

Université Batna 2

Année universitaire : 2022-2023

Faculté des Sciences de la nature et de la vie

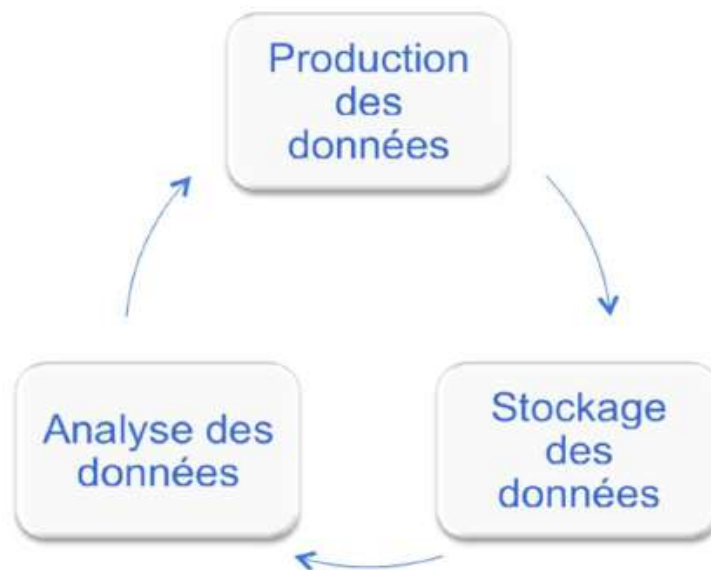
Département d'Ecologie et environnement

Cours 1 : Les banques de données biologiques

1. Définition de la bioinformatique

- La bioinformatique est définie comme l'utilisation de bases de données et d'algorithmes informatiques pour analyser, les gènes, les protéines, et la collection complète d'acide désoxyribonucléique (ADN) d'un organisme vivant (le génome).

- La bioinformatique est la discipline de l'analyse « *in silico*¹ » de l'information biologique renfermée dans les séquences nucléotidiques (séquences de nucléotides) et protéiques (séquence des acides aminés).



C'est une discipline complémentaire aux approches classiques de la biologie :

- *In vivo* (tests au sein des organismes vivants) ;
- *In situ* (tests dans les milieux naturels) ;
- *In vitro* (tests dans des tubes).

¹ *in silico* : se réfère à l'outil informatique. Lorsqu'on dit *in silico* cela veut dire l'utilisation des processeurs, logiciels informatiques pour gérer, traiter et analyser l'information biologique contenu essentiellement dans les séquences nucléiques et protéiques.

2. Domaines de la Bioinformatique

- **Stockage et Gestion des données** : Banques de données généralistes et spécialisées.
- **Structures moléculaires** : Visualisation, analyse, classification, prédiction.
- **Analyse de séquences** : Alignements, recherches de similarités, détection de motifs.
- **Génomique structurale** : Annotation des génomes, génomique comparative.
- **Génomique fonctionnelle** : Transcriptome², protéome³, interactome⁴.
- **Phylogénie** : Relations évolutives entre gènes, entre génomes, entre organismes ; Inférence de scénarios évolutifs.
- **Analyse des réseaux biomoléculaires** : Réseaux métaboliques, d'interactions protéiques, de régulation génétique, ...

Exemples d'applications

- Recherche en biologie
 - L'organisation moléculaire de la cellule / organisme
 - Développement
 - Mécanismes de l'évolution
- Médecine
 - Diagnostic de cancers
 - Détection des gènes impliqués dans le cancer
- La recherche pharmaceutique
 - mécanismes d'action des médicaments
 - identification de cibles pharmaceutiques
- Biotechnologie
 - La thérapie génique

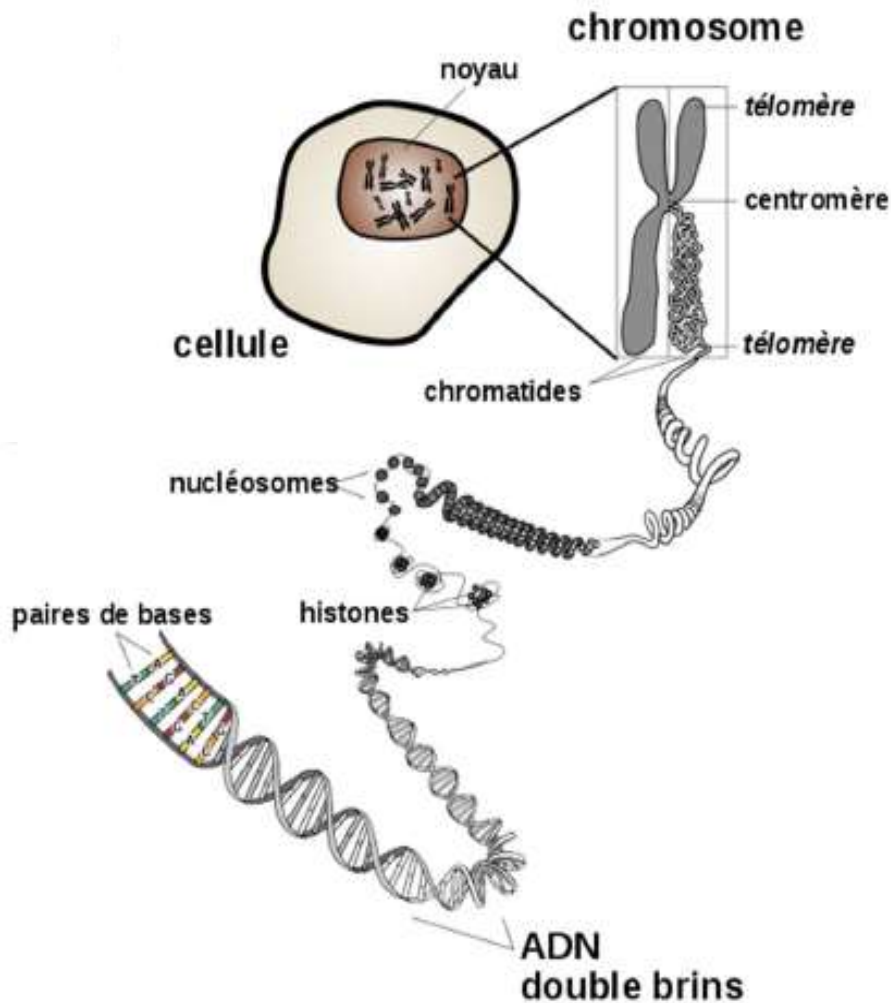
3. Rappel de biologie moléculaire

- L'information génétique est contenue dans les *chromosomes* situés dans le noyau des cellules
- Chaque cellule d'un être humain comporte **23 paires** de chromosomes
- Un chromosome est constitué de molécules d'ADN

² Le **transcriptome** : l'ensemble des transcrits ou ARNm

³ Le **protéome** : l'ensemble des protéines bio synthétisés dans une cellule, un tissu ou chez un organisme.

