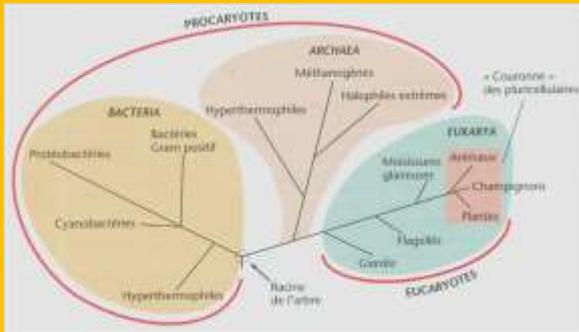
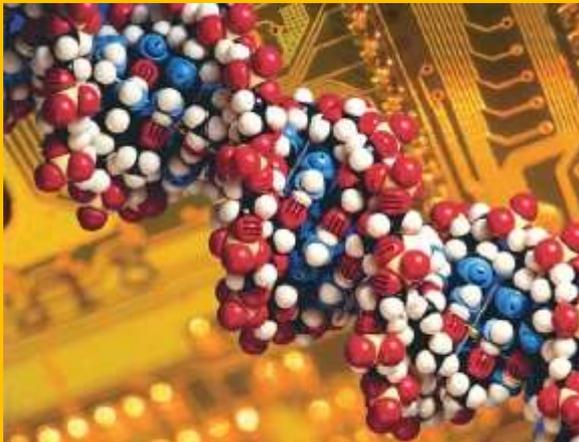


Université Frères Mentouri Constantine
Institut de la Nutrition, de l'Alimentation et des Technologies Agro-alimentaires (INATAA)
Master 1-Technologies Alimentaires
Année universitaire 201-2020



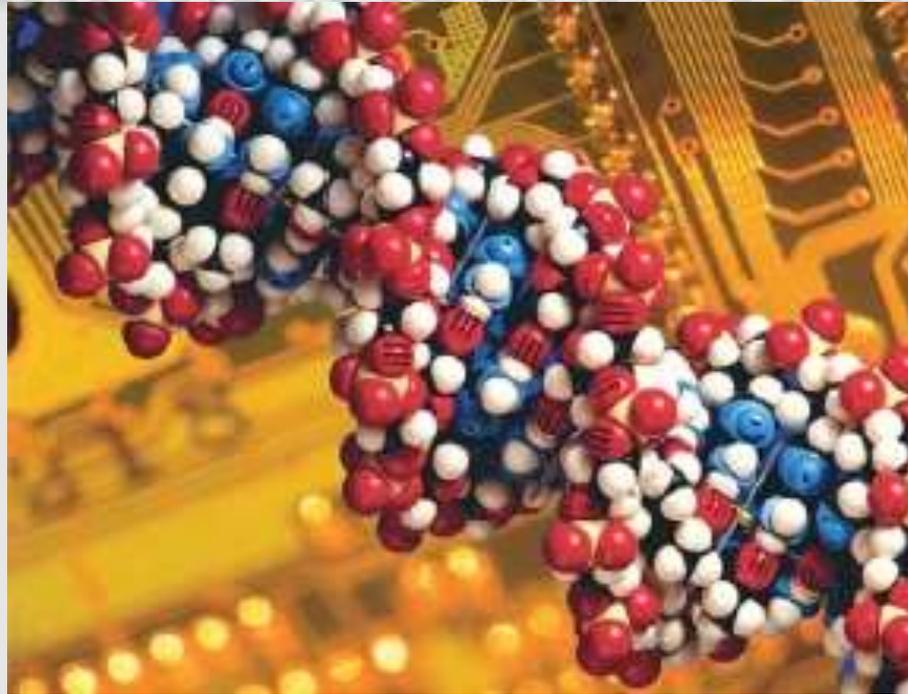
INITIATION À LA BIOINFORMATIQUE



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

Search: 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/Swiss-Prot](#)
- [UniProtKB/TrEMBL](#)

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 accompanying [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 Cochrane](#) as part of the [The Protein and Nucleotide Database Group \(PANDA\)](#) under [Ewan Birney](#).

Link	Explanation
Access	Database queries, Collocated services, webserver, FTP archives (EMBL release, alignments etc), EMBL resource version archive (SVA), Browse by geography
Submission	Primary sequence submissions, third party annotation, updates.
Documentation	Release notes , user manual , Information for Submitters , FAQ , Release information , Footprints Changes , EMBL database statistics , Feature lists , EMBL documentation , Search entry , Accession Number Prefix Codes , Examples of annotation , EMBL Features & Qualifiers , ORF line standards , Database Policies
Publications	Group publications
People	Group members
Contact	How to contact the EMBL Nucleotide Sequence Database
News	List of recent changes on this site

DDBJ - DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of the three nucleotide sequence databases that constitute the International Nucleotide Sequence Database Collaboration (INSDC), 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 DDBJ](#)
- Apr. 12, 2010 [Release of the raw and assembled sequence data set from ngs](#)
- Apr. 12, 2010 [DAD \(DDBJ amino acid database\) Ref. 51.0 Released](#)

Maintenance

- Apr. 21, 2010 [Suspension of some DDBJ activities in Japanese holidays\(A/29.5/1-5\)](#)
- Mar. 16, 2010 [\(Apr. 23\)ARSA database search \(DDBJ, DAD\) temporary unavailable](#)
- Feb. 03, 2010 [\(Important!\) Termination of a part of DDBJ 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**).

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/>

NCBI Resources How To Sign in to NCBI

GenBank Nucleotide Search

GenBank Submit Genomes WGS HTGs EST/GSS Metagenomes TPA TSA INSDC

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp](#) site. The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#), which is divided into three divisions: [CoreNucleotide](#) (the main collection), [dbEST](#) (Expressed Sequence Tags), and [dbGSS](#) (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using [NCBI e-utils](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

GenBank Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)

*En Décembre 2017, Genbank archivait plus de 206 millions 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>

EMBL-EBI

Services Research Training About us

ENA
European Nucleotide Archive

Search

Examples: [BN000065](#), [histone](#)

Advanced
Sequence

Home Search & Browse Submit & Update Software About ENA Support

European Nucleotide Archive

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](#)

Access to ENA data is provided through the browser, through search tools, large scale file download and through the API.

Text Search

Search

Examples: [BN000065](#), [histone](#)

Advanced search

Sequence Search

Enter or paste a nucleotide sequence or accession number

Search

Advanced search

Popular

- Submit and update
- Sequence submissions
- Genome assembly submissions
- Submitting environmental sequences
- Citing ENA data
- Rest URLs for data retrieval
- Rest URLs to search ENA

Latest ENA news

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

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 homepage. At the top, there is a navigation bar with links for 'About DDBJ', 'How to Use', 'Report/Statistics', 'FAQ', and 'Contact Us'. A search bar is located in the top right corner. The main content area is divided into several sections. On the left, there are social media links for RSS, DDBJ Twitter, and Mail Magazine. Below these are logos for INSDC, NCBI, and EMBL-EBI. The main content area features a 'DDBJ Service' section with four icons: 'Data Submission', 'Search / Analysis', 'Super Computer', and 'ftp.ddbj.nig.ac.jp'. Below this is a 'Hot Topics' section with a 'News Archive' filter and a list of recent news items. The left sidebar contains logos for NIG, JBIportal, NBDC, DBCLS, and PDBj.

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

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 shows the UniProt website homepage. At the top, there is a navigation bar with the UniProt logo and a search bar. Below the navigation bar, there is a main content area with several sections:

- UniProtKB:** UniProt Knowledgebase, Swiss-Prot (553,474) - Manually annotated and reviewed. Records with information extracted from literature and curator-evaluated computational analysis.
- TrEMBL (73,711,881):** Automatically annotated and not reviewed. Records that await full manual annotation.
- 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, Keywords.
- Getting started:** Text search, BLAST, Sequence alignments, Retrieve/ID mapping, Peptide search.
- UniProt data:** Download latest release, Statistics, How to cite us, Submit your data, SPARQL.
- Protein spotlight:** Out Of The Ordinary (January 2017) - Life depends on chemical signals. Without them our heart wouldn't know how to beat or our thoughts how to form, our eyes would be unable to see, our legs unable to walk and our mouths would be incapable of speech...
- News:** Forthcoming changes, Planned changes for UniProt, UniProt release 2017_01, UniProt release 2016_11, UniProt release 2016_10, News archive.

