INTRODUCTION A LA BIOINFORMATIQUE

Silvina GONZALEZ-RIZZO
silvina.gonzalez-rizzo@univ-antilles.fr

Bioinformatique pour des biologistes

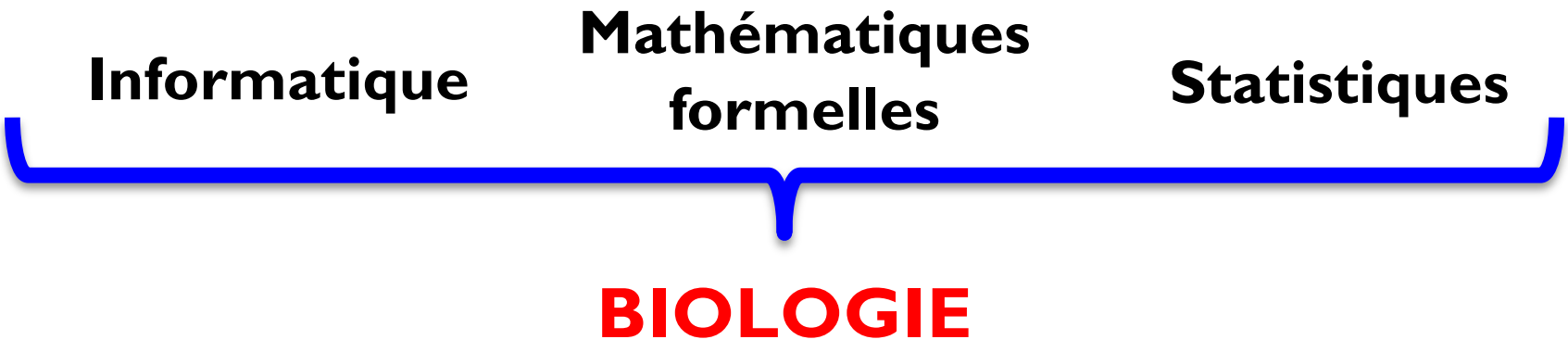
- Objectif du cours:
  - Vous montrer les taches courantes de la bioinformatique qu'un biologiste/biochimiste doit savoir traiter par lui-même sans avoir recours au spécialiste a fin de répondre à des questions usuelles comme:

1. Comment extraire des informations pertinentes dans une base de données biologiques?
2. Est-ce que ce gène appartient à une famille connue?
3. Comment designer des amorces pour amplifier un gène spécifique?
4. Existe-t-il d'autres gènes homologues??
5. ETC……

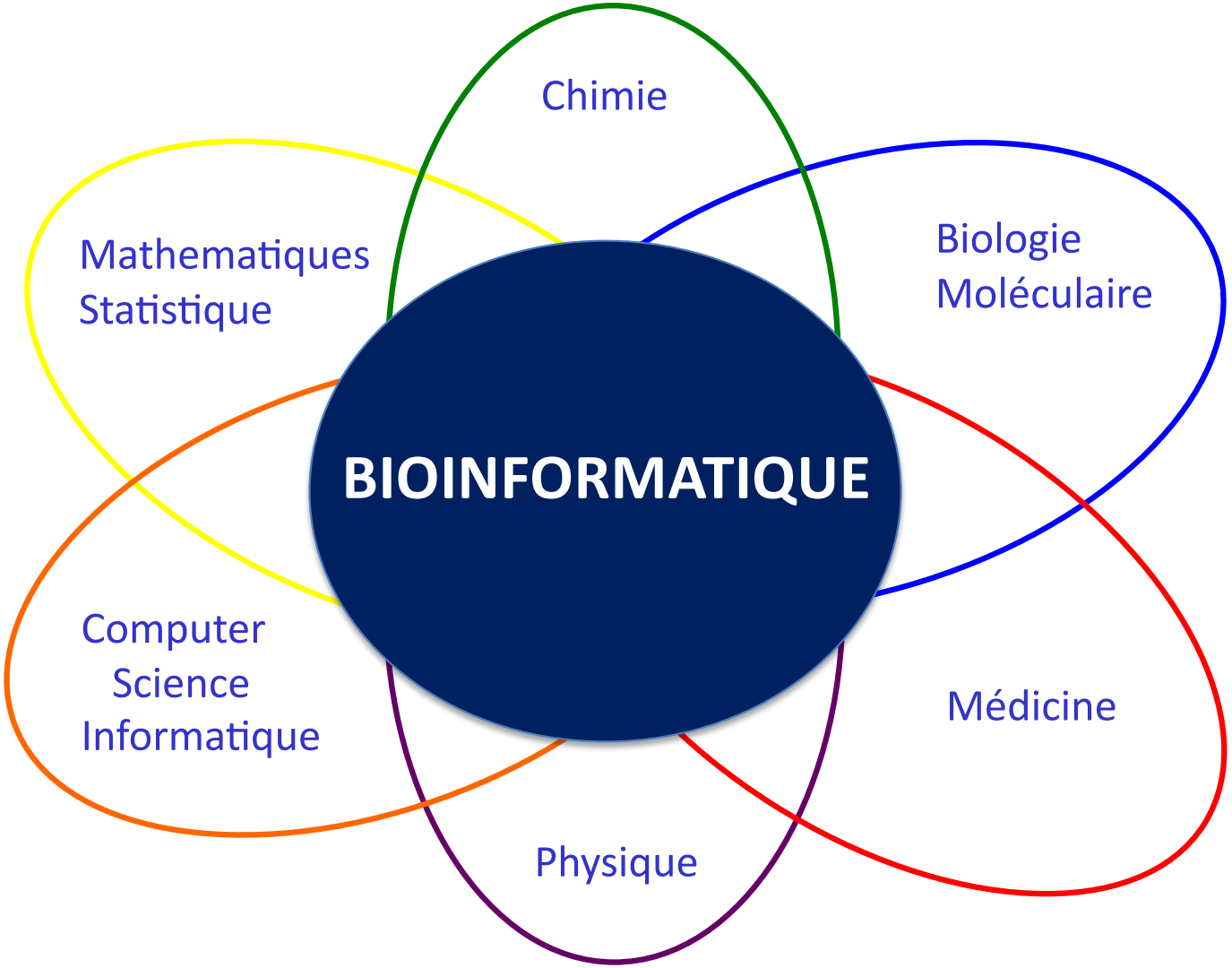
Bioinformatique?

- Ensemble de méthodes, de logiciels et d'applications en ligne qui permettent de gérer, manipuler, et analyser des données biologiques.

- La bioinformatique met en jeu plusieurs champs disciplinaires :



La bioinformatique est interdisciplinaire



Bioinformatique

- Le terme bioinformatique peut décrire toutes les applications informatiques résultant de ces recherches dans le but de résoudre un problème scientifique posé par la biologie.

- Cela va de l'analyse du génome à la modélisation de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'image, le séquençage du génome, la reconstruction d'arbres phylogénétiques (phylogénie), etc.

Bioinformatique

Comme le décrit très bien **Jean-Michel Claverie** :

"La bioinformatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation de l'information génétique (séquences) et structurale (repliement 3-D).

*C'est le décryptage de la « **bioinformation** » (« Computational Biology »). La bioinformatique est donc une branche théorique de la Biologie. Son but, comme tout volet théorique d'une discipline, est d'effectuer la **synthèse des données disponibles** (à l'aide de modèles et de théories), **d'énoncer des hypothèses généralisatrices** (p.e: comment les protéines se replient ou comment les espèces évoluent), et de formuler des prédictions (p.e : localiser ou prédire la fonction d'un gène)".*

Bioinformatique?

- **ADN (Génome)**
  - Séquences de nucléotides
  - Séquences de gènes
  - Banques de données
- **ARN (Transcriptome)**
  - Séquence
  - Structure
- **Protéines (Protéome)**
  - Séquence
  - Structure
  - Réseaux d'intéraction

Bioinformatique

On peut aussi appeler cette discipline:

- BIOLOGIE *in silico*

par analogie avec les termes:

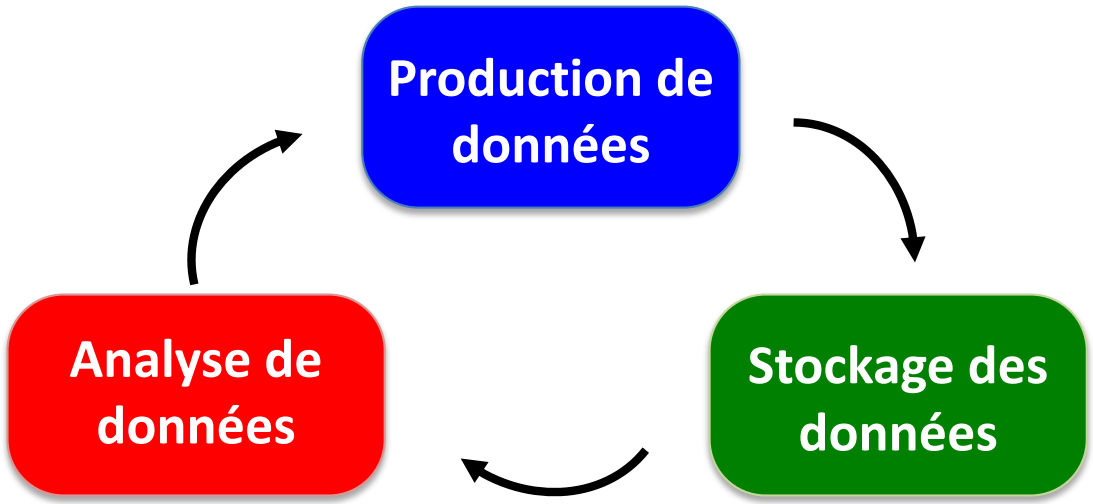
- Biologie *in vitro* → environnement artificiel

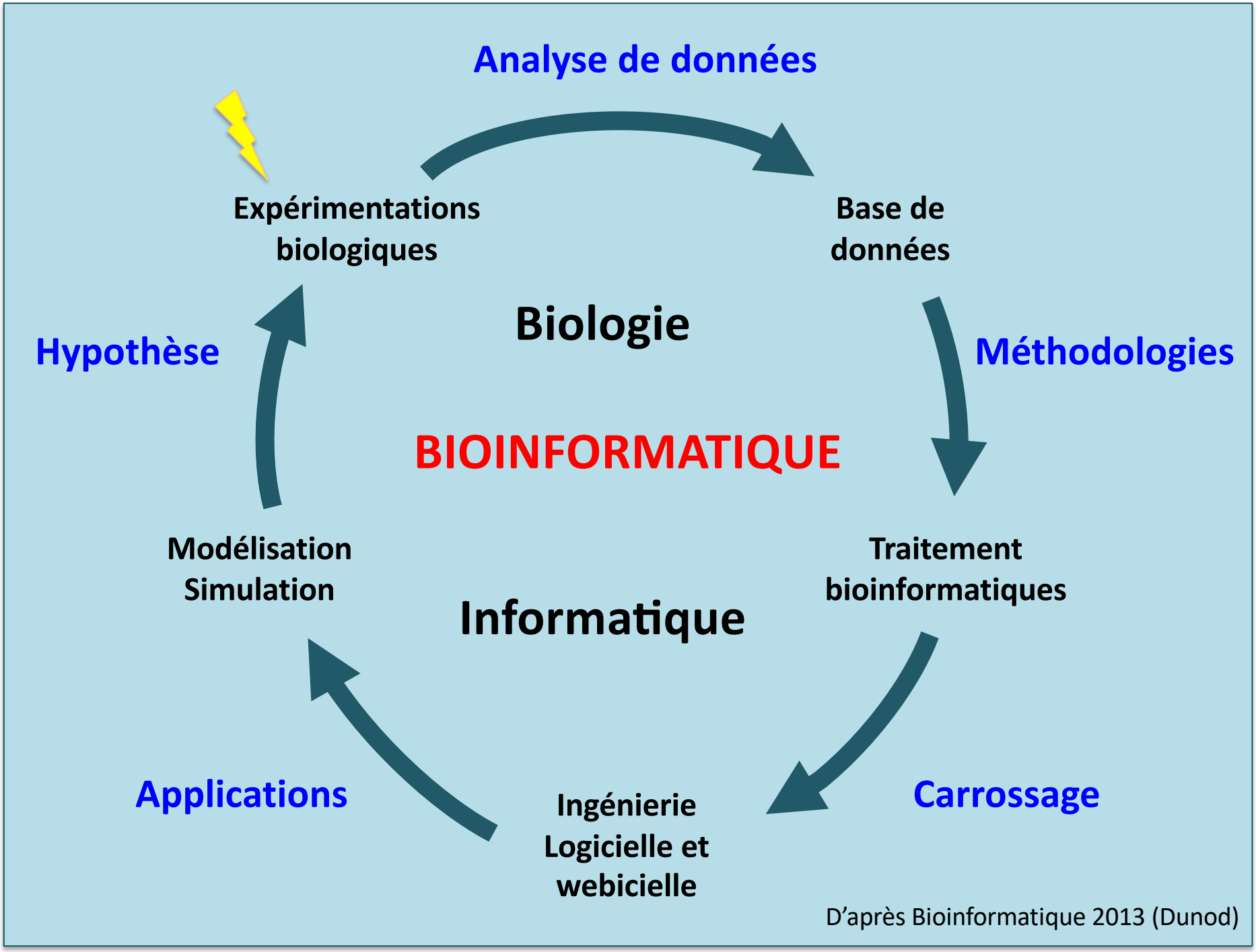
- Biologie *in vivo* → organismes vivants

Bioinformatique

Trois activités principales:

- Acquisition et organisation des données biologiques
- Conception de logiciels pour l'analyse, la comparaison et la modélisation des données
- Analyse des résultats produits par les logiciels



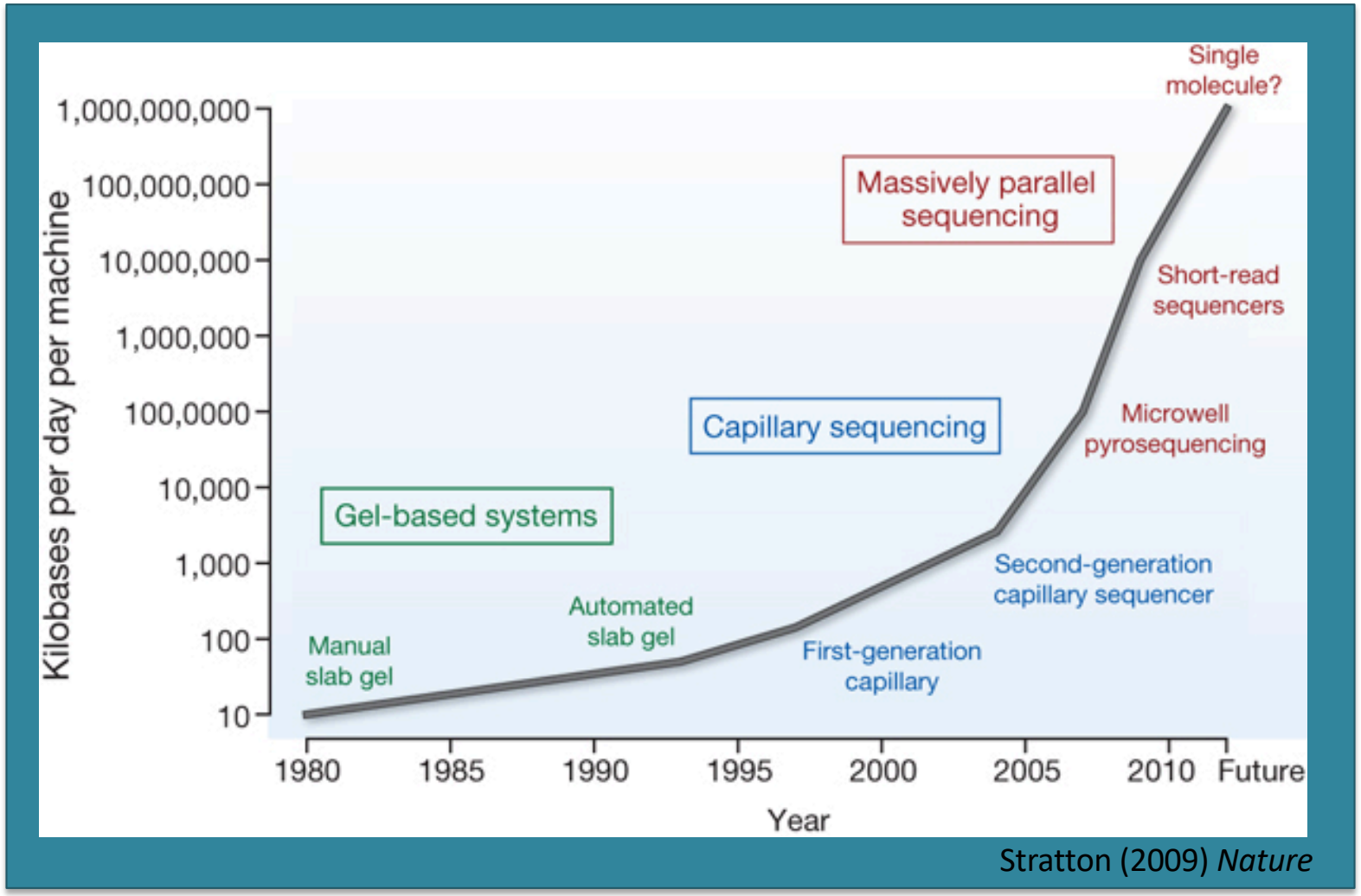

Analyse de données
Expérimentations biologiques
Base de données
Biologie
Hypothèse
Méthodologies
BIOINFORMATIQUE
Modélisation Simulation
Traitement bioinformatiques
Informatique
Applications
Ingénierie Logicielle et webicielle
Carrossage
D'après Bioinformatique 2013 (Dunod)

Bioinformatique

- Aujourd'hui pratiquement tout projet de biologie comporte une étape d'analyse bioinformatique des données.
- Un biologiste passe environ 20-30%* de son temps à utiliser des outils bioinformatiques.

L'évolution des technologies de séquençage

- En biologie, de nouveaux types de données issus des progrès technologiques émergent constamment



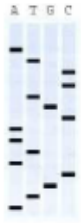
Stratton (2009) *Nature*

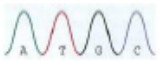
Comment on obtient ces séquences?

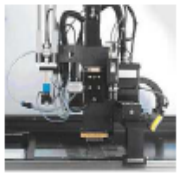
Historique (rapide) des technologies….

Rappel

Découverte de la structure de l'ADN 1953 (Watson et Crick)







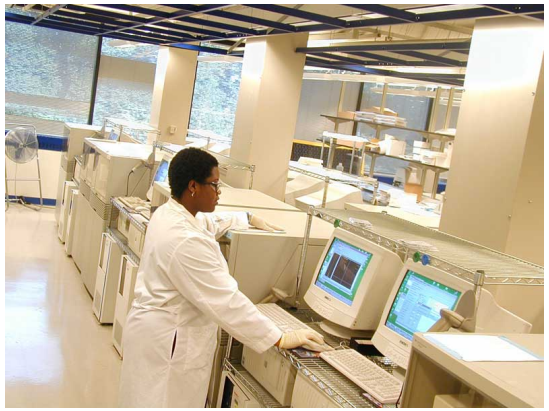
1975	1977	1990	1995	1999	2000	2007	2012
Southern Blot	Séquençage - Sanger - Gilbert	Séquençage par mesure de la fluorescence	Puce à ADN (microarray)		Séquenceur à capillaire	Séquençage à haut débit (NGS)	N-NGS

1ère GENERATION SEQUENCEUR	2ème GENERATION SEQUENCEUR	3ème GENERATION SEQUENCEUR

Génomique: la révolution en biologie?

- **1900 – La seconde révolution industrielle**

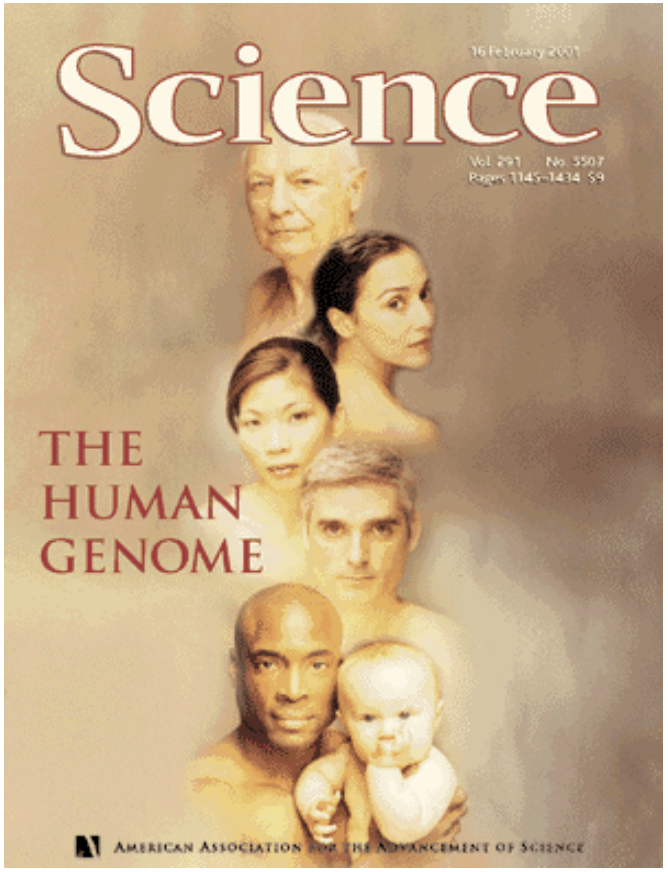
- **2000 - La révolution biologique**



Génomique: la révolution en biologie?

2000

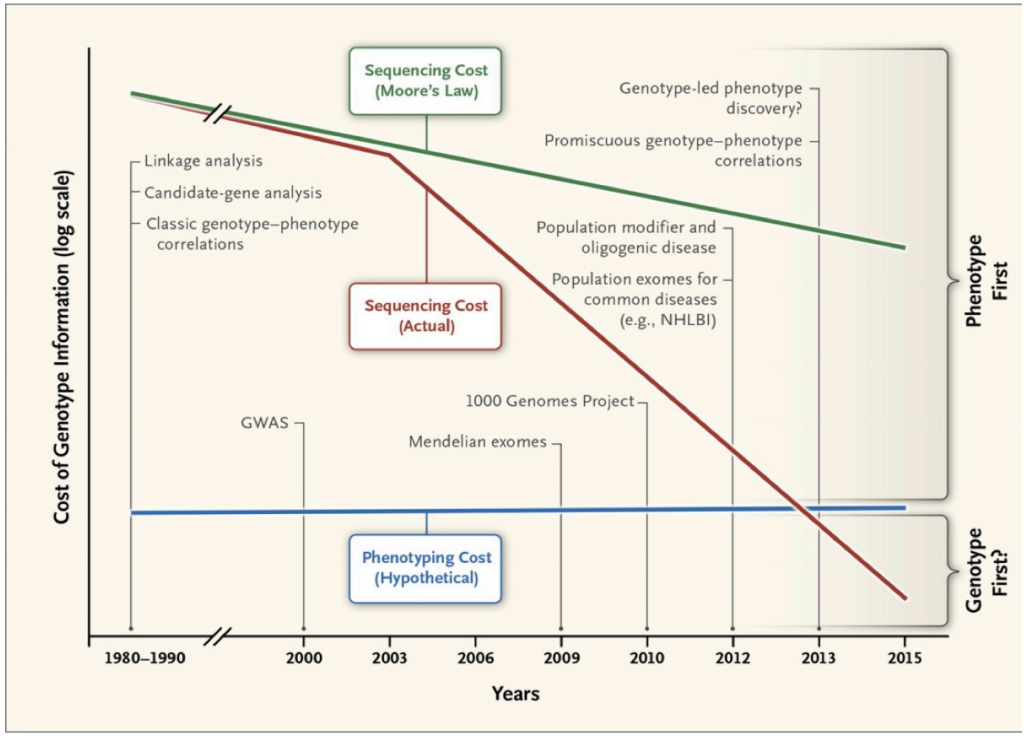
- **Le décryptage du génome humain**



La révolution NGS (*next-generation sequencing*)

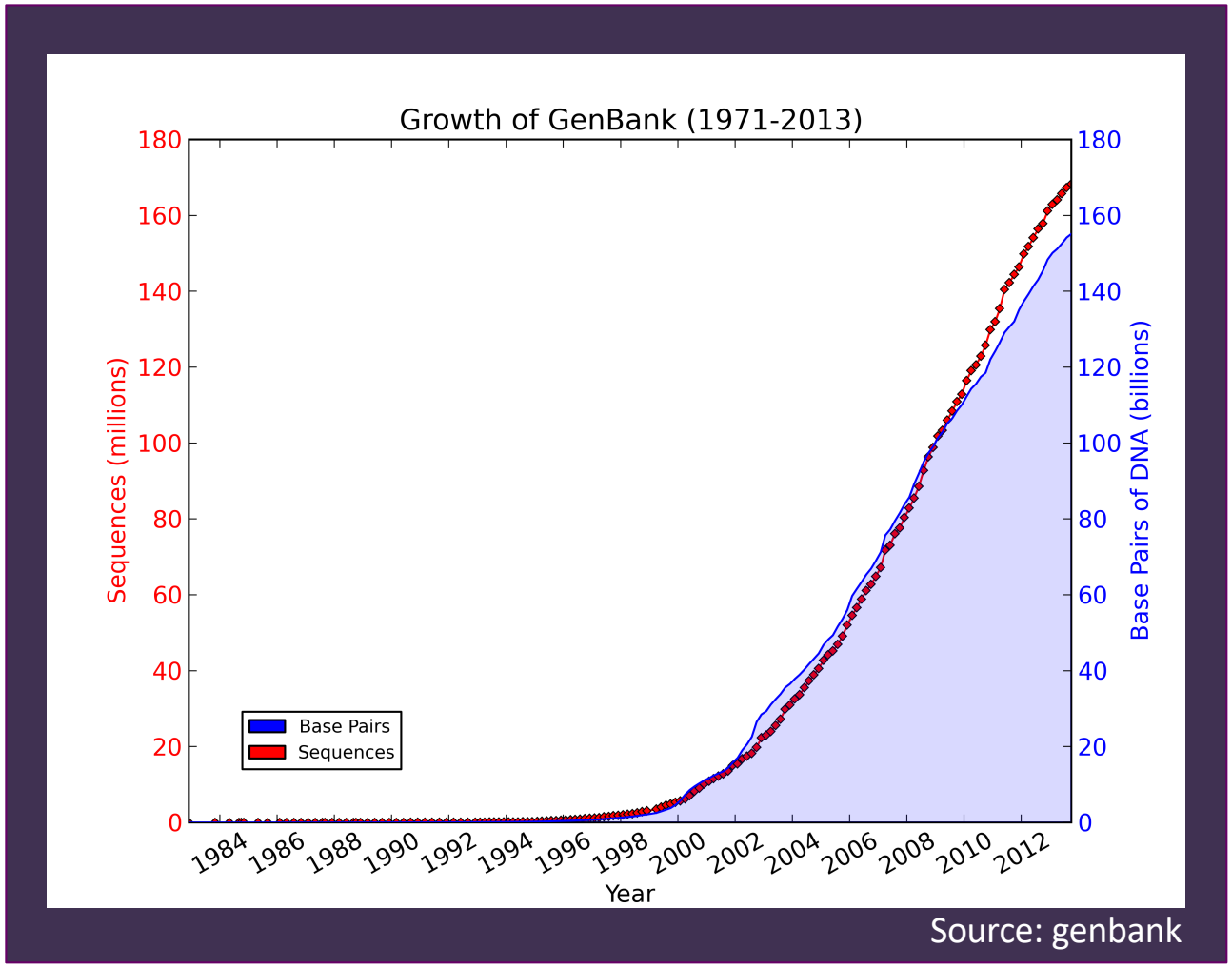
Le séquençage haut débit (HTS pour *high-throughput sequencing*) aussi appelé NGS pour *next-generation sequencing* désigne un ensemble de méthodes apparues à partir de 2005 produisant des millions de séquences en un *run* et à faibles coûts

