République Algérienne Démocratique et populaire

Université Batna 2

Cours de Bioinformatique
(M1 Biotechnologie végétale + M1 EDP)

Cours 1 :

Les Banques de séquences biologiques

Réaliser par:

Mr. Ghedadba N

Contenu de la matière

Chapitre 1- Les banques de séquences biologiques

Chapitre 2- Alignement de séquences biologiques

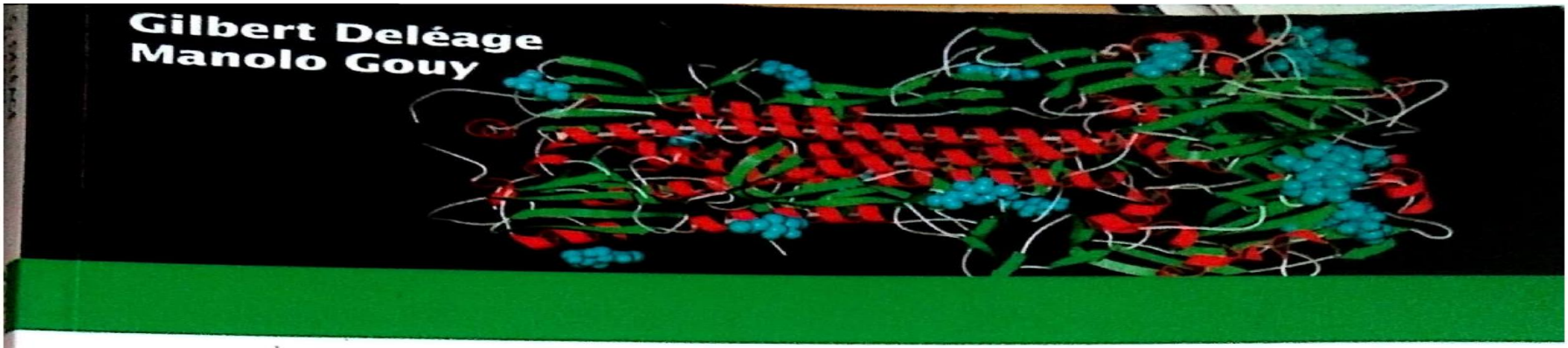
Chapitre 3- La phylogénie moléculaire et l'annotation des génomes

Objectif

- Initier les étudiants aux problématiques bio-informatiques liées à l'émergence des nouvelles biotechnologies.

- Donner aux étudiants la connaissance et les moyens pour utiliser les logiciels existants sur le Web qui permettent déjà de traiter de manière puissante les données biologiques générées par les nouvelles biotechnologies (bases de données, logiciels de traitement de séquence, logiciels statistiques).

Références



Gilbert Deléage
Manolo Gouy

Bioinformatique

Cours et cas pratique

Avec ce livre, des bonus sur le web

- Licence 3
- Master
- Écoles d'ingénieurs

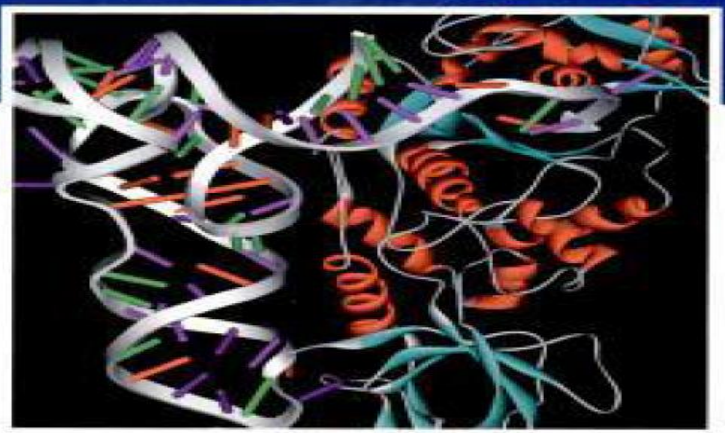
DUNOD

BIOSCIENCES ET TECHNIQUES

Collection dirigée par J. Figarella et A. Calas

3e édition

G. Coutouly, E. Klein, E. Barbieri,
M. Kriat

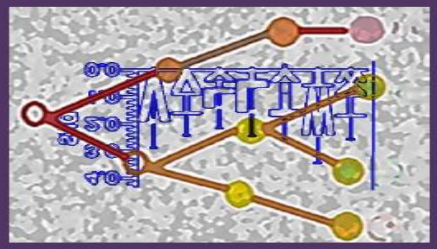


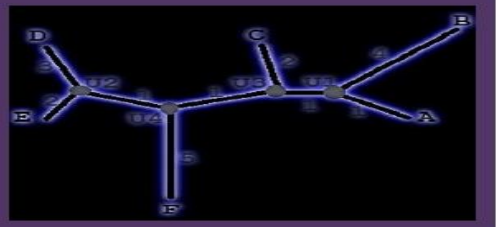
Travaux dirigés de biochimie, biologie moléculaire et bioinformatique

Cours de phylogénie moléculaire

Distances et constructions phylogénétiques

Support pédagogique de phylogénie moléculaire destiné aux étudiants du système LMD de Master ‘M1 et M2) et doctorants de Biotechnologie végétale, Biochimie et Microbiologie.





UNIVERSITE
CONSTANTINE 1

Faculté des Sciences de

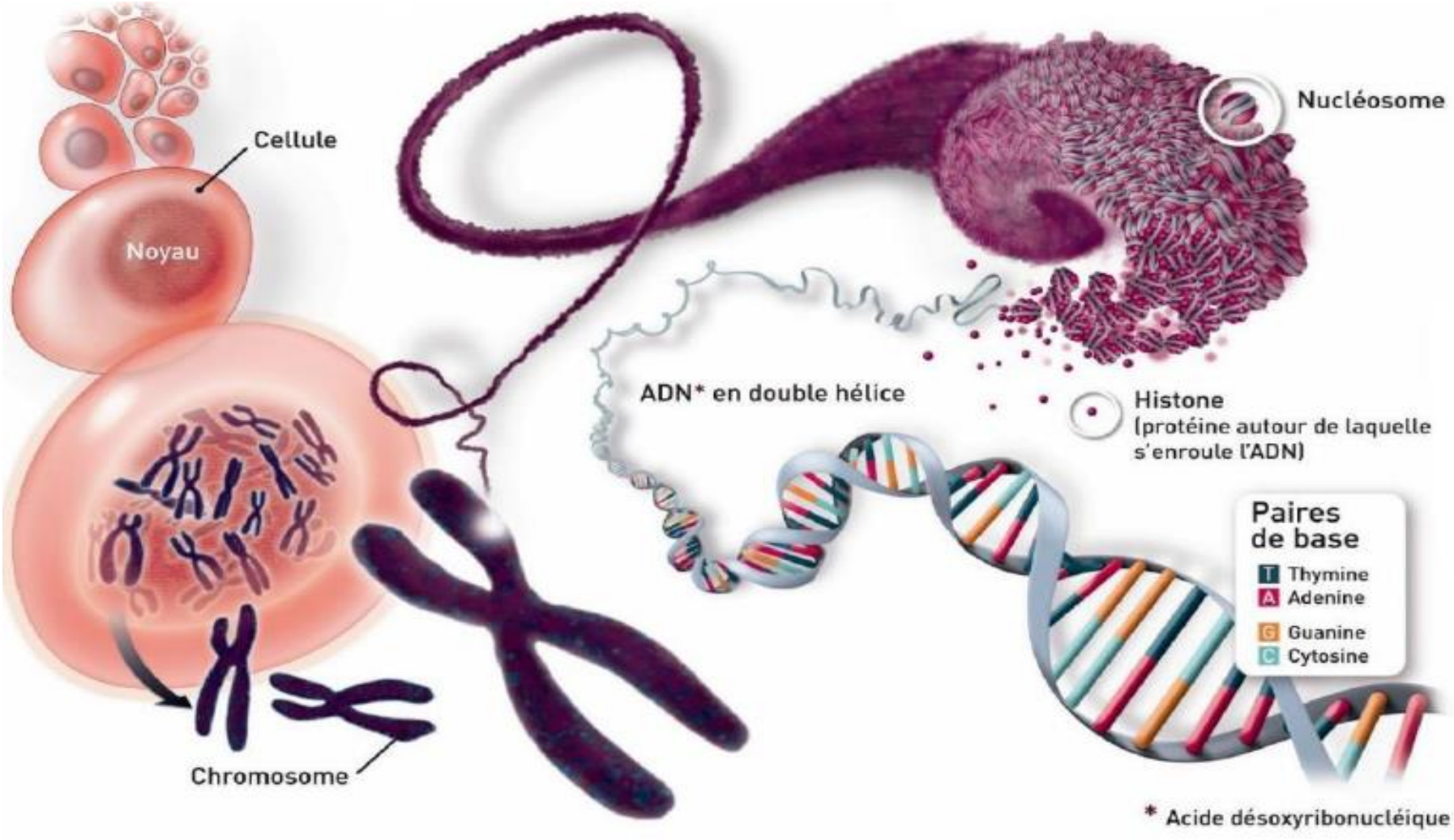
la Nature et de la Vie

Pr DJEKOUN A.

Pr HAMIDECHI M. A.

Dogme central de la biologie moléculaire

Le génome: c'est-à-dire l'ensemble des gènes d'un organisme, est structurellement défini par les chromosomes au sein de chaque cellule, selon la théorie chromosomique de l'hérédité. Chaque chromosome est constitué d'une molécule d'ADN (Figure 1), support physique des gènes, et de protéines associées



L'ADN

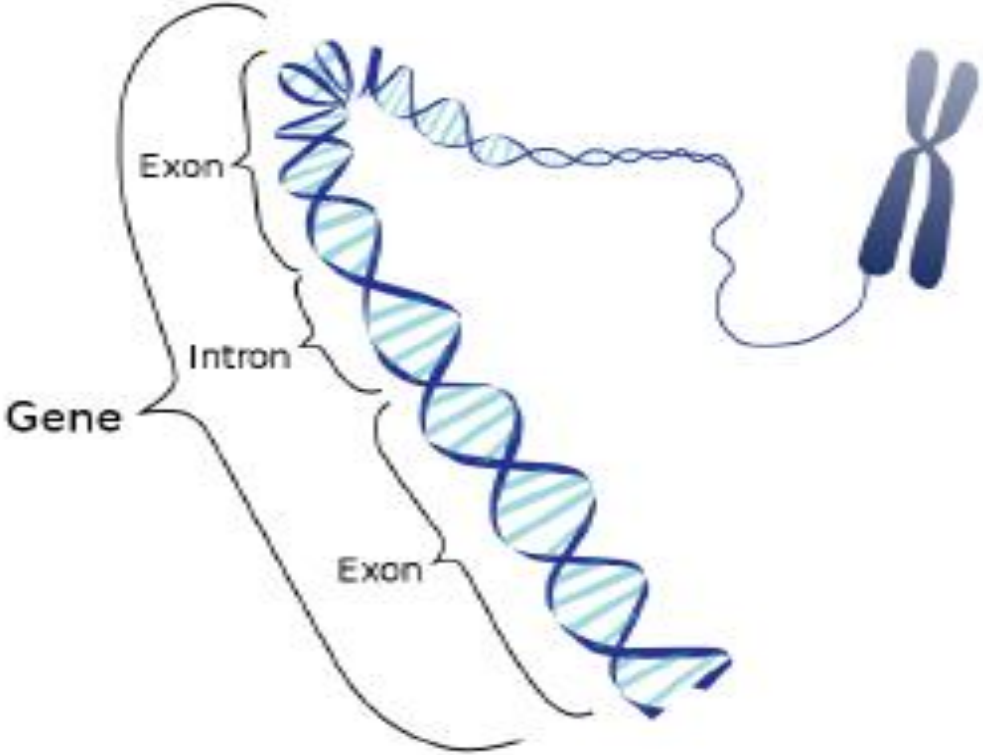
- L'ADN est le **support de l'information génétique.**
- L'ADN est une **longue molécule, faite de deux brins s'enroulant en une double hélice.**
- Les deux brins de la double hélice suggèrent un mécanisme de réplication de l'ADN
- Chaque brin est le support d'une **succession de nucléotides**
- **Quatre types de nucléotides : (Adénine, Cytosine, Guanine, Thymine).**

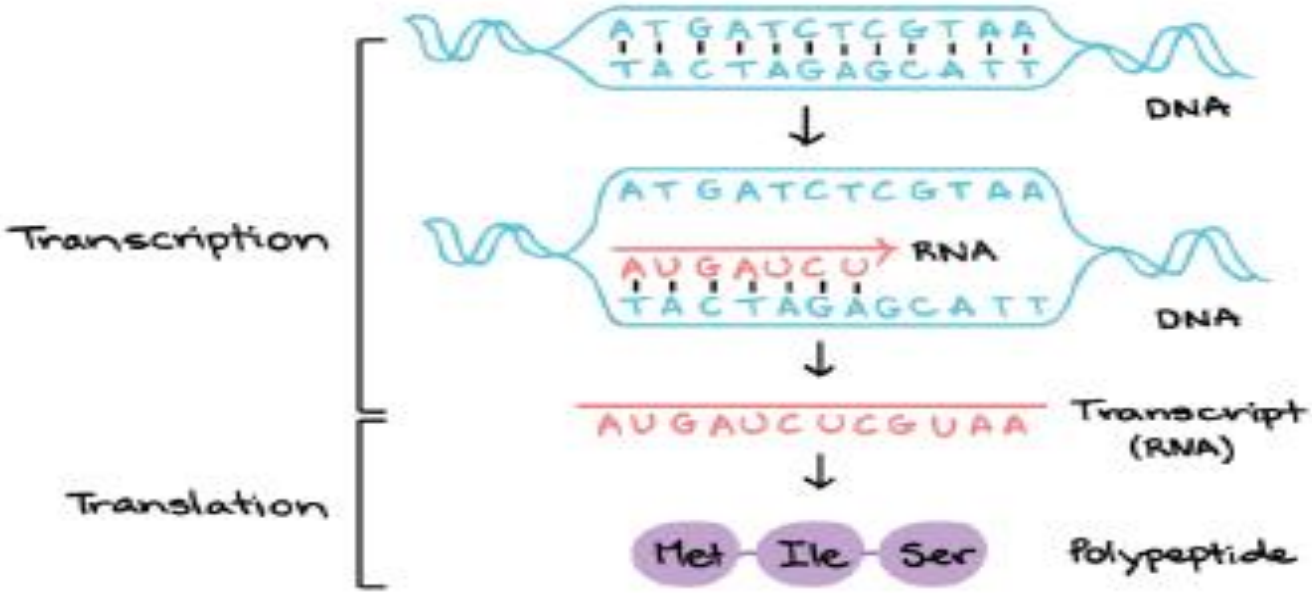
La structure de l'ADN représenté par la figure 2, les deux brins sont orientés, on parle de l'orientation 5' → 3' par des considérations biochimiques.

