HMSN204
Info. Biologique et Outils bioinformatiques

Banques de données biologiques
(3h de Cours/TD + 4.5h de TP/TD)

Anne-Muriel Chifolleau

http ://www.lirmm.fr/∼arigon/enseignement/HMSN204/

LIRMM - UM

Plan

Sources de données biologiques (1/2)

- Comprendre ce qui constitue le “livre de la vie” , comprendre comment une collection de simple nucléotides est à la base de la vie, ...

  $\rightarrow$ collecter d'immense quantité de données de séquences et les stocker de manière à pouvoir les analyser et les retrouver facilement.

- Volume de données généré trop important pour être géré par les techniques de publication traditionnelle.

Création et maintenance des sources de données biologiques pour l'archivage, le stockage, la diffusion et l'exploitation des données biologiques

Sources de données biologiques (2/2)

- Source de données en biologie = collection organisée d'informations mémorisées sur un support permanent et pouvant être restituées au biologiste de manière opportune

  $\Rightarrow$ Construire des banques de données ou des bases de données pour collecter les informations, stocker et organiser les données, distribuer l'information, faciliter l'exploitation des données

  $\rightarrow$ Banque de données : collection de fichiers structurés
  $\rightarrow$ Base de données : collection de données structurées

Différence = élément de base considéré
Fichier de données ou Donnée

- Séquence = élément central autour duquel les banques et bases de données se sont constituées

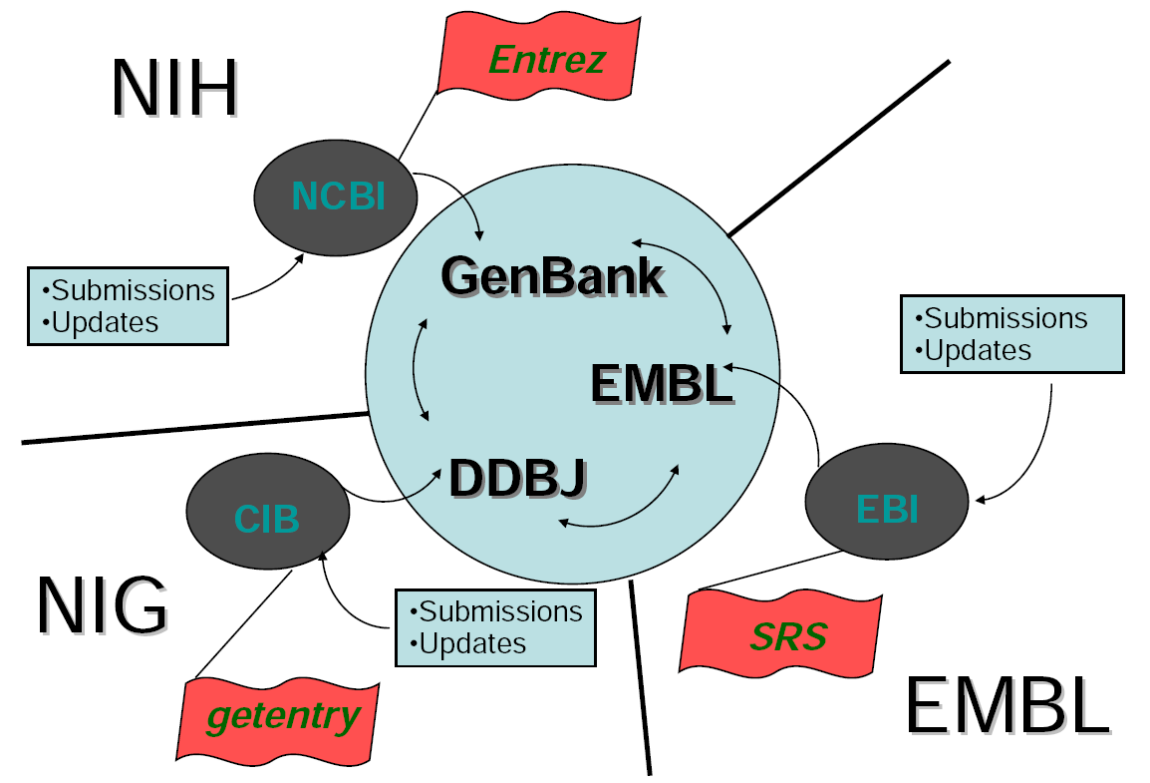
Bref historique (1/3)