Elles se caractérisent par l'utilisation d'approches massivement parallèles, permettant de séquencer des millions de fragments simultanément



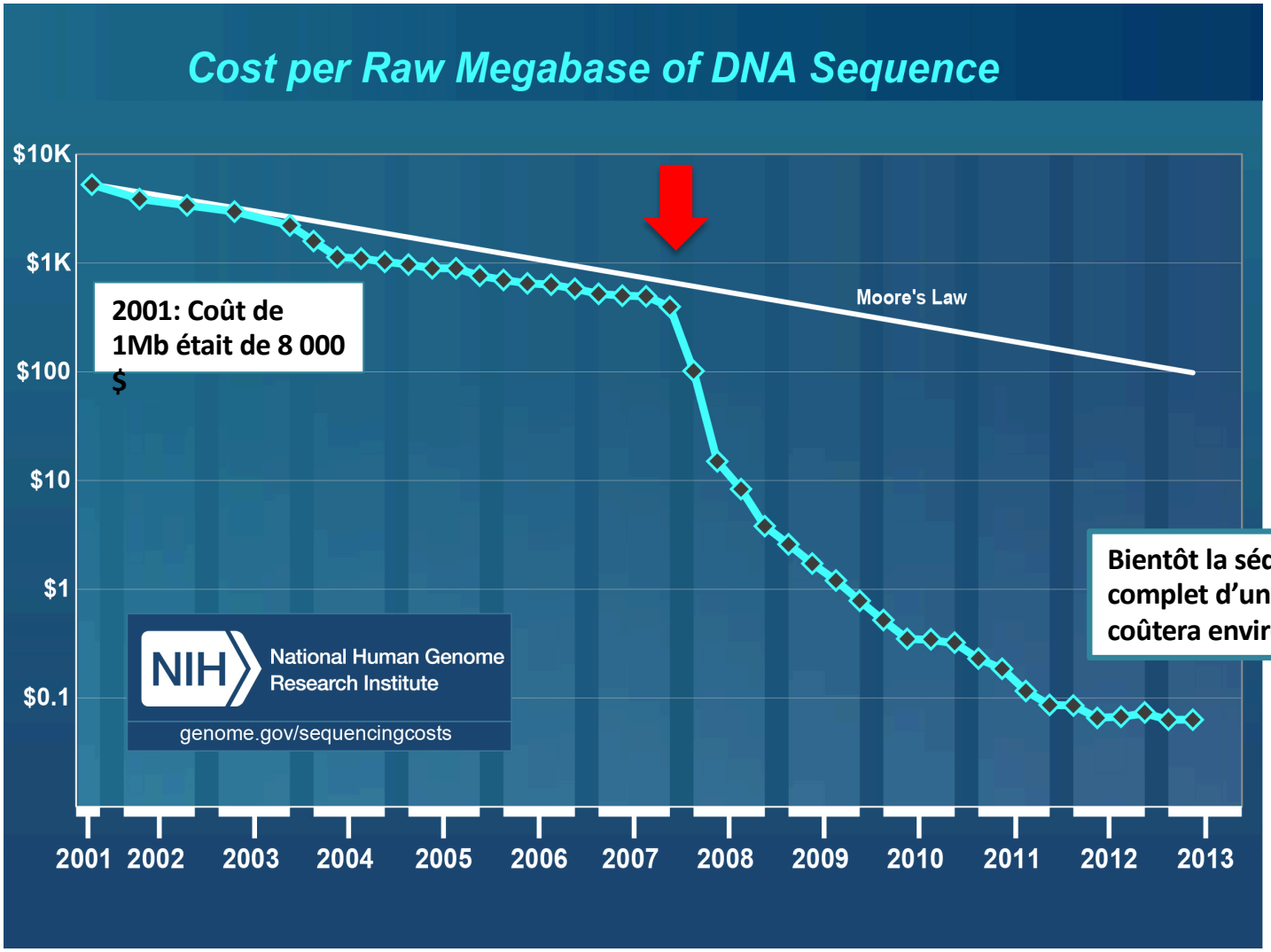
Lu JT *et al*. *N Engl J Med* 2014,371:593-6

Le volume de données croît de manière exponentielle



L'explosion de la quantité de données biologiques nécessite des outils de stockage adaptés

La baisse des coûts est corrélée à l'augmentation des débits



2012

Vers la troisième génération de séquenceur…

- **Séquenceur à haut débit de « paillasse »**

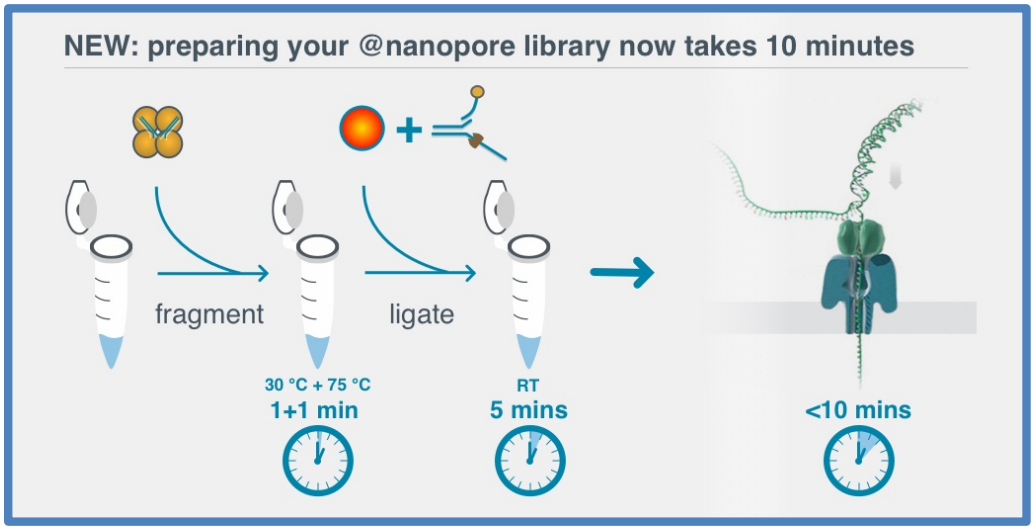
- **2nd génération d'appareils à haut débit a fin de résoudre deux des limitations des premières générations de machines:**

  – **D'une part la nécessité d'amplifier l'échantillon à analyser:**
  Nouvelles machines capables de lire directement la séquence d'ADN.

  – **D'autre part le temps d'acquisition des appareils (par cycle):**
  Les nouveaux séquenceurs sont capables d'effectuer la lecture de l'ADN en temps réel.

Séquenceur de poche?

Oxford NANOPORE Technologies



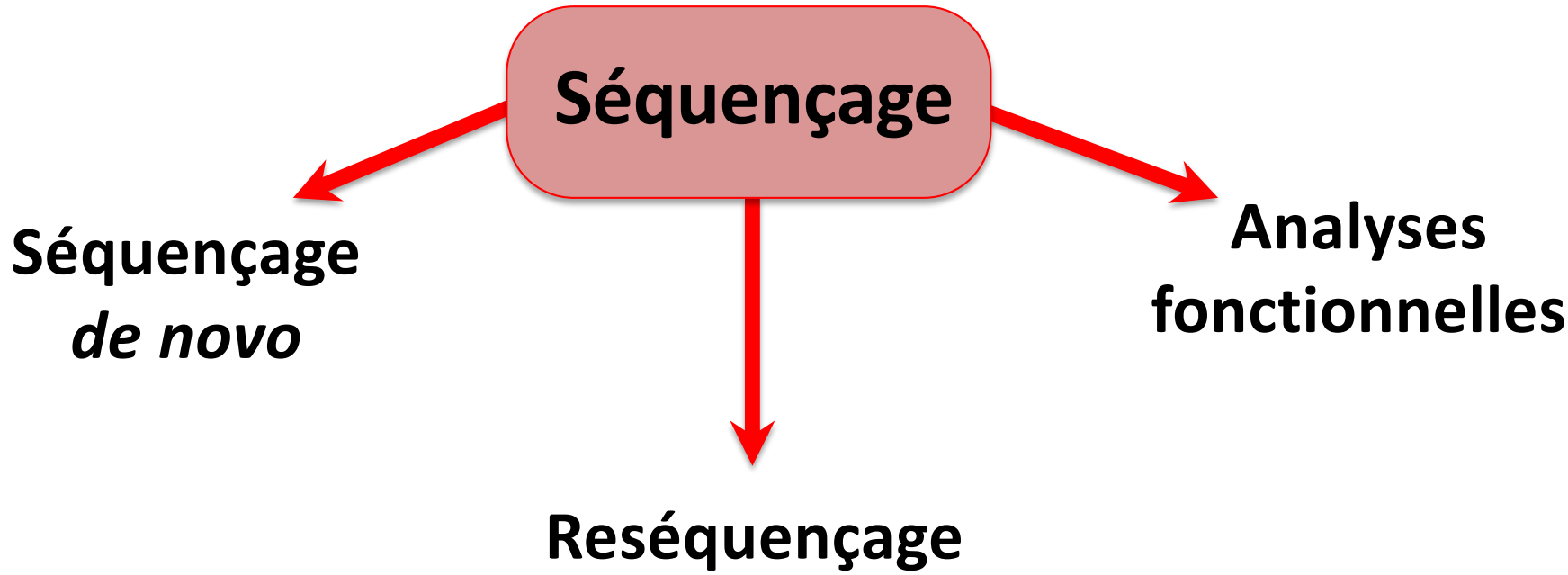
Evolution technologique et nouvelles approches

- **Les avancées technologiques ont entraînées une baisse significative des coûts de séquençage.**

- **La robotisation permet de traiter plusieurs dizaines ou centaines de milliers de clones avec peu d'interventions humaines et en réduisant les erreurs de manipulation**

→ **Développement d'outils de génomique**

Les applications du séquençage



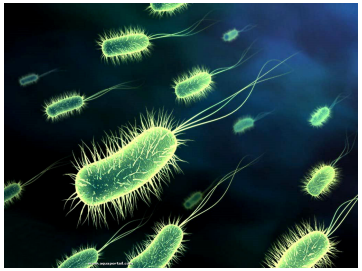
Les applications du séquençage

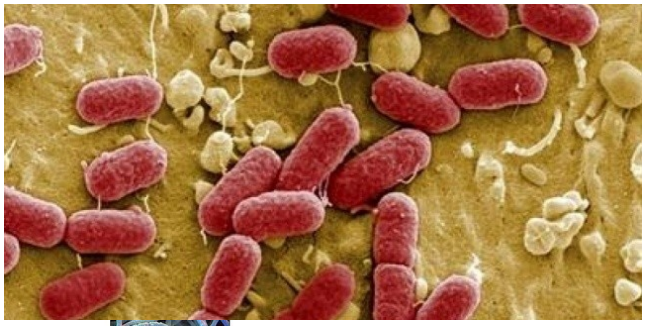
- **Le séquençage *de novo***
  - **Fournir la séquence de génome inconnu**
  - **Combinaison de plusieurs méthodes pour obtenir des version de génome de bonne qualité**
  - **Utilisés dans le domaine médical pour la découverte de génomes d'agents pathogènes inconnu ou de nouveaux virus**

Les applications du séquençage

- **Le reséquençage**

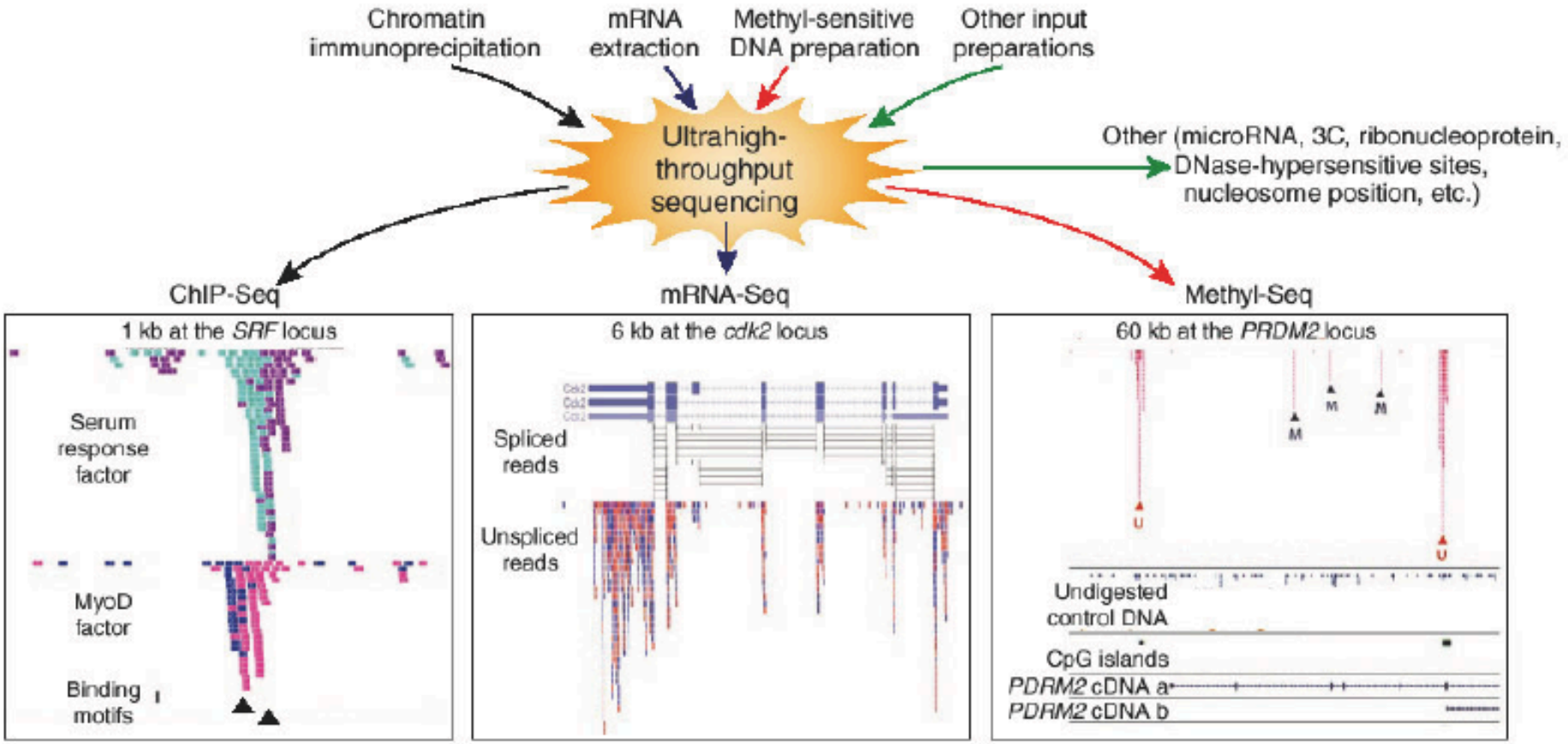
- Buts : analyser différents génomes en les **comparant à une souche de référence**.
- Recherche de **polymorphismes** dans une population, d'identification de **mutations** en biotechnologie, d'analyse d'**évolution** d'organismes, de différenciation d'une cellule au cours du temps, de la découverte d'**ADN anciens** …
- **Métagénomique** : caractériser les différents génomes présents dans un échantillon.
- Exemple : caractériser les micro-organismes pathogènes présents chez un patient (sang, tissus, …), définir l'ensemble des espèces présents dans l'environnement (écologie, dépollution, …), comprendre l'évolution des espèces, …





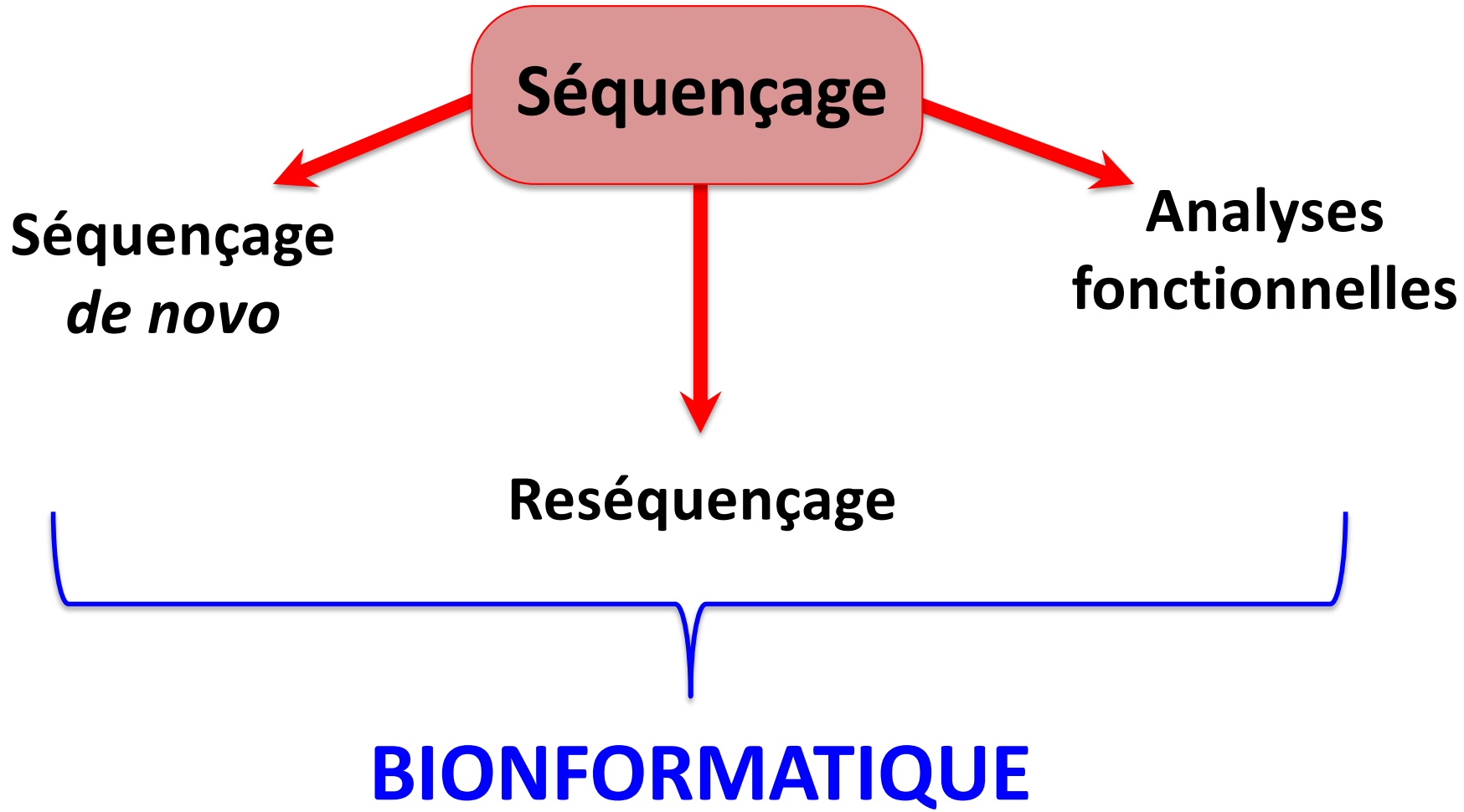
Les applications du séquençage

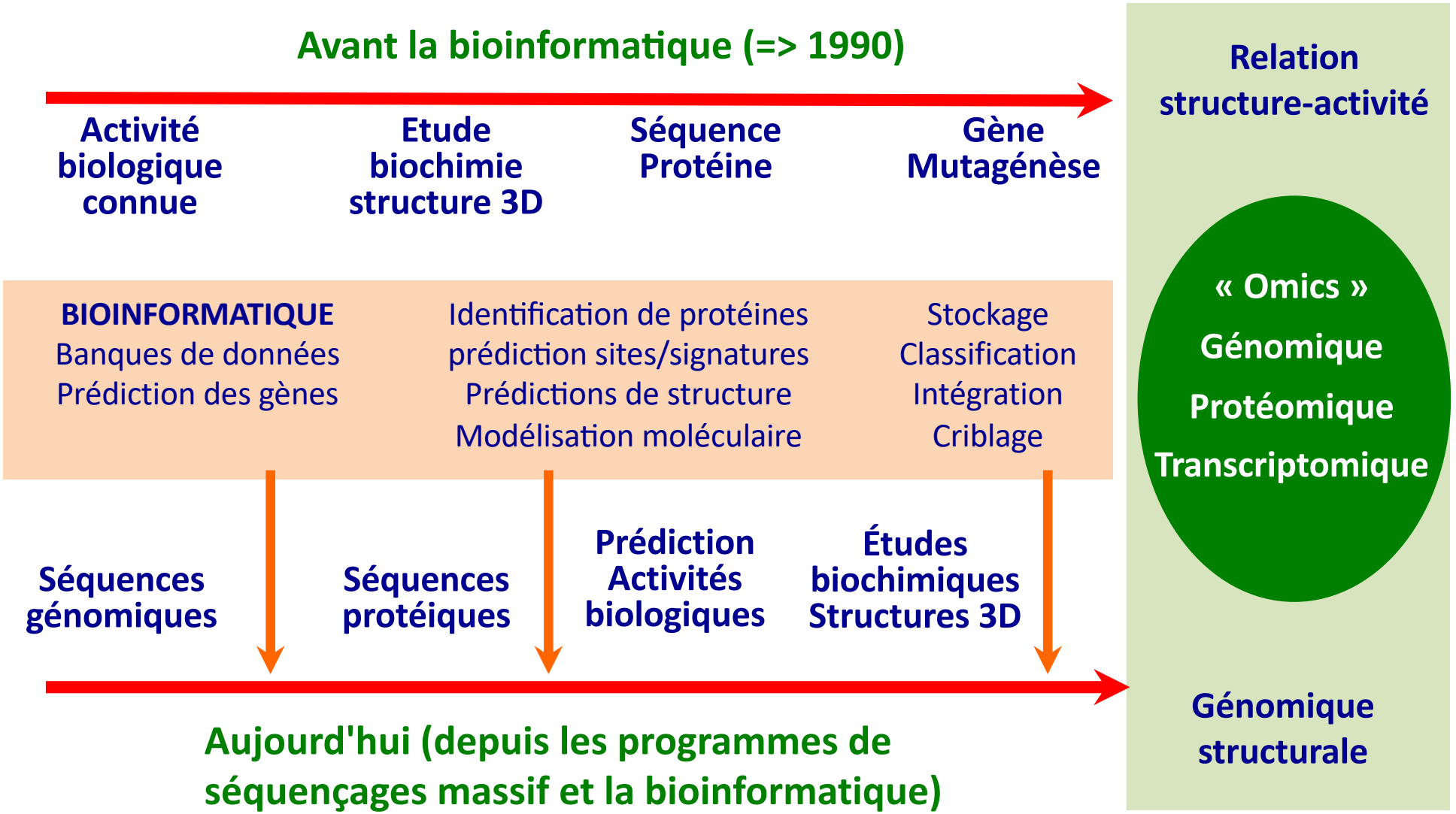
- **Les applications fonctionnelles**



Wold et al. (2008) *Nat. Methods*

Les applications du séquençage

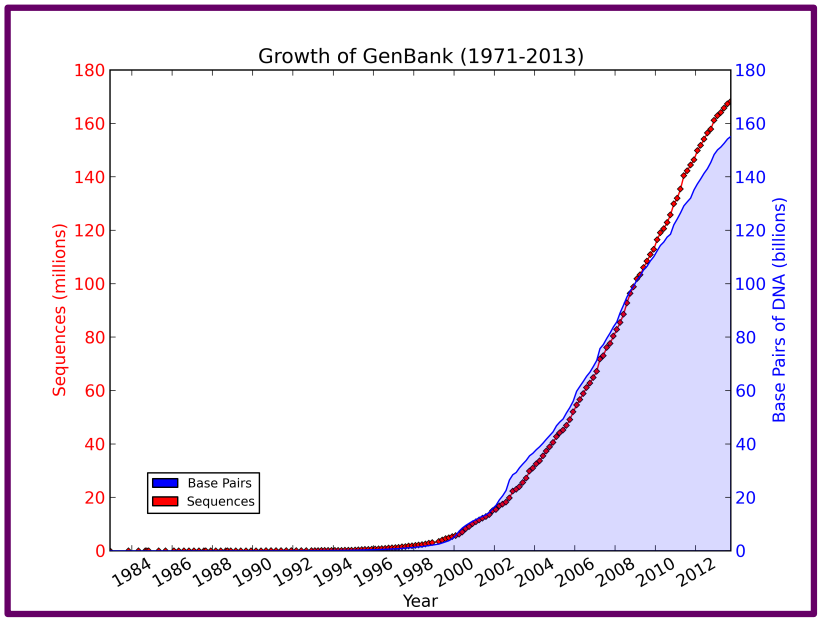




(Deleage et Gouy: Bioinformatique 2013 ed. Dunod)

Objectifs et applications de la bioinformatique

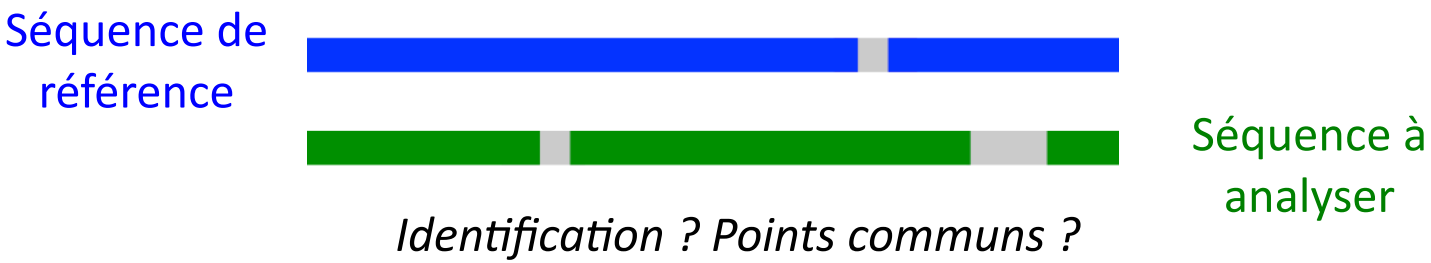
1. Collecter et stocker des informations dans des bases de données, accessibles en ligne.



