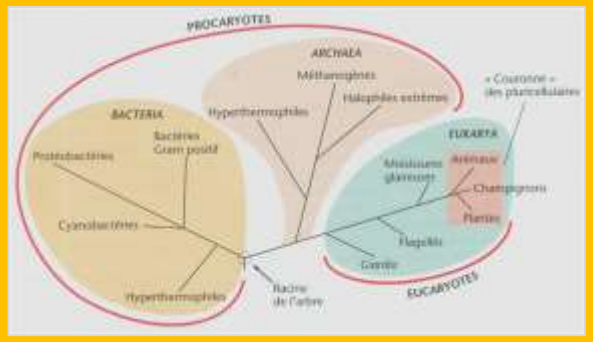
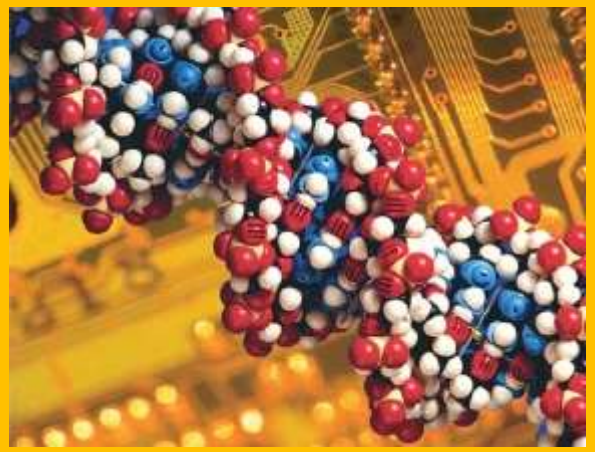
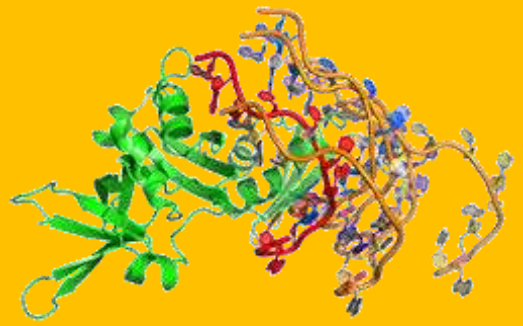
Université Frères Mentouri Constantine
Institut de la Nutrition, de l'Alimentation et des Technologies Agro-alimentaires (INATAA)
Master 1-Technologies Alimentaires
Année universitaire 201-2020



INITIATION À LA BIOINFORMATIQUE

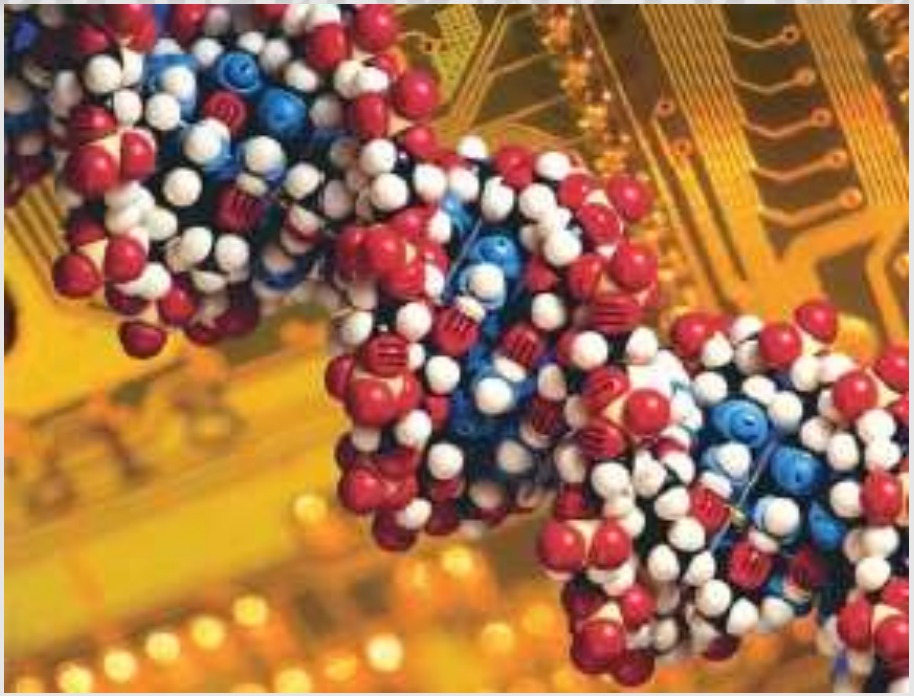




CE QUE NOUS ALLONS VOIR:

- ✓ 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*.

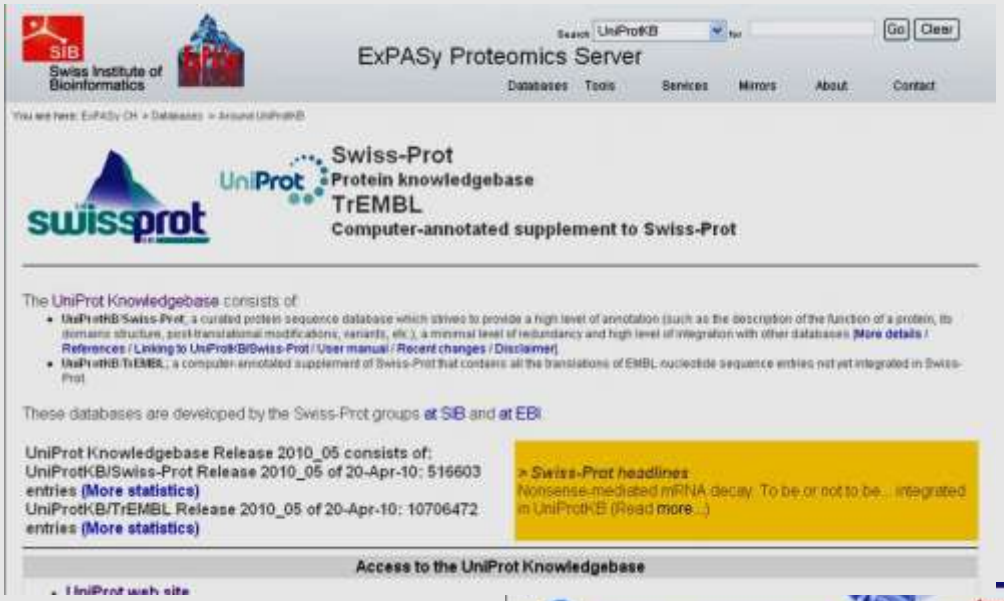
*la bioinformatique tire sa définition de deux concepts importants : la biologie et l'information car le suffixe informatique n'a rien à voir avec l'utilisation des ordinateurs pour la biologie.

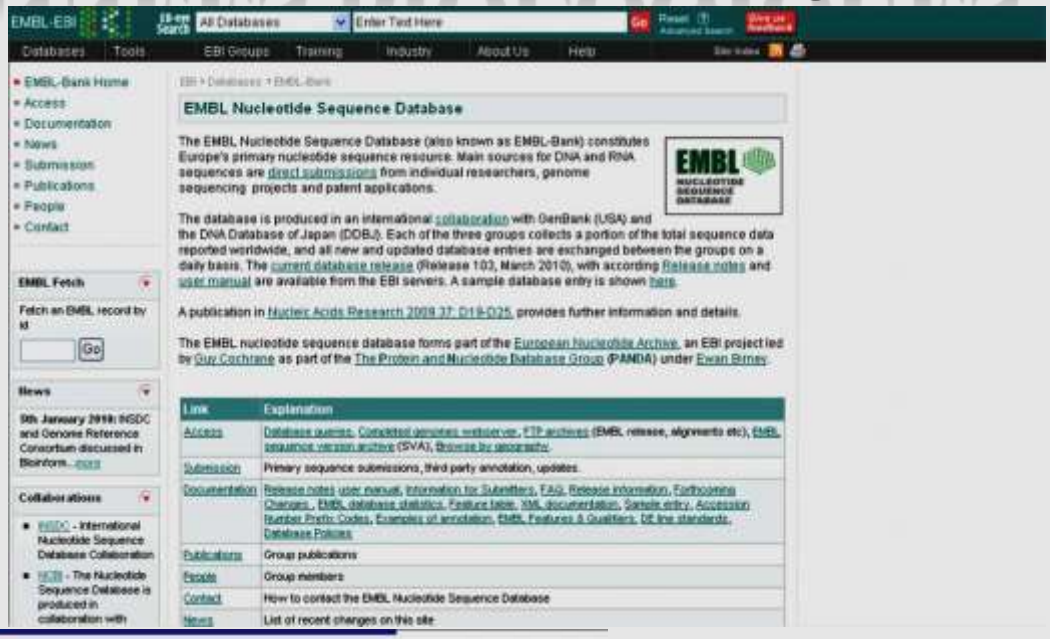
L'INFORMATION BIOLOGIQUE

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





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 gènomes, 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**).

