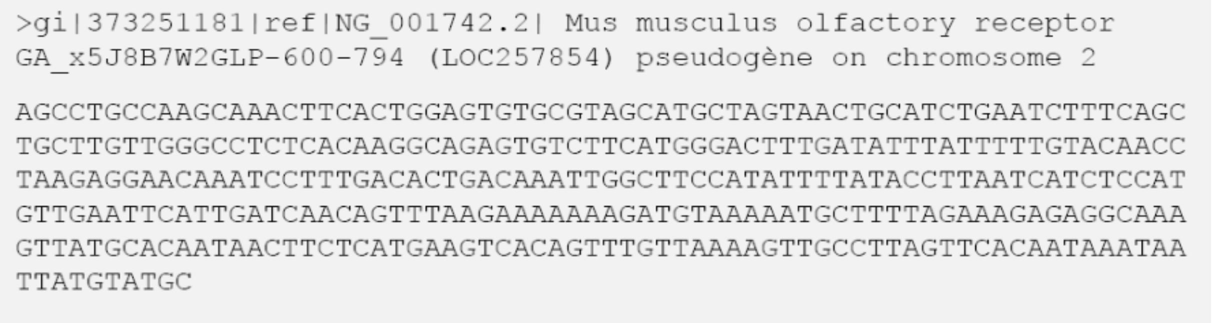


```
> Identifiant|Commentaire

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Typical examples:

Here is an example of a nucleic acid sequence:

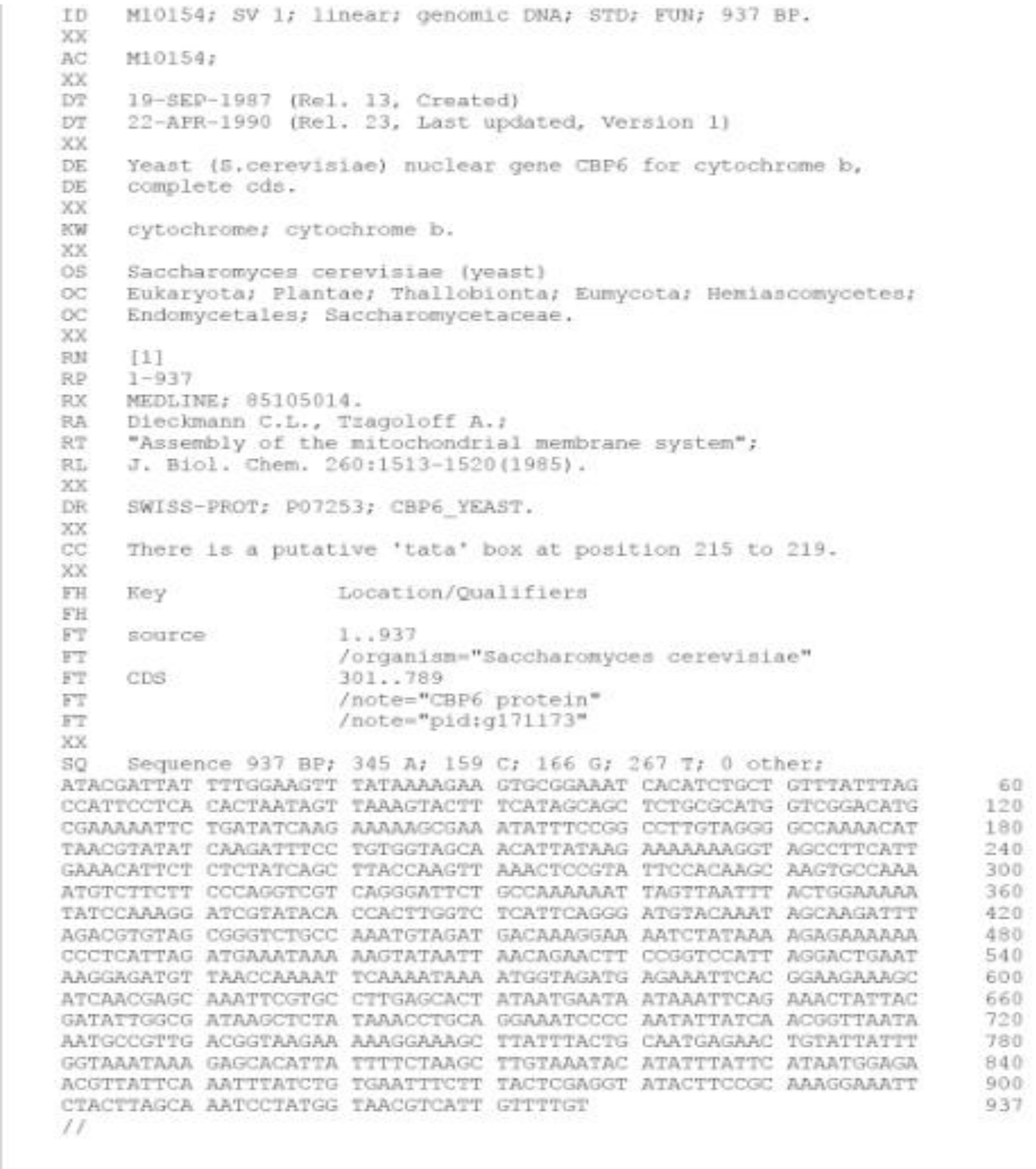


```
>gi|373251181|ref|NG_001742.2| Mus musculus olfactory receptor
GA_x5J8B7W2GLP-600-794 (LOC257854) pseudogène on chromosome 2

AGCCTGCCAAGCAAACTTCACTGGAGTGTGCGTAGCATGCTAGTAACTGCATCTGAATCTTTCAGC
TGCTTGTTGGGCCTCTCACAAGGCAGAGTGTCTTCATGGGACTTTGATATTTATTTTTGTACAACC
TAAGAGGAACAAATCCTTTGACACTGACAAATTGGCTTCCATATTTTATACCTTAATCATCTCCAT
GTTGAATTCATTGATCAACAGTTTAAGAAAAAAAGATGTAAAAATGCTTTTAGAAAGAGAGGCAAA
GTTATGCACAATAACTTCTCATGAAGTCACAGTTTGTTAAAAGTTGCCTTAGTTCACAATAAATAA
TTATGTATGC
```

3.1.2 The EMBL format

https://www.ebi.ac.uk/ena: An example of a genomic DNA sequence from the microorganism Saccharomyces cerevisiae



```
ID   M10154; SV 1; linear; genomic DNA; STD; FUN; 937 BP.
XX
AC   M10154;
XX
DT   19-SEP-1987 (Rel. 13, Created)
DT   22-APR-1990 (Rel. 23, Last updated, Version 1)
XX
DE   Yeast (S.cerevisiae) nuclear gene CBP6 for cytochrome b,
DE   complete cds.
XX
KW   cytochrome; cytochrome b.
XX
OS   Saccharomyces cerevisiae (yeast)
OC   Eukaryota; Plantae; Thallobionta; Eumycota; Hemiascomycetes;
OC   Endomycetales; Saccharomycetaceae.
XX
RN   [1]
RP   1-937
RX   MEDLINE; 85105014.
RA   Dieckmann C.L., Tzagoloff A.;
RT   "Assembly of the mitochondrial membrane system";
RL   J. Biol. Chem. 260:1513-1520(1985).
XX
DR   SWISS-PROT; P07253; CBP6_YEAST.