L'explosion de la quantité de données biologiques nécessite des outils de stockage adaptés

Bioinformatique

2. Fournir des outils de comparaison de séquences (protéiques ou nucléotidiques).



- **Identifier une séquence par rapport à une base de données**
- **Déterminer le degré de similitudes entre deux séquences ou plus**
- **Repérer des motifs structuraux :**

  -gènes, promoteurs, etc. pour des séquences nucléotidiques.

  -zone de repliement, site actif, etc. pour une séquence protéique

Bioinformatique

3. Fournir des outils de traduction de séquences

- **Simplifier les taches de traduction**
- **Proposer plusieurs possibilités de protéines pour une même séquence**
- **Repérer exons / introns**

Bioinformatique

4. Fournir des outils de prédiction

Prédiction physiologique et fonctionnelle

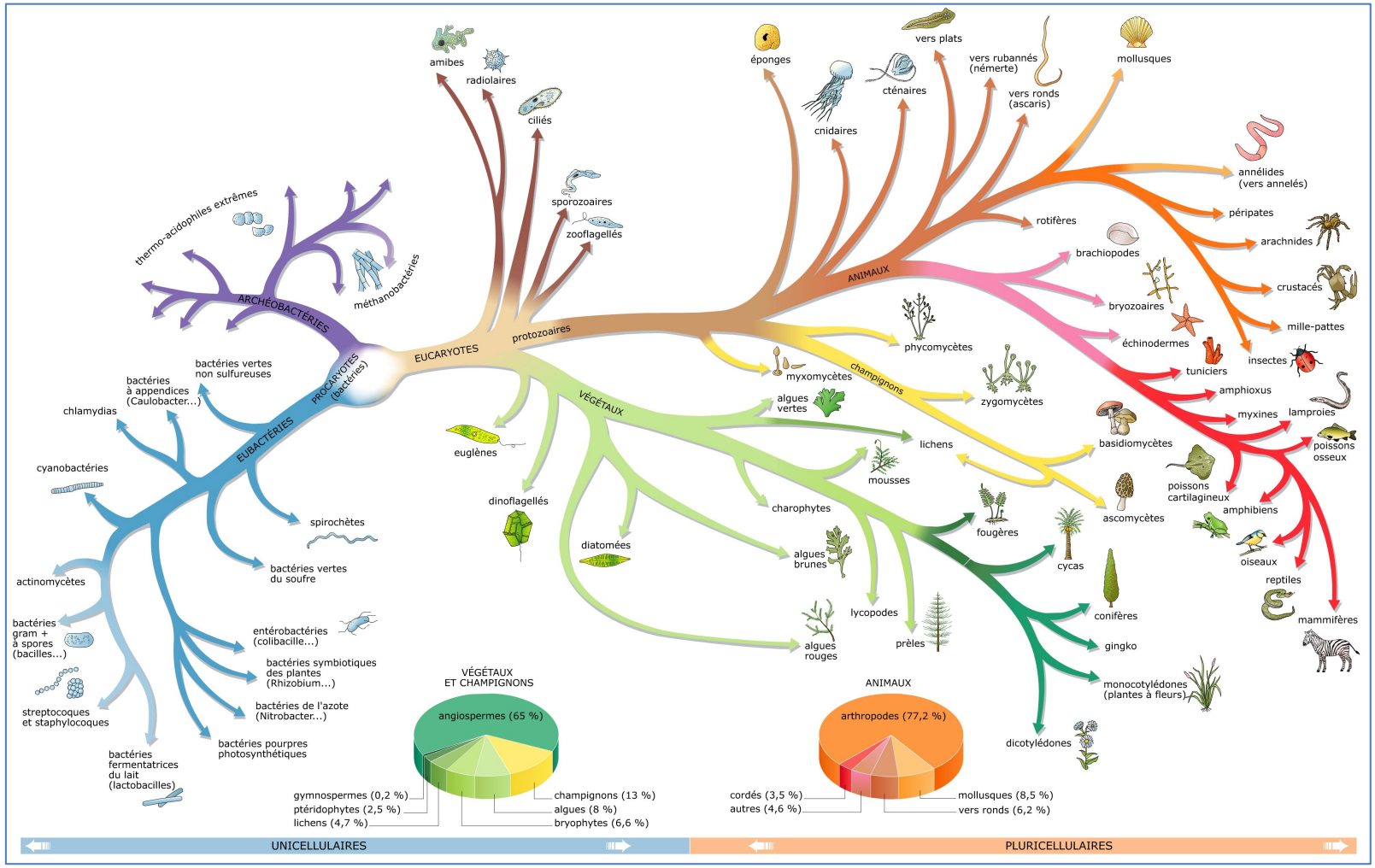
- Repérer un opéron
- Repérer un gène ou une protéine anormale
- Prévoir la structure 3D d'une protéine
- Repérer des mutations
- Prédire une pathologie…

Prédiction expérimentale

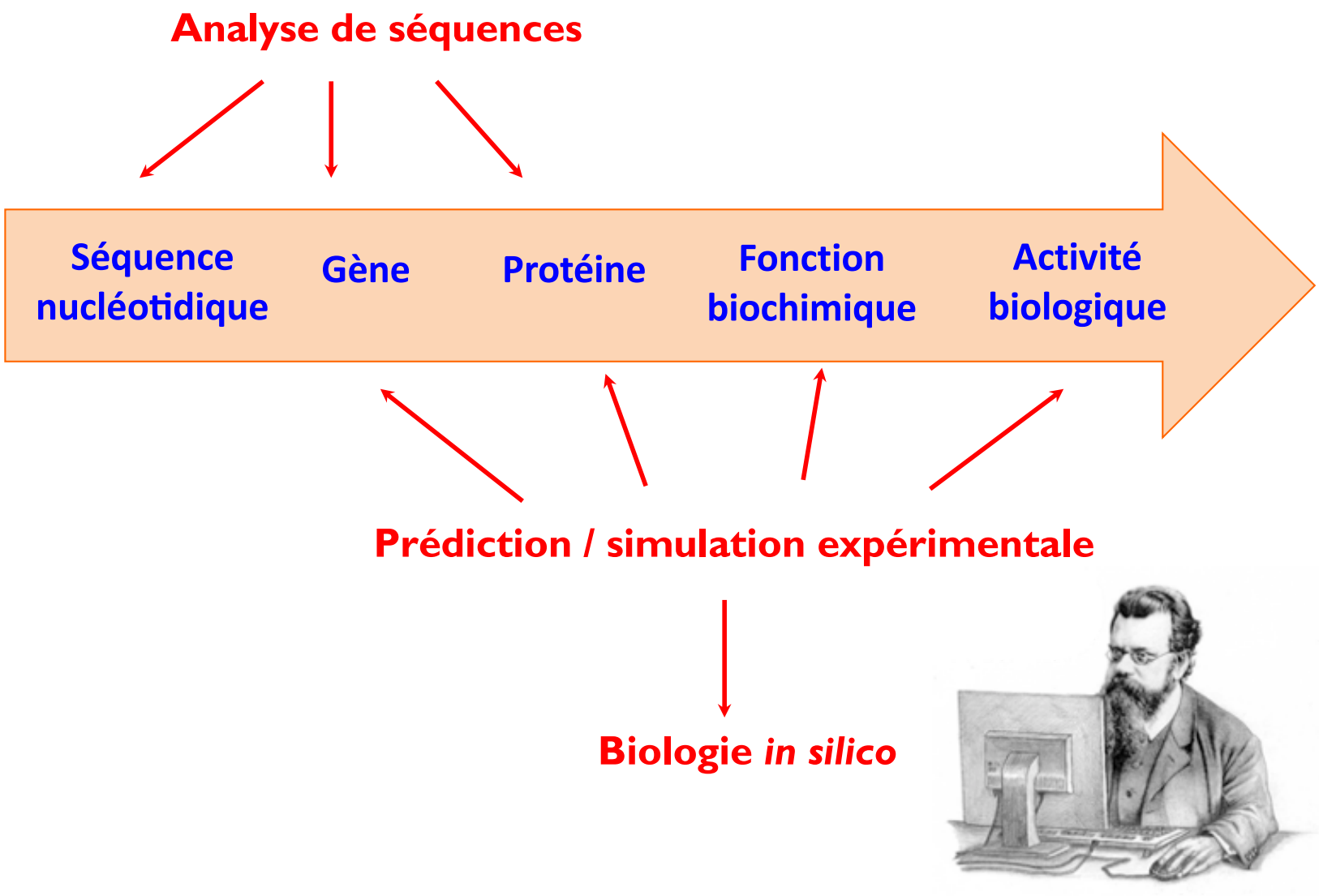
- Repérer des sites de restriction
- Prévoir la digestion d'un nucléotide
- Prévoir / simuler la migration de fragments nucléotidiques ou protéiques lors d'une électrophorèse…

Bioinformatique

5. Études phylogénétiques et l'évolution moléculaire des êtres vivant



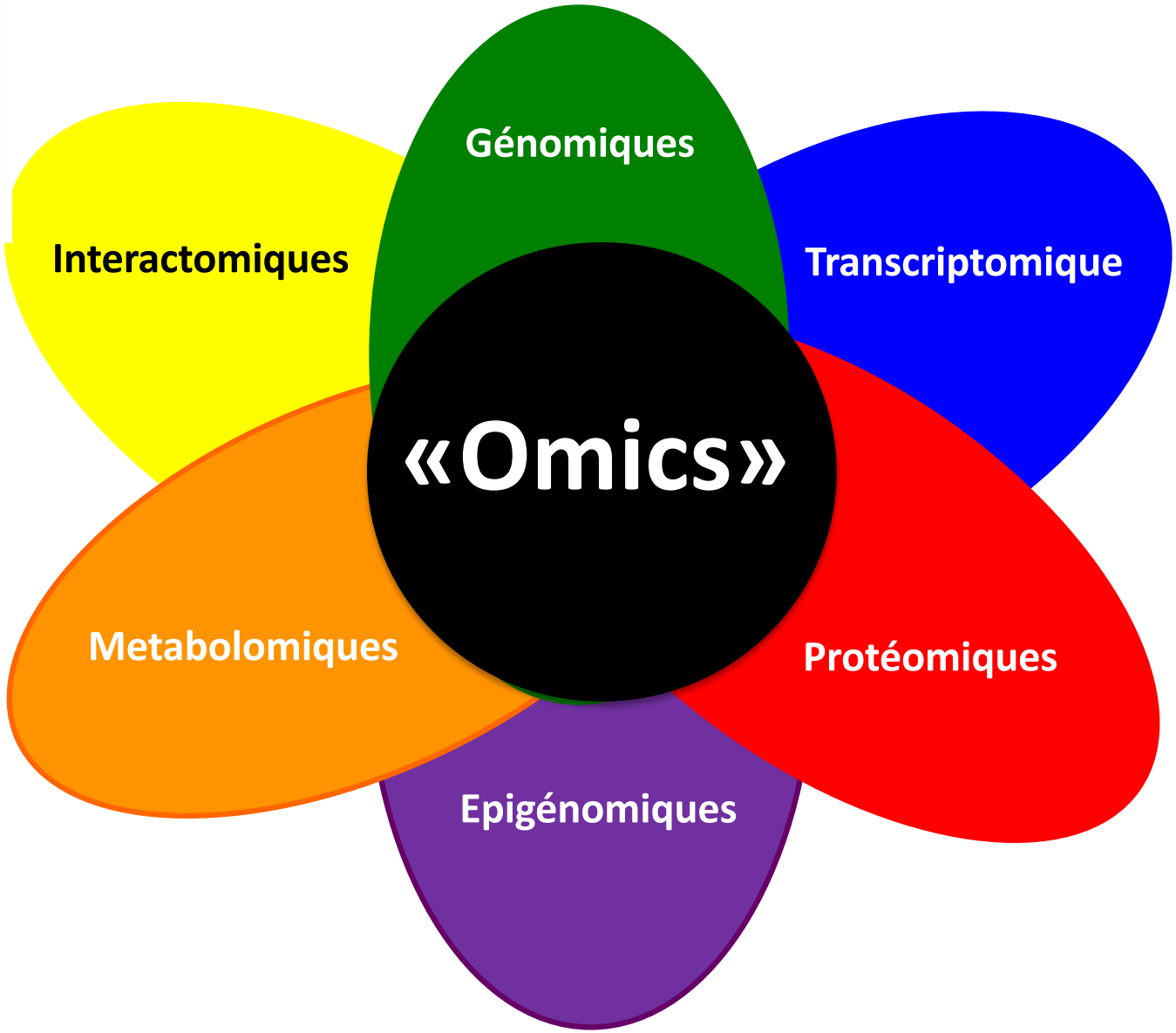
Bioinformatique



L'ERE D' « OMICS »

L'ère des omiques

Transcriptomique, protéomique, métabolomique, épigénomique...
Chacun de ces termes désigne un ensemble massif d'informations récoltées en une fois sur un échantillon donné (ARN, protéines, cellule, population de cellules, organe..).
En plein avènement depuis les années 2000, l'étude de ces échantillons est au cœur d'une évolution radicale de la biologie et ouvre des perspectives nombreuses et prometteuses en médecine.



BASES DE DONNÉES

Bases de données

- **Pour aboutir à la formulation de ces modèles et à ces prédictions, il est indispensable de tout d'abord collecter et organiser les données à travers la création de bases de données.**

LES BASES DE DONNÉES

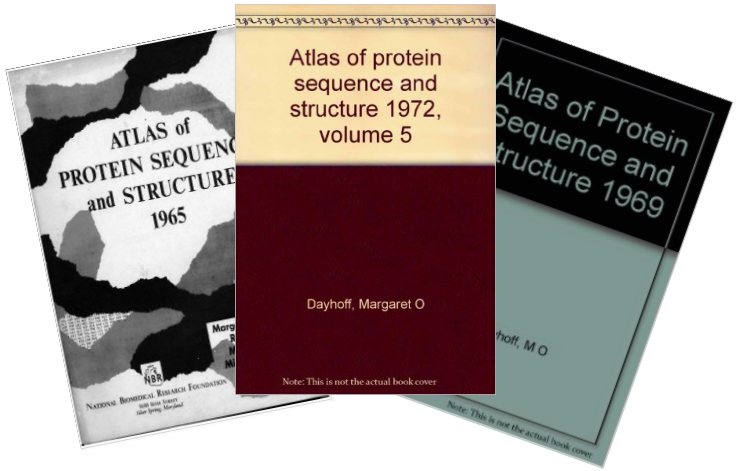
- **Une base de données est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter leur utilisation (ajout, mise à jour, recherche et éventuellement analyse).**

Bases de données ≠ Banques de données

- **Une banque de données est un ensemble de fichiers textes sans relation entre eux (fichier « plats »). Une base de données est un ensemble de relation entre des données gérées avec un système de gestion de base de données (SGBD) et interrogeable par SQL (Structure Query Langage).**

Bases de données

- **Les séquences nucléiques et protéiques étaient déposées dans un livre édité par M. Dayhoff**
- **1980: Premières banques informatisées de données de séquence biologiques**



- **Différentes catégories de bases de données :**
  - **Bases généralistes**
  - **Bases spécialisées**

Peuvent générer leurs propres données (privés)→ accessibilité restreinte ou
Obtenir les données d'autres labo (public)→ « libre service »

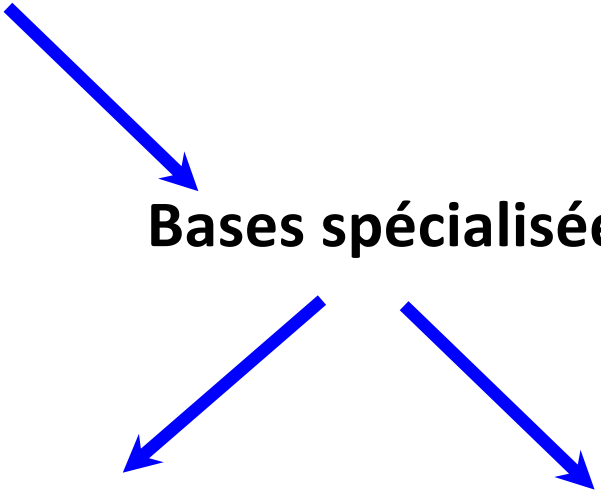
Bases de données

- **Différentes catégories de bases de données :**

Bases généralistes

-banque N

-banque P



Bases spécialisées

Banques génomiques

Banques fonctionnelles :

-transcriptome

-protéome

-métabolome…

-maladies…

Bases de données

Les deux plus grands centres de bioinformatique du monde sont:

- **Institut Européen de Bioinformatique (EBI) →**
  http://www.ebi.ac.uk

- **National Center for Biotechnology Information (NCBI)→**
  http://ncbi.nlm.nih.gov/

Banque de données de séquences de protéines SWISS-PROT→
http://www.uniprot.org/

Bases de données généralistes: publique

Pour collecter l'ensemble de séquences nucléiques:

- EMBL (Europe) devenue aujourd'hui l'ENA= European Nucleotide Archive
  → http://www.ebi.ac.uk/ena
- GenBank (USA) → http://www.ncbi.nlm.nih.gov/genbank/
- DDBJ (Japon) DNA DATA BANQUE OF JAPAN→ http://www.ddbj.nig.ac.jp

CONSORTIUM INTERNATIONAL→ INSDC

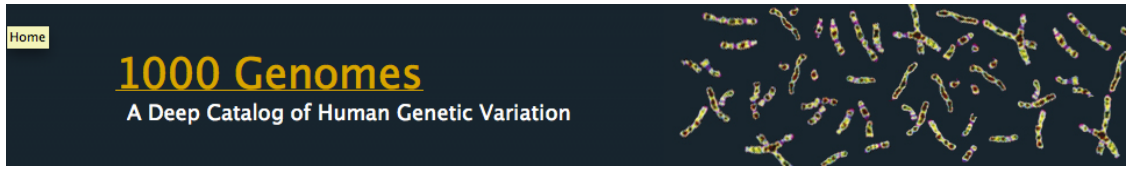
Permet d'accéder à des nombreuses données et aucun d'entre eux génère ces données

Soumettre des séquences → numéro d'accession

Bases de données généralistes: privé

- 1000 genomes project

http://www.1000genomes.org/



- http://www.jcvi.org/

J. Craig Venter®
INSTITUTE

Bases de données généralistes

Bases de données de séquences protéiques:

Les deux plus importantes :

- **SwissProt (1986) : banque manuellement annotée et «nettoyée »**
- **PIR/NBRF (1984) : banque américaine fournissant une classification des protéines basée sur la similarité entre les séquences.**

CONSORTIUM *Universal Protein Resource* → UniProt

http://www.uniprot.org

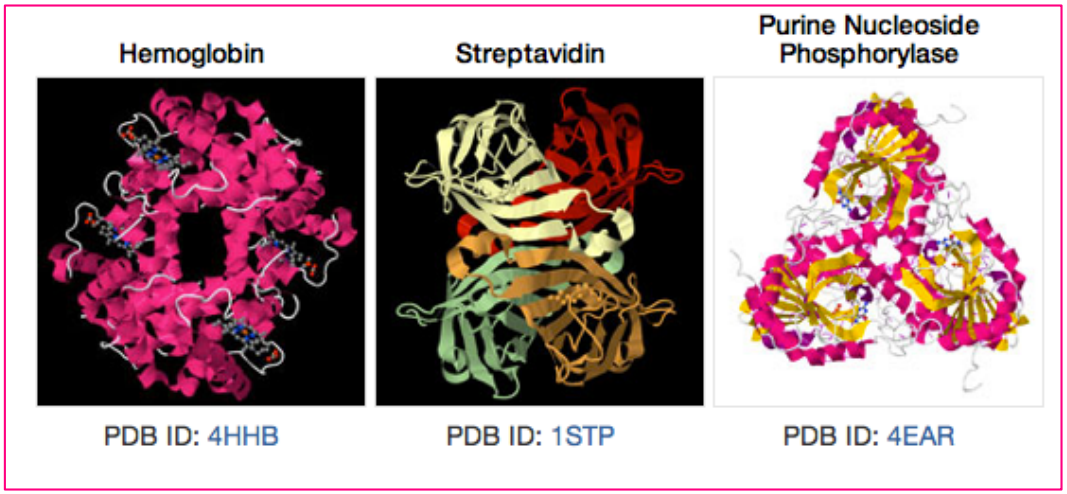
http://www.expasy.org

Bases de données généralistes

Bases de données de structure:

- La Protein Database (PDB) stockent les structures protéiques obtenues par RMN ou cristallographie
- Une entrée contient donc les coordonnées de tous les atomes de la structure

http://mm.rcsb.org



Bases de données spécialisées

- **Chaque année, en janvier, le journal Nucleic Acids Research publie un numéro spécial dédié aux bases de données « database »**

The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection

Michael Y. Galperin and Xosé M. Fernández-Suárez

The 19th annual Database Issue of *Nucleic Acids Research* features descriptions of **92 new online databases** covering various areas of molecular biology and 100 papers describing recent updates to the databases previously described in *NAR* and other journals. The highlights of this issue include, among others, a description of neXtProt, a knowledgebase on human proteins; a detailed explanation of the principles behind the NCBI Taxonomy Database; NCBI and EBI papers on the recently launched BioSample databases that store sample information for a variety of database resources; descriptions of the recent developments in the Gene Ontology and UniProt Gene Ontology Annotation projects; updates on Pfam, SMART and InterPro domain databases; update papers on KEGG and TAIR, two universally acclaimed databases that face an uncertain future; and a separate section with 10 wiki-based databases, introduced in an accompanying editorial. The *NAR* online Molecular Biology Database Collection, available at http://www.oxfordjournals.org/nar/database/a/, has been updated and now lists **1380 databases**. Brief machine-readable descriptions of the databases featured in this issue, according to the BioDBcore standards, will be provided at the http://biosharing.org/biodbcore web site. The full content of the Database Issue is freely available online on the *Nucleic Acids Research* web site (http://nar.oxfordjournals.org/).

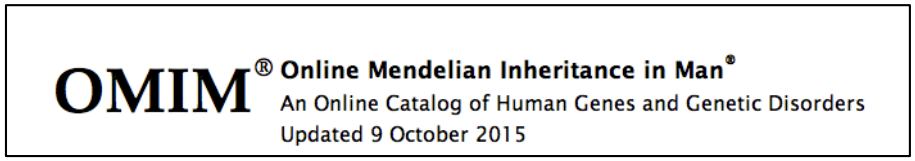
Bases de données spécialisée

Dédiées à un organisme :

- Flybase : Drosophile http://flybase.org
- HIV database : www.hiv.lanl.gov/
- Porteco: Escherichia coli http://www.porteco.org
- *Arabidopsis thaliana:* TAIR https://www.arabidopsis.org

Bases de données spécialisée

Dédiées à un type de séquences particulier :



- IMGT : données d'immunologie http://www.imgt.org
- EPD : Eukaryotic Promoter Database http://epd.vital-it.ch
- The European ribosomal RNA database

http://bioinformatics.psb.ugent.be/webtools/rRNA

- Online Mendelian Inheritance in Man http://www.omim.org
- GOLD: Genome Online Database https://gold.jgi.doe.gov

