

## Sequences searching

*Bioinformatics teachings*

*<http://bioinfomed.fr> - Olivier Croce -*

# Généralités

Bioinformatique => L'approche in silico de la biologie

## Activités principales

- \* Acquisition, organisation, stockage des données biologiques (ex. bases de données)
- \* Utilisation ou conception de logiciels
- \* Données => Analyse, comparaison ou modélisation
- \* Production de nouvelles données biologiques

## Quelques conseils

- \* Méfiez-vous des résultats donnés par les logiciels :
  - => un logiciel modeste bien utilisé, donnera toujours de meilleurs résultats qu'un bon logiciel mal utilisé
  - => La qualité des résultats est parfois diminuée au profit de la rapidité
  - => Beaucoup de logiciels ne font que de la prédiction
- \* Méfiez-vous des banques de données (séquences par exemple) :
  - => Les données se sont pas toujours fiables
  - => La mise à jour n'est pas toujours récente

La réalité mathématique n'est pas la réalité biologique

Les ordinateurs/logiciels sont des outils, qu'il faut apprendre à bien utiliser

## Quelques liens utiles en bioinformatique



\* La Société Française de BioInformatique (SFBI)  
<http://sfbi.impg.prd.fr/>



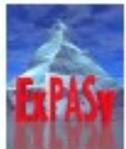
\* Logiciels pour la biologie de l'Institut Pasteur  
<http://bioweb.pasteur.fr/>



\* Le Pôle Bioinformatique Lyonnais (PBIL)  
<http://pbil.univ-lyon1.fr/pbil.html> <http://npsa-pbil.ibcp.fr/>



