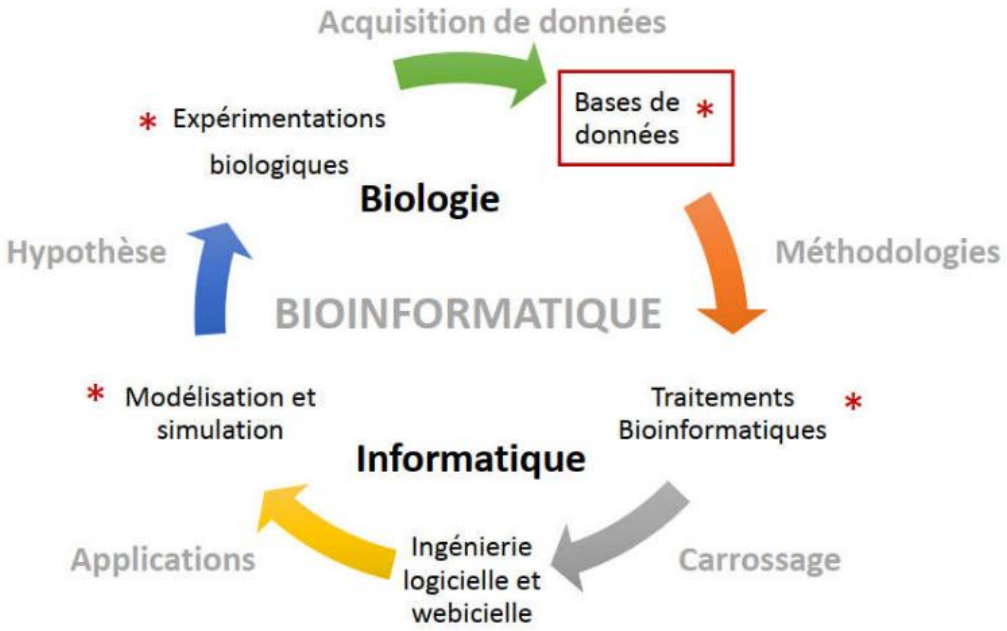
Cour 03 : Stockage de la bioinformation biologique (les bases de données et les banques de données)

1. Introduction et définitions :

Les techniques récentes de biologie moléculaire génèrent une quantité massive de données qui ne sont gérables par les techniques de publication traditionnelle. Dans ce contexte, les banques et les bases de données sont maintenant une source d'information majeure pour la communauté scientifique.



Souvent les termes de banque ou base sont utilisées sans distinction particulière. Toutefois il existe une différence non seulement pour l'utilisateur mais aussi pour l'implantation informatique de ces dernières :

- Banque de données :

Ensemble de données relatif à un domaine défini des connaissances et organisé pour être offert aux consultations d'utilisateurs

- Base de données :

Ensemble de données organisées en vue de son utilisation par des programmes correspondant à des applications distinctes et de manière à faciliter l'évolution indépendante des données et des programmes.

Il existe un grand nombre de banques ou bases de données d'intérêt biologique. Cette introduction sera limitée à une présentation des principales banques de données publiques, basées sur la structure primaire des séquences. Nous distinguerons deux types de banques généralistes et spécialisées.

Ces bases de données peuvent contenir des informations : (ADN, protéines, gènes et génomes, taxonomie, autres, …etc.). On y trouve également une bibliographie et une expertise biologique directement liées aux séquences traitées.

-Différence entre bases de données et banques de données

Il convient de dire qu'une banque de données est une base de données (car tableau structuré) mais qui contient des informations biologiques hétérogènes (virus, bactéries, champignons, végétaux, animaux) alors qu'une base de données est plus spécialisée (base spécifique à E. coli, à Bacillus, etc.).spécifique à E. coli, à Bacillus, etc.).

-Rôle des banques/bases de données

-Collecter les informations (séquences, cartographie physique, génétique…, données structurales, relationnelles…, - auprès de: biologistes, littératures, autres bases de données)

-Stocker et organiser

-Distribuer l'information

-Faciliter l'exploitation

2. Les types de banques de données

Nous distinguerons deux types de banques, celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations (**banques de données généralistes**) et celles qui correspondent à des données plus homogènes établies autour d'une thématique (**banques de données spécialisées**) et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe de scientifique.