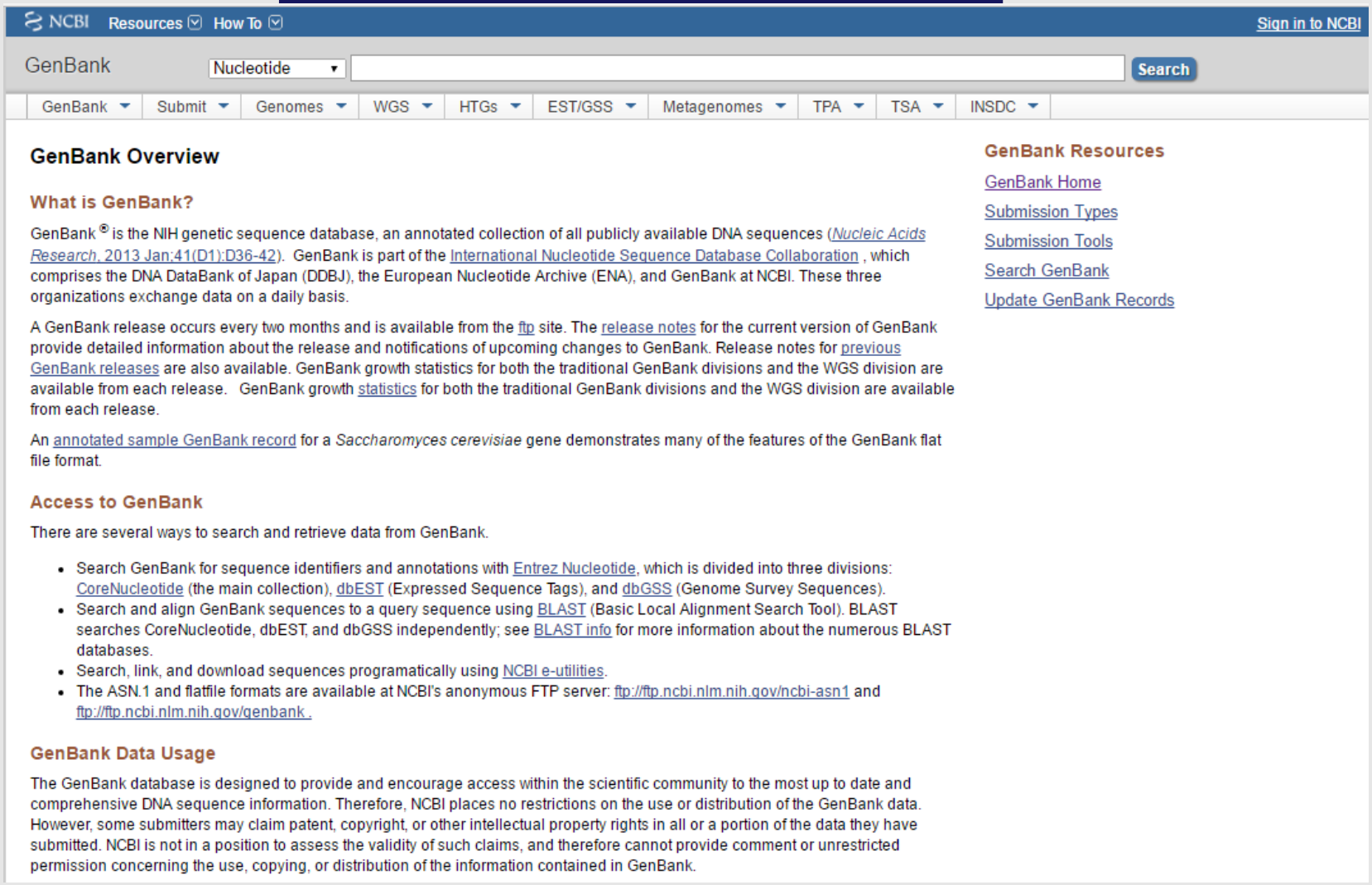
BANQUES DE DONNÉES GÉNÉRALISTES

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/



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

BANQUES DE DONNÉES GÉNÉRALISTES

BD DE SÉQUENCES NUCLÉIQUES

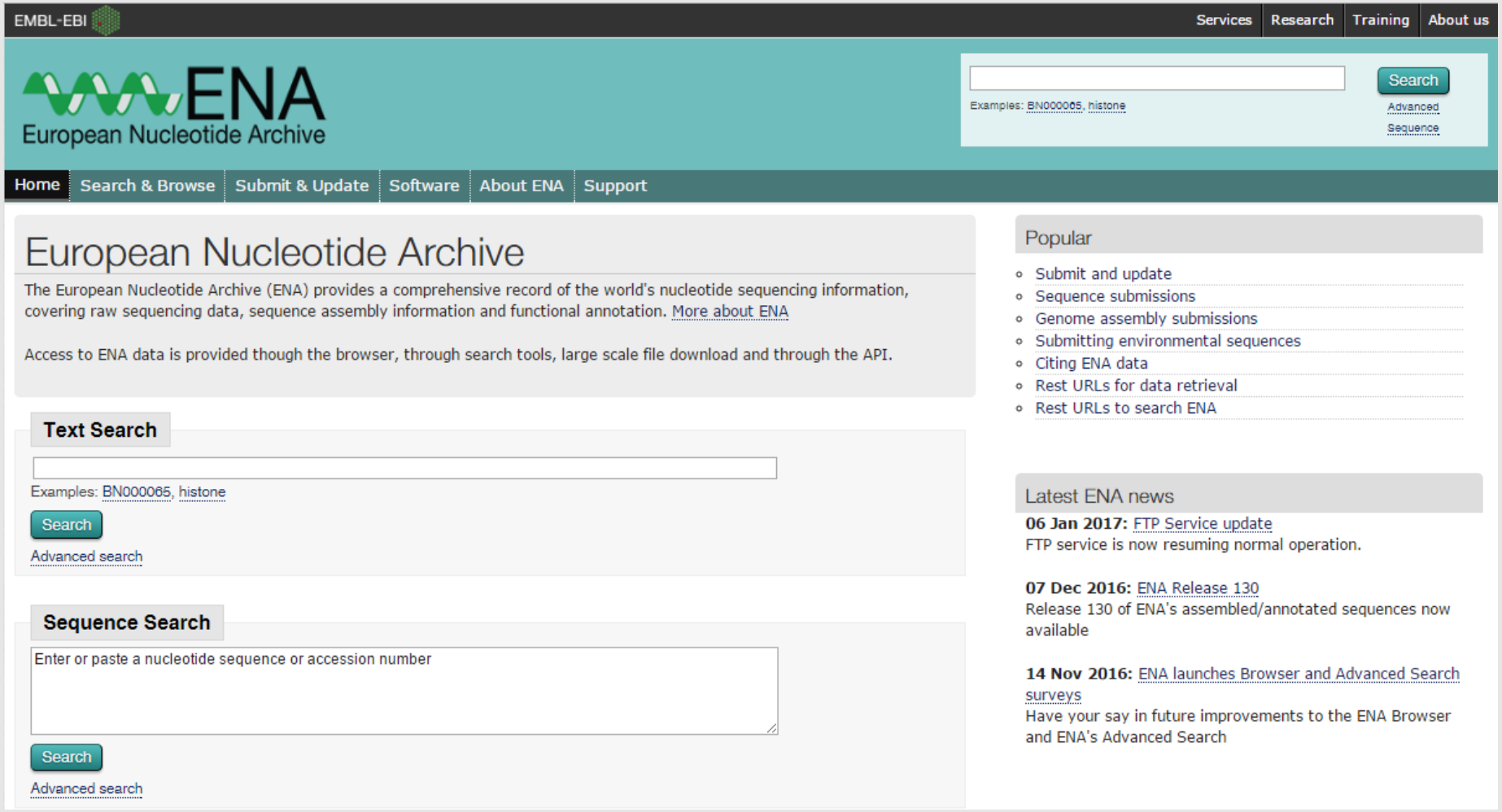
ENA

(*European Nucleotide Archive*)

banque européenne * diffusée par **EMBL-EBI** (Cambridge, UK)

(*European Molecular Biology Library-European Bioinformatics Institute*)

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



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

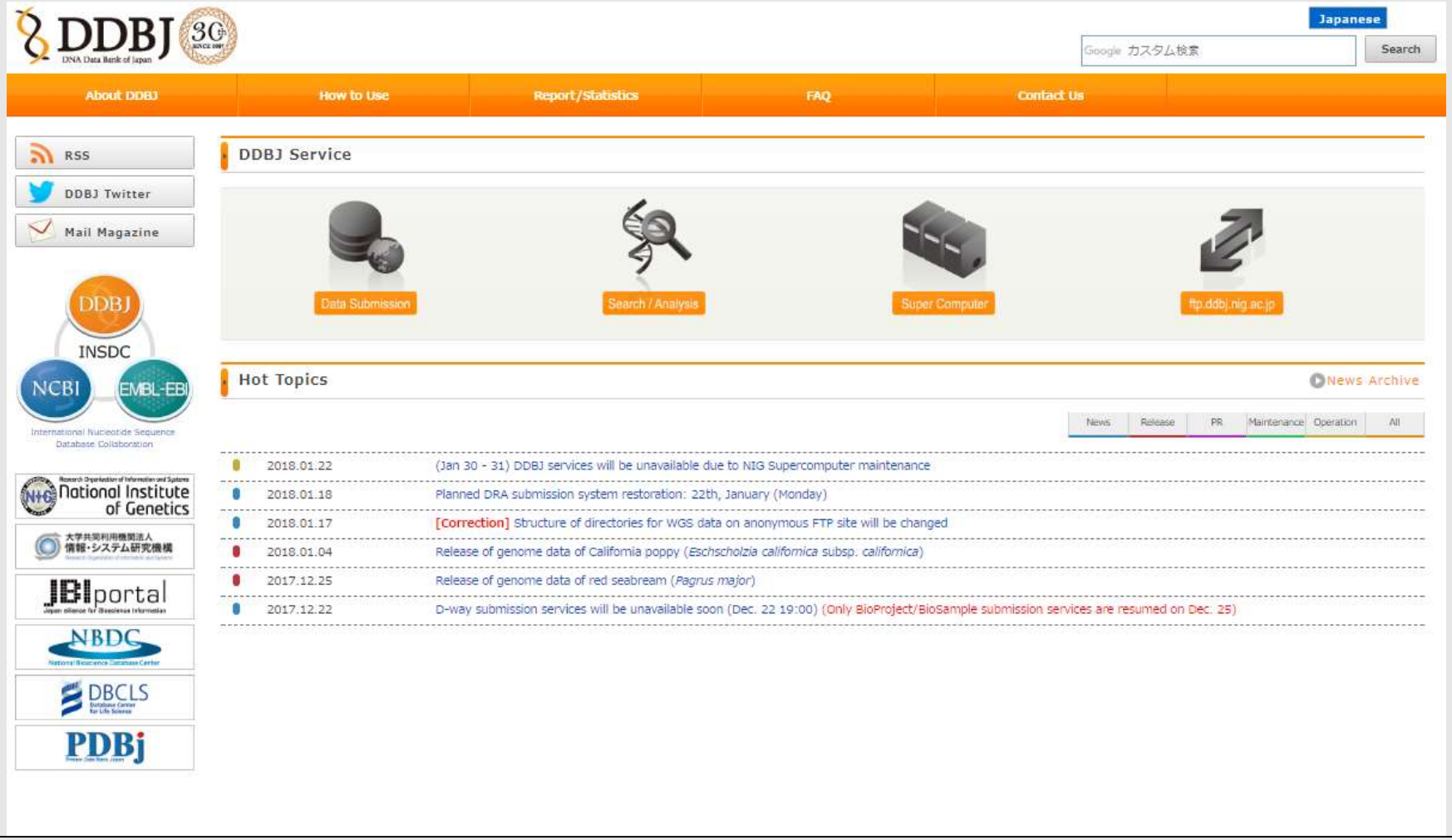
BANQUES DE DONNÉES GÉNÉRALISTES

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



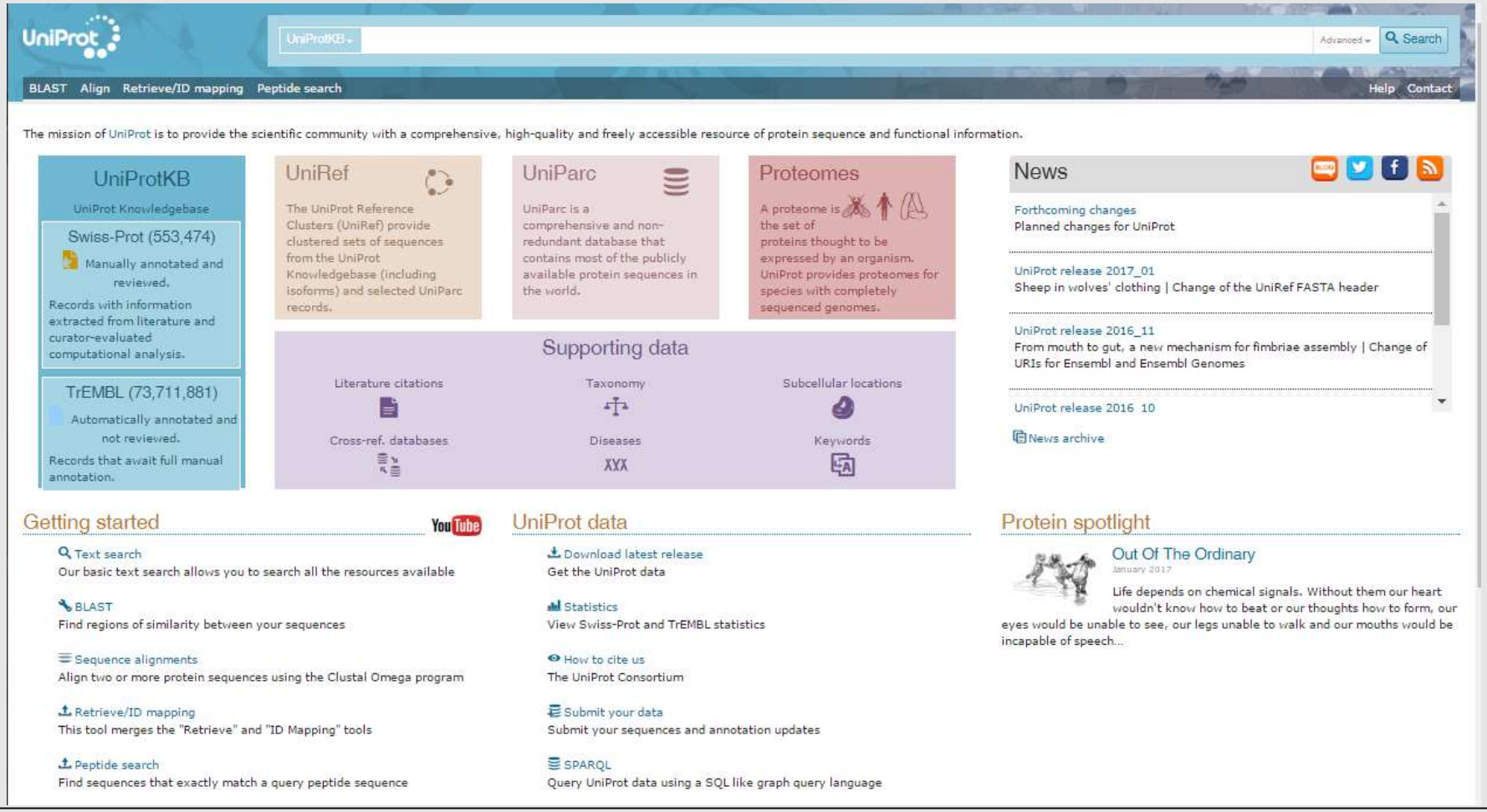
*En Décembre 2017, DDBJ archivait plus de 948 millions de séquences nucléiques