- $\sim$ 1960s Margaret Dayhoff et ses collaborateurs collectent toutes les séquences de protéines connues et les regroupent dans l'Atlas of protein
  $\rightarrow$ 50 entrées
  $\rightarrow$ Version papier jusqu'en 78, puis version électronique
  $\rightarrow$ En 1984, l'Atlas devient PIR (Protein Information Ressource)

- A partir de 1982 - EMBL (European Molecular Biology Laboratory) et NCBI (National Center for Biotechnology Information) : transcription et interprètation des séquences publiées dans les journaux au format électronique
  $\rightarrow$ Véritable explosion de la quantité de séquences disponibles
  $\rightarrow$ Rapidement, la DDBJ (DNA DataBank of Japan) se joint à la collaboration pour la collecte de données

Bref historique (2/3)

- 1988 Regroupement de ces 3 groupes (appelés *International Nucleotide Sequence Database Collaboration*, INSDC)
  - → Accord pour un format commun
  - → Échange des données journalier
  - → Chaque groupe gère les mises à jour des séquences qu'il a créées



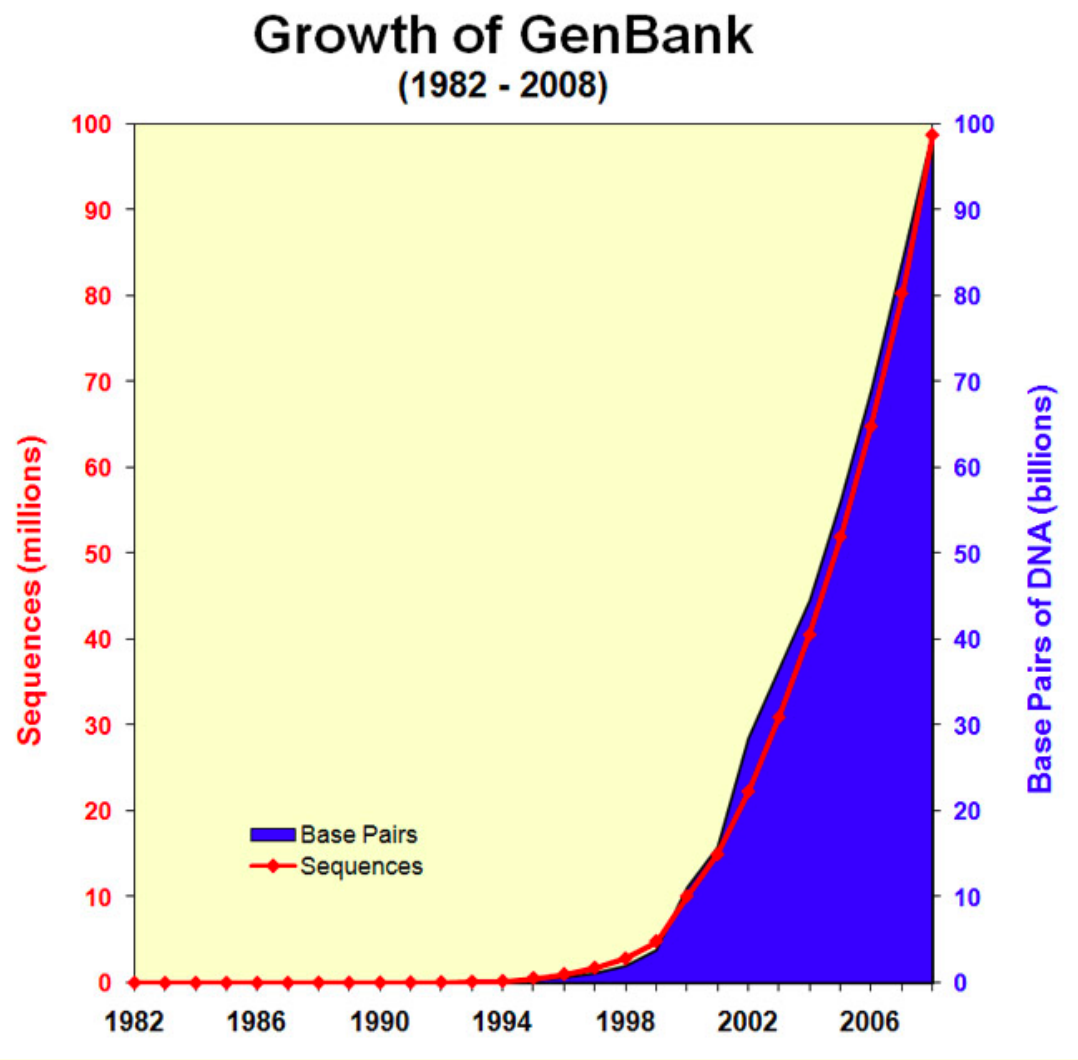
Bref historique (3/3)