XX
CC   There is a putative 'tata' box at position 215 to 219.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..937
FT                   /organism="Saccharomyces cerevisiae"
FT   CDS             301..789
FT                   /note="CBP6 protein"
FT                   /note="pid:g171173"
XX
SQ   Sequence 937 BP; 345 A; 159 C; 166 G; 267 T; 0 other;
ATACGATTAT TTTGGAAGTT TATAAAAGAA GTGCGGAAAT CACATCTGCT GTTTATTTAG        60
CCATTCCTCA CACTAATAGT TAAAGTACTT TCATAGCAGC TCTGCGCATG GTCGGACATG       120
CGAAAAATTC TGATATCAAG AAAAAGCGAA ATATTTCCGG CCTTGTAGGG GCCAAAACAT       180
TAACGTATAT CAAGATTTCC TGTGGTAGCA ACATTATAAG AAAAAAAGGT AGCCTTCATT       240
GAAACATTCT CTCTATCAGC TTACCAAGTT AAACTCCGTA TTCCACAAGC AAGTGCCAAA       300
ATGTCTTCTT CCCAGGTCGT CAGGGATTCT GCCAAAAAAT TAGTTAATTT ACTGGAAAAA       360
TATCCAAAGG ATCGTATACA CCACTTGGTC TCATTCAGGG ATGTACAAAT AGCAAGATTT       420
AGACGTGTAG CGGGTCTGCC AAATGTAGAT GACAAAGGAA AATCTATAAA AGAGAAAAAA       480
CCCTCATTAG ATGAAATAAA AAGTATAATT AACAGAACTT CCGGTCCATT AGGACTGAAT       540
AAGGAGATGT TAACCAAAAT TCAAAATAAA ATGGTAGATG AGAAATTCAC GGAAGAAAGC       600
ATCAACGAGC AAATTCGTGC CTTGAGCACT ATAATGAATA ATAAATTCAG AAACTATTAC       660
GATATTGGCG ATAAGCTCTA TAAACCTGCA GGAAATCCCC AATATTATCA ACGGTTAATA       720
AATGCCGTTG ACGGTAAGAA AAAGGAAAGC TTATTTACTG CAATGAGAAC TGTATTATTT       780
GGTAAATAAA GAGCACATTA TTTTCTAAGC TTGTAAATAC ATATTTATTC ATAATGGAGA       840
ACGTTATTCA AATTTATCTG TGAATTTCTT TACTCGAGGT ATACTTCCGC AAAGGAAATT       900
CTACTTAGCA AATCCTATGG TAACGTCATT GTTTTGT                                937
//
```

An explanation of the EMBL format organization is given below:

ID : Identifier, this is the name of the entry containing the sequence. This line has the following structure:

entry name ; data class; molecule; division; length. The name is followed by the indication of the data class, then the type of molecule DNA, RNA or cDNA (XXX if the entry has not been annotated); then the division to which the entry belongs and finally the length of the sequence in base pairs (bp).

AC : The accession number of the entry, which does not change across successive versions of the database. There can be multiple accession numbers for the same entry. Indeed, when two entries are merged into one, a new number may be assigned to the new entry, and those from the former independent entries are retained.

DT : Gives the date of incorporation into the database (1st line) and the date of the last update of the entry (2nd line).

FROM : This line contains descriptive information about the sequence such as the gene name, the region of the genome from which it originates, etc... It is in fact the title of the sequence.

KW : Provide the keyword(s) designated by the authors. These can be used to retrieve the entry in the database. Keywords separated by semicolons are listed in alphabetical order.

OS : Specifies the organism from which the sequence originates; most often, the Latin name is given followed by the English common name in parentheses. In the case of hybrids, the OS/OC lines are specified for each organism in the hybrid.

RN : Unique number assigned to each bibliographic reference in the entry. This number is used to designate the reference in the comments (CC comments) and the biological characteristics field (FT features).

RP : Gives the region of the gene for which the bibliographic reference is associated.

RX : Provides the MEDLINE reference associated with the bibliography. MEDLINE is a bibliographic database that compiles literature related to the biological and biomedical sciences. The database is managed and updated by the National Library of Medicine (NLM).

RA : Indicates the authors of the cited article or work. Authors are cited in the order given in the publication.

RT : Indicates the article title. If the sequence has been submitted to the database but not published, the line will contain only a semicolon.