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

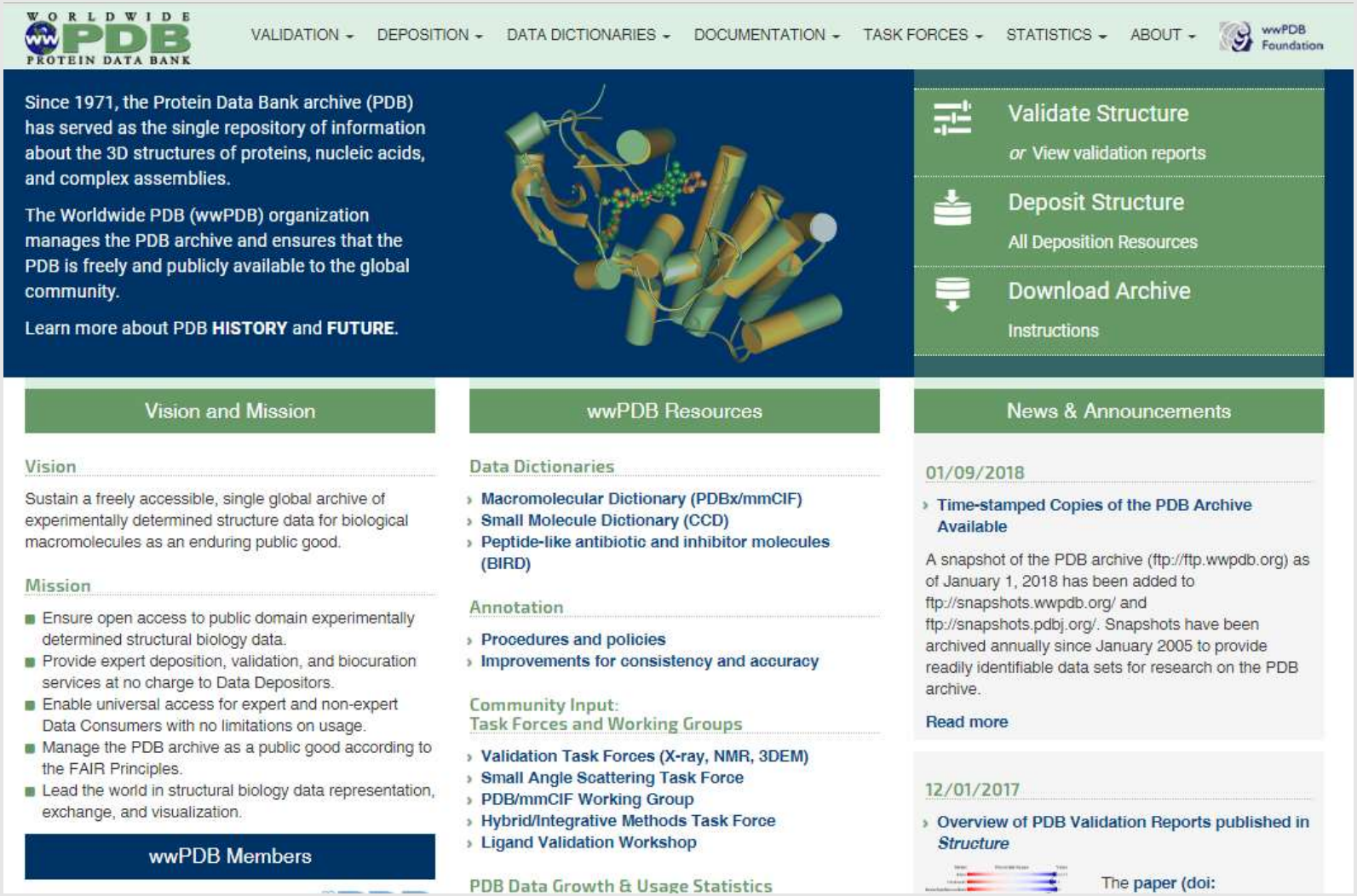


*En Décembre 2017, Uniprot archivait: >102 millions (TrEMBL) et > 556 mille (UniProtKB/Swiss-Prot) séquences protéiques annotées automatiquement et manuellement, non revues et revues, respectivement.

PDB (*Protein Data Bank*) *:

Structure 3D de protéines, acides nucléiques et autres molécules

http://www.wwpdb.org/

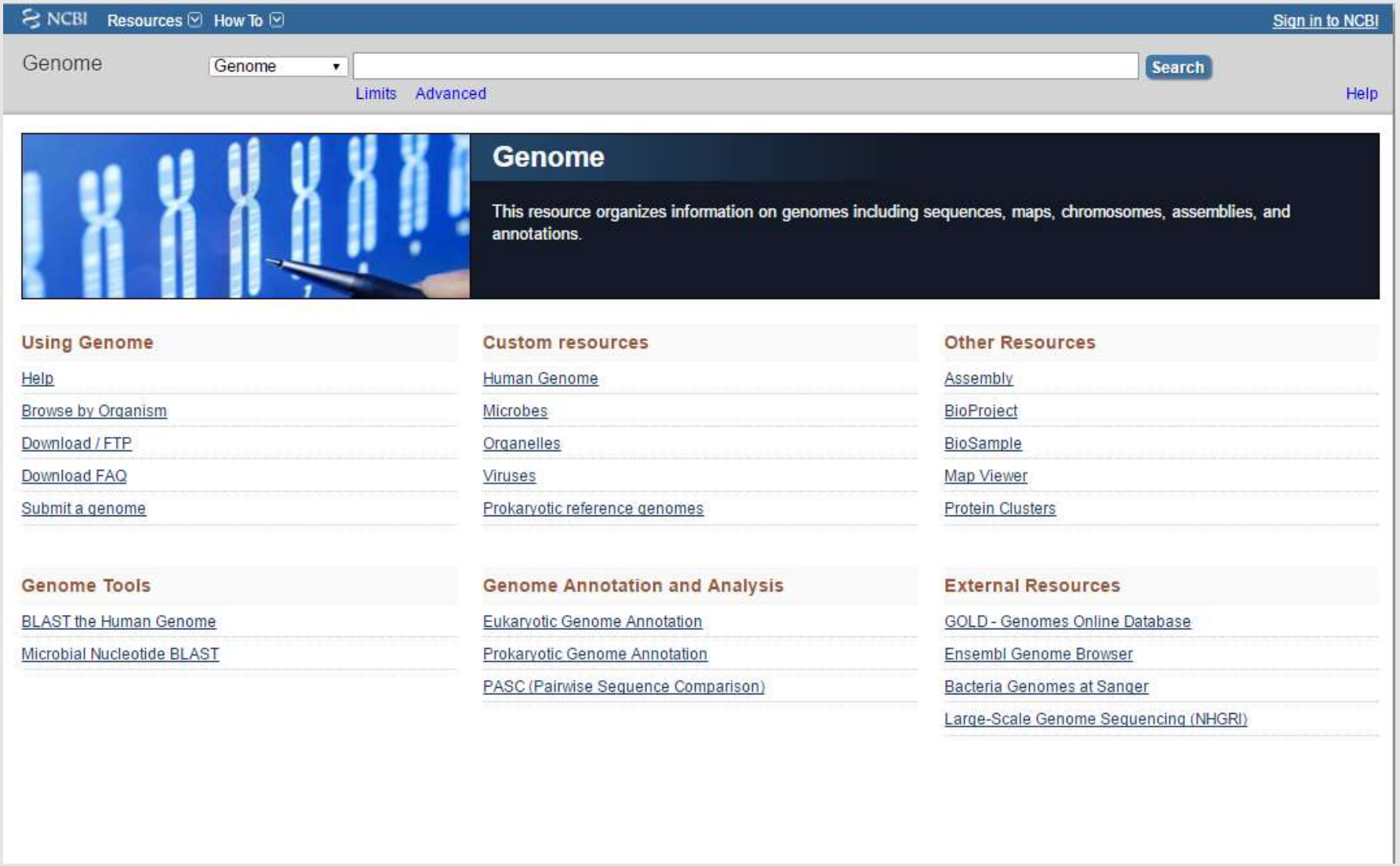


*Le 16 Janvier 2018, PDB archivait près de 141 mille structures 3D, obtenues essentiellement par expérimentation (cristallographie à rayon X, spectroscopie RMN, Cryo-microscopie électronique, etc.).

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

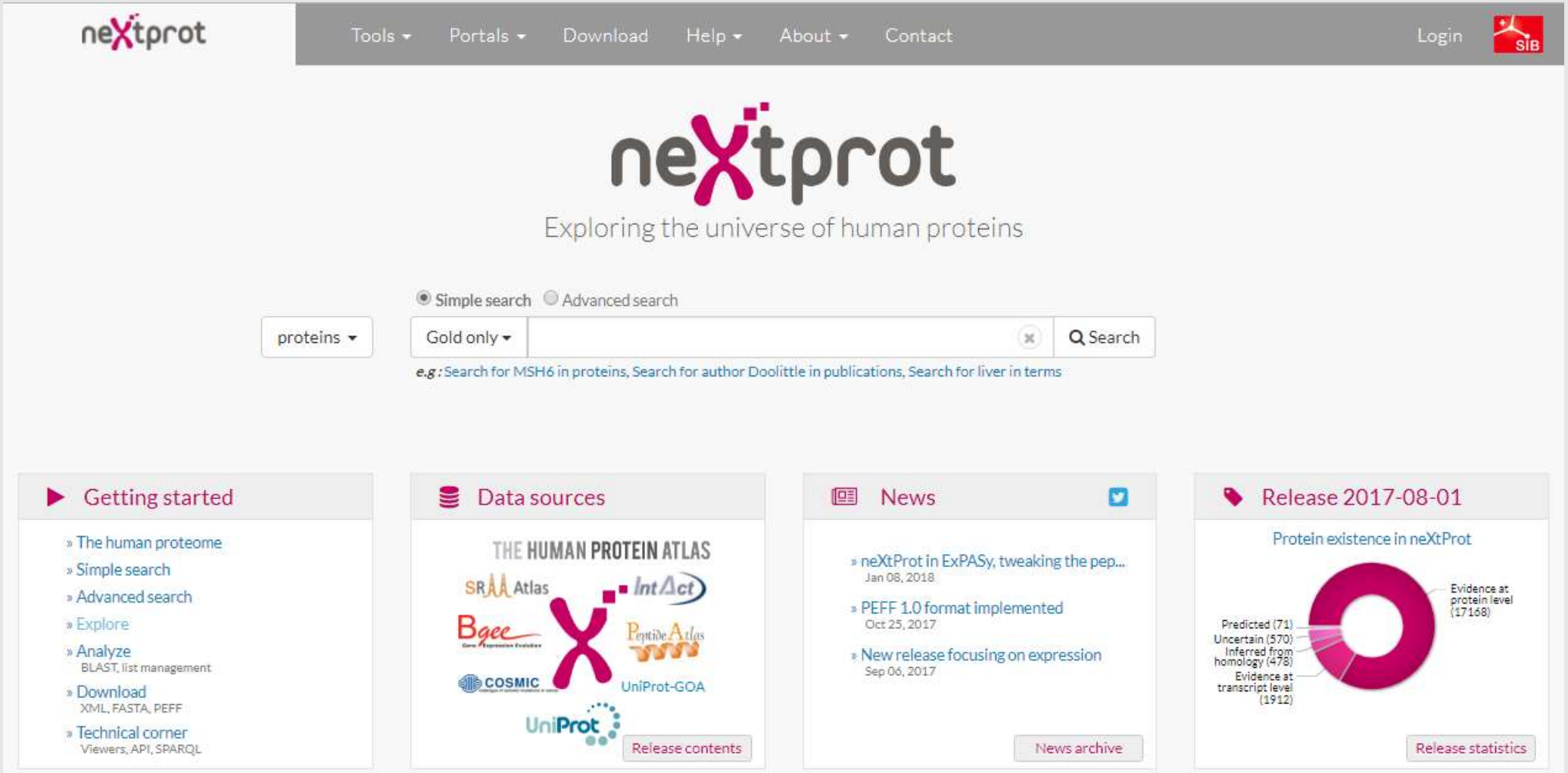


2. Les banques de données spécialisées

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.

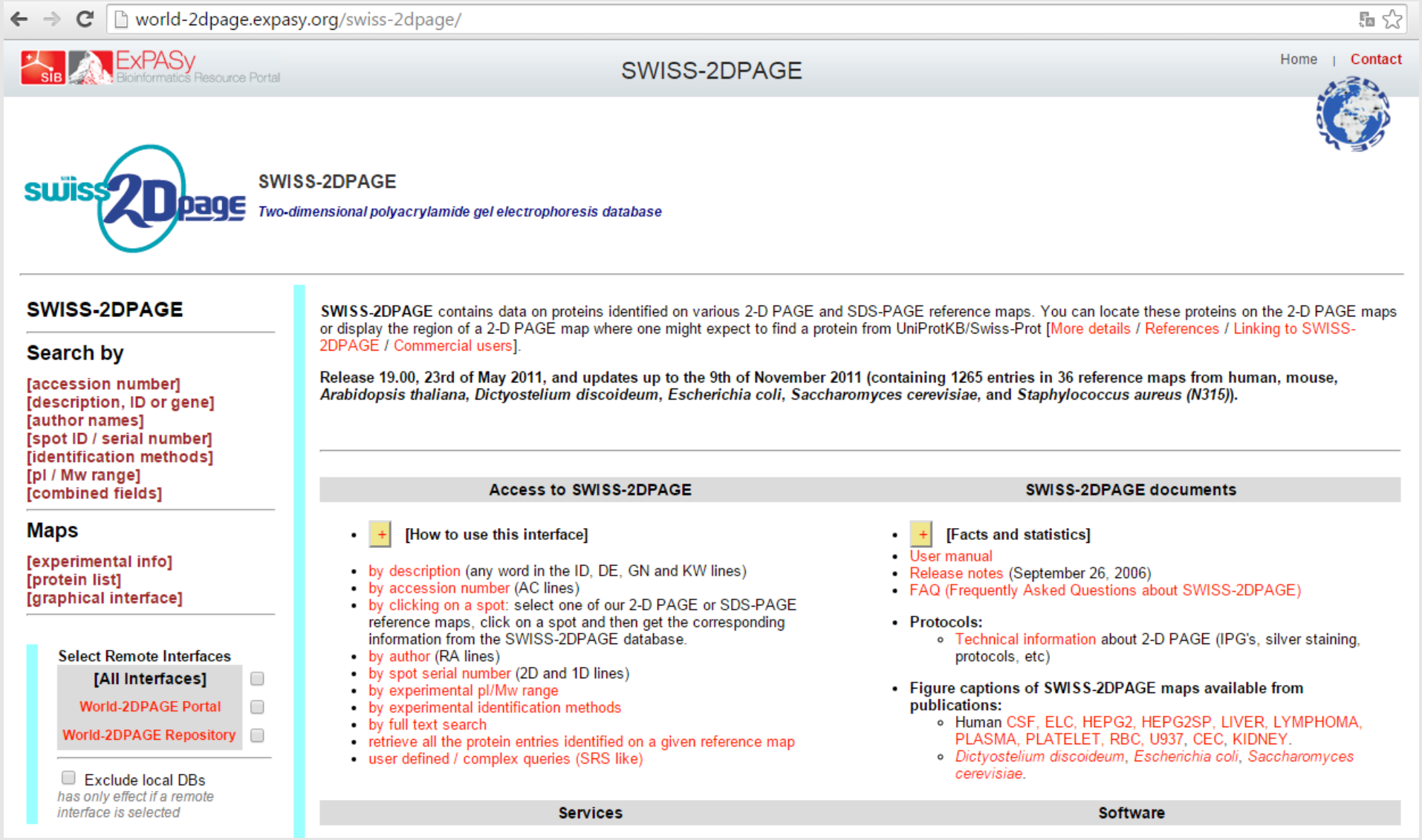
neXtprot :
Banque de protéines humaines
https://www.nextprot.org/