\* European Bioinformatics Institute (EBI)  
<http://www.ebi.ac.uk/>

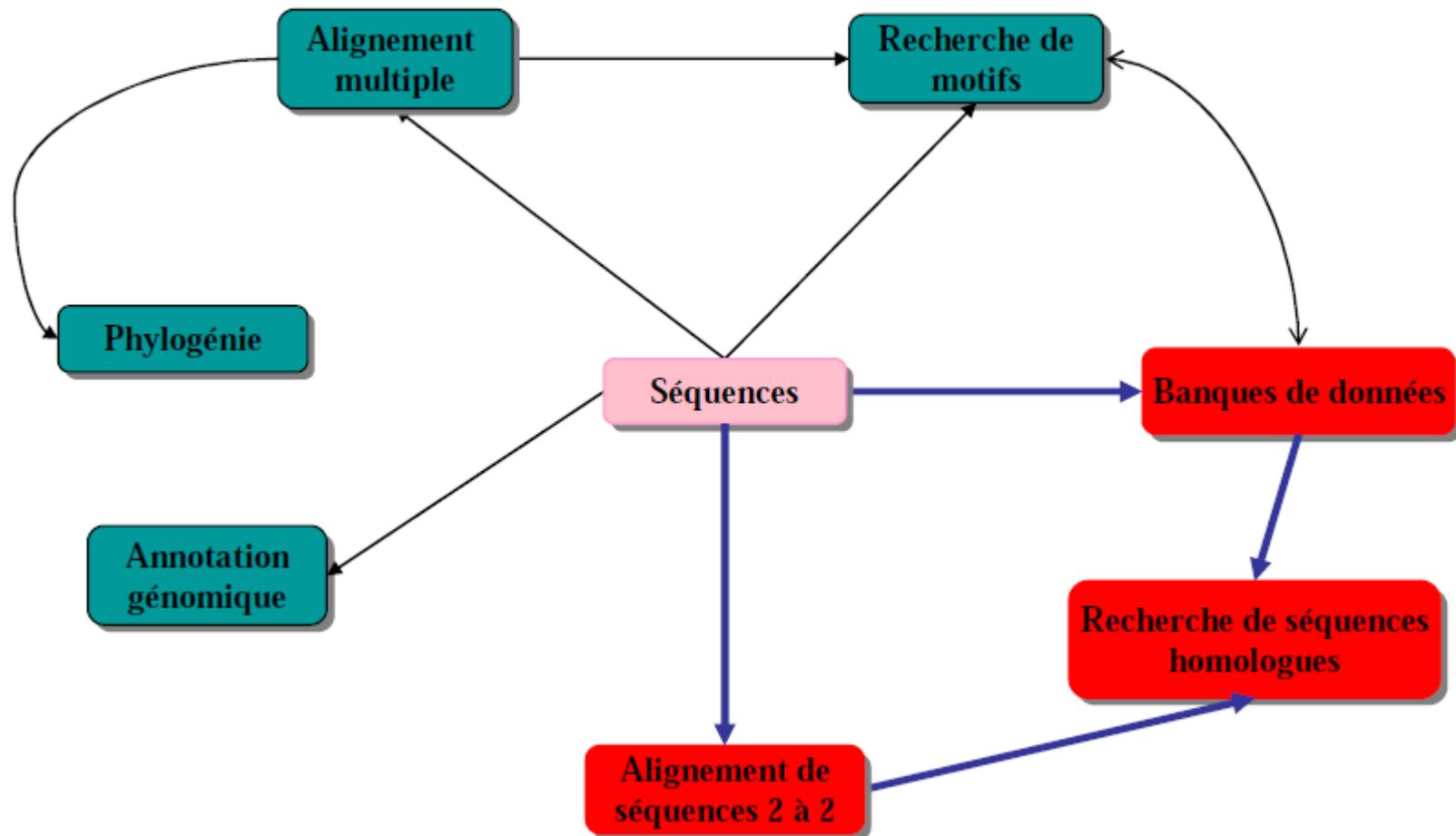


\* Les outils de protéomique d'ExPASy  
<http://www.expasy.org/tools/>



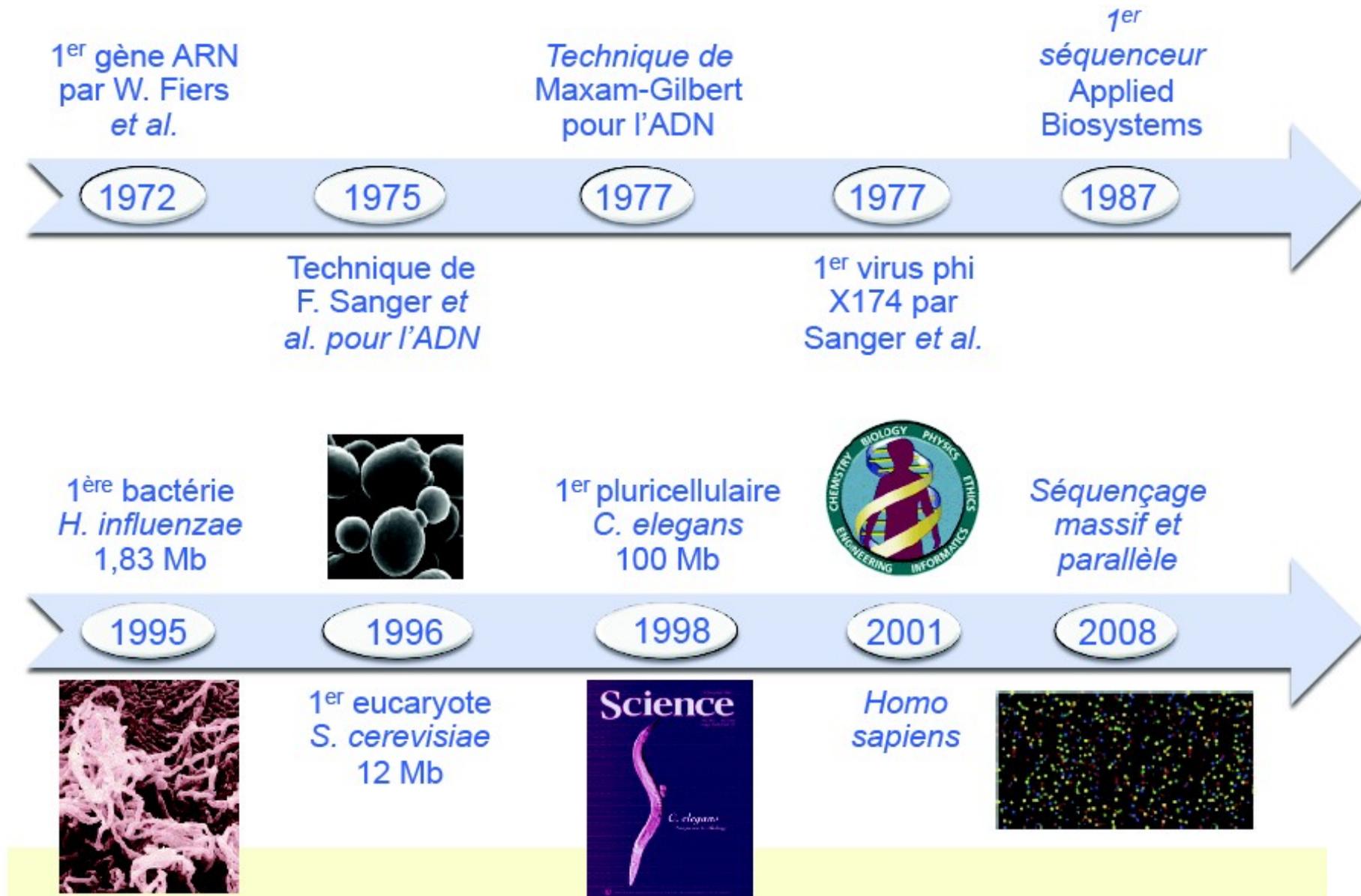
\* National Center for Biotechnology Information (NCBI)  
<http://www.ncbi.nlm.nih.gov/>

# Analyse de séquences, pourquoi ?



# Bases de données de séquences nucléotidiques

- => Origine des données : Séquençage de molécules d'ADN ou d'ARN
- => Nécessité de stocker beaucoup de données et de manière organisée.



# Bases de données de séquences nucléotidiques

Banques nucléiques, collaboration => International Nucleotide Sequence Database Collaboration

## \* Association des 3 banques nucléiques :

EMBL-Bank (European Molecular Biology Laboratory) – EBI  
<http://www.ebi.ac.uk/embl/>

GenBank (banque des Etats-Unis d'Amérique) – NCBI  
<http://www.ncbi.nlm.nih.gov/Genbank/>

DDBJ (DNA DataBank of Japan) – CIB  
<http://www.ddbj.nig.ac.jp/>

- Echange quotidien des données
- Répartition de la collecte des données
- Chaque banque (est supposée) collecter les données de son continent

## \* Les données stockées :

- 1 séquence + ses annotations = 1 entrée
- Fragments de génomes : un ou plusieurs gènes, un bout de gène, séquence intergénique
- Génomes complets
- ARNm, ARNt, ARNr, ... (fragments ou entiers)

# Bases de données de séquences nucléotidiques

## Banques nucléiques, format d'une entrée :

- 3 parties :

Description générale  
de la séquence

« Features »

Description des objets  
biologiques présents  
sur la séquence

La séquence

```
ctccggcagc ccgaggtcat cctgctagac tcagacctgg atgaacctat agacttgcgc      60
tcggtcaaga gccgcagcga ggccggggag ccgccagct ccctccaggt gaagcccagag    120
acaccggcgt cggcggcggt ggcggtggcg gcggcagcgg caccaccac gacggcggag    180
```

- Chaque ligne commence par un mot-clé
  - Deux lettres pour EMBL
  - Maximum 12 lettres pour Genbank et DDBJ
- Fin d'une entrée : //

## EMBL, description générale de la séquence

- ID : toujours la 1ère ligne d'une entrée

Accession	Version	Topologie	Molécule	Classe	Taxonomie	Taille seq
M71283	SV 1	linear	genomic DNA	STD	BCT	1322 BP

- AC : numéros d'accension

- Un n°acc principal pour chaque entrée, unique
- Une liste de n°acc secondaires (historique de l'entrée)

- DT : dates de création et de dernière version

- DE : description du contenu de l'entrée

- KW : mots-clés ; peu renseigné

- OS, OC : organisme contenant la séq. et sa taxonomie

- RN, RC, RX, RP, RA, RT, RL : réf. bibliographiques

- Uniquement les références données par les auteurs de l'entrée

## GenBank et DDBJ, description générale

- LOCUS : toujours la première ligne d'une entrée

Locus name	Taille seq	Molécule	Topologie	Division	Date
BACCOMQP	1322 bp	DNA	linear	BCT	26-APR-1993

- DEFINITION = DE
- ACCESSION = AC
- VERSION ~ DT
- KEYWORDS = KW
- SOURCE, ORGANISM = OS, OC
- REFERENCE, AUTHORS, TITLE, JOURNAL, ... = R...