→ **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

Tableau qui présente quelques banques de données généralistes

→ 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

Tableau qui présente quelques banques de données spécialisées

- **Les banques de données généralistes**

• Ces banques contiennent des données hétérogènes :

- Collecte la plus exhaustive possible

-Enorme richesse de séquences en un seul ensemble ;

-Grande diversité d'organismes ;

-Nombreuses informations qui accompagnent les séquences

- Banques de séquences nucléiques

- Banques de séquences protéiques

• **Avantage :** tout est consultable en une fois

• **Inconvénients :** difficiles à maintenir, difficiles à interroger

- **Les banques de données spécialisées**

• Ces banques contiennent des données homogènes

• Collecte établie autour d'une thématique particulière

•Elles ont pour but :

- de recenser des familles de séquences autour de caractéristiques biologiques précises comme les signaux de régulation, les promoteurs de gènes, les signatures peptidiques ou les gènes identiques issus d'espèces différentes.

- de regrouper des classes spécifiques de séquences comme les vecteurs de clonage, les enzymes de restriction, ou toutes les séquences d'un même génome.

• Pour des besoins spécifiques liés à l'activité d'un groupe de personnes, ou encore par compilations bibliographiques

• **Avantages :** facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée,…

• **Inconvénients :** ne cible pas toujours ce que l'on veut ; toutes les banques possibles n'existent pas

2.1 Banques de séquences généralistes :

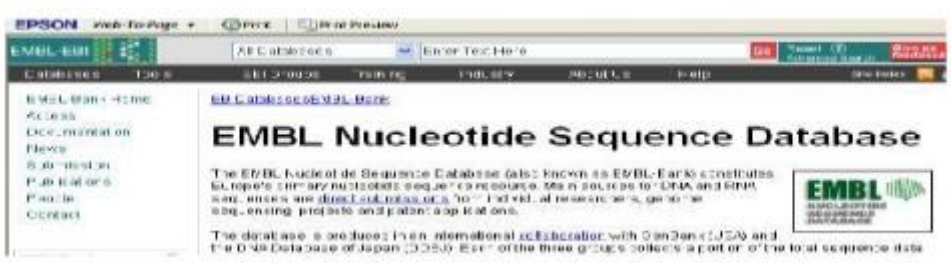
2.1.1 Les banques de séquences nucléiques

Les données stockées dans ce type de banques sont des données issues de séquençage d'ADN et ARN. Trois banques nucléiques connues, elles partagent des informations et donc contiennent des ensembles presque identiques de séquences. Ces trois banques s'échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes : « DDBJ/EMBL/GenBank »:

• **La banque EMBL:** créée en 1980 et financée par l'EMBO (European Moleculary Biology

Organization), développée au sein du Laboratoire Européen de Biologie Moléculaire situé à

Heidelberg (Allemagne), elle est maintenant diffusée par l'EBI: http://www.ebi.ac.uk/embl/. En 24 février 2014, la banque contient 369.5 millions séquences.



• **La banque GenBank (Genetic Sequence Databank):** créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information) : http://www.ncbi.nlm.nih.gov/. En février 2014 la banque contient 171.123.749 séquences. GenBank contient une sous-banque de protéines, traduction des séquences nucléiques, appelée GenPept.

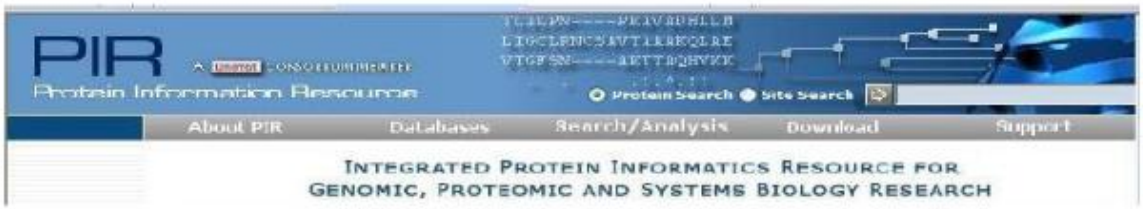
• **La banque DDBJ (DNA Databank of Japan) :** créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon), a enregistré un total de 81.994.905 de séquences ADN le mois de décembre 2019 (DDBJ 2019).

2.1.2. Les banques protéiques

Les données stockées dans ces bases sont issues d'une traduction de séquences d'ADN ou par le séquençage de protéines (rare car long et coûteux):

• **La banque SwissProt :** est une banque protéique crée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI. Celle-ci regroupe aussi des séquences annotées de la banque PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL. En février 2014 la banque contient 542503 séquences compressant 192888369 acides aminés.