- L'extrémité 5' se termine par un groupement phosphate.
- L'extrémité 3' se termine par un groupement hydroxyle.

Il faut retenir qu'il y a un brin sens orienté 5' → 3' et un brin anti-sens orienté 3' → 5', et généralement lorsqu'on veut connaitre la séquence des nucléotides le long d'un brin, on va la lire traditionnellement dans le sens 5' → 3'.

Figure 2. Représentation schématique d'une succession de nucléotides d'un fragment d'ADN double brins.


Exon
Intron
Gene
Exon


ATGATCTCGTAA
TACTAGAGCATT
DNA
ATGATCTCGTAA
AUGAUCU
RNA
TACTAGAGCATT
DNA
Transcription
AUGAUCUCGUAA
Transcript (RNA)
Translation
Met
Ile
Ser
Polypeptide

- Le texte génomique est écrit dans un alphabet de 4 lettres : A, C, G, T

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)

I. Introduction à la Bioinformatique

I. Définition de la Bio-informatique

- C’est un traitement de **l’information Biologique** stockée sous forme de données accessibles aisément et exploitable

- 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)

I. Définition moderne de la Bio-informatique

- Est une multi discipline (discipline hybride) théorique de l'analyse *in Silico* de l'information biologique contenu dans les séquences nucléiques (ADN et ARN) et protéiques

La Bio-information

- La Bio-informatique s'intéresse aux données :

- **Le génome**: est la totalité de l'ADN d'un individu ou d'une espèce.
- **Le transcriptome**: est l'ensemble des ARN messagers transcrits à partir du génome.
- Le **protéome** (l'ensemble des protéines biosynthétisées).
- le **sécrétome** (l'ensemble des protéines sécrétées)
- le **métabolome:** l 'ensemble des métabolites produits par une cellule (molécules organiques qui ne sont ni ADN, ni protéines

Domaines de la Bio-informatique (Applications)

- **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 fonctionnelle :** Transcriptome, protéome,

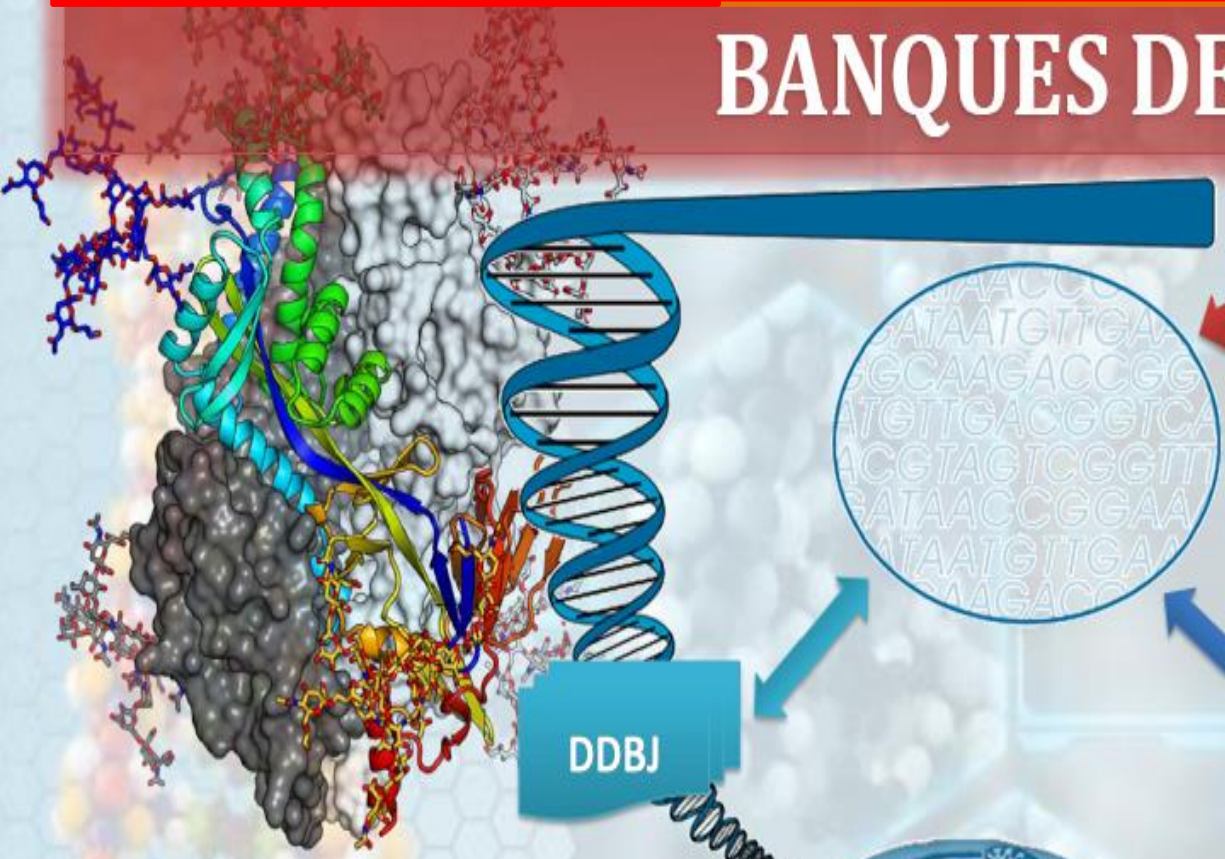
- Phylogénie : Relations évolutives entre gènes, entre génomes, entre organismes ; Inférence de scénarios évolutifs.

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

CHAPITRE I :

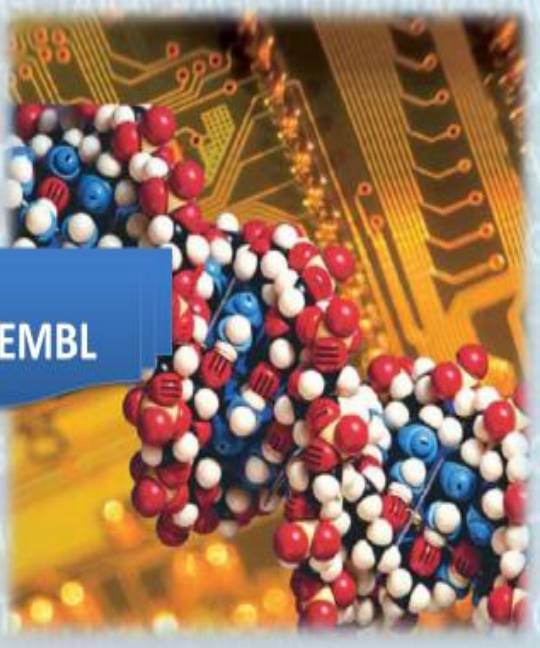
BANQUES DE DONNEES



GenBank

DDBJ

EMBL



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. Contenues 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- Types de Bases des données

- ***Banques de données généralistes :***

- Celles qui correspondent à une collecte des données la plus exhaustive possible et la plus large possible.

- ***Banque ou base de données spécialistes :***

– Celles qui correspondent à des données plus homogènes établies autour d'une thématique

1. Banques de séquences généralistes

– Données globales (pas de focus sur une application ou organisme particulier) informations hétérogènes

- Banques de séquences nucléiques (ADN et ARN)
- Banques de séquences protéiques
- Banques de structure 3D de macromolécules
- Banques d'articles scientifiques (Bibliographique)

- **Avantage :** tout est consultable en une fois
- **Inconvénients :** difficiles à maintenir, difficiles à interroger, problèmes de redondances

1.1. Les banques de séquences nucléiques

- **Origine des données**

- Séquençage d'ADN et d'ARN

- 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)

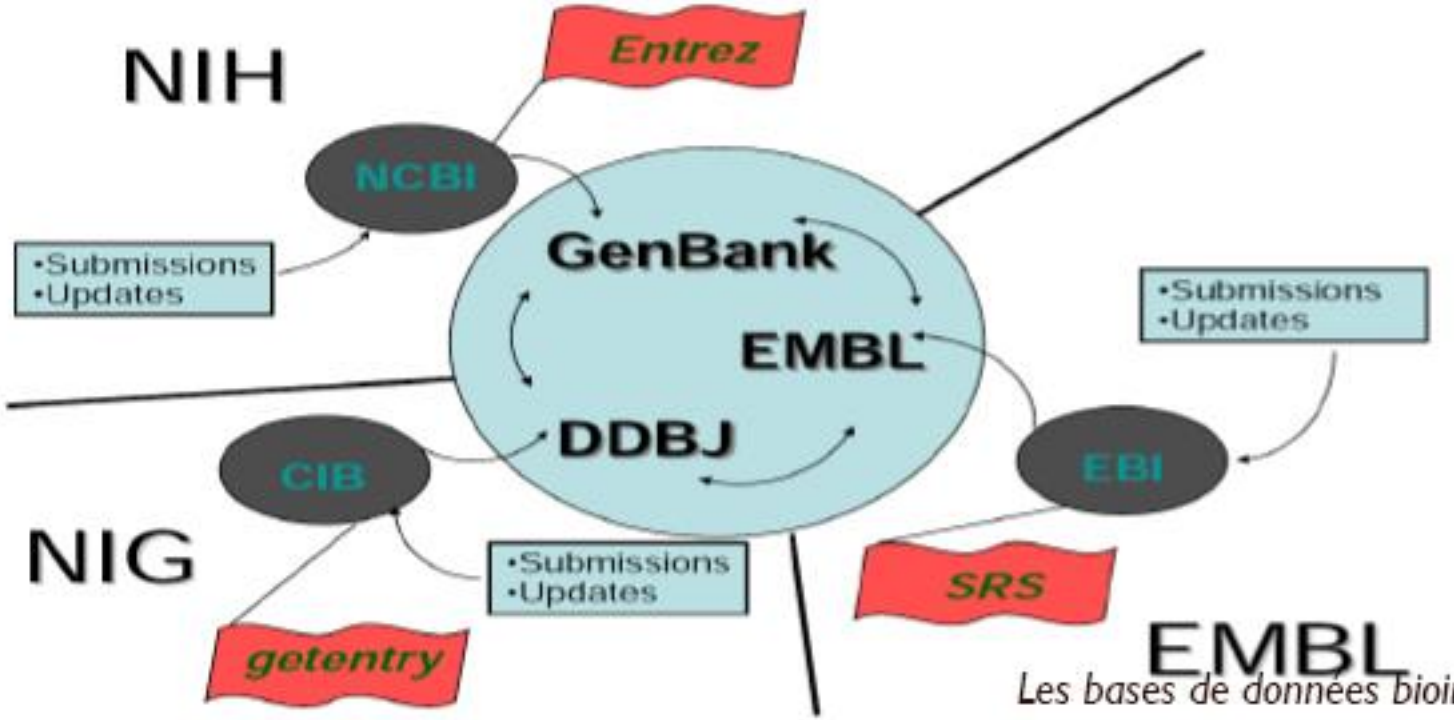
[Remarque 1] : toutes les séquences (ADN ou ARN) sont écrites avec des T

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

- Il existe sur internet 3 principales banques qui sont interconnectés

USA: ***GenBank*** (NCBI) https://www.ncbi.nlm.nih.gov/genbank/
Europe: ***ENA*** *(EMBL-EBI)* https://www.ebi.ac.uk/ena
Japon : ***DDBJ*** http://www.ddbj.nig.ac.jp/index-e.html

- 1988 Regroupement de ces 3 groupes (appelés International Nucleotide Sequence Database Collaboration, INSDC)
  - ✓ Accord pour un format commun
  - ✓ Echange des données journalier
  - ✓ Chaque groupe gère les mises a jour des séquences qu'il a créées.



Genbank

http://www.ncbi.nlm.nih.gov/Genbank/

NCBI

GenBank Overview

PubMed | Entrez | BLAST | OMIM | Books | Taxonomy | Structure

Search Entrez for Go

- NCBI Home
- NCBI Site Map
- Submit to GenBank
- Submit an update
- Search GenBank
- GenBank and RefSeq: a comparison
- BLAST

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2008 Jan;36(Database issue):D25-30). There are approximately 85,759,586,764 bases in 82,853,685 sequence records in the traditional GenBank divisions and 108,635,736,141 bases in 27,439,206 sequence records in the WGS division as of February 2008.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank record may be viewed for a *Saccharomyces cerevisiae* gene.

In The News: Platypus Genome

Explore Platypus Genome resources.