*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

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

PDB Data Growth & Usage Statistics

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*

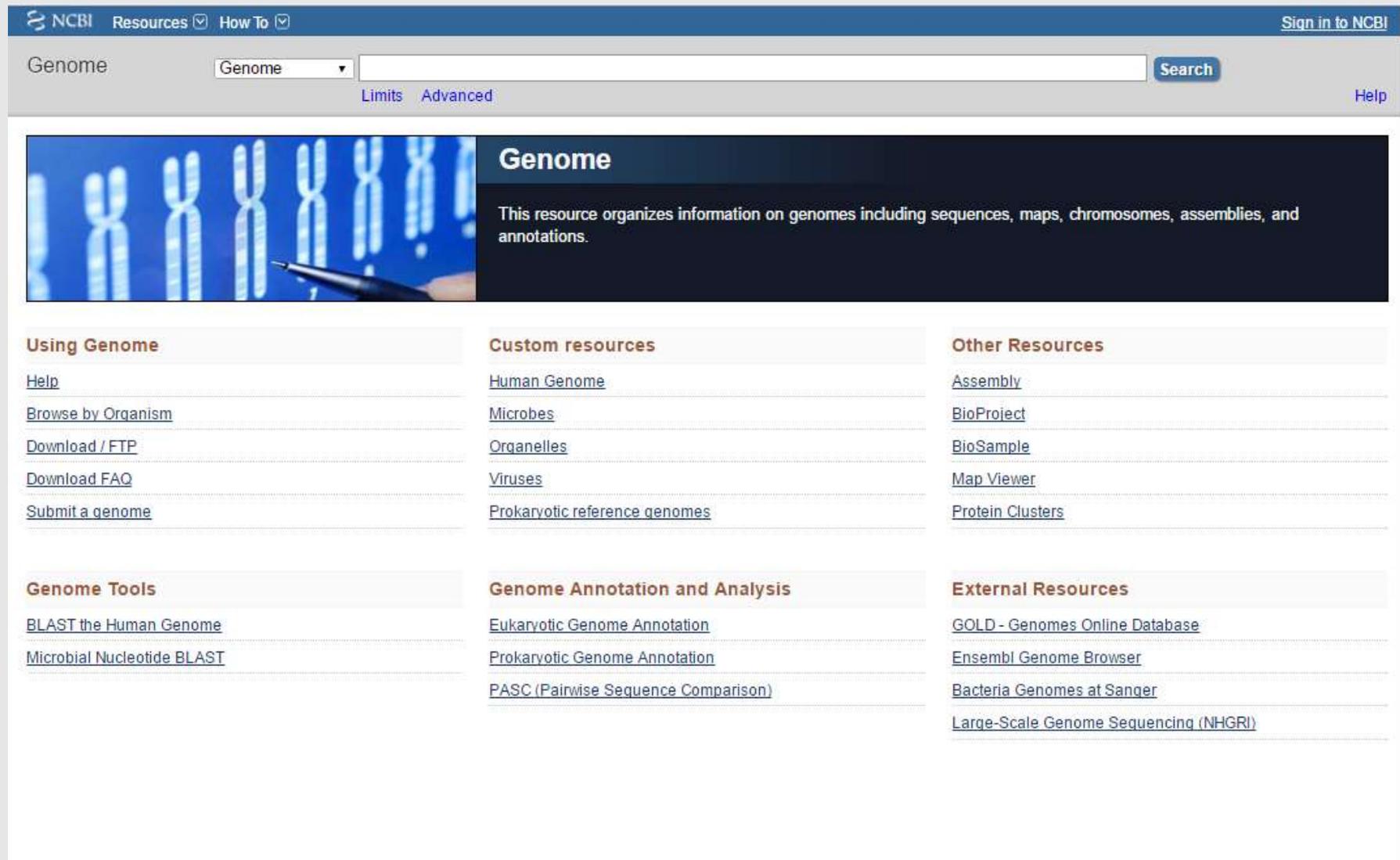
The paper (doi: [10.1016/j.str.2017.11.001](#))

*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.). 4

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 Search Limits Advanced 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 Sanqer](#)
- [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.

neXtprot :

Banque de protéines humaines

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



Tools ▾ Portals ▾ Download Help ▾ About ▾ Contact

Login



neXtprot

Exploring the universe of human proteins

Simple search Advanced search

proteins ▾

Gold only ▾



Q Search

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



Release contents

📰 News

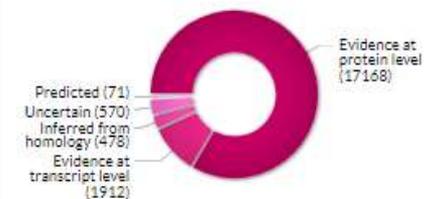


- » neXtProt in ExPASy, tweaking the pep...
Jan 08, 2018
- » PEFF 1.0 format implemented
Oct 25, 2017
- » New release focusing on expression
Sep 06, 2017

News archive

📅 Release 2017-08-01

Protein existence in neXtProt



Release statistics

SWISS-2DPAGE :

Base de données de protéines identifiées par électrophorèse bidimensionnelle

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

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

 **ExPASy**
Bioinformatics Resource Portal

Home | [Contact](#)

 **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

-  [\[How to use this interface\]](#)
- [by description](#) (any word in the ID, DE, GN and KW lines)
- [by accession number](#) (AC lines)
- [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.
- [by author](#) (RA lines)
- [by spot serial number](#) (2D and 1D lines)
- [by experimental pI/Mw range](#)
- [by experimental identification methods](#)
- [by full text search](#)
- [retrieve all the protein entries identified on a given reference map](#)
- [user defined / complex queries](#) (SRS like)

SWISS-2DPAGE documents

-  [\[Facts and statistics\]](#)
- [User manual](#)
- [Release notes](#) (September 26, 2006)
- [FAQ](#) (Frequently Asked Questions about SWISS-2DPAGE)
- **Protocols:**
 - [Technical information](#) about 2-D PAGE (IPG's, silver staining, protocols, etc)
- **Figure captions of SWISS-2DPAGE maps available from publications:**
 - Human [CSF](#), [ELC](#), [HEPG2](#), [HEPG2SP](#), [LIVER](#), [LYMPHOMA](#), [PLASMA](#), [PLATELET](#), [RBC](#), [U937](#), [CEC](#), [KIDNEY](#).
 - *Dictyostelium discoideum*, *Escherichia coli*, *Saccharomyces cerevisiae*.

Services

Software

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

MGI Mouse Genome Informatics

Search Download More Resources Submit Data Find Mice (IMSR) Analysis Tools Contact Us Browsers

Keywords, Symbols, or IDs **QuickSearch**

Or use topic specific search and analysis tools:

- 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
- Nomenclature

Getting Started:

- Introduction to mouse genetics
- How to use MGI (Text & Video)
- Cre Portal Tutorial

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.

[About Us](#) [MGI Publications](#)

ALLIANCE of GENOME RESOURCES **AGR 1.0 release:** Search for gene, disease & QO data from 6 Model Organism Databases & the GO Consortium

Parkinson's disease (DOID:143303)

Definition: A genetic disorder that has, in part, a degeneration of the central nervous system that often impairs motor skills, speech, and other functions. [HUGO Nomenclature \(nomenclature\)](#)