## Banques nucléiques, lignes FT (Features)

Format (partagé par toutes les banques) :

- **Key** : un seul mot indiquant un groupe fonctionnel
  - Vocabulaire contrôlé, hiérarchique
    - gene : séquence complète du gène (y compris les introns)
    - CDS : séquence codante (sans les introns, entre ATG et Stop)
- **Location** : instructions pour trouver l'objet sur la séquence de l'entrée
- **Qualifiers** : description précise du groupe fonctionnel
  - Format : /qualifier="commentaires libres"
    - /gene="comQ" : nom du gène concerné
    - /note="competence regulation" : information concernant la fonction

## Banques nucléiques, localisation des « keys »

**467** : l'annotation ne concerne qu'une seule base

**109..1105** : entre les positions 109 et 1105 (incluses)  
Toujours la position la plus petite en premier

**<1..21 ou 1275..>1322** : « Keys » tronqués

Commence avant le premier nt de l'entrée

Se termine après le dernier nt de l'entrée (taille seq = 1322)

**<234..888** : début réel inconnu, mais avant 234

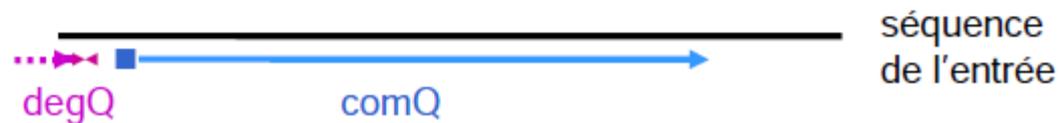
**234..>888** : fin réelle inconnue, mais après 888

**complement(340..565)** : séquence complémentaire inversée à celle de l'entrée (brin -)

**join(12..78,134..202)** : fragments indiqués mis bout à bout (concaténés) ; nombre de fragments illimité

## Exemple de « Feature » d'une séquence ADN

```
FT  CDS                <1..21
FT                      /codon_start=1
FT                      /db_xref="SWISS-PROT:Q99039"
FT                      /transl_table=11
FT                      /gene="degQ"
FT                      /protein_id="AAA22322.1"
FT                      /translation="YAMKIS"
FT  terminator         21..47
FT                      /gene="degQ"
FT  promoter           109..140
FT                      /gene="comQ"
FT  mRNA               146..1105
FT                      /partial
FT                      /gene="comQ"
```



# Bases de données de séquences nucléotidiques

## Le format FASTA

- Utilisé par les logiciels d'analyse de séquences
- Fichier texte simple, créé avec n'importe quel éditeur de texte (ex. blocnote, textpad, etc.) : toto.fas, toto.fna, toto.fa, etc..
- Une ligne débutant avec « > » + un identifiant unique (obligatoire) + des commentaires (optionel)

La séquence brute (pas d'espace, ni de nombre), sur une ou plusieurs lignes

```
>AC25300 Human Polycomb 2 homolog (hPc2) mRNA, partial cds
ctccggcagccccgaggtcatcctgctagactcagacctggatgaaccac
Ctccggcagccccgaggtcatcctgctagactcagacctggatgaaccat
agacttgcgctcgggtcaagagccgcagcaggccggggagccgcccagct
ccctccaggtgaagcccagacaccggcgtcggcggcgggtggcgggtggcg
Gcggcagcggcaccaccacgacggcgggagaagcct
>ID23003 hPc2 gene
ggacgaacctgcagagtcgctgagcgagttcaagcccttctttgggaata
taattatcaccgacgtcaccggaactgcctcaccgttactttcaaggag
tacgtgacgggtg
```

## Banques nucléiques, inconvénients

- \* Difficulté de mise à jour des données

Version plus récente d'une séquence ou d'une annotation dans d'autres banques (ex : banques dédiées à un génome complet)

- \* Forte redondance

Un même fragment de séquence présent dans plusieurs entrées

- \* Annotations peu normalisées

Difficulté de recherche d'une information précise

- \* Annotations peu précises

Peu de descriptions sur les gènes et leurs produits

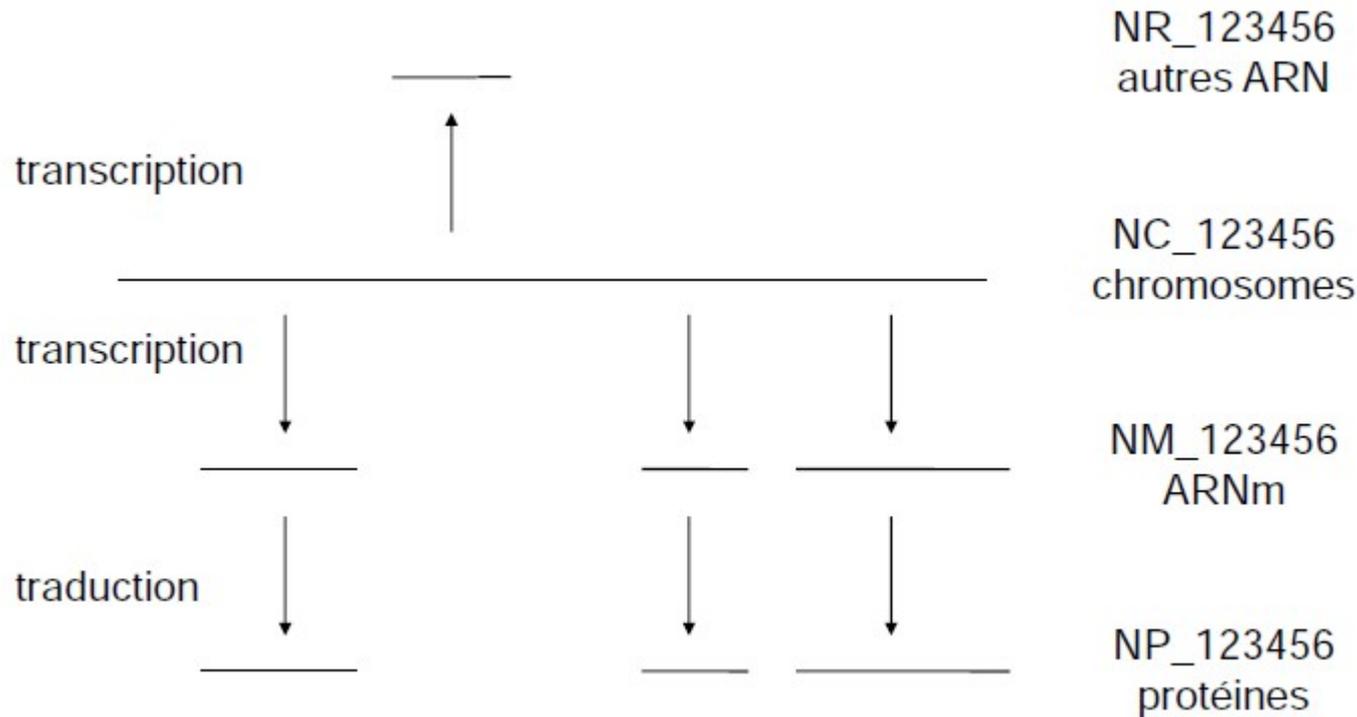
- \* Erreurs dans les annotations et dans les séquences

# Bases de données de séquences nucléotidiques

Bases connues : **nr / nt** et **RefSeq** (=Reference Sequence collection)

=> bases de séquences non-redondantes.