- Platypus Genome Project
- Platypus Taxonomic and Sequence Resources
- Platypus Genome Resource Guide
- Duck-Billed Platypus Genome Sequence Published (NIH Press Release)


European Bioinformatics Institute [GB] | https://www.ebi.ac.uk
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
EMBL-EBI
18% webpage
Usage via web interfaces in 2016.
The home for big data in biology
More about EMBL-EBI's impact in our annual report
Data from 2016
Our unique Search service helps you explore dozens of biological data resources.
More about EBI Search
Find a tool for your data analysis.
Share your scientific data with the world.
Deposit data
All
map kinase
Find a tool
Example searches: blast keratin bfl1
We are EMBL-EBI
The European Bioinformatics Institute (EMBL-EBI) is part of EMBL, Europe's flagship laboratory for the life sciences. More about EMBL-EBI and our impact.
Data resources
Explore our open data resources to enrich your research. Browse data, perform analyses or share your own results.
Research
Find out about our research groups, postdoctoral schemes and PhD Programme
This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use.
I agree, dismiss this banner


Serveur introuvable - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Adresse http://www.ddbj.nig.ac.jp/searches-e.html
Messenger
OK
Liens
DDBJ
DNA Data Bank of Japan
Accession DNA Protein Taxonomy Site Search
Accession numbers
Go
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)
Termination of providing SRS (Sequence Retrieval System) services (2008.8.15)
Database Search
getentry
Data retrieval by accession numbers, etc.
ARSA
All-round Retrieval of Sequence and Annotation
SRS
Data retrieval by key words
Terminated the service as of Dec. 26, 2008.
TXSearch
Retrieval of unified taxonomy database
Homology Search
FASTA | BLAST | PSI-BLAST | SSEARCH
HMMPFAM
A motif search program against Pfam on the basis of Hidden Markov Model (HMM)
DDBJ Vector Screening System
SQmatch
Matching a regular expression against sequences in Database
Genome Analyses
GIB
Genome information broker
GIB-V
GIB for Viruses
GTPS
Reannotation of bacterial genomes using a new common protocol
GTOP
Genome to protein structure and function
H-Invitational Database CIB-DDBJ Flat File Server
suspended for the system maintenance
Mirroring H-InvDB
Database of full-length cDNAs assigned functional annotation as a result of H-Invitational
Protein Database and Structure
PMD
Protein mutant database
Internet

Banques de données de séquences protéiques

Origine des informations sur les protéines

- **Traduction de séquences d'ADN**

  * **TrEMBL traduction automatique de EMBL**

    * **Genpept traduction automatique de GenBank**

- **Séquençage de protéines** (Rare car long et coûteux)

- **Protéines dont la structure 3D est connue**

• Les données stockées : séquences + annotations

- Protéines entières
- Fragments de protéines

Les banques de séquences protéiques

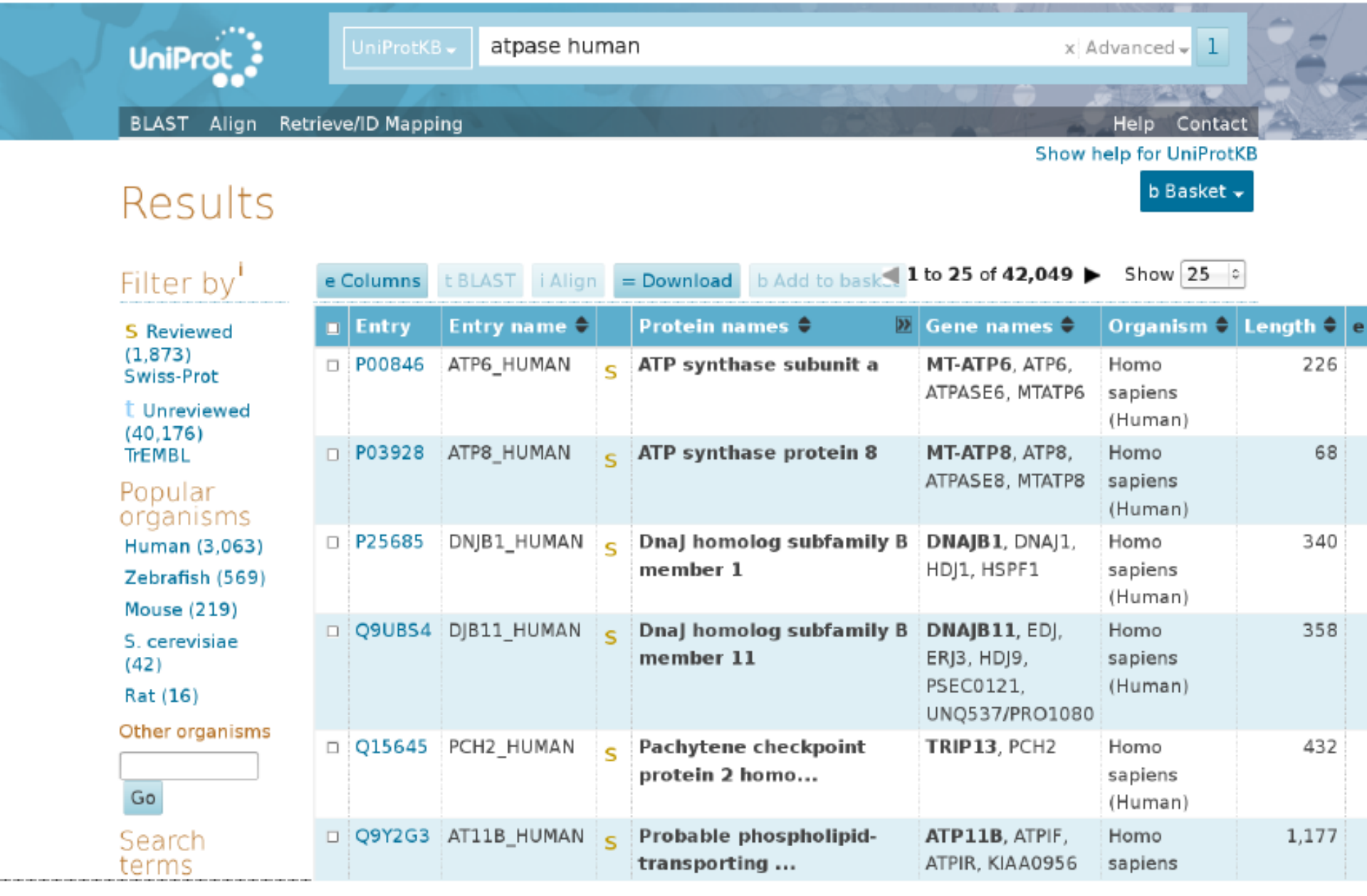
Swiss-Prot + PIR + TrEMBL-EBI

UniProt

(Universal Protein Ressource)

http://www.uniprot.org/

Uniprot


UniProt
UniProtKB
atpase human
Advanced
BLAST
Align
Retrieve/ID Mapping
Help
Contact
Show help for UniProtKB
Basket
Results
Filter by
Reviewed (1,873) Swiss-Prot
Unreviewed (40,176) TrEMBL
Popular organisms
Human (3,063)
Zebrafish (569)
Mouse (219)
S. cerevisiae (42)
Rat (16)
Other organisms
Go
Search terms
Columns
BLAST
Align
Download
Add to basket
1 to 25 of 42,049
Show 25
Entry
Entry name
Protein names
Gene names
Organism
Length
P00846 ATP6_HUMAN ATP synthase subunit a MT-ATP6, ATP6, ATPASE6, MTATP6 Homo sapiens (Human) 226
P03928 ATP8_HUMAN ATP synthase protein 8 MT-ATP8, ATP8, ATPASE8, MTATP8 Homo sapiens (Human) 68
P25685 DNJB1_HUMAN DnaJ homolog subfamily B member 1 DNAJB1, DNAJ1, HDJ1, HSPF1 Homo sapiens (Human) 340
Q9UBS4 DJB11_HUMAN DnaJ homolog subfamily B member 11 DNAJB11, EDJ, ERJ3, HDJ9, PSEC0121, UNQ537/PRO1080 Homo sapiens (Human) 358
Q15645 PCH2_HUMAN Pachytene checkpoint protein 2 homo... TRIP13, PCH2 Homo sapiens (Human) 432
Q9Y2G3 AT11B_HUMAN Probable phospholipid-transporting ... ATP11B, ATPIF, ATPIR, KIAA0956 Homo sapiens 1,177

Protein **Sodium/potassium-transporting ATPase subunit alpha-1**

Gene **ATP1A1**

Organism *Homo sapiens (Human)*

Status S Reviewed - ●●●●● - Experimental evidence at protein level[i]

Display None

FUNCTION
NAMES & TAXON
SUBCELL. LOCAT
PATHOL./BIOTECH
PTM / PROCESSI
EXPRESSION
INTERACTION
STRUCTURE
FAMILY & DOMAI
SEQUENCES (4)
CROSS-REFEREN
PUBLICATIONS
ENTRY INFORMA
MISCELLANEOUS

Function[i]

This is the catalytic component of the active enzyme, which catalyzes the hydrolysis of ATP coupled with the exchange of sodium and potassium ions across the plasma membrane. This action creates the electrochemical gradient of sodium and potassium ions, providing the energy for active transport of various nutrients.

Catalytic activity[i]

$ATP + H_2O + Na^+(In) + K^+(Out) = ADP + phosphate + Na^+(Out) + K^+(In)$.

Sites

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Active site[i]	376 - 376	1	4-aspartylphosphate intermediate By similarity			
Binding site[i]	487 - 487	1	ATP By similarity			
Metal binding[i]	717 - 717	1	Magnesium By similarity			
Metal binding[i]	721 - 721	1	Magnesium			

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 .

PDB

- PDB (BrookHaven Protein DataBank)
  - http://www.rcsb.org
  - **Séquences et structures des protéines**
  - Visualisation en 3D
  - Les données proviennent de cristallographie, de RMN,...
  - Pour certaines protéines, plusieurs structures sont disponibles
    - Structure de la protéine seule ou avec ligand
    - Structure de la protéine dans différents milieux
    - Structure obtenue avec des méthodes expérimentales différentes

Home Hide
News & Publications
Policies
FAQ
Contact
Feedback
About Us

Deposition Hide
All Deposit Services
Electron Microscopy
NMR
Validation Server
BioSync Beamline
Related Tools

Search Hide
Advanced Search
Latest Release
Latest Publications
Sequence Search
Ligand Search
Unreleased Entries
Browse Database
Histograms

Explorer:

315 Structure Hits | 9 Unreleased Structures | 158 Citations | 231 Ligand Hits | 107 Web Page Hits | GO Hits | SCOP Hits | CATH Hits

Advanced Keyword Query for: MAP KINASE

Query Options:

Display/Download: Generate Reports:

Sort by: Results per Page:

Displaying results 1 - 10 of 315 total | Page 1 of 32

3FXW — **High resolution crystal structure of mitogen-activated protein kinase-activated protein kinase 3/inhibitor 2 complex**

Characteristics Release Date: 26-Jan-2010 Exp. Method: X-RAY DIFFRACTION
Resolution: 2.00 Å

Classification Transferase

Compound
Molecule: MAP kinase-activated protein kinase 3
Polymer: 1 **Type:** polypeptide(L) **Length:** 336
Chains: A
EC#: 2.7.11.1
Fragment: Kinase domain, UNP residues 33-349

Authors Cheng, R.K.Y., Barker, J., Palan, S., Felicetti, B., Whittaker, M., Hesterkamp, T.

3KGA — **Crystal structure of MAPKAP kinase 2 (MK2) complexed with a potent 3-aminopyrazole ATP site inhibitor**

Summa | **Sequence** | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

Crystal structure of MAPKAP kinase 2 (MK2) complexed with a potent 3-aminopyrazole ATP site inhibitor

3KGA

Display Files | Download Files | Share this Page

Sequence Display

The sequence display provides a graphical representation of the UniProtKB, PDB - ATOM and PDB - SEQRES sequences. Different 3rd party annotations can be graphically mapped on the sequence and displayed in the Jmol viewer.

The structure **3KGA** has in total **1** chains.

Currently viewing **unique chains** only. show all chains

Sequence & Structure Relationships

Display Jmol

Enable Jmol to view annotations in 3D.

Chain A : MAP kinase-activated protein kinase 2

FASTA | Sequence & DSSP | Image

Polymer 1
Length: 299 residues
Chain Type: polypeptide(L)
Reference: UniProtKB P49137

Display Parameters

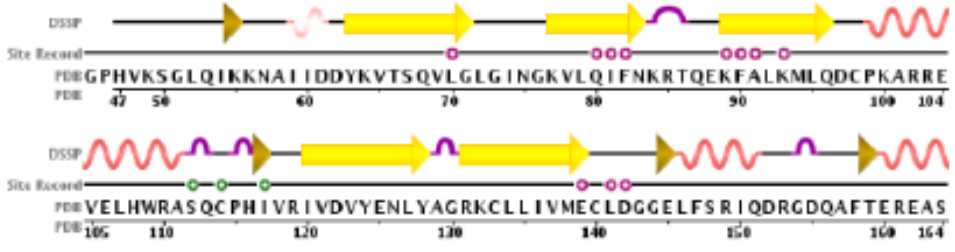
Currently displayed: **SEQRES sequence.**
Display external (UniProtKB) sequence

Mouse over an annotation to see more details. Click annotation to enable Jmol.

Annotations *Add Annotations* Select

Secondary Structure:**DSSP** [hide] [reference]
38% helical (15 helices; 116 residues)
19% beta sheet (14 strands; 57 residues)

Structural Feature:**Site Record** [hide] [reference]
3KGA_A_AC2_16 BINDING SITE FOR RESIDUE LX9 A 365 (SOFTWARE)
3KGA_A_AC1_4 BINDING SITE FOR RESIDUE MG A 1 (SOFTWARE)



PDB : structure secondaires

PDB : séquence des protéines

Summary | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

Crystal structure of MAPKAP kinase 2 (MK2) complexed with a potent 3-aminopyrazole ATP site inhibitor

3KGA

Display Files
Download Files
Share this Page

Sequence Display

The sequence display provides a graphical representation of the UniProtKB, PDB - ATOM and PDB - SEQRES sequences. Different 3rd party annotations can be graphically mapped on the sequence and displayed in the Jmol viewer.

The structure **3KGA** has in total **1** chains.

Currently viewing **unique chains** only. **show all chains**

Sequence & Structure Relationships

Display Jmol

Enable Jmol to view annotations in 3D.

Chain A : MAP kinase-activated protein kinase 2

FASTA | Sequence & DSSP | Image

Polymer: 1
Length: 299 residues
Chain Type: polypeptide(L)
Reference: **UniProtKB** P49137

Display Parameters

Currently displayed: **SEQRES sequence.**
Display external (UniProtKB) sequence

3KGA_A.fasta.txt [Lecture seule] (/tmp) - gedit

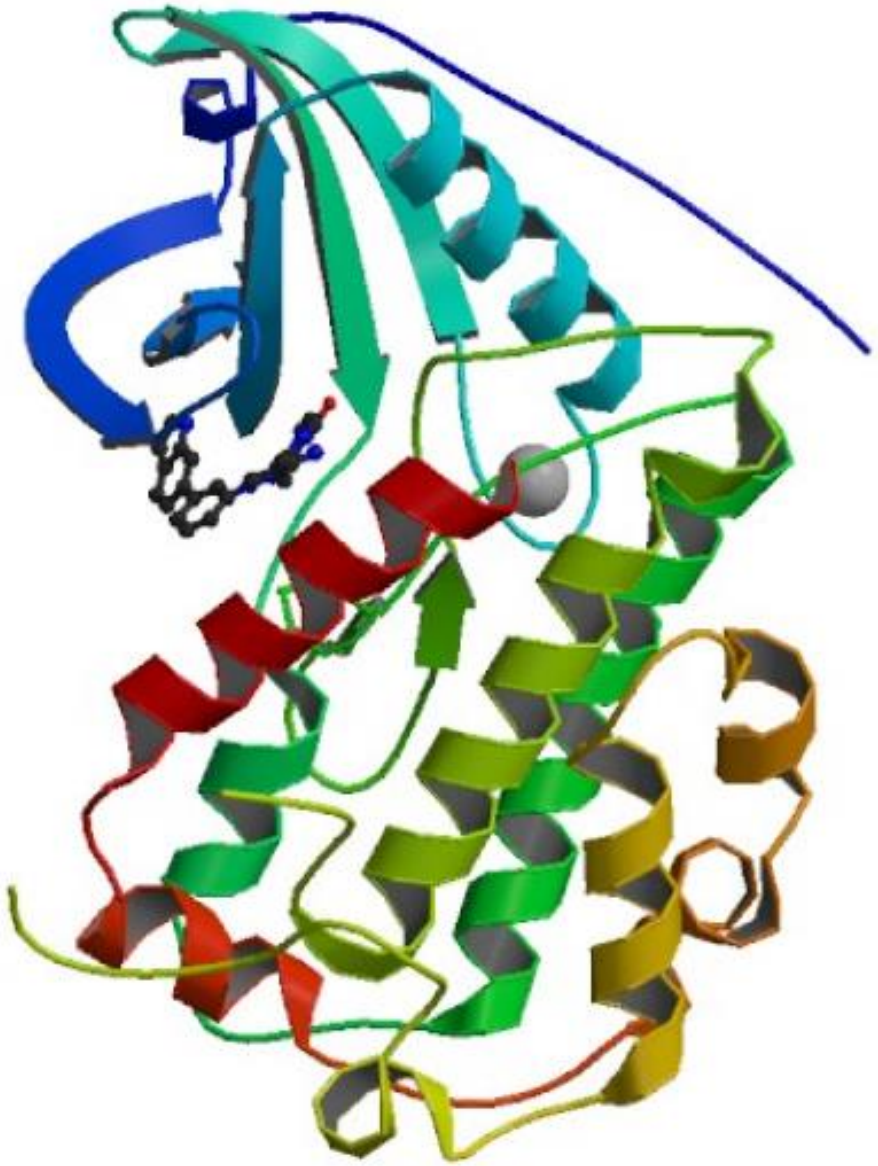
Fichier Édition Affichage Rechercher Outils Documents Aide