• **PIR-NBRF (Protein Information Resource-National Biomedical Research Foundation)** créée en 1984 par la NBRF (National Biomedical Research Foundation). Elle est maintenant un ensemble de données issues du MIPS (Martinsried Institute for Protein Sequences, Munich, Allemagne) et de la banque japonaise JIPID (Japan International Protein Information Database) ;



2.2 Bases spécialisées

• **La banque PDB (Protein Data Bank)** créée en 1971, c'est la banque de référence des structures protéiques obtenues expérimentalement par cristallographie rayon X, spectroscopie RMN et cryo- microscopie électronique (technique la plus récemment utilisée). Les coordonnées des atomes formant la structure d'une protéine, le détail de la séquence, les conditions de cristallisation sont les principales informations disponibles pour chaque structure de la banque. C'est à partir de cette banque que sont détectés les homologues structuraux.

ECD : base sur les séquences nucléiques d'Escherichia coli.

NRL3D : base de séquences protéiques dont la structure tridimensionnelle a été déterminée.

TFD : base de facteurs de transcription.

Prosite : bases de motifs protéiques. Elle peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique.

CATH : base sur les classifications hiérarchiques (ordonnées) des structures protéiques. IMGT, base de séquences des immunoglobulines et des récepteurs T.

GENATLAS : base d'informations issues de la cartographie des gènes humains.

KEGG : bases de voies métaboliques.

Les bases de motifs :

On sait que certains segments d'ADN ou de protéines sont déterminants dans l'analyse des séquences car ils correspondent à des sites précis d'activité biologique comme par exemple les éléments de régulation des gènes ou les signatures peptidiques. C'est pourquoi des bases spécialisées se sont naturellement constituées autour de ces séquences. L'utilisation des bases spécialisées comme les bases de motifs, est devenue un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéines inconnues ou savoir à quelle famille appartient une séquence non encore caractérisée.

• **TFD ou IMD :** sont employées pour des séquences promotrices des gènes

• **Prosite ou BLOCKS :** sont utilisées pour des protéines inconnues ou bien des séquences protéiques traduites à partir de cDNA ou de séquences génomiques.

Pour détecter une fonctionnalité sur une séquence, il suffit d'exécuter un programme qui s'appliquera à repérer la présence de certains motifs recensés dans ces bases et ainsi à prédire l'appartenance de la séquence testée à un groupe de séquences ayant une signature commune.

Banques immunologiques

Elles sont spécialisées dans les informations suivantes :

- Séquences
- Récepteur (cellule T, par exemple)
- Complex MHC (Major Histocompatibility Complex)
- Système HLA

Banques Structure 2D ou 3D

Elles sont spécialisées dans les informations suivantes :

- Coordonnées 3D de protéines *
- Structure secondaire des protéines
- Domaines structuraux
- Centre actif des enzymes
- Complexes récepteurs-ligands
- Atlas de topologie structurale des protéines

2.3 Bases de données Bibliographiques

Les bases de données bibliographiques répertorient toute catégorie d'objets bibliographiques : livres, journaux scientifiques, articles …

Ex : PubMed est une base de données bibliographiques en sciences biologiques et sciences biomédicales dont la couverture débute en 1946 et qui contient plus de 30 millions de références. En plus des articles indexés dans MEDLINE, PubMed contient aussi des références additionnelles, incluant les articles en accès libre de PubMed Central et les livres du NCBI. Il a été développé par le (NCBI), et est hébergé par la Bibliothèque nationale de médecine (NLM) américaine du National Institutes of Health.

3. Structuration et organisation

Les grandes banques de séquences généralistes telles que GenBank ou l'EMBL sont des projets internationaux qui constituent des leaders dans le domaine. Elles sont maintenant devenues indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique.

3.1. Fichiers et formats

Les séquences sont stockées en général sous forme de fichiers texte qui peuvent être soit des fichiers personnels (présents dans un espace personnel), soit des fichiers publics (séquences des banques) accessibles par des outils Web.

Le format correspond à l'ensemble des règles (contraintes) de présentation auxquelles sont soumises la ou les séquences dans un fichier donné. Le format permet :

- Une mise en forme automatisée

- Le stockage homogène de l'information

- Le traitement informatique ultérieur de l'information.

Une seule pièce d'informations dans une base de données est nommée "entrée"

Pour que l'utilisateur puisse se repérer, toutes ces informations sont mises à la disposition de la collectivité scientifique selon une organisation en rubriques ou en champs.

3.1.1. Le format FASTA

Il existe plusieurs formats dont le plus courant est le format FASTA :

Appelé aussi format (Pearson) est un format de fichier texte utilisé pour stocker des séquences biologiques de nature nucléique ou protéique.

La séquence, sous forme de lignes de 80 caractères maximum, est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">". Plusieurs séquences peuvent être ainsi mises dans un même fichier.