- 1986 - SWISSPROT (3900 prot.) : conversion du PIR en un format similaire à celui d'EMBL + ajout d'information pour chaque protéine
  ⇒ Notoriété de SWISSPROT comme banque de haute qualité

- Collaboration entre SWISSPROT et EMBL
  But : mettre à disposition rapidement les séquences de protéines non encore annotées par SWISSPROT
  ⇒ TrEMBL (Translation of EMBL nucleotide sequences)

- Fin 2003 - UNIPROT (UNIversal PROTein Ressource) : jonction des informations de SWISSPROT, TrEMBL et PIR
  → UniProtKB (*UniProt KnowledgeBase*) = SWISSPROT + TrEMBL
  → UniRef (*UniProt Reference Clusters*) : regroupement en 1 entrée des séquences similaires pour accélérer les recherches
  → UniParc (*UniProt Archive*) : utilisé pour garder une trace des séquences et de leur idendifiants

Un ensemble de données gigantesque



[https ://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/]

- En 1982 :
  606 séquences
  2 427 bases
- En décembre 2008 :
  98 868 465 séquences
  99 116 431 942 bases
- En décembre 2017 :
  206 293 625 séquences
  249 722 163 594 bases

Forme actuelle des banques de séquences

- Recherche publique donc données publiques
  - → Partage international des données
  - → Évite qu'une expérience soit refaite par différents laboratoires
  - → Permet la comparaison des résultats

- Beaucoup de données générées par l'expérimentation
  - → Outils adaptés à de grands volumes de données
  - → Gestion des grandes banques par des organismes spécialisés

- Accès fréquent à ces données
  - → Interrogation *via* internet

- Une source d'informations pour l'interprétation
  - → Ma séquence est-elle déjà présente dans les banques ?
  - → Quelles sont les séquences similaires à la mienne ?
  - → Quelles sont les protéines similaires à la mienne (structure, fonction) ?
  - → . . .

- Quelle est la qualité des données ?
  - → Erreurs de frappe
  - → Erreurs dans les séquences
  - → Erreurs d'annotation
  - → Redondance des données

- Où sont les données ?
  - → Dans les laboratoires
  - → Dans des banques ou bases de données spécialisées : souvent maintenues par un laboratoire
  - → Dans des banques de données généralistes : maintenues par des consortiums

- Différentiations possibles des banques de données
  - → Banque primaire ou généraliste
  - → Banque secondaire ou spécialisée

Banques de données primaires et secondaires

- Banques primaires/généralistes $\sim$ archives
- Banques secondaires/spécialisées $\sim$ données vérifiées (corrigées et annotées)

- La plus grande contribution des banques de données à la communauté des biologistes est de rendre les séquences accessibles

- Les banques primaires contiennent majoritairement des résultats expérimentaux (avec qqs interprétations), mais qui ne sont pas vérifiés, ni analysés
  $\rightarrow$ *les séquences nucléiques d'EMBL/GenBank/DDBJ sont issues du séquençage d'une molécule biologique qui existe dans un tube à essai, qqpart dans un labo ; elles ne représentent pas les séquences qui sont un consensus d'une population*

- Ceci a des conséquences sur l'interprétation de l'analyse des séquences : informations parfois très utiles ou trompeuses

- Ces banques contiennent des données hétérogènes (collecte la plus exhaustive possible)
- Banques de séquences nucléiques (GenBank, EMBL, DDBJ)
- Banques de séquences protéiques (PIR, SWISSPROT, UNIPROT)
- Banques d'articles scientifiques (PubMed)

Avantage : tout est consultable en une fois
Inconvénient : difficiles à maintenir, difficiles à interroger

Les banques de données spécialisées

- Ces banques contiennent des données homogènes
- Collecte établie autour d'une thématique particulière telle ques :
  - → Banques de localisation (*mapping databases* GeneLoc)
  - → Banques de structures 3D de macromolécules (PDB)
  - → Banques génomiques (UCSC, Ensembl)
  - → Banque spécialisée pour un génome spécifique, banques de séquences immunologiques, de voies métaboliques, de cartes génétiques, de motifs protéiques, d'expression de gènes, de structures, . . .