Synonyms: parkinson aphasia; Parkinson disease

Case References: EP13092309.07, AC11004828, KCPDM.032, KCONCH.032A, N000001212, HSH2000000.07, AC11004817, OAHM.00000007, OAHM.00000017, OAHM.00000027, OAHM.00000037, OAHM.00000047, OAHM.00000057, OAHM.00000067, OAHM.00000077, OAHM.00000087, OAHM.00000097, OAHM.00000107, OAHM.00000117, OAHM.00000127, OAHM.00000137, OAHM.00000147, OAHM.00000157, OAHM.00000167, OAHM.00000177, OAHM.00000187, OAHM.00000197, OAHM.00000207, OAHM.00000217, OAHM.00000227, OAHM.00000237, OAHM.00000247, OAHM.00000257, OAHM.00000267, OAHM.00000277, OAHM.00000287, OAHM.00000297, OAHM.00000307, OAHM.00000317, OAHM.00000327, OAHM.00000337, OAHM.00000347, OAHM.00000357, OAHM.00000367, OAHM.00000377, OAHM.00000387, OAHM.00000397, OAHM.00000407, OAHM.00000417, OAHM.00000427, OAHM.00000437, OAHM.00000447, OAHM.00000457, OAHM.00000467, OAHM.00000477, OAHM.00000487, OAHM.00000497, OAHM.00000507, OAHM.00000517, OAHM.00000527, OAHM.00000537, OAHM.00000547, OAHM.00000557, OAHM.00000567, OAHM.00000577, OAHM.00000587, OAHM.00000597, OAHM.00000607, OAHM.00000617, OAHM.00000627, OAHM.00000637, OAHM.00000647, OAHM.00000657, OAHM.00000667, OAHM.00000677, OAHM.00000687, OAHM.00000697, OAHM.00000707, OAHM.00000717, OAHM.00000727, OAHM.00000737, OAHM.00000747, OAHM.00000757, OAHM.00000767, OAHM.00000777, OAHM.00000787, OAHM.00000797, OAHM.00000807, OAHM.00000817, OAHM.00000827, OAHM.00000837, OAHM.00000847, OAHM.00000857, OAHM.00000867, OAHM.00000877, OAHM.00000887, OAHM.00000897, OAHM.00000907, OAHM.00000917, OAHM.00000927, OAHM.00000937, OAHM.00000947, OAHM.00000957, OAHM.00000967, OAHM.00000977, OAHM.00000987, OAHM.00000997, OAHM.00001007, OAHM.00001017, OAHM.00001027, OAHM.00001037, OAHM.00001047, OAHM.00001057, OAHM.00001067, OAHM.00001077, OAHM.00001087, OAHM.00001097, OAHM.00001107, OAHM.00001117, OAHM.00001127, OAHM.00001137, OAHM.00001147, OAHM.00001157, OAHM.00001167, OAHM.00001177, OAHM.00001187, OAHM.00001197, OAHM.00001207, OAHM.00001217, OAHM.00001227, OAHM.00001237, OAHM.00001247, OAHM.00001257, OAHM.00001267, OAHM.00001277, OAHM.00001287, OAHM.00001297, OAHM.00001307, OAHM.00001317, OAHM.00001327, OAHM.00001337, OAHM.00001347, OAHM.00001357, OAHM.00001367, OAHM.00001377, OAHM.00001387, OAHM.00001397, OAHM.00001407, OAHM.00001417, OAHM.00001427, OAHM.00001437, OAHM.00001447, OAHM.00001457, OAHM.00001467, OAHM.00001477, OAHM.00001487, OAHM.00001497, OAHM.00001507, OAHM.00001517, OAHM.00001527, OAHM.00001537, OAHM.00001547, OAHM.00001557, OAHM.00001567, OAHM.00001577, OAHM.00001587, OAHM.00001597, OAHM.00001607, OAHM.00001617, OAHM.00001627, OAHM.00001637, OAHM.00001647, OAHM.00001657, OAHM.00001667, OAHM.00001677, OAHM.00001687, OAHM.00001697, OAHM.00001707, OAHM.00001717, OAHM.00001727, OAHM.00001737, OAHM.00001747, OAHM.00001757, OAHM.00001767, OAHM.00001777, OAHM.00001787, OAHM.00001797, OAHM.00001807, OAHM.00001817, OAHM.00001827, OAHM.00001837, OAHM.00001847, OAHM.00001857, OAHM.00001867, OAHM.00001877, OAHM.00001887, OAHM.00001897, OAHM.00001907, OAHM.00001917, OAHM.00001927, OAHM.00001937, OAHM.00001947, OAHM.00001957, OAHM.00001967, OAHM.00001977, OAHM.00001987, OAHM.00001997, OAHM.00002007, OAHM.00002017, OAHM.00002027, OAHM.00002037, OAHM.00002047, OAHM.00002057, OAHM.00002067, OAHM.00002077, OAHM.00002087, OAHM.00002097, OAHM.00002107, OAHM.00002117, OAHM.00002127, OAHM.00002137, OAHM.00002147, OAHM.00002157, OAHM.00002167, OAHM.00002177, OAHM.00002187, OAHM.00002197, OAHM.00002207, OAHM.00002217, OAHM.00002227, OAHM.00002237, OAHM.00002247, OAHM.00002257, OAHM.00002267, OAHM.00002277, OAHM.00002287, OAHM.00002297, OAHM.00002307, OAHM.00002317, OAHM.00002327, OAHM.00002337, OAHM.00002347, OAHM.00002357, OAHM.00002367, OAHM.00002377, OAHM.00002387, OAHM.00002397, OAHM.00002407, OAHM.00002417, OAHM.00002427, OAHM.00002437, OAHM.00002447, OAHM.00002457, OAHM.00002467, OAHM.00002477, OAHM.00002487, OAHM.00002497, OAHM.00002507, OAHM.00002517, OAHM.00002527, OAHM.00002537, OAHM.00002547, OAHM.00002557, OAHM.00002567, OAHM.00002577, OAHM.00002587, OAHM.00002597, OAHM.00002607, OAHM.00002617, OAHM.00002627, OAHM.00002637, OAHM.00002647, OAHM.00002657, OAHM.00002667, OAHM.00002677, OAHM.00002687, OAHM.00002697, OAHM.00002707, OAHM.00002717, OAHM.00002727, OAHM.00002737, OAHM.00002747, OAHM.00002757, OAHM.00002767, OAHM.00002777, OAHM.00002787, OAHM.00002797, OAHM.00002807, OAHM.00002817, OAHM.00002827, OAHM.00002837, OAHM.00002847, OAHM.00002857, OAHM.00002867, OAHM.00002877, OAHM.00002887, OAHM.00002897, OAHM.00002907, OAHM.00002917, OAHM.00002927, OAHM.00002937, OAHM.00002947, OAHM.00002957, OAHM.00002967, OAHM.00002977, OAHM.00002987, OAHM.00002997, OAHM.00003007, OAHM.00003017, OAHM.00003027, OAHM.00003037, OAHM.00003047, OAHM.00003057, OAHM.00003067, OAHM.00003077, OAHM.00003087, OAHM.00003097, OAHM.00003107, OAHM.00003117, OAHM.00003127, OAHM.00003137, OAHM.00003147, OAHM.00003157, OAHM.00003167, OAHM.00003177, OAHM.00003187, OAHM.00003197, OAHM.00003207, OAHM.00003217, OAHM.00003227, OAHM.00003237, OAHM.00003247, OAHM.00003257, OAHM.00003267, OAHM.00003277, OAHM.00003287, OAHM.00003297, OAHM.00003307, OAHM.00003317, OAHM.00003327, OAHM.00003337, OAHM.00003347, OAHM.00003357, OAHM.00003367, OAHM.00003377, OAHM.00003387, OAHM.00003397, OAHM.00003407, OAHM.00003417, OAHM.00003427, OAHM.00003437, OAHM.00003447, OAHM.00003457, OAHM.00003467, OAHM.00003477, OAHM.00003487, OAHM.00003497, OAHM.00003507, OAHM.00003517, OAHM.00003527, OAHM.00003537, OAHM.00003547, OAHM.00003557, OAHM.00003567, OAHM.00003577, OAHM.00003587, OAHM.00003597, OAHM.00003607, OAHM.00003617, OAHM.00003627, OAHM.00003637, OAHM.00003647, OAHM.00003657, OAHM.00003667, OAHM.00003677, OAHM.00003687, OAHM.00003697, OAHM.00003707, OAHM.00003717, OAHM.00003727, OAHM.00003737, OAHM.00003747, OAHM.00003757, OAHM.00003767, OAHM.00003777, OAHM.00003787, OAHM.00003797, OAHM.00003807, OAHM.00003817, OAHM.00003827, OAHM.00003837, OAHM.00003847, OAHM.00003857, OAHM.00003867, OAHM.00003877, OAHM.00003887, OAHM.00003897, OAHM.00003907, OAHM.00003917, OAHM.00003927, OAHM.00003937, OAHM.00003947, OAHM.00003957, OAHM.00003967, OAHM.00003977, OAHM.00003987, OAHM.00003997, OAHM.00004007, OAHM.00004017, OAHM.00004027, OAHM.00004037, OAHM.00004047, OAHM.00004057, OAHM.00004067, OAHM.00004077, OAHM.00004087, OAHM.00004097, OAHM.00004107, OAHM.00004117, OAHM.00004127, OAHM.00004137, OAHM.00004147, OAHM.00004157, OAHM.00004167, OAHM.00004177, OAHM.00004187, OAHM.00004197, OAHM.00004207, OAHM.00004217, OAHM.00004227, OAHM.00004237, OAHM.00004247, OAHM.00004257, OAHM.00004267, OAHM.00004277, OAHM.00004287, OAHM.00004297, OAHM.00004307, OAHM.00004317, OAHM.00004327, OAHM.00004337, OAHM.00004347, OAHM.00004357, OAHM.00004367, OAHM.00004377, OAHM.00004387, OAHM.00004397, OAHM.00004407, OAHM.00004417, OAHM.00004427, OAHM.00004437, OAHM.00004447, OAHM.00004457, OAHM.00004467, OAHM.00004477, OAHM.00004487, OAHM.00004497, OAHM.00004507, OAHM.00004517, OAHM.00004527, OAHM.00004537, OAHM.00004547, OAHM.00004557, OAHM.00004567, OAHM.00004577, OAHM.00004587, OAHM.00004597, OAHM.00004607, OAHM.00004617, OAHM.00004627, OAHM.00004637, OAHM.00004647, OAHM.00004657, OAHM.00004667, OAHM.00004677, OAHM.00004687, OAHM.00004697, OAHM.00004707, OAHM.00004717, OAHM.00004727, OAHM.00004737, OAHM.00004747, OAHM.00004757, OAHM.00004767, OAHM.00004777, OAHM.00004787, OAHM.00004797, OAHM.00004807, OAHM.00004817, OAHM.00004827, OAHM.00004837, OAHM.00004847, OAHM.00004857, OAHM.00004867, OAHM.00004877, OAHM.00004887, OAHM.00004897, OAHM.00004907, OAHM.00004917, OAHM.00004927, OAHM.00004937, OAHM.00004947, OAHM.00004957, OAHM.00004967, OAHM.00004977, OAHM.00004987, OAHM.00004997, OAHM.00005007, OAHM.00005017, OAHM.00005027, OAHM.00005037, OAHM.00005047, OAHM.00005057, OAHM.00005067, OAHM.00005077, OAHM.00005087, OAHM.00005097, OAHM.00005107, OAHM.00005117, OAHM.00005127, OAHM.00005137, OAHM.00005147, OAHM.00005157, OAHM.00005167, OAHM.00005177, OAHM.00005187, OAHM.00005197, OAHM.00005207, OAHM.00005217, OAHM.00005227, OAHM.00005237, OAHM.00005247, OAHM.00005257, OAHM.00005267, OAHM.00005277, OAHM.00005287, OAHM.00005297, OAHM.00005307, OAHM.00005317, OAHM.00005327, OAHM.00005337, OAHM.00005347, OAHM.00005357, OAHM.00005367, OAHM.00005377, OAHM.00005387, OAHM.00005397, OAHM.00005407, OAHM.00005417, OAHM.00005427, OAHM.00005437, OAHM.00005447, OAHM.00005457, OAHM.00005467, OAHM.00005477, OAHM.00005487, OAHM.00005497, OAHM.00005507, OAHM.00005517, OAHM.00005527, OAHM.00005537, OAHM.00005547, OAHM.00005557, OAHM.00005567, OAHM.00005577, OAHM.00005587, OAHM.00005597, OAHM.00005607, OAHM.00005617, OAHM.00005627, OAHM.00005637, OAHM.00005647, OAHM.00005657, OAHM.00005667, OAHM.00005677, OAHM.00005687, OAHM.00005697, OAHM.00005707, OAHM.00005717, OAHM.00005727, OAHM.00005737, OAHM.00005747, OAHM.00005757, OAHM.00005767, OAHM.00005777, OAHM.00005787, OAHM.00005797, OAHM.00005807, OAHM.00005817, OAHM.00005827, OAHM.00005837, OAHM.00005847, OAHM.00005857, OAHM.00005867, OAHM.00005877, OAHM.00005887, OAHM.00005897, OAHM.00005907, OAHM.00005917, OAHM.00005927, OAHM.00005937, OAHM.00005947, OAHM.00005957, OAHM.00005967, OAHM.00005977, OAHM.00005987, OAHM.00005997, OAHM.00006007, OAHM.00006017, OAHM.00006027, OAHM.00006037, OAHM.00006047, OAHM.00006057, OAHM.00006067, OAHM.00006077, OAHM.00006087, OAHM.00006097, OAHM.00006107, OAHM.00006117, OAHM.00006127, OAHM.00006137, OAHM.00006147, OAHM.00006157, OAHM.00006167, OAHM.00006177, OAHM.00006187, OAHM.00006197, OAHM.00006207, OAHM.00006217, OAHM.00006227, OAHM.00006237, OAHM.00006247, OAHM.00006257, OAHM.00006267, OAHM.00006277, OAHM.00006287, OAHM.00006297, OAHM.00006307, OAHM.00006317, OAHM.00006327, OAHM.00006337, OAHM.00006347, OAHM.00006357, OAHM.00006367, OAHM.00006377, OAHM.00006387, OAHM.00006397, OAHM.00006407, OAHM.00006417, OAHM.00006427, OAHM.00006437, OAHM.00006447, OAHM.00006457, OAHM.00006467, OAHM.00006477, OAHM.00006487, OAHM.00006497, OAHM.00006507, OAHM.00006517, OAHM.00006527, OAHM.00006537, OAHM.00006547, OAHM.00006557, OAHM.00006567, OAHM.00006577, OAHM.00006587, OAHM.00006597, OAHM.00006607, OAHM.00006617, OAHM.00006627, OAHM.00006637, OAHM.00006647, OAHM.00006657, OAHM.00006667, OAHM.00006677, OAHM.00006687, OAHM.00006697, OAHM.00006707, OAHM.00006717, OAHM.00006727, OAHM.00006737, OAHM.00006747, OAHM.00006757, OAHM.00006767, OAHM.00006777, OAHM.00006787, OAHM.00006797, OAHM.00006807, OAHM.00006817, OAHM.00006827, OAHM.00006837, OAHM.00006847, OAHM.00006857, OAHM.00006867, OAHM.00006877, OAHM.00006887, OAHM.00006897, OAHM.00006907, OAHM.00006917, OAHM.00006927, OAHM.00006937, OAHM.00006947, OAHM.00006957, OAHM.00006967, OAHM.00006977, OAHM.00006987, OAHM.00006997, OAHM.00007007, OAHM.00007017, OAHM.00007027, OAHM.00007037, OAHM.00007047, OAHM.00007057, OAHM.00007067, OAHM.00007077, OAHM.00007087, OAHM.00007097, OAHM.00007107, OAHM.00007117, OAHM.00007127, OAHM.00007137, OAHM.00007147, OAHM.00007157, OAHM.00007167, OAHM.00007177, OAHM.00007187, OAHM.00007197, OAHM.00007207, OAHM.00007217, OAHM.00007227, OAHM.00007237, OAHM.00007247, OAHM.00007257, OAHM.00007267, OAHM.00007277, OAHM.00007287, OAHM.00007297, OAHM.00007307, OAHM.00007317, OAHM.00007327, OAHM.00007337, OAHM.00007347, OAHM.00007357, OAHM.00007367, OAHM.00007377, OAHM.00007387, OAHM.00007397, OAHM.00007407, OAHM.00007417, OAHM.00007427, OAHM.00007437, OAHM.00007447, OAHM.00007457, OAHM.00007467, OAHM.00007477, OAHM.00007487, OAHM.00007497, OAHM.00007507, OAHM.00007517, OAHM.00007527, OAHM.00007537, OAHM.00007547, OAHM.00007557, OAHM.00007567, OAHM.00007577, OAHM.00007587, OAHM.00007597, OAHM.00007607, OAHM.00007617, OAHM.00007627, OAHM.00007637, OAHM.00007647, OAHM.00007657, OAHM.00007667, OAHM.00007677, OAHM.00007687, OAHM.00007697, OAHM.00007707, OAHM.00007717, OAHM.00007727, OAHM.00007737, OAHM.00007747, OAHM.00007757, OAHM.00007767, OAHM.00007777, OAHM.00007787, OAHM.00007797, OAHM.00007807, OAHM.00007817, OAHM.00007827, OAHM.00007837, OAHM.00007847, OAHM.00007857, OAHM.00007867, OAHM.00007877, OAHM.00007887, OAHM.00007897, OAHM.00007907, OAHM.00007917, OAHM.00007927, OAHM.00007937, OAHM.00007947, OAHM.00007957, OAHM.00007967, OAHM.00007977, OAHM.00007987, OAHM.00007997, OAHM.00008007, OAHM.00008017, OAHM.00008027, OAHM.00008037, OAHM.00008047, OAHM.00008057, OAHM.00008067, OAHM.00008077, OAHM.00008087, OAHM.00008097, OAHM.00008107, OAHM.00008117, OAHM.00008127, OAHM.00008137, OAHM.00008147, OAHM.00008157, OAHM.00008167, OAHM.00008177, OAHM.00008187, OAHM.00008197, OAHM.00008207, OAHM.00008217, OAHM.00008227, OAHM.00008237, OAHM.00008247, OAHM.00008257, OAHM.00008267, OAHM.00008277, OAHM.00008287, OAHM.00008297, OAHM.00008307, OAHM.00008317, OAHM.00008327, OAHM.00008337, OAHM.00008347, OAHM.00008357, OAHM.00008367, OAHM.00008377, OAHM.00008387, OAHM.00008397, OAHM.00008407, OAHM.00008417, OAHM.00008427, OAHM.00008437, OAHM.00008447, OAHM.00008457, OAHM.00008467, OAHM.00008477, OAHM.00008487, OAHM.00008497, OAHM.00008507, OAHM.00008517, OAHM.00008527, OAHM.00008537, OAHM.00008547, OAHM.00008557, OAHM.00008567, OAHM.00008577, OAHM.00008587, OAHM.00008597, OAHM.00008607, OAHM.00008617, OAHM.00008627, OAHM.00008637, OAHM.00008647, OAHM.00008657, OAHM.00008667, OAHM.000