Ouvrir | Enregistrer | Annuler

3KGA_A.fasta.txt

```
>3KGA:A|PDBID|CHAIN|SEQUENCE
GPHVKSGLQIKKNAIIDDYKVTSQVLGLGINGKVLQIFNKRTQEKFALKMLQDCPKARREVELHWRASQCPHIVRIVDVY
ENLYAGRKCLLIVMECLDGGELFSRIQDRGDQAFTEREASEIMKSIGEAIQYLHSINIAHRDVKPENLLYTSKRPNAILK
LTDFGFAKETTGEKYDKSCDMWSLGVIMYILLCGYPPFYSNHGLAISPGMKTRIRMGQYEFPNPEWSEVSEEVKMLIRNL
LKTEPTQRMTITEFMNHPWIMQSTKVPQTPLHTSRVLKEDKERWEDVKEEMTSALATMR
```

PDB : structure tertiaires



Biological Assembly Image for 3KGA
Crystal structure of MAPKAP kinase 2 (MK2) complexed with a potent 3-aminopyrazole ATP site inhibitor

PDB 3KGA: Crystal structure of MAPKAP kinase 2 (MK2) complexed with a potent 3-aminopyrazole ATP site inhibitor

Velcicky, J., Feifel, R., Hawtin, S., Heng, R., Huppertz, C., et al.
Bioorg.Med.Chem.Lett. (2009)

Crystal structure of MAPKAP kinase 2 (MK2) complexed with a potent 3-aminopyrazole ATP site inhibitor

3KGA

Primary Citation and Related Literature

Primary Citation

Novel 3-aminopyrazole inhibitors of MK-2 discovered by scaffold hopping strategy.

Velcicky, J., Feifel, R., Hawtin, S., Heng, R., Huppertz, C., Koch, G., Kroemer, M., Moebitz, H., Revesz, L., Scheufler, C., Schlapbach, A.

(2009) Bioorg.Med.Chem.Lett.

PubMed: 20060294

DOI: 10.1016/j.bmcl.2009.10.138

Search Related Articles in PubMed

PubMed Abstract:

New, selective 3-aminopyrazole based MK2-inhibitors were discovered by scaffold hopping strategy. The new derivatives proved to inhibit intracellular phosphorylation of hsp27 as well as LPS-induced TNFalpha release in cells. In addition, selected derivative 14e also inhibited LPS-induced TNFalpha release in ... [Read More & Search PubMed Abstracts]

Publication Details

MeSH Terms (Primary Citation)

Literature Network

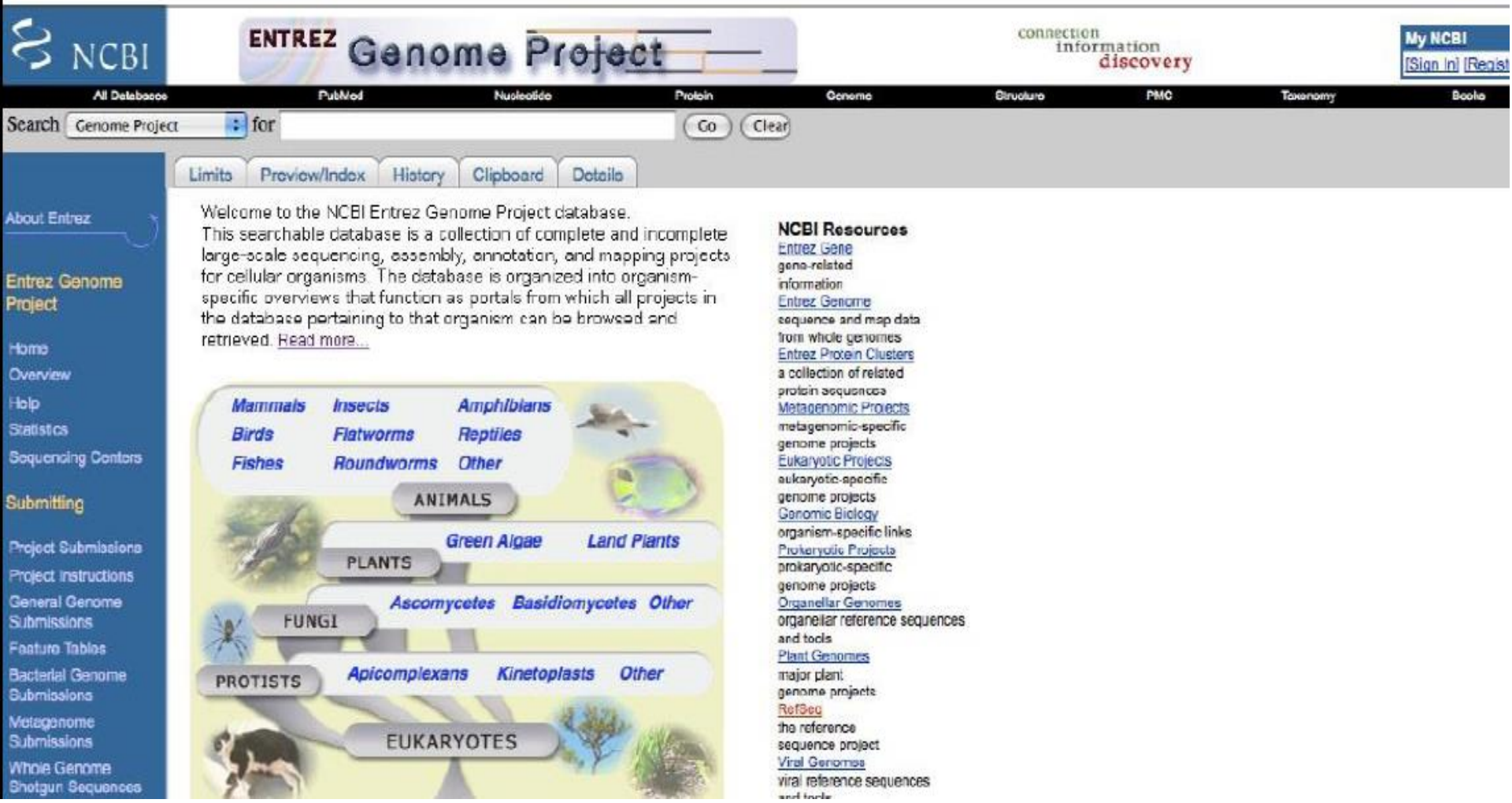
iHOP: 9261 GeneRIF: 9261 ;

Information provided by **BioLit**:

PubMedCentral articles found to contain 3KGA

No related articles could be found in PubMedCentral.

Bases des données sur les genomes



Banques de données spécialisées

Les banques de données spécialisées

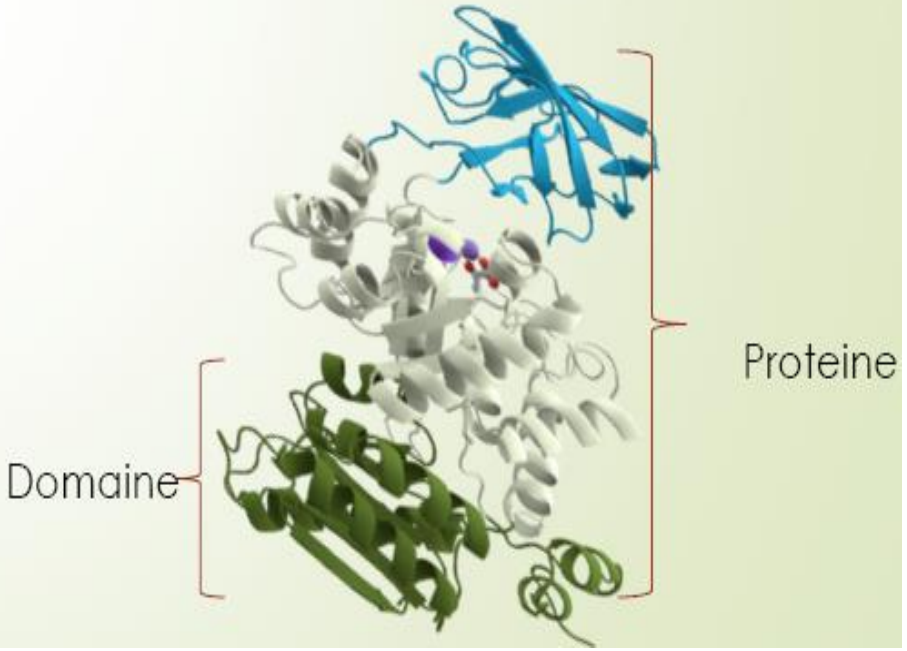
- Ces banques contiennent des données **homogènes**
  - Collecte, établie autour d'une thématique particulière
- **Avantages :** facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée, …
- **Inconvénients :** ne cible pas toujours ce que l'on veut; toutes les banques possibles n'existent pas
- **Exemples :** banques spécialisées pour un génome, banques de séquences d'immunologies, banques sur des séquences validées, …

Quelques examples de Bases de données spécialisés

1- Base de donnés PROSITE : http://www.expasy.ch/prosite/

définition :base de données sur les domaines des protéines, les familles protéiques et les fonctions biologiques associées .

➔ Un domaine = une région d'une protéine ayant une fonction biologique propre, que l'on retrouve sur plusieurs protéines.



1- Base de donnés PROSITE : http://www.expasy.ch/prosite/

Database of protein domains, families and functional sites

New SARS-CoV-2 relevant PROSITE motifs

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users].
PROSITE is complemented by ProRule , a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 2021_04 of 29-Sep-2021 contains 1895 documentation entries, 1311 patterns, 1326 profiles and 1338 ProRule.

Search

e.g. PDOC00022, PS50089, SH3, zinc finger

Search

Browse

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

Quick Scan mode of ScanProsite

Other tools

PROSITE

Search PROSITE for Go Clear

ExPASy Proteomics Server

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH > Databases > PROSITE

Search in PROSITE for: atpase

(Release 20.60, of 09-Feb-2010)

Enter search terms:

atpase

☐ Prefix and append wildcard '*' to words.

new search clear

By default, this search engine searches for complete words only. If you did not find what you expected, and would try to do a substring match, you should perform a new search and select 'prefix and append wildcard to words'.

Number of documents in PROSITE containing the search term:47

- PDOC00572 AAA-protein family signature
- PDOC00364 ABC transporter integral membrane type-1 domain profiles
- PDOC00420 ATP synthase a subunit signature
- PDOC00137 ATP synthase alpha and beta subunits signature
- PDOC00526 ATP synthase c subunit signature
- PDOC00327 ATP synthase delta (OSCP) subunit signature
- PDOC00138 ATP synthase gamma subunit signature
- PDOC00185 ATP-binding cassette, ABC transporter-type, signature and profile
- PDOC00017 ATP/GTP-binding site motif A (P-loop)
- PDOC00661 ArsR-type HTH domain signature and profile
- PDOC00610 Chaperonins TCP-1 signatures
- PDOC00576 Chaperonins cpn10 signature
- PDOC00268 Chaperonins cpn60 signature
- PDOC00196 Clathrin light chains signatures
- PDOC51413 DBINO domain profile

Identifiants :

PDOC... => domaine

P... => protéine

PROSITE

ABC transporter integral membrane type-1 domain profiles

Description:

ABC transporters belong to the ATP-Binding Cassette (ABC) superfamily which uses the hydrolysis of ATP to energize diverse biological import and export systems (see <PDOC00185>). ABC transporters are minimally constituted of two conserved regions: a highly conserved ATP binding cassette (ABC) and a less conserved transmembrane domain (TMD). These regions can be found on the same protein (mostly in eukaryotes and bacterial exporters) or on two different ones (mostly bacterial importers) [1,2,3]. The function of the integral inner-membrane protein is to translocate the substrate across the membrane. Studies of P-glycoprotein function indicate that residues lining the proposed chamber opening (residues of TM2, TM5 and TM6) play an important role in substrate recognition [4].

In exporters and eukaryotes, ABC transporters consist of a single polypeptide composed of an N-terminal domain of approximately 320 residues, apparently containing six transmembrane segments, fused to a highly conserved ABC-ATPase domain of approximately 260 residues [5,6,7]. In some cases an N-terminal peptidase domain of 130-150 residues appended to the TMD is also found, which may contain additional transmembrane segments as in the HlyB subfamily [8,9].

The 3D structure of the E. coli lipid A flippase MsbA homodimer reveals that association of the two transmembrane domains forms one chamber that adopt a cone-shape which extends along a pseudo two-fold axis perpendicular to the cell membrane (see <PDB:1JSQ>) [10]. The chamber has an opening on either side of the membrane to provide free access for the lipid substrate from the cytoplasmic leaflet of the lipid bilayer, while excluding molecules from the outer leaflet. The chamber openings are defined by intramolecular interactions between TM2 of one monomer and TM5 of the other. The residues lining the chamber are contributed by all 12 transmembrane α-helices [10,11].

