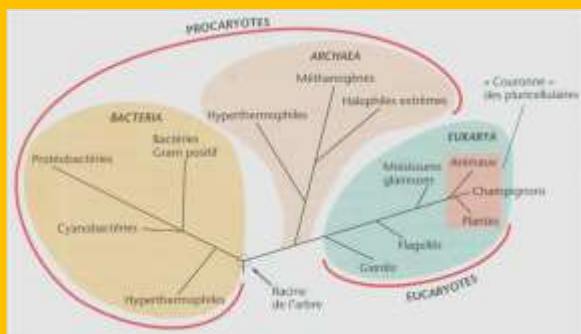
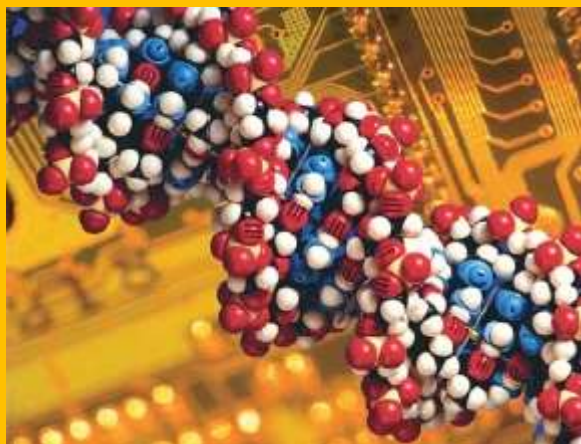


Université Frères Mentouri Constantine  
Institut de la Nutrition, de l'Alimentation et des Technologies Agro-alimentaires (INATAA)  
Master 1-Technologies Alimentaires  
Année universitaire 201-2020



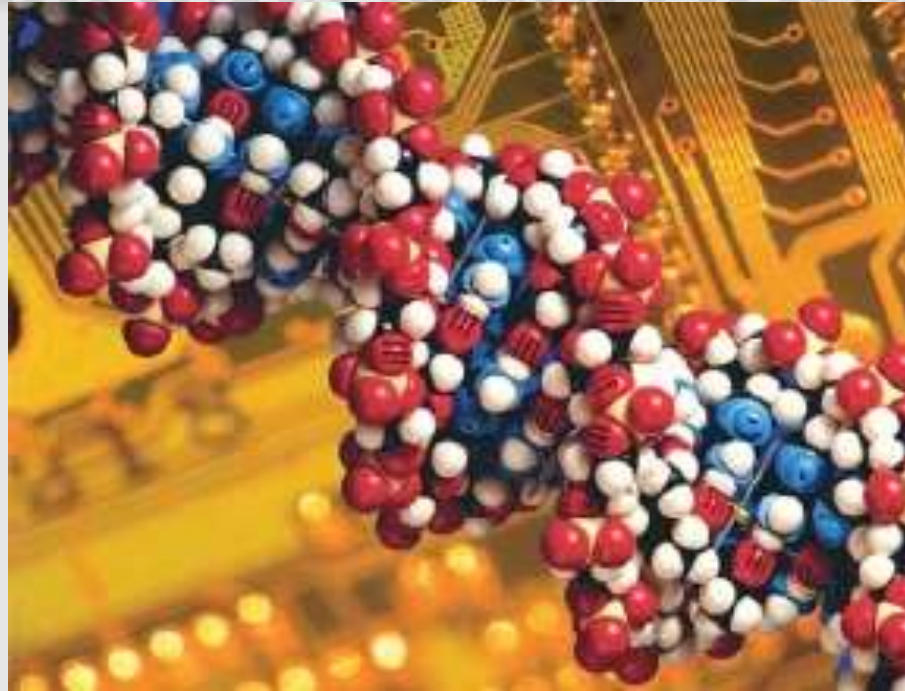
# INITIATION À LA BIOINFORMATIQUE



# CE QUE NOUS ALLONS VOIR:

- ✓ Les banques de données biologiques ;
- ✓ Les alignements ;
- ✓ Exemples d'application:
  - ☐ Phylogénie moléculaire
  - ☐ Annotation des séquences

# **INTRODUCTION À LA BIOINFORMATIQUE**



## DÉFINITION DE LA BIOINFORMATIQUE

- La bioinformatique est le **traitement de l'information biologique** sous forme de données accessibles aisément et exploitables.
- Elle est également définie comme étant la (multi) discipline théorique de l'analyse "*in silico*" de l'information biologique contenue dans les **séquences** nucléiques et protéiques\*.

---

\*la bioinformatique tire sa définition de deux concepts importants : la biologie et l'information car le suffixe informatique n'a rien à voir avec l'utilisation des ordinateurs pour la biologie.

**L'INFORMATION BIOLOGIQUE**

La bioinformatique s'intéresse aux données du :

- ✓ génome (totalité du matériel génétique de la cellule) ;
- ✓ transcriptome (ensemble des ARNm transcrits) ;
- ✓ protéome (l'ensemble des protéines bio-synthétisées) ;
- ✓ métabolome (molécules organiques -métabolites- impliquées dans les activités métaboliques de la cellule vivante).



# LES BANQUES DE DONNÉES BIOLOGIQUES

**ExPASy Proteomics Server**  
 Databases Tools Services Mirrors About Contact

You are here: ExPASy CH > Databases > Around UniProtKB

**Swiss-Prot**  
 Protein knowledgebase  
**TrEMBL**  
 Computer-annotated supplement to Swiss-Prot

The UniProt Knowledgebase consists of:

- **UniProtKB/Swiss-Prot**, a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. [More details / References / Linking to UniProtKB/Swiss-Prot / User manual / Recent changes / Disclaimer](#)
- **UniProtKB/TrEMBL**, a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups **at SIB** and **at EBI**

UniProt Knowledgebase Release 2010\_05 consists of:  
 UniProtKB/Swiss-Prot Release 2010\_05 of 20-Apr-10: 516603 entries ([More statistics](#))  
 UniProtKB/TrEMBL Release 2010\_05 of 20-Apr-10: 10706472 entries ([More statistics](#))

Access to the UniProt Knowledgebase

• UniProtKB with sites

**EMBL-EBI** EBI Search All Databases Enter Text Here

Databases Tools EBI Groups Training Industry About Us Help

• EMBL Bank Home  
 • Access  
 • Documentation  
 • News  
 • Submission  
 • Publications  
 • People  
 • Contact

**EMBL Nucleotide Sequence Database**

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

The database is produced in an international [collaboration](#) with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The [current database release](#) (Release 103, March 2010), with according [Release notes](#) and [user manual](#) are available from the EBI servers. A sample database entry is shown [here](#).

A publication in *Nucleic Acids Research* 2009 37: D18-D25 provides further information and details.

The EMBL nucleotide sequence database forms part of the [European Nucleotide Archive](#), an EBI project led by Guy Cauchemez as part of the [The Protein and Nucleotide Database Group \(PANDA\)](#) under Ewan Birney.

Link	Explanation
<a href="#">Access</a>	Database queries, Contextual genomes, webviewer, FTP archives (EMBL release, alignments etc), <a href="#">EMBL sequence version archive (SVA)</a> , <a href="#">Browse by geography</a>
<a href="#">Submission</a>	Primary sequence submissions, third party annotation, updates
<a href="#">Documentation</a>	Release notes, user manual, information for Submitters, FAQ, Release information, Formatting changes, EMBL database statistics, Feature table, XML documentation, Sequence entry, Submission, Feature table codes, Examples of annotations, EMBL Features & Qualifiers, ORF line standards, Database Policies
<a href="#">Publications</a>	Group publications
<a href="#">People</a>	Group members
<a href="#">Contact</a>	How to contact the EMBL Nucleotide Sequence Database
<a href="#">News</a>	List of recent changes on this site

**DDBJ**  
 DNA Data Bank of Japan

Accession DNA, Protein, AIDs, Taxonomy, Site Search  
 Accession numbers Go

DDBJ UniProt PDB DAD PRF Patent [History](#)

HOME Submission How to Use Search/Analysis FTP/WebAPI Report/Statistics Contact Us RSS Japanese

• About DDJ  
 • How to Use  
 • Q and A

**Sequence Submission**

- [SAKURA](#)
- [Mass Submission](#)
- [Data Updates](#)
- [DDJ Read Archive](#)
- [DDJ Trace Archive](#)

**Search**

- [Identity](#)
- [ARSA](#)
- [TVSearch](#)
- [BLAST](#)
- [PS-BLAST](#)
- [FASTA](#)
- [SSEARCH](#)

**Phylogenetics**

- [ClustalW](#)

**DDJ : DNA Data Bank of Japan**

DDJ (DNA Data Bank of Japan) is one of the three databanks that constitute DDJ/EMBL/GenBank International Nucleotide Sequence Database which was established through cooperative work with EBI in Europe and NCBI in the USA.

**Hot Topics**

- Apr. 15, 2010 [The Chinese Academy of Science professors visited DDJ](#)
- Apr. 12, 2010 [Release of the raw and assembled sequence data set from ngs](#)
- Apr. 12, 2010 [DAD \(DDJ amino acid database\) Rel. 51.0 Released](#)

**Maintenance**

- Apr. 21, 2010 [Suspension of some DDJ activities in Japanese holidays \(4/29-5/1-5\)](#)
- Mar. 16, 2010 [\(Apr. 23\) ARSA database search \(DDJ, DAD\) temporary unavailable](#)
- Feb. 03, 2010 [\(Important\) Termination of a part of DDJ services](#)

**Sequence Data Submission**

- [Submit my sequences](#)  
 Orientation for the data submission
- [Update my entries](#)

**FTP/Web API**

- [FTP \(ftp.ddbj.nig.ac.jp\)](#)  
 Download data files
- [Web API](#)

**CATÉGORIES DE BANQUES DE DONNÉES BIOLOGIQUES**

BD biologiques: grandes bibliothèques de données de biologie et des sciences de la vie obtenues par **expérimentation** ou par analyse des **simulations**. Ces banques sont **libres d'accès**, les données sont obtenues de manière collaborative.

Nous distinguerons deux types de banques:

1. **Banques de données généralistes:** correspondent à une collecte des données la plus exhaustive possible ;
2. **Banques de données spécialisées:** correspondent à des données plus homogènes établies autour d'une thématique particulière.

## CATÉGORIES DE BANQUES DE DONNÉES BIOLOGIQUES

**Les banques de données généralistes:** contiennent des données hétérogènes:

- Banques de séquences nucléiques (**EMBL**, **GenBank**, **DDBJ**) : gènes, fragments de gènes, de génomes, ADN non codant, ARN, etc. ;
- Banques de séquences protéiques (**Uniprot**, **PIR**, **SwissProt**) : traduction de séquences ADN codant des protéines complètes ou fragments, protéines séquencées, annotées ou pas ;
- Banques génomiques et de localisation (**Ensembl**) : génomes annotés ;
- Banques de structures 3D de macromolécules (**PDB**) ;
- Banques d'articles scientifiques (**Medline**).



## BANQUES DE DONNÉES GÉNÉRALISTES

## BD DE SÉQUENCES NUCLÉIQUES

Trois principales banques, **interconnectées entre elles** :

**Genbank**: banque de données américaine\*, diffusée par le NCBI (*National Center for Biotechnology Information*, Los Alamos, USA)

<http://www.ncbi.nlm.nih.gov/genbank/>

The screenshot shows the NCBI GenBank homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. Below this is a search bar with 'GenBank' and a dropdown menu set to 'Nucleotide'. A 'Search' button is to the right. Below the search bar is a horizontal menu with various database categories: GenBank, Submit, Genomes, WGS, HTGs, EST/GSS, Metagenomes, TPA, TSA, and INSDC. The main content area is divided into two columns. The left column is titled 'GenBank Overview' and contains a section 'What is GenBank?' which describes the database as a collection of publicly available DNA sequences. It also mentions the 'GenBank release' cycle and provides links to 'release notes' and 'GenBank releases'. Below this, it mentions an 'annotated sample GenBank record' for *Saccharomyces cerevisiae*. The right column is titled 'GenBank Resources' and lists links for 'GenBank Home', 'Submission Types', 'Submission Tools', 'Search GenBank', and 'Update GenBank Records'. At the bottom of the left column, there is a section 'Access to GenBank' which lists several ways to search and retrieve data, including using 'Entrez Nucleotide', 'BLAST', and 'NCBI e-utilities'. The bottom of the page features a 'GenBank Data Usage' section with a disclaimer about the database's design and the NCBI's policy on patent and copyright claims.

\*En Décembre 2017, Genbank archivait plus de 206 millions de séquences nucléiques.

## BANQUES DE DONNÉES GÉNÉRALISTES

## BD DE SÉQUENCES NUCLÉIQUES

## ENA

*(European Nucleotide Archive)*banque européenne \* diffusée par **EMBL-EBI** (Cambridge, UK)*(European Molecular Biology Library-European Bioinformatics Institute)*<http://www.ebi.ac.uk/ena>

The screenshot shows the ENA website with the EMBL-EBI logo in the top left. The top navigation bar includes links for Services, Research, Training, and About us. The main header features the ENA logo and a search bar with a 'Search' button and a link to 'Advanced Sequence'. Below the header is a secondary navigation bar with links for Home, Search & Browse, Submit & Update, Software, About ENA, and Support.

The main content area is titled 'European Nucleotide Archive' and describes the archive's purpose: 'The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#)'. It also states: 'Access to ENA data is provided through the browser, through search tools, large scale file download and through the API.'

There are two main search sections: 'Text Search' and 'Sequence Search'. Each has a search input field, a 'Search' button, and a link to 'Advanced search'. The 'Text Search' section includes examples: 'BN000065, histone'.

On the right side, there is a 'Popular' section with a list of links: 'Submit and update', 'Sequence submissions', 'Genome assembly submissions', 'Submitting environmental sequences', 'Citing ENA data', 'Rest URLs for data retrieval', and 'Rest URLs to search ENA'. Below this is a 'Latest ENA news' section with three entries: '06 Jan 2017: [FTP Service update](#) FTP service is now resuming normal operation.', '07 Dec 2016: [ENA Release 130](#) Release 130 of ENA's assembled/annotated sequences now available', and '14 Nov 2016: [ENA launches Browser and Advanced Search surveys](#) Have your say in future improvements to the ENA Browser and ENA's Advanced Search'.

\*Le 05 Janvier 2018, ENA archivait plus de 1 milliard 157 millions de séquences nucléiques.