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 Taxonomy Search 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 shows the PubMed website interface. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' links, and 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. The main content area is divided into three columns:

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

Below these columns are three sections:

- Latest Literature:** Lists new articles from highly accessed journals, such as '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 promotes 'PubMed Journals'.
- Trending Articles:** Lists PubMed records with recent increases in activity, such as 'Clinical outcomes of a scapular-focused treatment in patients with subacromial pain syndrome: a systematic review', 'Effectiveness of neuromuscular taping on painful hemiplegic shoulder: a randomised clinical trial', 'Rare and low-frequency coding variants alter human adult height', 'Moconnell'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', and 'Parvovirus B19 during pregnancy: a review'.
- PubMed Commons:** Lists featured comments, such as 'Circuits in reward & aversion: @gstuber posts journal club review of study identifying neuron populations', 'Migrating database: @odsouthan provides updated info for finding pharmacological data resource', 'Genes for lactation persistence in cattle: @Eric_Fauman critiques findings of a genome-wide association study', 'Reviewing replication: R Tibshirani critiques statistics; A Collings cross-posts comment from original study authors', and 'Evaluating connection between food energy supply & obesity: @JamesonVoss discusses ecologic study'.

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-U, 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

```
ctccggcagc cggaggtcat cctgctagac tcagacctgg atgaacccat agacttgcgc      60
tcggtcaaga gccgcagcga ggccggggag ccgccagct cctccaggt gaagcccagag     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"
     gene              1..1599
                        /gene="SUC2"
     mRNA              1..1599
                        /gene="SUC2"
     CDS               1..1599
                        /gene="SUC2"
                        /codon_start=1
                        /product="invertase"
                        /protein_id="AFN08663.1"
                        /db_xref="GI:393395466"
                        /translation="MLIQAFLEFLLAGFAAKIASMNYNETSDRPLVHFTFNKGMNDPFI
                        GLWYIEKDAKWHLYFYQYHMDYVMGTPLFWGHATSDDLTHWEDPIAIAPKRNDGGAF
                        SGRMVDYHNTSGFPNDTIDFNQKVAIWTNTPESEHQYISYELDGGTPTTKYQKRP
                        VLAANSTQPHDPKYPWEPSPQKWMITAASQOYKIEIYSSDOLSKSWLESAFANGFL
                        GYQRECPGLIEVPTKQOPSKSYWVFIISINDGAPAGSSFNQYFVGSYNGTHFEAFDNG
                        SRVDFGRDYALQTFYFDPTTYSALGLANASHWREYSAFVFTNPWRSSNSLVRKFSL
                        NTEYQANPTELINLKAEPILNINAGPMSRFATNITLTKANSYVDSNSTGTLEPE
                        LVYAVNTQTISKSVFFDLSEMFEGLEDPEYLSNGFEASASFFLDRGNSKVKFVKE
                        NPFTNPMVSNQPFKSERNDLSYKVIYGLLDQNLILEYFDNGDQVSTNTYFMTGNAL
                        GSVMMITGYNLFLYIDSEFQVDEK"
```

```
ORIGIN
1 atgcttttgc aagcttttct tttcttttgc gctggttttg cagccaaaat ctctgactca
61 atgcaaaaag aaactagcga tagacttttg gtccacttca cccccaaaaa ggcctggatg
121 aatgacccaa atgggttttg gtaactgaaa aagatgcca atggcactct gacttttcaa
181 tcaaacccaa atgacacnct atggggtacg ccttggtttt gggggccatg tacttccgat
241 gatttgactc attggaaga tgaacccatt gctatgctc ccaagcgtaa cgtattccgt
301 gcttttctcg gcttccatggt ggttggattca aacacacga gttggttttt caatgatact
361 attgatccaa gcaaacagat gcttgcgatt tgcacttata acactcctga aagtgaagag
421 caatcacata gctattctct tgatggctgt tcaactttta ctgaatacca aaagaacctt
481 gttttagctg ccaactccac tcaattcaga gatcccaagg tgttctggta tgaaccttct
541 caaaaatgga ttatgacgpc tgcacaatca caagatcca aaattgaat ttaactctrt
601 gatgacttga agtactggaa gctagaactc gcttttgcct atgaagtttt cttaggctac
661 caatgaaat gttccagttt gattgaagtc ccaactgagc aagatccttc caaatcctat
721 tgggtcattg ttatttctat caatccaggt gactctgctg ggggttcttt caaacceat
781 ttgttggat ctttccatgg tactcatttt gaagcgtttg acactcaatc tagagtgtta
841 gattttgcta aggactacta tgccttgcaa acttttctca acacagaccc aactgactgt
901 tcaactttag gattgctgt ggtctcaaan tggagtagca gttcctttgt ccaactaac
961 ccaatgagat cctccatgct ttgttccgca aagttttctt tgaactcctga atactceagct
1021 aatccagaga ctgaattgat caatttgaaa gccaaaccaa tattgaacct tagtaattct
1081 ggttccctgt ctgtttttgc tactaacaca acttcaacta aggcacttca ttaacaatgc
1141 gatttgaaga actgacttgg tactctagag tttaggtttg tttaagctgt taacacaaa
1201 caaacctat ccaaatcctt ctlttccgca ttctcacttt gtttcaaggg tttagaagat
1261 ctgagaagat atttaagaat ggtttttgaa gccagtgctt ctctcttctt ttggccctgt
1321 ggtactctca agtccagttt tgcacagagg aaccatcttt tcaacaaagc aatgtctgtc
1381 acaaaccaac ctttcaagtc tgaagaagca ctaagtact ataaagtga cgccttctgt
1441 gatcaaaaac ttttgaatt gacttcaaac gatggagatg ttgttctca aataccttac
1501 ttctagacca ccgtaaacgc tctaggatct gtgaactaga ccaactggtt cgtataattg
1561 ttctacatg acaagttcaa agtaagggaa gtaaatag
```