⁴ l'**interactome** : l'ensemble des protéines et/ou acides nucléiques interagissant entre eux

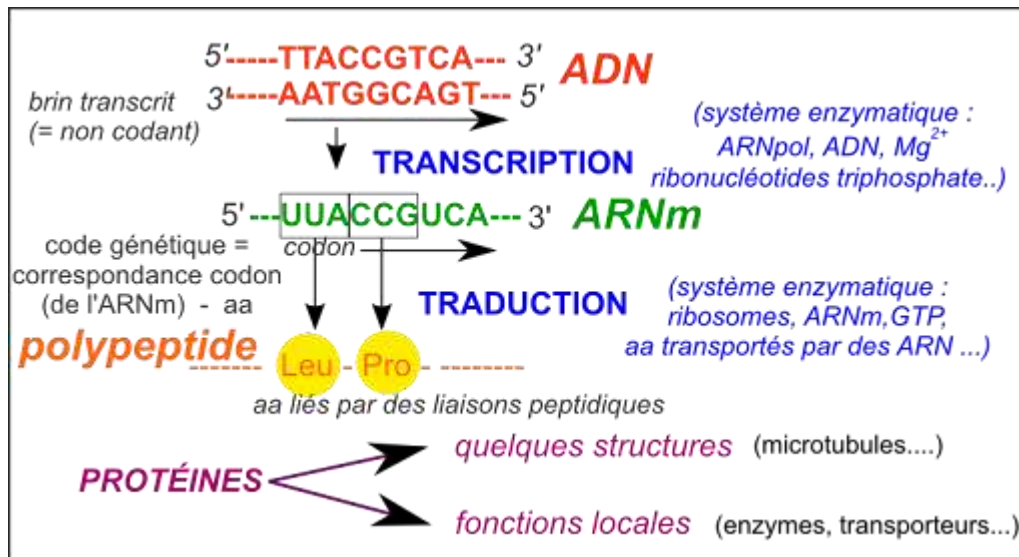


Par interaction avec l'environnement, l'ADN se transforme en protéines :

- _ La **transcription**, transfert de l'ADN vers une autre molécule, l'**ARN**
- _ La **traduction**, transfert depuis l'ARN vers des **protéines**
- _ L'activité des protéines détermine l'activité des cellules qui vont ensuite déterminer le fonctionnement des organes et de l'organisme
- _ **Traduction** de l'**ADN** en **protéine** :
- _ Les quatre lettres A, C, G et T s'associent en mots de trois lettres (GGA, CTA...) pour former un **codon**. Des ribosomes décodent ces codons en **acides aminés** combinées pour former des protéines.

Dogmes centraux de la biologie moléculaire :

L'ADN est le support de l'information génétique et constitué de deux brins antiparallèles et complémentaires. Son information est transcrite en ARN. Cet ARN peut être un ARN messager ou ARN fonctionnelle (ARNr, ARNt, ...). L'ARN (monobrin) est capable de traduire son information en protéine via une opération que l'on appelle traduction. On fait toutes ces données sont traitées par la bioinformatique autrement dit l'ADN et ses dérivés sont l'objet essentiel de l'étude bioinformatique.



Language protéique :

■ Acides aminés : codes à 1 et 3 lettres

- Acide aspartique (D, Asp)
- Acide glutamique (E, Glu)
- Alanine (A, Ala)
- Arginine (R, Arg)
- Asparagine (N, Asn)
- Cystéine (C, Cys)
- Glutamine (Q, Gln)
- Glycine (G, Gly)
- Histidine (H, His)
- Isoleucine (I, Ile)
- Leucine (L, Leu)
- Lysine (K, Lys)
- Méthionine (M, Met)
- Phénylalanine (F, Phe)
- Proline (P, Pro)
- Sérine (S, Ser)
- Thréonine (T, Thr)
- Tryptophane (W, Trp)
- Tyrosine (Y, Tyr)
- Valine (V, Val)

Chapitre I. LES BANQUE DE DONNÉES BIOLOGIQUES

Les bases de données contenant des informations biologiques et des données largement diffusées par le réseau Internet. Elles sont généralement reliées entre elles par des liens « links ».

1. Définition

Les bases de données biologiques sont des **bibliothèques électronique et informatisé** qui contiennent des informations sur les sciences de la vie, collectées grâce à des expériences scientifiques, à la littérature publiée, aux technologies expérimentales à haut débit, et aux analyses informatiques.

2. Rôle des banques et bases de données biologiques

Leur principale mission est de rendre publiques les séquences qui ont été déterminées, ainsi un des premiers intérêts de ces banques est la masse de séquences qu'elles contiennent. Entre autres ils ont pour mission l'archivage, le stockage, la diffusion et l'exploitation des données biologiques.

3. Contenus des bases de données biologiques

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.

4. Les types de banques de données

Il existe un grand nombre de bases de données d'intérêt biologique. Nous nous limiterons dans ce chapitre à une présentation des principales banques de données publiques, basées sur la structure primaire des séquences, qui sont largement utilisées dans l'analyse informatique des séquences.

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 scientifiques.

