Sequences analysis

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 résulats 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 appprendre 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/

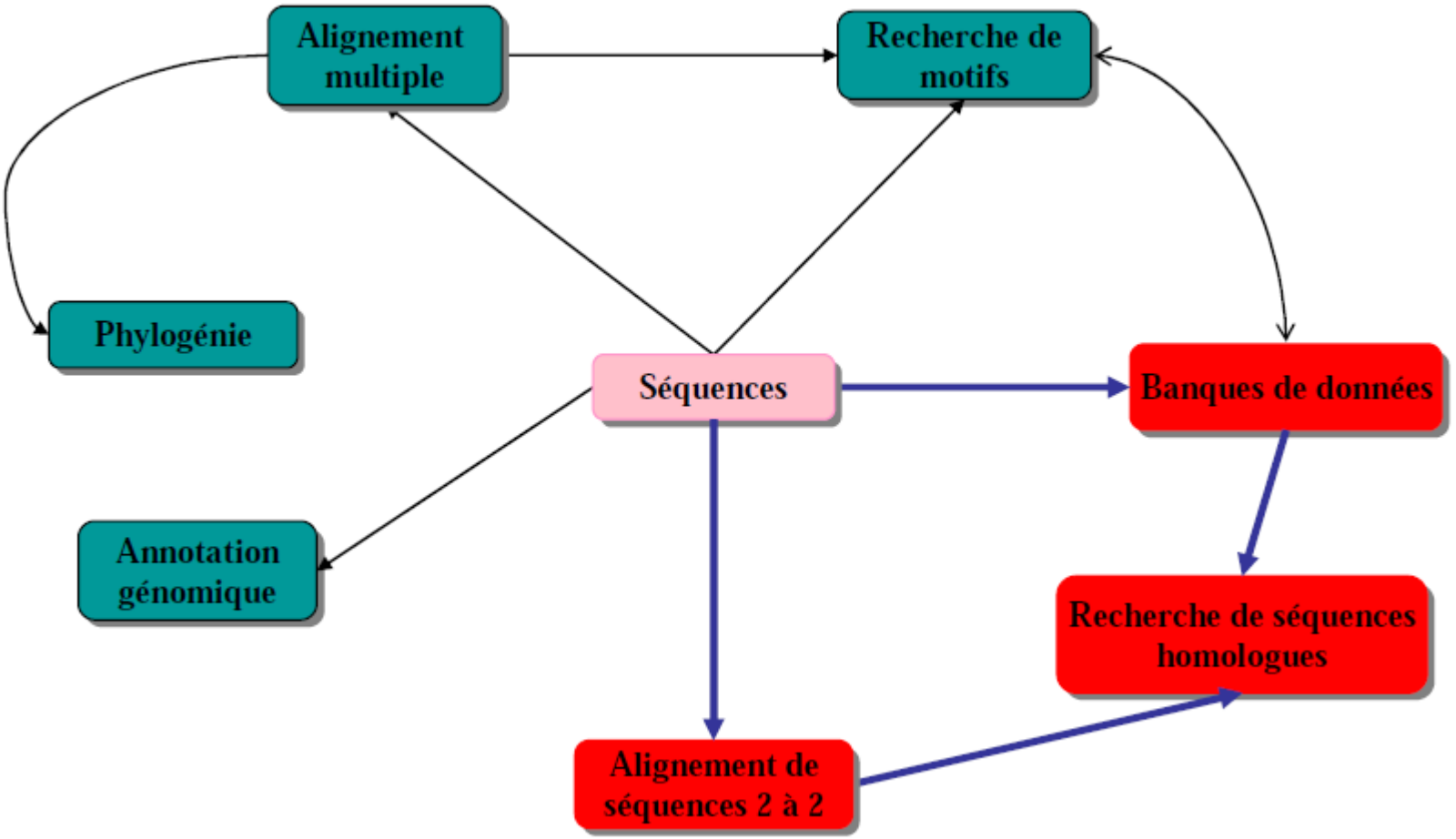
* Le Pôle Bioinformatique Lyonnais (PBIL)
http://pbil.univ-lyon1.fr/pbil.html http://npsa-pbil.ibcp.fr/

* 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.

1er gène ARN par W. Fiers *et al.* — 1972

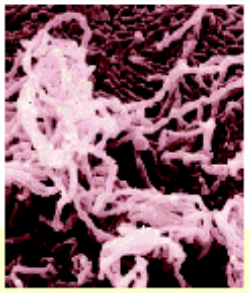
Technique de F. Sanger *et al. pour l'ADN* — 1975

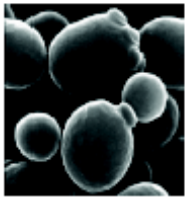
Technique de Maxam-Gilbert pour l'ADN — 1977

1er virus phi X174 par Sanger *et al.* — 1977

1er séquenceur Applied Biosystems — 1987

1ère bactérie *H. influenzae* 1,83 Mb — 1995



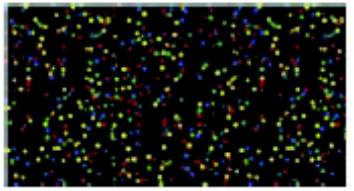


1er eucaryote *S. cerevisiae* 12 Mb — 1996

1er pluricellulaire *C. elegans* 100 Mb — 1998

Homo sapiens — 2001

Séquençage massif et parallèle — 2008



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

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

La séquence

```
ctccggcagc ccgaggtcat cctgctagac tcagacctgg atgaacccat agacttgcgc        60
tcggtcaaga gccgcagcga ggccggggag ccgcccagct ccctccaggt gaagcccgag       120
acaccggcgt cggcggcggt ggcggtggcg gcggcagcgg cacccaccac gacggcggag       180
```

Bases de données de séquences nucléotidiques

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'accession
  - 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

Bases de données de séquences nucléotidiques

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…

Bases de données de séquences nucléotidiques

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

Bases de données de séquences nucléotidiques

Banques nucléiques, localisation des « keys »

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

109..1105 : entre les positions 109 et 1105 (inclues)
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

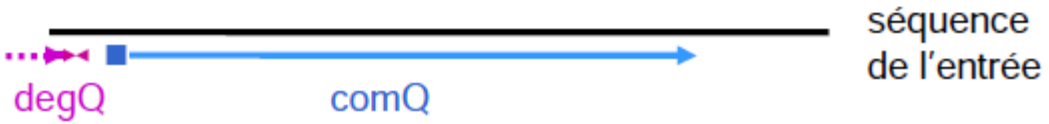
complement(340..565) : séquence complémentaire inversée à celle de l'entrée (brin -)