=> **nr/nt** plus complet que **RefSeq**, RefSeq + contrôlée et avec liens explicites entre les séquences nucléiques et protéiques.



# Bases de données de séquences protéiques

## => Origine des données

- Traduction automatique des séquences d'ADN ou d'ARNm
- Séquençage de protéines (rare car + long et coûteux)

=> Les données stockées : séquences + annotations = une entrée

- Protéines entières
- Fragments de protéines

1956 : F. Sanger établit la séquence en aa de l'insuline

1965 : Atlas of Protein Sequences, M. Dayhoff  
Version papier jusqu'en 78, puis version électronique

1984 : création de PIR-NBRF (Protein Information Resource - National Biomedical Research Foundation) Collaboration avec MIPS (Allemagne) et JIPID (Japon)

1986 : création de SwissProt  
Collaboration entre SIB (Swiss Institut of Bioinformatics) et EBI

Fin 2003 : UniProt (Universal Protein Resource)  
Mise en commun des informations de PIR et SwissProt/TrEMBL  
<http://www.expasy.uniprot.org/>

# Bases de données de séquences proteiques

UniProt est en 2 parties

## SwissProt

- Données corrigées et validées par des experts
- Haut niveau d'annotations
  - \* Description de la fonction (références associées)
  - \* Localisation des domaines fonctionnels
  - \* Modifications post-traductionnelles
  - \* Existence de variants, ...
- Redondance minimale
- Nombreux liens vers d'autres banques (60 BD)

## TrEMBL

- Entrées supplémentaires à SwissProt (pas encore annotées)
- Traduction automatique de l'EMBL

# Bases de données de séquences proteiques

## SwissProt/TrEMBL, format d'une entrée : format basé sur celui de l'EMBL

- Mot-clé de 2 lettres au début de chaque ligne
- Format différent pour les Features
- Mots-clés supplémentaires :
  - GN : les différents noms du gène qui code pour la protéine (OR) et les différents gènes qui codent pour la même protéine (AND)
  - OX : références croisées vers les banques taxonomiques
  - CC : commentaires, lignes très documentées dans SwissProt
  - KW : mots-clés issus d'un dictionnaire
  - DR : références vers d'autres banques de données
    - Vers les séquences nucléiques (EMBL/GenBank/DDBJ)
    - Vers les structures 3D
  - ...

Informations découpées en blocs pour plus de lisibilité

- CC -!- TOPIC: First line of a comment block;
- CC second and subsequent lines of a comment block.

De nombreux sujets sont abordés

- FUNCTION : description générale de la fonction de la protéine
- CATALYTIC ACTIVITY : description des réactions catalysées par les enzymes
- DEVELOPMENTAL STAGE : description du stade spécifique auquel la protéine est exprimée
- SUBUNIT : complexes dont fait partie la protéine (+ partenaires)
- ...

# Pourquoi interroger une banque ?

- Obtenir des informations nouvelles et pertinentes
- Aide à la mise au point d'expériences
- Validation des résultats d'une expérience
  
- Trouver si des séquences sont déposées dans les banques.
  
- Identifier des protéines homologues (avec ancêtre commun ):
  - orthologue : organismes différents
  - paralogue : organismes identiques
  
- Déterminer si des séquences ont une fonction similaire ou proche.
  
- Déterminer des familles de protéines ayant un domaine conservé.
  
- Localiser des régions codantes et non codantes (aligner des séquences génomiques ADN et des séquences exprimées (cDNAs, ESTs)).
  
- Etablir des relations entre les séquences.

=> Comment interroger une banque ?

# 1 - Interrogation par annotations

**Interrogation des annotations d'une banque => Recherche dans les annotations**

**Systèmes d'interrogation de banques de données : Entrez, Acnuc**

=> Recherche de mots ou expressions dans le texte des entrées

- \* Obtention de données pertinentes (pas trop de résultats, mais tous ceux relatifs à notre problématique)

- \* Simplicité d'utilisation (syntaxe d'interrogation intuitive)

- \* Réponse rapide

- \* Possibilité d'analyse des résultats (couplage à d'autres outils)

=> Systèmes d'interrogation de banques de données :

- \* **Entrez**, permet aussi d'interroger des banques de séquences (<http://www.ncbi.nlm.nih.gov>)

Même fonctionnement que pour interroger PubMed

- \* **ACNUC** ([http://doua.prabi.fr/search/query\\_fam](http://doua.prabi.fr/search/query_fam))

# 1 - Interrogation par annotations

## WWW-Query

[History](#) [Modify](#) [Retrieve](#) [Releases](#) [Help](#)

[Example](#)

### Sequence search selected

Search for sequences  Search for families, alignments and trees

### Protein sequences selected

Protein [databank](#)  Nucleotide [databank](#)

### Selection criteria:

1.	DEFAULT	Keyword	<input type="text"/>
2.	AND	Keyword	<input type="text"/>
3.	AND	Keyword	<input type="text"/>
4.	AND	Keyword	<input type="text"/>

List name:

[http://doua.prabi.fr/search/query\\_fam](http://doua.prabi.fr/search/query_fam)

## Nucleotide Advanced Search Builder

Use the builder below to create your search

[Edit](#)

### Builder

<input type="text" value="All Fields"/>	<input type="text"/>	<input type="button" value="-"/>	<a href="#">Show index list</a>
AND <input type="text" value="All Fields"/>	<input type="text"/>	<input type="button" value="-"/> <input type="button" value="+"/>	<a href="#">Show index list</a>

or [Add to history](#)