Tableau 1. Quelques banques de données généralistes

→ Banques de séquences nucléiques généralistes			
Nom	Lien	Date de création	Description
EMBL	http://www.ebi.ac.uk/embl/	1980	Banque européenne (European Molecular 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 2. Quelques banques de données spécialisées

→ Banques de données 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

4.1. Les banques de données généralistes

- Ces banques contiennent des données hétérogènes :
 - Données globales (pas de focus sur une application ou organisme particulier)
 - Collecte la plus exhaustive et la plus large des données possibles
 - Banques de séquences nucléiques (ADN et ARN)
 - Banques de séquences protéiques
 - Banques de structures 3 D de macromolécules
 - Banques d'articles scientifiques (Bibliographiques)
- **Avantage** : tout est consultable en une fois
- **Inconvénients** : difficiles à maintenir, difficiles à interroger, problèmes de redondance

- **Qualité des séquences des banques généralistes**

- Très riches
 - Grand nombre de séquences accessibles
 - Grande diversité des organismes représentés
 - Informations accompagnant les séquences (annotation, expertise, bibliographie, liens)
- Peu/pas de contrôles sur la qualité des entrées
 - Les auteurs sont responsables des entrées ! => Nombreux Problèmes/Erreurs
- Erreurs dans les séquences (contaminations, séquençage, méthodologie)

4.2. Les banques de données spécialisées

- Ces banques contiennent des données homogènes
- Collecte établie autour d'une thématique particulière
- **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

- **Exemples** :

Bases spécialisées pour un génome spécifique, bases de séquences immunologiques, de voies métaboliques, de cartes génétiques, de motifs protéiques, d'expression de gènes, de structures, . . .

Quelques exemples :

- **A) LEAPdb : Late Embryogenesis Abundant Proteins database** (G. Hunault & E. Jaspard) : cette base de données contient un grand nombre d'informations sur les protéines LEA⁵ impliqués dans la tolérance à de nombreux stress, notamment la déshydratation et le froid. Pour l'instant, on les a mises en évidence principalement chez les plantes.

LEAPDB
1642 proteins

[Home](#)

[Browse](#)

[Search](#)

[Blast](#)


[Statistical analysis](#)


[Export](#)

[Submit](#)

[Help](#)

[Contacts](#)





Admin

Late Embryogenesis Abundant Proteins

LEA proteins have been discovered in 1981. Although, they are almost associated with abiotic stress tolerance (particularly dehydration and cold stress), their actual function remains unknown.

The LEAP database provides useful data about the different CLASSES of LEA proteins for the analysis of their structure - function relationships. More...

Main classifications of LEAP with time - Introduction of CLAS nomenclature

Ref	Dufré et al.	Wiley	Furness and Olive	Battaglia et al.	Bendathia et al.	Handberg and Madsen	LEAPdb
Year	1981	1993	2007	2008	2008	2008	2010
FF1027	D11	Group 2	Group 2	Group 2	Group 2	Group 2	Classes 1 to 4
FF1027	D19 D12	Group 1	Group 1	Group 1	Group 1	LEA_3	Class 5
FF1267	D7	Group 3	Group 3	Group 3	Group 3	LEA_4	Class 6
FF1267	D25	Group 5	—	Group 5	—	—	—
FF1148	D8	—	—	Group 6	Group 7	LEA_7	Classes 7 & 8
FF1142	D12	—	LEA8	Group 8	Group 8	LEA_8	Class 9
FF1740	—	Group 4	Group 4	Group 4	Group 4	LEA_9	Class 10
FF1740	D13	—	—	Group 4	—	—	—
FF1497	D4	Group 6	Group 6	Group 6	Group 6	SNP	Class 11
FF1714	—	—	—	Group 9	Group 9	PLC419	Class 12

Gilles Hunault and Emmanuel Jaspard,
LEAPdb: a database for the late embryogenesis abundant proteins
[BMC Genomics 2010, 11:221](#)

Emmanuel Jaspard, David Macherel and Gilles Hunault,
Computational and Statistical Analyses of Amino Acid Usage and Physico-Chemical Properties of the Twelve Late Embryogenesis Abundant Protein Classes
[PLoS ONE 2012, 7\(5\)](#)

⁵ Les protéines abondantes de l'embryogenèse tardive (protéines LEA) sont des protéines des plantes et de certaines bactéries et invertébrés qui protègent contre l'agrégation des protéines due à la dessiccation ou aux stress osmotiques associés aux basses températures. Les protéines LEA ont été initialement découvertes s'accumulant tard dans l'embryogenèse des graines de coton.

Les protéines LEA fonctionnent par des mécanismes distincts de ceux présentés par les chaperons moléculaires de choc thermique. Bien que les causes de l'induction de la protéine LEA n'aient pas encore été déterminées, des changements conformationnels dans les facteurs de transcription ou les protéines membranaires intégrales dus à la perte d'eau ont été suggérés. Les protéines LEA protègent particulièrement les membranes mitochondriales contre les dommages dus à la déshydratation.

- La Fonction des protéines LEA :