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

Bases de données de séquences nucléotidiques

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
ctccggcagcccgaggtcatcctgctagactcagacctggatgaacccac
Ctccggcagcccgaggtcatcctgctagactcagacctggatgaacccat
agacttgcgctcggtcaagagccgcagcgaggccggggagccgcccagct
ccctccaggtgaagcccgagacaccggcgtcggcggcggtggcggtggcg
Gcggcagcggcacccaccacgacggcggagaagcct
>ID23003 hPc2 gene
ggacgaacctgcagagtcgctgagcgagttcaagcccttctttgggaata
taattatcaccgacgtcaccgcgaactgcctcaccgttactttcaaggag
tacgtgacggtg
```

Bases de données de séquences nucléotidiques

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

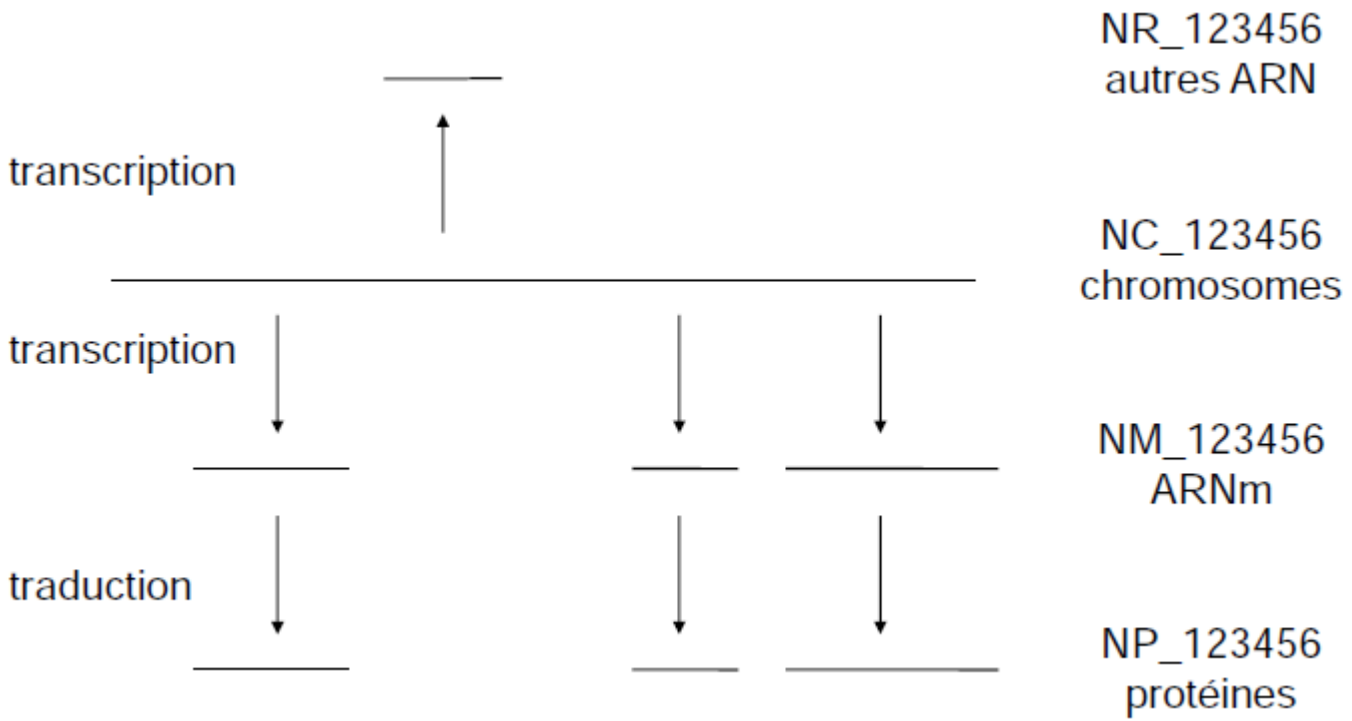
* Erreurs dans les annotations et dans les séquences

Bases de données de séquences nucléotidiques

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

=> bases de séquences non-redontantes.

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



Bases de données de séquences proteiques

=> 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 ancetre 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, SRS, 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

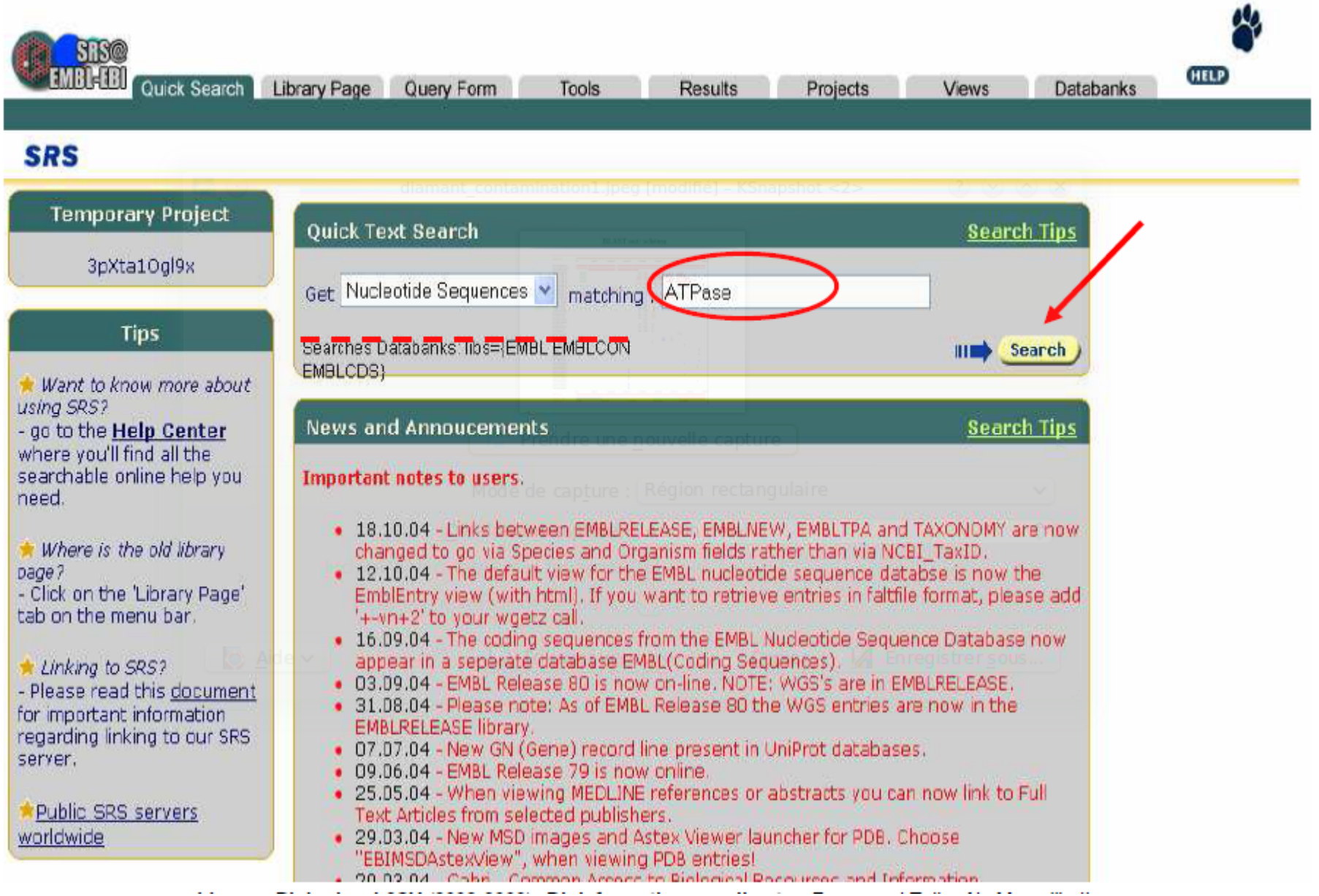
* **SRS** : un autre système d'interrogation, + élaboré, mais plus complexe (http://www.dkfz.de/srs/ ou liste à : http://bioblog.instem.com/download/srs-parser-and-software-downloads/public-srs-installations/)

* **ACNUC** (http://doua.prabi.fr/search/query_fam)

1 - Interrogation par annotations

Exemple de serveur SRS : **http://www.dkfz.de/srs/**

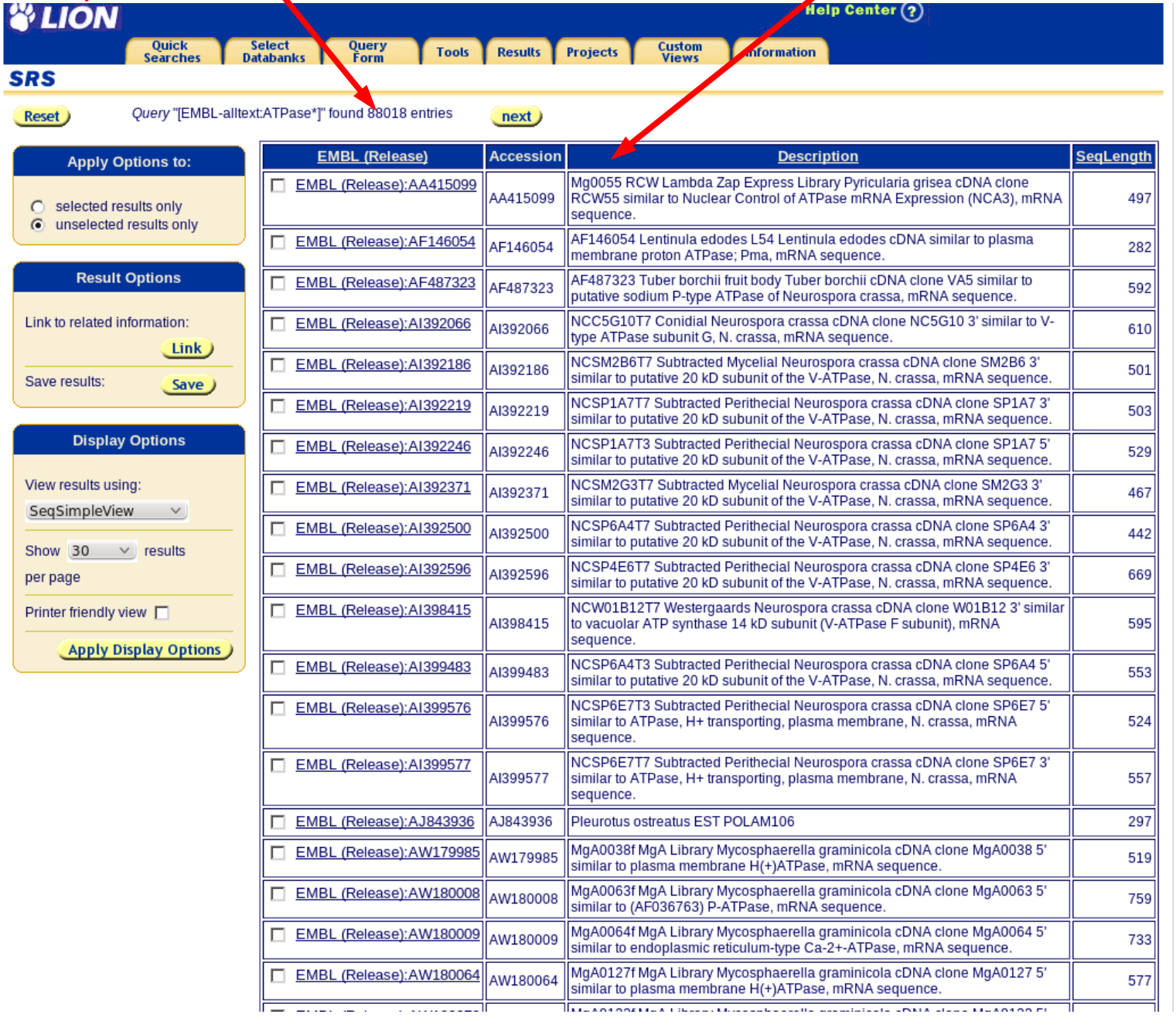
Pour une recherche simple



1 - Interrogation par annotations

Nombre d'entrées (séquences + annotations) trouvées

Entrées disponibles, téléchargeables



LION

Help Center

Quick Searches | Select Databanks | Query Form | Tools | Results | Projects | Custom Views | Information

SRS

Reset