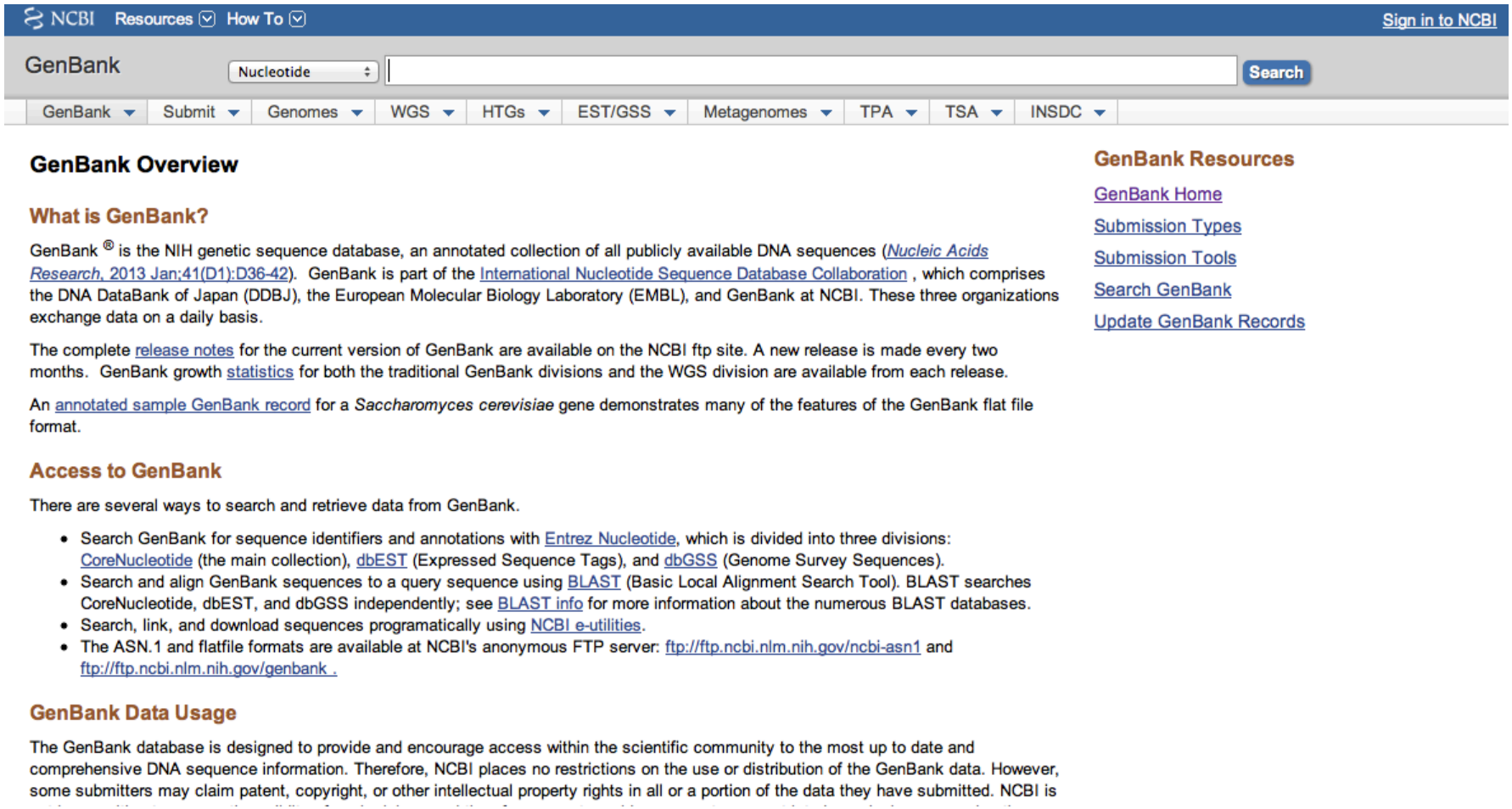
Bases de données: NCBI

C'est un de plus grande centre de bioinformatique du monde

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

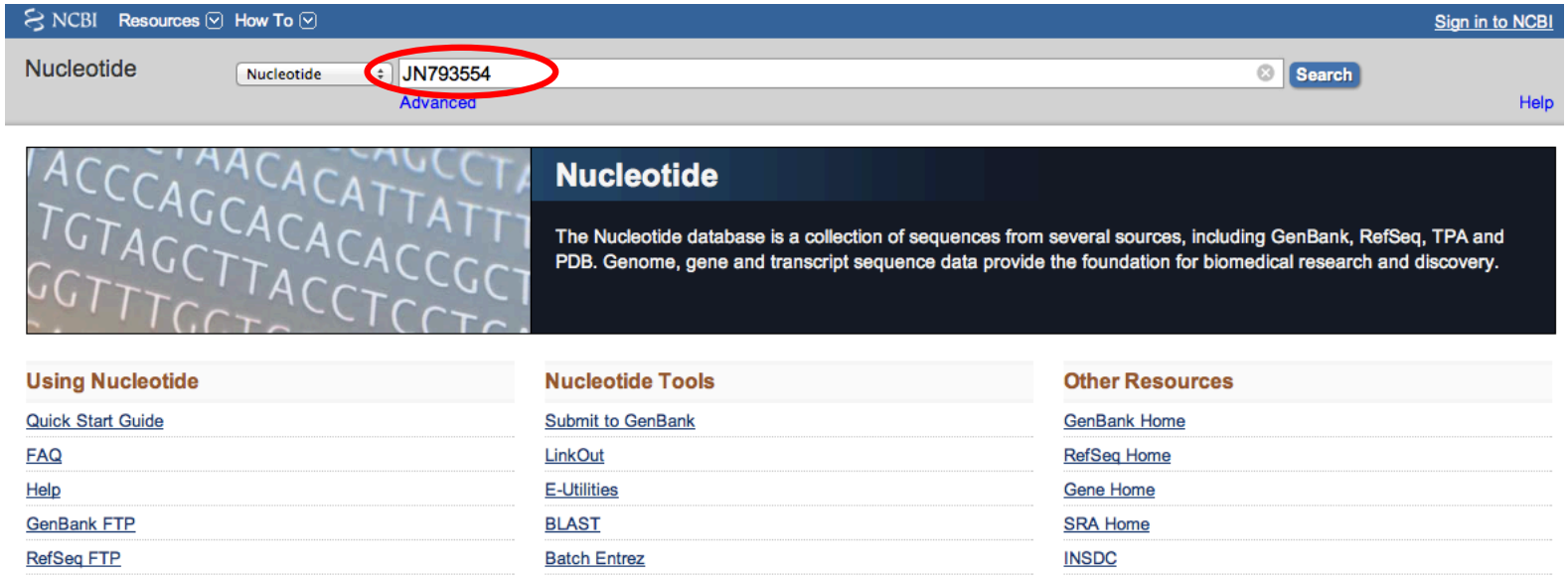
Comment on cherche dans les bases de données?
Exemple Genbank

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



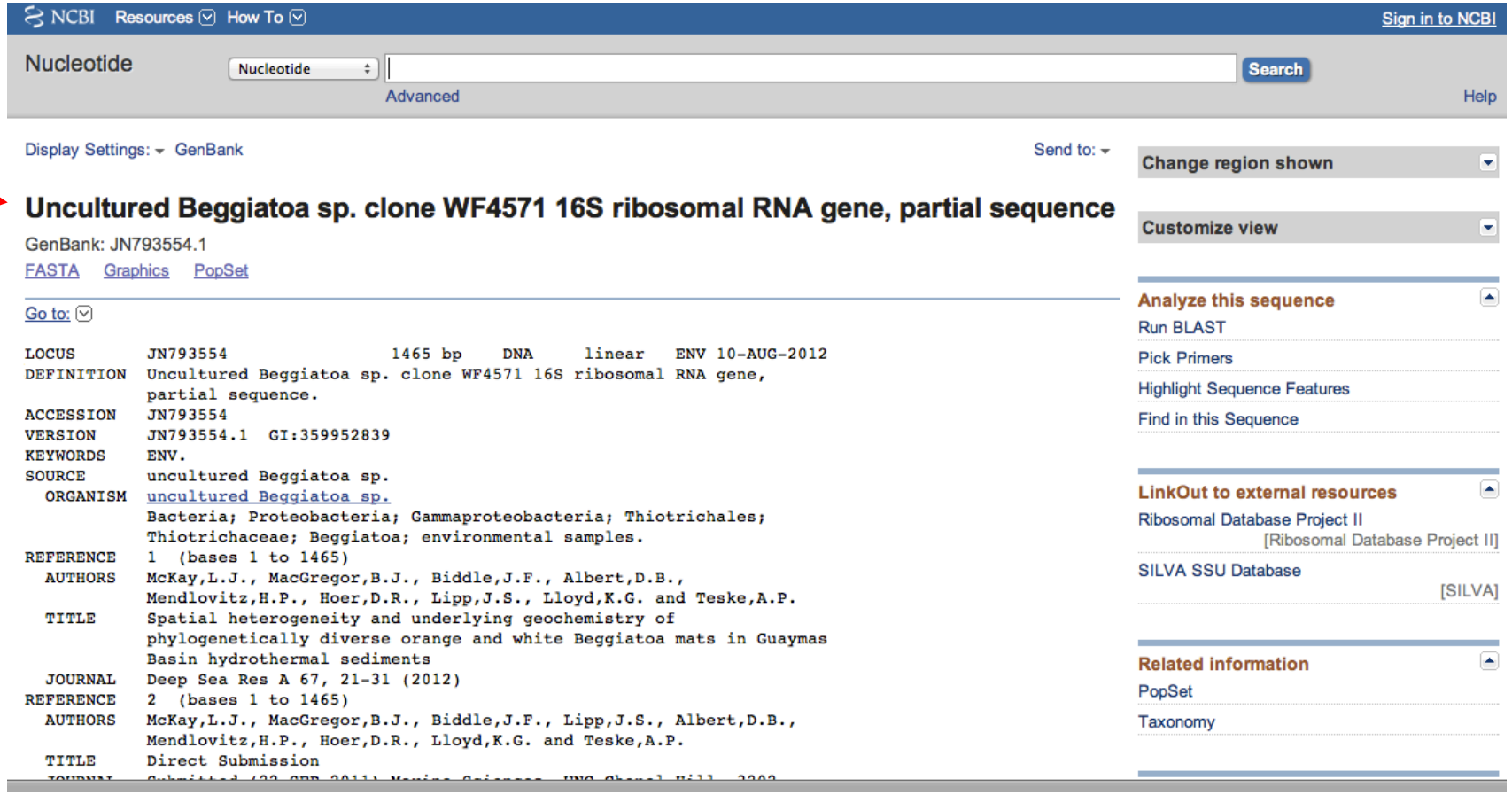
Exemple Genbank

Je connais le No. d'accession (paper)



Exemple Genbank

RESULTATS….



Exemple Genbank

```
LOCUS       JN793554                1465 bp    DNA     linear   ENV 10-AUG-2012
DEFINITION  Uncultured Beggiatoa sp. clone WF4571 16S ribosomal RNA gene,
            partial sequence.
ACCESSION   JN793554
VERSION     JN793554.1  GI:359952839
KEYWORDS    ENV.
SOURCE      uncultured Beggiatoa sp.
  ORGANISM  uncultured Beggiatoa sp.
            Bacteria; Proteobacteria; Gammaproteobacteria; Thiotrichales;
            Thiotrichaceae; Beggiatoa; environmental samples.
REFERENCE   1  (bases 1 to 1465)
  AUTHORS   McKay,L.J., MacGregor,B.J., Biddle,J.F., Albert,D.B.,
            Mendlovitz,H.P., Hoer,D.R., Lipp,J.S., Lloyd,K.G. and Teske,A.P.
  TITLE     Spatial heterogeneity and underlying geochemistry of
            phylogenetically diverse orange and white Beggiatoa mats in Guaymas
            Basin hydrothermal sediments
  JOURNAL   Deep Sea Res A 67, 21-31 (2012)
REFERENCE   2  (bases 1 to 1465)
  AUTHORS   McKay,L.J., MacGregor,B.J., Biddle,J.F., Lipp,J.S., Albert,D.B.,
            Mendlovitz,H.P., Hoer,D.R., Lloyd,K.G. and Teske,A.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-SEP-2011) Marine Sciences, UNC-Chapel Hill, 3202
            Venable Hall, Chapel Hill, NC 27599, USA
FEATURES             Location/Qualifiers
     source          1..1465
                     /organism="uncultured Beggiatoa sp."
                     /mol_type="genomic DNA"
                     /isolation_source="hydrothermal seep at Guaymas Basin"
                     /db_xref="taxon:204720"
                     /clone="WF4571"
                     /environmental_sample
                     /note="white filament collected on Alvin dive 4571"
     rRNA            <1..>1465
                     /product="16S ribosomal RNA"
```

Exemple Genbank

```
ORIGIN
        1 attgaacgct ggcggcatgc ttaatacatg caagtcgaac ggtaacatgc ccttcggggt
       61 gatgacgagt ggcggacggg tgagtaatgc ataggaatct acccagtaga cggggaacaa
      121 cttggggaaa ctcaagctaa taccgcgtaa gccttacggg ggatagcggg ggacttttta
      181 ggaagcctcg cgatattgga tgagcctatg ccggattagc tagttggtgg ggtaaaagct
      241 taccaaggct acgatccgta gctggtctga gaggacgatc agccacactg ggactgagac
      301 acggcccaga ctcctacggg aggcagcagt agggaatatt ggacaatggg cgaaagcttg
      361 atccagcaat gccgcgtgtg tgaagaaggc ctgcgggttg taaagcactt tcagttggga
      421 agaaaagctt taggttaata ccctgaagta ttgacgttac caacagaaga agcaccggct
      481 aactctgtgc cagcagccgc ggtaatacag agggtgcgag cgttaatcgg aattactggg
      541 cgtaaagcgt acgtaggcgg ttgagtcagt cggatgtgaa agcccaaggc ttaaccttgg
      601 aactgcattc gatactactc ggctagagta caggagaggg aagcggaatt cctagtgtag
      661 cggtgaaatg cgtagagatc aggagggaca ccagtggcga aggcggcttc ttggactgat
      721 actgacgctg aggtacgaaa gcgtggggag caaacaggat tagatgccct ggtagtccac
      781 gccctaaacg atgagaacta gatgttgggg gaatttaatc ccttagtatc gcagctaacg
      841 cgctaagttc tccgcctggg gagtacggcc gcaaggttaa aactcaaatg aattgacggg
      901 ggcccgcaca agcggtggag catgtggttt aattcgatgc aacgcgaaga accttacctg
      961 gccttgacat ccttggaacc tcgcagagat gtgagggtgc cttcgggaac cgagagacag
     1021 gtgctgcatg gctgtcgtca gctcgtgtcg tgagatgttg ggttaagtcc cgcaacgagc
     1081 gcaacccttа tccctagttg ccagcgattc ggtcgggaac tctagagaga ctgccggtga
     1141 caaaccggag gaaggtgggg acgacgtcaa gtcatcatgg cccttacggc cagggctaca
     1201 cacgtgctac aatggggtag tacaaagggt tgcgaacccg cgaggggtg ctaatctcac
     1261 aaaactactc gtagtccgga ttggagtctg caactcgact ccatgaagtt ggaatcgcta
     1321 gtaatcgcgg atctgcatgt cgcggtgaat acgttcccgg gccttgtaca caccgcccgt
     1381 cacaccatgg gagtgggctg taccagaagt aggtagtcta accgcaaggg ggacgcttac
     1441 cacggtatgg ttcatgactg gggtg
//
```

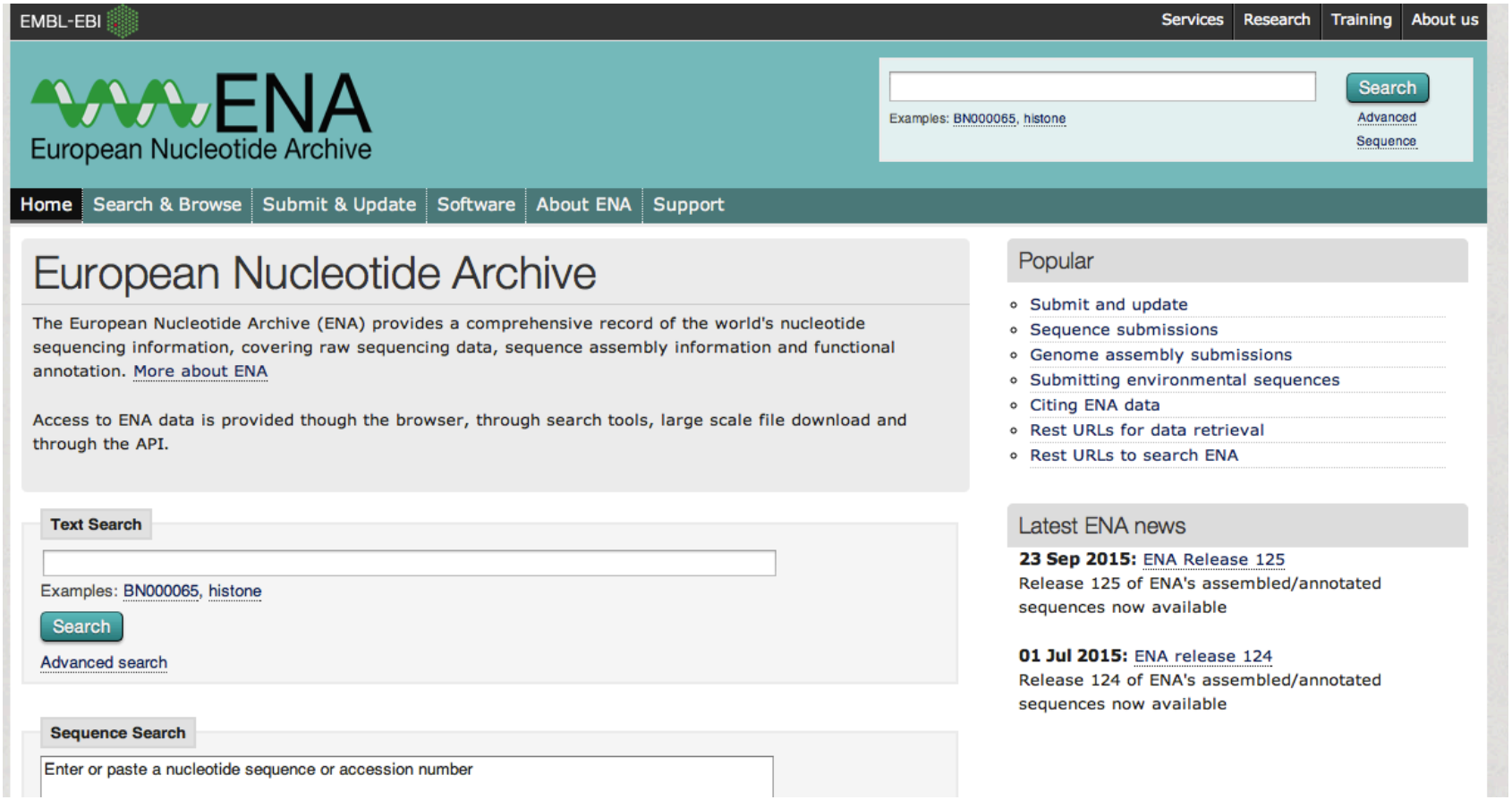
Exemple Genbank

En résumé, une fiche comporte de nombreuses informations :

Locus	Identificateur (nom et taille de la séquence)
Definition	Description de la séquence
Accession / version	Numéro d'accès dans la base
Keyword / Source / Organism / Reference / Authors / Title / Journal	Informations diverses (taxonomie, publications…)
Features	Caractéristiques de la séquence / produits d'expression (CDS)
Origin	Séquence (par blocs de caractères / par lignes)
//	Fin de l'entrée dans la base

Bases de données: ENA

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



FORMAT FASTA

Format commun de manipulation des données :
le format FASTA (Fast – alignment)

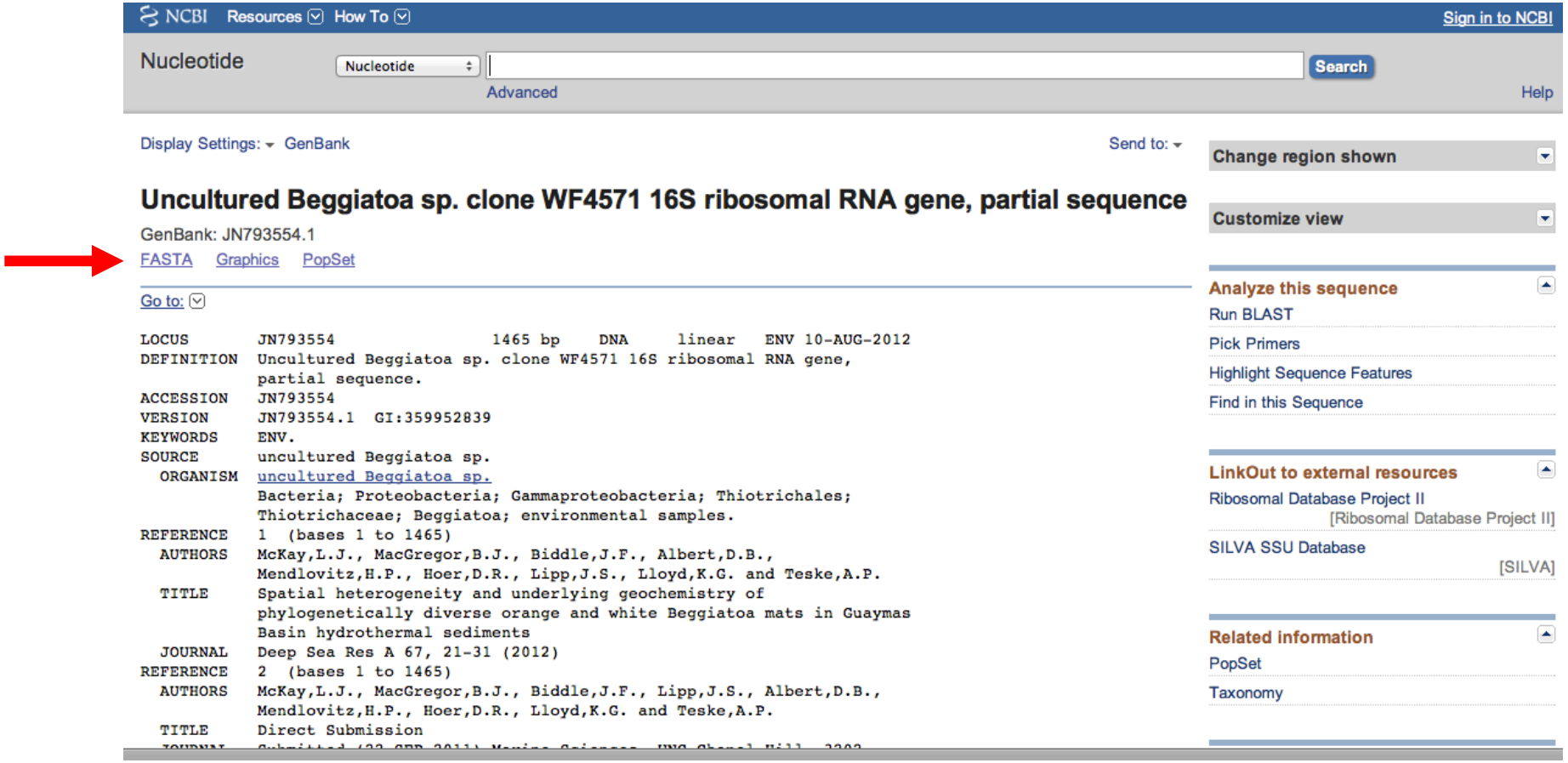
Objectif : **manipuler facilement** des séquences dans les bases de données, à l'aide d'un **format universel**, compatibles avec les traitements de texte (sous forme de fichier texte), ou par copier – coller.

Exemple de la fiche précédente de la séquence partiel du gene codant pour 16S ribosomal de Beggiatoa en format FASTA :

```
ORIGIN
        1 attgaacgct ggcggcatgc ttaatacatg caagtcgaac ggtaacatgc ccttcggggt
       61 gatgacgagt ggcggacggg tgagtaatgc ataggaatct acccagtaga cggggaacaa
      121 cttggggaaa ctcaagctaa taccgcgtaa gccttacggg ggatagcggg ggacttttta
      181 ggaagcctcg cgatattgga tgagcctatg ccggattagc tagttggtgg ggtaaaagct
      241 taccaaggct acgatccgta gctggtctga gaggacgatc agccacactg ggactgagac
      301 acggcccaga ctcctacggg aggcagcagt agggaatatt ggacaatggg cgaaagcttg
      361 atccagcaat gccgcgtgtg tgaagaaggc ctgcgggttg taaagcactt tcagttggga
      421 agaaaagctt taggttaata ccctgaagta ttgacgttac caacagaaga agcaccggct
      481 aactctgtgc cagcagccgc ggtaatacag agggtgcgag cgttaatcgg aattactggg
      541 cgtaaagcgt acgtaggcgg ttgagtcagt cggatgtgaa agcccaaggc ttaaccttgg
      601 aactgcattc gatactactc ggctagagta caggagaggg aagcggaatt cctagtgtag
      661 cggtgaaatg cgtagagatc aggagggaca ccagtggcga aggcggcttc ttggactgat
      721 actgacgctg aggtacgaaa gcgtggggag caaacaggat tagatgccct ggtagtccac
      781 gccctaaacg atgagaacta gatgttgggg gaatttaatc ccttagtatc gcagctaacg
      841 cgctaagttc tccgcctggg gagtacggcc gcaaggttaa aactcaaatg aattgacggg
      901 ggcccgcaca agcggtggag catgtggttt aattcgatgc aacgcgaaga accttacctg
      961 gccttgacat ccttggaacc tcgcagagat gtgagggtgc cttcgggaac cgagagacag
     1021 gtgctgcatg gctgtcgtca gctcgtgtcg tgagatgttg ggttaagtcc cgcaacgagc
     1081 gcaaccctta tccctagttg ccagcgattc ggtcgggaac tctagagaga ctgccggtga
     1141 caaaccggag gaaggtgggg acgacgtcaa gtcatcatgg cccttacggc cagggctaca
     1201 cacgtgctac aatgggtag tacaaagggt tgcgaacccg cgaggggtg ctaatctcac
     1261 aaaactactc gtagtccgga ttggagtctg caactcgact ccatgaagtt ggaatcgcta
     1321 gtaatcgcgg atctgcatgt cgcggtgaat acgttcccgg gccttgtaca caccgcccgt
     1381 cacaccatgg gagtgggctg taccagaagt aggtagtcta accgcaaggg ggacgcttac
     1441 cacggtatgg ttcatgactg gggtg
//
```

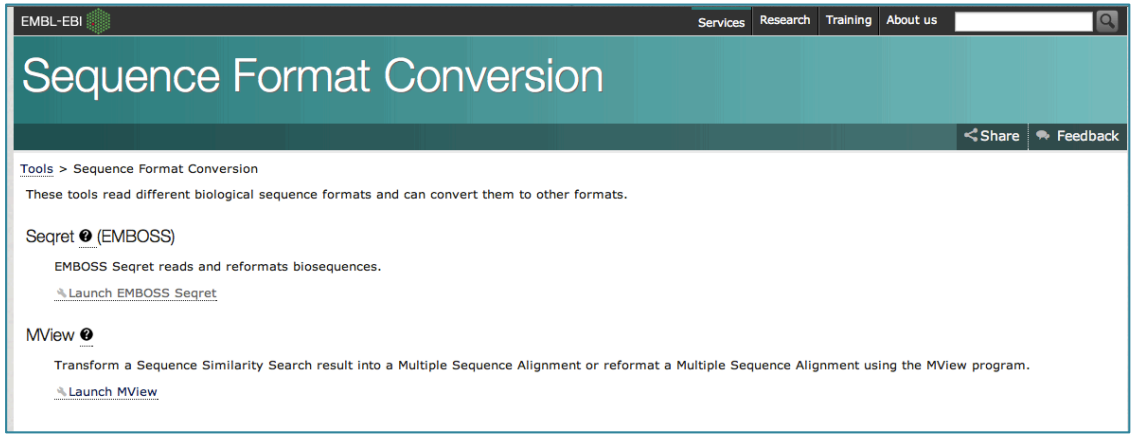
```
>gi|359952839|gb|JN793554.1| Uncultured Beggiatoa sp. clone WF4571 16S ribosomal
RNA gene, partial sequence
ATTGAACGCTGGCGGCATGCTTAATACATGCAAGTCGAACGGTAACATGCCCTTCGGGGTGATGACGAGT
GGCGGACGGGTGAGTAATGCATAGGAATCTACCCAGTAGACGGGGAACAACTTGGGGAAACTCAAGCTAA
TACCGCGTAAGCCTTACGGGGGATAGCGGGGGACTTTTTAGGAAGCCTCGCGATATTGGATGAGCCTATG
CCGGATTAGCTAGTTGGTGGGGTAAAAGCTTACCAAGGCTACGATCCGTAGCTGGTCTGAGAGGACGATC
AGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATATTGGACAATGGG
CGAAAGCTTGATCCAGCAATGCCGCGTGTGTGAAGAAGGCCTGCGGGTTGTAAAGCACTTTCAGTTGGGA
AGAAAAGCTTTAGGTTAATACCCTGAAGTATTGACGTTACCAACAGAAGAAGCACCGGCTAACTCTGTGC
CAGCAGCCGCGGTAATACAGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGG
TTGAGTCAGTCGGATGTGAAAGCCCAAGGCTTAACCTTGGAACTGCATTCGATACTACTCGGCTAGAGTA
CAGGAGAGGGAAGCGGAATTCCTAGTGTAGCGGTGAAATGCGTAGAGATCAGGAGGGACACCAGTGGCGA
AGGCGGCTTCTTGGACTGATACTGACGCTGAGGTACGAAAGCGTGGGGAGCAAACAGGATTAGATGCCCT
GGTAGTCCACGCCCTAAACGATGAGAACTAGATGTTGGGGGAATTTAATCCCTTAGTATCGCAGCTAACG
CGCTAAGTTCTCCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACA
AGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTGGCCTTGACATCCTTGGAACC
TCGCAGAGATGTGAGGGTGCCTTCGGGAACCGAGAGACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCG
TGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATCCCTAGTTGCCAGCGATTCGGTCGGGAAC
TCTAGAGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGACGACGTCAAGTCATCATGGCCCTTACGGC
CAGGGCTACACACGTGCTACAATGGGGTAGTACAAAGGGTTGCGAACCCGCGAGGGGGTGCTAATCTCAC
AAAACTACTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTTGGAATCGCTAGTAATCGCGG
ATCTGCATGTCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGCTG
TACCAGAAGTAGGTAGTCTAACCGCAAGGGGGACGCTTACCACGGTATGGTTCATGACTGGGGTG
```