## BANQUES DE DONNÉES GÉNÉRALISTES

## BD DE SÉQUENCES NUCLÉIQUES

Trois principales banques :

**DDBJ (*DNA Data Bank of Japan*):** Banque de données japonaise\*, diffusée par le NIG (*National Institute of Genetics*, Japon)

<http://www.ddoj.nig.ac.jp>

The screenshot shows the DDBJ website interface. At the top, there's a header with the DDBJ logo and a 30th anniversary badge. A search bar with 'Google カスタム検索' and a 'Search' button is on the right. Below the header is an orange navigation bar with links: 'About DDBJ', 'How to Use', 'Report/Statistics', 'FAQ', and 'Contact Us'. The main content area is divided into two columns. The left column has social media links (RSS, DDBJ Twitter, Mail Magazine) and a vertical stack of logos: DDBJ, INSDC, NCBI, EMBL-EBI, NIG, JBIportal, NBDC, DBCLS, and PDBj. The right column has a 'DDBJ Service' section with four icons: 'Data Submission' (database cylinders), 'Search / Analysis' (DNA helix with magnifying glass), 'Super Computer' (server rack), and 'ftp.ddbj.nig.ac.jp' (FTP icon). Below this is a 'Hot Topics' section with a 'News Archive' link and a table of recent news items.

Date	Topic
2018.01.22	(Jan 30 - 31) DDBJ services will be unavailable due to NIG Supercomputer maintenance
2018.01.18	Planned DRA submission system restoration: 22th, January (Monday)
2018.01.17	[correction] Structure of directories for WGS data on anonymous FTP site will be changed
2018.01.04	Release of genome data of California poppy ( <i>Eschscholzia californica</i> subsp. <i>californica</i> )
2017.12.25	Release of genome data of red seabream ( <i>Pagrus major</i> )
2017.12.22	D-way submission services will be unavailable soon (Dec. 22 19:00) (Only BioProject/BioSample submission services are resumed on Dec. 25)

\*En Décembre 2017, DDBJ archivait plus de 948 millions de séquences nucléiques

## BANQUES DE DONNÉES GÉNÉRALISTES

## BD DE SÉQUENCES PROTÉIQUES

**UniProt (*Universal Protein Resource*)\*:** Consortium regroupant les données de plusieurs banques de données protéiques (séquences protéiques annotées): SwissProt (banque suisse)- TrEMBL (*Traduced EMBL*) et PIR (*Protein Information Resource*)

<http://www.uniprot.org>

The screenshot displays the UniProt website interface. At the top, there's a navigation bar with the UniProt logo and a search bar. Below the navigation bar, a banner states the mission of UniProt. The main content area is divided into several sections:

- UniProtKB:** UniProt Knowledgebase, featuring Swiss-Prot (553,474) with manually annotated and reviewed records, and TrEMBL (73,711,881) with automatically annotated and not reviewed records.
- UniRef:** The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.
- UniParc:** UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.
- Proteomes:** A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.
- Supporting data:** Literature citations, Taxonomy, Subcellular locations, Cross-ref. databases, Diseases, and Keywords.
- Getting started:** Text search, BLAST, Sequence alignments, Retrieve/ID mapping, and Peptide search.
- UniProt data:** Download latest release, Statistics, How to cite us, Submit your data, and SPARQL.
- Protein spotlight:** Out Of The Ordinary (January 2017) featuring a protein involved in heart development.
- News:** Forthcoming changes, UniProt release 2017\_01, UniProt release 2016\_11, and UniProt release 2016\_10.

\*En Décembre 2017, Uniprot archivait: >102 millions (TrEMBL) et > 556 mille (UniProtKB/Swiss-Prot) séquences protéiques annotées automatiquement et manuellement, non revues et revues, respectivement. 4



## BANQUES DE DONNÉES GÉNÉRALISTES

## BD DE SÉQUENCES PROTÉIQUES

PDB (*Protein Data Bank*) \*:

Structure 3D de protéines, acides nucléiques et autres molécules

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

**WORLDWIDE PDB PROTEIN DATA BANK**

VALIDATION → DEPOSITION → DATA DICTIONARIES → DOCUMENTATION → TASK FORCES → STATISTICS → ABOUT → wwPDB Foundation

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Learn more about PDB **HISTORY** and **FUTURE**.

**Validate Structure**  
or View validation reports

**Deposit Structure**  
All Deposition Resources

**Download Archive**  
Instructions

**Vision and Mission**

**Vision**

Sustain a freely accessible, single global archive of experimentally determined structure data for biological macromolecules as an enduring public good.

**Mission**

- Ensure open access to public domain experimentally determined structural biology data.
- Provide expert deposition, validation, and biocuration services at no charge to Data Depositors.
- Enable universal access for expert and non-expert Data Consumers with no limitations on usage.
- Manage the PDB archive as a public good according to the FAIR Principles.
- Lead the world in structural biology data representation, exchange, and visualization.

**wwPDB Members**

**wwPDB Resources**

**Data Dictionaries**

- Macromolecular Dictionary (PDBx/mmCIF)
- Small Molecule Dictionary (CCD)
- Peptide-like antibiotic and inhibitor molecules (BIRD)

**Annotation**

- Procedures and policies
- Improvements for consistency and accuracy

**Community Input:**  
**Task Forces and Working Groups**

- Validation Task Forces (X-ray, NMR, 3DEM)
- Small Angle Scattering Task Force
- PDB/mmCIF Working Group
- Hybrid/Integrative Methods Task Force
- Ligand Validation Workshop

**News & Announcements**

**01/09/2018**

► **Time-stamped Copies of the PDB Archive Available**

A snapshot of the PDB archive (<ftp://ftp.wwpdb.org>) as of January 1, 2018 has been added to <ftp://snapshots.wwpdb.org/> and <ftp://snapshots.pdbj.org/>. Snapshots have been archived annually since January 2005 to provide readily identifiable data sets for research on the PDB archive.

[Read more](#)

**12/01/2017**

► **Overview of PDB Validation Reports published in Structure**

[The paper \(doi:\)](#)

**PDB Data Growth & Usage Statistics**

\*Le 16 Janvier 2018, PDB archivait près de 141 mille structures 3D, obtenues essentiellement par expérimentation (cristallographie à rayon X, spectroscopie RMN, Cryo-microscopie électronique, etc.).



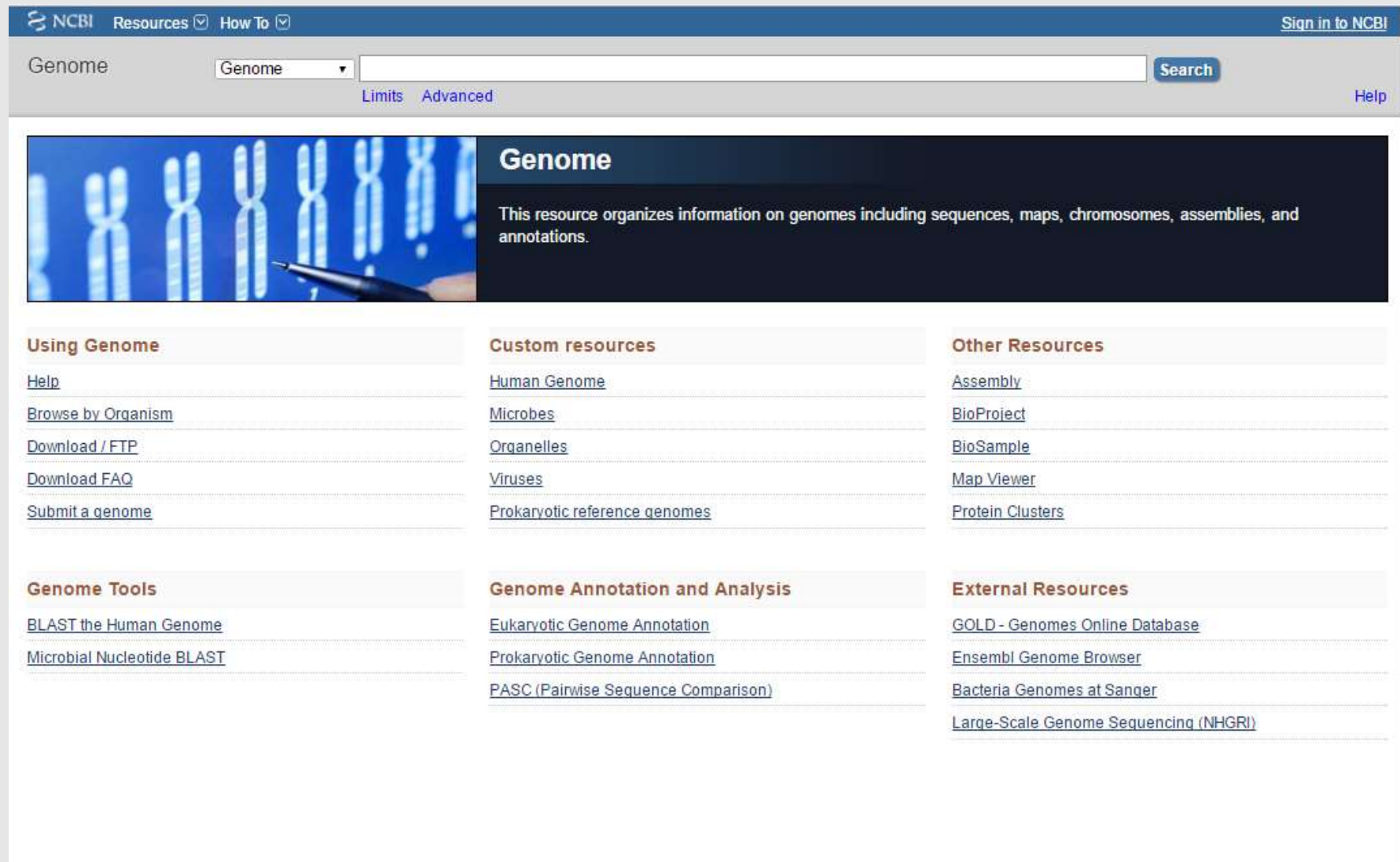
## BANQUES DE DONNÉES GÉNÉRALISTES

## BD DE SÉQUENCES GÉNOMIQUES

## Genome

Banque de génomes annotés appartenant à plus de 34 000 organismes (eucaryotes, procaryotes, virus, plasmides, organelles) (données de janvier 2018)

<https://www.ncbi.nlm.nih.gov/genome>



NCBI Resources How To Sign in to NCBI

Genome Genome Limits Advanced Search Help

## Genome

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

### Using Genome

- [Help](#)
- [Browse by Organism](#)
- [Download / FTP](#)
- [Download FAQ](#)
- [Submit a genome](#)

### Genome Tools

- [BLAST the Human Genome](#)
- [Microbial Nucleotide BLAST](#)

### Custom resources

- [Human Genome](#)
- [Microbes](#)
- [Organelles](#)
- [Viruses](#)
- [Prokaryotic reference genomes](#)

### Genome Annotation and Analysis

- [Eukaryotic Genome Annotation](#)
- [Prokaryotic Genome Annotation](#)
- [PASC \(Pairwise Sequence Comparison\)](#)

### Other Resources

- [Assembly](#)
- [BioProject](#)
- [BioSample](#)
- [Map Viewer](#)
- [Protein Clusters](#)

### External Resources

- [GOLD - Genomes Online Database](#)
- [Ensembl Genome Browser](#)
- [Bacteria Genomes at Sanger](#)
- [Large-Scale Genome Sequencing \(NHGRI\)](#)

## 2. Les banques de données spécialisées

Ces banques contiennent des données homogènes, sont établies autour :

### ☐ D'une thématique

- Banques spécialisées dans certaines voies métaboliques, de structures particulières, d'expression de gènes, etc.

### ☐ D'un organisme

- Génome d'*Arabidopsis thaliana*, génome d'*Escherichia coli*, etc.

## BANQUES DE DONNÉES SPÉCIALISÉES

## EXEMPLES

# neXtprot :

## Banque de protéines humaines

<https://www.nextprot.org/>

neXtprot

Tools ▾ Portals ▾ Download Help ▾ About ▾ Contact

Login

neXtprot

Exploring the universe of human proteins

☒ Simple search ☐ Advanced search

proteins ▾ Gold only ▾

e.g.: Search for MSH6 in proteins, Search for author Doolittle in publications, Search for liver in terms

► Getting started

- » The human proteome
- » Simple search
- » Advanced search
- » Explore
- » Analyze
  - BLAST, list management
- » Download
  - XML, FASTA, PEFF
- » Technical corner
  - Viewers, API, SPARQL

Data sources

THE HUMAN PROTEIN ATLAS

SRAtlas IntAct Bgee PeptideAtlas COSMIC UniProt-GOA UniProt

Release contents

News

- » neXtProt in ExPASy, tweaking the pep...  
Jan 08, 2018
- » PEFF 1.0 format implemented  
Oct 25, 2017
- » New release focusing on expression  
Sep 06, 2017

News archive

Release 2017-08-01

Protein existence in neXtProt

Predicted (71)  
 Uncertain (570)  
 Inferred from homology (478)  
 Evidence at transcript level (1912)  
 Evidence at protein level (17168)