Query "[EMBL-alltext:ATPase*]" found 88018 entries

next

Apply Options to:

- selected results only
- unselected results only

Result Options

Link to related information: Link

Save results: Save

Display Options

View results using: SeqSimpleView

Show 30 results per page

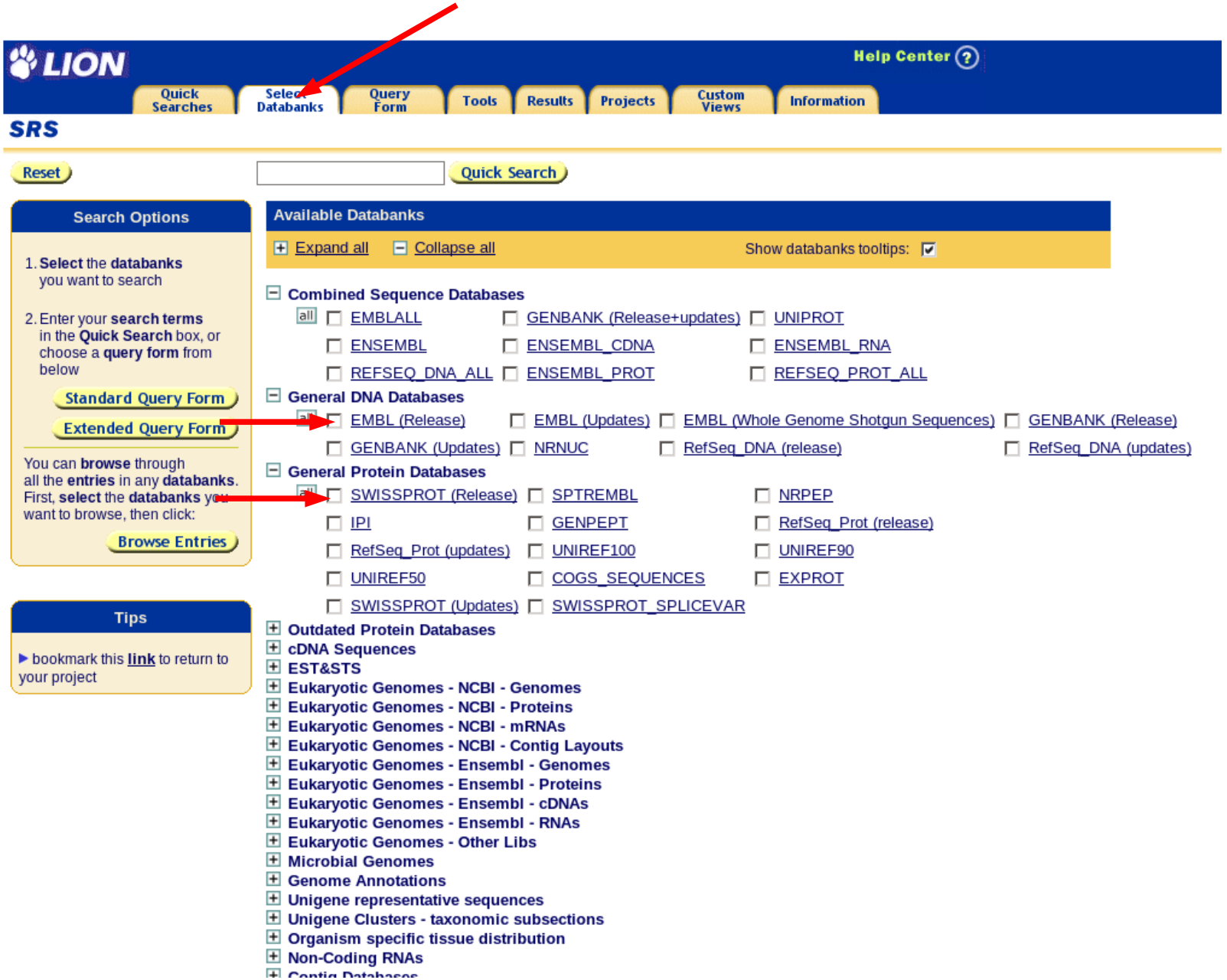
Printer friendly view

Apply Display Options

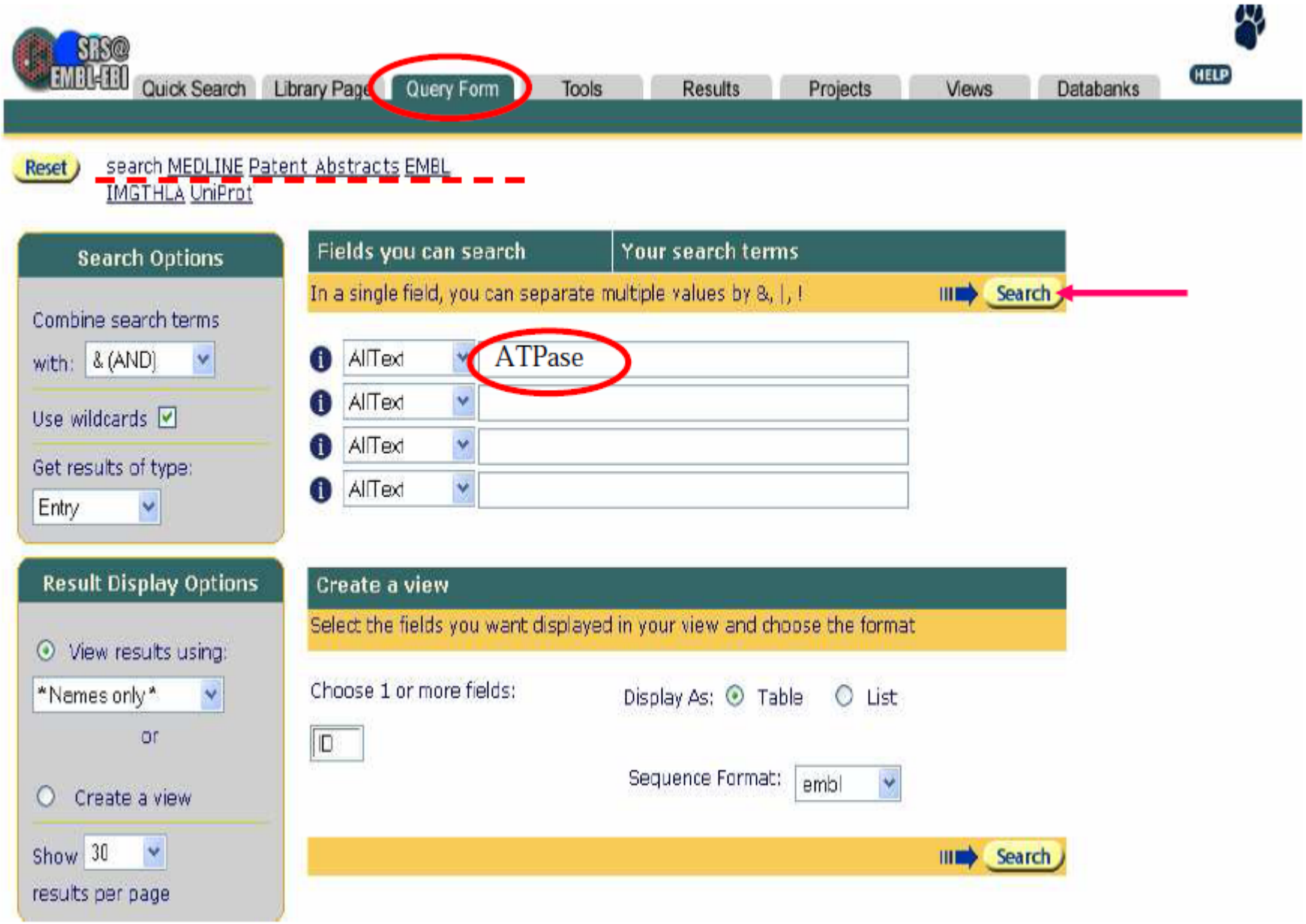
EMBL (Release)	Accession	Description	SeqLength
EMBL (Release):AA415099	AA415099	Mg0055 RCW Lambda Zap Express Library Pyricularia grisea cDNA clone RCW55 similar to Nuclear Control of ATPase mRNA Expression (NCA3), mRNA sequence.	497
EMBL (Release):AF146054	AF146054	AF146054 Lentinula edodes L54 Lentinula edodes cDNA similar to plasma membrane proton ATPase; Pma, mRNA sequence.	282
EMBL (Release):AF487323	AF487323	AF487323 Tuber borchii fruit body Tuber borchii cDNA clone VA5 similar to putative sodium P-type ATPase of Neurospora crassa, mRNA sequence.	592
EMBL (Release):AI392066	AI392066	NCC5G10T7 Conidial Neurospora crassa cDNA clone NC5G10 3' similar to V-type ATPase subunit G, N. crassa, mRNA sequence.	610
EMBL (Release):AI392186	AI392186	NCSM2B6T7 Subtracted Mycelial Neurospora crassa cDNA clone SM2B6 3' similar to putative 20 kD subunit of the V-ATPase, N. crassa, mRNA sequence.	501
EMBL (Release):AI392219	AI392219	NCSP1A7T7 Subtracted Perithecial Neurospora crassa cDNA clone SP1A7 3' similar to putative 20 kD subunit of the V-ATPase, N. crassa, mRNA sequence.	503
EMBL (Release):AI392246	AI392246	NCSP1A7T3 Subtracted Perithecial Neurospora crassa cDNA clone SP1A7 5' similar to putative 20 kD subunit of the V-ATPase, N. crassa, mRNA sequence.	529
EMBL (Release):AI392371	AI392371	NCSM2G3T7 Subtracted Mycelial Neurospora crassa cDNA clone SM2G3 3' similar to putative 20 kD subunit of the V-ATPase, N. crassa, mRNA sequence.	467