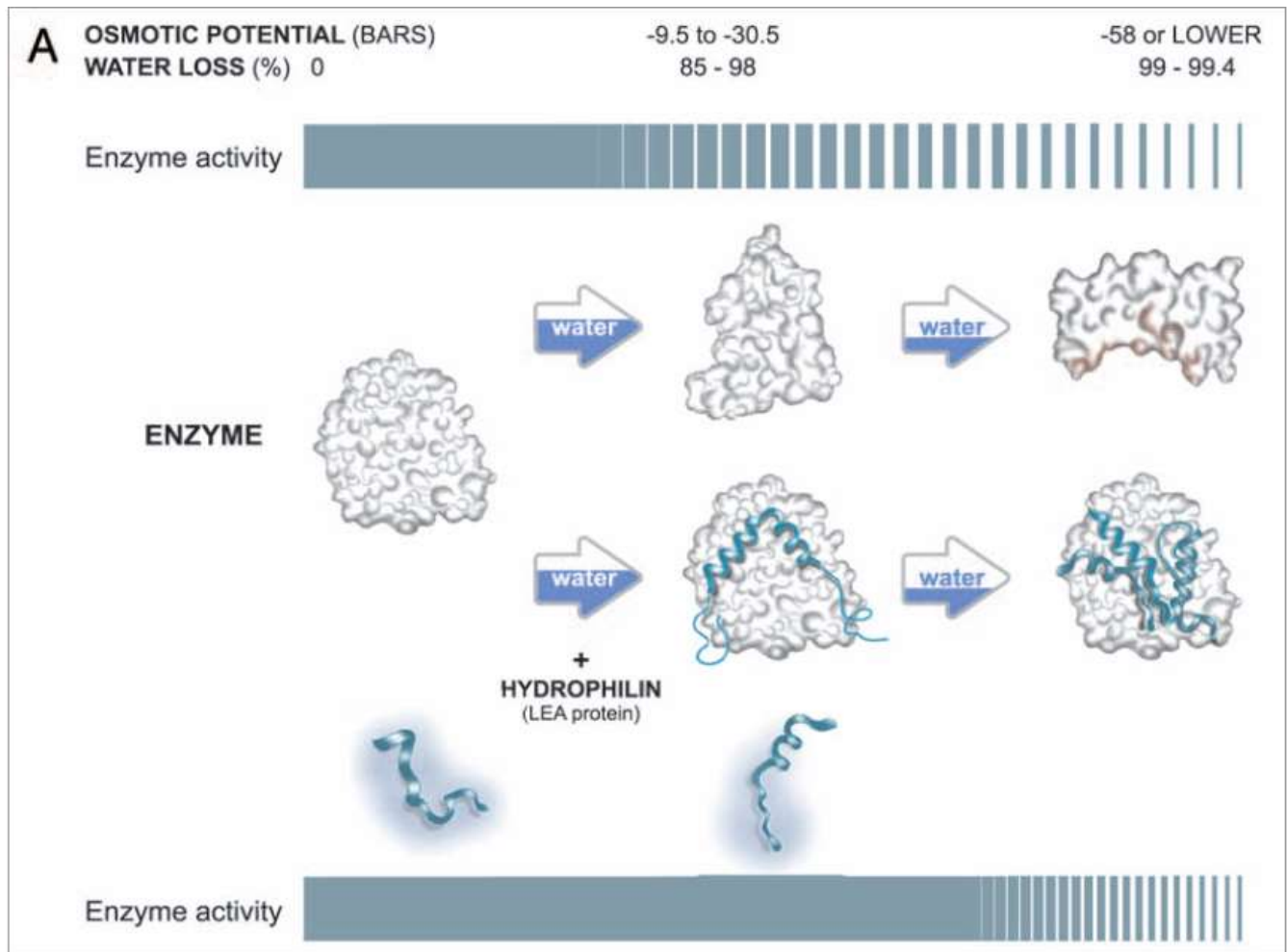
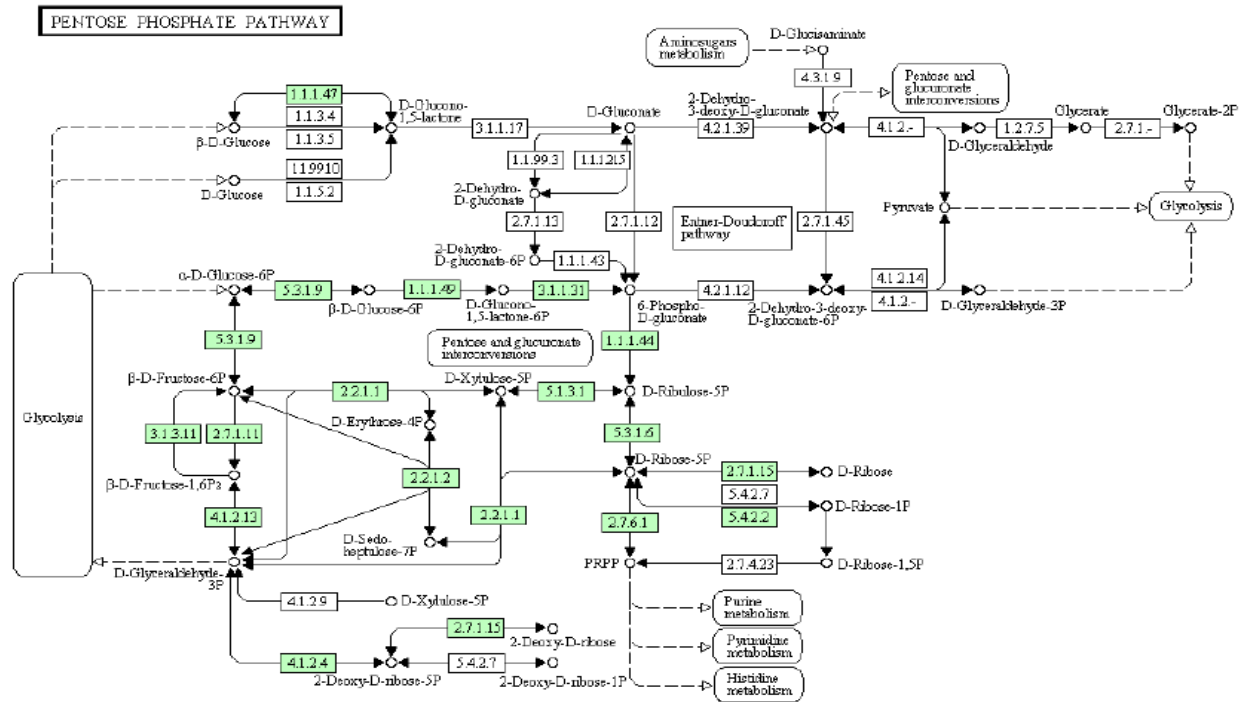