Release statistics

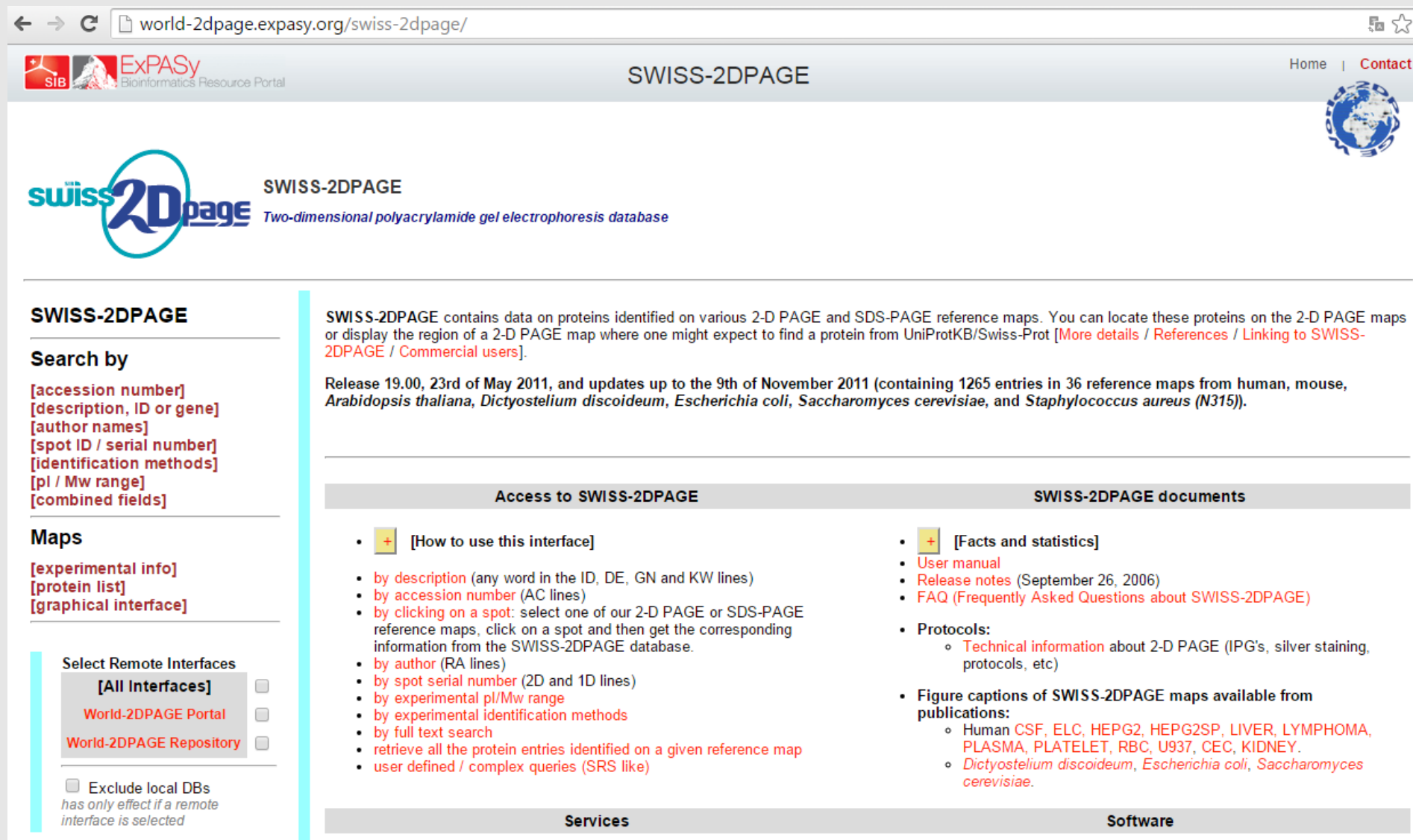
## CATÉGORIES DE BANQUES DE DONNÉES BIOLOGIQUES

## EXEMPLES

## SWISS-2DPAGE :

Base de données de protéines identifiées par électrophorèse bidimensionnelle

<http://world-2dpage.expasy.org/swiss-2dpage/>



← → ↻ world-2dpage.expasy.org/swiss-2dpage/

SIB ExPASy Bioinformatics Resource Portal

SWISS-2DPAGE

Home | Contact

**swiss2Dpage** SWISS-2DPAGE  
Two-dimensional polyacrylamide gel electrophoresis database

---

**SWISS-2DPAGE**

**Search by**

- [accession number]
- [description, ID or gene]
- [author names]
- [spot ID / serial number]
- [identification methods]
- [pI / Mw range]
- [combined fields]

**Maps**

- [experimental info]
- [protein list]
- [graphical interface]

**Select Remote Interfaces**

**[All Interfaces]** ☐

World-2DPAGE Portal ☐

World-2DPAGE Repository ☐

☐ Exclude local DBs  
has only effect if a remote interface is selected

SWISS-2DPAGE contains data on proteins identified on various 2-D PAGE and SDS-PAGE reference maps. You can locate these proteins on the 2-D PAGE maps or display the region of a 2-D PAGE map where one might expect to find a protein from UniProtKB/Swiss-Prot [[More details](#) / [References](#) / [Linking to SWISS-2DPAGE](#) / [Commercial users](#)].

Release 19.00, 23rd of May 2011, and updates up to the 9th of November 2011 (containing 1265 entries in 36 reference maps from human, mouse, *Arabidopsis thaliana*, *Dictyostelium discoideum*, *Escherichia coli*, *Saccharomyces cerevisiae*, and *Staphylococcus aureus* (N315)).

---

Access to SWISS-2DPAGE	SWISS-2DPAGE documents
<ul style="list-style-type: none"> <li> <b>[How to use this interface]</b> <ul style="list-style-type: none"> <li>by description (any word in the ID, DE, GN and KW lines)</li> <li>by accession number (AC lines)</li> <li>by clicking on a spot: select one of our 2-D PAGE or SDS-PAGE reference maps, click on a spot and then get the corresponding information from the SWISS-2DPAGE database.</li> <li>by author (RA lines)</li> <li>by spot serial number (2D and 1D lines)</li> <li>by experimental pI/Mw range</li> <li>by experimental identification methods</li> <li>by full text search</li> <li>retrieve all the protein entries identified on a given reference map</li> <li>user defined / complex queries (SRS like)</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li> <b>[Facts and statistics]</b> <ul style="list-style-type: none"> <li>User manual</li> <li>Release notes (September 26, 2006)</li> <li>FAQ (Frequently Asked Questions about SWISS-2DPAGE)</li> </ul> </li> <li><b>Protocols:</b> <ul style="list-style-type: none"> <li>Technical information about 2-D PAGE (IPG's, silver staining, protocols, etc)</li> </ul> </li> <li><b>Figure captions of SWISS-2DPAGE maps available from publications:</b> <ul style="list-style-type: none"> <li>Human CSF, ELC, HEPG2, HEPG2SP, LIVER, LYMPHOMA, PLASMA, PLATELET, RBC, U937, CEC, KIDNEY.</li> <li><i>Dictyostelium discoideum</i>, <i>Escherichia coli</i>, <i>Saccharomyces cerevisiae</i>.</li> </ul> </li> </ul>

---

Services	Software



## BANQUES DE DONNÉES SPÉCIALISÉES

## EXEMPLES

**MGI (*Mouse Gene Informatics*) :**  
 banque de séquences nucléotidiques de la souris  
<http://www.informatics.jax.org/>

The screenshot displays the MGI website homepage. At the top, the MGI logo (a mouse head) and the text "Mouse Genome Informatics" are visible. A navigation bar includes links for "About", "Help", "FAQ", "Search", "Download", "More Resources", "Submit Data", "Find Mice (IMSR)", "Analysis Tools", "Contact Us", and "Browsers".

The main content area is divided into several sections:


- Search Section:** A "QuickSearch" box for "Keywords, Symbols, or IDs". Below it, a list of "topic specific search and analysis tools" includes: Genes, Phenotypes & Mutant Alleles, Human-Mouse: Disease Connection, Gene Expression Database (GXD), Recombinase (cre), Function, Strains, SNPs & Polymorphisms, Vertebrate Homology, Mouse Models of Human Cancer, Pathways, Batch Data and Analysis Tools, and Nomenclature.
- Getting Started:** A section with links for "Introduction to mouse genetics", "How to use MGI (Text & Video)", and "Cre Portal Tutorial".
- Community Interest:** A section at the bottom for community updates.
- Right Sidebar:**
  - A descriptive paragraph: "MGI is the international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease." with links to "About Us" and "MGI Publications".
  - A banner for the "ALLIANCE of GENOME RESOURCES" announcing the "AGR 1.0 release: Search for gene, disease & QO data from 6 Model Organism Databases & the GO Consortium". It features a detailed entry for "Parkinson's disease (DOID:143307)".
  - A "What's new at MGI" section, updated October 20, 2017, listing recent updates and new data imports.
  - Links for "MGI Statistics" and "More MGI news".




## Taxonomy

Banque de données taxonomiques de plus de 547 000 organismes génétiquement identifiées  
(données de Janvier 2018)

<https://www.ncbi.nlm.nih.gov/taxonomy>

 NCBI Resources ☒ How To ☒ [Sign in to NCBI](#)

Taxonomy    
[Limits](#) [Advanced](#) [Help](#)



### Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

#### Using Taxonomy

- [Quick Start Guide](#)
- [FAQ](#)
- [Handbook](#)
- [Taxonomy FTP](#)

#### Taxonomy Tools

- [Browser](#)
- [Common Tree](#)
- [Statistics](#)
- [Name/ID Status](#)
- [Genetic Codes](#)
- [Linking to Taxonomy](#)
- [Extinct Organisms](#)

#### Other Resources

- [GenBank](#)
- [LinkOut](#)
- [E-Utilities](#)
- [Batch Entrez](#)
- [INSDC](#)

## BANQUES DE DONNÉES GÉNÉRALISTES

## BD BIBLIOGRAPHIQUES

- ✓ **Medline** (*Medical Literature Analysis and Retrieval System Online*) : la plus grande base de données bibliographiques de littérature relative aux sciences biologiques et médicales (biologie, biochimie, médecine clinique, pharmacologie, psychiatrie, toxicologie, etc.), gérée par la bibliothèque nationale de médecine des Etats Unis d'Amérique (NLM).

<http://www.ncbi.nlm.nih.gov/pubmed>

The screenshot displays the PubMed homepage. At the top, there's a navigation bar with 'NCBI Resources' and 'How To' links, along with a 'Sign in to NCBI' button. Below this is the 'PubMed' logo and a search bar with a dropdown menu set to 'PubMed' and a 'Search' button. A 'Help' link is also present. The main content area features a large banner image of books and a text box stating: 'PubMed comprises more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.'

Below the banner, there are three columns of links and featured content:

- Using PubMed:**
  - PubMed Quick Start Guide
  - Full Text Articles
  - PubMed FAQs
  - PubMed Tutorials
  - New and Noteworthy
- PubMed Tools:**
  - PubMed Mobile
  - Single Citation Matcher
  - Batch Citation Matcher
  - Clinical Queries
  - Topic-Specific Queries
- More Resources:**
  - MeSH Database
  - Journals in NCBI Databases
  - Clinical Trials
  - E-Utilities (API)
  - LinkOut

At the bottom, there are three sections:

- Latest Literature:** Lists new articles from highly accessed journals, including 'Am J Clin Nutr (8)', 'Am J Orthod Dentofacial Orthop (35)', 'Cochrane Database Syst Rev (7)', 'J Biol Chem (44)', 'JAMA (1)', 'Lancet (7)', 'N Engl J Med (8)', 'Nature (44)', 'Proc Natl Acad Sci U S A (25)', and 'Science (85)'. It also mentions 'Try PubMed Journals, our new experimental feature for following journals of interest to you.' and 'PubMed Journals'.
- Trending Articles:** Lists PubMed records with recent increases in activity, including 'Clinical outcomes of a scapular-focused treatment in patients with subacromial pain syndrome: a systematic review. Br J Sports Med. 2016.', 'Effectiveness of neuromuscular taping on painful hemiplegic shoulder: a randomised clinical trial. Disabil Rehabil. 2016.', 'Rare and low-frequency coding variants alter human adult height. Nature. 2017.', 'McConnell's patellar taping does not alter knee and hip muscle activation differences during proprioceptive exercises: A randomized placebo-controlled trial in women with patellofemoral pain syndrome. J Electromyogr Kinesiol. 2016.', and 'Parvovirus B19 during pregnancy: a review. J Prenat Med. 2010.'.
- PubMed Commons:** Lists featured comments, including 'Circuits in reward & aversion: @gstuber posts journal club review of study identifying neuron populations. bit.ly/2ZG9mN Feb 2', 'Migrating database: @odsouthan provides updated info for finding pharmacological data resource. bit.ly/2kidWnC Jan 31', 'Genes for lactation persistence in cattle: @Eric\_Fauman critiques findings of a genome-wide association study. bit.ly/2RLtsB Jan 30', 'Reviewing replication: R Tibshirani critiques statistics; A Collings cross-posts comment from original study authors bit.ly/24U3H Jan 27', and 'Evaluating connection between food energy supply & obesity: @JamesonVoss discusses ecologic study. bit.ly/2PxpU Jan 29'.

Fin 2016, cette base de données contenait plus de 23,5 millions de citations, publiées depuis 1948 dans environ 5100 revues biomédicales en 60 langues différentes

**AUTRES EXEMPLES DE BANQUES DE DONNÉES BIOLOGIQUES**

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZy, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DGP, DictyDb, Picty\_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, GeneCards, Genline, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HeXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISSMODEL Repository, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, YEPD, YPD, YPM, etc.

## Définition d'un format

**Ensemble des règles de présentation** auxquelles sont soumises la ou les séquences dans un fichier donné. Ce qui permet :

- une mise en forme automatisée;
- le stockage homogène de l'information;
- le traitement informatique ultérieur de l'information par les logiciels de gestion des banques.