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:4932"
    gene              <1..>1599
                     /gene="SUC2"
    mRNA              <1..>1599
                     /gene="SUC2"
                     /product="invertase"
    CDS               1..1599
                     /gene="SUC2"
                     /codon_start=1
                     /product="invertase"
                     /protein_id="AFN08663.1"
                     /db_xref="GI:393395466"
                     /translation="MLLQAFLLFLLAGFAAKISASMTNETSDRPLVHFTPNKGWMNDPN
GLWYDEKDAKWHLYFQYNPNDTVWGTPLFWGHATSDDLTHWEDEPIAIAIPKRNDSGAF
SGSMVVDYNNNTSGFFNDTIDPRQRCVAIWTYNTPESEEQYISYSLDGGYTFTEYQKNP
VLAANSTQFRDPKVFVWYEPSQKWIMTAAKSQDYKIEIYSSDDLKSWKLESAFANEGFL
GYQYECPLIEVPTEQDPSKSYWVMFISINPGAPAGGSFNQYFVGSFNGTHFEAFDNO
SRVVDVDFGKDYYALQTFNTDPTYGSALGIAWASNWEYSAFVPTNPWRSMSLVKRFSL
NTEYQANPETELINLKAEPILNISNAGPWSRFATNTTLTKANSYNVDLSNSTGTLEFE
LVYAVNTTQTISKSVFPDLSLWFKGLEDPEEYLRMGFEASASSFFLDRGNSKVKEVKE
NPYFTNRMSVNNQPFKSENDLSYYKVYGLLDQNIILELYFNDGDVVSTNTYFMTTGNAL
GSVNMTTGVDNLFYIDKQVREVK"
```

La séquence (format Genbank)

ORIGIN

```
1 atgcttttgc aagctttcct tttccttttg gctggttttg cagccaaaat atctgcatca
61 atgacaaacg aaactagcga tagaccttg gtccacttca cacccaacaa gggctggatg
121 aatgacccaa atgggttgtg gtacgatgaa aaagatgcca aatggcatct gtactttcaa
181 tacaacccaa atgacaccgt atggggtagc ccattgtttt ggggccatgc tacttccgat
241 gatttgactc attgggaaga tgaaccatt gctatcgctc ccaagcgtaa cgattcaggt
301 gctttctctg gctccatggg ggttgattac aacaacacga gtgggttttt caatgatact
361 attgatccaa gacaaagatg cgttgcgatt tggacttata aactcctga aagtgaagag
421 caatacatta gctattctct tgatgggtgg tacactttta ctgaatacca aaagaaccct
481 gtttttagctg ccaactccac tcaattcaga gatccaaagg tgttctggta tgaaccttct
541 caaaaatgga ttatgacggc tgccaaatca caagactaca aaattgaaat ttactcctct
601 gatgacttga agtcctggaa gctagaatct gcatttgcta atgaaggttt cttaggctac
661 caatatgaat gtccagggtt gattgaagtc ccaactgagc aagatccttc caaatcctat
721 tgggtcatgt ttatttctat caatccaggt gcacctgctg gcggttcctt caaccaatat
781 tttgttggat cettcaatgg tactcatttt gaagcgtttg acaatcaatc tagagtggta
841 gattttggta aggactacta tgccttgcaa actttcttca acacagacc aacgtacggg
901 tcagcattag gtattgctg ggcttcaaac tgggagtaca gtgcctttgt cccaactaac
961 ccatggagat catccatgtc tttggtcgcg 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 caaacatat ccaaatccgt ctttcccgcac ttatcacttt ggttcaaggg tttagaagat
1261 cctgaagaat atttaagaat gggttttgaa gccagtgcct cttccttctt tttggaccgt
1321 ggtaactcta aggtcaagtt tgtcaaggag aacctatatt tcacaaacag aatgtctgtc
1381 aacaaccaac cattcaagtc tgagaacgac ctaagttact ataaagtgta cggcctactg
1441 gatcaaaaca tcttggaatt gtacttcaac gatggagatg tggtttctac aaatacctac
1501 ttcatgacca ccggtaacgc tctaggatct gtgaacatga cactgggtgt 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

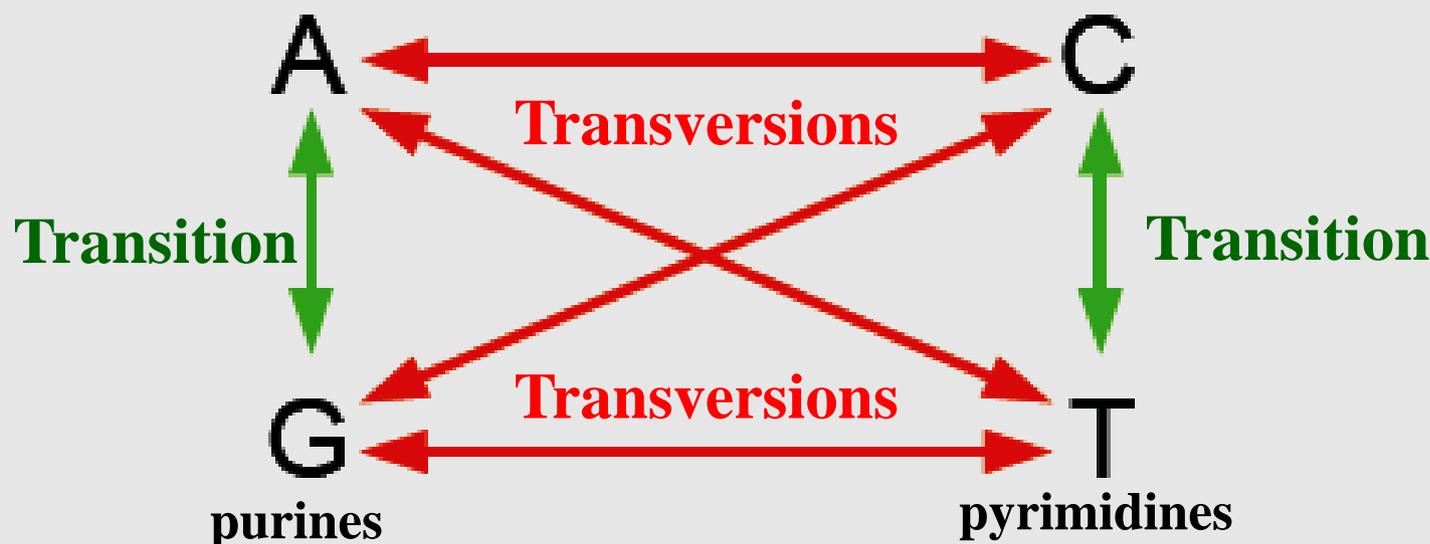
- ✓ Une ligne de commentaires précédé de « > »
- ✓ La séquence brute (pas d'espace, ni de nombre)

```
>Human Polycomb 2 homolog (hPc2) mRNA, partial cds
ctccggcagcccgagggtcatcctgctagactcagacctggatgaacccat
agacttgcgctcgggtcaagagccgcagcgcaggccggggagccgcccagct
ccctccagggtgaagcccgagacaccggcgtcggcggcgggtggcgggtggcg
gcggcagcggcaccaccacgacggcgggagaagcctccagccgaggccca
ggacgaacctgcagagtcgctgagcgcagttcaagcccttctttgggaata
taattatcaccgacgtcaccgcgaactgcctcaccgttactttcaaggag
tacgtgacgggtg
```


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);
ATCTCGNCTATC
 - ✓ 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 ?

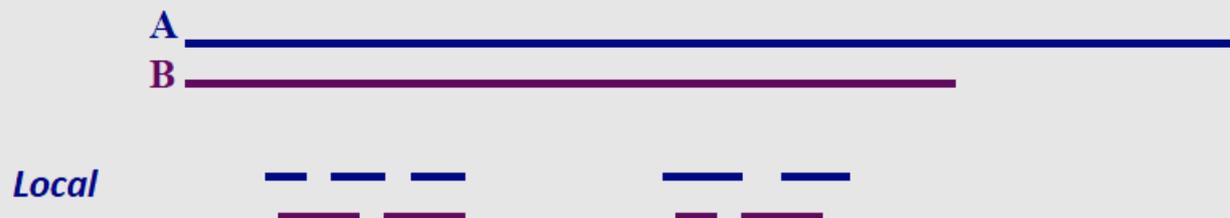


Finalités différentes

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.