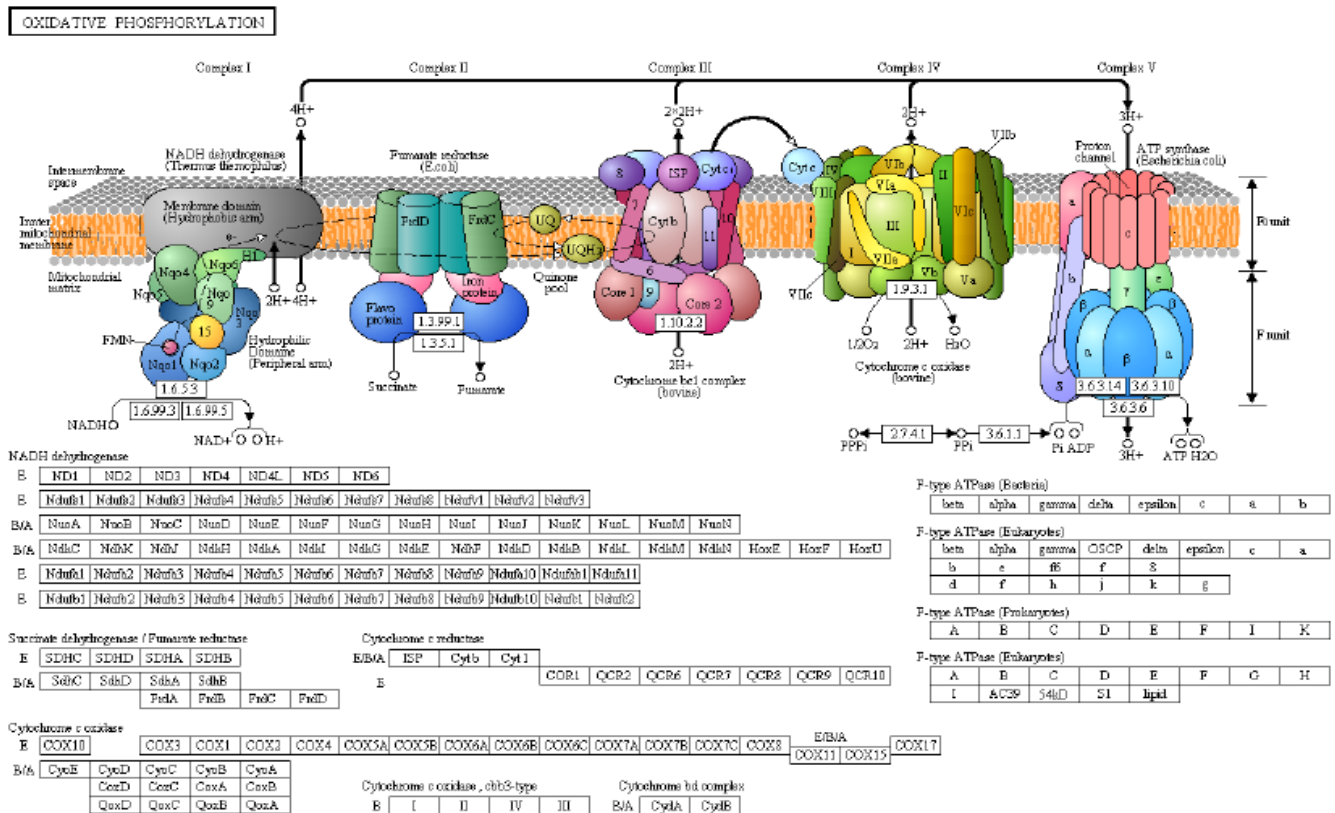
Figure. Ce schéma illustre un modèle hypothétique pour la fonction des protéines LEA et d'autres hydrophilines. Dans cet exemple, sous modération déficit hydrique, une enzyme subite des changements conformationnels qui conduisent à une diminution de son activité et, dans des conditions de stress plus sévères, plus critiques les modifications structurelles entraînent l'exposition de résidus hydrophobes (ombrage rouge). La présence de protéines LEA (hydrophilines) (brin vert) empêche les modifications de la conformation de l'enzyme, à la suite desquelles l'enzyme conserve son activité, dans des conditions de limitation en eau. Cet effet peut être atteint à un rapport hydrophiline:enzyme de 1:1 sous un stress hydrique modéré; cependant, en cas de déshydratation sévère, l'action de plus d'un l'hydrophiline par molécule d'enzyme pourrait éviter d'autres changements conformationnels pouvant conduire à l'agrégation des protéines.

B) KEEG Pathway : base de données spécialisées dans les voies métaboliques

Exemple (1) : La voie de pentose phosphate



Exemple (2) : La phosphorylation oxydative



00190 2/0/10
© Kanislas Laboratoire

4.3. Les banques de séquences nucléiques

• Origine des données

- Séquençage d'ADN et d'ARN
- Traduction inverse de séquences protéiques en séquences ADN
- Les données stockées : séquences + annotations
- Fragments de génomes
- Un ou plusieurs gènes, un bout de gène, séquence intergénique, ...
- Génomes complets
- ARNm, ARNt, ARNr, ... (fragments ou entiers)

[Remarque 1] : toutes les séquences (ADN ou ARN) sont écrites avec des T

[Remarque 2] : les séquences sont toujours orientées 5' → 3'.

- Types de Banques généralistes de séquences nucléotidiques

• ENA (European Nucleotide Archive) ou EMBL (European Molecular Biology Laboratory) :

- Création 1980 par l'European Molecular Biology Organisation
- Diffusée par European Bioinformatics Institute (EBI)

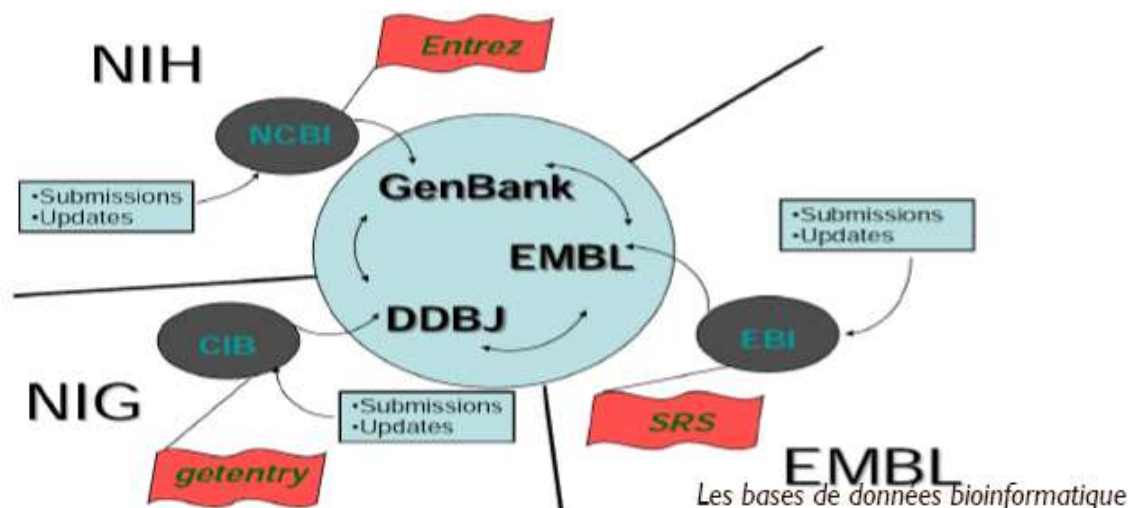
• Genbank

- Création 1982 par IntelliGenetics
- Diffusée par National Center for Biotechnology Information (NCBI)