## LES BANQUES DE DONNÉES BIOLOGIQUES

## SYNTAXE D'UNE ENTRÉE

Exemple des banques de données nucléiques

Contient trois parties :

- 1- Description générale de la séquence
- 2- Features : Description des objets biologiques présents sur la séquence, destinées au système de gestion
- 3- La séquence

**Description générale de la  
séquence**

**« Features »**  
**Description des objets  
biologiques présents sur la  
séquence**

### La séquence

```
ctcggcagc ccgaggtcat cctgctagac tcagacctgg atgaacccat agacttgccg      60
tcggtcaaga gccgcagcga ggccggggag ccgccagct cctccaggt gaagcccag      120
Acaccggcgt Cggcggcggt Ggcggtggcg Gcggcagcgg Caccaccac Gacggcggag      180
```



# Saccharomyces cerevisiae strain JZ1C invertase (SUC2) gene, complete cds

GenBank: JQ836661.1

[FASTA](#) [Graphics](#)

LOCUS JQ836661 1599 bp DNA linear PLN 26-DEC-2012  
DEFINITION Saccharomyces cerevisiae strain JZ1C invertase (SUC2) gene, complete cds.  
ACCESSION JQ836661  
VERSION JQ836661.1 GI:393395465  
KEYWORDS .  
SOURCE Saccharomyces cerevisiae (baker's yeast)  
ORGANISM [Saccharomyces cerevisiae](#)  
Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomycetes.  
REFERENCE 1 (bases 1 to 1599)  
AUTHORS Wang,S.A. and Li,F.L.  
TITLE Invertase SUC2 is the Key Hydrolase for Inulin Degradation in Saccharomyces cerevisiae  
JOURNAL Appl. Environ. Microbiol. 79 (1), 403-406 (2013)  
PUBMED [23104410](#)  
REFERENCE 2 (bases 1 to 1599)  
AUTHORS Wang,S.-A. and Li,F.-L.  
TITLE Direct Submission  
JOURNAL Submitted (27-MAR-2012) Key Laboratory of Biofuels, Qingdao Institute of Bioware and Bioprocess Technology, 189 Songling Road, Qingdao, Shandong 266101, China

FEATURES Location/Qualifiers  
source 1..1599  
/organism="Saccharomyces cerevisiae"  
/mol\_type="genomic DNA"  
/strain="JZ1C"  
/db\_xref="taxon:4932"  
<1..>1599  
/gene="SUC2"  
<1..>1599  
/gene="SUC2"  
/product="invertase"  
1..1599  
/gene="SUC2"  
/codon\_start=1  
/product="invertase"  
/protein\_id="AFN08663.1"  
/db\_xref="GI:393395466"  
/translation="MLIQAFLLAGFAAKISASMYNETSDRPLVHFTPNKGMNDPNGLWYEDKDAKWHLYFYHNDTVMGTPLFWGHATSDDLTHWDEPIAIAPKRNDGGAFSGSMVVDYHNTSGFFNDTIDPQKVAIWNTDSESEHQYISYSLDGGGTFPTYQKNPVLAASTQPHDPKYPWFPSQKWMIAKSSQOYKIEIYSSDOLSKWLESAPANEGLGTQYECPLIEVPTGQPSKSYWVFI SINDGAPAGGSSFNQYFVGSYNGTHFAFDNGSRVVDPGEDYALQTFPTDSTYGSALCIANASHWIYSAFVFTNFWRSMSLVRKFSLENTYEQANPTELINLKAEPILNISNAGPMSRFATNTITLHANSYNVDLSNSTGTLEFELVYAVHTQTITISKSVFPDLSEMFEGLEDPEYLSNGFEASASSFFLDGNSKVFVEKNPFTNRMSSVNNQPFKSENDLSYKVKYGLLDQNLILELYFNDGQVVTNTYFMTGNALGSVMMITGYNLI FYLSEFQVDEK"

ORIGIN  
1 atgctttttg aagcttttct tttctttttg gctggttttg cagccaaaat atctgcatca  
61 atgcaaaaag aaactagcga tagacatttg gtccacttca cccccaaa cggctggatg  
121 aatgacccaa atgggttttg gtagcatgaa aaagatgcca atggcatctc gtactttcaa  
181 tacaacccaa atgacacagt atgggttagc cctttgtttt ggggccatgc tacttcgat  
241 gatttgactc attggaaga tgaacccatt gctatgcttc ccaagcgtaa cgaattcagg  
301 gcttttctcg gctccatggt ggttgattac aacacacga gttgtttt caatgatac  
361 attgacccaa gacaaagatg cgtttcgatt tgcatttata acactcttga aagtgaagag  
421 caatacattg gctattctct tcatggttgc taccatttta ctgaatacca aaagaacctt  
481 gtttttagctt ccaactccac tcaattcaga gatccaaagg tgtttcggta tgaaccttct  
541 caaaaatgga ttatgacgpc tgcacaaatc caagactaca aaattgaat ttactctct  
601 gatgacttga agtactggaa gctgaactct gcttttgata atgaaggttt cttaggctac  
661 caatatgaat gtccaggttt gatgaaagt ccaactgagc aagatctctc caaatcttat  
721 tgggtcatgt ttatttctat caatccaggt gacactgctg gcggttcttt caacctaat  
781 ttgtttggtt ctttcaatgg tactcatttt gaagcgtttg acactcaatc tagagtgtga  
841 gatttttgga aggaactata tgcattgcaa acttttctta acacagacc accgtacggt  
901 tcaagcattag gattgctgtt ggtttcaaac tgggagtaca gtgcttttgt cccaactaac  
961 ccatggagat catccatgtc ttgtgtccgc aagttttctt tgaadactga atactcaagt  
1021 aatccagaga ctgaattgat caatttgaaa gccgaaccaa tattgaacct tagtaattgt  
1081 ggttacttgt ctgtttttgc tactaacaca acttcaacta aggcacattc ttacaaatgc  
1141 gatttgagca actcagctgg taactatagg tttaggttgg tttaagctgt taacacaaa  
1201 caaacctat ccaactcgtt ctltcccgac ttatcacttt ggttcaaggg tttagaagat  
1261 cttagaagat atttaagaat ggggttttga gccagtgctt ctctcttctt tttagccggt  
1321 ggttaactta aggtcaagtt tgcacagagg aacccatttt tcaacaaag aatgtctgtc  
1381 aacacaccaa ctttcaagtc tgaagacgac ctgaattact ataaagtga cgccttactg  
1441 gatcaaaaaa ttttgaattt gtacttcaac gatggagatg tggtttctac aatatcctac  
1501 ttctagacca ccgtaaacgc tctaggatct gtgaactaga ccactggtgt cgataatttg  
1561 ttctacattg acaagttcca agtaagggaa tgaacatag

## Description générale de la séquence

### « Features »

## Description des objets biologiques présents sur la séquence

- Chaque ligne commence par un mot-clé

## La séquence

- Deux lettres pour EMBL
- Maximum 12 lettres pour Genbank et DDBJ

- Fin d'une entrée : //

## Description générale de la séquence (Genbank)

LOCUS JQ836661 1599 bp DNA linear PLN 26-DEC-2012  
 DEFINITION *Saccharomyces cerevisiae* strain JZ1C invertase (SUC2) gene,  
 complete cds.  
 ACCESSION JQ836661  
 VERSION JQ836661.1 GI:393395465  
 KEYWORDS .  
 SOURCE *Saccharomyces cerevisiae* (baker's yeast)  
 ORGANISM [Saccharomyces cerevisiae](#)  
 Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;  
 Saccharomycetes; Saccharomycetales; Saccharomycetaceae;  
 Saccharomyces.  
 REFERENCE 1 (bases 1 to 1599)  
 AUTHORS Wang,S.A. and Li,F.L.  
 TITLE Invertase SUC2 Is the Key Hydrolase for Inulin Degradation in  
*Saccharomyces cerevisiae*  
 JOURNAL Appl. Environ. Microbiol. 79 (1), 403-406 (2013)  
 PUBMED [23104410](#)  
 REFERENCE 2 (bases 1 to 1599)  
 AUTHORS Wang,S.-A. and Li,F.-L.  
 TITLE Direct Submission  
 JOURNAL Submitted (27-MAR-2012) Key Laboratory of Biofuels, Qingdao  
 Institute of Bioenergy and Bioprocess Technology, 189 Songling  
 Road, Qingdao, Shandong 266101, China

## « Features »

### Description des objets biologiques présents sur la séquence

FEATURES	Location/Qualifiers
source	1..1599 /organism="Saccharomyces cerevisiae" /mol_type="genomic DNA" /strain="JZ1C" /db_xref="taxon: <a href="#">4932</a> "
<a href="#">gene</a>	<1..>1599 /gene="SUC2"
<a href="#">mRNA</a>	<1..>1599 /gene="SUC2" /product="invertase"
<a href="#">CDS</a>	1..1599 /gene="SUC2" /codon_start=1 /product="invertase" /protein_id=" <a href="#">AFN08663.1</a> " /db_xref="GI:393395466" /translation="MLLQAFLLFLLAGFAAKISASMTNETSDRPLVHFTPNKGWMNDPN GLWYDEKDAKWHLYFQYNPNDTVWGTPLEFWGHATSDDLTHWEDEPIAIAPKRND SGAF SGSMVVDYNNNTSGFFNDTIDPRQRCVAIWYNTPESEEQYISYSLDGGYTFTEYQKNP VLAANSTQFRDPKVFWYEPSQKWIMTAAKSQDYKIEIYSSDDLKSWKLESAFANEGFL GYQYECPLIEVPTEQDPSKSYWVMFISINPGAPAGGSFNQYFVGSFNGTHFEAFDNQ SRVVDGKDYALQTFNTDPTYGSALGIAWASNWEYS AFVPTNPWRSSMSLVKRFSL NTEYQANPETELINLKAEPILNISNAGPWSRFATNTTLTKANSYNVDLSNSTGTLEFE LVYAVNTTQTISKSVFPDLSLWFKGLEDP E EYLRMGFEASASSFFLDRGNSKVKE NPYFTNRMSVNNQPFKSENDLSYYKVYGLLDQNILELYFNDGDVVSTNTYFMTTGNAL GSVNMTTGVDNLFYIDKFQVREVK"



## La séquence (format Genbank)

ORIGIN

```

1  atgctttttgc aagcttttcct tttcctttttg gctgggttttg cagccaaaat atctgcatca
61  atgacaaacg aaactagcga tagacctttg gtccacttca caccacaaca gggctggatg
121  aatgacccaa atgggtttgtg gtacgatgaa aaagatgcca aatggcatct gtactttcaa
181  tacaacccaa atgacaccgt atgggggtacg ccattgtttt ggggccatgc tacttccgat
241  gatttgactc attgggaaga tgaaccatt gctatcgctc ccaagcgtaa cgattcaggt
301  gctttctctg gctccatggg ggttgattac aacaacacga gtgggttttt caatgatact
361  attgatccaa gacaaagatg cgttgcgatt tggacttata acactcctga aagtgaagag
421  caatacatta gctattctct tgatgggtgg taccctttta ctgaatacca aaagaaccct
481  gtttttagctg ccaactccac tcaattcaga gatccaaagg tgttctggta tgaaccttct
541  caaaaatgga ttatgacggc tgccaaatca caagactaca aaattgaaat ttactcctct
601  gatgacttga agtcctggaa gctagaatct gcatttgcta atgaagggtt cttaggctac
661  caatatgaat gtccagggtt gattgaagtc ccaactgagc aagatccttc caaatcctat
721  tgggtcatgt ttatttctat caatccagggt gcacctgctg gcggttcctt caaccaatat
781  tttgttggtt ccttcaatgg tactcatttt gaagcggttg acaatcaatc tagagtggta
841  gatttttggtt aggactacta tgccttgcaa actttcttca acacagaccc aacgtacggg
901  tcagcattag gtattgcctg ggcttcaaac tgggagtaca gtgcctttgt cccaactaac
961  ccatggagat catccatgtc tttgggtccgc aagttttctt tgaacactga atatcaagct
1021  aatccagaga ctgaattgat caatttgaaa gccgaaccaa tattgaacat tagtaatgct
1081  ggtccctggg ctcgttttgc tactaacaca actctaacta aggccaattc ttacaatgtc
1141  gatttgagca actcgactgg taccctagag tttgagttgg tttacgctgt taacaccaca
1201  caaaccatat ccaaatccgt ctttcccgac ttatcacttt ggttcaaggg tttagaagat
1261  cctgaagaat atttaagaat gggttttgaa gccagtgtt cttccttctt tttggaccgt
1321  ggtaactcta aggtcaagtt tgtcaaggag aacccatatt tcacaaacag aatgtctgtc
1381  aacaaccaac cattcaagtc tgagaacgac ctaagttact ataaagtgtc cggcctactg
1441  gatcaaaaca tcttggaatt gtacttcaac gatggagatg tggtttctac aaatacctac
1501  ttcatgacca ccggtaacgc tctaggatct gtgaacatga ccactgggtg cgataatttg
1561  ttctacattg acaagttcca agtaagggaa gtaaaatag

```

## Quelques formats de données biologiques

- ✓ Format des banques, exemples :
  - Séquences ADN/ARN : EMBL, GenBank et DDBJ
  - Séquences protéiques : Uniprot, SwissProt et TrEMBL, PIR...
  - Format PHYLIP (*PHYLogeny Inference Package*), FOSN (*Files Of Sequence Names*), RSF (*Rich Sequence Format files*), RSF (*Rich Sequence Format files*), MSF (*Multiple Sequence Format*), etc.
- ✓ Formats lus par la plupart des outils en bioinformatique
  - FASTA
  - Séquence brute (*raw sequence*)

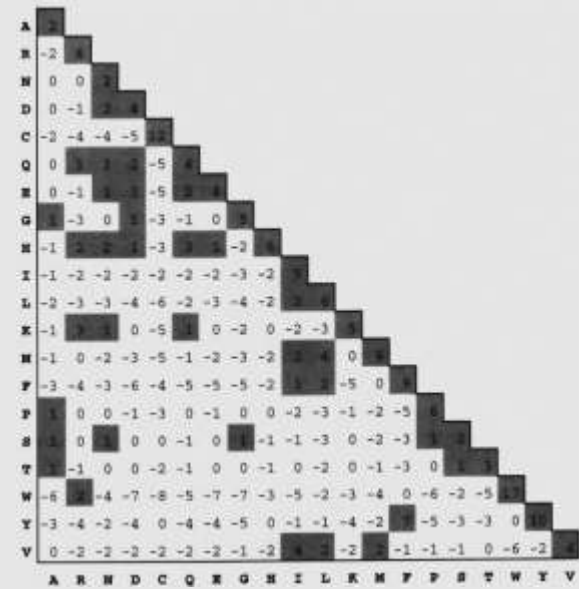
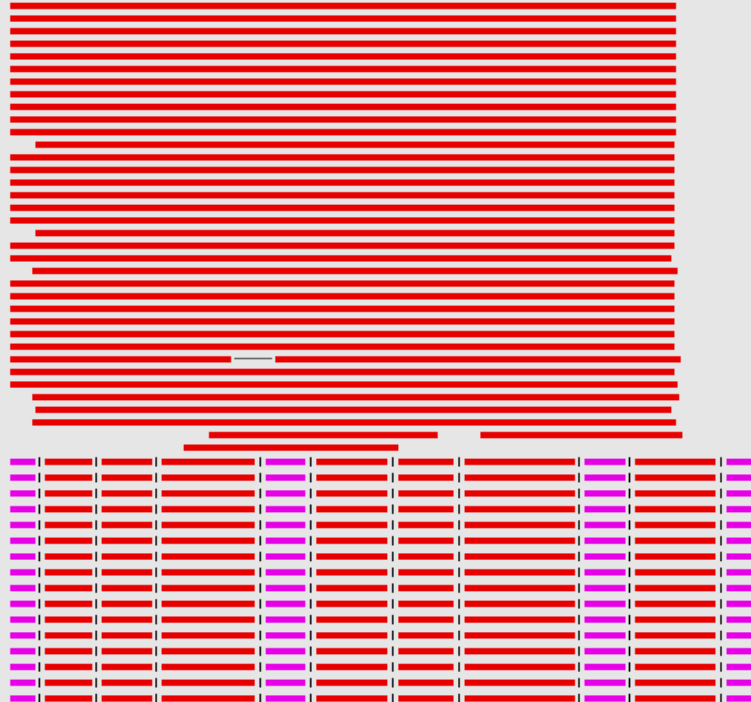
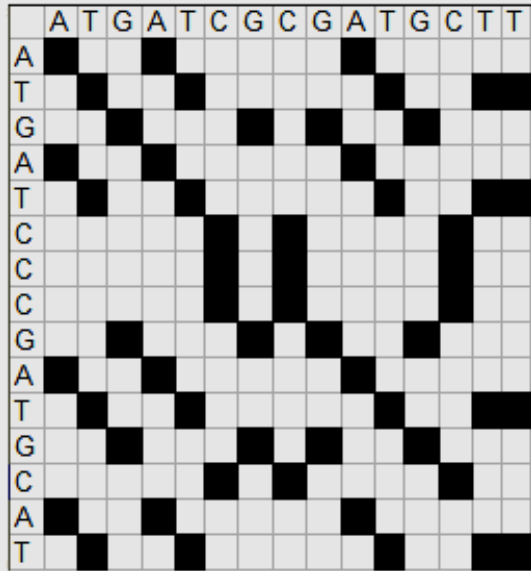