EMBL (Release):AI392500	AI392500	NCSP6A4T7 Subtracted Perithecial Neurospora crassa cDNA clone SP6A4 3' similar to putative 20 kD subunit of the V-ATPase, N. crassa, mRNA sequence.	442
EMBL (Release):AI392596	AI392596	NCSP4E6T7 Subtracted Perithecial Neurospora crassa cDNA clone SP4E6 3' similar to putative 20 kD subunit of the V-ATPase, N. crassa, mRNA sequence.	669
EMBL (Release):AI398415	AI398415	NCW01B12T7 Westergaards Neurospora crassa cDNA clone W01B12 3' similar to vacuolar ATP synthase 14 kD subunit (V-ATPase F subunit), mRNA sequence.	595
EMBL (Release):AI399483	AI399483	NCSP6A4T3 Subtracted Perithecial Neurospora crassa cDNA clone SP6A4 5' similar to putative 20 kD subunit of the V-ATPase, N. crassa, mRNA sequence.	553
EMBL (Release):AI399576	AI399576	NCSP6E7T3 Subtracted Perithecial Neurospora crassa cDNA clone SP6E7 5' similar to ATPase, H+ transporting, plasma membrane, N. crassa, mRNA sequence.	524
EMBL (Release):AI399577	AI399577	NCSP6E7T7 Subtracted Perithecial Neurospora crassa cDNA clone SP6E7 3' similar to ATPase, H+ transporting, plasma membrane, N. crassa, mRNA sequence.	557
EMBL (Release):AJ843936	AJ843936	Pleurotus ostreatus EST POLAM106	297
EMBL (Release):AW179985	AW179985	MgA0038f MgA Library Mycosphaerella graminicola cDNA clone MgA0038 5' similar to plasma membrane H(+)ATPase, mRNA sequence.	519
EMBL (Release):AW180008	AW180008	MgA0063f MgA Library Mycosphaerella graminicola cDNA clone MgA0063 5' similar to (AF036763) P-ATPase, mRNA sequence.	759
EMBL (Release):AW180009	AW180009	MgA0064f MgA Library Mycosphaerella graminicola cDNA clone MgA0064 5' similar to endoplasmic reticulum-type Ca-2+-ATPase, mRNA sequence.	733
EMBL (Release):AW180064	AW180064	MgA0127f MgA Library Mycosphaerella graminicola cDNA clone MgA0127 5' similar to plasma membrane H(+)ATPase, mRNA sequence.	577

1 - Interrogation par annotations

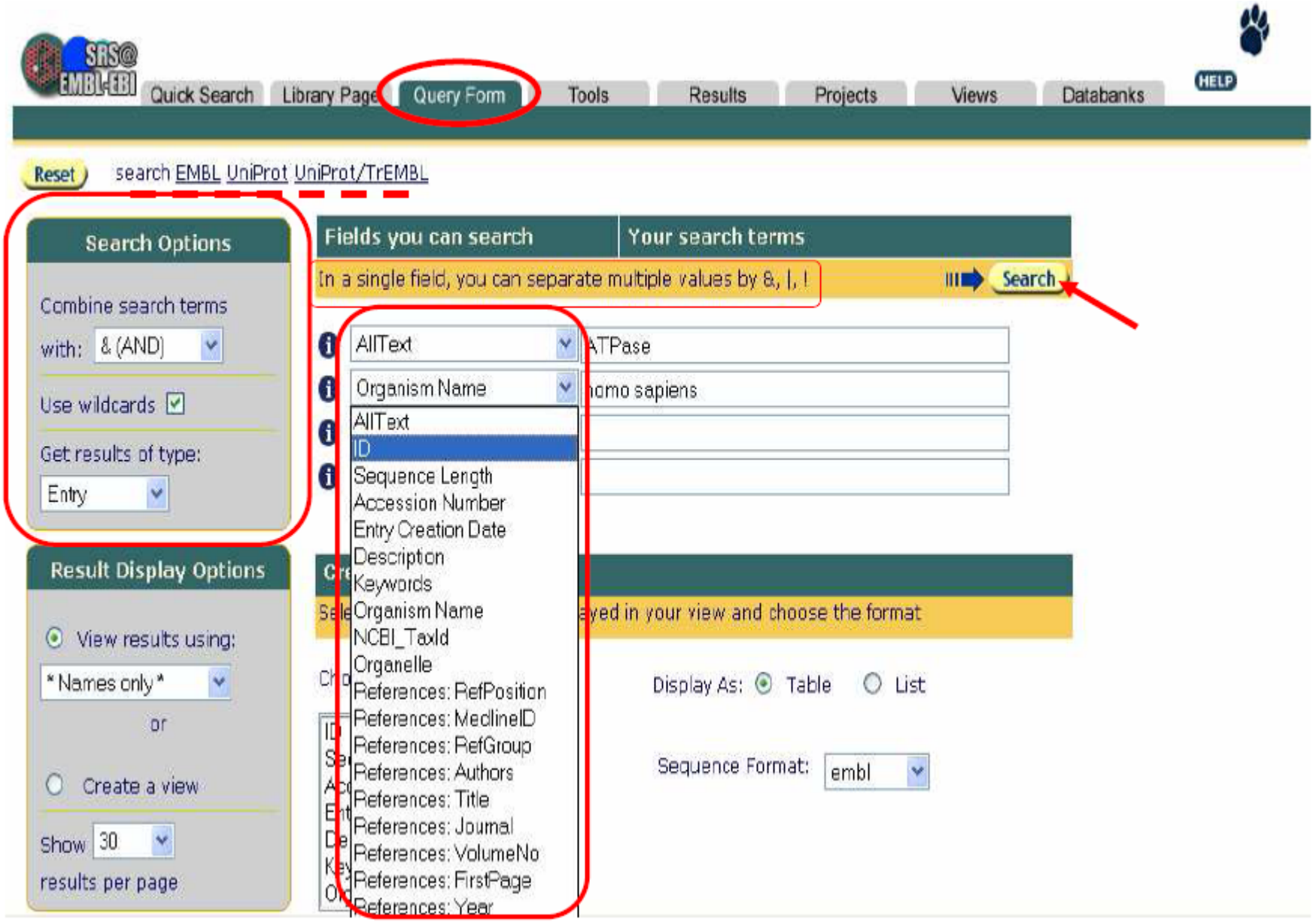