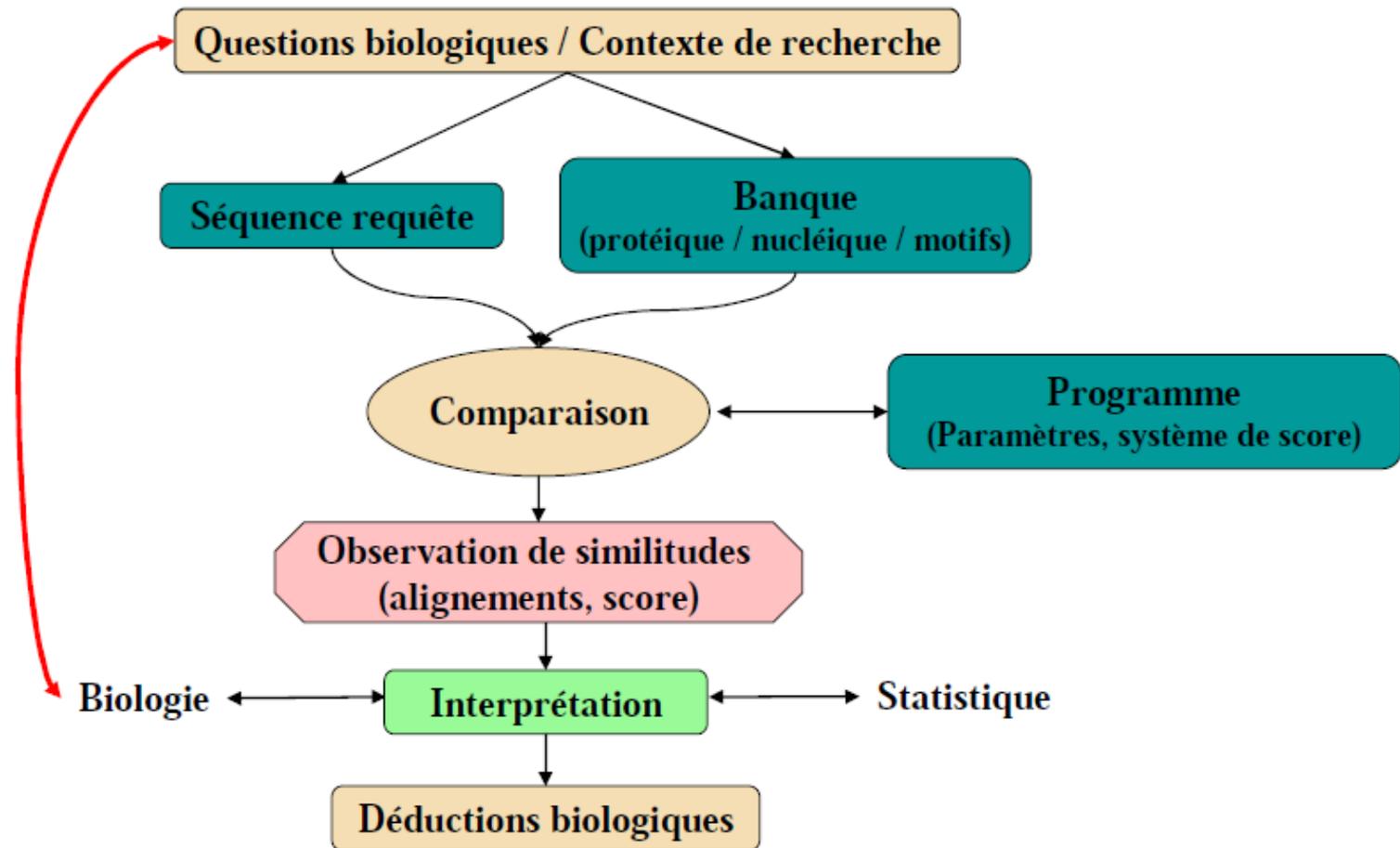
<https://www.ncbi.nlm.nih.gov/nucore/advanced>

[https://www.ncbi.nlm.nih.gov/books/NBK44863/#sequencesquickstart.How\\_do\\_I\\_use\\_a\\_simpl](https://www.ncbi.nlm.nih.gov/books/NBK44863/#sequencesquickstart.How_do_I_use_a_simpl)

## 2 - Interrogation par similarité - Recherche dans les séquences

**Pourquoi ?** => Savoir si ma séquence ressemble à d'autres déjà connues ; Trouver toutes les séquences d'une même famille ; Rechercher toutes les séquences qui contiennent un motif donné

**Comment ?** => Comparaison d'une séquence aux séquences de la banque : BLAST, FASTA, Blat, YASS, ...



Programme = alignement + statistique



analyse biologique

## 2 - Interrogation par similarité – les alignements

Notion clé : l'alignement – 2 types d'alignements :



→ utilisé en phylogénie

**Alignement global:** alignement de toute la séquence A avec toute la séquence B

Méthode employée pour aligner des séquences dont on soupçonne l'homologie.

L'alignement est optimisé sur toute la longueur des séquences.

L'algorithme de référence est celui de **Needleman & Wunsch**.



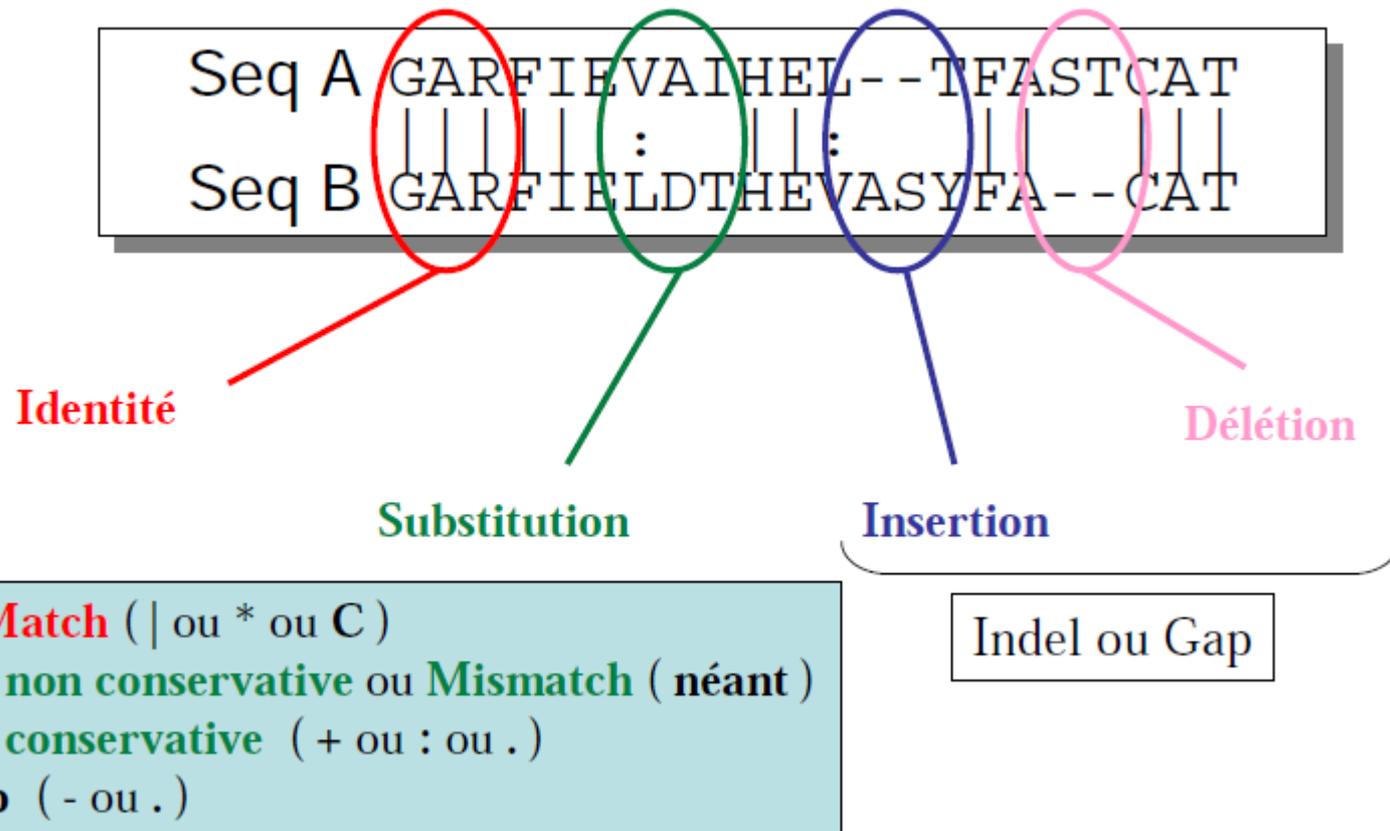
→ utilisé pour la recherche de similarité