## Le format FASTA

- ✓ Une ligne de commentaires précédé de « > »
- ✓ La séquence brute (pas d'espace, ni de nombre)

```
>Human Polycomb 2 homolog (hPc2) mRNA, partial cds  
ctccggcagcccgagggtcatcctgctagactcagacctggatgaacccat  
agacttgcgctcgggtcaagagccgcagcgaggccggggagccgcccagct  
ccctccagggtgaagcccgagacaccggcgctcggcgggcggtggcggtggcg  
gcggcagcgggcaccacacgacggcgggagagaagcctccagccgaggccca  
ggacgaacctgcagagtcgctgagcgagttcaagcccttctttgggaata  
taattatcacccgacgtcacccgcgaactgcctcacccgttactttcaaggag  
tacgtgacggtg
```

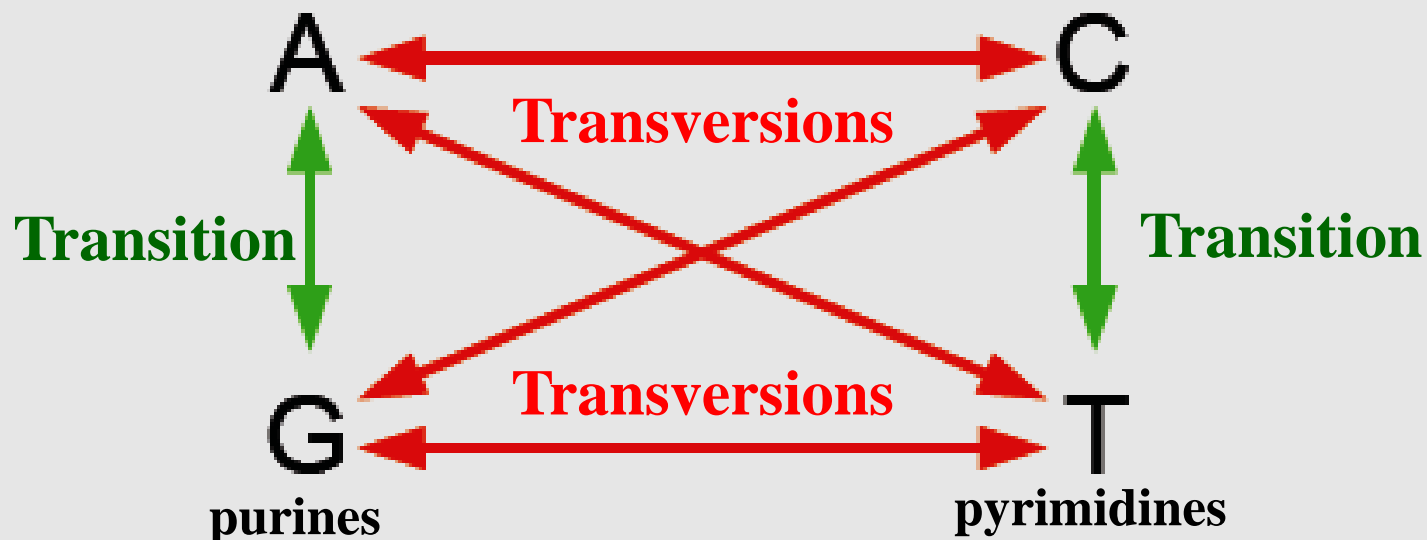
# LES ALIGNEMENTS



## ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

## INTRODUCTION

- L'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leurs séquences
- Evolution des gènes par mutations, deux types « locaux » sont intéressants:
  - ✓ les insertions-délétions ou indels (apparition ou disparition d'une nucléotide);  
ATCTCG**N**CTATC
  - ✓ les substitutions (remplacement d'une nucléotide par une autre)



**ALIGNEMENT DES SÉQUENCES: DÉFINITIONS**

- ✓ En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines) repose essentiellement sur la notion d'**alignement**;
- ✓ L'**alignement** est une opération qui vise à identifier des zones communes entre un groupe de  $k$  séquences;

Ce qui pourrait indiquer que :

- ☐ La structure (primaire, secondaire ou tertiaire) des séquences est semblable;
- ☐ La fonction biologique est proche ou différente;
- ☐ L'origine des séquences alignées est commune ou éloignée (notion homologie).

## ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

- ✓ Selon la taille des séquences comparées on distingue deux types d'alignements:

## *Alignement local ou global ?*

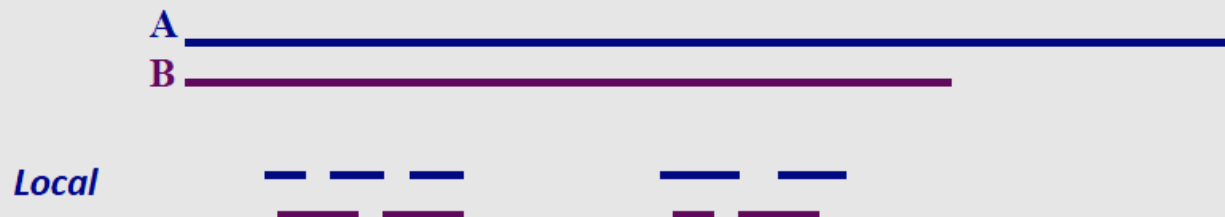


Finalités différentes

L'alignement **global** est conçu pour comparer des séquences apparentées sur toute leur longueur.



L'alignement **local** est conçu pour rechercher dans la séquence A des régions semblables à la séquence B (ou à des parties de la séquence B)





**ALIGNEMENT DES SÉQUENCES: DÉFINITIONS**

✓ Selon le nombre de séquences comparées on distingue deux types:

- par paires : on aligne 2 séquences
- multiple : on aligne plus de 2 séquences

# Applications

- étude phylogénétique;
- étude comparative des génomes;
- prédiction de gène;
- prédiction de la structure et fonction des ARN;
- prédiction de la structure 2D/3D des protéines;
- caractérisation de la fonction des protéines;
- .....

## ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

## NOTION DE SIMILARITÉ, D'IDENTITÉ ET D'HOMOLOGIE

Il existe plusieurs termes permettant de nommer la ressemblance entre deux séquences biologiques:

- ✓ La **similarité** est une quantité qui se mesure en % d'**identité**, l'identité elle-même peut être définie comme une ressemblance parfaite entre deux séquences.
- ✓ L'**homologie** une propriété (évolutive) des séquences: deux séquences sont dites homologues si elles possèdent un ancêtre commun. L'homologie présente la particularité d'être transitive. Si A est homologue à B et B homologue à C, alors A est homologue à C même si A et C se ressemblent très peu.

## ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

## NOTION DE SIMILARITÉ, D'IDENTITÉ ET D'HOMOLOGIE

✓ L'homologie se mesure par la similarité. On considère qu'une similarité significative (à partir de 30%) est signe d'homologie sauf si les séquences présentent une faible complexité. L'inverse n'est par contre pas vrai. Une absence totale de similarité ne veut pas dire non-homologie.



## ALIGNEMENT DE DEUX SÉQUENCES

## ALIGNEMENT OPTIMAL DE DEUX SÉQUENCES

- ✓ Afin de comparer deux séquences d'une manière objective (indépendante de l'observateur), on doit d'abord les aligner d'une manière optimale. L'alignement **optimal** est obtenu quand la **coïncidence des lettres** composant les deux séquences est **maximale** ;
- ✓ Un alignement optimal est une analyse qui permet de trouver le nombre minimum de mutations ponctuelles (insertion-délétion, substitution) qui permettent de transformer une séquence en une autre;

# LA DISTANCE

- Elle correspond au nombre d'indels et de substitutions séparant deux séquences A et B;
- En fonction de la quantité de mutations ponctuelles, la distance entre deux séquences A et B prend la forme suivante :

$$d(A, B) = \# \text{substitutions} + \# \text{indels}$$

# LE SCORE DE SIMILARITÉ

- Le score exprime le degré de similitude entre deux séquences :

$$S(A, B) = \# \text{identité} - (\# \text{substitutions} + \# \text{indels})$$

$$S(A, B) = \# \text{identité} - d(A, B)$$

- La construction de l'alignement consiste donc à identifier le meilleur alignement possible entre deux séquences, celui qui minimise la distance d'édition  $d(A, B)$  ou qui maximise le score  $S(A, B)$ .

## ALIGNEMENT DE DEUX SÉQUENCES

## EXEMPLES DE MATRICES DE SCORES

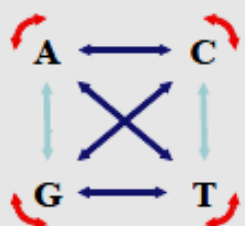
## Matrices de scores pour l'ADN

## ➤ La matrice identité

match → 1  
mismatch → 0

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

## ➤ La matrice de transition/transversion (substitutions)



Identité: 3  
Transition: 1  
Transversion: 0

	A	C	G	T
A	3	0	1	0
C	0	3	0	1
G	1	0	3	0
T	0	1	0	3

## ➤ La matrice identité dans BLAST

	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5



- ✓ La **programmation dynamique** est un outil facile et efficace pour trouver l'alignement **optimal** parmi tous les alignements **possibles**.
- ✓ Cette méthode utilise un algorithme développé par Needleman et Wunch (1970).

## ALIGNEMENT DE DEUX SÉQUENCES

## LA PROGRAMMATION DYNAMIQUE

✓ L'algorithme de Needleman et Wunsch permet de réaliser un alignement global entre deux séquences nucléiques. Son expression est de la forme :

$$S(i, j) = \text{Max} \begin{cases} S(i-1, j-1) + s(i, j) \\ S(i-1, j) \\ S(i, j-1) \end{cases}$$

## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH : EXEMPLE PRATIQUE

✓ Pour réaliser un alignement global des deux séquences suivantes de taille  $m$  et  $n$  respectivement ( $n$  et  $m$  peuvent être inégales):

$S1 = \text{TAAGTCCG}$   $m=8$  et  $S2 = \text{TAAGTACG}$   $n=8$

✓ Pour calculer l'alignement entre les deux séquences  $S1$  et  $S2$ , quatre étapes sont nécessaires :

## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH : EXEMPLE PRATIQUE

## ETAPE 1: CALCULE DE LA MATRICE INITIALE

- Il s'agit d'insérer les deux séquences S1 et S2 dans une matrice de sorte que S1 soit à l'horizontal et S2 à la verticale du tableau, puis remplir les cases par 1 (identité des deux nucléotides de S1 et de S2) ou 0 (sinon) :

	T	A	A	G	T	C	C	G
T	<b>1</b>	0	0	0	<b>1</b>	0	0	0
A	0	<b>1</b>	<b>1</b>	0	0	0	0	0
A	0	<b>1</b>	<b>1</b>	0	0	0	0	0
G	<b>0</b>	0	0	<b>1</b>	0	0	0	<b>1</b>
T	<b>1</b>	0	0	0	<b>1</b>	0	0	0
A	0	<b>1</b>	<b>1</b>	0	0	0	0	0
C	0	0	0	0	0	<b>1</b>	<b>1</b>	0
G	0	0	0	<b>1</b>	0	0	0	<b>1</b>

(matrice d'identité)



## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH : EXEMPLE PRATIQUE

**ETAPE 2: CALCULE DE LA MATRICE TRANSFORMÉE: INITIALISATION DE LA MATRICE**

- Nouvelle matrice ( $m+2$ ,  $n+2$ ) dans laquelle la 1ère ligne et la 1ère colonne sont initialisées à zéro ) :

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0								
A	0								
A	0								
G	0								
T	0								
A	0								
C	0								
G	0								

## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH : EXEMPLE PRATIQUE

## ETAPE 2: CALCULE DE LA MATRICE TRANSFORMÉE

- ✓ L'application de l'algorithme de Needleman et Wunsh permet de remplir les cases de cette matrice. Le résultat est le suivant :

$$S(i, j) = \text{Max} \begin{cases} S(i-1, j-1) + s(i, j) \\ S(i-1, j) \\ S(i, j-1) \end{cases}$$

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2
A	0	1	2	3	3	3	3	3	3
G	0	1	2	3	4	4	4	4	4
T	0	1	2	3	4	5	5	5	5
A	0	1	2	3	4	5	5	5	5
C	0	1	2	3	4	5	6	6	6
G	0	1	2	3	4	5	6	6	7

## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH : EXEMPLE PRATIQUE

## ETAPE 3: PARCOURS DE LA MATRICE TRANSFORMÉE

- ✓ Parcourir la matrice transformée depuis le plus haut score calculé (ici  $S=7$ ) jusqu'au score le plus petit (ici  $S=1$ ) :

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2
A	0	1	2	3	3	3	3	3	3
G	0	1	2	3	4	4	4	4	4
T	0	1	2	3	4	5	5	5	5
A	0	1	2	3	4	5	5	5	5
C	0	1	2	3	4	5	6	6	6
G	0	1	2	3	4	5	6	6	7