• DDBJ (DNA Databank of Japan)

- Création 1986 par National Institute of Genetics (NIG)
- Diffusée par National Institute of Genetics (NIG)

❖ Ces trois banques sont interconnectées et elles échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes « The DDBJ/EMBL/Genbank Feature Table Definition »



4.4. Les banques de séquences protéiques

- **Origine des données**

- La Traduction automatisée de séquences d'ADN en séquences protéiques
- Séquençage de protéines (Chimique : Méthode d'Edman, enzymatique) (Rare car long et coûteux)
- Protéines dont la structure 3D est connue

- Les données stockées : séquences + annotations

- Protéines entières
- Fragments de protéines

- **Types de Banques généralistes de séquences protéiques**

- TrEMBL : traduction automatique de EMBL
- Genpept : traduction automatique de GenBank
- PIR (Protein Information Ressource) :
 - Première banque des protéines (1965) = Atlas of proteins publié par Margaret Dayhoff
 - Banque américaine (NBRF- National Biomedical Research Fondation)
 - Protéines regroupés en familles
- SwissProt
 - 1986 à l'université de Genève
 - Origine des séquences TrEMBL

Swiss-Prot + PIR + TrEMBL-EBI



UniProt

(Universal Protein Ressource)

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



3 - Banques de structure 3D de macromolécules (PDB) ;

PDB (Protein DataBank)

<http://www.rcsb.org>

- Séquences et structures 3D des protéines et des acides nucléiques macromolécules .
- Visualisation en 3D .

5. 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

5.1. 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**.

5.1.1. 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
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

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
AGCCTGCCAAGCAAACCTTCACTGGAGTGTGCGTAGCATGCTAGTAACTGCATCTGAATCTTTCAGC
TGCTTGTTGGGCCTCTCACAAGGCAGAGTGTCTTCATGGGACTTTGATATTTATTTTGTACAACC
TAAGAGGAACAAATCCTTTGACACTGACAAATTTGGCTTCCATATTTTATACTTAATCATCTCCAT
GTTGAATTCATTGATCAACAGTTTAAGAAAAAAGATGTAAAAATGCTTTTAGAAAGAGAGGCAAA
GTTATGCACAATAACTTCTCATGAAGTCACAGTTTGTAAAAGTTGCCTTAGTTTACAATAAATAA
TTATGTATGC
```

5.1.2. 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:gl71173"
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 AACTCCGTA 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 TAACCAAAT  TCAAATAAAA 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'accèsion de l'entrée qui ne varie pas au cours des versions successives de la banque. Il peut y avoir plusieurs numéros d'accèsions 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'accèsion 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.

5.1.3. 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="MSSSQVVRDSAKKLVNLLLEKYPKDRIHHLVSFRDQIARFRRVA
            GLPNVDDKKGKSIKEKKPSLDEIKSIIINRTSGPLGLNKEMLTQNKMVDEKFTEESIN
            EQIRALSTIMNNKFRNYDIGDKLYKPAGNPQYYQRLINAVDGKKKESLFTAMRTVLF
            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 gccaaaaacat
            181 taacgtatat caagatttcc tgtggtagca acattataag aaaaaaaggt agccttcatt
            241 gaaacattct ctctatcagc ttaccaagtt aaactccgta ttccacaagc aagtgcctaaa
            301 atgtcttctt cccagggtcgt cagggattct gccaaaaaat tagttaattt actggaaaaa
            361 tatccaaagg atcgtataca ccacttggtc tcattcaggg atgtacaaat agcaagattt
            421 agacgtgtag cgggtctgcc aaatgtagat gacaaaggaa aatctataaa agagaaaaaa
            481 ccctcattag atgaaataaa aagtataatt aacagaactt cgggtccatt aggactgaat
            541 aaggagatgt taacccaaaat tcaaaataaa atggtagatg agaaattcac ggaagaaagc
            601 atcaacgagc aaattcgtgc cttgagcact ataataaata 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

```

5-1-4- Format Swiss Prot

ID TCPB_YEAST Reviewed; 527 AA.
 AC P39076; D6VVE5;
 DT 01-FEB-1995, integrated into UniProtKB/Swiss-Prot.
 DT 01-FEB-1995, sequence version 1.
 DT 12-SEP-2018, entry version 165.
 DE RecName: Full=T-complex protein 1 subunit beta;
 DE Short=TCP-1-beta;
 DE AltName: Full=CCT-beta;
 GN Name=CCT2; Synonyms=BIN3, TCP2; OrderedLocusNames=YIL142W;
 OS *Saccharomyces cerevisiae* (strain ATCC 204508 / S288c) (Baker's yeast).
 OC Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
 OC Saccharomycetes; Saccharomycetales; Saccharomycetaceae; *Saccharomyces*.
 OX NCBI_TaxID=559292;
 RN [1]
 RP NUCLEOTIDE SEQUENCE [GENOMIC DNA].
 RC STRAIN=ATCC 204511 / S288c / AB972;
 RX PubMed=7908441; DOI=10.1073/pnas.91.7.2743;
 RA Miklos D., Caplan S., Mertens D., Hynes G., Pitluk Z., Kashi Y.,
 RA Harrison-Lavoie K., Stevenson S., Brown C., Barrell B.G.,
 RA Horwich A.L., Willison K.;
 RT "Primary structure and function of a second essential member of the
 RT heterooligomeric TCPI chaperonin complex of yeast, TCPI beta."
 RL Proc. Natl. Acad. Sci. U.S.A. 91:2743-2747(1994).