RL : Provides an abbreviated list of the journal's references. For an article currently in press, the volume and page number will be 0.

DR : Establishes links with other databases that contain information related to this entry. For example, if the protein translation of a sequence exists in the SWISS-PROT database, the DR line will point to the corresponding entry in SWISS-PROT. This line is composed of several fields, which are as follows:

- Database identifier: The database identifier is the common short name given to this database.

- Primary identifier: points to the entry in this database and depends on the referenced database. It points to the accession number if the database is SWISS-PROT, to the ID field if the database is TFD or FLYBASE, and to the entry code if the database is EPD (Eukaryotic). Promoter Database)

- Secondary identifier: complements the information given by the primary identifier and depends on the referenced database, for example it is the name of the entry for UniProt .

CC : Provides comments on the sequence.

FH : This line is used to improve the readability of an entry when it is printed or displayed on the terminal screen: it is the header of the FT (feature) field.

FT : Sequence characteristics (features).

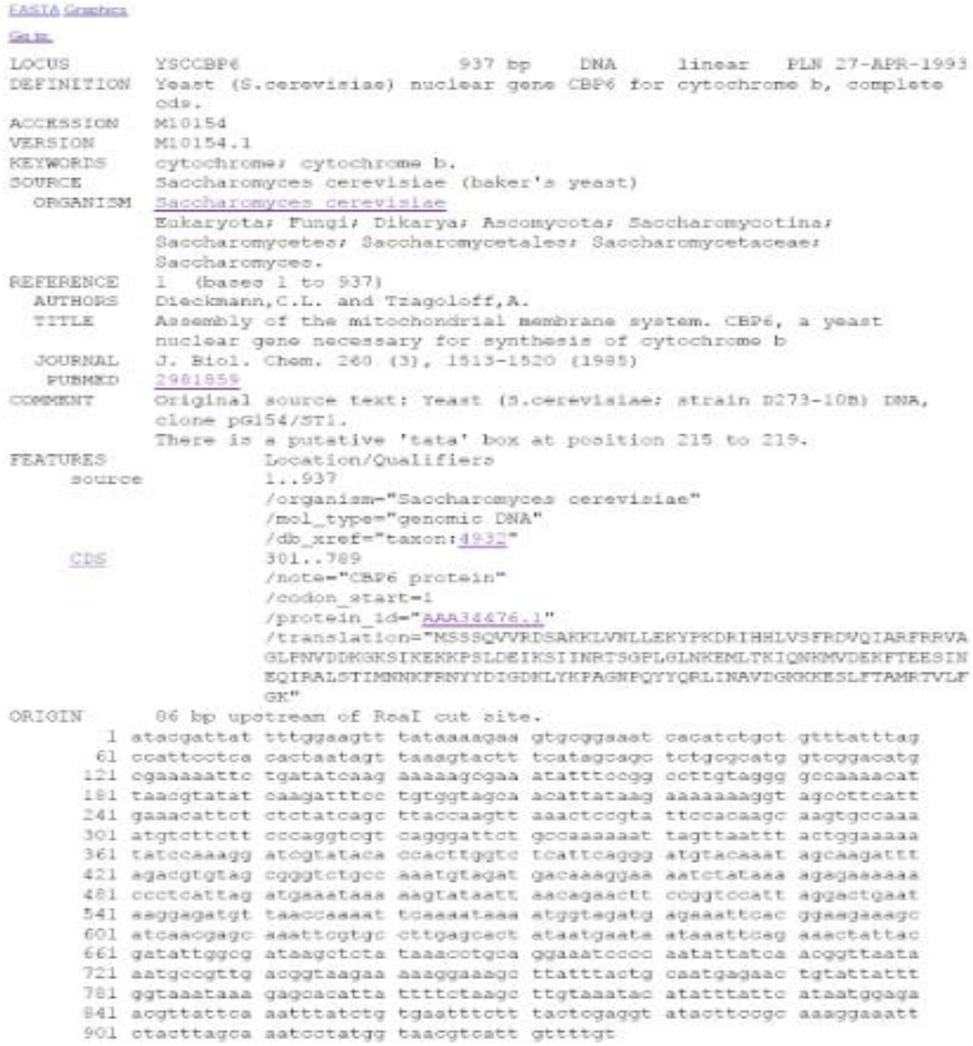
SQ : Sequence (60 nucleotides per line in the direction 5'--->3').