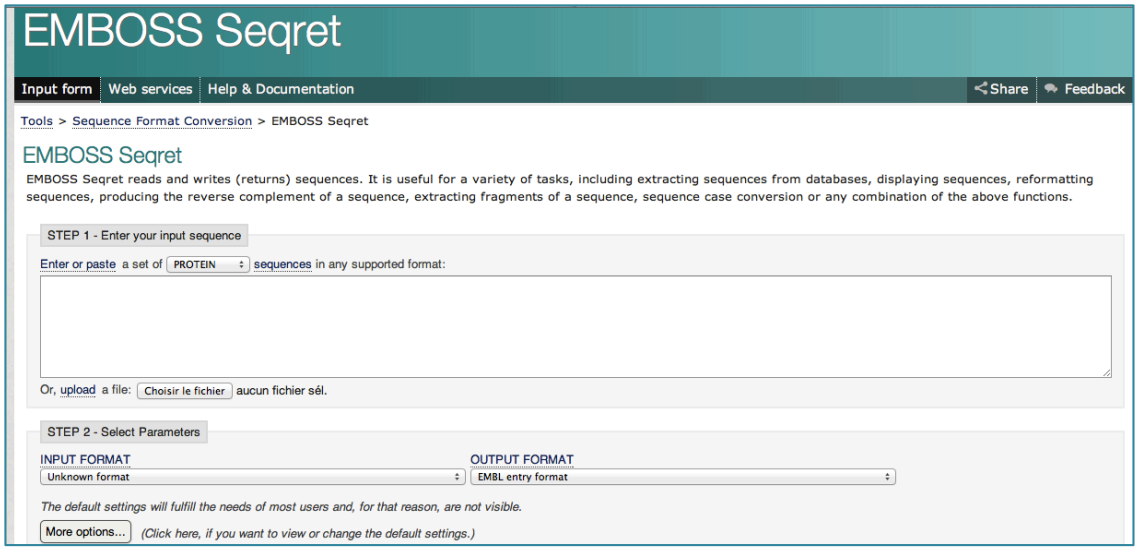
Exemple Genbank



Conversion de Format

- http://www.ebi.ac.uk/Tools/sfc/





FORMAT FASTA

➢ Format commun de manipulation des données :
le format FASTA (Fast – alignment)

- Les bases nucléotidiques ne référencient que des **monobrins d'ADN** (même si la séquence soumise est de l'ADN bicaténaire ou de l'ARN)
  → **la séquence est toujours dans le sens 5'P – 3'OH**

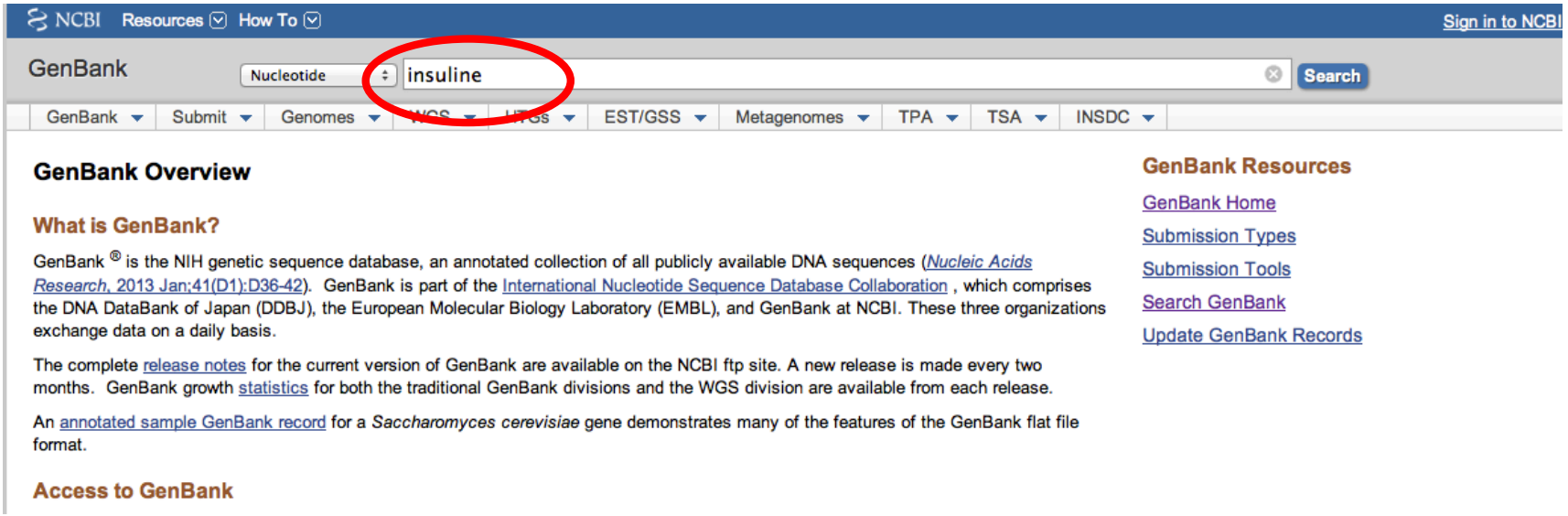
- **Les séquences nucléotidiques** selon le degré de précision de l'enregistrement seront écrites le plus souvent avec **A,T, C et G** et/ou avec **R,Y** (base pu**R**ique A et G / base p**Y**rimidique C et T) et/ou **K,M** (base **K**eto G et T / base a**M**ino A et C).

- **Les bases protéiques** sont référencées :
  → **avec la séquence dans le sens N vers C terminal**
  → **avec le symboles d'acides aminés à 1 lettre**

Un autre exemple Genbank

Le nom d'une protéine ou d'un gène…
exemple insuline

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

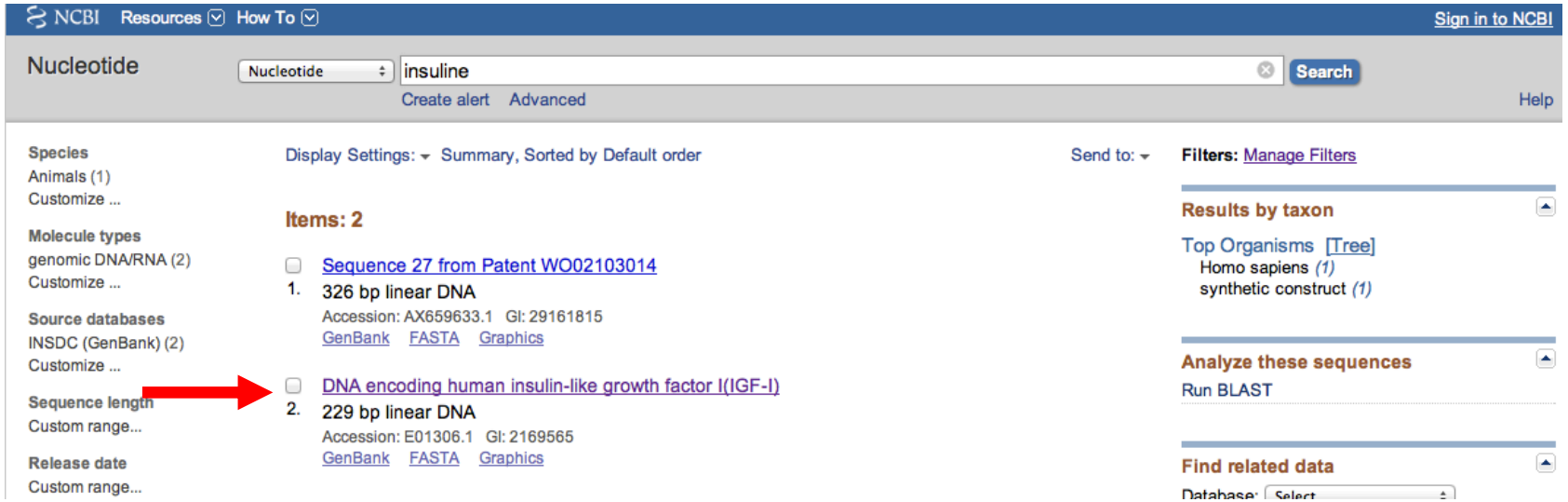


Un autre exemple Genbank

Le nom d'une protéine ou d'un gène…

exemple insuline

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

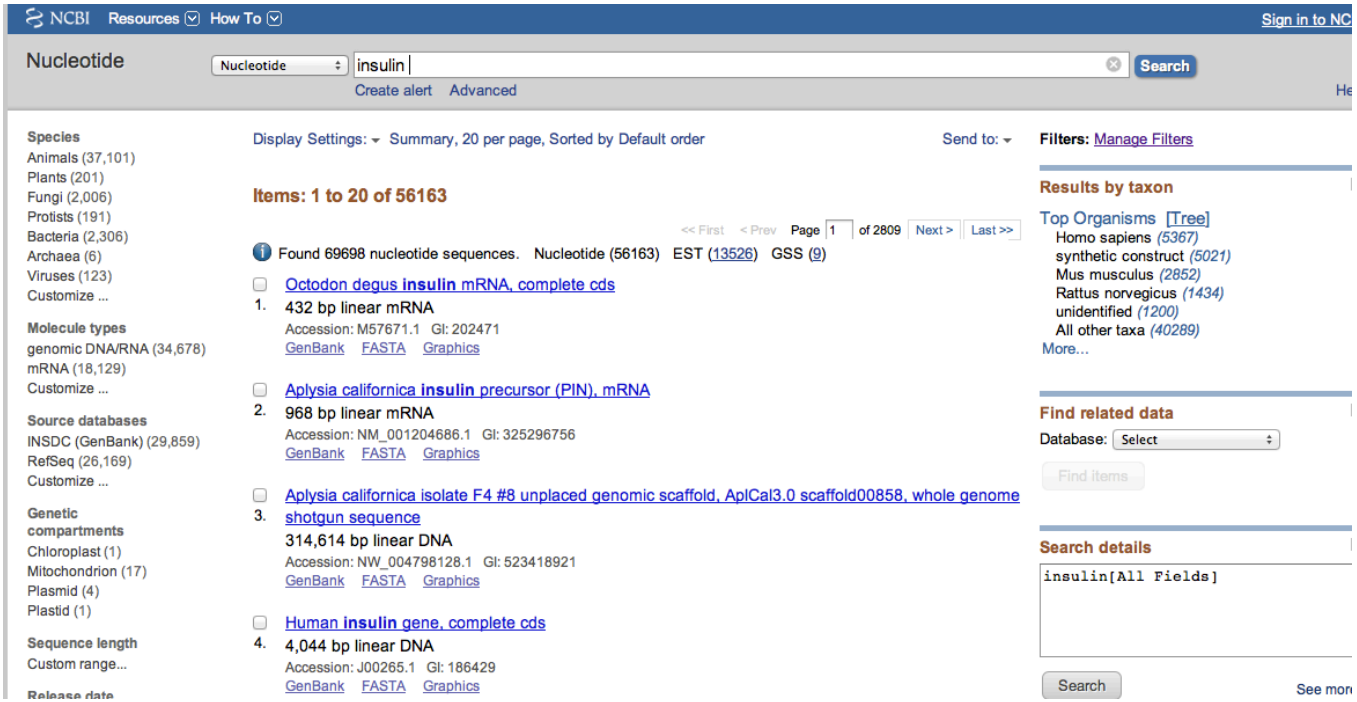


Un autre exemple Genbank

Le nom d'une protéine ou d'un gène…

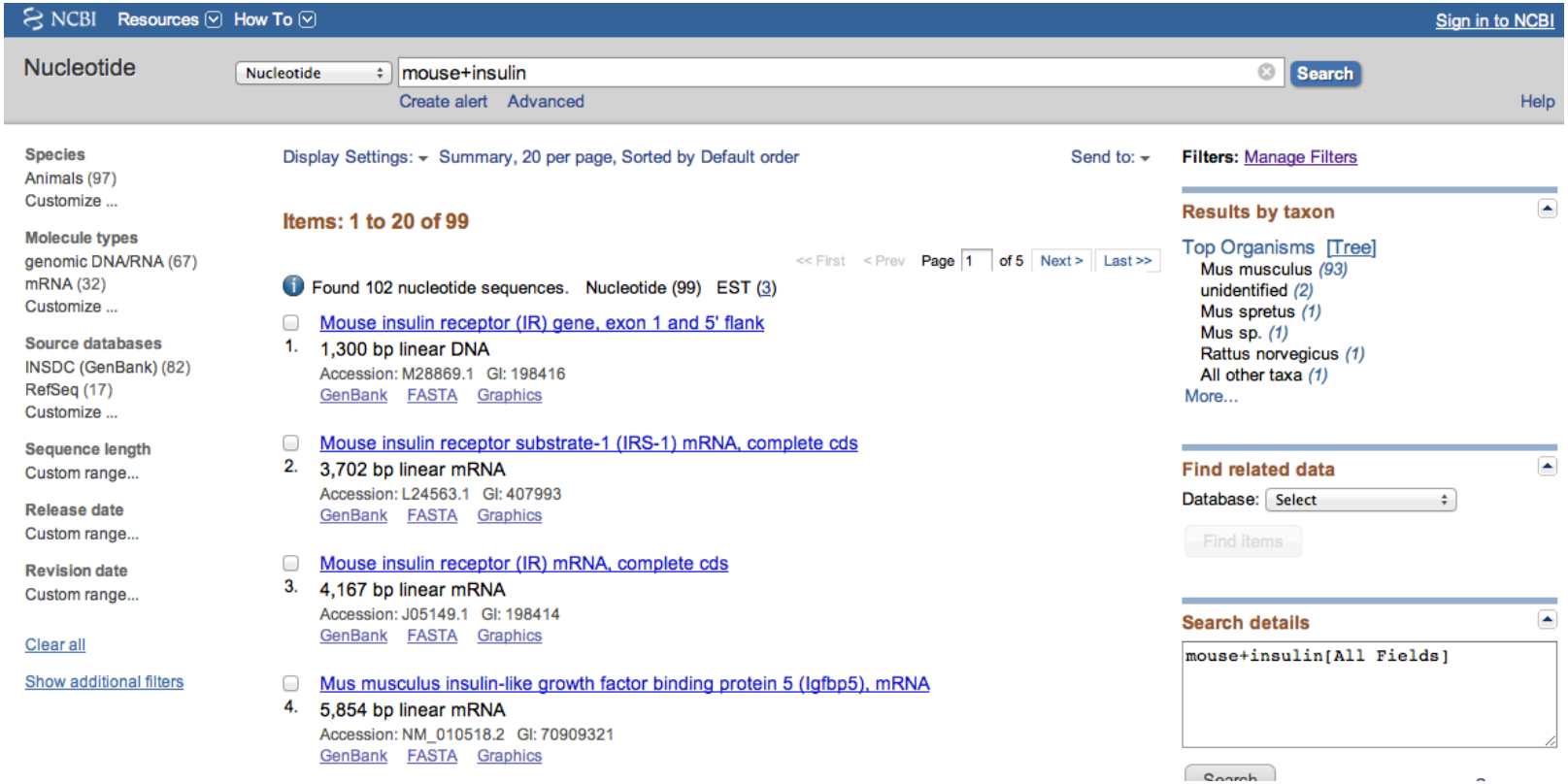
exemple insulin

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



Un autre exemple Genbank

Le nom d'une protéine ou gène…
exemple mouse +insulin



Base de données: Genbank

Il ne faut pas hésiter a « naviguer » dans les portails de base de données…

Ils concentrent beaucoup d'information utiles!!!!

MAIS ATTENTION!!!!

- Méfiez-vous des banques de données :
- Les données ne sont pas toujours fiables
- La mise à jour n'est pas toujours récente
- La réalité mathématique n'est pas la réalité biologique :
- Les ordinateurs ne font pas de biologie, ils calculent … vite !

Comment s'assurer de la qualité de l'information ?

Autorité →

- **Source de l'information, auteurs,**
- **statut, …**

Péremption→

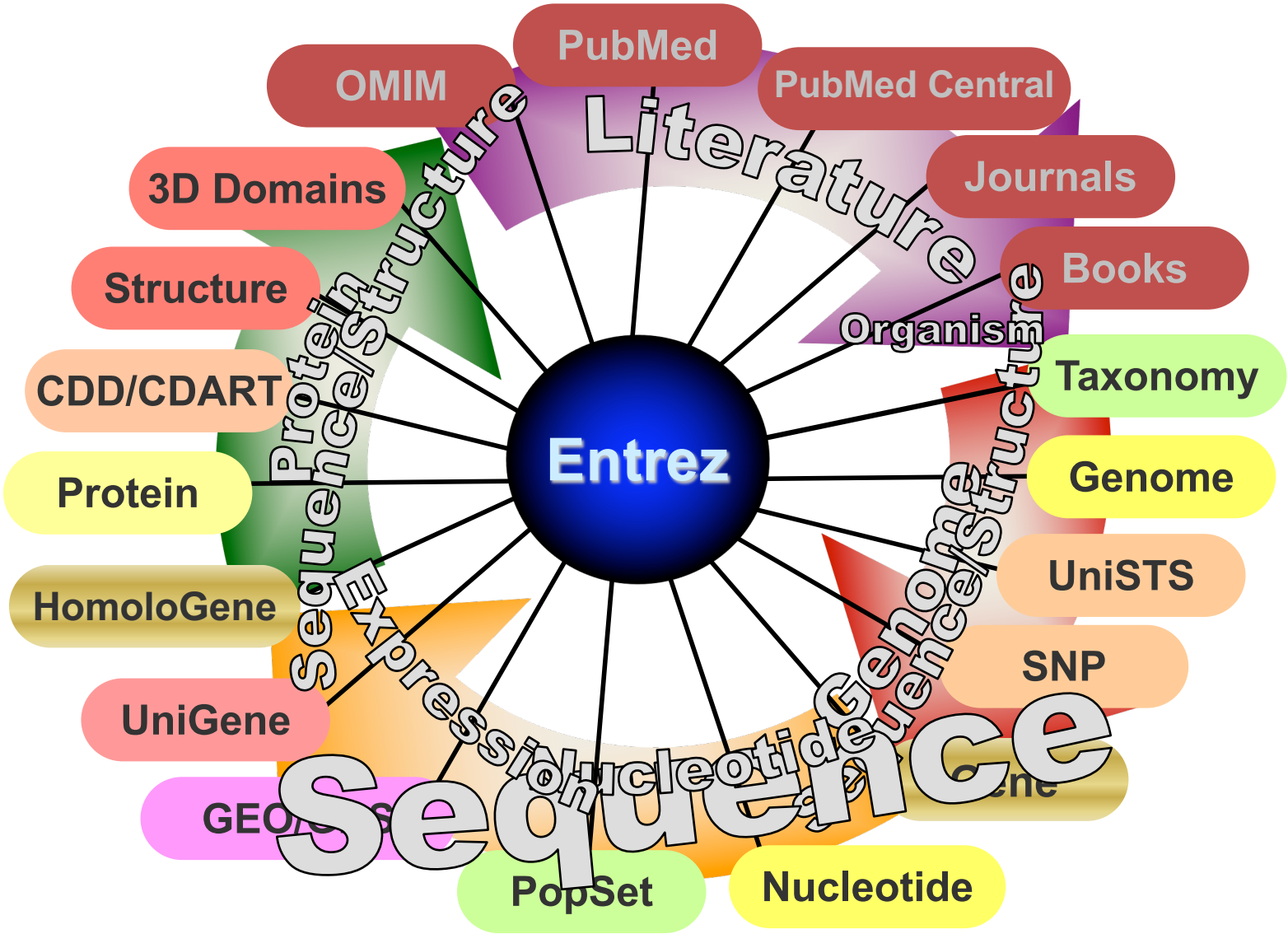
- **Date de création, de mise à jour, …**
- **Attention, ce qui est validé un jour peut être démenti par la suite !**

Transparence →

- **Documentation disponible**

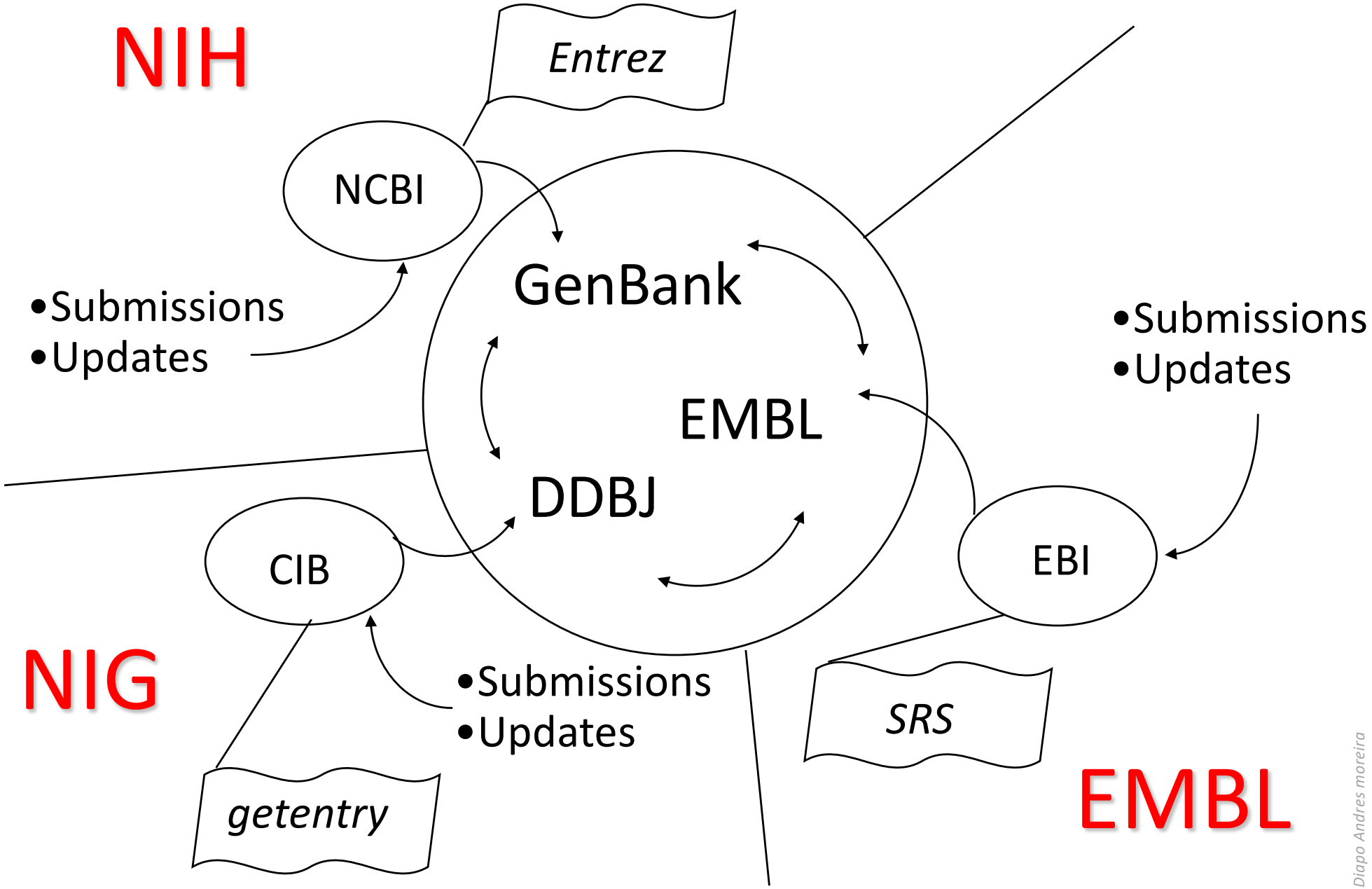
Règles valables aussi bien pour des bases de données, que pour un logiciel, un site web, …

NCBI


PubMed
OMIM
PubMed Central
Literature
Journals
3D Domains
Books
Structure
Protein Sequence/Structure
Organism
CDD/CDART
Taxonomy
Entrez
Protein
Genome
Genome Sequence/Structure
UniSTS
HomoloGene
SNP
UniGene
Expression
Nucleotide Sequence
PopSet
Nucleotide

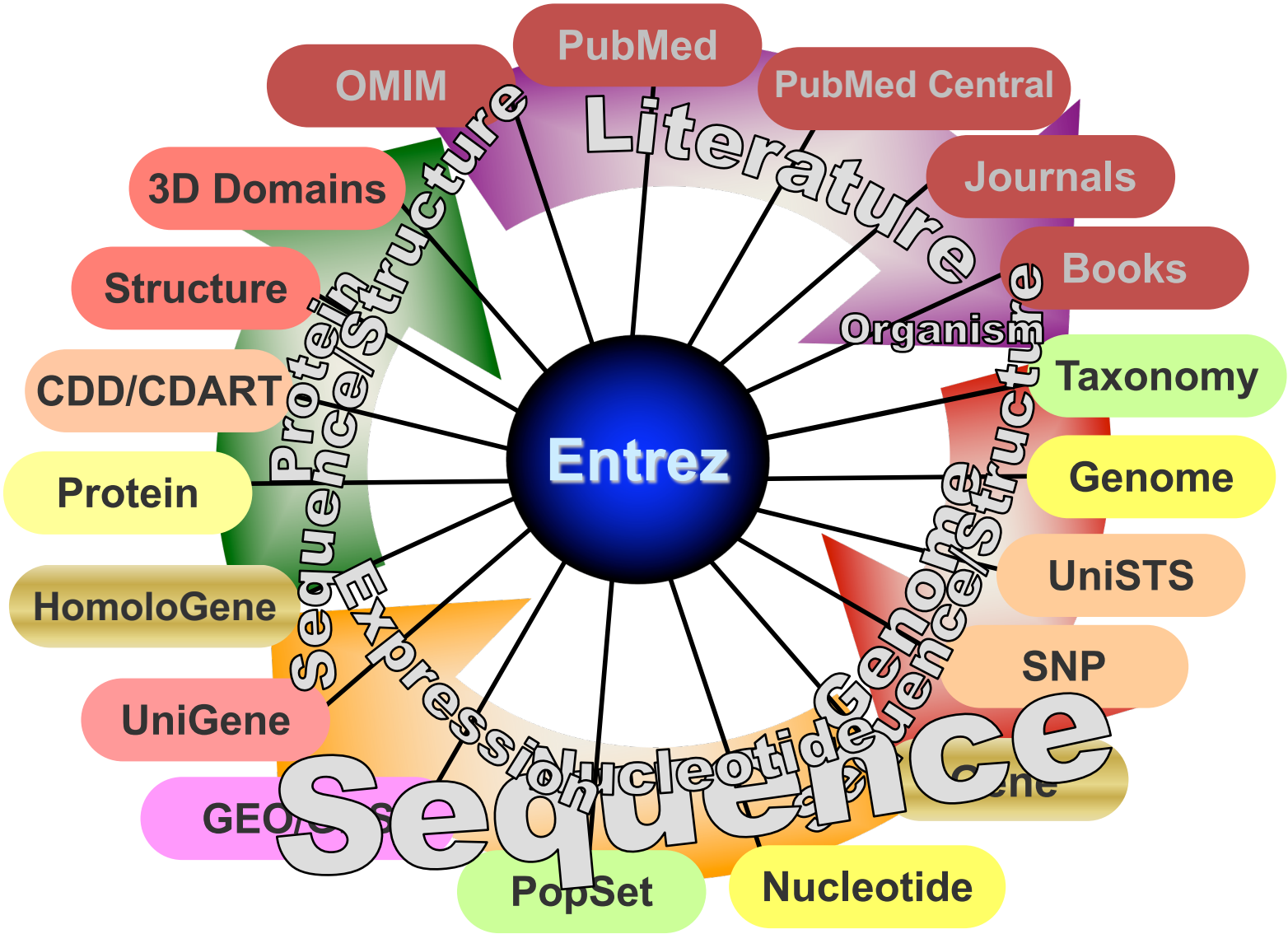
Principes de la comparaison de séquences

Bases de données



Diapo Andres moreira

NCBI


PubMed
OMIM
PubMed Central
Literature
Journals
3D Domains
Books
Structure
Protein Sequence/Structure
Organism
CDD/CDART
Taxonomy
Entrez
Protein
Genome
Genome Sequence/Structure
UniSTS
HomoloGene
SNP
Expression
Nucleotide Sequence
UniGene
GEO
PopSet
Nucleotide

Principes de la comparaison des séquences

- **Pourquoi comparer des séquences ?**
- **Principes de base et initiation de méthodes de l'alignement pour la comparaison des séquences**
- **Initiation à l'utilisation de quelques logiciels bio-informatiques d'alignement**

Concepts de recherche de similarité de séquences

La prémisse:

- **Une séquence n'est pas informative par elle même; elle doit être analysée par des méthodes comparatives contre des bases de données existantes pour développer des hypothèses concernant leur relation et leur fonction.**

Comparaison de séquences biologiques: Pour quoi faire?