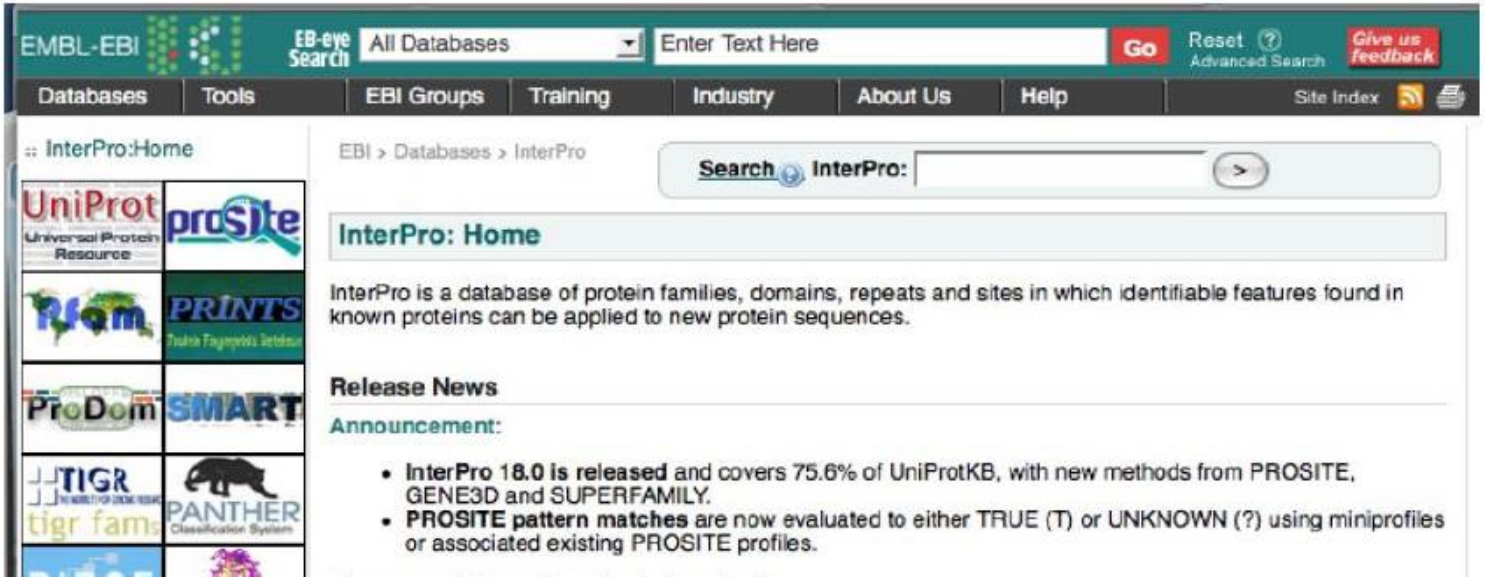
In importers (found only in prokaryotes or archaea) most ABC transporters consist of four domains usually encoded by independent polypeptides, two ABC modules and two TMDs which are thought to contain six transmembrane regions [12,13]. The approximately 30 kD TMD displays a distinctive signature, the EAA motif, a 20 amino acid conserved sequence located about 100 residues from the C-terminus. The motif is hydrophilic and has been found to reside in a cytoplasmic loop located between the penultimate and the antepenultimate transmembrane segment in all proteins with a known topology [14,15,16]. It appears to play an important role in ensuring the correct assembly of the prokaryotic ABC transport complex [17] and constituting an interaction site with the so-called helical domain of the ABC module [18,19]. The TMDs form either homo-oligomeric channels or associate with another TMD to form hetero-oligomers.

InterPro : est une base de données utilisée pour la classification et l'annotation automatique de protéines[1].

Base spécialisées dans les domaines de proteins

INTERPRO

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



Base de données dédiée au métabolisme

KEGG

http://www.genome.jp/kegg/

KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions, and relations

KEGG2 ATLAS PATHWAY BRITE GENES SSDB LIGAND DBGET

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:

1. **Metabolism**
   Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid
   Glycan PK/NRP Cofactor/vitamin Secondary metabolite Xenobiotics
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

6. **Drug Development**

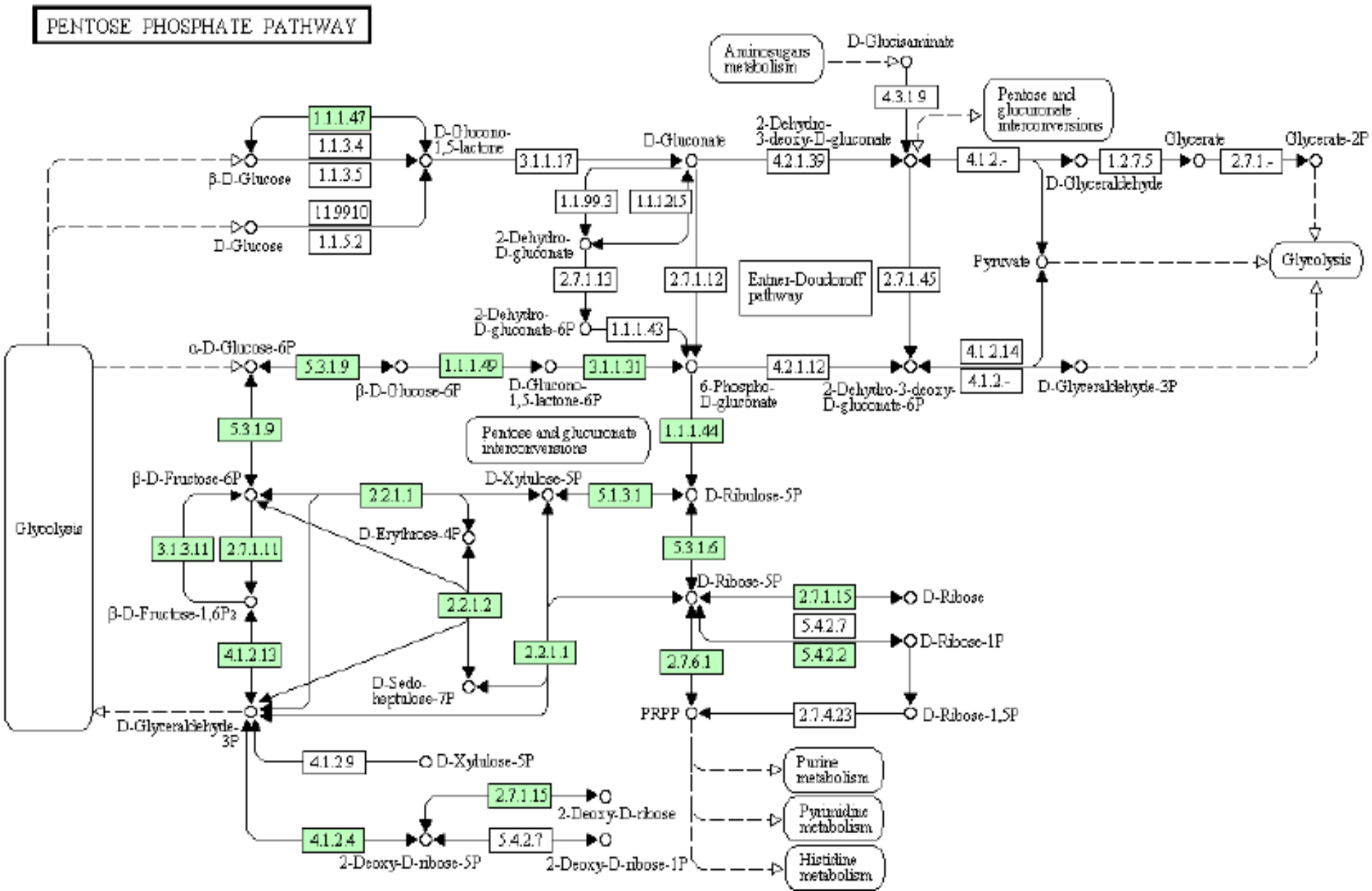
Pathway Modules

KEGG MODULE is a new collection of pathway modules, molecular complexes, and other functional units, each represented as a list of KEGG Orthology (KO) identifiers. KEGG MODULE can be accessed through a BRITE hierarchy:

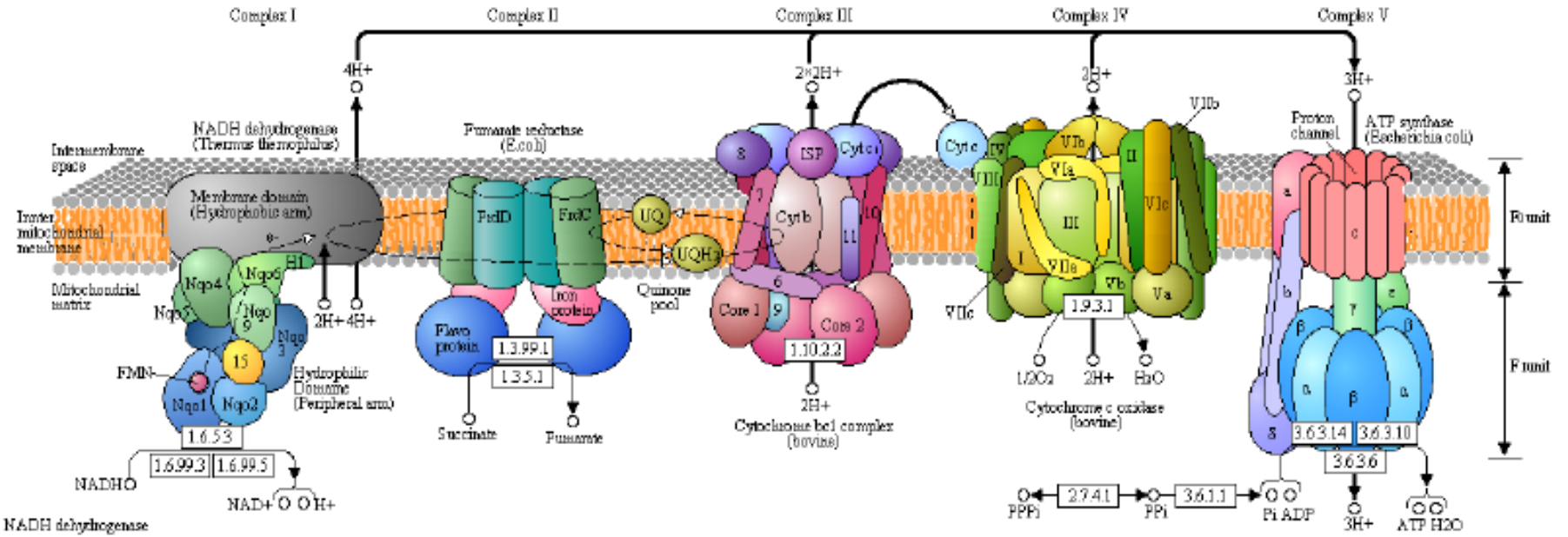
KEGG pathway modules

or by the DBGET search.

Exemple de voies metaboliques dans KEG



OXIDATIVE PHOSPHORYLATION



NADH dehydrogenase

B	ND1	ND2	ND3	ND4	ND4L	ND5	ND6
B	Ndufs1	Ndufs2	Ndufs3	Ndufs4	Ndufs5	Ndufs6	Ndufs7	Ndufs8	Ndufv1	Ndufv2	Ndufv3
B/A	NuoA	NuoB	NuoC	NuoD	NuoE	NuoF	NuoG	NuoH	NuoI	NuoJ	NuoK	NuoL	NuoM	NuoN
B/A	NdhC	NdhK	NdhJ	NdhH	NdhA	NdhI	NdhG	NdhE	NdhF	NdhD	NdhB	NdhL	NdhM	NdhN	HoxE	HoxF	HoxU
B	Ndufa1	Ndufa2	Ndufa3	Ndufa4	Ndufa5	Ndufa6	Ndufa7	Ndufa8	Ndufa9	Ndufa10	Ndufsb1	Ndufa11
B	Ndufb1	Ndufb2	Ndufb3	Ndufb4	Ndufb5	Ndufb6	Ndufb7	Ndufb8	Ndufb9	Ndufb10	Ndufb1	Ndufb2

Succinate dehydrogenase / Fumarate reductase

E	SDHC	SDHD	SDHA	SDHB
B/A	SdhC	SdhD	SdhA	SdhB
			FrdA	FrdB	FrdC	FrdD

Cytochrome c reductase

E/B/A	ISP	Cyt b	Cyt 1						
E				COR1	QCR2	QCR6	QCR7	QCR8	QCR9	QCR10

Cytochrome c oxidase

E	COX10		COX3	COX1	COX2	COX4	COX5A	COX5B	COX6A	COX6B	COX6C	COX7A	COX7B	COX7C	COX8	E/B/A		COX17
																COX11	COX15	
B/A	CyoE	CyoD	CyoC	CyoB	CyoA													
		CcoD	CcoC	CcoA	CcoB													
		QoxD	QoxC	QoxB	QoxA													

Cytochrome c oxidase, cbb3-type

B	I	II	IV	III

Cytochrome bd complex

B/A	CydA	CydB

F-type ATPase (Bacteria)

beta	alpha	gamma	delta	epsilon	c	a	b

F-type ATPase (Eukaryotes)

beta	alpha	gamma	OSCP	delta	epsilon	c	a
b	e	f6	f	8			
d	f	h	j	k	g		

P-type ATPase (Prokaryotes)

A	B	C	D	E	F	I	K

P-type ATPase (Eukaryotes)