La simplicité du format FASTA rend la manipulation et la lecture (ou analyse syntaxique) des séquences aisées par l'utilisation d'outils de traitement de texte et de langages de programmations tels que C++, Java, Python, R, Matlab ou Perl.

Ainsi un fichier FASTA se présente sous la forme suivante (les X représentant acides nucléiques ou aminés) :

```
> Identifiant|Commentaire

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Exemples types :

Voici un exemple de séquence nucléique :

```
>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 Le format EMBL

https://www.ebi.ac.uk/ena : L'exemple d'une séquence d'ADN génomique d'un micro organisme 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
//
```

Une explication de l'organisation du format EMBL est donnée ci-dessous :

ID : Identificateur, c'est le nom de l'entrée contenant la séquence. Cette ligne a la structure suivante :

nom de l'entrée ; classe de la donnée ; molécule ; division ; longueur. Le nom est suivi de l'indication de la classe de donnée, puis du type de molécule ADN, ARN ou ADNc (XXX si l'entrée n'a pas été annotée) ; ensuite la division à laquelle l'entrée appartient et enfin la longueur de la séquence en paires de bases (bp).

AC : Numéro d'accession de l'entrée qui ne varie pas au cours des versions successives de la banque. Il peut y avoir plusieurs numéros d'accessions pour une même entrée. En effet lorsque deux entrées sont fusionnées en une seule, un nouveau numéro peut être attribué à la nouvelle entrée et ceux provenant des ex-entrées indépendantes sont conservés.

DT : Donne la date d'incorporation dans la base (1ère ligne) et la date de la dernière mise à jour de l'entrée (2ème ligne).

DE : Cette ligne contient des informations descriptives sur la séquence comme le nom du gène, la région du génome dont elle est issue etc... C'est en fait le titre de la séquence.

KW : Donne-le(s) mot(s)-clé(s) désignés par les auteurs. Ils peuvent être utilisés pour retrouver l'entrée dans la base. Les mots-clés séparés par des ; sont rangés par ordre alphabétique.

OS : Spécifie l'organisme d'où provient la séquence ; le plus souvent, on donne le nom latin suivi du nom commun anglais entre parenthèses. Dans le cas d'hybrides les lignes OS/OC sont spécifiées pour chaque organisme de l'hybride.

RN : Numéro unique attribué à chaque référence bibliographique de l'entrée. Ce numéro est utilisé pour désigner la référence dans les commentaires (CC comments) et le champ des caractéristiques biologiques (FT features).

RP : Donne la région du gène pour laquelle la référence bibliographique est associée.

RX : Donne la référence MEDLINE associée à la bibliographie. MEDLINE Est une base de données bibliographiques regroupant la littérature relative aux sciences biologiques et biomédicales. La base est gérée et mise à jour par la Bibliothèque américaine de médecine (NLM).

RA : Indique les auteurs de l'article ou du travail cité. Les auteurs sont cités dans l'ordre donné dans la publication.

RT : Indique le titre de l'article. Si la séquence a été soumise à la base et non publiée, la ligne ne contiendra qu'un ;

RL : Donne d'une manière abrégée les références du journal. Pour un article sous presse le numéro du volume et des pages sera de 0.

DR : Etablit des liaisons avec d'autres bases de données qui contiennent une information en relation avec cette entrée. Par exemple, si la traduction protéique d'une séquence existe dans la banque de données SWISS-PROT, la ligne DR pointera sur l'entrée correspondante dans SWISS-PROT. Cette ligne est composée de plusieurs champs qui sont les suivants :

• Identificateur de la banque de données : L'identificateur de la base de données est le nom abrégé courant que l'on donne à cette base.

• Identificateur primaire : pointe sur l'entrée de cette base et dépend de la base référencée. Il pointe sur le numéro d'accession si la base est SWISS-PROT, sur le champ ID si la base est TFD ou FLYBASE et sur le code d'entrée si la base est EPD (Eucaryotic Promoter Database)

• Identificateur secondaire : complète l'information donnée par l'identificateur primaire et dépend de la base référencée, par exemple c'est le nom de l'entrée pour UniProt.

CC : Donne les commentaires sur la séquence.

FH : Cette ligne sert à améliorer la lecture d'une entrée lorsqu'elle est imprimée ou affichée sur l'écran du terminal : c'est l'en-tête du champ FT (feature)

FT : Caractéristiques de la séquence (features).

SQ : Séquence (60 nucléotides par ligne dans le sens 5'--->3').

CC : Commentaires

// Fin de l'entrée.

3.1.3. Le format Genbank

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      Q6 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 tgaattttct tactcgaggt atacttccgc aaaggaaatt
      901 ctacttagca aatcctatgg taacgtcatt gttttgt
```