- **Cette comparaison est nécessaire dans différents types d'études :**
  - **Localiser un gène sur un génome**
  - **Identification de gènes homologues**
  - **Recherche de similarité de séquences**
  - **Recherche de domaines ou motifs conservées: Identification des résidus important pour la structure ou la fonction**
  - **Recherche de contraintes fonctionnelles communes à un ensemble de gènes ou de protéines.**

Comparaison de séquences biologiques: Pour quoi faire?

- Prédiction de structure (ARN, protéine) et/ou fonction
- Étude sur la variabilité entre séquences.
- Reconstitution des relations évolutives entre séquences.
- Choix d'amorces PCR
- Construction de contigs (séquençage)

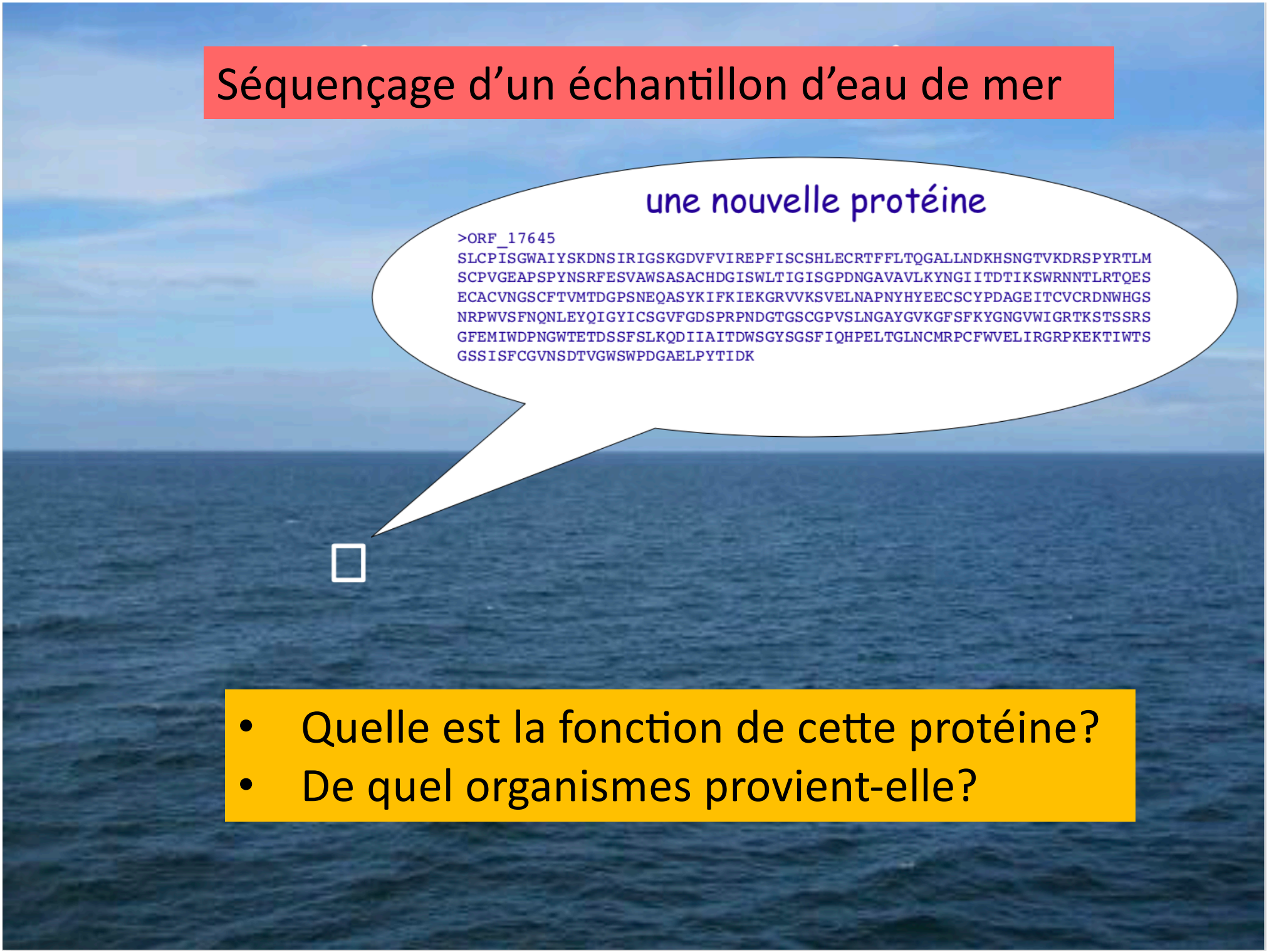
- En fait, la liste des question est illimité….

La déduction par homologie, ou le «dogme central» de la bioinformatique

L' évolution laisse une trace parfaitement visible lorsque on compare des séquences de gènes

- Les gènes des organismes modernes sont issus de remaniement de gènes ancestraux.
- Evolution des gènes=mutations, insertions, délétion
- Les régions fonctionnelles des gènes (sites catalytique, de fixation, etc.) sont soumises à sélection. Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.
- Les régions non fonctionnelles ne subissent aucune pression de sélection et divergent rapidement à mesure que s'accumulent les mutations.

Séquençage d'un échantillon d'eau de mer



- Quelle est la fonction de cette protéine?
- De quel organismes provient-elle?

EXAMPLE AVEC UN MOT

« FINESTRA »??

Alignements des mots avec « Finestra »

Certaines lettres sont identiques

```
finestra
|||:::
affinité
```

```
finestra
|||||
estragon
```

```
finestra
|||||
fines
```

```
finestra
|:||:::
fenêtre
```

Alignements des mots avec « Finestra »

Certaines lettres sont différentes

Alignements des mots avec « Finestra »

Certaines lettres sont différentes

Certaines lettres n'ont pas de correspondant

Alignements des mots avec « Finestra »

Plusieurs alternatives

```
i <> e      finestra
s <> t      |:||:::
t <> r      fenêtre
r <> e
```

ou

```
            finestra
i <> e      |:||  ||:
a <> e      fenê-tre
```

Alignements des mots avec « Finestra »

Modifications crédibles vs. improbables

- consonne /voyelle: **r <> a** 
- voyelle/voyelle: **e <> a** 

Modifications qui gardent la nature de la lettre sont plus probables/crédibles

Alignements des mots avec « Finestra »

Plus il y a de lettres orphelines, moins la correspondance est probable

Alignements des mots avec « Finestra »

5 matchs
0 modifications
6 orphelins

```
finestra
|||:::
affinité
```

3 matchs
3 modifications (2+1)
4 orphelins

```
finestra
|||||
estragon
```

5 matchs
0 modifications
3 orphelins

```
finestra
|||||
fines
```

Comment trancher?

```
finestra
|:|| ||:
fenê tre
```

5 matchs
2 modifications (2+0)
1 orphelin

Alignements des mots avec « Finestra »

SCORE

- plus il y a de matchs, **plus** le score est élevé
- plus il y a de modifications, **moins** le score est élevé
  - modifications conservatrices: -
  - modifications non-conservatrices: ---
- plus il y a d'orphelins, **moins** le score est élevé

$$S = (+1) \times \text{matchs} + (-0.5) \times MC + (-1) \times MNC + (-1) \times \text{orphelins}$$

Alignements des mots avec « Finestra »

SCORE

$$S = (+1) \times \text{matchs} + (-0.5) \times MC + (-1) \times MNC + (-1) \times \text{orphelins}$$

```
finestra
|||:::
affinité
```

S= -3

```
finestra
|||||
estragon
```

S= -1

```
finestra
|||||
fines
```

S= 2

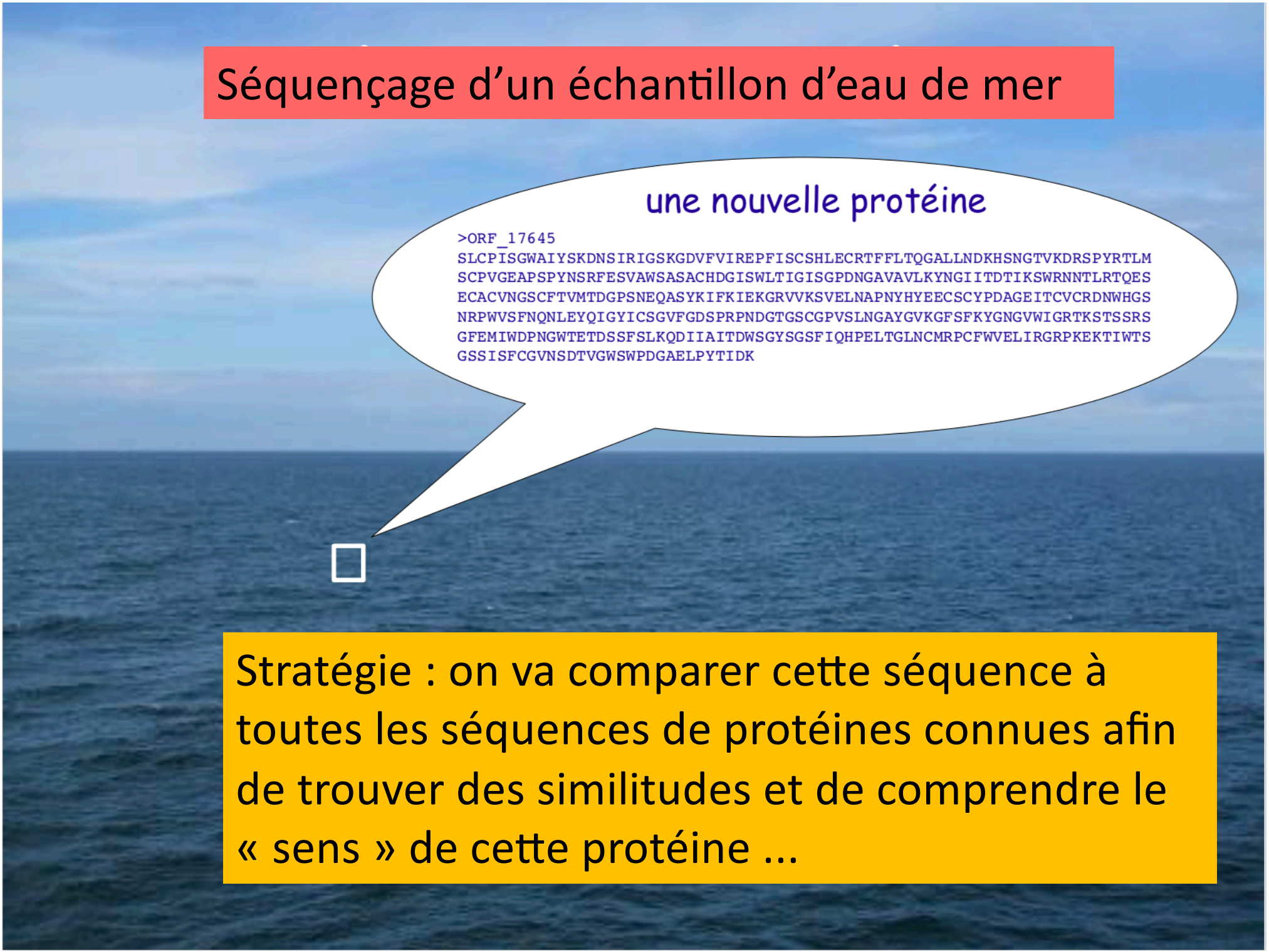
```
finestra
|:|| ||:
fenê tre
```

S= 3

CONCLUSION SUR UN ALIGNEMENT

- Recherche de similarités pour trouver une signification
  - **similarité orthographique ≈ similarité de sens**
- Nécessité d'une **origine commune** !!
- Certaines substitutions sont crédibles, d'autres moins
- certaines lettres sont insérées, d'autres disparaissent

Séquençage d'un échantillon d'eau de mer



Stratégie : on va comparer cette séquence à toutes les séquences de protéines connues afin de trouver des similitudes et de comprendre le « sens » de cette protéine ...

Comparaisons de séquences biologiques

- comparer 2 séquences, c'est chercher des **similarités** qui

- (1) reflètent des fonctions semblables
- (2) sont la trace d'une origine évolutive commune **= homologie**

Comparaisons de séquences: homologie

Homologie: désigne un lien évolutif entre deux traits

Il y a deux types de homologie, les séquences peuvent être:

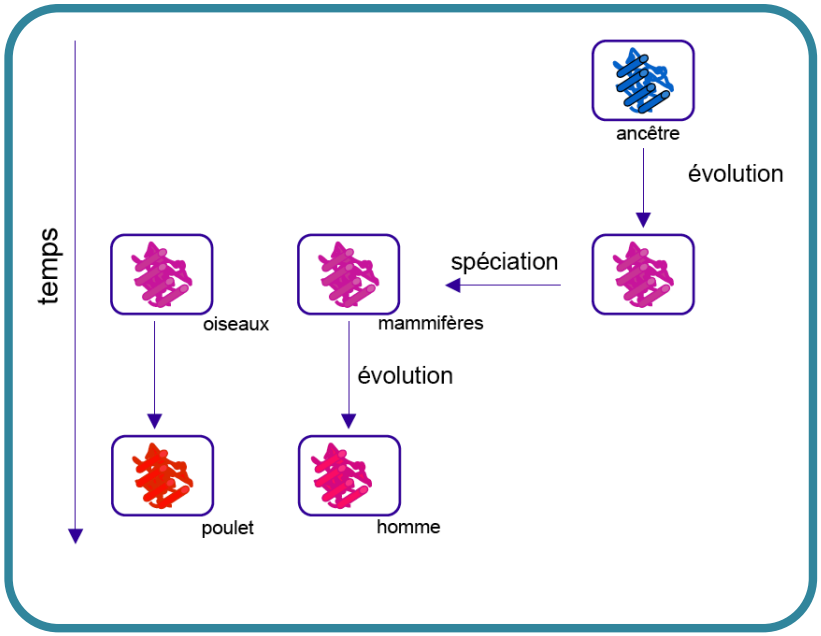
Orthologues ou Paralogues

- **Orthologues: séquences homologues sur des espèces différentes, que ont divergées par spéciation.**

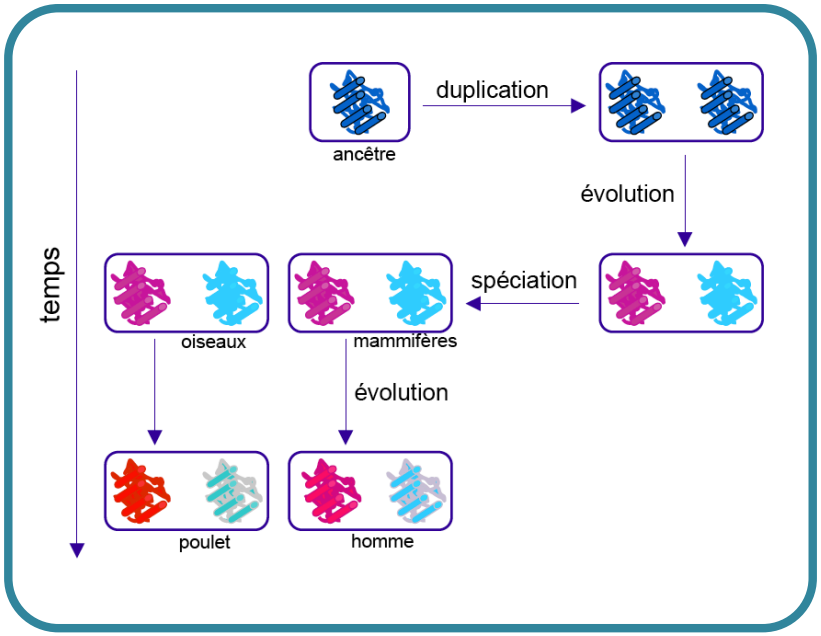
- **Paralogues: séquences homologues sur une même espèce, qui ont divergées après la duplication d'un gène.**

Evolution d'un gène

Spéciation

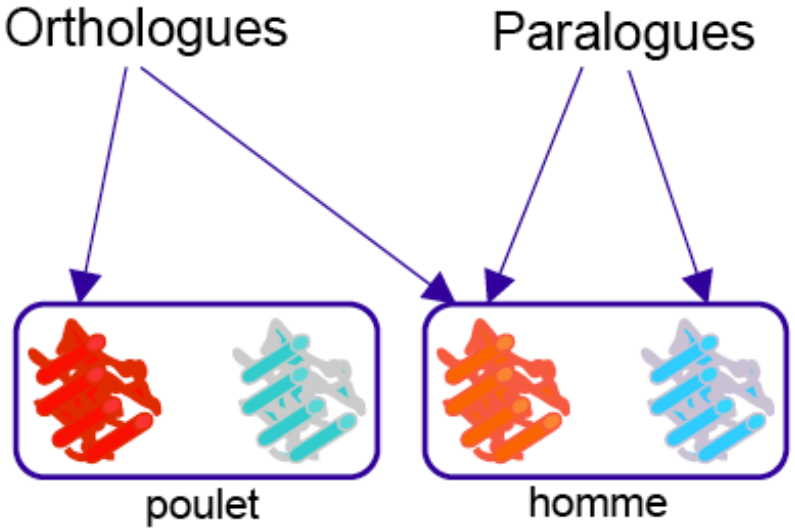


Apparition de nouveaux gènes par duplication



Paralogues ou Orthologues?

- **Homologues:** gènes provenant d'un ancêtre commun



Orthologues:
gènes homologues issus de la spéciation

Paralogues:
gènes homologues issus d'un phénomène de duplication

L'homologie de séquence

En bioinformatique: Homologie = parenté = ancêtre commun

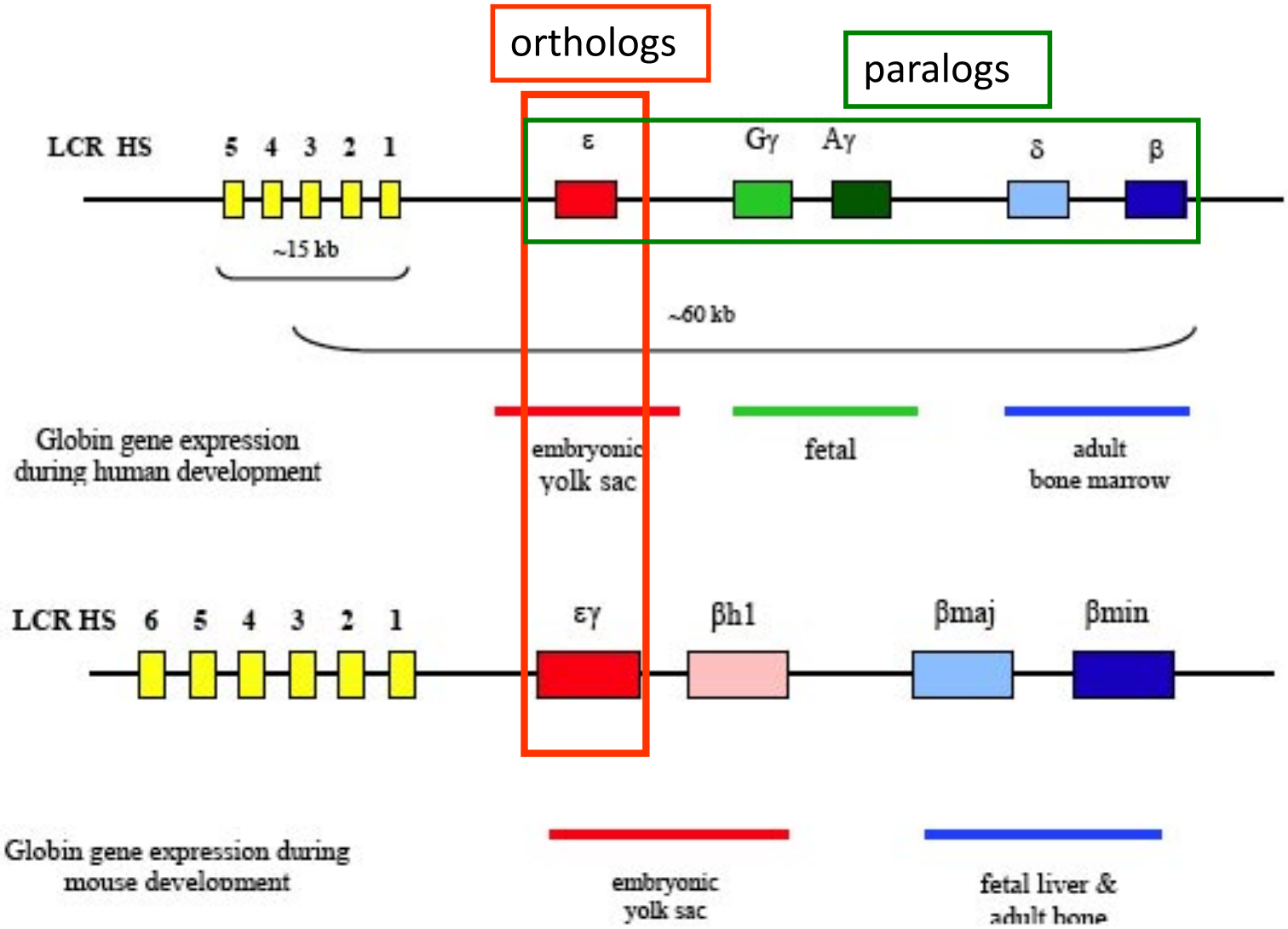
Le bras humain est homologue à l'aile de l'oiseau
Le bras humain n'est pas homologue à l'aile de la mouche

On est homologue ou on ne l'est pas

Donc on ne dit pas: «très homologue», «faible homologie», «28% d'homologie», etc.

Pour une notion quantitative… on parle de similitude ("très similaire", etc.) ou d'identité (28% d'identité)

ORTHOLOGS AND PARALOGS INTO LOCUS ß FROM GLOBINS



Fonction et homologie

- **L'homologie n'implique pas même fonction**
- **Des orthologues rapprochés (p. ex. homme/souris) ont le plus souvent la même fonction dans l'organisme.**
- **Des orthologues distants (p. ex. homme/mouche) ont plus rarement le même rôle *phénotypique*, mais peuvent exercer le même rôle dans une *voie* donnée.**
- **Les paralogues acquièrent rapidement des fonctions différentes**

Homologie/ Similarité

- **Homologie→ Existence d'une origine évolutive commune**
  - **Orthologues**
  - **Paralogues**
- **Binaire: oui/non**

2 séquences sont ou ne sont pas homologues

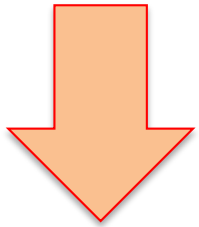
- **Similarité→ Mesure de la ressemblance entre 2 séquences**
- **Exprimée en % d'identité**
- **(30%, 50%, 80%)**

2 séquences sont plus ou moins similaires

NOTION QUANTITATIVE ENTRE SEQUENCES

Comment détecter une homologie?