Changer de bases de données



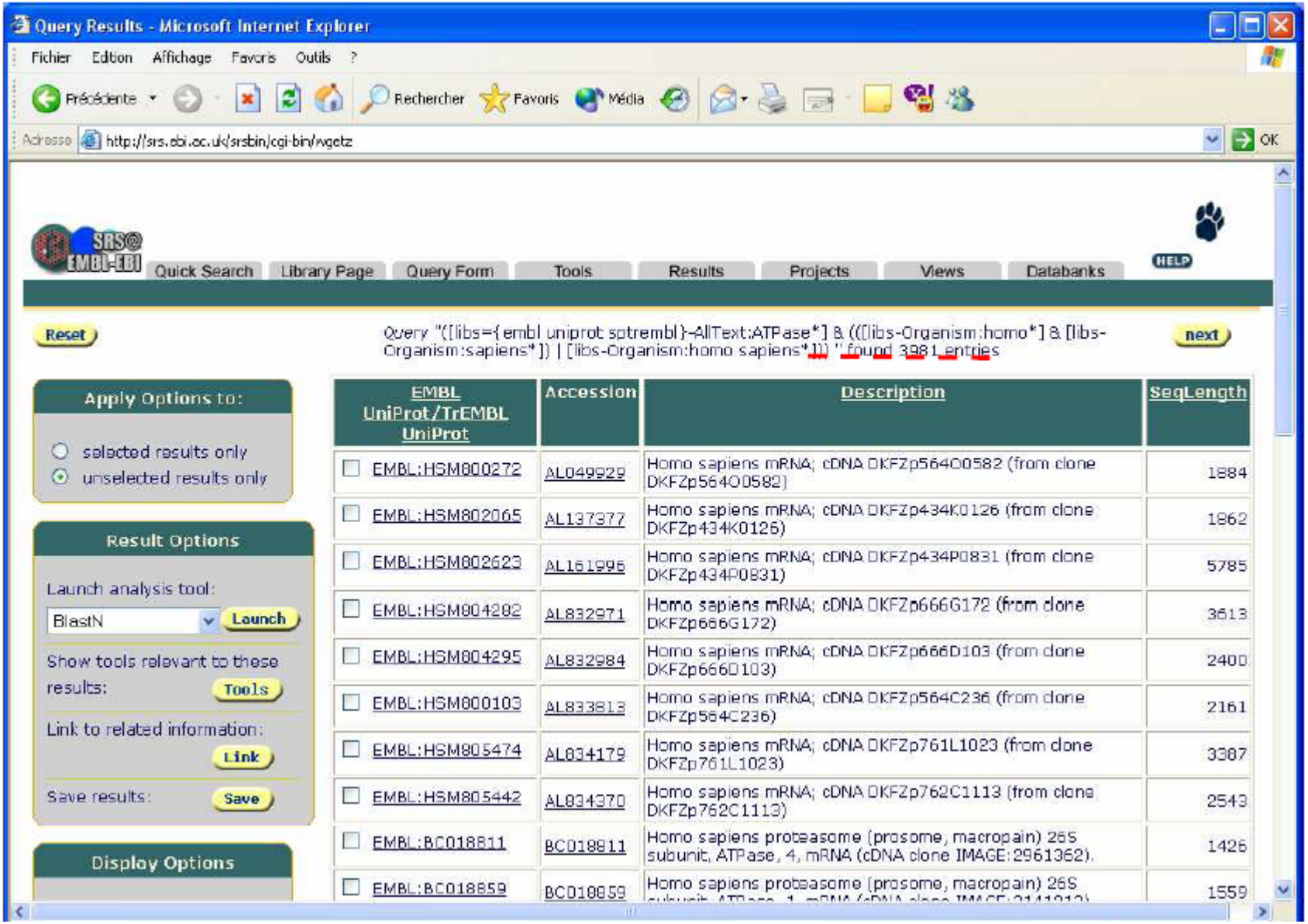
1 - Interrogation par annotations



1 - Interrogation par annotations



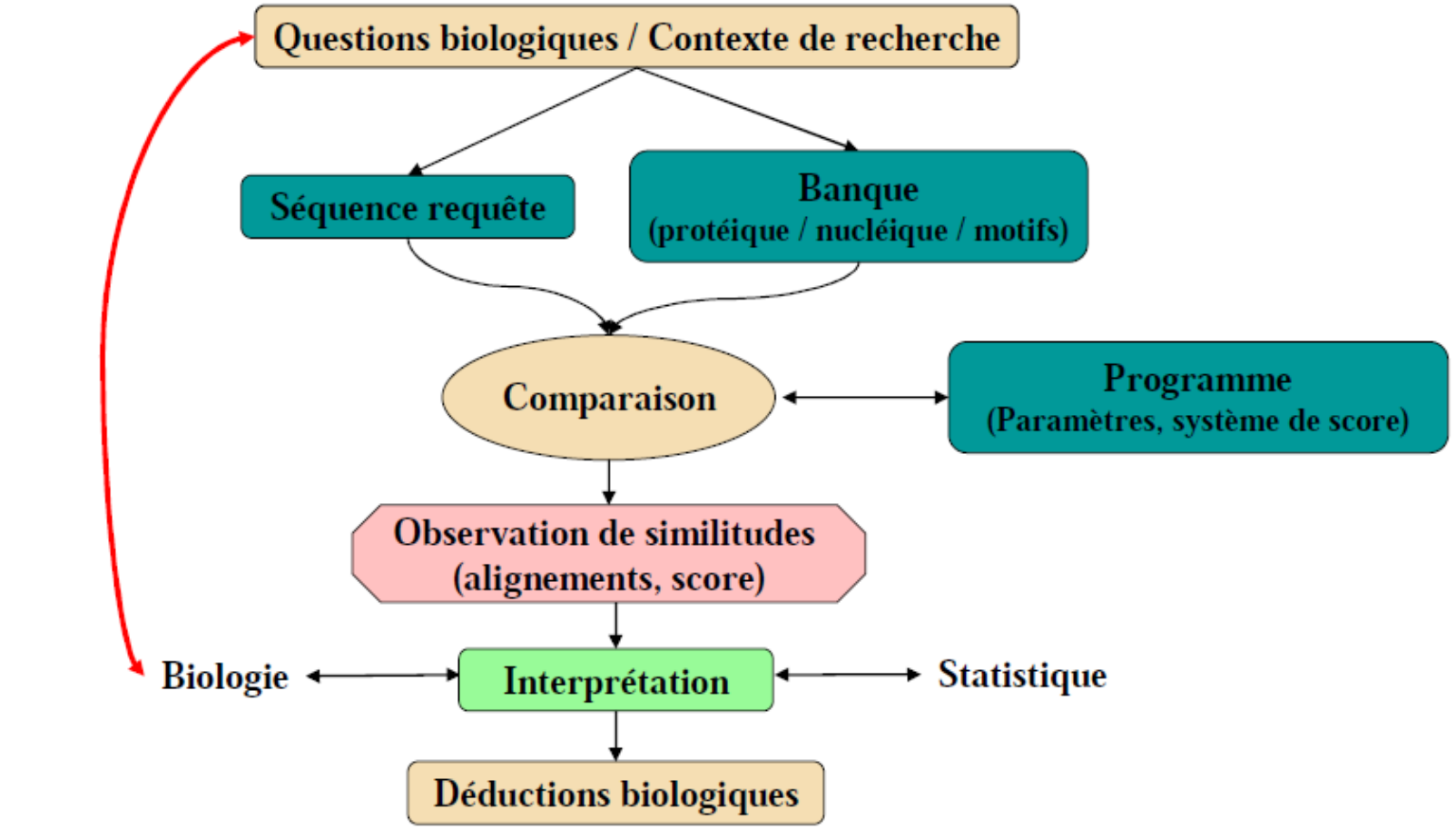
1 - Interrogation par annotations



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, …



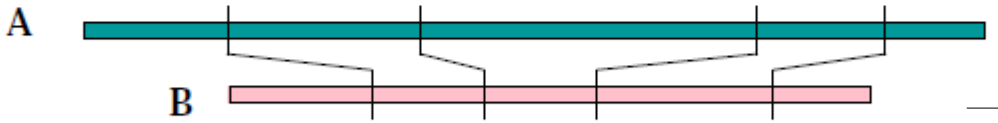
2 - Interrogation par similarité – les alignements

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 & Wunsh.

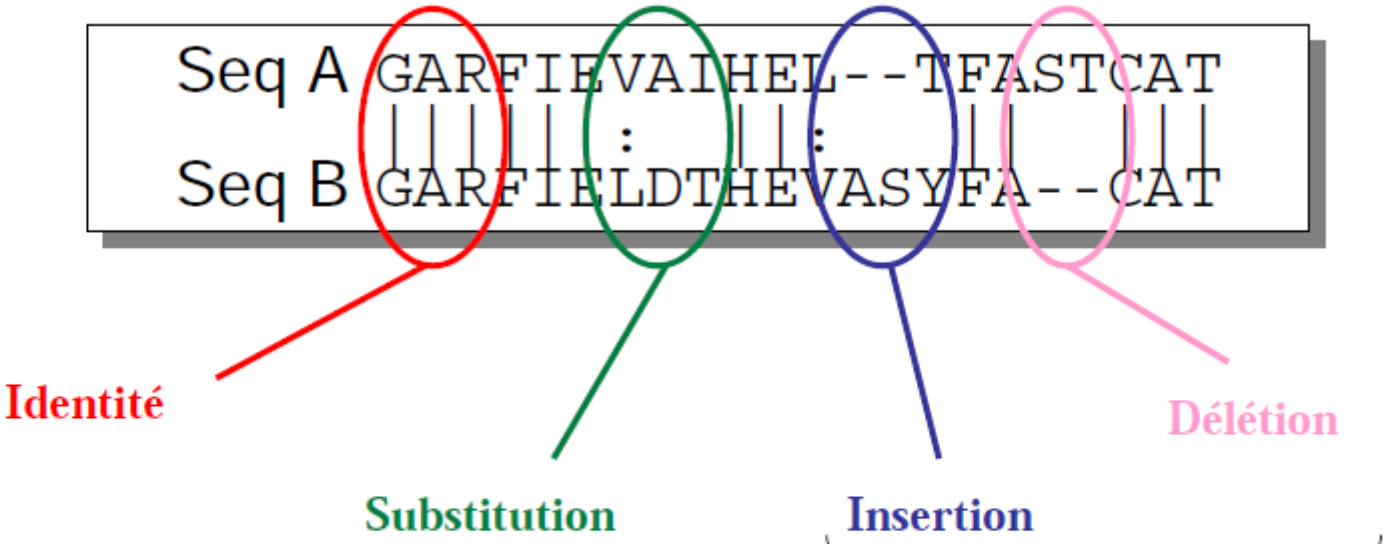


→ utilisé pour la recherche de similarité

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



Indel ou Gap

Identité ou **Match** (| ou * ou **C**)
Substitution non conservative ou **Mismatch** (**néant**)
Substitution conservative (+ ou : ou .)
Indel ou **Gap** (- ou .)

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 caractère 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.

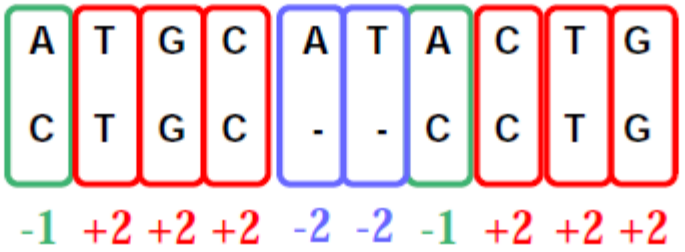
A noter aussi que des séquence sans ressemblance apparente peuvent aussi être homologues (on le retrouve par ex. au niveau 3D) =>Des séquences homologues ne sont pas nécessairement similaires.

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.**

Sc (identité) = 2
Sc (substitution) = -1
Sc (gap) = -2



Score total

+ 6

=> 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)

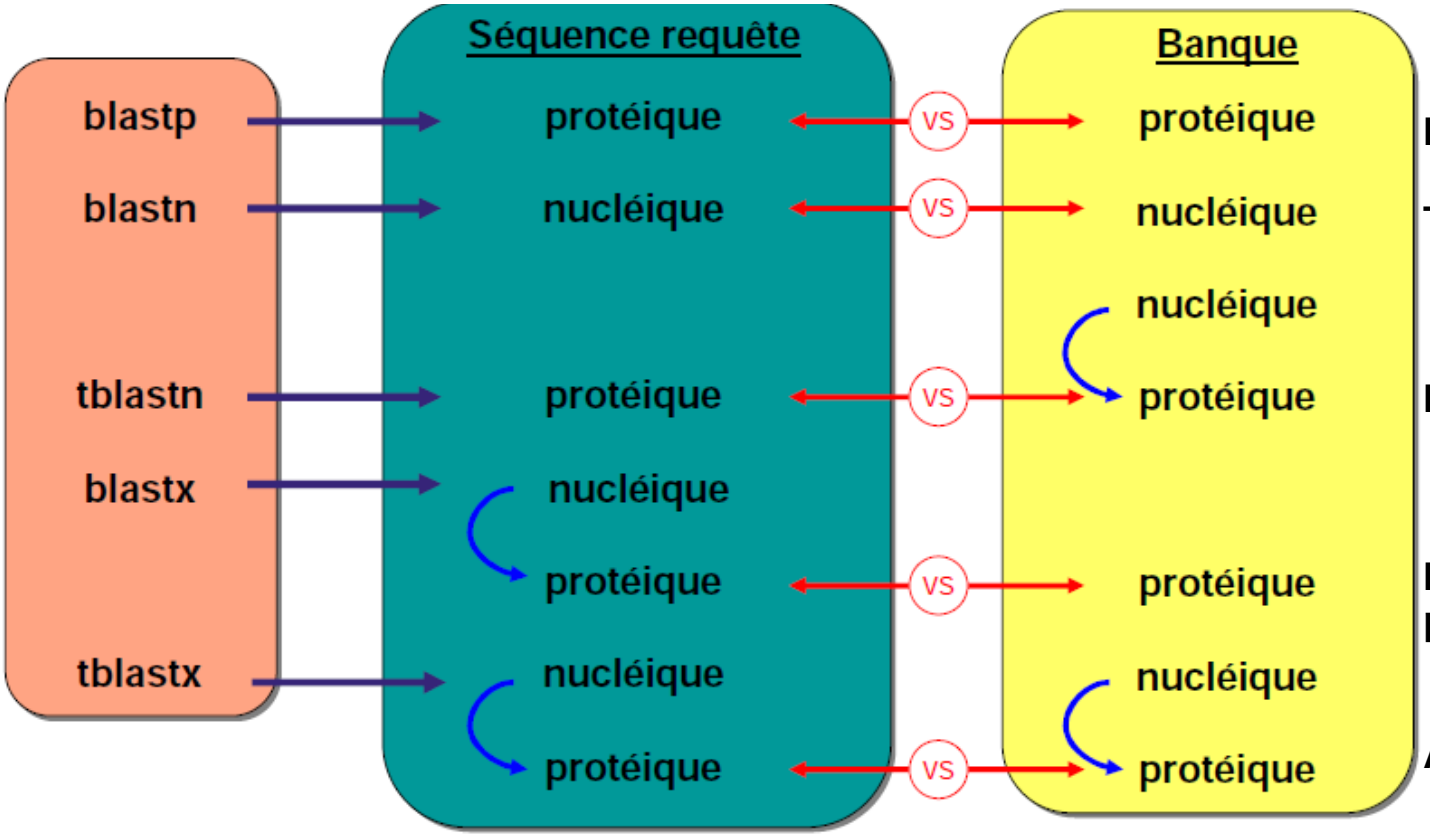
=> exemples suivants avec "Blast" (avec interface graphique du NCBI)

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

2 - Interrogation par similarité – Blast

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

Différents programmes :



Utilisation types :

Etude de familles de proteines

Test existance séquences, arn, etc.

Prédiction de gènes et de CDS

Localisation du gène codant la proteine dans le génome

Annotations, autres

BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