✓ 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 $S(A, B)$.

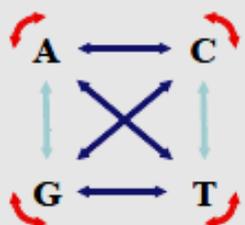
Matrices de scores pour l'ADN

➤ 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).

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

(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 WUNSH : 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 :

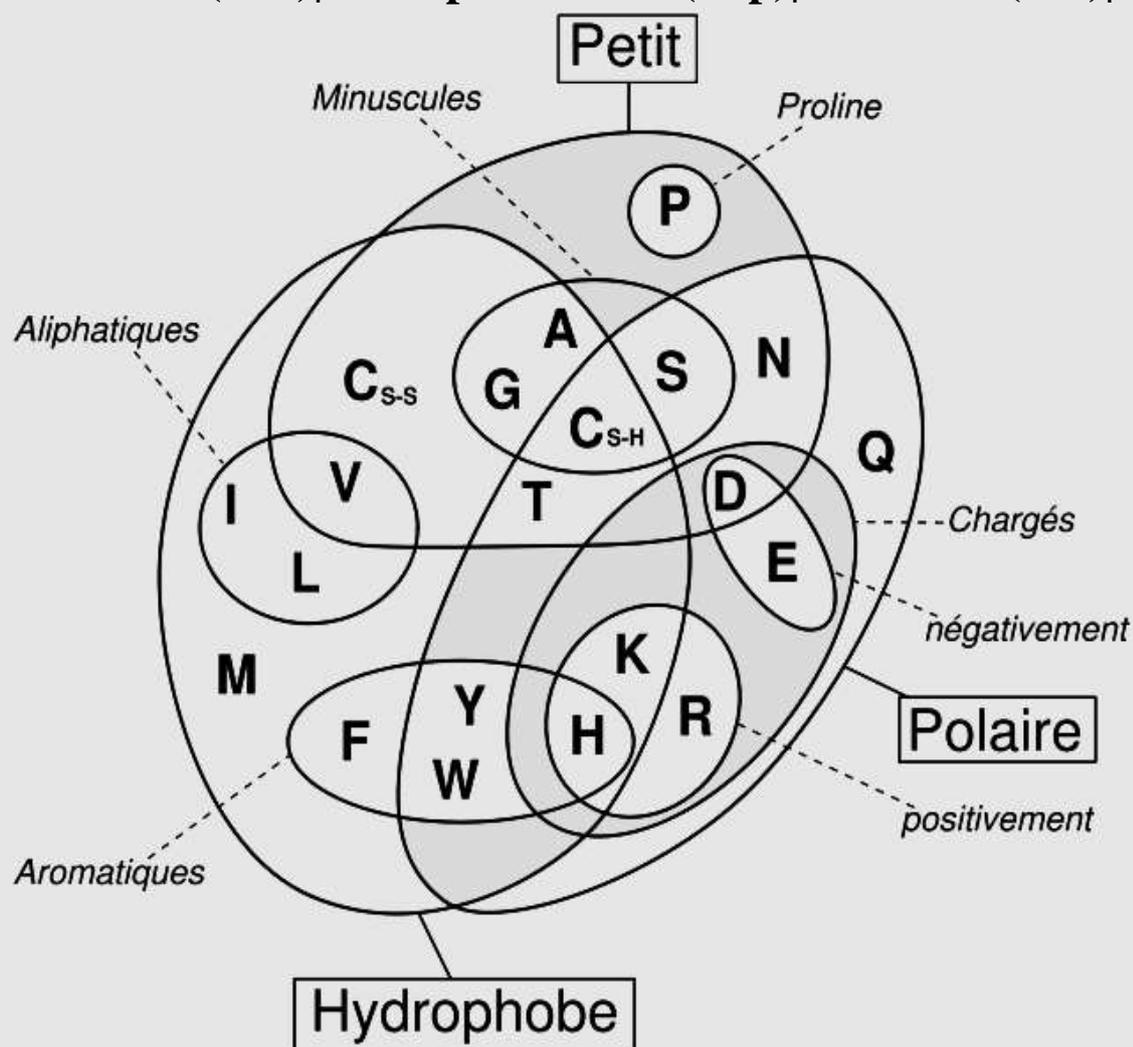
$$\%id = (7/9) * 100 = 77,78\%$$

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.

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 :

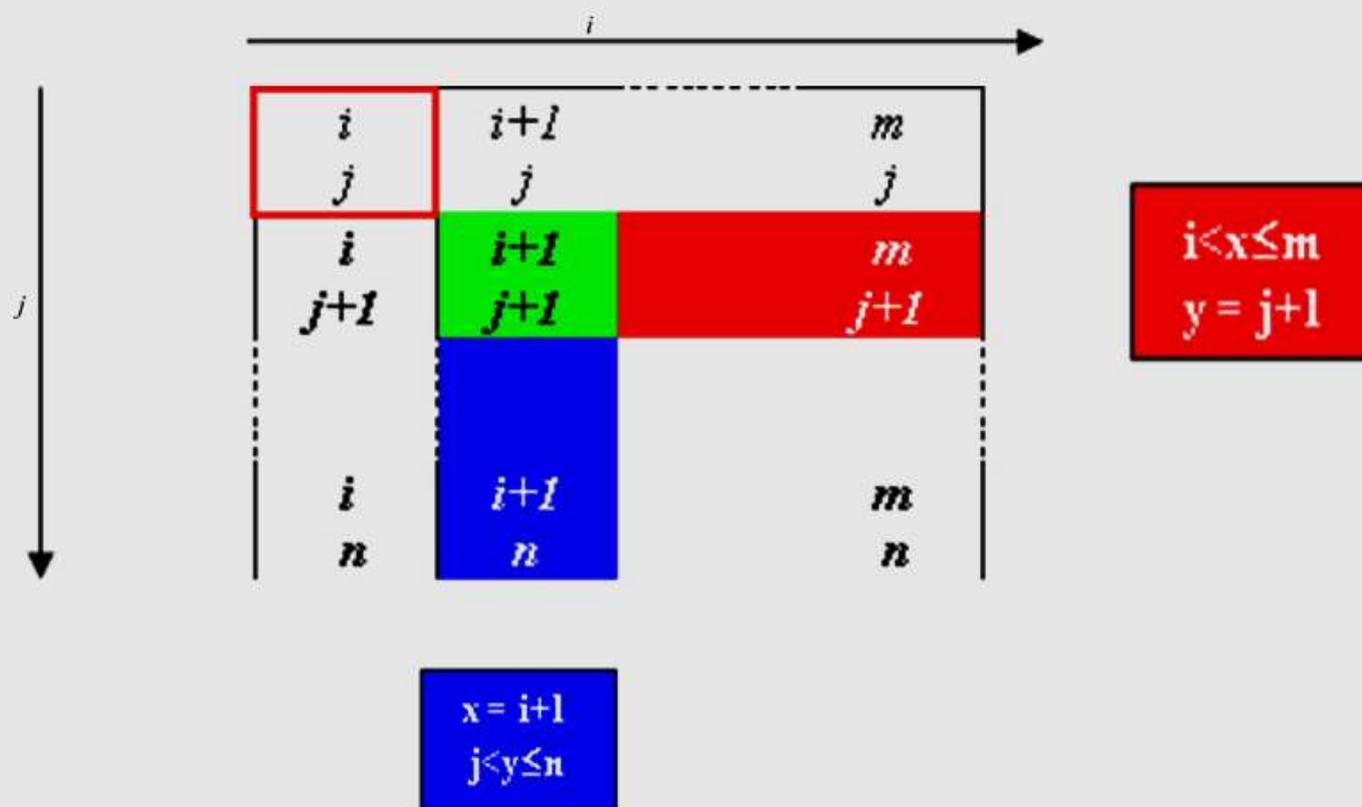
$$i < x \leq m \text{ et } y = j+1$$

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



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

	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

(matrice de substitution)

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 _(max)	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

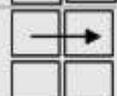
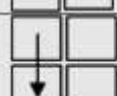
ALIGNEMENT DE DEUX SÉQUENCES

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^{ème} 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.