CC : Comments

// End of entry.

3.1.3. The Genbank format

GenBank : M10154.1



FASTA Graphics

Go to:

```
LOCUS       YSCCBP6                  937 bp    DNA     linear   PLN 27-APR-1993
DEFINITION  Yeast (S.cerevisiae) nuclear gene CBP6 for cytochrome b, complete
            cds.
ACCESSION   M10154
VERSION     M10154.1
KEYWORDS    cytochrome; cytochrome b.
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1  (bases 1 to 937)
  AUTHORS   Dieckmann,C.L. and Tzagoloff,A.
  TITLE     Assembly of the mitochondrial membrane system. CBP6, a yeast
            nuclear gene necessary for synthesis of cytochrome b
  JOURNAL   J. Biol. Chem. 260 (3), 1513-1520 (1985)
   PUBMED   2981859
COMMENT     Original source text: Yeast (S.cerevisiae; strain D273-10B) DNA,
            clone pG154/ST1.
            There is a putative 'tata' box at position 215 to 219.
FEATURES             Location/Qualifiers
     source          1..937
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:4932"
     CDS             301..789
                     /note="CBP6 protein"
                     /codon_start=1
                     /protein_id="AAA34476.1"
                     /translation="MSSSQVVRDSAKKLVNLLEKYPKDRIHHLVSFRDVQIARFRRVA
                     GLPNVDDKGKSIKEKKPSLDEIKSIINRTSGPLGLNKEMLTKIQNKMVDEKFTEESIN
                     EQIRALSTIMNNKFRNYYDIGDKLYKPAGNPQYYQRLINAVDGKKKESLFTAMRTVLF
                     GK"
ORIGIN      86 bp upstream of RsaI cut site.
        1 atacgattat tttggaagtt tataaaagaa gtgcggaaat cacatctgct gtttatttag
       61 ccattcctca cactaatagt taaagtactt tcatagcagc tctgcgcatg gtcggacatg
      121 cgaaaaattc tgatatcaag aaaaagcgaa atatttccgg ccttgtaggg gccaaaacat
      181 taacgtatat caagatttcc tgtggtagca acattataag aaaaaaaggt agccttcatt
      241 gaaacattct ctctatcagc ttaccaagtt aaactccgta ttccacaagc aagtgccaaa
      301 atgtcttctt cccaggtcgt cagggattct gccaaaaaat tagttaattt actggaaaaa
      361 tatccaaagg atcgtataca ccacttggtc tcattcaggg atgtacaaat agcaagattt
      421 agacgtgtag cgggtctgcc aaatgtagat gacaaaggaa aatctataaa agagaaaaaa
      481 ccctcattag atgaaataaa aagtataatt aacagaactt ccggtccatt aggactgaat
      541 aaggagatgt taaccaaaat tcaaaataaa atggtagatg agaaattcac ggaagaaagc
      601 atcaacgagc aaattcgtgc cttgagcact ataatgaata ataaattcag aaactattac
      661 gatattggcg ataagctcta taaacctgca ggaaatcccc aatattatca acggttaata
      721 aatgccgttg acggtaagaa aaaggaaagc ttatttactg caatgagaac tgtattattt
      781 ggtaaataaa gagcacatta ttttctaagc ttgtaaatac atatttattc ataatggaga
      841 acgttattca aatttatctg tgaatttctt tactcgaggt atacttccgc aaaggaaatt
      901 ctacttagca aatcctatgg taacgtcatt gttttgt
```