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

NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more.

Enter Query Sequence

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

Query subrange

From

To

Or, upload file Choisir...

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

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)

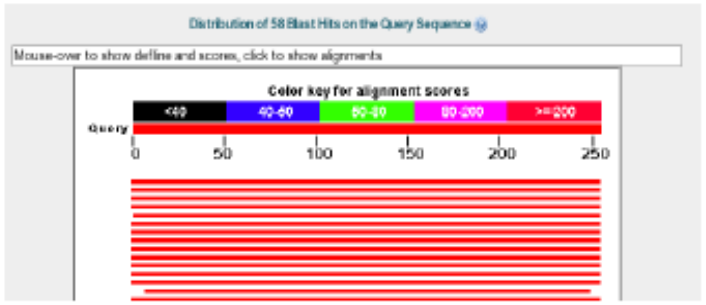
Mask
- ☑ Mask for lookup table only
- ☐ Mask lower case letters

BLAST

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

☐ Show results in a new window

1. récapitulatif de la requête

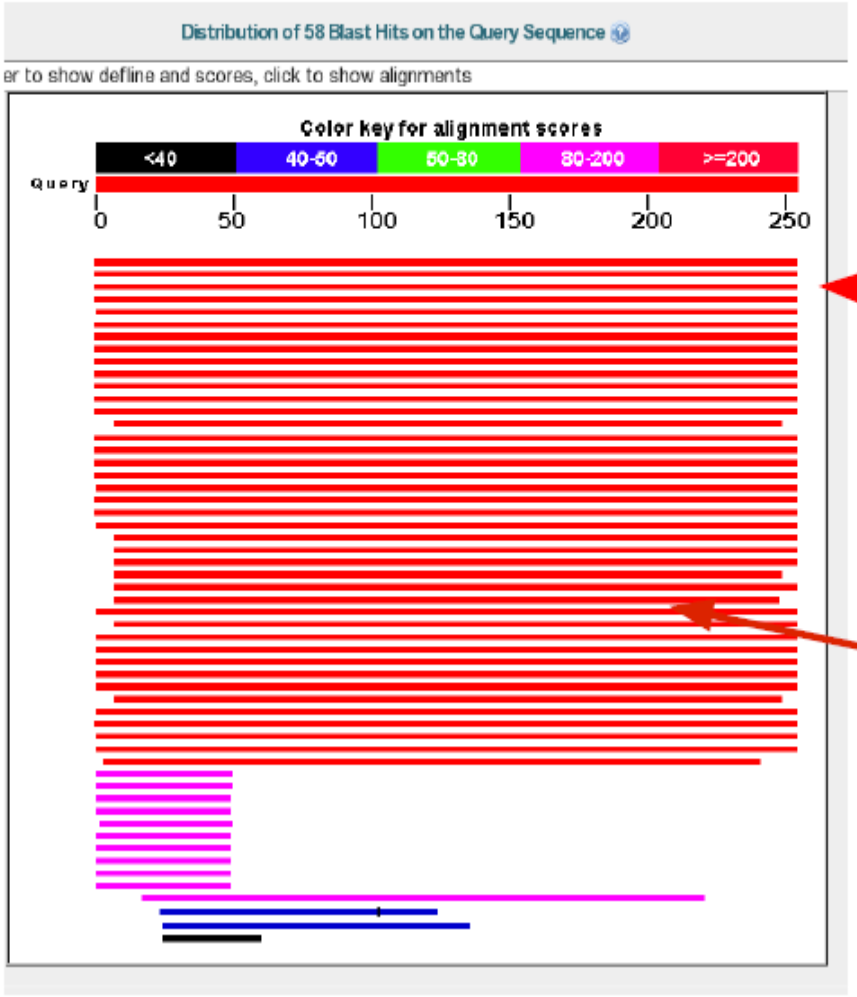


2. représentation graphique des résultats

Accession	Description	Max score	Total score	Query coverage	E value
P36914.2	glucoamylase [Aspergillus oryzae RIB40] >gi|94730359|sp|P36914.2|AMY	1245	1245	100%	0.0
P22832.1	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucoh	845	845	99%	0.0
P69327.1	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucoh	843	843	99%	0.0
P23176.1	RecName: Full=Glucoamylase I; AltName: Full=1,4-alpha-D-glucan glucol	842	842	99%	0.0
P14804.3	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucoh	631	631	95%	0.0
Q03045.1	RecName: Full=Glucoamylase P; AltName: Full=1,4-alpha-D-glucan gluc	560	560	90%	0.0
Q60087.1	RecName: Full=Probable glucoamylase; AltName: Full=1,4-alpha-D-gluca	294	294	73%	4e-92
P07883.2	RecName: Full=Glucoamylase 1; Short=Gluc 1; AltName: Full=1,4-alpha-D	287	287	68%	3e-80
P40042.1	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucoh	245	245	67%	1e-71
P78989.2	RecName: Full=Glucoamylase GLA1; AltName: Full=1,4-alpha-D-glucan g	224	224	70%	6e-65
P08017.1	RecName: Full=Glucoamylase GLU1; AltName: Full=1,4-alpha-D-glucan g	220	220	70%	2e-63
P08019.2	RecName: Full=Glucoamylase, intracellular sporulation-specific; AltName:	209	209	70%	8e-57
P04062.2	RecName: Full=Glucoamylase S1; AltName: Full=1,4-alpha-D-glucan gluc	183	183	54%	3e-48