Avantage : facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée, . . .
Inconvénient : ne cible pas toujours exactement ce que l'on veut, toutes les banques possibles n'existent pas

- La revue Nucleic Acids Research consacre un numéro spécial "database issue" chaque année
  (http ://www.oxfordjournals.org/nar/database/c/).

Les banques nucléiques

- Les trois principales banques :
  → EMBL (Europe, 82)
  → GenBank (É-U, 82)
  → DDBJ (Japon, 86)

- Échange quotidien des données entre ces banques depuis 88

- Répartition de la collecte des données : chaque banque collecte les données de son continent

- Mise à jour : nouvelles version disponibles plsr fois par an (date et num de version), mise à disposition des "Updates"

- Format de stockage similaire : les fichiers (*flatfiles*) représentent l'unité d'information élémentaire

Format des fichiers

- Chaque entrée (séquence + informations) est stockée dans un fichier (*flatfile*)

- Ces fichiers se composent de trois parties :
  → Entête (*header*) : description générale de l'entrée
  → Les caractéristiques (*features*) : objets biologiques présents sur la séquence
  → La séquence elle-même

- Les formats de DDBJ et de GenBank sont très similaires

- Chaque ligne commence par un mot clé
  → Deux lettres pour EMBL
  → Maximum 12 lettres pour GenBank et DDBJ

- Fin d'une entrée par //

Exemple d'entrée EMBL


```
ID   U49845; SV 1; linear; genomic DNA; STD; FUN; 5028 BP.
XX
AC   U49845;
XX
DT   07-MAY-1996 (Rel. 47, Created)
DT   17-APR-2005 (Rel. 83, Last updated, Version 4)
XX
DE   Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and
DE   Rev7p (REV7) genes, complete cds.
XX
KW   .
XX
OS   Saccharomyces cerevisiae (baker's yeast)
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
XX
RN   [1]
RP   1-5028
RX   PUBMED; 7871890.
RA   Torpey L.E., Gibbs P.E., Nelson J., Lawrence C.W.;
RT   "Cloning and sequence of REV7, a gene whose function is required for DNA
RT   damage-induced mutagenesis in Saccharomyces cerevisiae";
RL   Yeast 10(11):1503-1509(1994).
XX
RN   [2]
RP   1-5028
RX   PUBMED; 8846915.
RA   Roemer T., Madden K., Chang J., Snyder M.;
RT   "Selection of axial growth sites in yeast requires Axl2p, a novel plasma
RT   membrane glycoprotein";
RL   Genes Dev. 10(7):777-793(1996).
XX
RN   [3]
RP   1-5028
RA   Roemer T.;
RT   ;
RL   Submitted (22-FEB-1996) to the EMBL/GenBank/DDBJ databases.
RL   Terry Roemer, Biology, Yale University, New Haven, CT, USA
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..5028
FT                   /organism="Saccharomyces cerevisiae"
FT                   /chromosome="IX"
FT                   /map="9"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:4932"
```

Exemple d'entrée GenBank/DDBJ


```
LOCUS       SCU49845                5028 bp    DNA     linear   PLN 21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
   PUBMED   7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
   PUBMED   8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
```

Les *flatfiles* en détails (1/8)

1. Les différentes informations de l'entête (*header*) :

- Première ligne : Locus/ID

→ Embl

```
ID       U49845 ;       SV 1 ; linear ; genomic DNA ; STD ; FUN ; 5028 BP.
```

→ DDBJ/GenBank

```
LOCUS       SCU49845          5028 bp    DNA     linear   PLN 21-JUN-1999
```

- Le champ date (chez EMBL uniquement)

```
DT   07-MAY-1996 (Rel. 47, Created)
DT   17-APR-2005 (Rel. 83, Last updated, Version 4)
```

- Lignes de définition : synthèse du contenu biologique

```
DE   Saccharomyces cerevisiae TCP1-beta gene, partial cds ; and Axl2p (AXL2)
and
DE   Rev7p (REV7) genes, complete cds.
```

```
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds ; and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
```