→		insertion dans i
		délétion dans j
↓		insertion dans j
		délétion dans i

## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH : EXEMPLE PRATIQUE

## ETAPE 4: ALIGNEMENT DES DEUX SÉQUENCES ET CALCUL DE SCORE

Séquence S1	T	A	A	G	T	—	C	C	G
						*		*	
Séquence S2	T	A	A	G	T	A	C	—	G

- ✓ Le score global de cet alignement est de 7.
- ✓ Le pourcentage de l'identité (la similarité) entre les deux séquences S1 et S2 est :

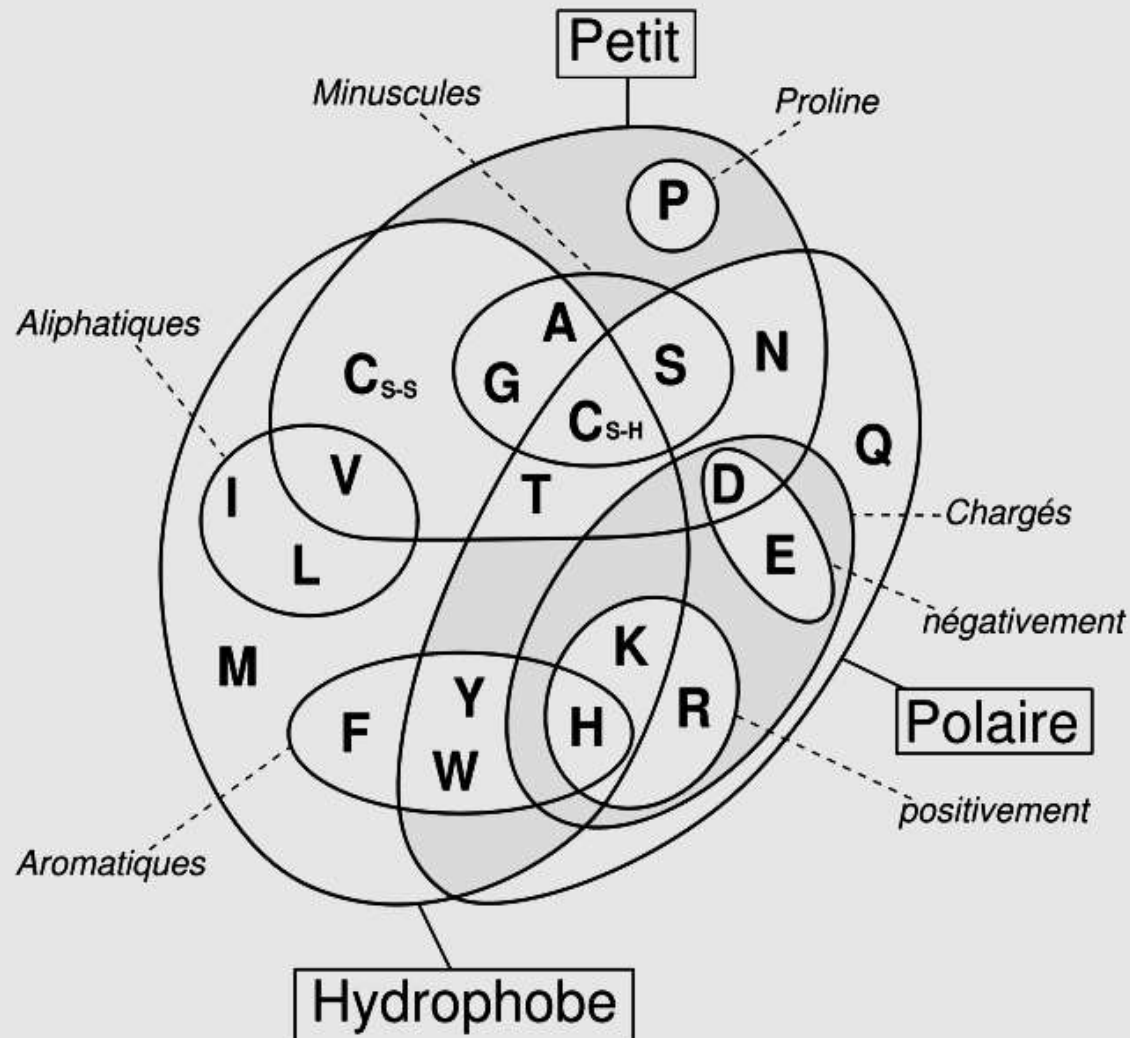
$$\%id = (7/9) * 100 = 77,78\%$$

## ALIGNEMENT DE DEUX SÉQUENCES

## PROGRAMMATION DYNAMIQUE POUR LES SÉQUENCES PROTÉIQUES

Code international des acides aminés selon l'IUPAC (*INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY*)

G - Glycine (Gly)|P - Proline (Pro)|A - Alanine (Ala)|V - Valine (Val)|L - Leucine (Leu)|I - Isoleucine (Ile)|M - Methionine (Met)|C - Cysteine (Cys)|F - Phenylalanine (Phe)|Y - Tyrosine (Tyr)|W - Tryptophan (Trp)|H - Histidine (His)|K - Lysine (Lys)|R - Arginine (Arg)|Q - Glutamine (Gln)|N - Asparagine (Asn)|E - Glutamic Acid (Glu)|D - Aspartic Acid (Asp)|S - Serine (Ser)|T - Threonine (Thr)





✓ La programmation dynamique dans le cas des séquences protéiques utilisent des matrices dites de **substitution**, elles sont basées soit sur la capacité de substitution entre acides aminés des mêmes groupes **physico-chimiques** (**matrices physico-chimiques**) soit en se basant sur la probabilité de substitution d'un acide aminé au cours du **temps** (**matrices évolutionnistes**).

## ALIGNEMENT DE DEUX SÉQUENCES

## PROGRAMMATION DYNAMIQUE POUR LES SÉQUENCES PROTÉIQUES

## LA MATRICE DE SUBSTITUTION PAM250 (MATRICE ÉVOLUTIONNISTE)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	2	4

- ✓ Score Positif : les résidus sont similaires, les mutations entre eux arrivent plus souvent qu'attendues par hasard ;
- ✓ Score Négatif : les résidus sont dissimilaires, les mutations entre eux arrivent moins souvent qu'attendus par hasard ;
- ✓ La matrice PAM-250 est très largement utilisée.

## ALIGNEMENT DE DEUX SÉQUENCES

## PROGRAMMATION DYNAMIQUE POUR LES SÉQUENCES PROTÉIQUES

**L'ALGORITHME DE NEEDLEMAN ET WUNSCH POUR LE CAS DES PROTÉINES**

- ✓ Une matrice est d'abord constituée avec une séquence disposée verticalement et l'autre horizontalement.
- ✓ On commence par attribuer à chaque cellule de la matrice  $(i,j)$  la valeur correspondant au maximum du score dans la ligne  $(i+1)$  et la colonne  $(j+1)$  à laquelle on additionne la valeur d'échange des acides aminés appariés en  $(i,j)$  ;
- ✓ À la fin de ce processus, chaque cellule  $(i,j)$  contient donc le score maximal pour toutes les sous-séquences jusqu'au point  $(i,j)$  ;
- ✓ La dernière étape consiste à retracer l'alignement à partir des cellules contenant les scores les plus élevés ;
- ✓ L'équation suivante résume le principe de calcul d'une case de la matrice transformée :

## ALIGNEMENT DE DEUX SÉQUENCES

## PROGRAMMATION DYNAMIQUE POUR LES SÉQUENCES PROTÉIQUES

## L'ALGORITHME DE NEEDLEMAN ET WUNSCH POUR LE CAS DES PROTÉINES

$$S(i,j) = se(i,j) + \max(S(x,y))$$

avec :

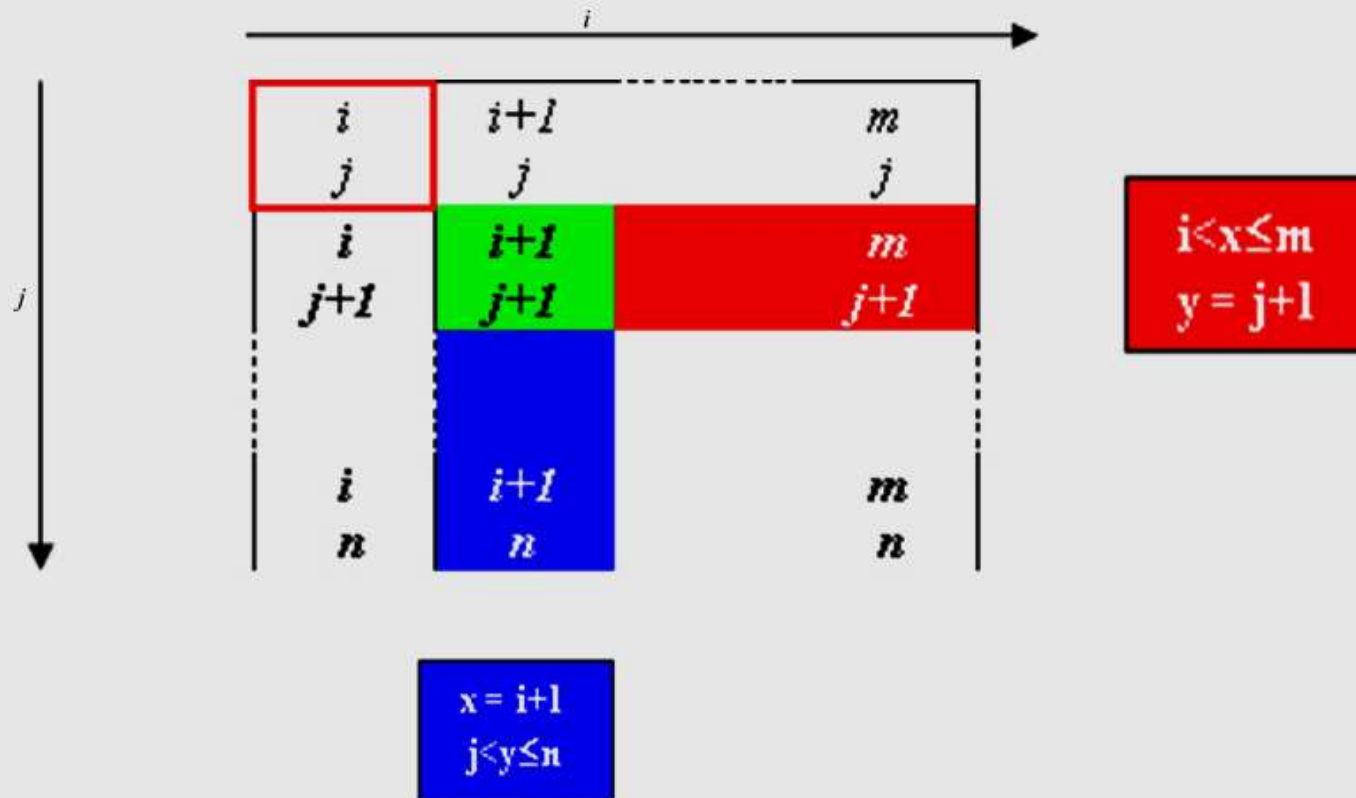
$$i < x \leq m \text{ et } y = j+1$$

ou

$$x = i+1 \text{ et } j < y \leq n$$

$S(i,j)$  est le score somme de la case d'indice  $i$  et  $j$  ;

$se$  le score élémentaire de la case d'indice  $i$  et  $j$  de la matrice initiale (score issu de la matrice de substitution)  $m$  et  $n$  sont les longueurs des deux séquences



## ALIGNEMENT DE DEUX SÉQUENCES

NEEDLEMAN ET WUNCH POUR LES SÉQUENCES PROTÉIQUES

# Exemple d'alignement avec utilisation de la matrice de substitution PAM250 :

On considère les deux séquences suivantes:

Séq 1 = VTEERDAF et Séq 2 = LTSHEAL



## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH POUR LES SÉQUENCES PROTÉIQUES

## ETAPE 1: CALCULE DE LA MATRICE INITIALE À PARTIR DE PAM250

(matrice de substitution)

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-2
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH POUR LES SÉQUENCES PROTÉIQUES

## ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

on commence par noter les valeurs de la dernière colonne et de la dernière ligne:

	V	T	E	E	R	D	A	F
L								2
T								-2
S								-3
H								-2
E								-5
A								-4
L	2	-2	-3	-3	-3	-4	-2	2

## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH POUR LES SÉQUENCES PROTÉIQUES

## ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

On applique l'algorithme de Needleman et Wunch:

	V	T	E	E	R	D	A	F
L								2
T								-2
S								-3
H								-2
E								-5
A							4	-4
L	2	-2	-3	-3	-3	-4	-2	2

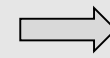
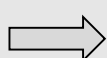


	V	T	E	E	R	D	A	F
L								2
T								-2
S								-3
H								-2
E								-5
A						2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

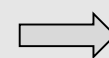
	V	T	E	E	R	D	A	F
L				6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9 <sub>(max)</sub>	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2



....



....



## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH POUR LES SÉQUENCES PROTÉIQUES

## ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

On applique l'algorithme de Needleman et Wunch:

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2



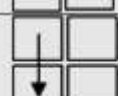
## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH POUR LES SÉQUENCES PROTÉIQUES

## ETAPE 3: PARCOURS DE LA MATRICE TRANSFORMÉE

- ✓ Le parcours s'effectue du plus haut score vers le plus petit. Si les trois cases ont des valeurs de scores égales, alors le chemin vers la diagonale est favorisé :

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

	Substitution
	insertion dans i déletion dans j
	insertion dans j déletion dans i

## ALIGNEMENT DE DEUX SÉQUENCES

## NEEDLEMAN ET WUNCH POUR LES SÉQUENCES PROTÉIQUES

## ETAPE 4: ALIGNEMENT DES DEUX SÉQUENCES ET CALCUL DE SCORE

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

		Substitution
		insertion dans i
		délétion dans j
		insertion dans j
		délétion dans i

Séq1	V	T	—	E	E	R	D	A	F
Séq2	L	T	S	H	E	—	—	A	L

✓ Le score global de cet alignement est  $S = 69$ .

✓ Il y a trois identités : T-T, E-E et A-A et trois similarités (substitutions): V-L, E-H et F-L

On peut supposer que la valine a été substituée en leucine dans la 2<sup>ème</sup> séquence (ou *vis versa*) par besoin d'adaptation de l'organisme à partir du quel a été isolée cette séquence. Le même raisonnement concernera les substitutions E-H et F-L.