P26790.1	RecName: Full=Glucoamylase S2; AltName: Full=1,4-alpha-D-glucan gluc	182	182	54%	4e-48
P22998.1	RecName: Full=Alpha-amylase; AltName: Full=1,4-alpha-D-glucan glucan	80.1	80.1	14%	2e-14
Q30585.1	RecName: Full=Cyclomaltodextrin glucanotransferase; AltName: Full=Cyc	74.7	74.7	16%	9e-13
P26790.1	RecName: Full=Alpha-amylase; AltName: Full=1,4-alpha-D-glucan glucan	71.2	71.2	12%	1e-11
P05618.1	RecName: Full=Cyclomaltodextrin glucanotransferase; AltName: Full=Cyc	71.2	71.2	14%	1e-11

3. résumé des résultats

4. les alignements



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

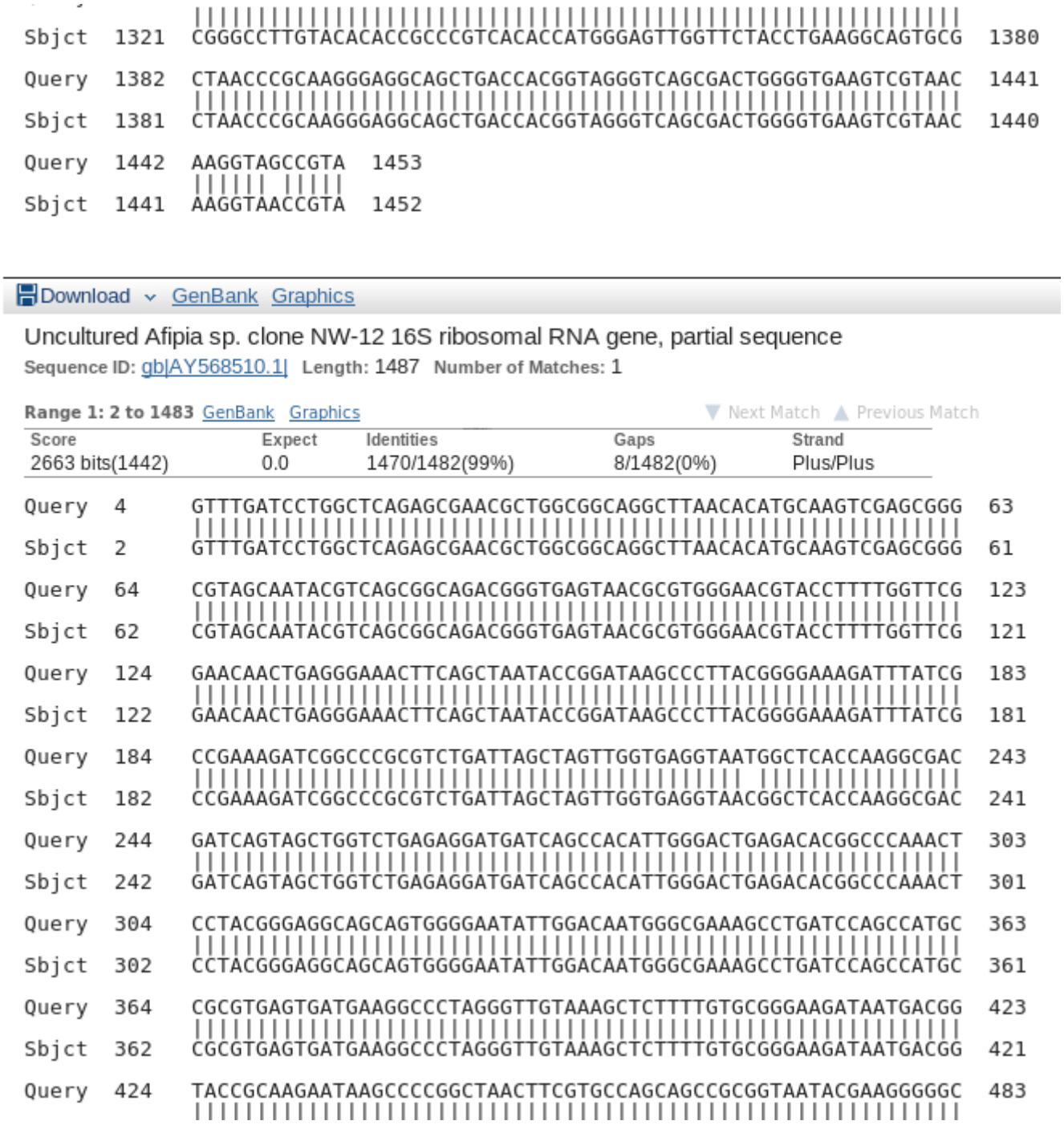
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

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

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    CGTAGCAATACGTCAGCGGCAGACGGGTGAGTAACGCGTGGGAACGTACCTTTTGGTTCG  123
Sbjct  62    CGTAGCAATACGTCAGCGGCAGACGGGTGAGTAACGCGTGGGAACGTACCTTTTGGTTCG  121
Query  124   GAACAACTGAGGGAAACTTCAGCTAATACCGGATAAGCCCTTACGGGGAAAGATTTATCG  183
Sbjct  122   GAACAACTGAGGGAAACTTCAGCTAATACCGGATAAGCCCTTACGGGGAAAGATTTATCG  181
Query  184   CCGAAAGATCGGCCCGCGTCTGATTAGCTAGTTGGTGAGGTAATGGCTCACCAAGGCGAC  243
Sbjct  182   CCGAAAGATCGGCCCGCGTCTGATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCGAC  241
Query  244   GATCAGTAGCTGGTCTGAGAGGATGATCAGCCACATTGGGACTGAGACACGGCCCAAACT  303
Sbjct  242   GATCAGTAGCTGGTCTGAGAGGATGATCAGCCACATTGGGACTGAGACACGGCCCAAACT  301
Query  304   CCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGC  363
Sbjct  302   CCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGC  361
Query  364   CGCGTGAGTGATGAAGGCCCTAGGGTTGTAAAGCTCTTTTGTGCGGGAAGATAATGACGG  423
Sbjct  362   CGCGTGAGTGATGAAGGCCCTAGGGTTGTAAAGCTCTTTTGTGCGGGAAGATAATGACGG  421
Query  424   TACCGCAAGAATAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGGGC  483
```