SWISS-2DPAGE :

Base de données de protéines identifiées par électrophorèse bidimentionnelle

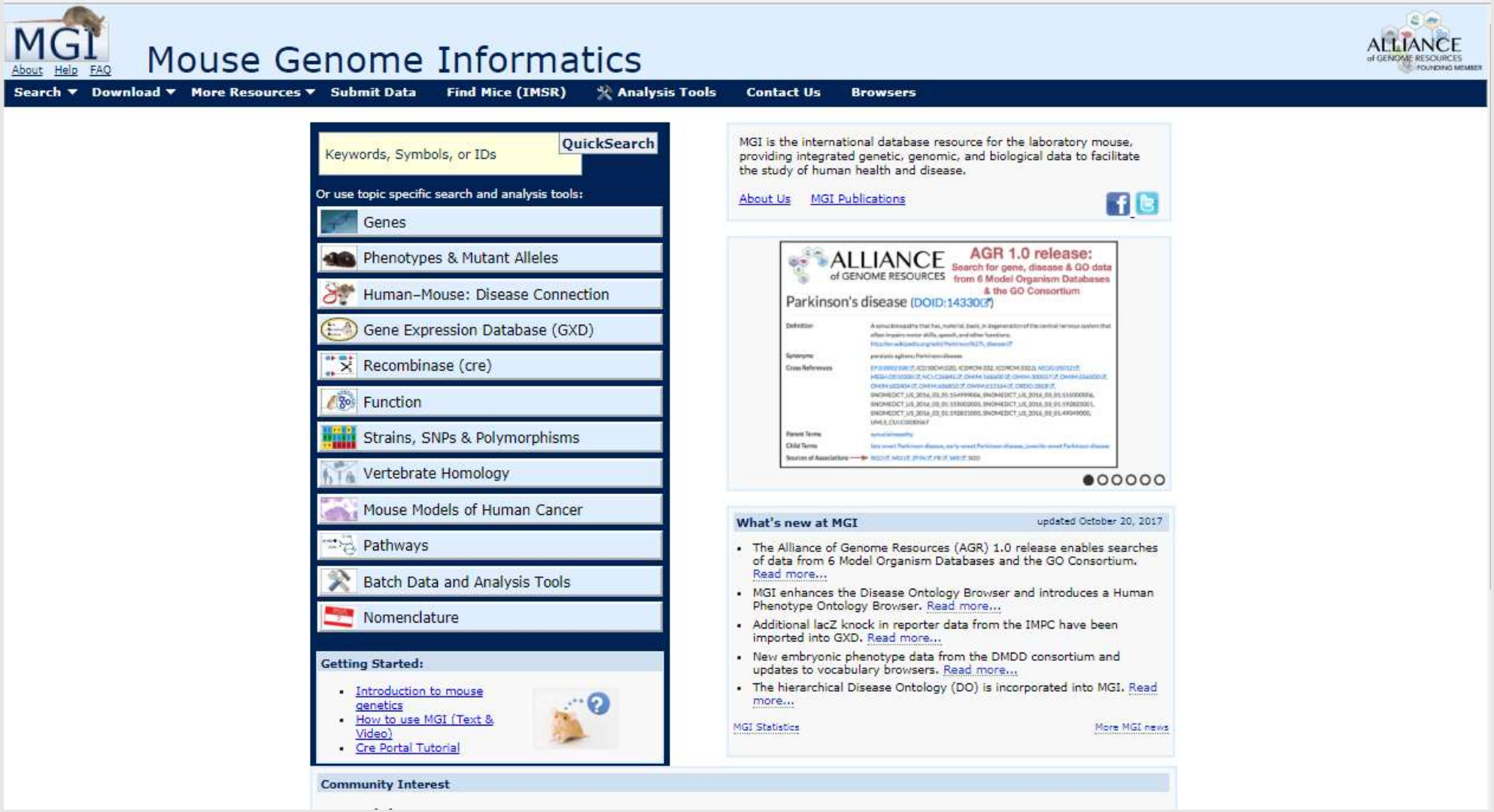
http://world-2dpage.expasy.org/swiss-2dpage/



MGI (*Mouse Gene Informatics*) :

banque de séquences nucléotidiques de la souris

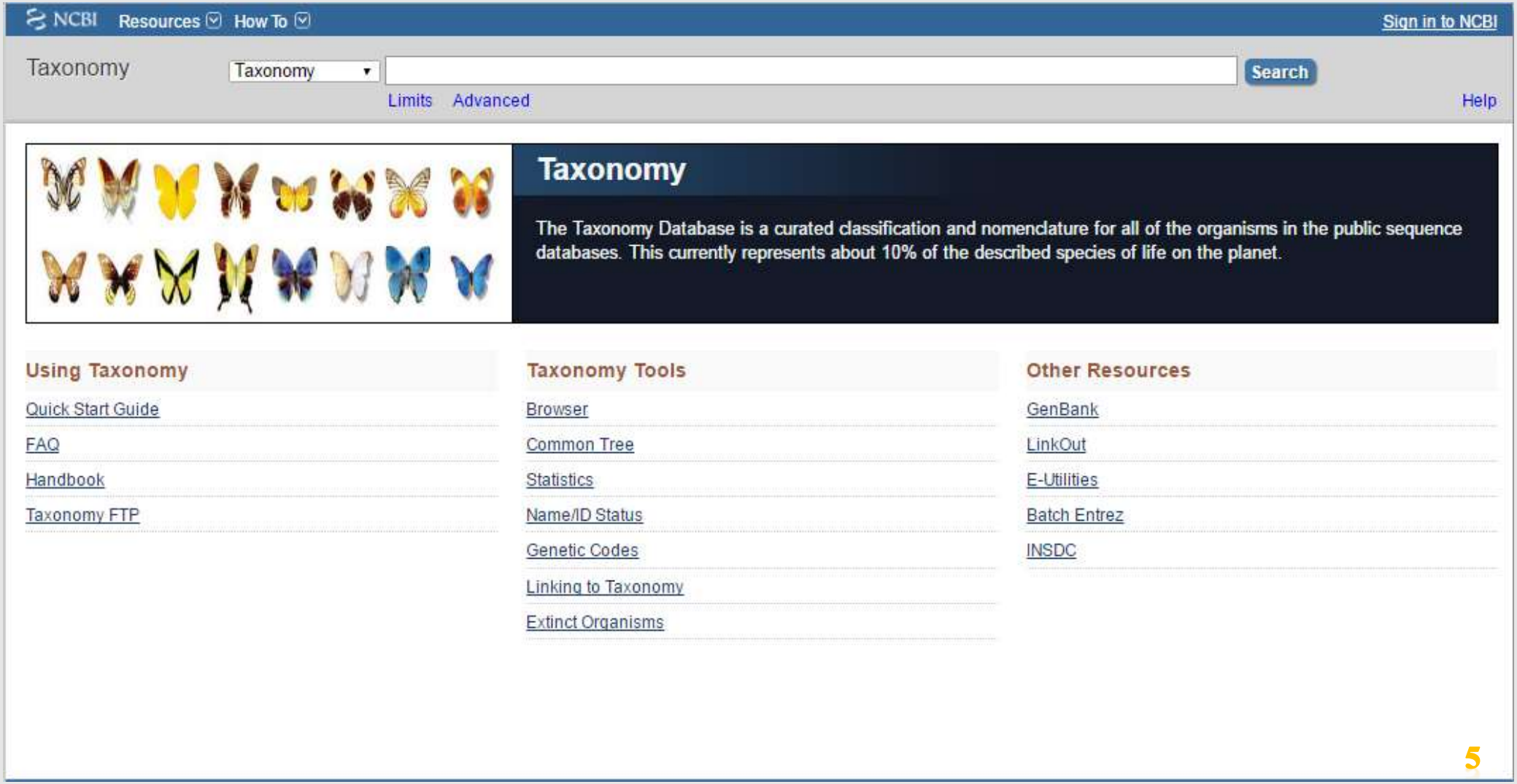
http://www.informatics.jax.org/



Taxonomy

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

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

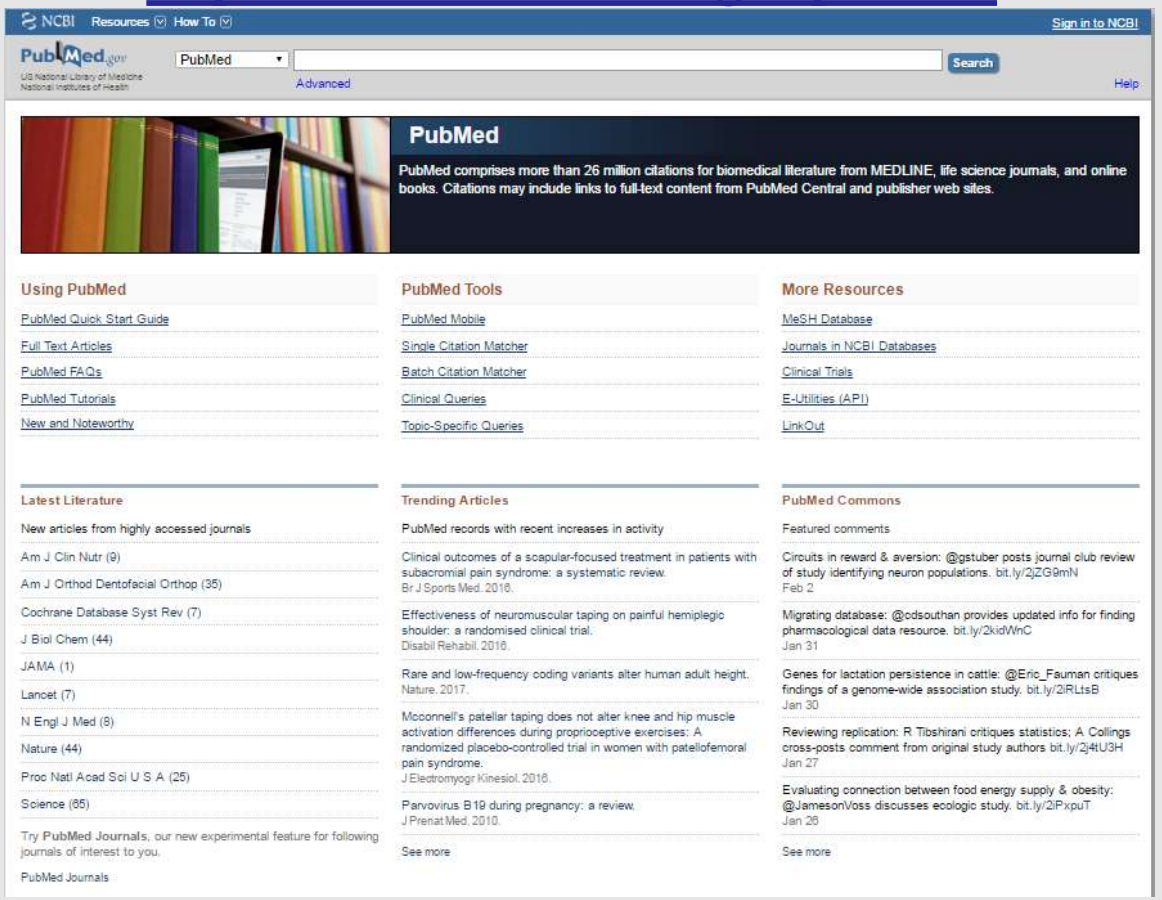


BANQUES DE DONNÉES GÉNÉRALISTES

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



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, DPlnteract, 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, HlVdb, 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, 0-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