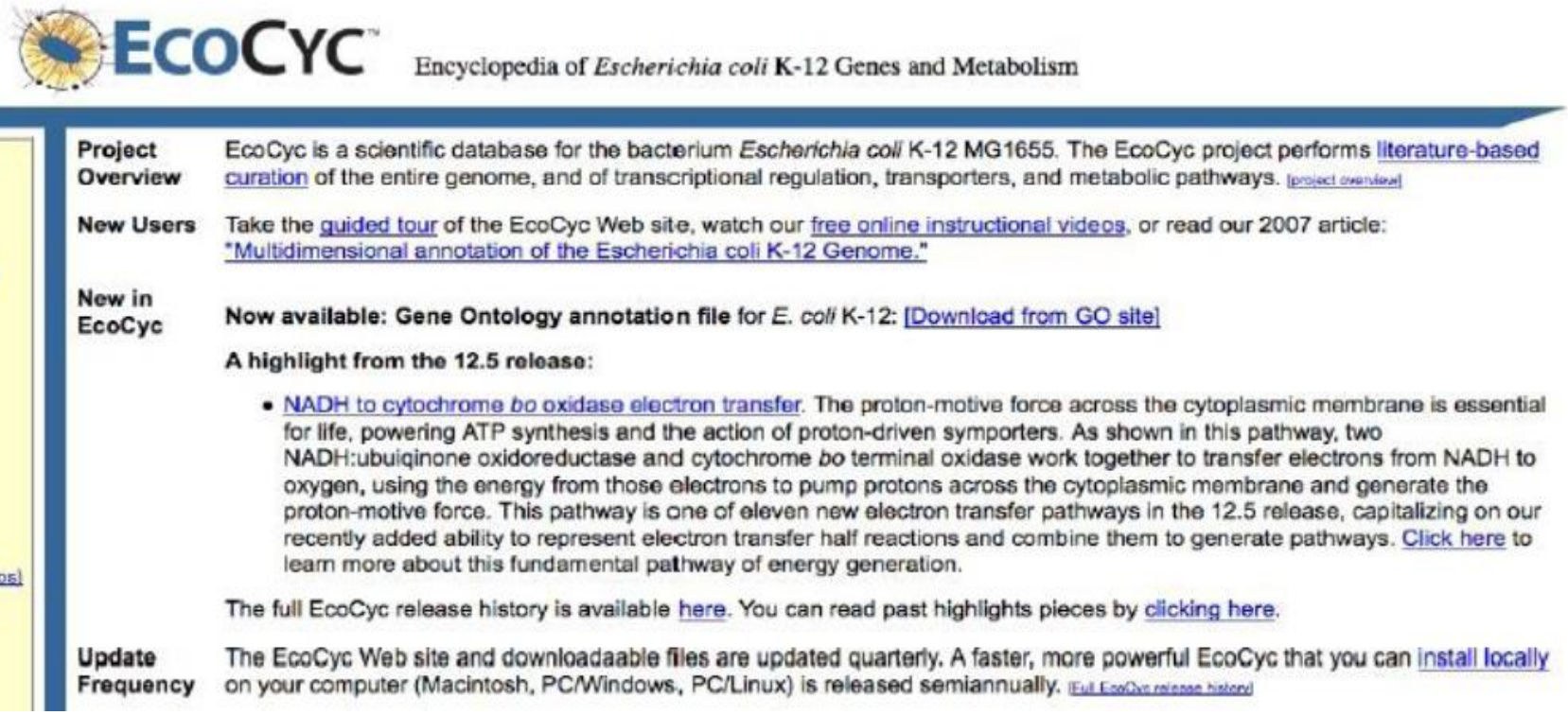
A	B	C	D	E	F	G	H
I	AC39	54kD	S1	lipid			

00190 2/8/10

Base consacrée uniquement sur *escherichia coli*

ECOCYC

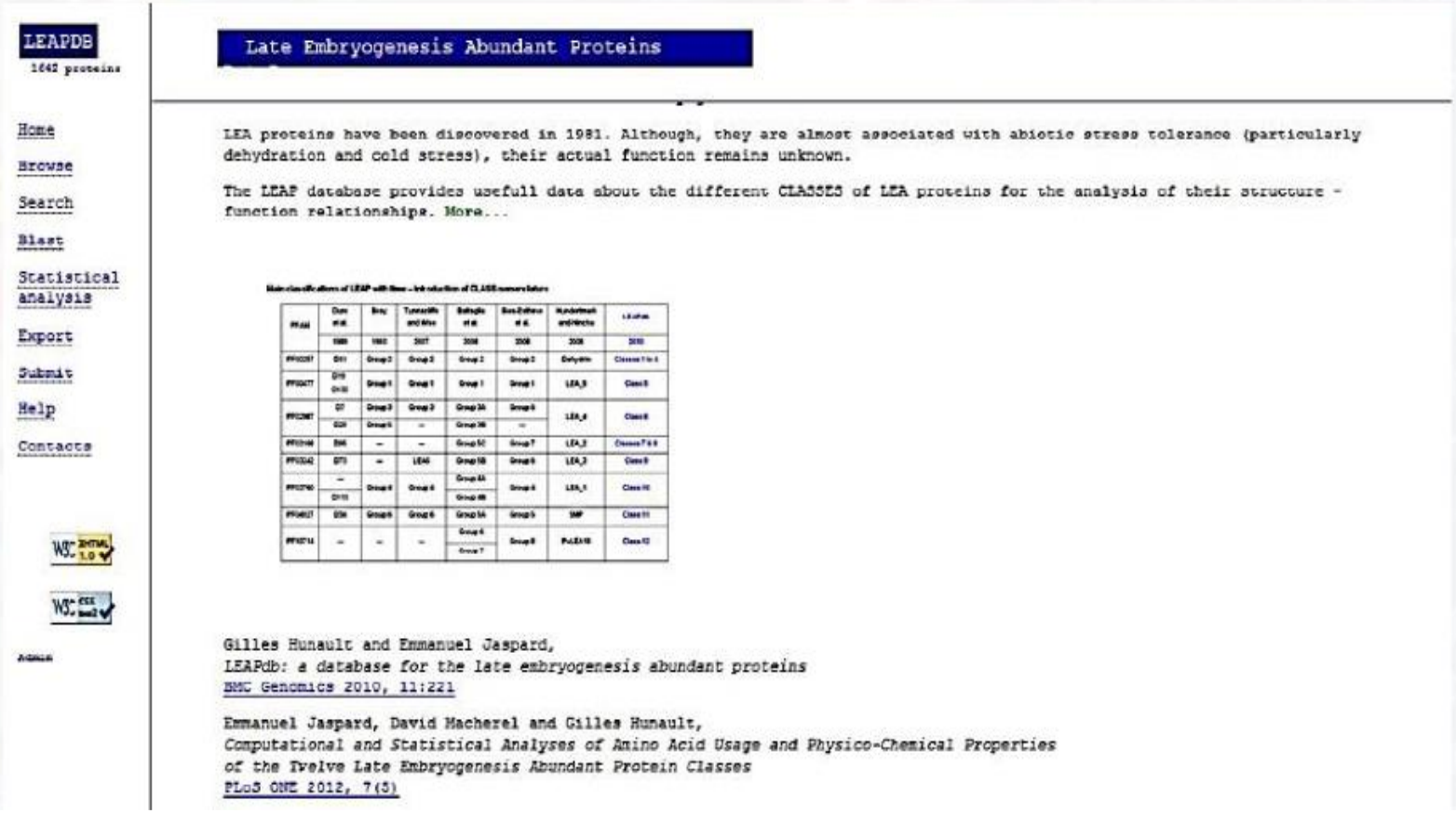
http://ecocyc.org/



https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-221/figures/1

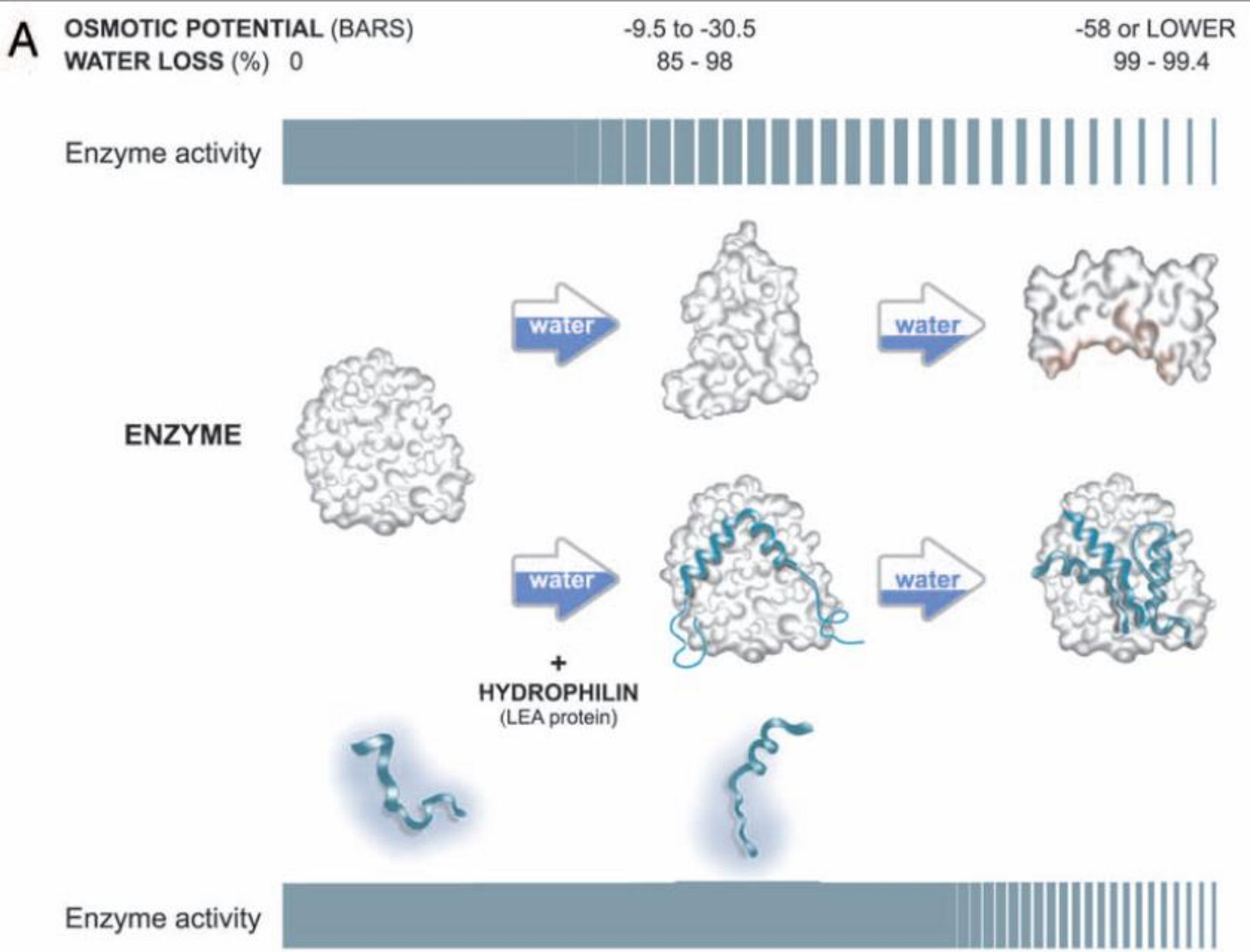
(LEAPdb)

Late Embryogenesis Abundant Proteins database (G. Hunault & E. Jaspard) : cette base de données contient un grand nombre d'informations sur les protéines LEA 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.



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.

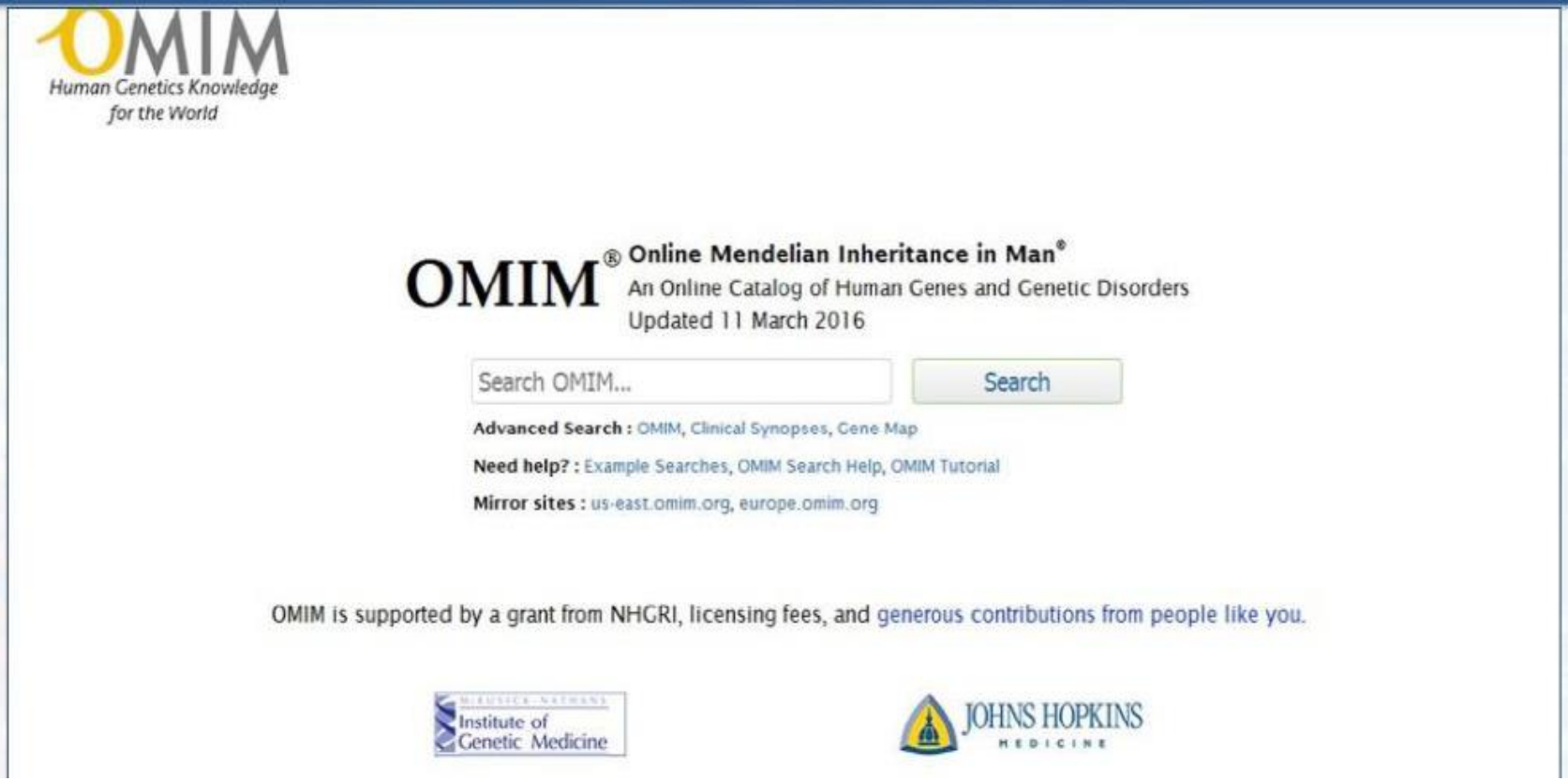
Les protéines LEA fonctionnent par des mécanismes distincts de ceux présentés par les chaperons moléculaires de choc thermique. Bien que les causes de l'induction de la protéine LEA n'aient pas encore été déterminées, des changements conformationnels dans les facteurs de transcription ou les protéines 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.


A
OSMOTIC POTENTIAL (BARS)
WATER LOSS (%) 0
-9.5 to -30.5
85 - 98
-58 or LOWER
99 - 99.4
Enzyme activity
water
water
ENZYME
water
water
+
HYDROPHILIN
(LEA protein)
Enzyme activity

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 subit 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.