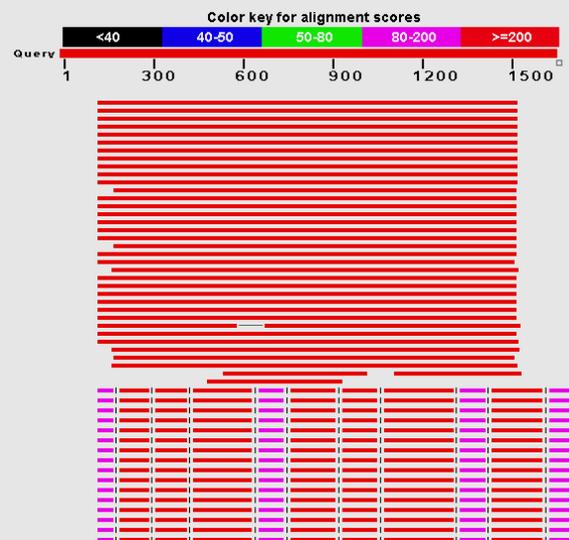
- ✓ Parmi les méthodes employant ces heuristiques progressives, citons: **CLUSTAL W**, **Dialign**, **DCA**, **MSA**, **PIMA**, **MULTALIGN**, **PILEUP**, **Coffee**, **HMMT**, **T-Coffee** , **POA** , **ProbCons** , **Multi-LAGAN**, **Muscle**, **MAFFT**, **SAGA**, etc.
- ✓ **Il n'existe pas de méthode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

ALIGNEMENT MULTIPLE

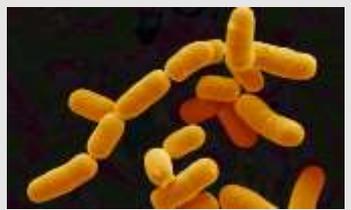
ALIGNEMENT D'UNE SÉQUENCE AVEC UNE BANQUE

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

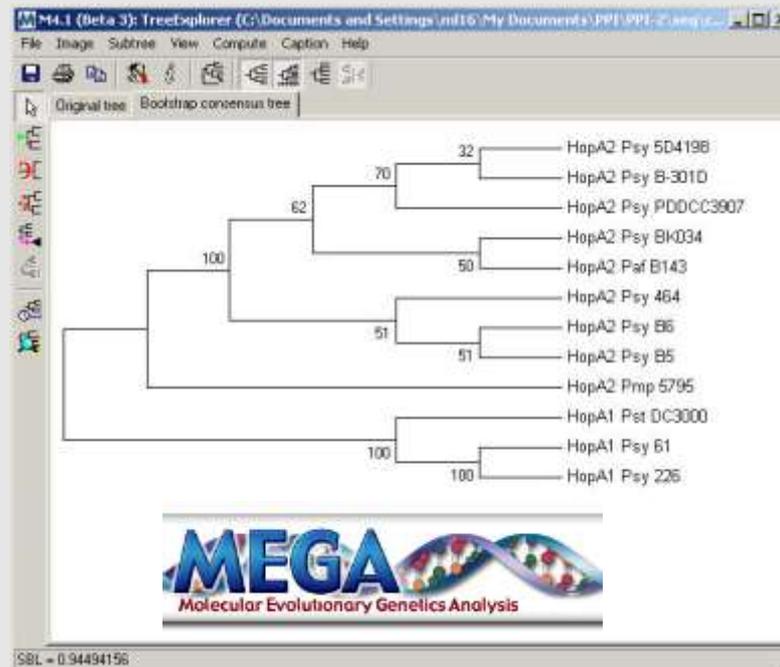
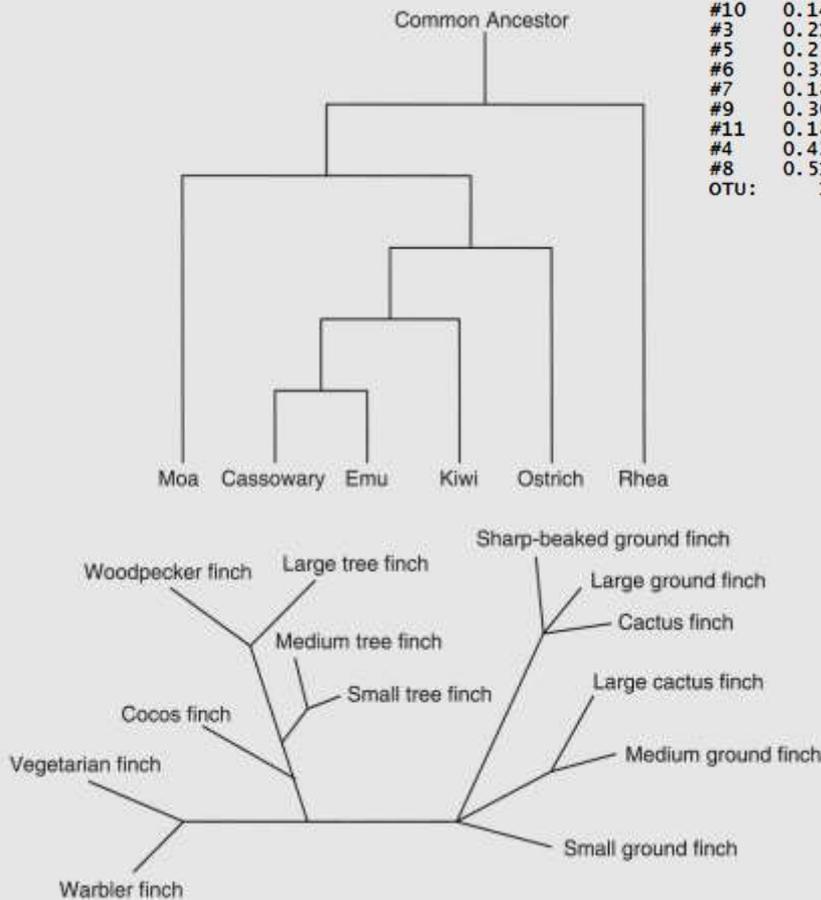


0.000 0.547 11 -LOG2(S(SM)) before clustering

#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).