**Alignement local:** alignement de sous-séquences de A avec des sous-séquences de B

Aligne seulement les régions dont le score est supérieur à un seuil donné. Utilisé lorsque l'on veut aligner deux séquences de taille très différente. (par ex. dans une recherche de sous-séquence). Beaucoup plus rapide que l'alignement global. Algorithme de **Smith & Waterman**.

## 2 - Interrogation par similarité – types d'alignements



3 situations sont possibles pour une position donnée de l'alignement :

- les caractères sont les mêmes: Identité ou match
- les caractères ne sont pas les mêmes: Substitution
- les caractères existent dans un cas et pas dans l'autre : Insertion ou Délétion

## 2 - Interrogation par similarité - Définitions

### **Identité (globale) :**

Proportion des paires de résidus identiques entre deux séquences alignées. (Exprimé généralement en %).

### **Similitude (=similarité) :**

Mesure de la ressemblance entre deux séquences. Le degré de similitude est quantifié par un score basé sur le % de similarité (% identité + % substitutions conservatives) des séquences. De 100% à quelques nucléotides/aminoacides en commun.

### **Gaps ou InDels :**

Proportion d'Indels entre deux séquences alignées (Exprimé en %).

### **Homologie :**

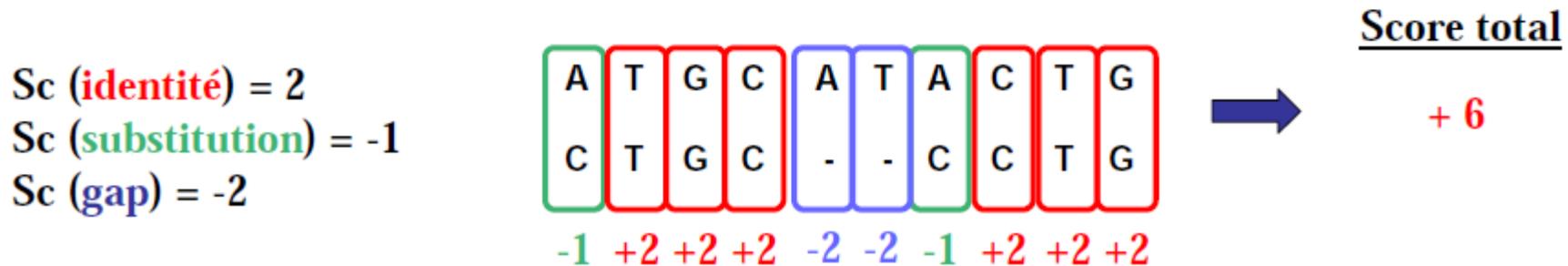
Deux séquences sont homologues si elles ont un ancêtre commun. Il n'y a pas vraiment de degré d'homologie (remarque : on ne dit pas: tres homologue, faible homologie, etc.) Il n'y a pas vraiment de limite, mais en dessous de 20-25% (twilight zone), il devient très difficile de distinguer une homologie d'une ressemblance fortuite.

A noter aussi que des séquence sans ressemblance apparente peuvent aussi être homologues (on le retrouve par ex. au niveau 3D) =>Des séquences homologues ne sont pas nécessairement similaires.

## 2 - Interrogation par similarité – Qualité alignements : Le score

Le score de l'alignement est la somme de toutes les positions 2 à 2.

➤ **Exemple:** On peut associer une récompense (positive) à des symboles alignés identiques et une pénalité (négative) à un substitution ou à un gap.



=> bon alignement donnera le score maximum entre 2 sequences

## 2 - Interrogation par similarité – alignement sur Bases de seq.

### Comparaison d'une séquence aux séquences de la banque

=> Identifier les [fragments de] séquences de la banque ayant une forte ressemblance avec la séquence requête

\* Requête : une ou plusieurs séquences (ADN/ARN ou protéine)

\* Résultat : une liste de séquences ressemblant à la séquence entrée (classées par score, identité, couverture, e-value)

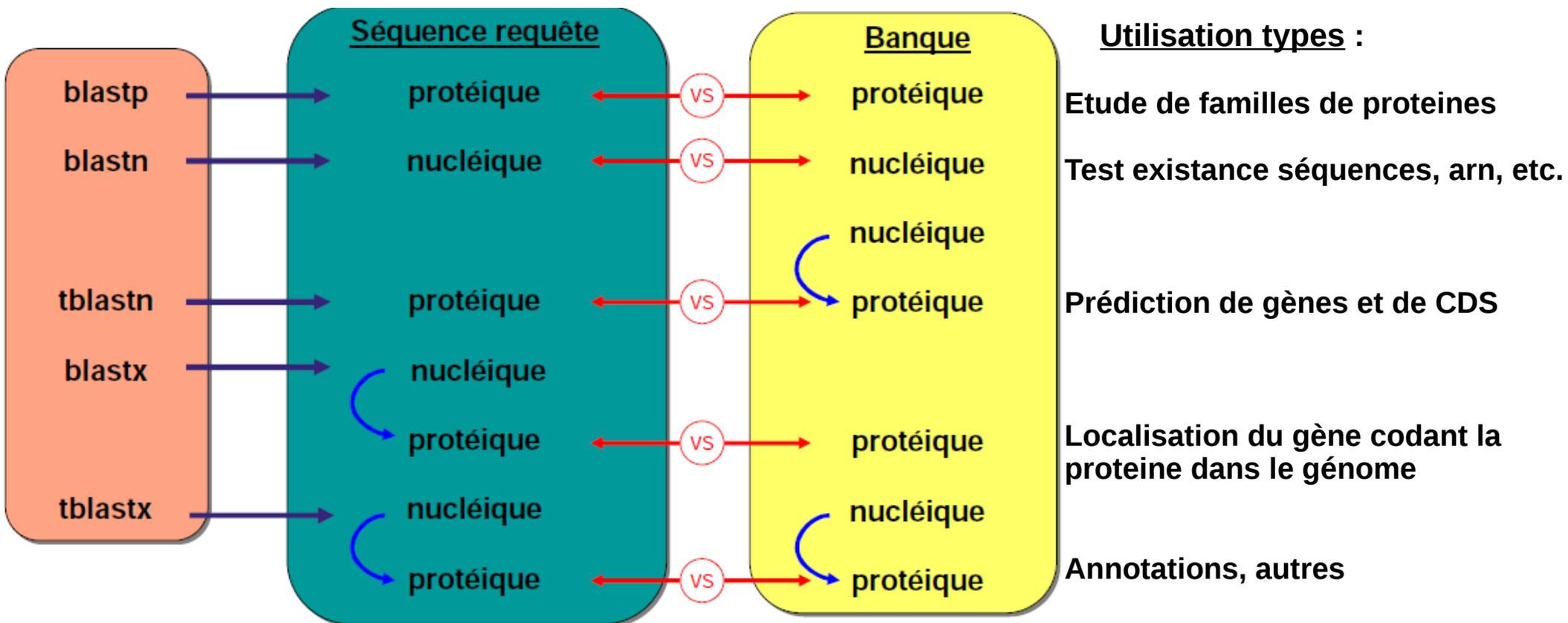
=> exemples suivants avec "Blast" (avec interface graphique du NCBI)