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 ccgaggtcat cctgctagac tcagacctgg atgaacccat agacttgcgc 60

tcggtcaaga gccgcagcga ggccggggag ccgcccagct ccctccaggt gaagcccgag 120

Acaccggcgt Cggcggcggt Ggcggtggcg Gcggcagcgg Cacccaccac 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;
            Saccharomyces.
REFERENCE   1  (bases 1 to 1599)
  AUTHORS   Wang,S.A. and Li,Y.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             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="MLLQAFLFLLAGFAAKISASMTNETSDRPLVHFTPNKGWMNDPN
                     GLWYDEKDAKWHLYFQYNPNDTVWGTPLFWGHATSDDLTNWEDEPIAIAPKRNDSGAF
                     SGSMVVDYNNTSGFFNDTIDPRQRCVAIWTYNTPESEEQYISYSLDGGYTFTEYQKNP
                     VLAANSTQFRDPKVFWYEPSQKWIMTAAKSQDYKIEIYSSDDLKSWKLESAFANEGFL
                     GYQYECPGLIEVPTEQDPSKSYWVMFISINPGAPAGGSFNQYFVGSFNGTHFEAFDNQ
                     SRVVDFGKDYYALQTFFNTDPTYGSALGIAWASNWEYSAFVPTNPWRSSMSLVRKFSL
                     NTEYQANPETELINLKAEPILNISNAGPWSRFATNTTLTKANSYNVDLSNSTGTLEFE
                     LVYAVNTTQTISKSVFFDLSLWFKGLEDPEEYLSMGFEASASSFFLDRGNSKVKFVKE
                     NPYFTNRMSVNNQPFKSENDLSYYKVYGLLDQNILELYFNDGDVVSTNTYFMTTGNAL
                     GSVNMTTGVDNLFYIDKFQVREVK"
ORIGIN
        1 atgcttttgc aagctttcct tttcctttg gctggttttg cagccaaaat atctgcatca
       61 atgacaaacg aaactagcga tagacctttg gtccacttca cacccaacaa gggctggatg
      121 aatgacccaa atgggttgtg gtacgatgaa aaagatgcca aatggcatct gtactttcaa
      181 tacaacccaa atgacaccgt atgggtacg ccattgtttt ggggccatgc tacttccgat
      241 gatttgactc attgggaaga tgaacccatt gctatcgctc ccaagcgtaa cgattcaggt
      301 gctttctctg gctccatggt ggttgattac aacaacacga gtgggttttt caatgatact
      361 attgatccaa gacaaagatg cgttgcgatt tggacttata acactcctga aagtgaagag
      421 caatacatta gctattctct tgatggtggt tacacttttа ctgaatacca aaagaaccct
      481 gttttagctg ccaactccac tcaattcaga gatccaaagg tgttctggta tgaaccttct
      541 caaaaatgga ttatgacggc tgccaaatca caagactaca aaattgaaat ttactcctct
      601 gatgacttga agtcctggaa gctagaatct gcatttgcta atgaaggttt cttaggctac
      661 caatatgaat gtccaggttt gattgaagtc ccaactgagc aagatccttc caaatcctat
      721 tgggtcatgt ttatttctat caatccaggt gccctgctg gcggttcctt ccaacaatat
      781 tttgttggat ccttcaatgg tactcatttt gaagcgtttg acaatcaatc tagagtggta
      841 gattttggta aggactacta tgccttgcaa actttcttca acacagaccc aacgtacggt
      901 tcagcattag gtattgcctg ggcttcaaac tgggagtaca gtgcctttgt cccaactaac
      961 ccatggagat catccatgtc tttggtccgc aagttttctt tgaacactga atatcaagct
     1021 aatccagaga ctgaattgat caatttgaaa gccgaaccaa tattgaacat tagtaatgct
     1081 ggtcctggt ctcgttttgc tactaacaca actctaacta aggccaattc ttacaatgtc
     1141 gatttgagca actcgactgg taccctagag tttgagttgg tttacgctgt taacaccaca
     1201 caaaccatat ccaaatccgt ctttccgtac ttatcacttt ggttcaaggg tttagaagat
     1261 cctgaagaat atttaagaat gggttttgaa gccagtgctt cttccttctt tttggaccgt
     1321 ggtaactcta aggtcaagtt tgtcaaggag aacccatatt tcacaaacag aatgtctgtc
     1381 aacaaccaac cattcaagtc tgagaacgac ctaagttact ataaagtgta cggcctactg
     1441 gatcaaaaca tcttggaatt gtacttcaac gatggagatg tggtttctac aaatacctac
     1501 ttcatgacca ccgttaacgc tctaggatct gtgaacatga ccactggtgt cgataatttg
     1561 ttctacattg acaagttcca agtaagggaa gtaaaatag
//
```

Description générale de la séquence

« Features »

Description des objets biologiques présents sur la séquence

La séquence

- **Chaque ligne commence par un mot-clé**
  - **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="MLLQAFLFLLAGFAAKISASMTNETSDRPLVHFTPNKGWMNDPN
                     GLWYDEKDAKWHLYFQYNPNDTVWGTPLFWGHATSDDLTHWEDEPIAIAPKRNDSGAF
                     SGSMVVDYNNTSGFFNDTIDPRQRCVAIWTYNTPESEEQYISYSLDGGYTFTEYQKNP
                     VLAANSTQFRDPKVFWYEPSQKWIMTAAKSQDYKIEIYSSDDLKSWKLESAFANEGFL
                     GYQYECPGLIEVPTEQDPSKSYWVMFISINPGAPAGGSFNQYFVGSFNGTHFEAFDNQ
                     SRVVDFGKDYYALQTFFNTDPTYGSALGIAWASNWEYSAFVPTNPWRSSMSLVRKFSL
                     NTEYQANPETELINLKAEPILNISNAGPWSRFATNTTLTKANSYNVDLSNSTGTLEFE
                     LVYAVNTTQTISKSVFPDLSLWFKGLEDPEEYLRMGFEASASSFFLDRGNSKVKFVKE
                     NPYFTNRMSVNNQPFKSENDLSYYKVYGLLDQNILELYFNDGDVVSTNTYFMTTGNAL
                     GSVNMTTGVDNLFYIDKFQVREVK"
```

La séquence (format Genbank)

```
ORIGIN
        1 atgcttttgc aagctttcct ttccttttg gctggttttg cagccaaaat atctgcatca
       61 atgacaaacg aaactagcga tagacctttg gtccacttca cacccaacaa gggctggatg
      121 aatgacccaa atgggttgtg gtacgatgaa aaagatgcca aatggcatct gtactttcaa
      181 tacaacccaa atgacaccgt atggggtacg ccattgtttt ggggccatgc tacttccgat
      241 gatttgactc attgggaaga tgaacccatt gctatcgctc ccaagcgtaa cgattcaggt
      301 gctttctctg gctccatggt ggttgattac aacaacacga gtgggttttt caatgatact
      361 attgatccaa gacaaagatg cgttgcgatt tggacttata acactcctga aagtgaagag
      421 caatacatta gctattctct tgatggtggt tacactttta ctgaatacca aaagaaccct
      481 gttttagctg ccaactccac tcaattcaga gatccaaagg tgttctggta tgaaccttct
      541 caaaaatgga ttatgacggc tgccaaatca caagactaca aaattgaaat ttactcctct
      601 gatgacttga agtcctggaa gctagaatct gcatttgcta atgaaggttt cttaggctac
      661 caatatgaat gtccaggttt gattgaagtc ccaactgagc aagatccttc caaatcctat
      721 tgggtcatgt ttatttctat caatccaggt gcacctgctg gcggttcctt caaccaatat
      781 tttgttggat ccttcaatgg tactcatttt gaagcgtttg acaatcaatc tagagtggta
      841 gattttggta aggactacta tgccttgcaa actttcttca acacagaccc aacgtacggt
      901 tcagcattag gtattgcctg ggcttcaaac tgggagtaca gtgcctttgt cccaactaac
      961 ccatggagat catccatgtc tttggtccgc aagttttctt tgaacactga atatcaagct
     1021 aatccagaga ctgaattgat caatttgaaa gccgaaccaa tattgaacat tagtaatgct
     1081 ggtccctggt ctcgttttgc tactaacaca actctaacta aggccaattc ttacaatgtc
     1141 gatttgagca actcgactgg taccctagag tttgagttgg tttacgctgt taacaccaca
     1201 caaaccatat ccaaatccgt ctttcccgac ttatcacttt ggttcaaggg tttagaagat
     1261 cctgaagaat atttaagaat gggttttgaa gccagtgctt cttccttctt tttggaccgt
     1321 ggtaactcta aggtcaagtt tgtcaaggag aacccatatt tcacaaacag aatgtctgtc
     1381 aacaaccaac cattcaagtc tgagaacgac ctaagttact ataaagtgta cggcctactg
     1441 gatcaaaaca tcttggaatt gtacttcaac gatggagatg tggtttctac aaatacctac
     1501 ttcatgacca ccggtaacgc tctaggatct gtgaacatga ccactggtgt cgataatttg
     1561 ttctacattg acaagttcca agtaagggaa gtaaaatag
//
```

Quelques formats de données biologiques

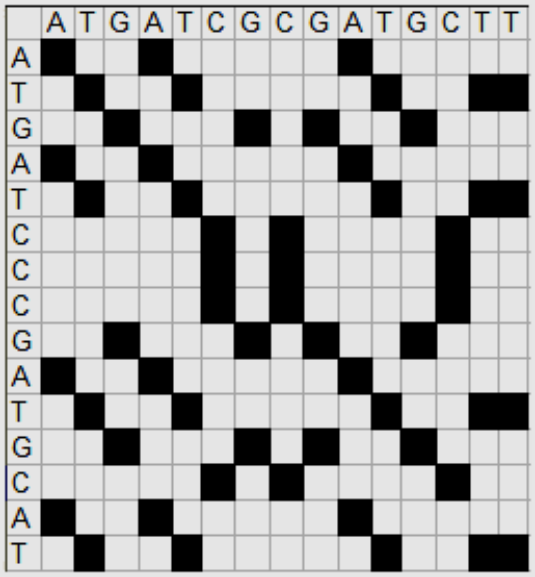
- ✓ Format des banques, exemples :
  - Séquences ADN/ARN : EMBL, GenBank et DDBJ
  - Séquences protéiques : Uniprot, SwissProt et TrEMBL, PIR…
  - Format PHYLYP (*PHYLogeny Inference Package*), FOSN (*Files Of Sequence Names*), RSF (*Rich Sequence Format files*), RSF (*Rich Sequence Format files*), MSF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

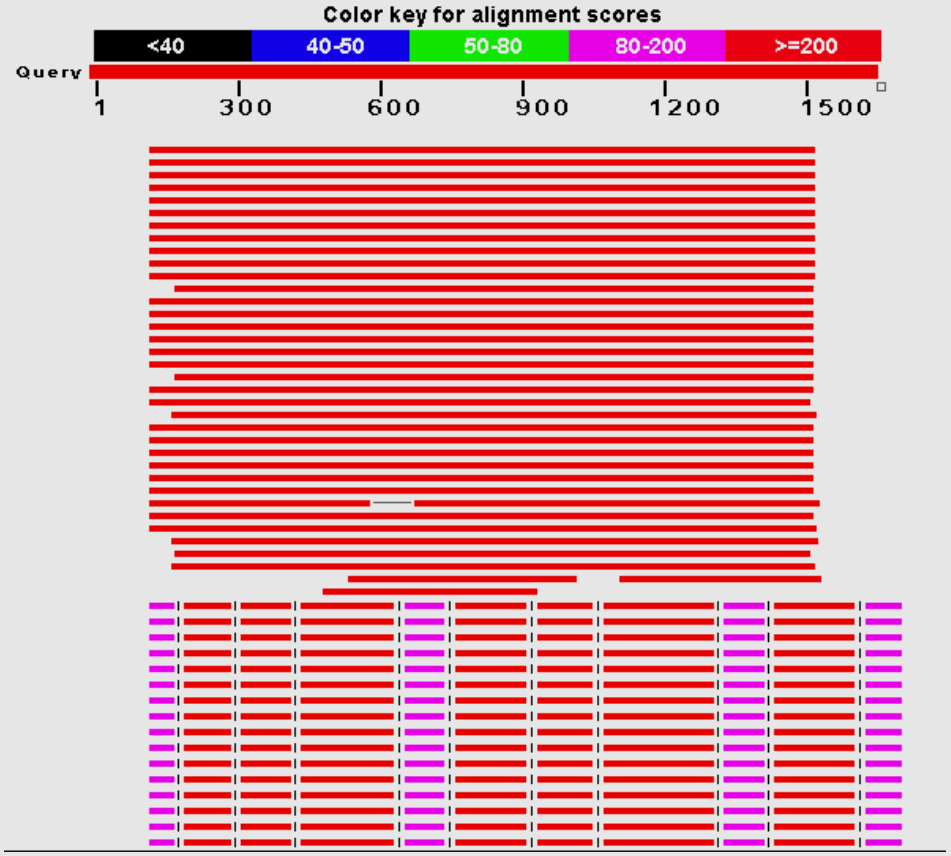
Le format FASTA

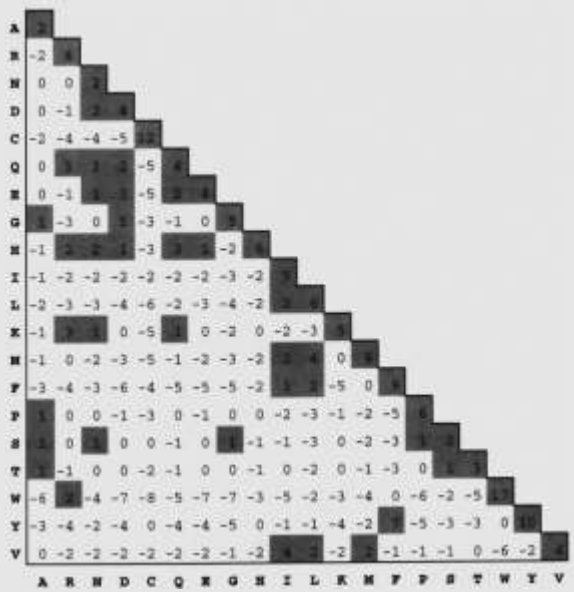
- ✓ Une ligne de commentaires précédé de « > »
- ✓ La séquence brute (pas d'espace, ni de nombre)

```
>Human Polycomb 2 homolog (hPc2) mRNA, partial cds

ctccggcagcccgaggtcatcctgctagactcagacctggatgaacccat

agacttgcgctcggtcaagagccgcagcgaggccggggagccgcccagct

ccctccaggtgaagcccgagacaccggcgtcggcggcggtggcggtggcg

gcggcagcggcacccaccacgacggcggagaagcctccagccgaggccca

ggacgaacctgcagagtcgctgagcgagttcaagcccttctttgggaata

taattatcaccgacgtcaccgcgaactgcctcaccgttactttcaaggag

tacgtgacggtg
```