E

pere

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

F

noeud

symbolise des ancêtres
hypothétiques partagés par
les UTO

temps

fils

A

fils

B

C

D

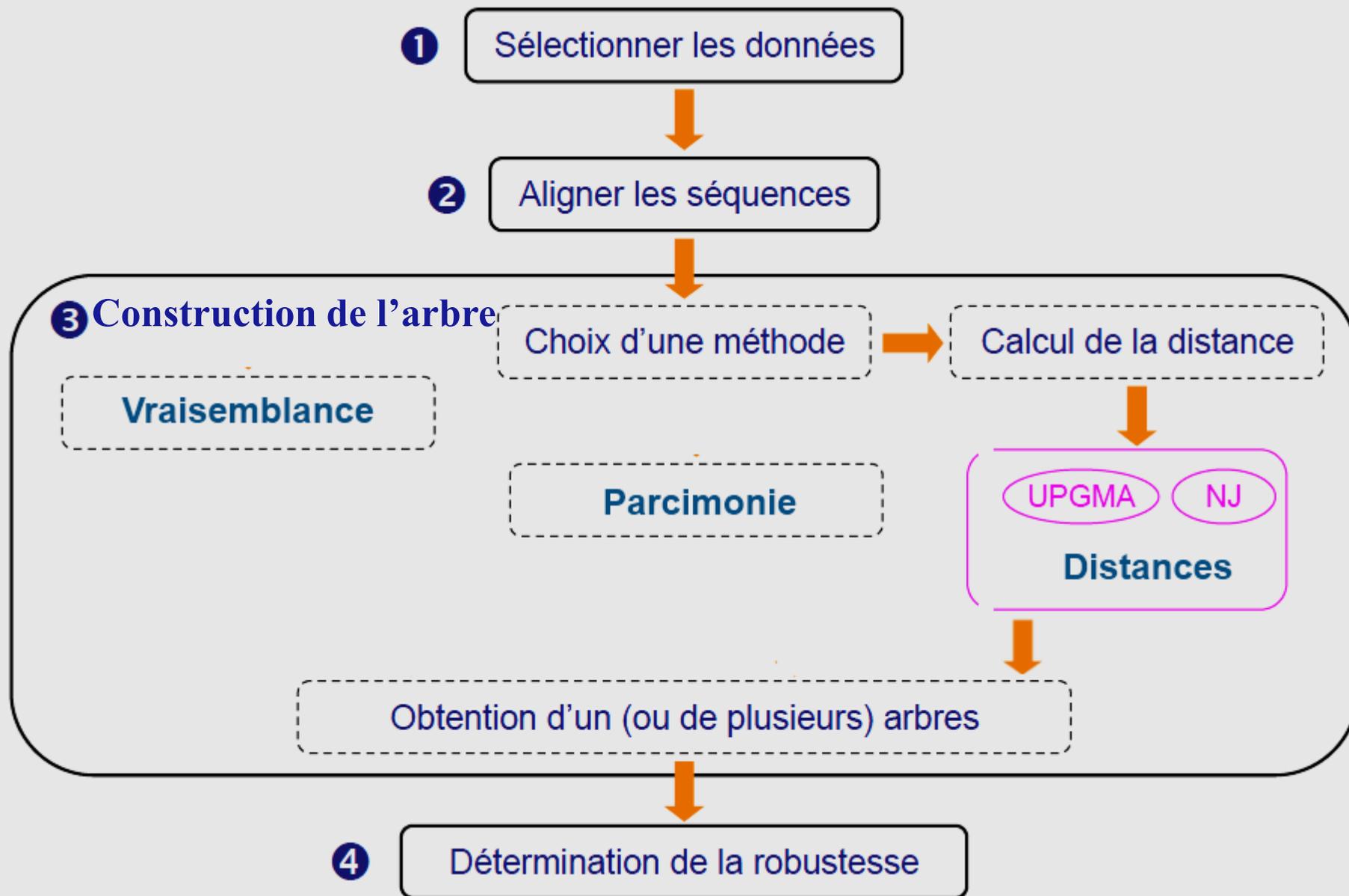
feuille

**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



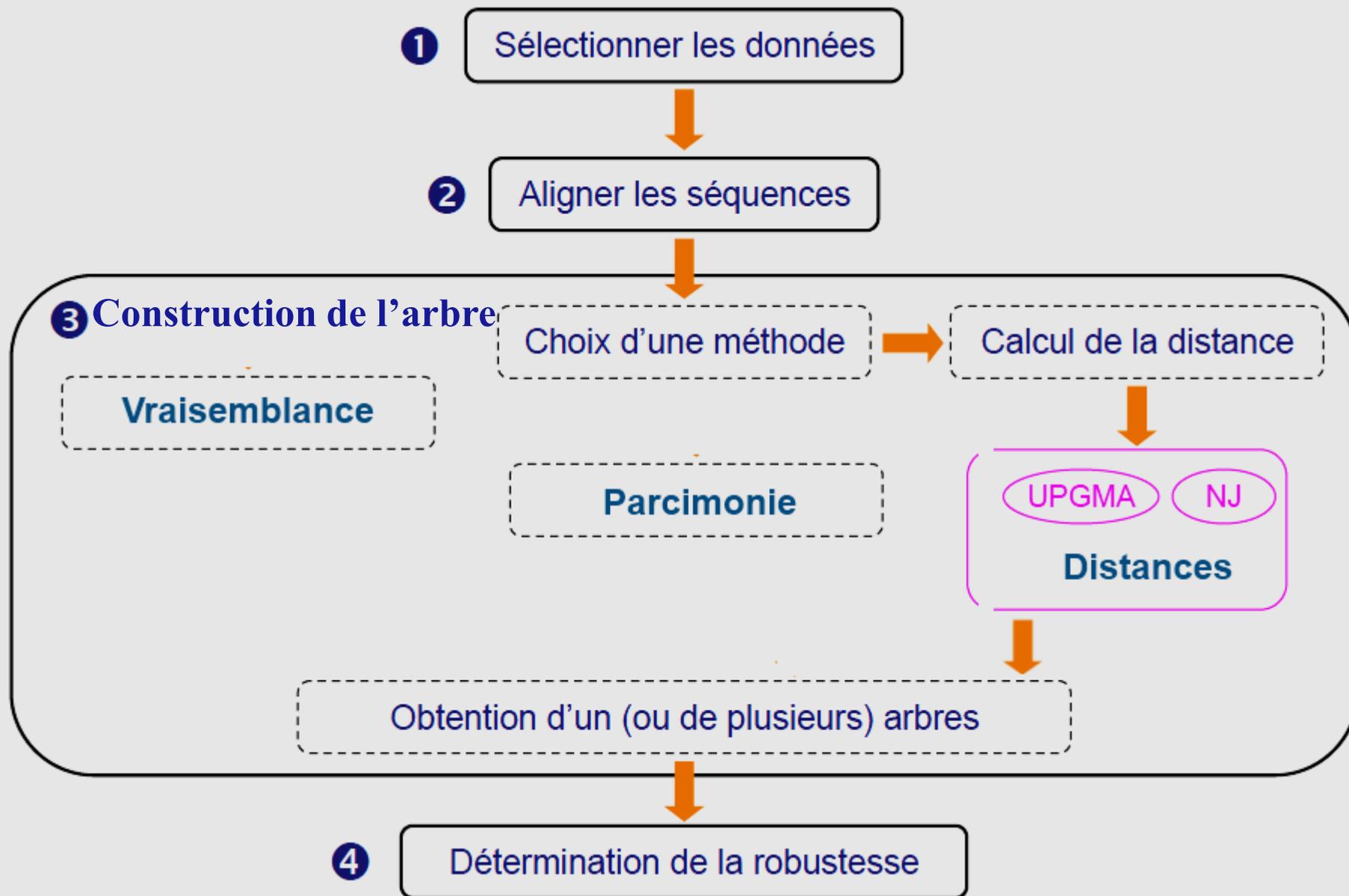
NEIGHBOR-JOINING (Saitou et Nei, 1987):

- ✓ 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



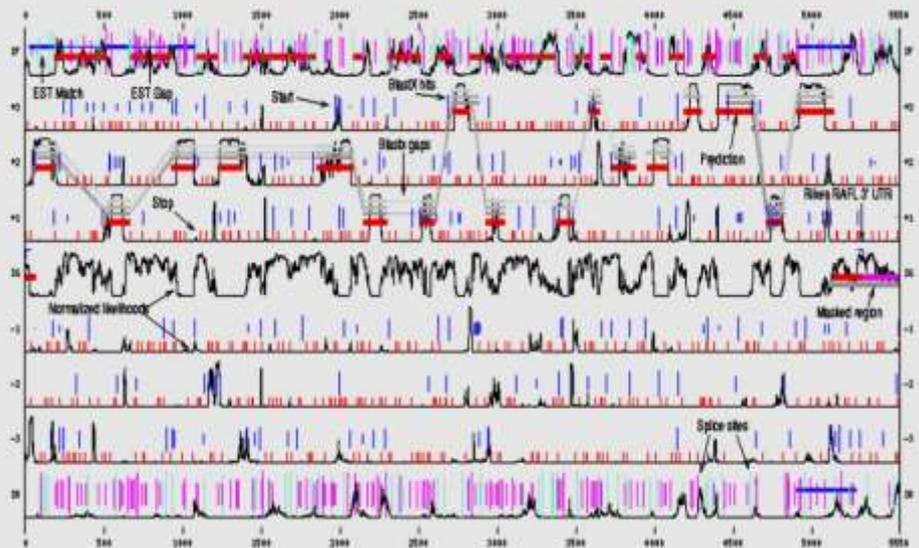
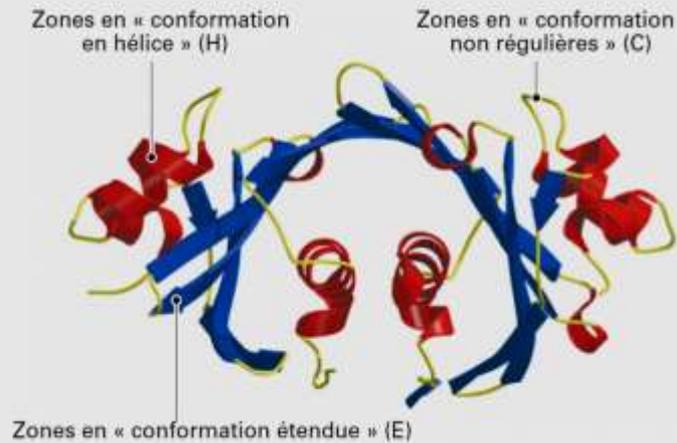
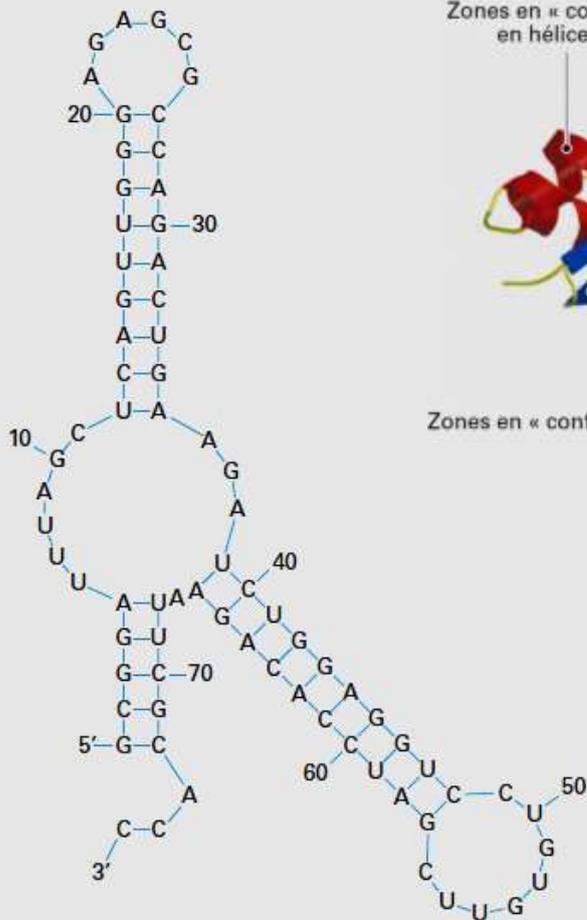
FIABILITÉ D'UN ARBRE

- ✓ 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 structurale (Where)

Permet de positionner les gènes et leurs produits :

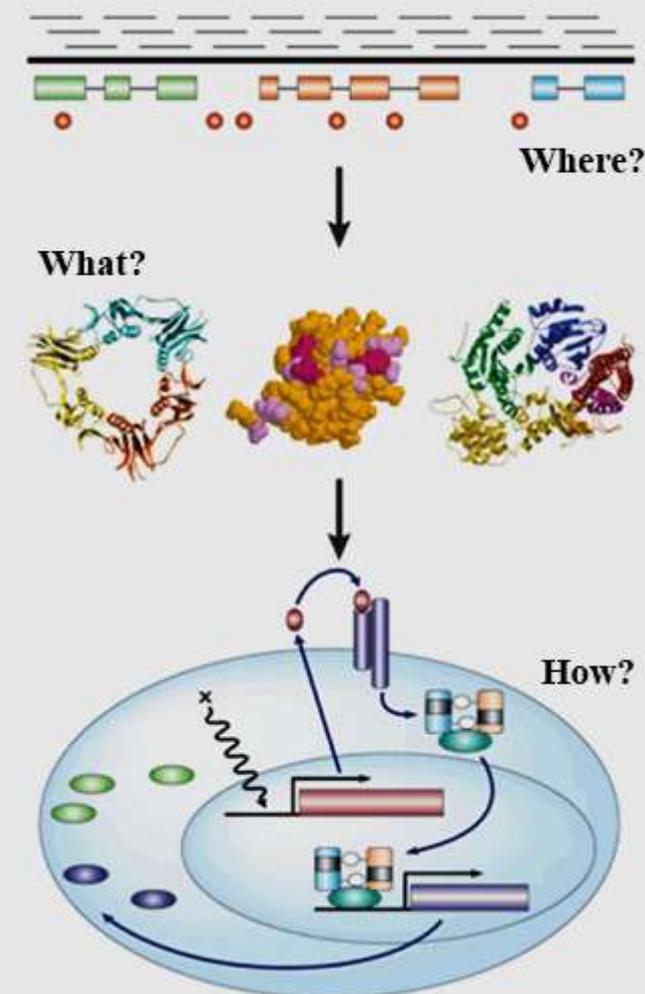
- éléments répétés (éléments transposables, satellites, etc.)
- gènes des ARN stables (ARNr, ARNt...)
- gènes protéiques
- régions régulatrices, etc.

✓ 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_c 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.

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 aboutit au développement de logiciels de prédiction plus ou moins efficaces:

Prédiction des structures secondaires:

- **PSIPRED, SSPRO, 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