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

## 2 - Interrogation par similarité – Blast

L'algorithme est basé sur un modèle statistique (Karlin et Altschul, 1990) qui s'applique aux comparaisons de séquences sans insertion / délétion.

Différents programmes :



[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#)

## Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. [more.](#)

Enter accession number(s), gi(s), or FASTA sequence(s)

[Clear](#)

Query subrange

From To 

Or, upload file

[Choir...](#)

Job Title

Enter a descriptive title for your BLAST search

 Align two or more sequences 

## Choose Search Set

Database

 Human genomic + transcript  Mouse genomic + transcript  Others (nr etc.):

Nucleotide collection (nr/nt)

Organism

Optional

 Enter organism name or id--completions will be suggested  Exclude 

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude

Optional

 Models (XM/XP)  Uncultured/environmental sample sequences

Entrez Query

Optional

Enter an Entrez query to limit search

## Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

**BLAST**Search **database Nucleotide collection (nr/nt)** using **Blastn (Optimize for somewhat similar sequences)** Show results in a new window [Algorithm parameters](#)

General Parameters

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

- Max target sequences: 100 (Select the maximum number of aligned sequences to display)
- Short queries:  Automatically adjust parameters for short input sequences
- Expect threshold: 10
- Word size: ♦ 16 (highlighted in yellow)
- Max matches in a query range: 0

Scoring Parameters

- Match/Mismatch Scores: 1,-2
- Gap Costs: Linear

Filters and Masking

- Filter:  Low complexity regions;  Species-specific repeats for: Homo sapiens (Human)
- Mask:  Mask for lookup table only;  Mask lower case letters

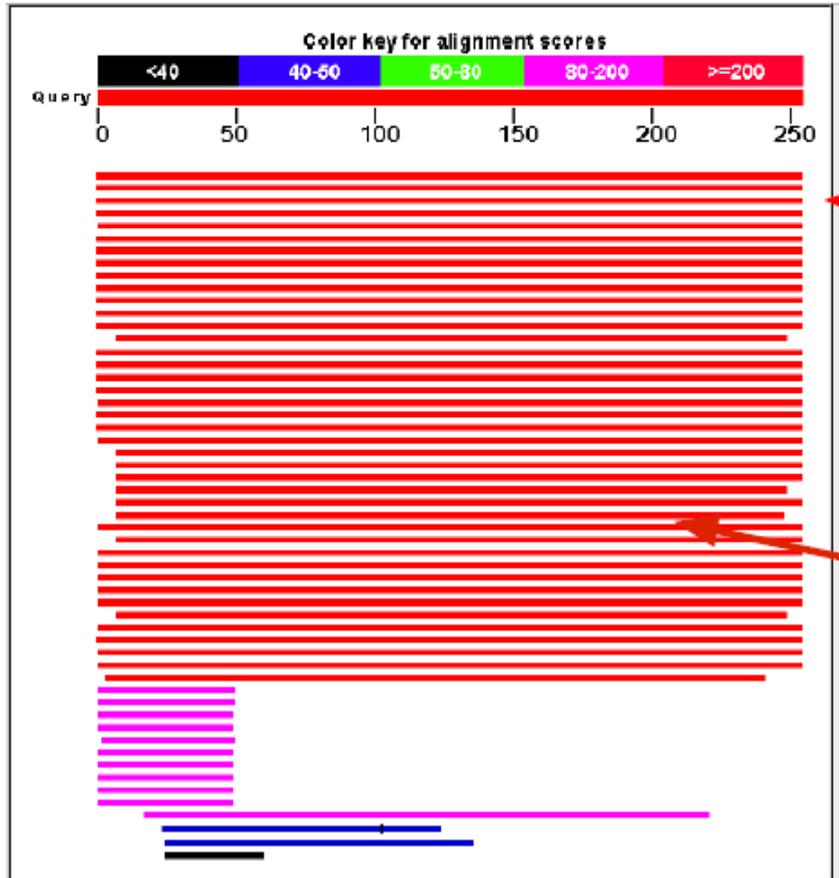
**BLAST**

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window



er to show define and scores, click to show alignments



## 2. représentation graphique des résultats

ce trait représente  
la séquence soumise (long. 253 AA)

chaque trait de couleur représente  
un alignement entre la séquence de  
départ et une séquence de la  
banque de donnée sélectionnée  
couleur → score  
longueur → taille de l'alignement

= HSP ("high scoring pair")