## ALIGNEMENT MULTIPLE

## PRINCIPE

- ✓ Dans l'alignement multiple, il est question de comparer plusieurs séquences à la fois.
- ✓ l'utilisation des algorithmes de la programmation dynamique pour un alignement multiple n'est pas recommandé en raison de la quantité d'information à analyser ;
- ✓ C'est pourquoi les alignements multiples seront le plus souvent effectués au moyen de méthodes **heuristiques** qui produiront une **approximation** de l'alignement optimal ;
- ✓ Une heuristique est une méthode de calcul qui fournit rapidement une solution réalisable, pas nécessairement optimale, pour un problème d'optimisation difficile ;
- ✓ Les heuristiques les plus utilisées sont dites **progressives**, elles débutent par l'alignement des deux séquences les plus proches, ensuite les séquences de plus en plus distantes sont ajoutées au fur et à mesure.

## ALIGNEMENT MULTIPLE

## ALGORITHME PROGRESSIF

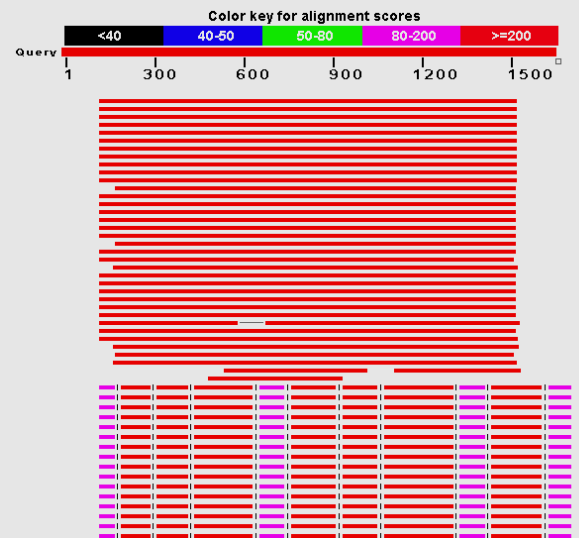
- ✓ Parmi les méthodes employant ces heuristiques progressives, citons: **CLUSTAL W**, **Dialign**, **DCA**, **MSA**, **PIMA**, **MULTALIGN**, **PILEUP**, **Coffee**, **HMMT**, **T-Coffee** , **POA** , **ProbCons** , **Multi-LAGAN**, **Muscle**, **MAFFT**, **SAGA**, etc.
- ✓ Il n'existe pas de méthode universelle.

## ALIGNEMENT MULTIPLE

## ALIGNEMENT D'UNE SÉQUENCE AVEC UNE BANQUE

# ALIGNEMENT D'UNE SÉQUENCE AVEC UNE BANQUE: EXEMPLE DE L'OUTIL BLAST (*BASIC LOCAL ALIGNMENT SEARCH TOOL*) (Altschul *et al.*, 1990)

- ✓ Outil utilisant une heuristique d'alignement local ;
- ✓ recherche dans une base de données de séquence des segments qui sont localement similaires à une séquence-requête fournie par l'utilisateur (*query sequence*) ;
- ✓ Le programme délivre des résultats similaires à la séquence requête, ces résultats sont accompagnés d'un score et d'un pourcentage de similarité;



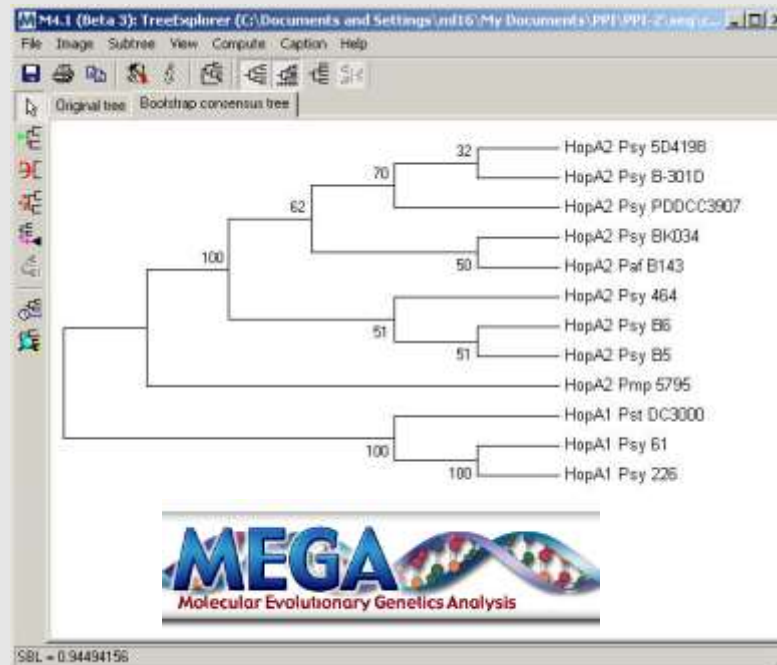
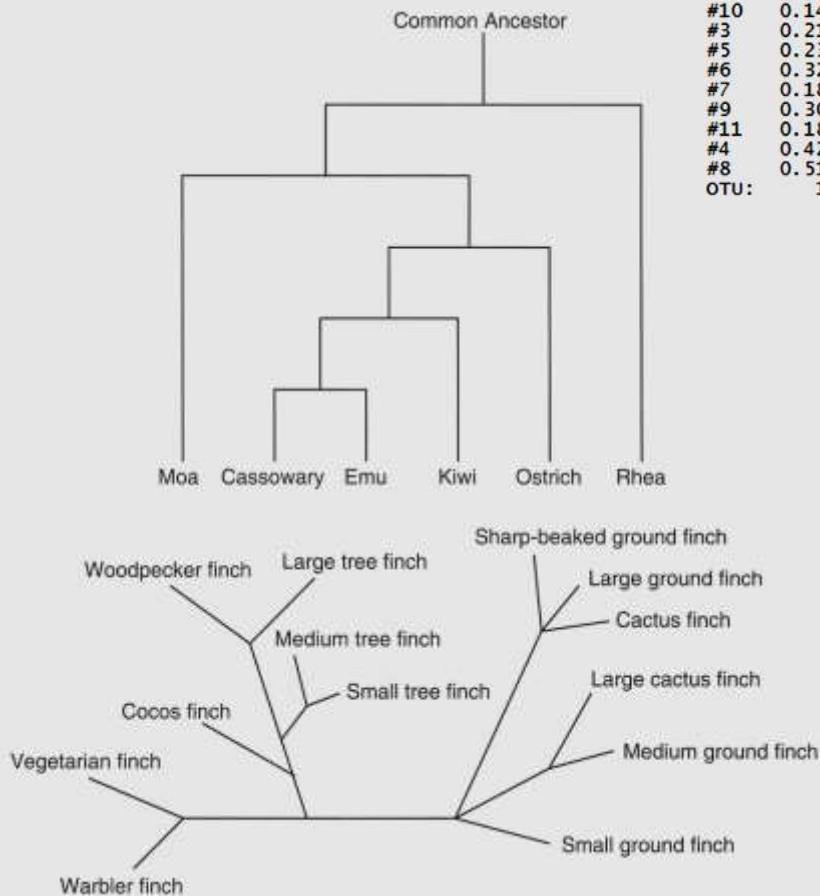
# PHYLOGÉNIE

0.000 0.547 11 -LOG2(S(SM)) before clustering

#1	0.000											
#2	0.187	0.000										
#3	0.216	0.148	0.000									
#4	0.428	0.298	0.389	0.000								
#5	0.232	0.163	0.041	0.447	0.000							
#6	0.327	0.252	0.148	0.346	0.120	0.000						
#7	0.187	0.252	0.312	0.298	0.298	0.206	0.000					
#8	0.517	0.382	0.292	0.546	0.259	0.259	0.489	0.000				
#9	0.303	0.322	0.337	0.322	0.371	0.371	0.184	0.462	0.000			
#10	0.143	0.120	0.263	0.346	0.206	0.346	0.252	0.489	0.322	0.000		
#11	0.187	0.206	0.263	0.396	0.298	0.396	0.206	0.489	0.141	0.206	0.000	
OTU:	1	2	3	4	5	6	7	8	9	10	11	

0.000 0.547 11 -LOG2(S(SM)) UNWEIGHTED AVERAGE LINKAGE

#1	0.000											
#2	0.187	0.000										
#10	0.143	0.120	0.000									
#3	0.216	0.148	0.263	0.000								
#5	0.232	0.163	0.206	0.041	0.000							
#6	0.327	0.252	0.346	0.148	0.120	0.000						
#7	0.187	0.252	0.252	0.312	0.298	0.206	0.000					
#9	0.303	0.322	0.322	0.337	0.371	0.371	0.184	0.000				
#11	0.187	0.206	0.206	0.263	0.298	0.396	0.206	0.141	0.000			
#4	0.428	0.298	0.346	0.389	0.447	0.346	0.298	0.322	0.396	0.000		
#8	0.517	0.382	0.489	0.292	0.259	0.259	0.489	0.462	0.489	0.546	0.000	
OTU:	1	2	10	3	5	6	7	9	11	4	8	



## DÉFINITION

**Phylogénie:** Du grec ancien *phýlon* (« tribu, race ») et *géneia* « qui engendre »

- La **phylogénie** est l'étude des relations de parentés entre différents êtres vivants en vue de comprendre l'évolution de ces organismes.
- La phylogénie trouve ses applications dans plusieurs domaines : systématique, génétique des populations, écologie, épidémiologie, phylogéographie, etc.
- Les relations mises en évidence par la phylogénie sont représentés **graphiquement** sous la forme d'**arbres phylogénétiques**.

**LA STRUCTURE D'UN ARBRE PHYLOGÉNÉTIQUE**

racine

ancêtre commun à tous les  
objets de l'arbre.

G

branche

dont la longueur est proportionnelle  
aux **distances évolutives** (nombre de  
mutations ou temps d'évolution).

noeud

symbolise des ancêtres  
**hypothétiques** partagés par  
les UTO

temps

pere

l'ensemble des  
branches définit la  
**Topologie** de l'arbre  
(sa forme).