LA BASE DE DONNÉES OMIM (ONLINE MENDELIAN INHERITANCE IN MAN)

Donne de nombreuses informations sur la classification des maladies génétiques, des présentations cliniques et la cartographie génomique du localisation de la maladie.
La base de données est mise à jour continuellement et offre probablement le meilleur lors de la recherche d'information sur les maladies héréditaires.



#219700

ICD+

CYSTIC FIBROSIS; CF

OMIM : maladie

Alternative titles; symbols

MUCOVISCIDOSIS

Phenotype Gene Relationships

Location	Phenotype	Phenotype MIM number	Gene/Locus	Gene/Locus MIM number
7q31.2	Cystic fibrosis	219700	CFTR	602421
19q13.2	{Cystic fibrosis lung disease, modifier of}	219700	TGFB1	190180

Lien vers les gènes ou les portions de chromosome responsables de la maladie

Clinical Synopsis

TEXT

A number sign (#) is used with this entry because the disorder is caused by mutations in the cystic fibrosis conductance regulator gene (CFTR; 602421), located on chromosome 7.

Description

Formerly known as cystic fibrosis of the pancreas, this entity has increasingly been labeled simply 'cystic fibrosis.' Manifestations relate not only to the disruption of exocrine function of the pancreas but also to intestinal glands (meconium ileus), biliary tree (biliary cirrhosis), bronchial glands (chronic bronchopulmonary infection with emphysema), and sweat glands (high sweat electrolyte with depletion in a hot environment). Infertility occurs in males and females.

Tree of life

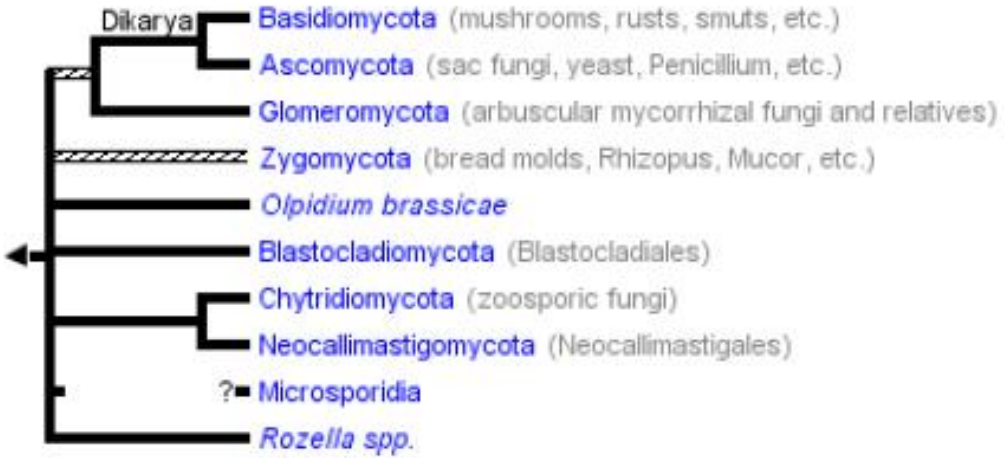
- **Tree of life :**
  - http://tolweb.org
  - Base de données de taxonomie
    - Classification des êtres vivants
  - Avec des photos !

Tree of life

Fungi

Eumycota: mushrooms, sac fungi, yeast, molds, rusts, smuts, etc.

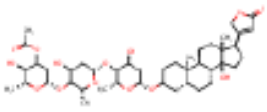
Meredith Blackwell, Rytas Vilgalys, Timothy Y. James, and John W. Taylor



DrugBank

- Base de données sur les médicaments
  - http://www.drugbank.ca
  - Information sur les cibles des médicaments
  - Attention : base américaine
    - => médicaments américains !

targets (1) Show Drugs with Similar Structures for All drugs

Identification	
Name	**Acetyldigitoxin**
Accession Number	**DB00511** (APRD01334)
Type	small molecule
Groups	approved
Description	Cardioactive derivative of lanatoside A or of digitoxin used for fast digitalization in congestive heart failure.
Structure	 Download: MOL | SDF | SMILES | InChI Display: 2D Structure | 3D Structure
Synonyms	Not Available
Brand names	• Acigoxin • Acylanid • Crystodigin
Brand name mixtures	Not Available
Categories	• Enzyme Inhibitors • Cardiotonic Agents • Anti-Arrhythmia Agents

DrugBank

Targets

1. **Sodium/potassium-transporting ATPase alpha-1 chain**

Pharmacological action: **yes**
Actions: **inhibitor**

This is the catalytic component of the active enzyme, which catalyzes the hydrolysis of ATP coupled with the exchange of sodium and potassium ions across the plasma membrane. This action creates the electrochemical gradient of sodium and potassium ions, providing the energy for active transport of various nutrients

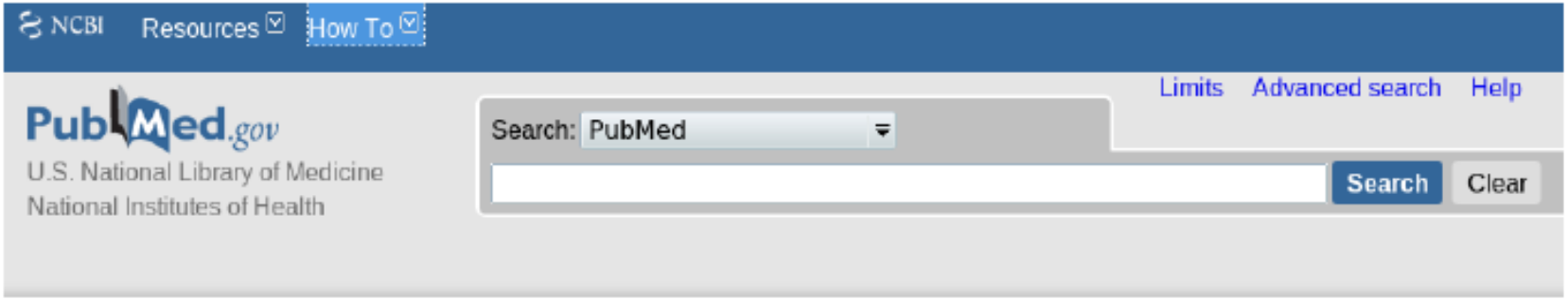
Organism class: **human**
UniProt ID: P05023
Gene: ATP1A1
Protein Sequence: FASTA
Gene Sequence: FASTA
SNPs: SNPJam Report

References:

1. Gonzalez-Garcia C, Cena V, Klein DC: Characterization of the alpha -like Na,K+-ATPase which mediates ouabain inhibition of adrenergic induction of N-acetyltransferase (EC 2.3.1.87) activity: studies with isolated pinealocytes. Mol Pharmacol. 1987 Dec;32(6):792-7. Pubmed

Pubmed / Medline

- http://www.ncbi.nlm.nih.gov/pubmed
- Medline : base de données bibliographiques en médecine / biologie
- Pubmed : interface permettant de consulter la base
- Lien vers le texte des articles disponibles en ligne



PubMed est une base de données bibliographiques, développé par le National Center for Biotechnology Information (NCBI), centrée sur la documentation en sciences biologiques.

Nucleic Acids Research

Database Categories List

- Major Sequence Repositories
- Comparative Genomics
- Gene Expression
- Gene Identification and Structure
- Genetic and Physical Maps
- Genomic Databases
- Intermolecular Interactions
- Metabolic Pathways and Cellular Regulation
- Mutation Databases
- Pathology
- Protein Databases
- Protein Sequence Motifs
- Proteome Resources
- RNA Sequences
- Retrieval Systems and Database Structure
- Structure
- Transgenics
- Varied Biomedical Content

Formats de fichiers en bioinformatique

Définition d'un format:

- **Les séquences sont stockées en général sous forme de fichiers texte, accessibles par des systèmes d'interrogations (Entrez pour GenBank, ACNUC pour EMBL, SRS pour UniProt, …)**
- **Le Format: correspond à l'ensemble des règles de présentation auxquelles sont soumises la ou les séquences dans un fichier donné.**

- **Le Format permet:**
- **Une mise en forme automatisée des données;**
- **Le stockage et la gestion homogène de l'information;**
- **Le traitement informatique ultérieur de l'information.**

Syntaxe d'une Entrée:

- **Une Entrée: est une fiche signalétique d'une séquence donnée.**

  *** Contient 3 parties:**

- ❑ **Entête (header) : Description générale de la séquence;**

- ❖ **Les caractéristiques (Features): Description des objets biologiques présents sur la séquences**

- ➢ **La séquence**

- Les formats de DDBJ et de GenBank sont très similaires
- Chaque ligne commence par un mot clé
  - Deux lettres pour EMBL
  - Maximum 12 lettres pour GenBank et DDBJ
- Fin d'une entrée par //

Partie 01: Header = entête

Thermus aquaticus DNA polymerase (PolI) gene

GenBank: J04639.1

FASTA Graphics

Go to:

```
LOCUS       TTHTAQP1A               2626 bp    DNA     linear   BCT 26-APR-1993
DEFINITION  Thermus aquaticus DNA polymerase (PolI) gene.
ACCESSION   J04639 M26480
VERSION     J04639.1
KEYWORDS    DNA polymerase.
SOURCE      Thermus aquaticus
  ORGANISM  Thermus aquaticus
            Bacteria; Deinococcus-Thermus; Deinococci; Thermales; Thermaceae;
            Thermus.
REFERENCE   1  (bases 1 to 2626)
  AUTHORS   Lawyer,F.C., Stoffel,S., Saiki,R.K., Myambo,K., Drummond,R. and
            Gelfand,D.H.
  TITLE     Isolation, characterization, and expression in Escherichia coli of
            the DNA polymerase gene from Thermus aquaticus
  JOURNAL   J. Biol. Chem. 264 (11), 6427-6437 (1989)
   PUBMED   2649500
COMMENT     Original source text: T.aquaticus (strain YT1; ATCC 25104) DNA.
            Draft entry and computer readable copy of sequence [1] kindly
            provided by D.H.Gelfand, 23-JUN-1989.
```

Partie 02: features = caractéristiques

```
FEATURES             Location/Qualifiers
     source          1..2626
                     /organism="Thermus aquaticus"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:271"
     CDS             121..2619
                     /EC_number="2.7.7.7"
                     /standard_name="Taq polymerase"
                     /codon_start=1
                     /transl_table=11
                     /product="DNA polymerase"
                     /protein_id="AAA27507.1"
                     /translation="MRGMLPLFEPKGRVLLVDGHHLAYRTFHALKGLTTSRGEPVQAV
                     YGFAKSLLKALKEDGDAVIVVFDAKAPSFRHEAYGGYKAGRAPTPEDFPRQLALIKEL
                     VDLLGLARLEVPGYEADDVLASLAKKAEKEGYEVRILTADKDLYQLLSDRIHVLHPEG
                     YLITPAWLWEKYGLRPDQWADYRALTGDESDNLPGVKGIGEKTARKLLEEWGSLEALL
                     KNLDRLKPAIREKILAHMDDLKLSWDLAKVRTDLPLEVDFAKRREPDRERLRAFLERL
                     EFGSLLHEFGLLESPKALEEAPWPPPEGAFVGFVLSRKEPMWADLLALAAARGGRVHR
                     APEPYKALRDLKEARGLLAKDLSVLALREGLGLPPGDDPMLLAYLLDPSNTTPEGVAR
                     RYGGEWTEEAGERAALSERLFANLWGRLEGEERLLWLYREVERPLSAVLAHMEATGVR
                     LDVAYLRALSLEVAEEIARLEAEVFRLAGHPFNLNSRDQLERVLFDELGLPAIGKTEK
                     TGKRSTSAAVLEALREAHPIVEKILQYRELTKLKSTYIDPLPDLIHPRTGRLHTRFNQ
                     TATATGRLSSSDPNLQNIPVRTPLGQRIRRAFIAEEGWLLVALDYSQIELRVLAHLSG
                     DENLIRVFQEGRDIHTETASWMFGVPREAVDPLMRRAAKTINFGVLYGMSAHRLSQEL
                     AIPYEEAQAFIERYFQSFPKVRAWIEKTLEEGRRRGYVETLFGRRRYVPDLEARVKSV
                     REAAERMAFNMPVQGTAADLMKLAMVKLFPRLEEMGARMLLQVHDELVLEAPKERAEA
                     VARLAKEVMEGVYPLAVPLEVEVGIGEDWLSAKE"
```

Partie 03: origin= séquence

```
ORIGIN
        1 ggcatgaaag tcagggcaga gccatctatt gcttacattt gcttctgaca caactgtgtt
       61 cactagcaac ctcaaacaga caccatggtg cacctgactc ctgaggagaa gtctgccgtt
      121 actgccctgt ggggcaaggt gaacgtggat gaagttggtg gtgaggccct gggcaggttg
      181 gtatcaaggt tacaagacag gtttaaggag accaatagaa actgggcatg tggagacaga
      241 gaagactctt gggtttctga taggcactga ctctctctgc ctattggtct attttcccac
      301 ccttaggctg ctggtggtct acccttggac ccagaggttc tttgagtcct ttggggatct
      361 gtccactcct gatgctgtta tgggcaaccc taaggtgaag gctcatggca agaaagtgct
      421 cggtgccttt agtgatggcc tggctcacct ggacaacctc aagggcacct ttgccacact
      481 gagtgagctg cactgtgaca agctgcacgt ggatcctgag aacttcaggg tgagtctatg
      541 ggacccttga tgttttcttt ccccttcttt tctatggtta agttcatgtc ataggaaggg
      601 gagaagtaac agggtacagt ttagaatggg aaacagacga atgatt
//
```

Définition et numéro d'accession

Organisme et taxonomie

Référence bibliographique

Caractéristiques de la séquence: Domaines, Motifs, RBS, CDS...