LES ALIGNEMENTS



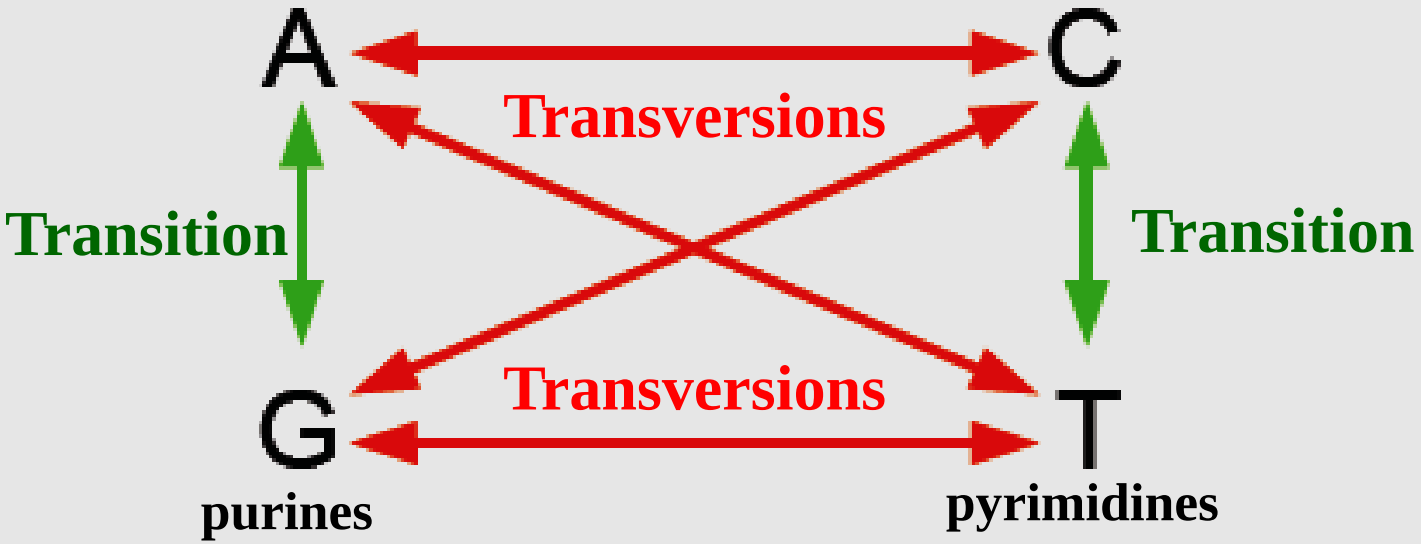




ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

- L'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leurs séquences
- Evolution des gènes par mutations, deux types « locaux » sont intéréssants:
  - ✓ 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.

- ✓ 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 :

$$d(A, B) = \sharp substitutions + \sharp indels$$

LE SCORE DE SIMILARITÉ

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

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

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

- 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).

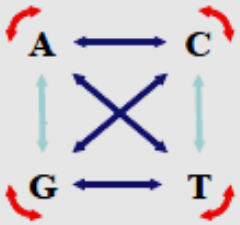
Matrices de scores pour l'ADN

- 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 :

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

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

S1 = TAAGTCCG m=8 et S2 = TAAGTACG n=8

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

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) :

(matrice d'identité)

	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**

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								

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 :

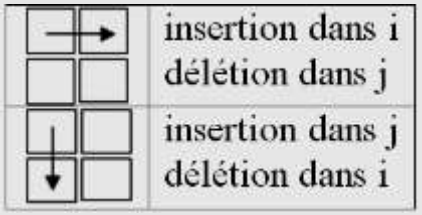
$$S(i,j) = Max \begin{cases} S(i-1,j-1) + s(i,j) \\ S(i-1,j) \\ S(i,j-1) \end{cases}$$

		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

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**



ETAPE 4: ALIGNEMENT DES DEUX SÉQUENCES ET CALCUL DE SCORE

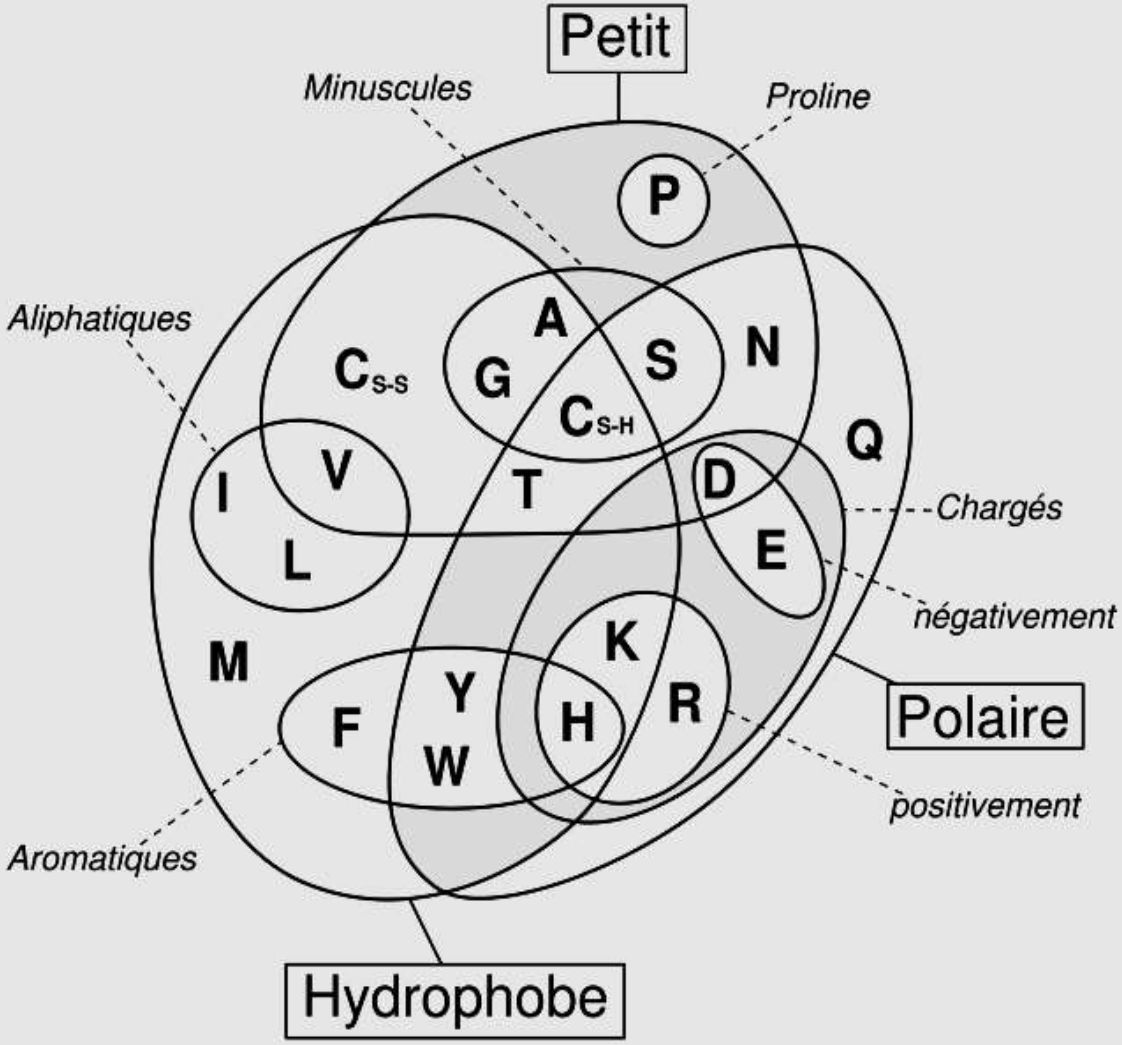
Séquence S1	**T**	**A**	**A**	**G**	**T**	**—**	**C**	**C**	**G**
	|	|	|	|	|	*	|	*	|
Séquence S2	**T**	**A**	**A**	**G**	**T**	**A**	**C**	**—**	**G**

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

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

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**).

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 :

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

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

avec :

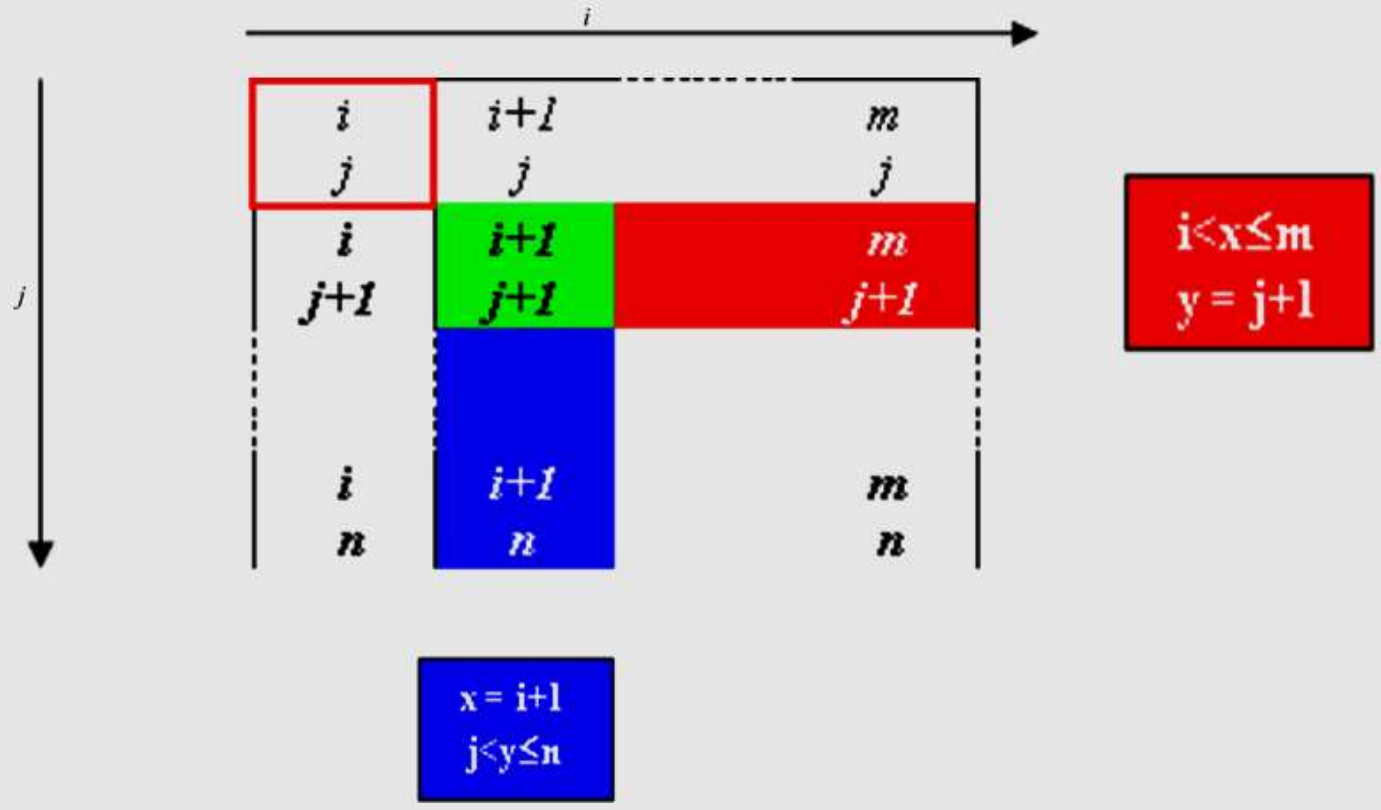
$i < x \leq m$ et $y = j+1$

ou

$x = i+1$ 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

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

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

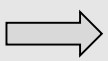
ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

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

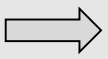
	V	T	E	E	R	D	A	F
L								2
T								-2
S								-3
H								-2
E								-5
A						2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

....



	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

....



ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

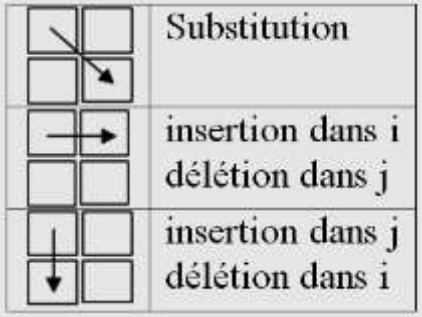
On applique l'algorithme de Needleman et Wunch:

	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

ETAPE 3: PARCOURS DE LA MATRICE TRANSFORMÉE

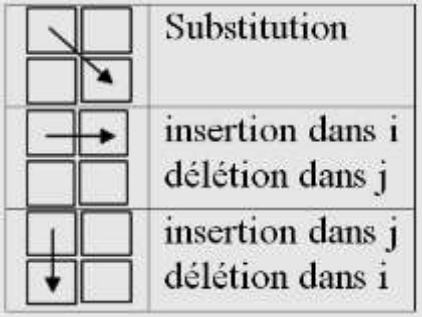
✓ Le parcours s'effectue du plus haut score vers le plus petit. Si les trois cases ont des valeurs de scores égales, alors le chemin vers la diagonale est favorisé :

	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**



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



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

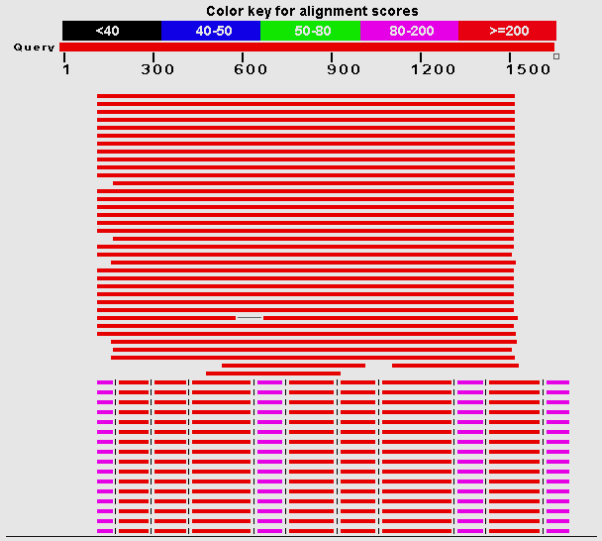
On peut supposer que la valine a été substituée en leucine dans la 2ème séquence (ou *vis versa*) par besoin d'adaptation de l'organisme à partir du quel a été isolée cette séquence. Le même raisonnement concernera les substitutions E-H et F-L.

- ✓ 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.**

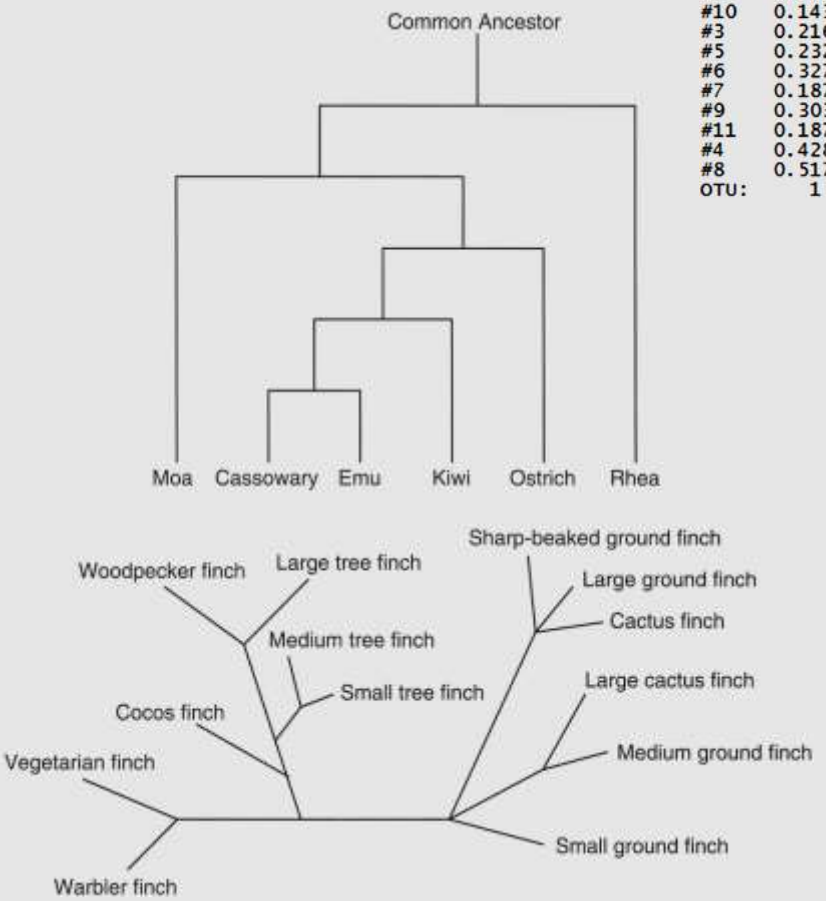
ALIGNEMENT D'UNE SÉQUENCE AVEC UNE BANQUE: EXEMPLE DE L'OUTIL BLAST (*BASIC LOCAL ALIGNMENT SEARCH TOOL*) (Altschul *et al.*, 1990

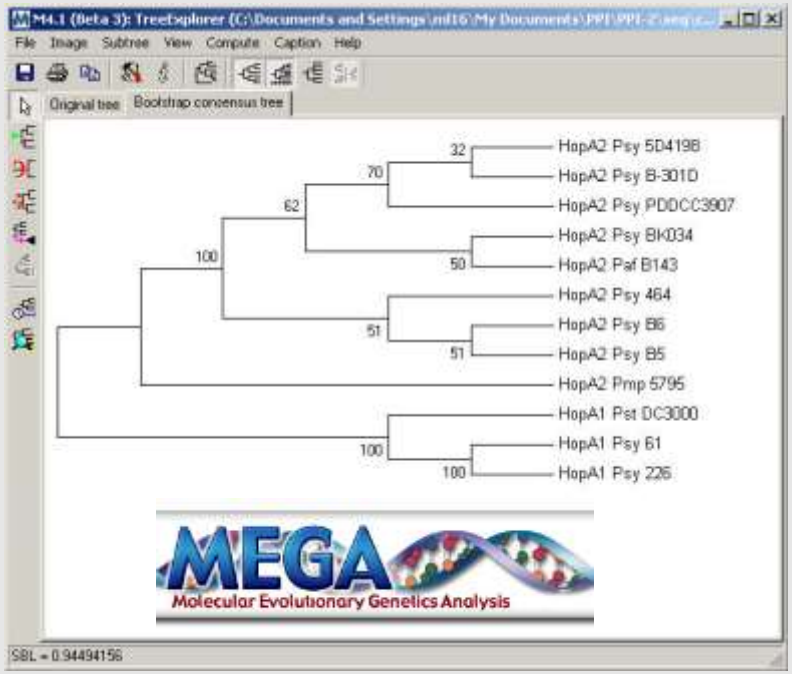
- ✓ Outil utilisant une heuristique d'alignement local ;
- ✓ recherche dans une base de données de séquence des segments qui sont localement similaires à une séquence-requête fournie par l'utilisateur (*query sequence*) ;
- ✓ Le programme délivre des résultats similaires à la séquence requête, ces résultats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

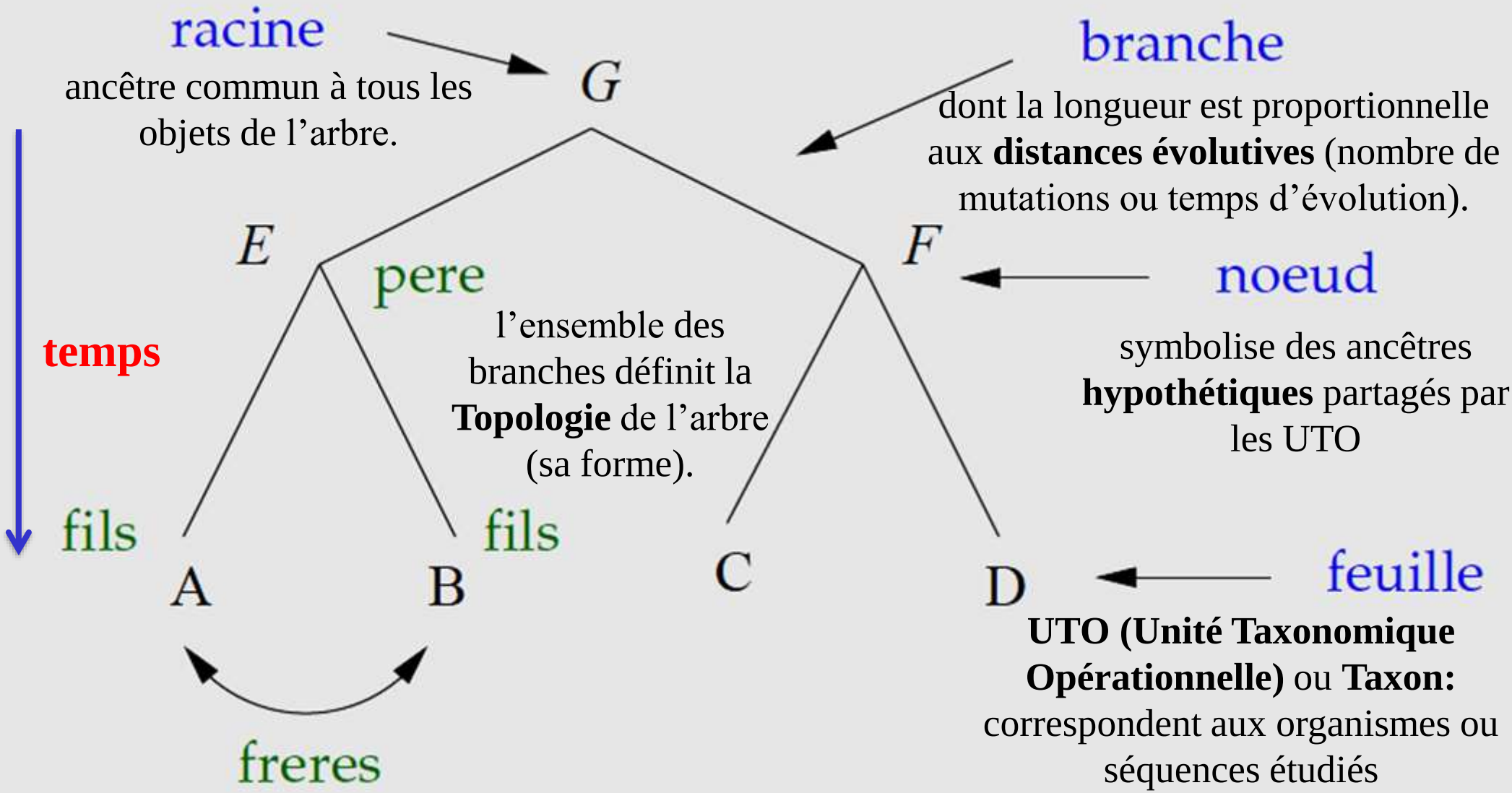




Phylogénie: Du grec ancien *phŷlon* (« tribu, race ») et *géneia* « qui engendre»

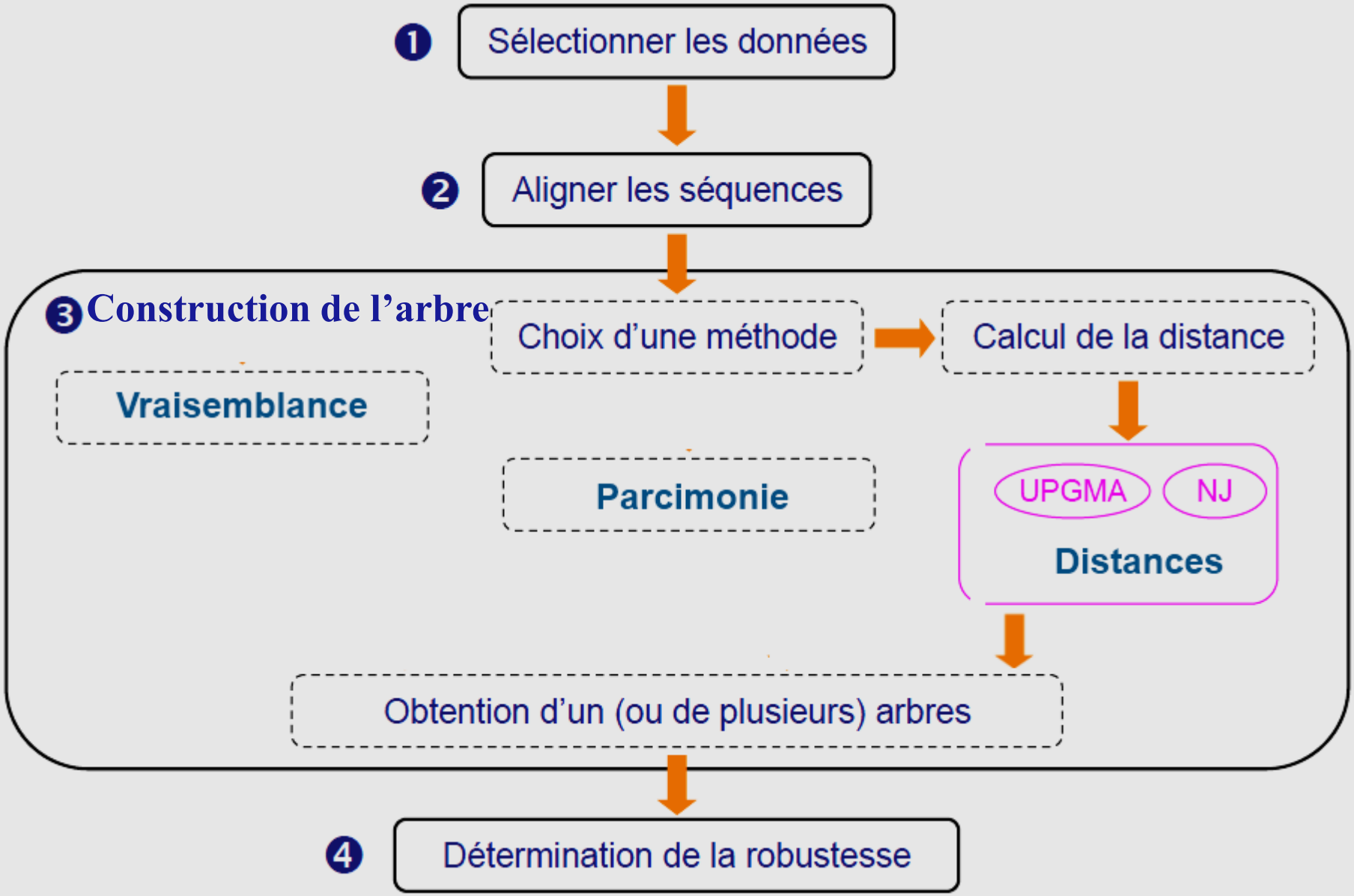
- La **phylogénie** est l'étude des relations de parentés entre différents êtres vivants en vue de comprendre l'évolution de ces organismes.
- La phylogénie trouve ses applications dans plusieurs domaines : systématique, génétique des populations, écologie, épidémiologie, phylogéographie, etc.
- Les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