### 3. Résumé des résultats

**Descriptions**

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#)
[Download](#)
[GenBank](#)
[Graphics](#)
[Distance tree of results](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	<a href="#">Uncultured bacterium clone Z206 16S ribosomal RNA gene, partial sequence</a>	2676	2676	98%	0.0	99%	<a href="#">GQ388829.1</a>
<input type="checkbox"/>	<a href="#">Uncultured Afipia sp. clone NW-12 16S ribosomal RNA gene, partial sequence</a>	2663	2663	99%	0.0	99%	<a href="#">AY568510.1</a>
<input type="checkbox"/>	<a href="#">Uncultured Afipia sp. 16S ribosomal RNA gene, partial sequence</a>	2658	2658	98%	0.0	99%	<a href="#">FJ572667.1</a>
<input type="checkbox"/>	<a href="#">Uncultured bacterium clone MA-78-I98C 16S ribosomal RNA gene, partial sequence</a>	2656	2656	97%	0.0	99%	<a href="#">HM141889.1</a>
<input type="checkbox"/>	<a href="#">Uncultured bacterium clone BL 16S ribosomal RNA gene, partial sequence</a>	2643	2643	97%	0.0	99%	<a href="#">FJ154970.1</a>
<input type="checkbox"/>	<a href="#">Afipia sp. BAC308 16S ribosomal RNA gene, partial sequence</a>	2643	2643	98%	0.0	99%	<a href="#">EU130950.1</a>
<input type="checkbox"/>	<a href="#">Afipia sp. OHSU_I 16 ribosomal RNA, tRNA-Ile, tRNA-Ala, 23S ribosomal RNA, and 5S riboso</a>	2615	2615	100%	0.0	99%	<a href="#">KC677616.1</a>
<input type="checkbox"/>	<a href="#">Afipia sp. OHSU_II 16 ribosomal RNA, tRNA-Ile, tRNA-Ala, 23S ribosomal RNA, and 5S riboso</a>	2615	2615	100%	0.0	99%	<a href="#">KC677615.1</a>
<input type="checkbox"/>	<a href="#">Uncultured Afipia sp. clone AV_6J-C05 16S ribosomal RNA gene, partial sequence</a>	2606	2606	96%	0.0	99%	<a href="#">EU341227.1</a>
<input type="checkbox"/>	<a href="#">Uncultured bacterium clone SPCUL1A1 16S small subunit ribosomal RNA gene, partial sequ</a>	2601	2601	95%	0.0	99%	<a href="#">AY186080.1</a>
<input type="checkbox"/>	<a href="#">Uncultured Afipia sp. clone C-12 16S ribosomal RNA gene, partial sequence</a>	2595	2595	99%	0.0	98%	<a href="#">AY568503.2</a>

#### 4. Présentation des alignements obtenus pour chaque séquence sélectionnée de la banque

```

Sbjct 1321 |CGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTTGGTTCTACCTGAAGGCAGTGCG| 1380
Query 1382 |CTAACCCGCAAGGGAGGCAGCTGACCACGGTAGGGTCAGCGACTGGGGTGAAGTCGTAAC| 1441
Sbjct 1381 |CTAACCCGCAAGGGAGGCAGCTGACCACGGTAGGGTCAGCGACTGGGGTGAAGTCGTAAC| 1440
Query 1442 |AAGGTAGCCGTA| 1453
Sbjct 1441 |AAGGTAACCGTA| 1452

```

 Download [GenBank](#) [Graphics](#)

Uncultured Afipia sp. clone NW-12 16S ribosomal RNA gene, partial sequence

Sequence ID: [gb|AY568510.1](#) Length: 1487 Number of Matches: 1

Range 1: 2 to 1483 [GenBank](#) [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Identities	Gaps	Strand
2663 bits(1442)	0.0	1470/1482(99%)	8/1482(0%)	Plus/Plus
Query 4	GTTTGATCCTGGCTCAGAGCGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGGG	63		
Sbjct 2	GTTTGATCCTGGCTCAGAGCGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGGG	61		
Query 64	CGTAGCAATACGTCAGCGGCAGACGGGTGAGTAACGCGTGGGAACGTACCTTTTGGTTTCG	123		
Sbjct 62	CGTAGCAATACGTCAGCGGCAGACGGGTGAGTAACGCGTGGGAACGTACCTTTTGGTTTCG	121		
Query 124	GAACAACGAGGGAACTTCAGCTAATACCGGATAAGCCCTTACGGGGAAAGATTTATCG	183		
Sbjct 122	GAACAACGAGGGAACTTCAGCTAATACCGGATAAGCCCTTACGGGGAAAGATTTATCG	181		
Query 184	CCGAAAGATCGGCCCGCTCTGATTAGCTAGTTGGTGAGGTAATGGCTCACCAAGGCGAC	243		
Sbjct 182	CCGAAAGATCGGCCCGCTCTGATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCGAC	241		
Query 244	GATCAGTAGCTGGTCTGAGAGGATGATCAGCCACATTGGGACTGAGACACGGCCCAAAC	303		
Sbjct 242	GATCAGTAGCTGGTCTGAGAGGATGATCAGCCACATTGGGACTGAGACACGGCCCAAAC	301		
Query 304	CCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGC	363		
Sbjct 302	CCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGC	361		
Query 364	CGCGTGAGTGATGAAGGCCCTAGGGTTGTAAGCTCTTTTGTGCGGGAAGATAATGACGG	423		
Sbjct 362	CGCGTGAGTGATGAAGGCCCTAGGGTTGTAAGCTCTTTTGTGCGGGAAGATAATGACGG	421		
Query 424	TACCGCAAGAATAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGTAATACGAAGGGGGC	483		



# Exercice

**1 – “séquences cibles”** : récupération des séquences (homologues) de « *Bacillus anthracis* », gene « *cya* »

**2 - ”séquences non-cibles”** : récupération des séquences proches, mais n'étant pas de *Bacillus anthracis* »

## Methodes à utiliser :

**1.1** En effectuant une recherche par mots-clefs grâce aux applications suivantes : Entrez (<http://www.ncbi.nlm.nih.gov>) ou WWWQuery (=ACNUC) ([http://doua.prabi.fr/search/query\\_fam](http://doua.prabi.fr/search/query_fam)), récupérer sur votre disque dur toutes les séquences codantes (CDS) relatives au gène “*cya*” de *Bacillus anthracis*. Sauvegarder dans un fichier fasta l'ensemble de ces séquences cibles.

**1.2** Pour compléter (ou vérifier) le jeu de séquences cibles, à l'aide d'une séquence représentative de votre résultat précédent, effectuer une recherche par similarité des séquences du gène, en utilisant le programme BLASTn (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>).

Ce programme renvoie pour chaque séquence trouvée un score, une e-value et une identité. Servez-vous de ces valeurs pour sélectionner les séquences qui pourraient être ajoutées à votre jeu de séquences cibles.

**2** Toujours par similarité (BLASTn donc), récupérer les séquences proches, mais n'étant pas de « *Bacillus anthracis* », afin de constituer un jeu de séquences proches «non-cibles ». Sauvegarder ces séquences dans un autre fichier fasta.

Pour ceux qui sont à l'aise avec les notions précédentes (ou qui ont terminé), effectuer la même recherche par Blast, mais en ligne de commande en utilisant le serveur du laboratoire  
=> [http://www.bioinfomed.fr/\\_\\_Teachings/bioinfo\\_urmite/blast\\_URMITE\\_tuto/blast\\_URMITE\\_tuto.html](http://www.bioinfomed.fr/__Teachings/bioinfo_urmite/blast_URMITE_tuto/blast_URMITE_tuto.html)