Séquence

Descripteur de fin de fichier

```
LOCUS       AAB59426                 490 aa            linear   PRI 31-DEC-1994
DEFINITION  cytochrome.
ACCESSION   AAB59426
VERSION     AAB59426.1  GI:181344
DBSOURCE    locus HUMCYPC219 accession M61854.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 490)
  AUTHORS   Romkes,M., Faletto,M.B., Blaisdell,J.A., Raucy,J.L. and
            Goldstein,J.A.
  TITLE     Cloning and expression of complementary DNAs for multiple members
            of the human cytochrome P450IIC subfamily
  JOURNAL   Biochemistry 30 (13), 3247-3255 (1991)
   PUBMED   2009263
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..490
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
     Protein         1..490
                     /product="cytochrome"
     Region          30..487
                     /region_name="p450"
                     /note="Cytochrome P450; pfam00067"
                     /db_xref="CDD:40168"
     Region          30..480
                     /region_name="CypX"
                     /note="Cytochrome P450 [Secondary metabolites
                     biosynthesis, transport, and catabolism]; COG2124"
                     /db_xref="CDD:32307"
     CDS             1..490
                     /gene="CYP2C19"
                     /coded_by="M61854.1:6..1478"
ORIGIN
        1 mdpfvvlvlc lscllllsiv rqssgrgklp pgptpipvig nilqidikdv skslsnlski
       61 ygpvftlyfg lkmvvihgy evvkealidl geefsgrghf plaeranrgf givfsngkrw
      121 keirrfslmt lrnfgmgkrs iedrvqeear clveelrktk aspcdptfil gcapcnvics
      181 iifqkrfdyk dqqflnlmek lneniriust pwiqicnnfp tiidyfpgth nklkknlafm
      241 esdilekvke hqesmdinnp rdfidcflik mekekhqnqqs eftienlvit aadllgagte
      301 ttsttlryal lllkhpevt akvqeeierv igrnrspcmq drgimpytda vvhevqryid
      361 liptslphav tcdvkfrnyl ipkgttilts lisvlhdnke fpnpemfdpr hfldeggnfk
      421 ksnyfmpfsa gkricvgegl armelflflt fiiqnfnlks lidpkdldtt pvvngfasvp
      481 pfyqlcfipv
//
```

GenBank/DDBJ

EMBL Flat File

```
ID   AF115338   standard; DNA; PRO; 591 BP.
AC   AF115338;
SV   AF115338.1
DT   03-JUN-1999 (Rel. 59, Created)
DT   23-AUG-1999 (Rel. 60, Last updated, Version 2)
DE   Pseudomonas fluorescens ECF sigma factor SigX (sigX) gene, complete cds.
KW   .
OS   Pseudomonas fluorescens
OC   Bacteria; Proteobacteria; gamma subdivision; Pseudomonadaceae; Pseudomonas.
RN   [1]
RP   1-591
RX   MEDLINE; 99369842.
RA   Brinkman F.S., Schoofs G., Hancock R.E., De Mot R.;
RT   "Influence of a putative ECF sigma factor on expression of the major outer
RT   membrane protein, OprF, in Pseudomonas aeruginosa and Pseudomonas
RT   fluorescens";
RL   J. Bacteriol. 181(16):4746-4754(1999).
RN   [2]
RP   1-591
RA   De Mot R.;
RT   ;
RL   Submitted (04-DEC-1998) to the EMBL/GenBank/DDBJ databases.
RL   F.A. Janssens Laboratory of Genetics, Applied Plant Sciences, K.
RL   Mercierlaan 92, Heverlee B-3001, Belgium
DR   SPTREMBL; Q9X4L7; Q9X4L7.
FH   Key             Location/Qualifiers
FH
FT   source 1..591
FT                   /db_xref="taxon:294"
FT                   /organism="Pseudomonas fluorescens"
FT                   /strain="M114"
FT   CDS 1..591
FT                   /codon_start=1
FT                   /db_xref="SPTREMBL:Q9X4L7"
FT                   /transl_table=11
FT                   /gene="sigX"
FT                   /product="ECF sigma factor SigX"
FT                   /protein_id="AAD34329.1"
FT                   /translation="MNKAQTLSTRYDPRELSDEELVARSHTELFHVTRAYEELMRRYQR
FT                   TLFNVCARYLGNDRDADDVCQEVMLKVLYGLKNLEGKSKFKTWLYSITYNECITQYRKE
FT                   RRKRRLMDALSLDPLEEASEEKALQPEEKGGLDRWLVYVNPIDRGILVLRFVAELEFQE
FT                   IADIMHMGLSATKMRYKRALDKLREKFAGETET"
SQ   Sequence 591 BP; 157 A; 133 C; 170 G; 131 T; 0 other;
     atgaataaag cccaaacgct atccacgcgc tacgaccccc gcgagctctc tgatgaggag        60
     ttggtcgcgc gctcgcatac cgagcttttt cacgtaacgc gcgcctatga agaactgatg       120
     cggcgttacc agcgaacatt atttaacgtt tgtgcgagat atcttgggaa cgatcgcgac       180
     gcagacgatg tctgtcagga agtcatgttg aaggtgctgt atggcctgaa gaacctcgag       240
     gggaaatcga agttcaaaac gtggctctac agcatcacgt acaacgaatg tattacgcag       300
     tatcggaagg aacggcgaaa gcgtcgcttg atggacgcat tgagtcttga ccccctcgag       360
```

Header

- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

Les *flatfiles* en détails (1/8)

1. Les différentes informations de l'entête (*header*) :

- Première ligne : Locus/ID

→ Embl

```
ID   U49845;   SV 1; linear; genomic DNA; STD; FUN; 5028 BP.
```

→ DDBJ/GenBank

```
LOCUS   SCU49845   5028 bp   DNA   linear   PLN 21-JUN-1999
```

- Le champ date (chez EMBL uniquement)

```
DT   07-MAY-1996 (Rel. 47, Created)
DT   17-APR-2005 (Rel. 83, Last updated, Version 4)
```

- Lignes de définition : synthèse du contenu biologique

```
DE   Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2)
and
DE   Rev7p (REV7) genes, complete cds.
```

```
DEFINITION   Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
             (AXL2) and Rev7p (REV7) genes, complete cds.
```

Division en organisme

Utilisée dans les 3 principales banques de séquences nucléiques

Sequence division	Database
Organismal	
BCT Bacterial	DDBJ, GenBank
PRO Prokaryotic	EMBL
FUN Fungal	EMBL
HUM Human	DDBJ, EMBL
PRI Prinate	DDBJ, EMBL, GenBank
ROD Rodent	DDBJ, EMBL, GenBank
MAM Other mammalian	DDBJ, EMBL, GenBank
VRT Other vertebrate	DDBJ, EMBL, GenBank
INV Invertebrate	DDBJ, EMBL, GenBank
PLN Plant	DDBJ, EMBL, GenBank
ORG Organelle	EMBL
VRL Viral	DDBJ, EMBL, GenBank
PHG Phage	DDBJ, EMBL, GenBank
RNA Structural RNA	DDBJ, EMBL, GenBank
SYN Synthetic and chimeric	DDBJ, EMBL, GenBank
UNA Unannotated	DDBJ, EMBL, GenBank
Functional	
EST Expressed sequence tag	DDBJ, EMBL, GenBank
STS Sequence tagged site	DDBJ, EMBL, GenBank
GSS Genome survey	DDBJ, EMBL, GenBank
HTG High-throughput genomic	DDBJ, EMBL, GenBank
PAT Patent	DDBJ, EMBL, GenBank
CON Virtual contigs of segmented sequences	DDBJ, EMBL, GenBank

Les *flatfiles* en détails (2/8)

- Le numéro d'accession : un id unique (commun aux 3 banques)

```
AC   U49845;
```

```
ACCESSION   U49845
```

- La version (équivalent à SV dans la 1re ligne d'EMBL)

```
VERSION     U49845.1  GI :1293613
```

- Lignes avec des mots-clés (KEYWORDS ou KW)

- Lignes de taxonomie

```
OS   Saccharomyces cerevisiae (baker's yeast)
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
```

```
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
```

Les *flatfiles* en détails (3/8)

- Les références : publication ou origine de la soumission

```
RN    [1]
RP    1-5028
RX    PUBMED; 7871890.
RA    Torpey L.E., Gibbs P.E., Nelson J., Lawrence C.W.;
RT    "Cloning and sequence of REV7, a gene whose function is required for DNA
RT    damage-induced mutagenesis in Saccharomyces cerevisiae";
RL    Yeast 10(11) :1503-1509(1994).
XX
RN    [2]
RP    1-5028
RX    PUBMED; 8846915.
RA    Roemer T., Madden K., Chang J., Snyder M.;
RT    "Selection of axial growth sites in yeast requires Axl2p, a novel plasma
RT    membrane glycoprotein";
RL    Genes Dev. 10(7) :777-793(1996).
XX
RN    [3]
RP    1-5028
RA    Roemer T.;
RT    ;
RL    Submitted (22-FEB-1996) to the EMBL/GenBank/DDBJ databases.
RL    Terry Roemer, Biology, Yale University, New Haven, CT, USA
```

Les *flatfiles* en détails (4/8)

REFERENCE	1 (bases 1 to 5028)
AUTHORS	Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE	Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL	Yeast 10 (11), 1503-1509 (1994)
PUBMED	7871890
REFERENCE	2 (bases 1 to 5028)
AUTHORS	Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE	Selection of axial growth sites in yeast requires Axl2p, a novel plasma membrane glycoprotein
JOURNAL	Genes Dev. 10 (7), 777-793 (1996)
PUBMED	8846915
REFERENCE	3 (bases 1 to 5028)
AUTHORS	Roemer,T.
TITLE	Direct Submission
JOURNAL	Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New Haven, CT, USA

Les *flatfiles* en détails (5/8)

2. Les caractéristiques (*features*)

   - Features table : annotation des séquences, gestion du transfert d'information, localisation des éléments biologiques
   - Mise à disposition d'un vocabulaire étendu controlé (Gene Ontology) pour décrire les caractéristiques biologiques des séquences (annotations)
   - Annotation = information additionnelle venant enrichir un document d'intérêt

     → Annotation structurale : caractériser les gènes au travers, notamment, de leur structure intron-exon, signaux de régultaion, sites d'épissage, ...

     → Annotation fonctionnelle : anticiper sur les motifs fonctionnels qui seront retrouvés au niveau protéiques, caractérisation des fonctions biologiques des produits d'expression des gènes

Les *flatfiles* en détails (6/8)

```
FT   source    1..5028
FT             /organism="Saccharomyces cerevisiae"
FT             /chromosome="IX"
FT             /map="9"
FT             /mol_type="genomic DNA"
FT             /db_xref="taxon :4932"
FT   CDS       <1..206
FT             /codon_start=3
FT             /product="TCP1-beta"
FT             /db_xref="GOA :P39076"
FT             /db_xref="UniProtKB/Swiss-Prot :P39076"
FT             /protein_id="AAA98665.1"
FT             /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEAA
FT             EVLLRVDNIIRARPRTANRQHM"
[...]
FT   CDS       complement(3300..4037)
FT             /codon_start=1
FT             /gene="REV7"
FT             /product="Rev7p"
FT             /db_xref="GOA :P38927"
FT             /db_xref="InterPro :IPR003511"
FT             /db_xref="SGD :S000001401"
FT             /db_xref="UniProtKB/Swiss-Prot :P38927"
FT             /protein_id="AAA98667.1"
FT             /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQF
FT             VPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVDKD
FT             DQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNRRVD
FT             SLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEKLISG
FT             DDKILNGVYSQYEEGESIFGSLF"
```

Les *flatfiles* en détails (8/8)

3. La séquence de nucléotides

→ EMBL

```
SQ   Sequence 5028 BP; 1510 A; 1074 C; 835 G; 1609 T; 0 other;
     gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg       60
     ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct      120
     ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa      180
     gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg      240
     ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa      300
     agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa      360
     attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat      420
     aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga      480
     gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc      540
```

→ GenBank

```
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
```

Swiss-Prot http://www.uniprot.org/uniprot/P39076.txt

```
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 TCP1 chaperonin complex of yeast, TCP1 beta.";
RL   Proc. Natl. Acad. Sci. U.S.A. 91:2743-2747(1994).
```

Formats de fichiers

Il existe plusieurs formats de stockage du texte :

- Spécifiques aux séquences :
  - Séquence brut (plain raw sequence)
  - Format FASTA
  - Format EMBL
  - Format GenBank
  - Format PDB

Formats de fichiers : Texte brut

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGC
CACGGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCCGAGAC
AGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTCCTGA
CTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCC
CCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCA
CCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCT
TCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTC
ACGCAAGTTTAATTACAGACCTGAA
```

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

Formats de fichiers : FASTA

- Format commun de manipulation des données :
  - Objectif : **manipuler facilement** des séquences dans les bases de données, à l'aide d'un **format universel**, compatibles avec les traitements de texte (sous forme de fichier texte), ou par copier – coller.
  - Le format FASTA : une 1re ligne de définition introduite par un >, contenant un identifiant sans espace et une description suivie de la séquence elle-même, sur plusieurs lignes (de taille 60 ou 80 classiquement)

```
>emb|U49845|U49845 Saccharomyces cerevisiae TCP1-beta
gene, partial cds; and Axl2p (AXL2) and Rev7p (REV7) genes,
complete
gatcctccatatacaacggtatctccacctcaggtttagatctcaacaacggaaccattg
ccgacatgagacagttaggtatcgtcgagagttacaagctaaaacgagcagtagtcagct
ctgcatctgaagccgctgaagttctactaagggtggataacatcatccgtgcaagaccaa
gaaccgccaatagacaacatatgtaacatatttaggatatacctcgaaaataataaaccg
ccacactgtcattattataattagaaacagaacgcaaaaattatccactatataattcaa
agacgcgaaaaaaaaagaacaacgcgtcatagaacttttggcaattcgcgtcacaaataa
attttggcaacttatgtttcctcttcgagcagtactcgagccctgtctcaagaatgtaat
aatacccatcgtaggtatggttaaagatagcatctccacaacctcaaagctccttgccga
```

Formats de fichiers : EMBL

```
ID AB000263 standard; RNA; PRI; 368 BP.
XX
AC AB000263;
XX
DE Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ Sequence 368 BP;
acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg 60
ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg 120
caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc 180
aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag 240
gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga 300
agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca 360
gacctgaa 368
//
```

- Peut contenir plusieurs séquences
- Chaque séquence commence par « ID » suivi par des descriptions.
- La séquence suit une ligne marquée de « SQ » et est terminée par « // »

Formats de fichiers : GenBank

```
LOCUS AB000263 368 bp mRNA linear PRI 05-FEB-1999
DEFINITION Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
ACCESSION AB000263
ORIGIN
        1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
       61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
      121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

- Peut contenir plusieurs séquences (séparées par « // »)
- Contient d'abord des mots clés (LOCUS, ACCESSION …) suivis de leurs valeurs, puis le mot clé ORIGIN suivi de la séquence (60 caractères par ligne)

Formats de fichiers : IG

```
; comment
; comment
AB000263
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACC
GCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAA
GGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCT
CATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCC
GGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATG
CTCACGCAAGTTTAATTACAGACCTGAA1
```

- Peut contenir plusieurs séquences
- Commence par des lignes de commentaires commençant par « ; », puis une ligne indiquant le nom de la séquence (sans espaces), puis la séquence.
- La séquence se termine par « 1 » si elle est linéaire, « 2 » si elle est circulaire.

Formats de fichiers : PDB (1/2)

```
HEADER    OXIDOREDUCTASE                  27-OCT-03                  1UR5
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
```

Formats de fichiers : PDB (1/2)

```
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