L'alignement des séquences est la principale méthode de comparaison. Elle permet d'identifier des régions conservées. On en déduit l'homologie



Comparer des séquences serait relativement simple si elles avaient toutes la même longueur. Comme ce n'est pas le cas, il faut les aligner, c'est à dire trouver où se trouvent les insertions et délétions, représentées par des « indels » (« gaps »)

Principe de la comparaison de séquences

La comparaison de séquences est l'outil central en bioinformatique

Repose sur des calculs matriciels ou des algorithmes complexes qui rendent des résultats sous forme de données statistiques (% match, score, e-value…)

Logiciel d'alignement le plus connu = BLAST (Basic Local Alignment Search Tool)

Comparaison de séquences: Démarche globale

Alignement de séquences

↓

Score de similitude

↓

homologie

↓

Identification, prédiction de structure de propriétés, de fonction

Type d'alignement

- **Alignement 2 à 2 (Pairwise alignement):**
  - **p.e: recherche de similarité dans une banque→ fasta, BLAST**
- **Alignement multiple:**
  - **p.e: famille de proteines (ClustalO, MUSCLE)**
- **Alignement global:**
  - **Sur la totalité de la longueur d'une sequence**
- **Alignement local:**
  - **Alignement de la ou des régions les plus fortement conservées**

Alignement

- **Aligner 2 séquences→**

C'est rechercher le maximum d'appariements entre les lettres qui les composent (nucléotides ou aa) avec le minimum de mésappariements et de gaps.

La recherche de similitude entre séquences nécessite la détermination d'un score de similarité

Comment trouver le meilleur alignement ?

Selon ce concept, le bon alignement est celui qui minimise les opérations à réaliser pour passer d'un séquence à l'autre.

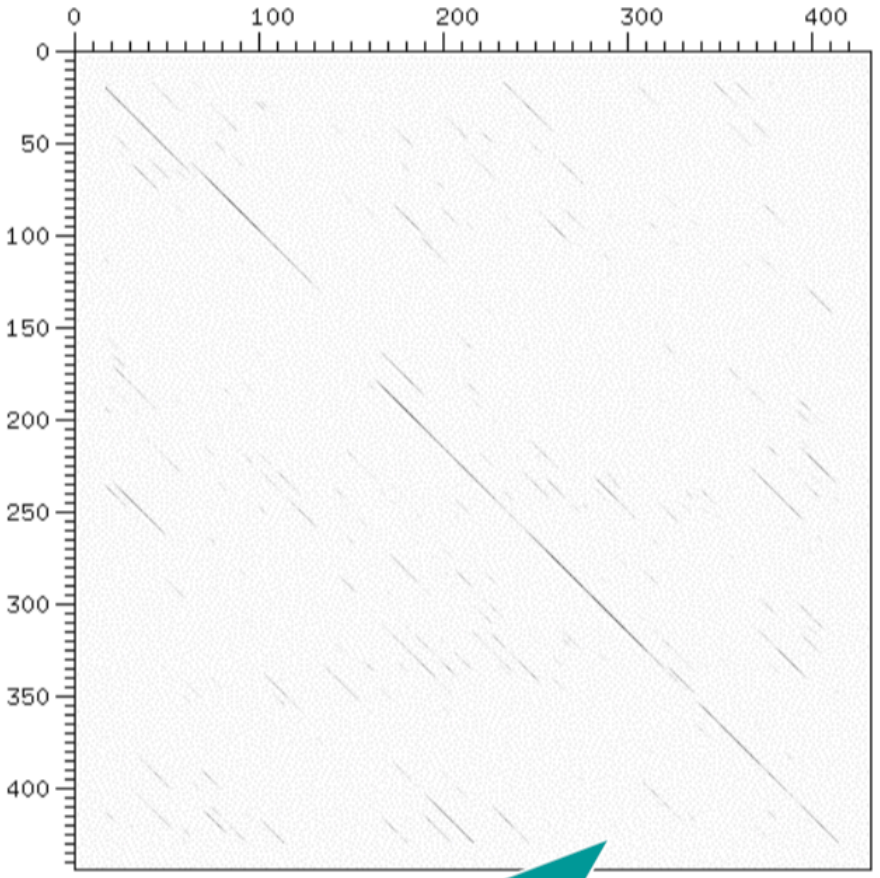
- **Opérations: conservation, remplacement/mutation, délétion, insertion. Une pénalité peut être affectée à chaque opération et la distance finale entre les deux séquences (distance d'édition) est la somme de ces pénalités.**

- **Le nombre d'alignements possibles est trop élevé: on ne peut pas les essayer tous pour trouver celui qui minimise la distance (ou maximise la ressemblance).**

Les « dot plots »?

Le dot plot est une représentation graphique simple des résidus identiques entre les deux séquences.

- Deux séquences à comparer sont représentées.
- On dessine ensuite un point dans la matrice lorsque les deux positions correspondantes sont identiques. Lorsque des régions se ressemblent, on voit apparaitre une diagonale.
- Les décalages entre les diagonales correspondent à des insertions ou délétions. Plusieurs diagonales parallèles indiquent une répétition.



Alignement: trouver le meilleur chemin dans ce graphe

Alignement: représentation

- **Les résidus (nucléotides, acides-aminés) sont superposés de façon à maximiser la similarité entre les séquences.**

```
G T T A A G G C G – G G A A A
T T T – – – G C C A G G A C A
  * *       * *     * * *   *
```

- **Il existe deux sortes de mutations :**
  - **Substitution (*mismatches*)**
  - **Insertion et Délétion (indels ou *gaps*).**

Quel est le bon alignement ?

```
G T T A C G A
G T T - G G A
* * *     * *
```

OU

```
G T T A C G A
G T T G - G A
* * *     * *
```

OU

```
G T T A C - G A
G T T - - G G A
* * *       * *
```

- **Pour le biologiste, généralement, le bon alignement est celui qui représente le scénario évolutif le plus probable**

→ **Qui minimise le nombre de changement évolutifs**

→ **Qui implique les changements évolutifs les plus probables**

=> Calcul d'un score pour évaluer la qualité de l'alignement

Comment calculer le meilleur alignement?

- **On utilise →**
  - **une matrice de substitution**
  - **un modèle de gap**
  - **une fonction de score (somme des paires)**

Le meilleur alignement est l'optimum pour la somme des paires

Détermination d'un score

- **Utilisation de matrice de substitution**

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

→ **Le score élémentaire (se) = la valeur donnée directement dans la matrice**

Calcul score global→ la somme des scores élémentaires **Score = Σ se**

- **Introduction de gap (avec pénalité)**

Pénalité pour l'insertion d'un gap (x)

Pénalité pour l'extension d'un gap (y)

P = coût global du gap de longueur L

P = x + yL

Score = Σ se - Σ P

Fonction de score de similarité

```
G T T A A G G C G – G G A A A
G T T – – – G C G A G G A C A
* * *       * * *   * * *   *
```

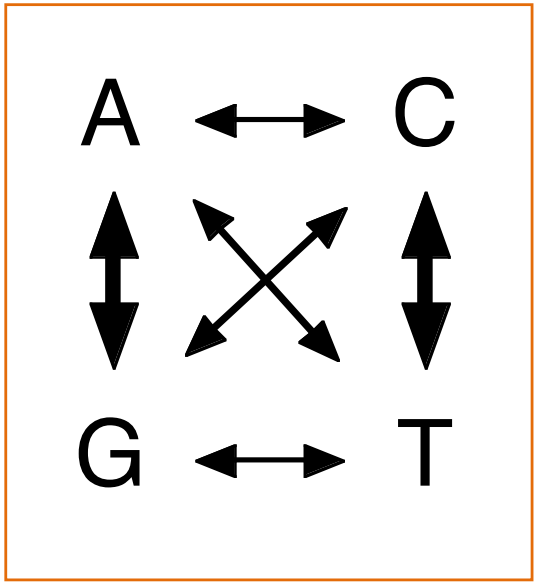
Score alignement = Score Identités - Score Différences

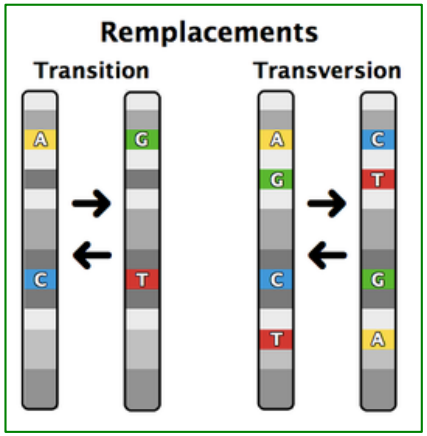
$$\text{Score} = \Sigma \text{ se} - \Sigma \text{ P}$$

Identité = 1
Mismatch = 0
Gap = -1

Score = 10 - 4 = 6

Modèle d'évolution (ADN)





- **Transition: A <-> G T <-> C**
- **Transversions : autres substitutions**
- **p(transition) > p(transversion)**

```
G T T A C G A       G T T A C G A
G T T - G G A   <   G T T G - G A
* * *     * *       * * * .   * *
```

Matrice de substitution (ADN)

	A	C	G	T
A	1	0	0.5	0
C	0	1	0	0.5
G	0.5	0	1	0
T	0	0.5	0	1

Exemple:

- **(identique)= 1**
- **(Transition)= 0,5**
- **(Transversion)= 0**

Gap = -1

```
G  T  T  A  C  G  A          G  T  T  A  C  G  A
G  T  T  -  G  G  A          G  T  T  G  -  G  A
-------------------          -------------------
1  1  1 -1  0  1  1          1  1  1 .5 -1  1  1
```

score = 4

score = 4.5

Diferents types de matrices

- **Matrices basées sur des comparaisons par paires utilisant des alignements locaux :**

  → BLOSUM (Henikoff et Henikoff, 1992)

- **Matrices basées sur des arbres construits en utilisant le maximum de parcimonie (alignement globaux) :**

  → PAM (Dayhoff *et al.*, 1978)

  → JTT (Jones *et al.*, 1992)

- **Matrices basées sur des arbres construits en utilisant le maximum de vraisemblance :**

  → WAG (Whelan et Goldman, 2001)

Matrices de substitution

Les matrices BLOSUM (*BLOcks Substitution Matrix*) sont déduites d'alignements de fragments (blocks) de protéines très éloignées.

BLOSUM 62

matrice construite en comparant des protéines orthologues ayant en moyenne 62% de similarité
il existe des matrices → BLOSUM30, BLOSUM80,...

Ces matrices sont bien adaptées aux recherches de séquences dans les banques de données (BLAST)

Matrices de substitution

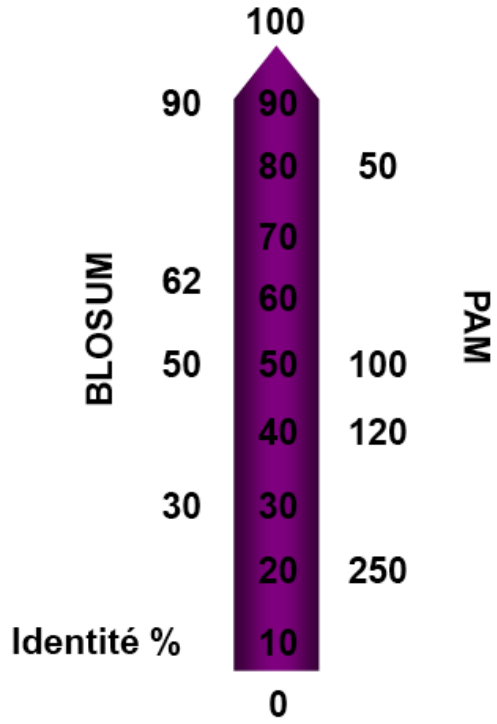
- **PAM250**

Déduite d'alignements globaux de familles de protéines très proches
matrice construite en comparant des protéines orthologues ayant subi en moyenne 250 mutations/100 acides aminés
→ il existe des matrices PAM100, PAM150,...

Choix d'une matrice??

Choix d'une matrice

Pas de consensus mais ce qui est généralement reconnu:
Meilleurs résultats avec les matrices utilisant des modèles d'évolution
→ BLOSUM globalement meilleures que PAM.



Degré de similarité des séquences.

BLOSUM 80	**BLOSUM 62**	BLOSUM 45
PAM 1	PAM 120	PAM 250
Moins divergent	⟷	**Plus divergent**

Il est recommandé d'expérimenter !

Comparaison de séquences

En pratique, plus le score d'alignement est élevé, plus les séquences sont similaires et présenteront des propriétés et des fonctions proches.

→ plus de 70% de similarité permettent d'affirmer qu'il y a homologie

Attention pas confondre les concept de similitude, identité et homologie

La similarité est une quantité qui se mesure en % d'identité, identité elle même définie comme une ressemblance parfaite entre deux séquences. L'homologie quand à elle est une propriété de séquences qui a une connotation évolutive.

Le cas des acide aminés

- Plus difficile à modéliser que celui des nucléotides :

Un acide aminé peut être remplacé par un autre de différentes façons (code génétique):

`Asp (GAC) → Tyr (UAC, UAU) 1 ou 2 mutations`

- Le nombre de substitutions requises pour passer d'un acide aminé à un autre diffère.

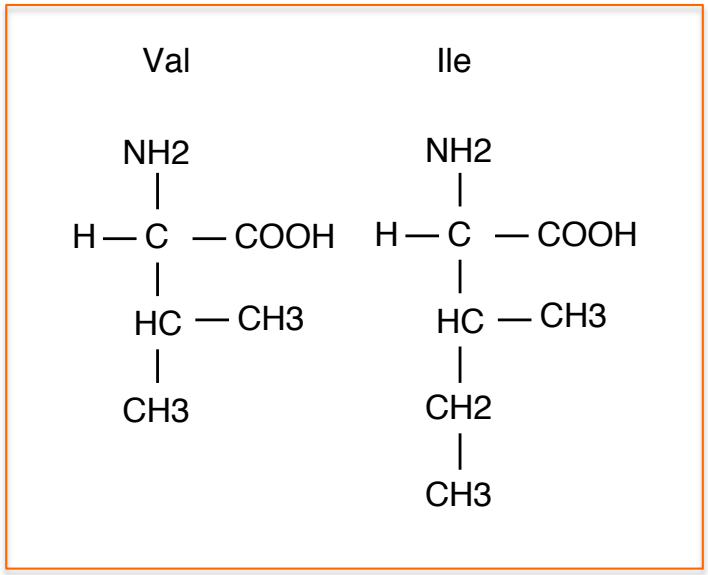
```
Asp (GAC, GAU)→ Tyr (UAC, UAU)   1 mutation
Asp (GAC, GAU)→ Cys (UGC, UGU)   2 mutations
Asp (GAC, GAU)→ Trp (UGG)        3 mutations
```

Le cas des acide aminés

- **Certaines substitutions peuvent avoir plus ou moins d'effet sur la fonction des protéines.**

- **Propriétés physico-chimiques des acides-aminés (acidité, hydrophobicité, encombrement stérique, etc.)**

Substitutions conservatrices



Le cas des acide aminés

- **On notes que:**

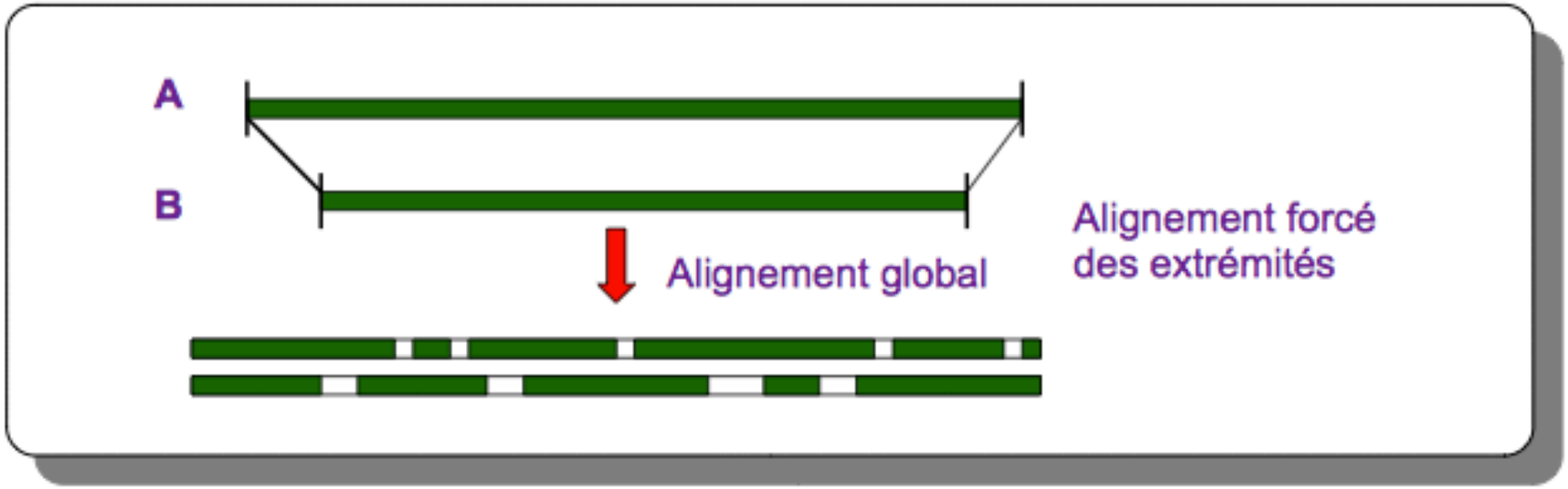
- **Les AA rares ont des scores élevés (Trp, Cys, Hist)**
- **Les AA communs ont des scores faibles (Ala, Leu, Ile)**
- **Les substitutions conservatives entre AA similaires sont peu pénalisantes**
- **Ces substitutions peuvent se produire sans affecter l'activité de la protéine (Ex.: Lys→ Arg)**

Matrices de substitution de protéines plus complexes

Différents types d'alignement

- **Des finalités très différentes→**
  - **Alignement global**
  - **Alignement local**

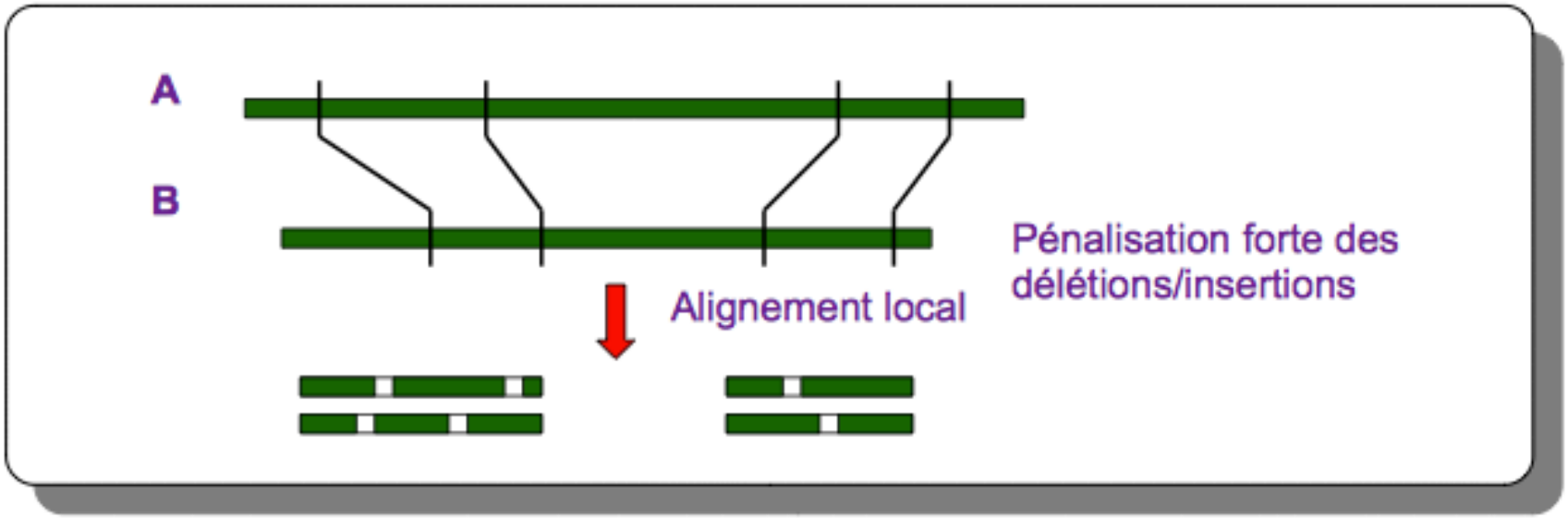
Alignement global



- **Utilisé pour aligner des séquences homologues (gènes, protéines, chromosomes) afin de déterminer les mutations évolutives**

- **La base des alignements multiples (ClustalW,...)**

Alignement local



- **Utilisé pour identifier des séquences homologues, (ex. dans les banques de données)**
- **L'homologie peut être restreinte à une portion de séquence (domaine protéique)**
- **Algo. le plus répandu: BLAST (blastp, blastn,...)**

Deux formes d'aligner

- « Pairwise » alignement→

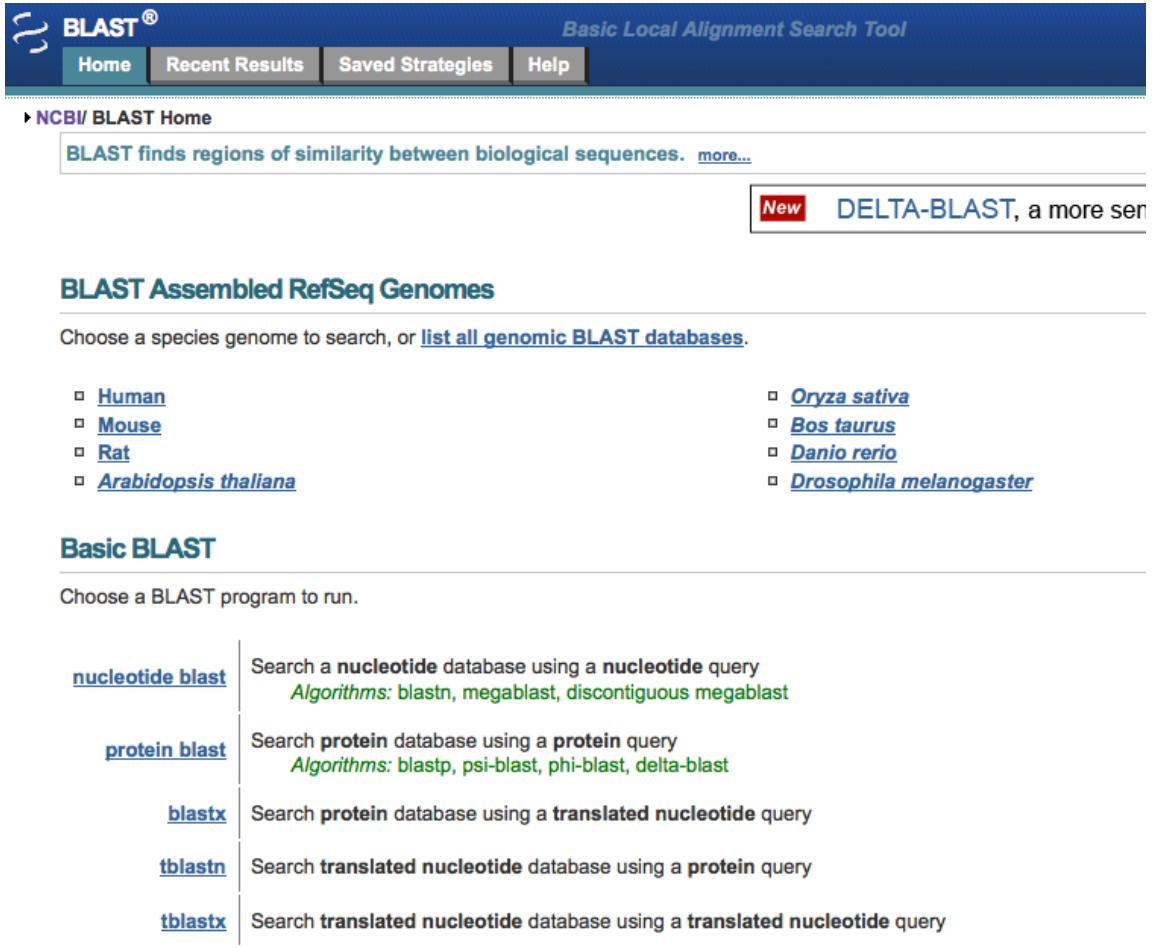
ou alignement de 2 séquences

- Alignement multiple→ plus de 2 séquences

BLAST

http://blast.ncbi.nlm.nih.gov/Blast.cgi

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.



Basic Local Alignment Search Tool

- **BLAST recherche dans une base de données de séquences de segments qui sont localement homologues à une séquence-test fournie par l'utilisateur (*query sequence*).**

- **BLAST utilise une matrice de substitution pour calculer des scores d'alignement et donner une pertinence statistique a ce alignement.**

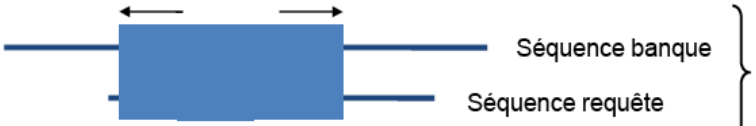
Principe du BLAST

1

Longueur du mot = *w*
Score ≥ *T*

Les séquences sont coupés en mots

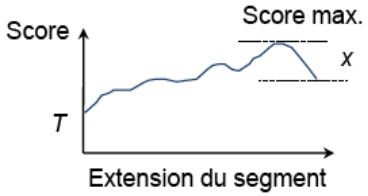
2



HSP : *High Scoring Pair*

Extension du segment à partir de “mots” trouvé (ou “seed”) dans un HSP

3



Extension stoppée quand :
- la fin d’une des deux séquences est atteinte
- score ≤ 0
- score ≤ score_max - *x*

Evaluation de l’alignement donnant un score final

→ **La première étape corresponde a un alignement local**

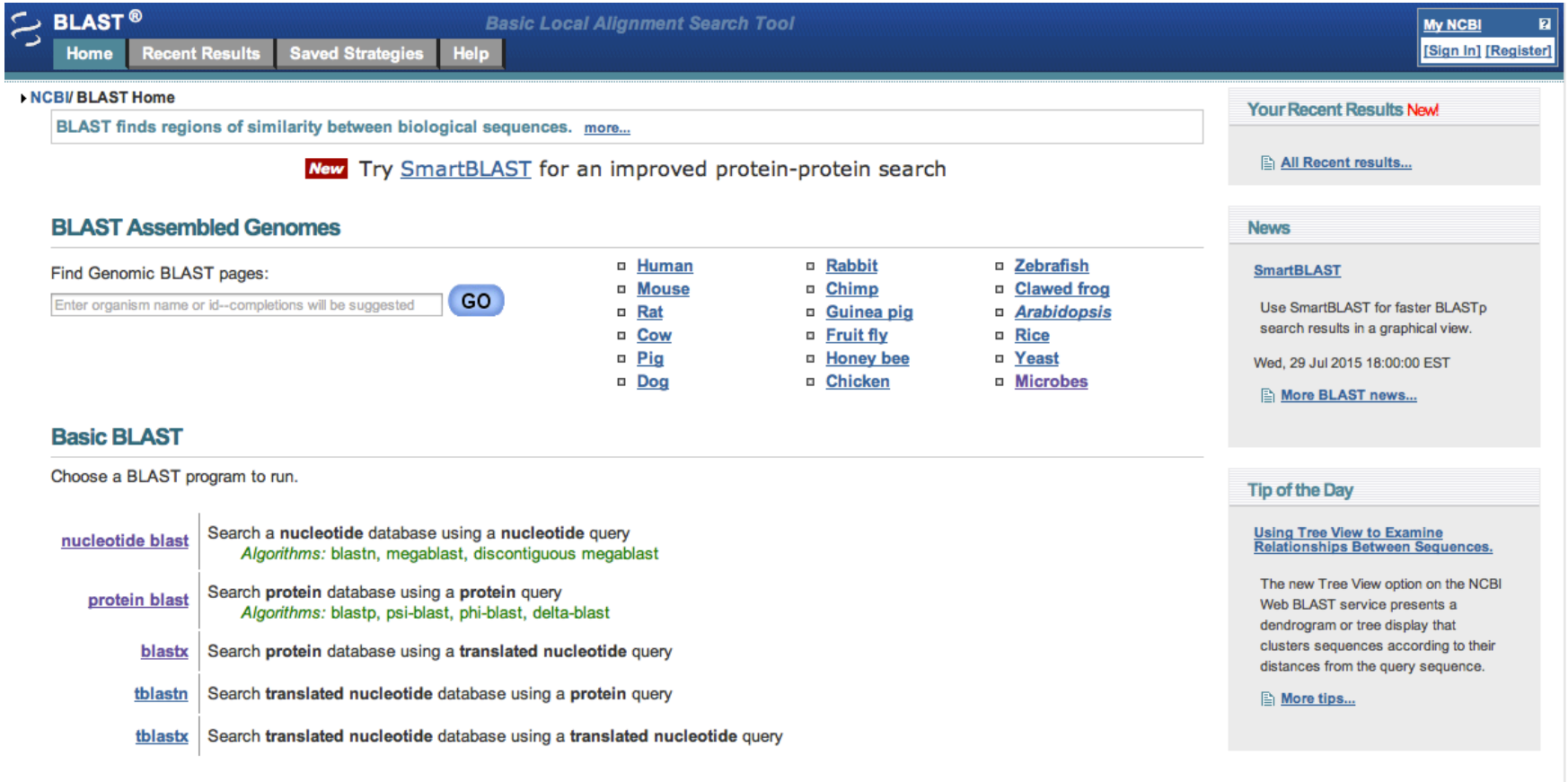
D’où vient le score?