=> La signification statistique de reflète pas forcément la signification biologique et inversement !

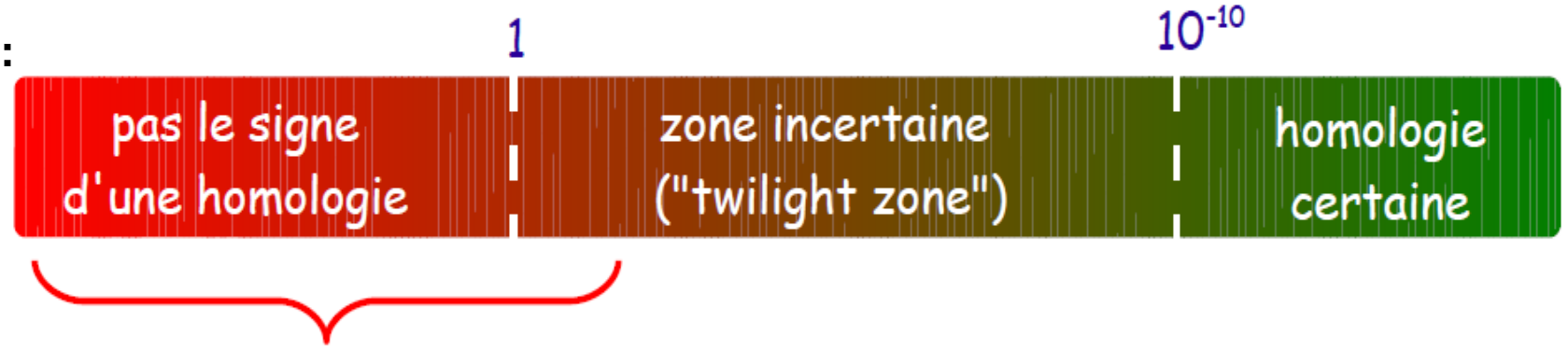
- **En générale, lorsque l'alignement est fait sur au moins 70% de la séquence:**

  - 30 < ID <= 50 ➔ Séquences faiblement similaires
  - 50 < ID <= 70 ➔ Séquences similaires
  - 70 < ID <= 100 ➔ Séquences fortement similaires

-On peut déjà parler de séquences homologues au delà de 70% de similarité, mais cela reste à confirmer par d'autres hypothèses: présence de motifs communs, etc….

-Si la Evalue est très faible (<10^{-20}), c'est probablement le signe d'une similarité entre les séquences. Mais, il ne faut jamais se fier uniquement à la Evalue.

e-value :



faux-positifs: on a un alignement, mais les séquences ne sont pas homologues

=> similarité due au hasard

Exercice

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

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

<u>Methodes à utiliser</u> :

1.1 En effectuant une recherche par mots-clefs grâce aux applications suivantes : SRS (http://www.dkfz.de/srs/), Entrez (http://www.ncbi.nlm.nih.gov) ou WWWQuery (=ACNUC) (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 etre 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 consituter 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 précedentes (ou qui ont terminé), effectuer la meme 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