5-1-5- Format de fichier texte brut (*Plain Raw Sequence*)

- ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGC
 CACGGCCACCGCTGCCCTGCCCTGGAGGGTGGCCCCACCGGCCGAGAC
 AGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTCCTGA
 CTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCAGGCCAGTGCCGGGCC
 CCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCA
 CCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCT
 TCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTC
 ACGCAAGTTTAAATTACAGACCTGAA

- Ne contient que des lettres désignant la séquence (acides aminés ou ADN)
- Une seule séquence est représentée

5-1-6- Format PDB

```

HEADER      OXIDOREDUCTASE                27-OCT-03                IUR5
TITLE      STABILIZATION OF A TETRAMERIC MALATE DEHYDROGENASE BY
TITLE      2 INTRODUCTION OF A DISULFIDE BRIDGE AT THE DIMER/DIMER
TITLE      3 INTERFACE
COMPND     MOL_ID: 1;
COMPND     2 MOLECULE: MALATE DEHYDROGENASE; COMPND 3 CHAIN:A, C;
COMPND     4 EC: 1.1.1.37; COMPND 5 ENGINEERED:YES;
COMPND     6 MUTATION:YES
SOURCE     MOL_ID: 1;
SOURCE     2 ORGANISM_SCIENTIFIC: CHLOROFLEXUS AURANTIACUS;
SOURCE     3 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE     4 EXPRESSION_SYSTEM_STRAIN: DH5A
KEYWDS     OXIDOREDUCTASE, TRICARBOXYLIC ACID CYCLE,
KEYWDS     2 MALATE DEHYDROGENASE
EXPDTA     X-RAY DIFFRACTION
AUTHOR     A.BJORK,B.DALHUS,D.MANTZILAS,V.G.H.EIJSINK,R.SIREVAG

```

```

SEQRES    1 A  309 MET  ARG LYS  LYS  ILE  SER  ILE  ILE  GLY  ALA  GLY  PHE  VAL
SEQRES    2 A  309 GLY  SER  THR  THR  ALA  HIS  TRP  LEU  ALA  ALA  LYS  GLU  LEU
SEQRES    3 A  309 GLY  ASP  ILE  VAL  LEU  LEU  ASP  ILE  VAL  GLU  GLY  VAL  PRO
SEQRES    4 A  309 GLN  GLY  LYS  ALA  LEU  ASP  LEU  TYR  GLU  ALA  SER  PRO  ILE
SEQRES    5 A  309 GLU  GLY  PHE  ASP  VAL  ARG  VAL  THR  GLY  THR  ASN  ASN  TYR
SEQRES    6 A  309 ALA  ASP  THR  ALA  ASN  SER  ASP  VAL  ILE  VAL  VAL  THR  SER
...
SEQRES    1 C  309 MET  ARG LYS  LYS  ILE  SER  ILE  ILE  GLY  ALA  GLY  PHE  VAL
SEQRES    2 C  309 GLY  SER  THR  THR  ALA  HIS  TRP  LEU  ALA  ALA  LYS  GLU  LEU
SEQRES    3 C  309 GLY  ASP  ILE  VAL  LEU  LEU  ASP  ILE  VAL  GLU  GLY  VAL  PRO
SEQRES    4 C  309 GLN  GLY  LYS  ALA  LEU  ASP  LEU  TYR  GLU  ALA  SER  PRO  ILE
SEQRES    5 C  309 GLU  GLY  PHE  ASP  VAL  ARG  VAL  THR  GLY  THR  ASN  ASN  TYR
SEQRES    6 C  309 ALA  ASP  THR  ALA  ASN  SER  ASP  VAL  ILE  VAL  VAL  THR  SER
...

```

- Contient une seule séquence
- Commence par des lignes de descriptions. La séquence suit les lignes débutant par « SEQRES »
- Acides aminés codés par 3 lettres, acides nucléiques par DA, DC, DG, DT, DI

Annexe (Résumé du cours)

DEUX TYPES DE BANQUES

-Celles qui correspondent à une collecte des données **plus exhaustive** possible et qui offrent finalement un ensemble plutôt **hétérogène** d'informations.
-Traitent des thématiques générales

"Banques de données"

OU

Banques de données ou bases de données **GÉNÉRALISTES**

-Celles qui correspondent à des données **plus homogènes et spécifiques** .
-Traitent des thématiques particulières

"Bases de données",

OU

Banques de données ou bases de données **SPÉCIALISÉES**

Banque Nucléiques

Il existe trois banques nucléique internationales

(1) GenBank

la banque américaine gérée par le National Center for Biotechnology Information (NCBI)

(2) EMBL (European Molecular Biology Laboratory)

La banque européenne maintenue à l'**E**uropean **B**ioinformatic **I**nstitute (**E**BI)

(3) DDBJ

La banque japonaise ou DNA DataBase of Japan

Ces trois banques gèrent l'ensemble des séquences nucléique et leurs annotations : elles coopèrent et échange quotidiennement leurs données afin de garantir une cohérence maximale dans la mise à disposition des séquences de la communauté scientifique.

Ces séquences sont organisées dans les banque sous forme des **entrées** .

Ces trois banques (GenBank, DDBJ, EMBL) sont interconnectées⁶ (inter-reliées) du fait qu'elles échangent leurs informations. Il suffit de consulter le contenu d'une de ces 3 banques pour accéder au contenu de ces 3 banques en même temps.

Banque Protéiques (exemples)

Swissprot & TrEMBL⁷ : Elle a été constituée à l'Université de Genève à partir de 1986. Elle est maintenant développée par le SIB ([Swiss Institute of Bioinformatics](#)) et l'EBI. Elle regroupe (entre autres) des séquences annotées de la PIR-NBRF ainsi que les séquences codantes traduites de l'EMBL (TrEMBL⁸).

UniProt ("Universal Protein Resource") : c'est la base de données des protéines

II. Les banques spécialisées : elles regroupent des données plus homogènes établies autour d'une thématique ou d'une méthode spécifique de production des données.