LA STRUCTURE D'UN ARBRE PHYLOGÉNÉTIQUE



Le temps s'écoule de la racine vers les feuilles

ÉTAPES DE CONSTRUCTION D'UN ARBRE PHYLOGÉNÉTIQUE



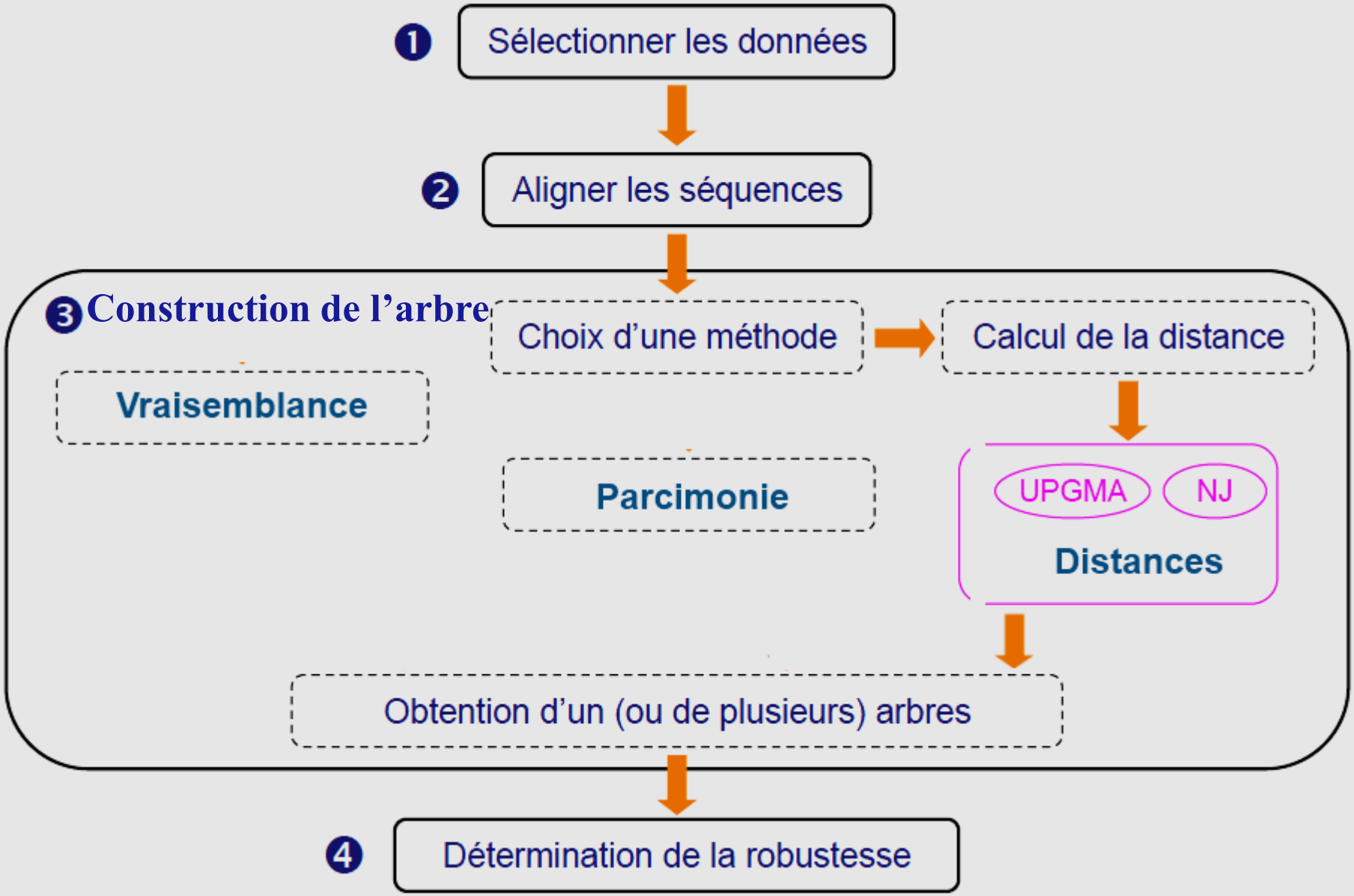
NEIGHBOR-JOINING (Saitou et Nei, 1987):

- ✓ 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

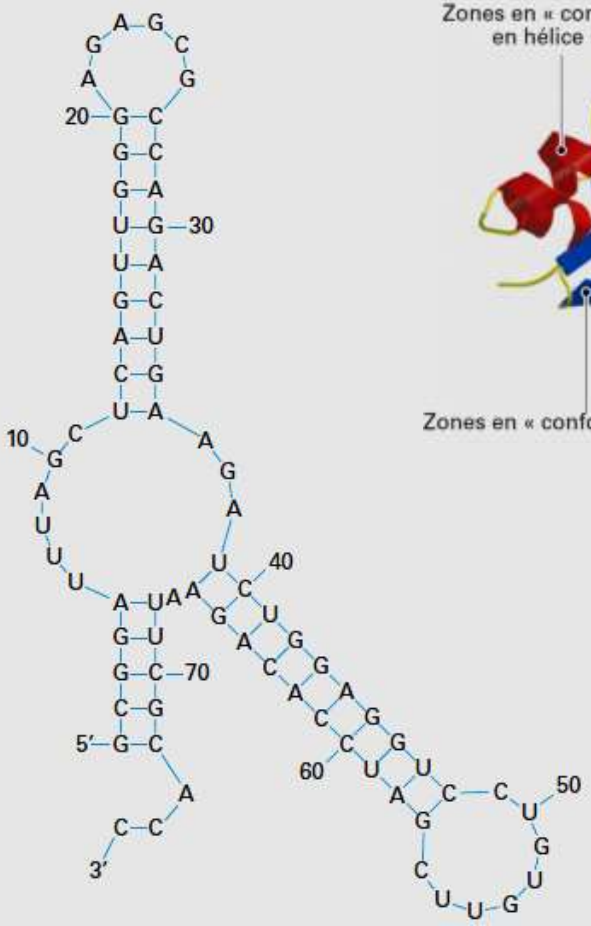


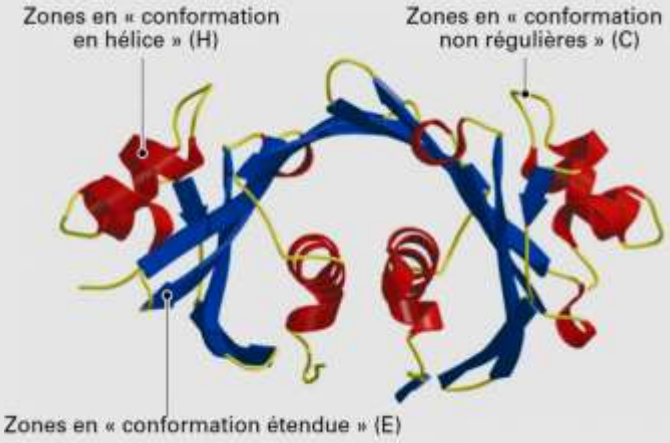
FIABILITÉ D'UN ARBRE

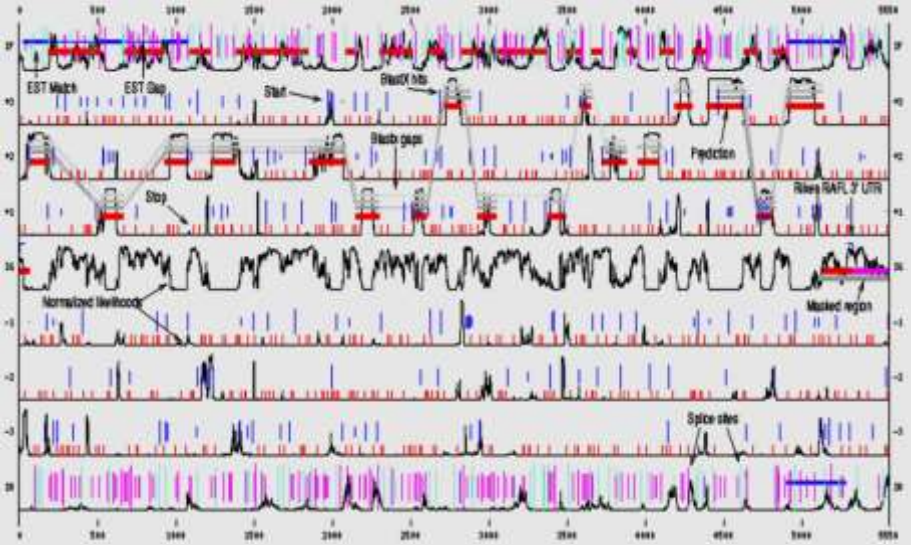
- ✓ La construction des arbres phylogénétiques est basée sur des hypothèses. Cela induit des erreurs au niveau topologique et conduit à des erreurs d'interprétation qui nécessitent d'être rectifiées.
- ✓ La méthode du **bootstrap** est la plus utilisée pour vérifier la **fiabilité** (**robustesse**)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

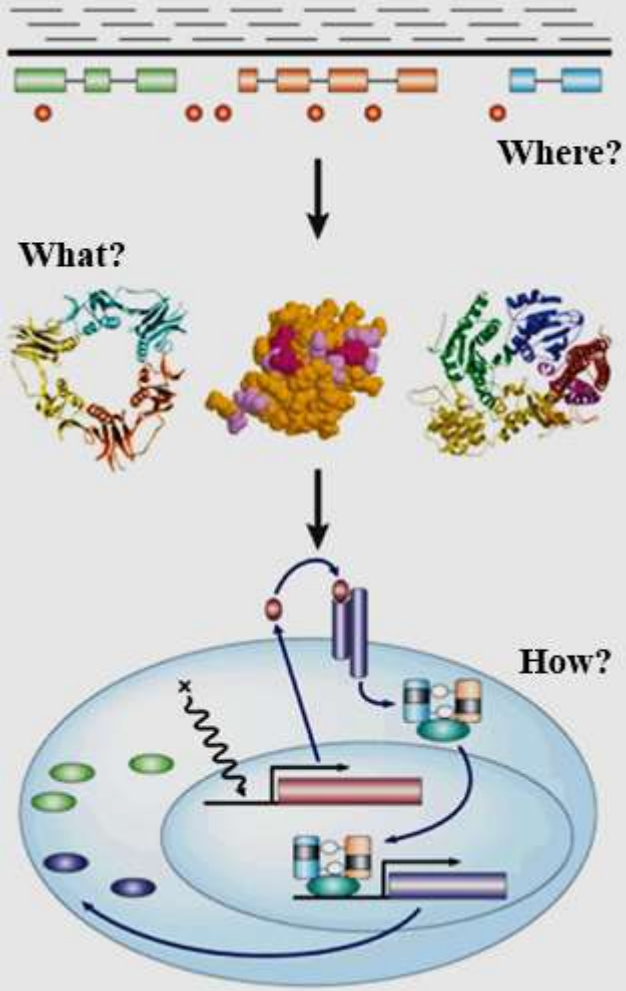
- ❑ éléments répétés (éléments transposables, satellites, etc.)
- ❑ gènes des ARN stables (ARNr, ARNt...)
- ❑ gènes protéiques
- ❑ régions régulatrices, etc.

✓ **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 Readin 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.

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 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**