- La qualité de chaque paire d’alignement est représenté par un score et les scores sont classés.
- Les matrices de score sont utilisés pour calculer le score de l’alignement (base par base (l'ADN) ou AA par AA (la protéine)).

- Le score d'alignement sera la somme de scores pour chaque position.

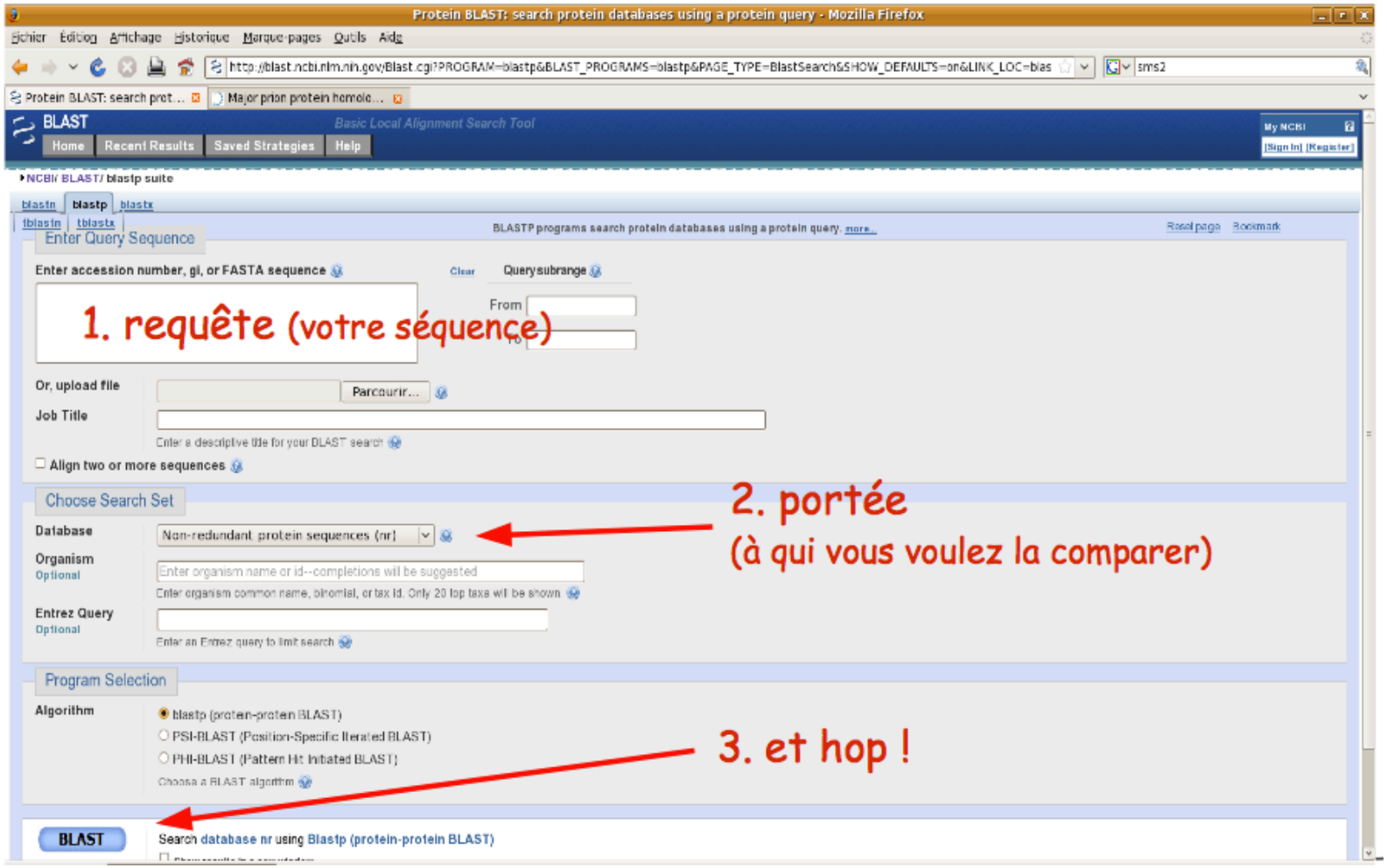
BLAST en pratique

http://blast.ncbi.nlm.nih.gov/Blast.cgi



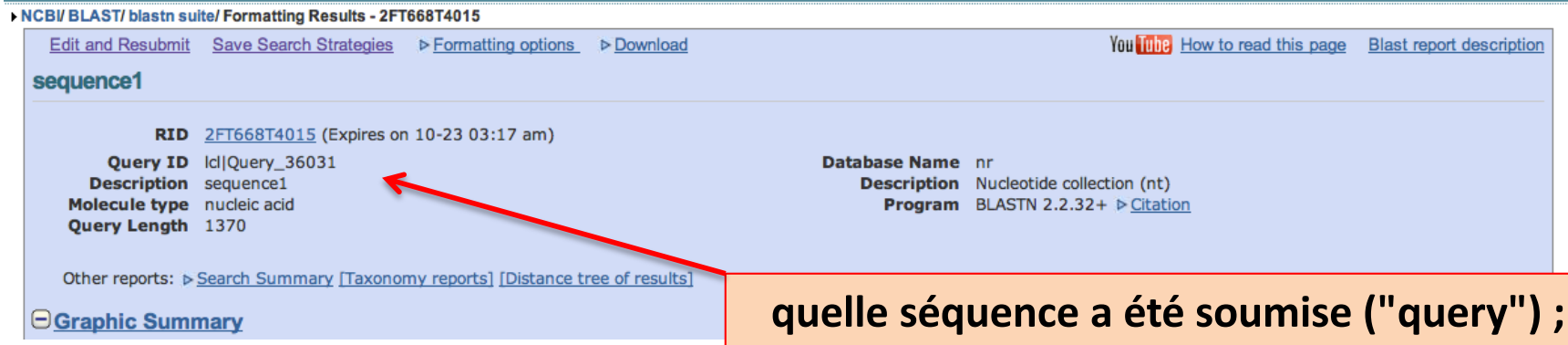
BLAST en pratique

http://blast.ncbi.nlm.nih.gov/Blast.cgi



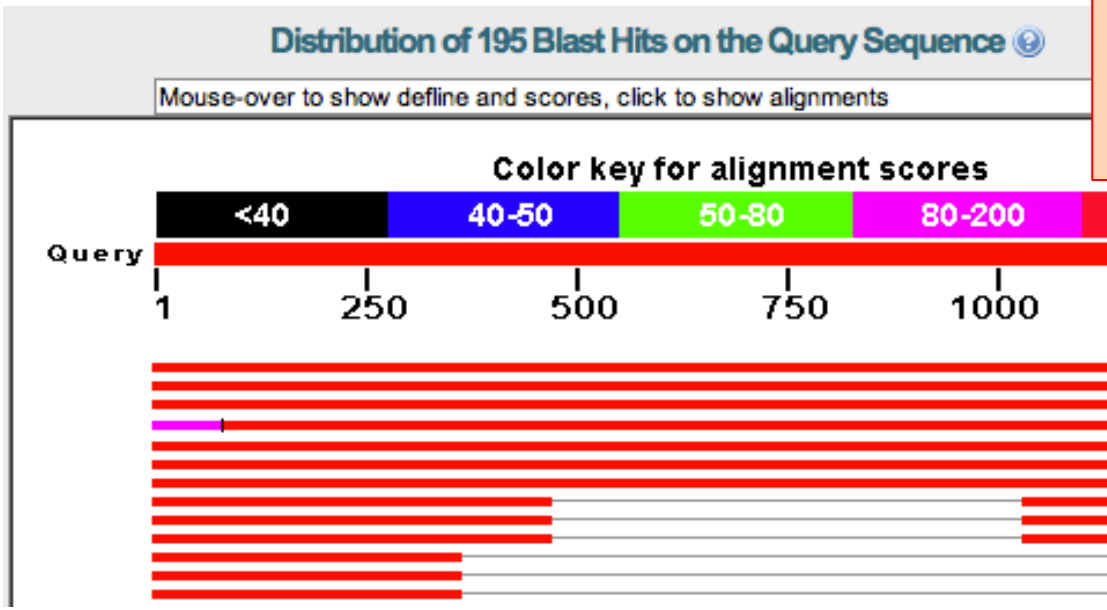
Résultats de BLAST

1. Récapitulatif de la requête



quelle séquence a été soumise ("query") ; identifiant, longueur, type

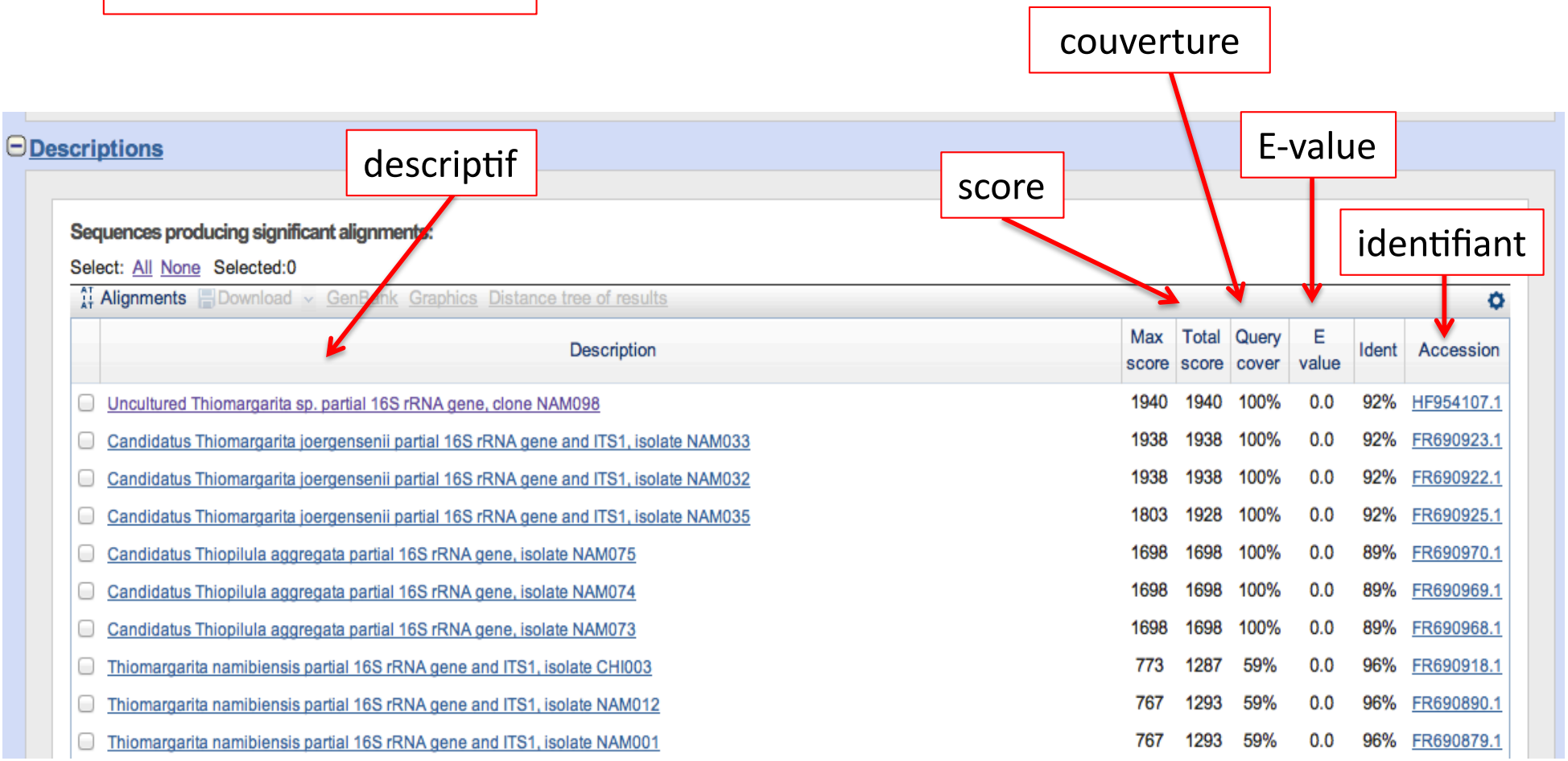
2. Représentation graphique des résultats



chaque trait de couleur représente un alignement entre la séquence « query » et une séquence de la banque de donnée sélectionnée couleur

Résultats de BLAST

3. Résumé des résultats

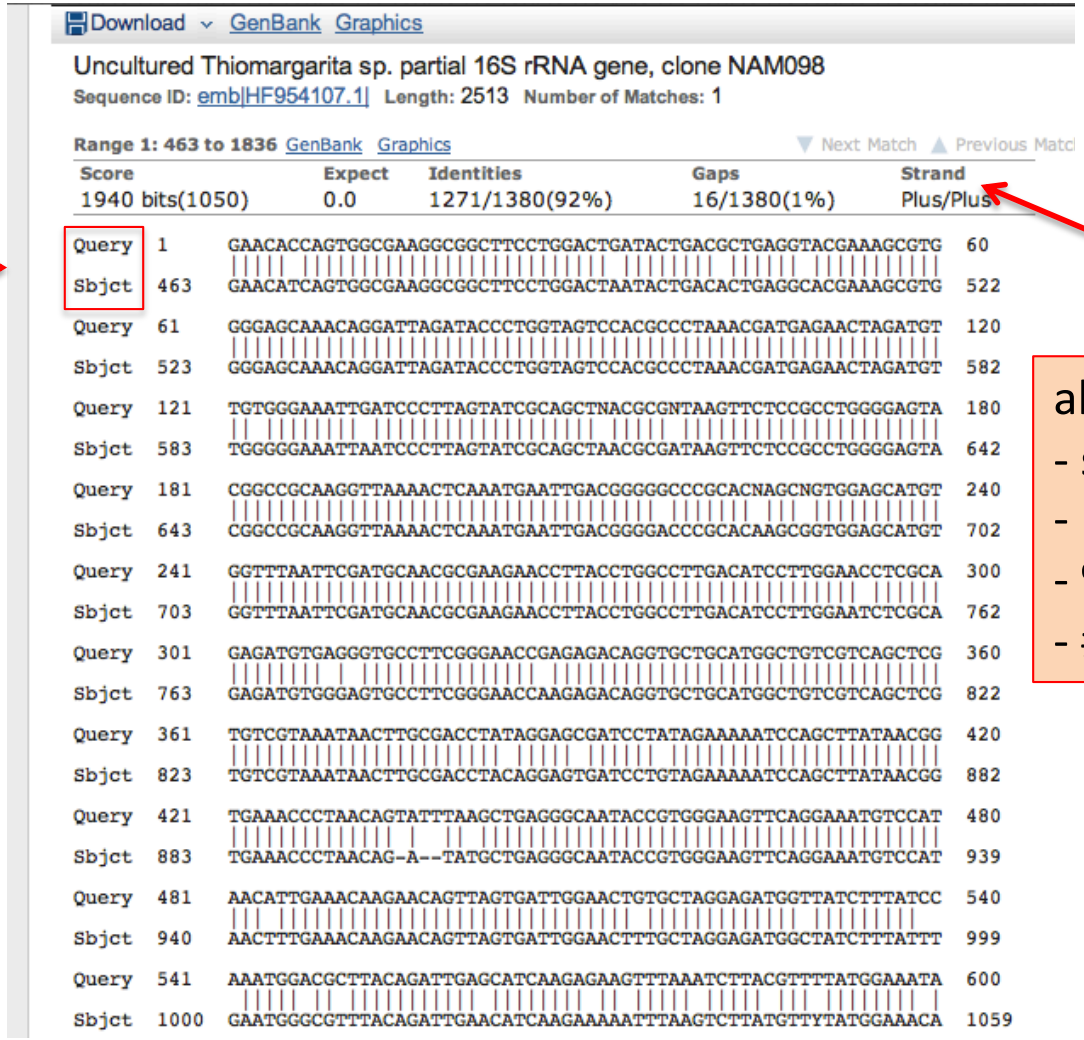


Description	Max score	Total score	Query cover	E value	Ident	Accession
Uncultured Thiomargarita sp. partial 16S rRNA gene, clone NAM098	1940	1940	100%	0.0	92%	HF954107.1
Candidatus Thiomargarita joergensenii partial 16S rRNA gene and ITS1, isolate NAM033	1938	1938	100%	0.0	92%	FR690923.1
Candidatus Thiomargarita joergensenii partial 16S rRNA gene and ITS1, isolate NAM032	1938	1938	100%	0.0	92%	FR690922.1
Candidatus Thiomargarita joergensenii partial 16S rRNA gene and ITS1, isolate NAM035	1803	1928	100%	0.0	92%	FR690925.1
Candidatus Thiopilula aggregata partial 16S rRNA gene, isolate NAM075	1698	1698	100%	0.0	89%	FR690970.1
Candidatus Thiopilula aggregata partial 16S rRNA gene, isolate NAM074	1698	1698	100%	0.0	89%	FR690969.1
Candidatus Thiopilula aggregata partial 16S rRNA gene, isolate NAM073	1698	1698	100%	0.0	89%	FR690968.1
Thiomargarita namibiensis partial 16S rRNA gene and ITS1, isolate CHI003	773	1287	59%	0.0	96%	FR690918.1
Thiomargarita namibiensis partial 16S rRNA gene and ITS1, isolate NAM012	767	1293	59%	0.0	96%	FR690890.1
Thiomargarita namibiensis partial 16S rRNA gene and ITS1, isolate NAM001	767	1293	59%	0.0	96%	FR690879.1

Résultats de BLAST

4. Les alignements

query → la séquence soumise
subject → la séquence trouvée dans la bdd



alignement = outil QUANTITATIF
- scores
- Expect (ou E-value)
- % identité
- # de gaps

Résultats de BLAST

BLAST= programme d'alignements locaux, permettant d'interroger une base de données de séquences à partir d'une séquence ("query »)

Résultat:

Liste d'HSP (high scoring pairs = alignements) avec
% identité, % gaps score brut, score en bits, E-value

E-value = valeur statistique : combien d'HSP de même score aurions nous obtenu au hasard contre une base de données aléatoire de même taille ?

E-value diminue lorsque le score augmente:
E-val < 10-10: homologie très probable;

Notes sur le E-value

- E-values bas suggère que les séquences sont homologues
- **IMPORTANT!!!** La taille des séquences, la taille de la BDD ainsi que le score modifie le E-value

→ E-value augmente en fonction de la taille de la base de données

→ E-value diminue en fonction de la taille de l'alignement

Alignement

- **Un alignement est d'autant plus significatif quand le score est élevé et le « E-values » faible.**

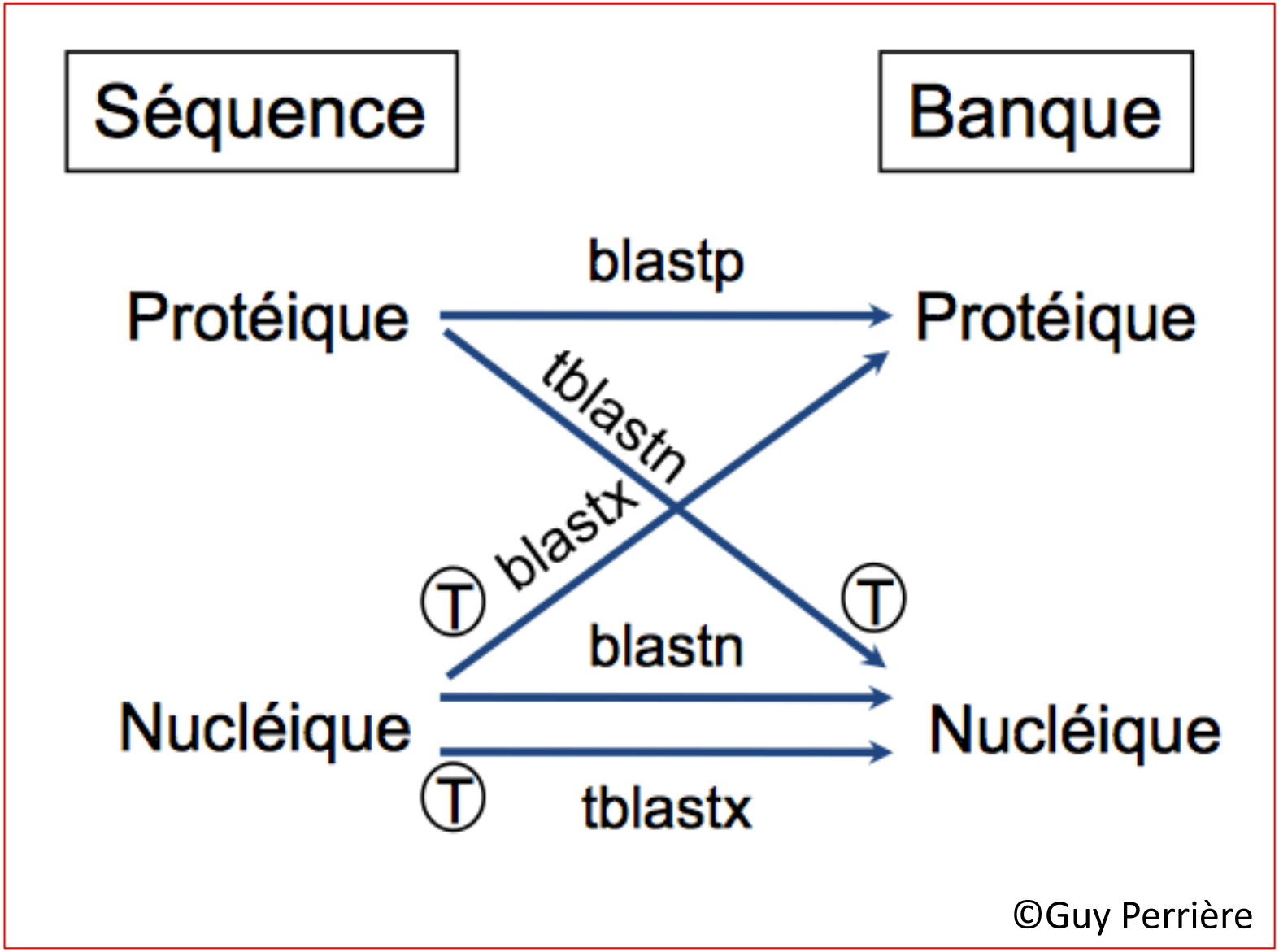
- **Un résultat significatif d'un BLAST indique une similitude entre les séquences et une présomption d'homologie**

Une absence de similarité ne signifie pas une absence d'homologie entre les deux séquences

Programme de BLAST

Program	Description
blastn	Search a **nucleotide** database using a **nucleotide** query
blastp	Search **protein** database using a **protein** query
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Programme de BLAST



Comparer une séquences d'ADN ou une séquence protéique ??

- **Similarité moyenne entre 2 séquences d'ADN de longueur égale: 25%**
- **Similarité moyenne entre 2 de séquences d'AA longueur égale: 5%**

il est plus fréquent d'avoir une bonne similarité due au hasard entre 2 séquences d'ADN que d'AA

Si la séquence d'ADN est potentiellement codante (présence d'ORF), on compare les séquences d'acides aminés plutôt que les séquences d'ADN

Vaut-il mieux comparer les protéines ou l'ADN pour rechercher des homologues d'une séquence?

- La meilleure façon de détecter des similitudes entre séquences est généralement la *comparaison au niveau protéique.*

1. Il existe 20 aa contre 4 bases.
2. Plusieurs codons produisent le même aa. 134 / 549 substitutions de bases sont synonymes. Les séquences protéiques sont plus informatives.
3. La principale raison: l'existence d'outils de comparaison plus puissants pour les aa: utilisation des propriétés physicochimiques ou des substitutions observées dans l'évolution. Même lorsque les aa sont différents, on est capable de retrouver des similitudes. On en est tout à fait incapable au niveau des bases.

Les comparaisons avec les séquences protéiques ne permettent de détecter que les régions codantes. Evidemment, on utilisera toujours la séquence ADN/ARN pour analyser ce qui n'est pas traduit!

N’oubliez pas

- **Méfiez-vous des résultats donnés par les logiciels :**
  - **La qualité des résultats est parfois diminuée au profit de la rapidité**
  - **Certains problèmes admettent un ensemble infini de possibilités**
  - **Ce n'est pas toujours la solution la meilleure qui est trouvée**