fils

A

fils

B

C

D

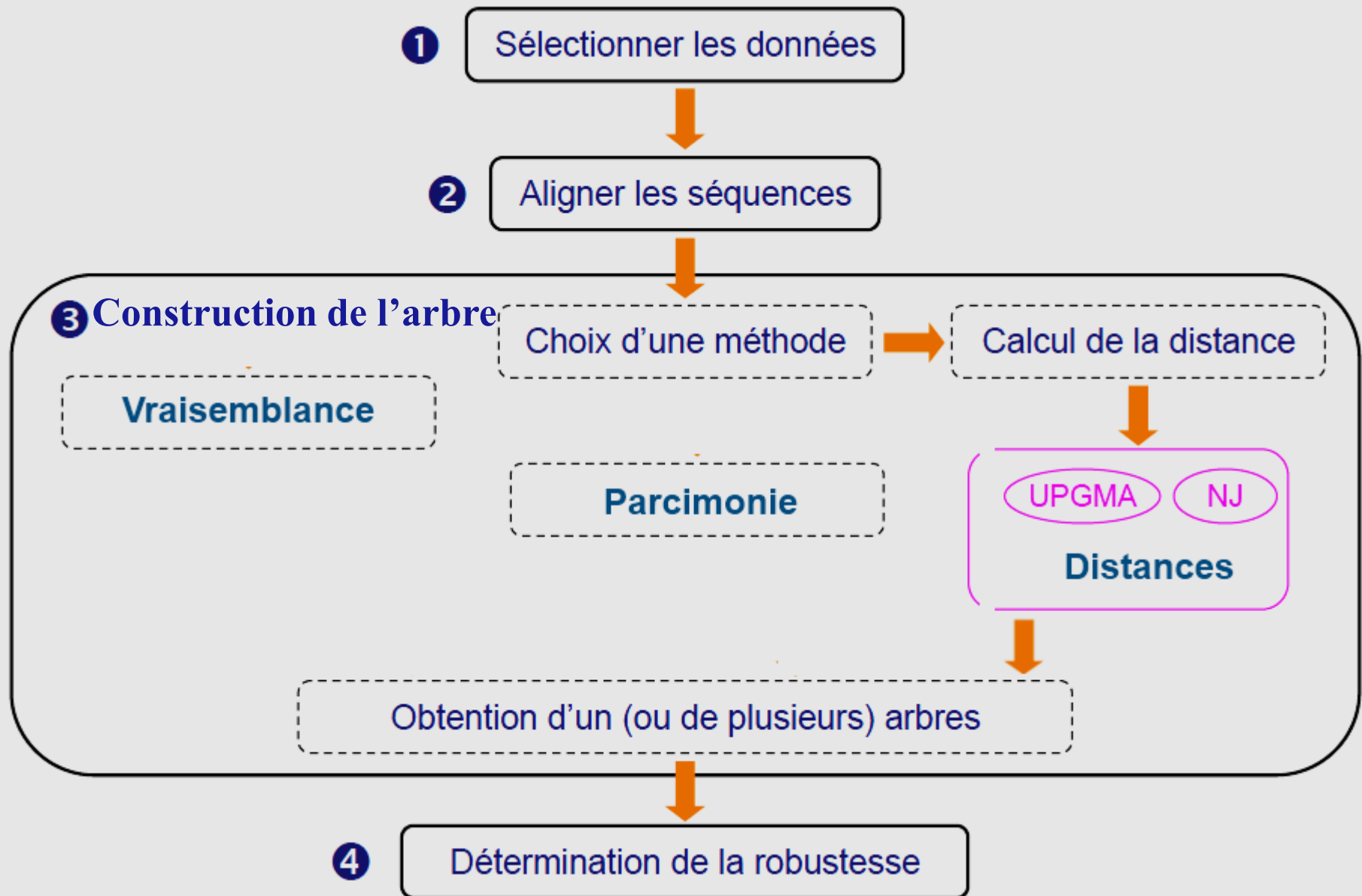
feuille

**UTO (Unité Taxonomique  
Opérationnelle)** ou **Taxon**:  
correspondent aux organismes ou  
séquences étudiés

freres

Le temps s'écoule de la racine vers les feuilles



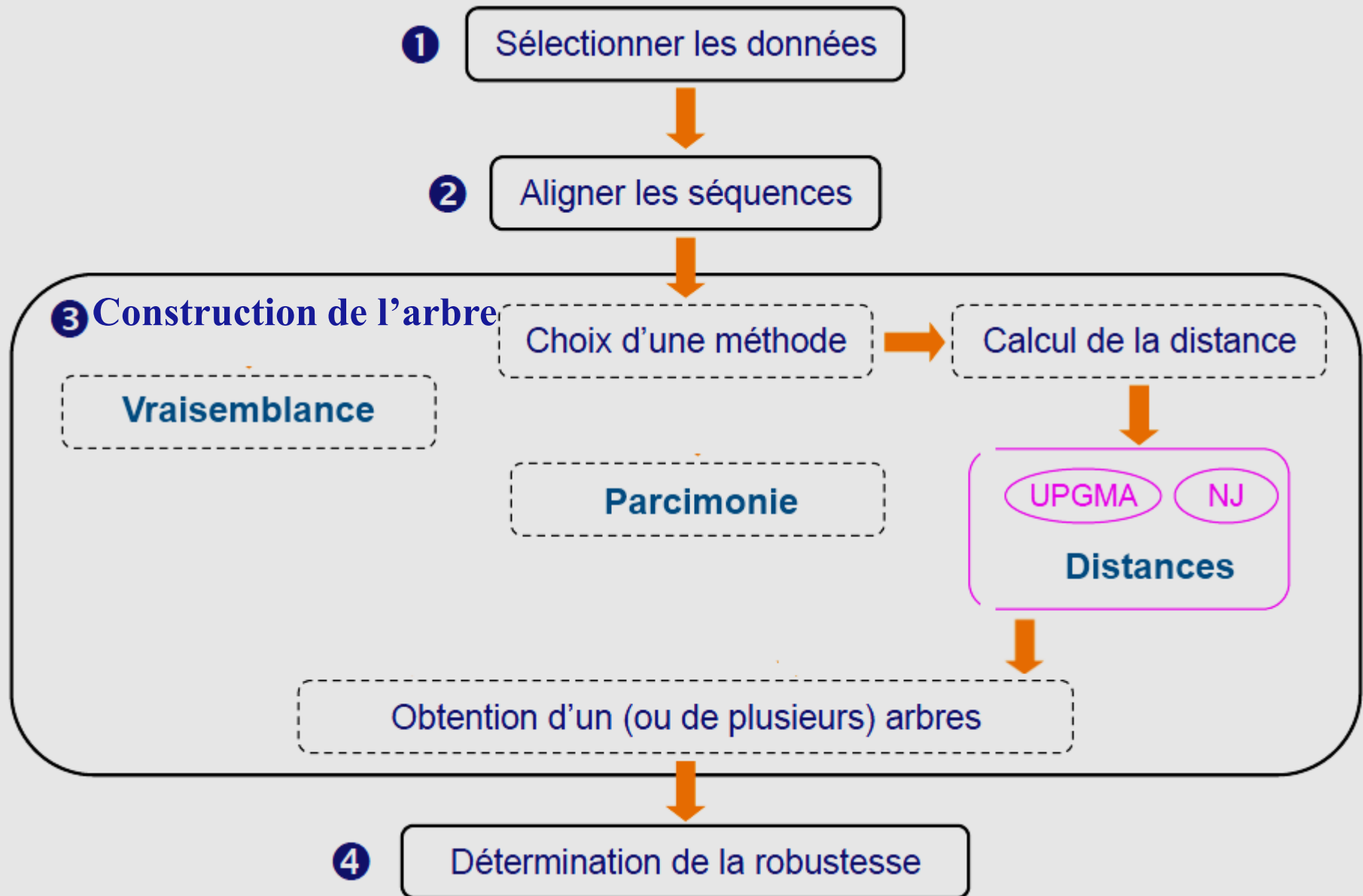
**ÉTAPES DE CONSTRUCTION D'UN ARBRE PHYLOGÉNÉTIQUE**

## NEIGHBOR-JOINING (Saitou et Nei, 1987):

- ✓ Méthode rapide, la plus utilisée;
- ✓ Le principe de NJ consiste en le calcul des longueurs des branches de l'arbre de sorte qu'elles soient la plus petites possibles
- ✓ Les distances arborées générées par cette méthode sont dites **additives**: les longueurs des chemins allant de la racine à n'importe quelle feuille ne sont pas égales, ce qui stipule que les caractères des UTO évoluent indépendamment les uns des autres (taux de mutations variables) ;

## ALGORITHME

- ✓ NJ utilise un algorithme de clustérisation séquentielle.

**ÉTAPES DE CONSTRUCTION D'UN ARBRE PHYLOGÉNÉTIQUE**

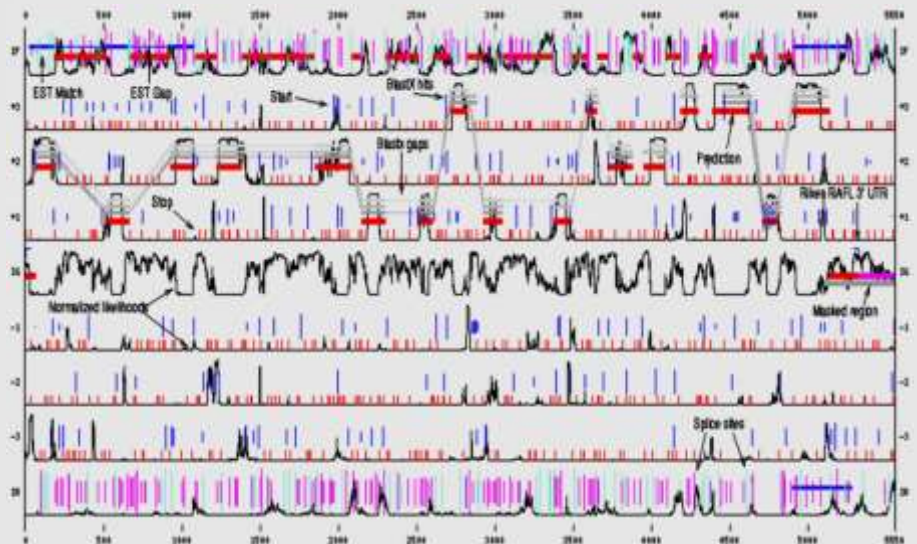
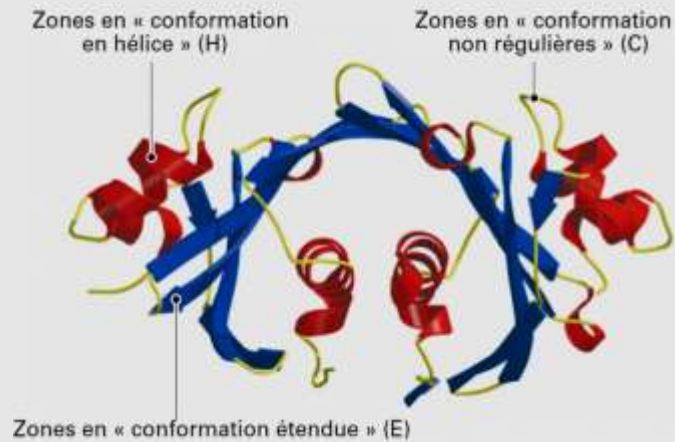
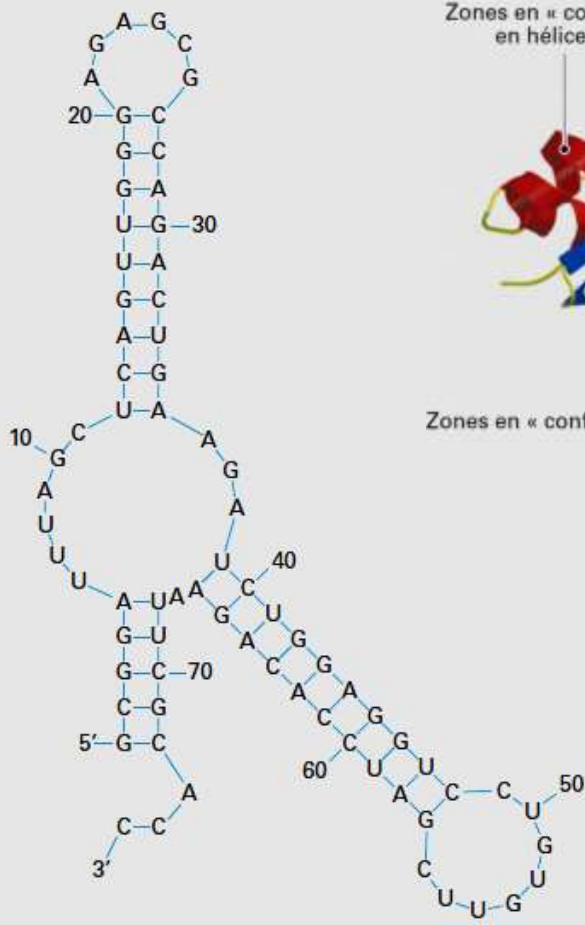
## FIABILITÉ D'UN ARBRE

- ✓ La construction des arbres phylogénétiques est basée sur des hypothèses. Cela induit des erreurs au niveau topologique et conduit à des erreurs d'interprétation qui nécessitent d'être rectifiées.
- ✓ La méthode du **bootstrap** est la plus utilisée pour vérifier la **fiabilité** (**robustesse**) de l'arbre obtenu avec telle ou telle méthode de construction.

# ANNOTATION DES SÉQUENCES

>ARN

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGA  
GGUCCUGUGUUCGAUCCACAGAAUUCGCACCA



**L'annotation est la recherche d'informations  
(position, structure, fonction) sur une séquence  
(nucléique, protéique) inconnue**



## ANNOTATION

## DÉFINITION

✓ **Annotation structurale (Where)**

Permet de positionner les gènes et leurs produits :

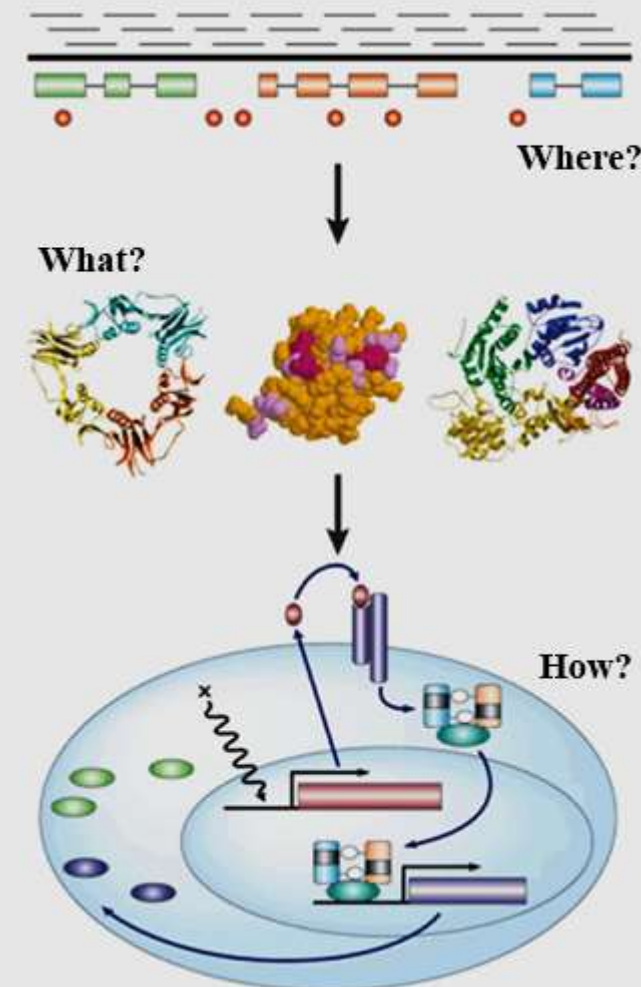
- ☐ éléments répétés (éléments transposables, satellites, etc.)
- ☐ gènes des ARN stables (ARNr, ARNt...)
- ☐ gènes protéiques
- ☐ régions régulatrices, etc.

✓ **Annotation fonctionnelle (What)**

Permet de prédire les fonctions et produits potentiels des gènes préalablement identifiés, leurs structures 2D et 3D (ARN et protéines) ainsi que leurs positions cellulaires (pour les protéines).

✓ **Annotation relationnelle (How)**

Permet de déterminer les interactions que les objets biologiques préalablement identifiés sont susceptibles d'entretenir (familles de gènes, réseaux de régulation, réseaux métaboliques, etc.).



**L'annotation est basée essentiellement sur la reconnaissance des différentes structures d'un gène (motifs, *ORF* “*Open Reading Frame*, etc.).**

## 1. MÉTHODE EXPÉRIMENTALE

Par alignement de la séquence de l'ADN génomique ( $ADN_g$ ) avec la séquence de l'ADN<sub>c</sub> complet isolé du même organisme ;

## 2. MÉTHODES INTRINSÈQUES (*Ab initio*)

Par l'emploi de modèles statistiques capables de localiser des séquences codantes et non codantes à partir d'exemples (séquences de gènes modèles) ;

## 3. MÉTHODES EXTRINSÈQUES (COMPARATIVES)

- ✓ Par la recherche des ORF's (*Open Reading Phase*) et des motifs;
- ✓ Par comparaison aux banques de données biologiques (avec les outils BLAST, FASTA) à la recherche des séquences d'ARNm et de protéines qui ressemblent à la séquence étudiée ;

## 4. MÉTHODES INTÉGRATIVES

Par la combinaison des trois dernières approches.

## ANNOTATION

## OUTILS

Exemple de logiciels d'annotation:

- ✓ Pour les séquences nucléiques
  - GeneFinder
  - Artemis
  - Geneious
- ✓ Pour les séquences protéiques
  - InterproScan

Exemple de banque de séquences annotées:

- ✓ Pour les séquences nucléiques
  - Ensembl
  - Genatlas
- ✓ Pour les séquences protéiques
  - Prosite
  - Pfam
  - Interpro

- ✓ Afin de résoudre ces problèmes, différentes méthodes algorithmiques et statistiques ont été développées, chacune ayant abouti au développement de logiciels de prédiction plus ou moins efficaces:

☐ **Prédiction des structures secondaires:**

- **PSIPRED, SSPPRO, PROF\_King**

☐ **Recherche d'informations:**

- **Ponts disulfure: DIANA, GDAP ;**
- **Homologie: 3DJury, HHPred**
- **Fonction: PFP, ProtFun**

☐ **Prédiction de la structure tertiaire (3D)**

- **SWISS-MODEL, GenTHREADER, HMMStr/Rosetta**

☐ **Evaluation des modèles: Eval123D, Procheck, Verify3D**