Exemples de banques spécialisées : La base de données KEEG pathway (voies métaboliques), Flybase, Prosite (domaines des protéines), Pfam (proteins family), TRANSFAC, SWISS 2D PAGE,

Exemple : bases spécialisée pour un génome spécifique, bases de séquences immunologiques, de voies métaboliques, de cartes génétiques, de motifs protéiques, d'expression de gènes, de structures, ...

⁶ Ces banques s'échangent systématiquement leur contenu depuis 1987 et adoptent un système de conventions communes (The DDBJ/EMBL/GenBank Feature Table Definition).

⁷ SwissProt et TrEMBL sont toutes les deux des banques généralistes contenant des séquences protéiques. La différence réside dans le fait que les données introduites dans la banque de données SwissProt sont manuellement expertisées avec des ajouts de commentaire décrivant la fonction de la protéine, sa localisation cellulaire etc., et des annotations dans la partie feature de certaines caractéristiques comme la présence de fragments transmembranaires, de motifs, de domaines fonctionnels. Ces annotations peuvent être extraites de publications ou obtenu à partir d'analyses réalisées par les annotateurs.

TrEMBL contient les séquences protéiques obtenues par traduction automatique des CDS (régions codantes) des données présentes dans EMBL.

⁸ **Attention** : il faut distinguer entre EMBL et TrEMBL : EMBL est la banque de données européenne généraliste de séquences d'acides nucléiques maintenue à l'EBI. Les banques généralistes d'acides nucléiques contiennent toutes les séquences d'acides nucléiques produites dans les laboratoires publiques. TrEMBL est elle aussi une banque de données généraliste mais elle contient des séquences protéiques. Elle est construite par traduction automatique de toutes les CDS de la banque EMBL. Les CDS (Coding Sequence) correspondent aux régions codantes des gènes (du codon initiateur au codon stop).

Exemples de formats liés aux logiciels de traitement des séquences

1. Format FASTA

Sans doute le plus répandu et l'un des plus pratiques car très simple. 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 mises dans un même fichier.

Avec le format FASTA, un seul fichier peut contenir plusieurs enregistrements (séquences). Chaque enregistrement commence par ">".

```
>J00265.1 HUMINS01 Human insulin gene, complete Cds
CTCGAGGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCACACACCCTCCAGGCTTGACCGGCCAGGGT
GTCCCCTTCTTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGTCTGGGACACCAGCTGGC
CTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGTCTGGGATCCTGGATCTCAGCTCCCT
GGCCGACAACACTGGCAAACCTCCTACTCATCCACGAAGGCCCTCCTGGGCATGGTGGTCTTCCCAGC
CTGGCAGTCTGTTCTCACACACCTTGTAGTGCCAGCCCTGAGGTTGCAGCTGGGGGTGTCTCTG
AAGGGCTGTGAGCCCCCAGGAAGCCCTGGGGAAGTGCCTGCCTTGCCTCCCCCGGCCCTGCCAGCGC
CTGGCTCTGCCCTCCTACCTGGGCTCCCCCATCCAGCCTCCCTCCCTACACACTCCTCTCAAGGAGG
CACCCATGTCTCTCCAGCTGCCGGGCCCTCAGAGCACTGTGGCGTCTGGGGCAGCCACCGCATGTCC
TGCTGTGGCATGGCTCAGGGTGGAAAGGGCGGAAGGGAGGGGTCTGCAGATAGCTGGTGCCCACTAC
CAAACCCGCTCGGGGCAGGAGAGCCAAAGGCTGGGTGTGTGCAGAGCGGCCCCGAGAGGTTCCGAGGC
TGAGGCCAGGGTGGGACATAGGGATGCGAGGGGCCGGGGCACAGGATACTCCAACCTGCCTGCCCCCA
TGGTCTCATCCTCCTGCTTCTGGGACCTCCTGATCCTGCCCTGGTGTAAAGAGGCAGGTAAGGGGCT
GCAGGCAGCAGGGCTCGGAGCCCATGCCCCCTCACCATGGGTGAGGCTGGACCTCCAGGTGCCTGTTC
TGGGGAGCTGGGAGGGCCGGAGGGGTGTACCCAGGGGCTCAGCCAGATGACACTATGGGGGTGATG
GTGTCATGGGACCTGGCCAGGAGAGGGG
```

→ EMBL

```
50 Sequence 5028 BP; 1510 A; 1074 C; 835 G; 1099 T; 0 other;
gatctccat atacaecgt atctccact caggtttaga tctcaaac ggaaccattg 60
ccgacatgag acagttagg atctctgaga gttacaagct aaaacgagca gtagtcaqct 120
ctgcatctga agccgctgaa gtctactaa gggggataa catcatcctg gcaagaccaa 180
gaaccgcaa tagacaacat atgtaacata tttaggat atctcgaaa taataaaccg 240
ccacactgtc attattataa ttagaacag aacgcaaaa ttatccacta tataattcaa 300
agacgcaaaa aaaaaagaac aacgctcat agaactttg gcaattcgg tcacaataa 360
atcttgcaa ettatgttc ctctcgagc agtactegag cctgtctca agaattgaa 420
aataccatc gtaggtatg ttaaagatg catctccaca acctcaaac tcttgccga 480
gagtcgccc cctttgtga gtaatttca ctttccat gagaactat tttcttattc 540
```

→ GenBank

```
ORIGIN
1 gatctccat atacaecgt atctccact caggtttaga tctcaaac ggaaccattg
61 ccgacatgag acagttagg atctctgaga gttacaagct aaaacgagca gtagtcaqct
121 ctgcatctga agccgctgaa gtctactaa gggggataa catcatcctg gcaagaccaa
181 gaaccgcaa tagacaacat atgtaacata tttaggat atctcgaaa taataaaccg
241 ccacactgtc attattataa ttagaacag aacgcaaaa ttatccacta tataattcaa
301 agacgcaaaa aaaaaagaac aacgctcat agaactttg gcaattcgg tcacaataa
361 atcttgcaa ettatgttc ctctcgagc agtactegag cctgtctca agaattgaa
421 aataccatc gtaggtatg ttaaagatg catctccaca acctcaaac tcttgccga
481 gagtcgccc cctttgtga gtaatttca ctttccat gagaactat tttcttattc
541 tttctctca catctgtg taggtgac tcacaagcc acctccacta gaagaacaga
```