Utilisée dans les 3 principales banques de séquences nucléiques

Sequence division	**Database**
Organismal	
BCT Bacterial	DDBJ, GenBank
PRO Prokaryotic	EMBL
FUN Fungal	EMBL
HUM Human	DDBJ, EMBL
PRI Primate	DDBJ, EMBL, GenBank
ROD Rodent	DDBJ, EMBL, GenBank
MAM Other mammalian	DDBJ, EMBL, GenBank
VRT Other vertebrate	DDBJ, EMBL, GenBank
INV Invertebrate	DDBJ, EMBL, GenBank
PLN Plant	DDBJ, EMBL, GenBank
ORG Organelle	EMBL
VRL Viral	DDBJ, EMBL, GenBank
PHG Phage	DDBJ, EMBL, GenBank
RNA Structural RNA	DDBJ, EMBL, GenBank
SYN Synthetic and chimeric	DDBJ, EMBL, GenBank
UNA Unannotated	DDBJ, EMBL, GenBank
Functional	
EST Expressed sequence tag	DDBJ, EMBL, GenBank
STS Sequence tagged site	DDBJ, EMBL, GenBank
GSS Genome survey	DDBJ, EMBL, GenBank
HTG High-throughput genomic	DDBJ, EMBL, GenBank
PAT Patent	DDBJ, EMBL, GenBank
CON Virtual contigs of segmented sequences	DDBJ, EMBL, GenBank

Les *flatfiles* en détails (2/8)

- Le numéro d'accession : un id unique (commun aux 3 banques)

```
AC       U49845 ;
```

```
ACCESSION       U49845
```

- La version (équivalent à SV dans la 1re ligne d'EMBL)

```
VERSION         U49845.1    GI :1293613
```

- Lignes avec des mots-clés (KEYWORDS ou KW)

- Lignes de taxonomie

```
OS      Saccharomyces cerevisiae (baker's yeast)
OC      Eukaryota ; Fungi ; Dikarya ; Ascomycota ; Saccharomycotina ;
Saccharomycetes ;
OC      Saccharomycetales ; Saccharomycetaceae ; Saccharomyces.
```

```
SOURCE          Saccharomyces cerevisiae (baker's yeast)
  ORGANISM      Saccharomyces cerevisiae
                Eukaryota ; Fungi ; Dikarya ; Ascomycota ; Saccharomycotina ;
                Saccharomycetes ; Saccharomycetales ; Saccharomycetaceae ;
                Saccharomyces.
```

- Les références : publication ou origine de la soumission

```
RN    [1]
RP    1-5028
RX    PUBMED ; 7871890.
RA    Torpey L.E., Gibbs P.E., Nelson J., Lawrence C.W. ;
RT    "Cloning and sequence of REV7, a gene whose function is required for DNA
RT    damage-induced mutagenesis in Saccharomyces cerevisiae" ;
RL    Yeast 10(11) :1503-1509(1994).
XX
RN    [2]
RP    1-5028
RX    PUBMED ; 8846915.
RA    Roemer T., Madden K., Chang J., Snyder M. ;
RT    "Selection of axial growth sites in yeast requires Axl2p, a novel plasma
RT    membrane glycoprotein" ;
RL    Genes Dev. 10(7) :777-793(1996).
XX
RN    [3]
RP    1-5028
RA    Roemer T. ;
RT    ;
RL    Submitted (22-FEB-1996) to the EMBL/GenBank/DDBJ databases.
RL    Terry Roemer, Biology, Yale University, New Haven, CT, USA
```

```
REFERENCE   1 (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
   PUBMED   7871890
REFERENCE   2 (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
   PUBMED   8846915
REFERENCE   3 (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
```

2. Les caractéristiques (*features*)

- Features table : annotation des séquences, gestion du transfert d'information, localisation des éléments biologiques
- Mise à disposition d'un vocabulaire étendu controlé (Gene Ontology) pour décrire les caractéristiques biologiques des séquences (annotations)
- Annotation = information additionnelle venant enrichir un document d'intérêt

  → Annotation structurale : caractériser les gènes au travers, notamment, de leur structure intron-exon, signaux de régultaion, sites d'épissage, . . .

  → Annotation fonctionnelle : anticiper sur les motifs fonctionnels qui seront retrouvés au niveau protéiques, caractérisation des fonctions biologiques des produits d'expression des gènes

Les *flatfiles* en détails (6/8)


```
FT    source    1..5028
FT              /organism="Saccharomyces cerevisiae"
FT              /chromosome="IX"
FT              /map="9"
FT              /mol_type="genomic DNA"
FT              /db_xref="taxon :4932"
FT    CDS       <1..206
FT              /codon_start=3
FT              /product="TCP1-beta"
FT              /db_xref="GOA :P39076"
FT              /db_xref="UniProtKB/Swiss-Prot :P39076"
FT              /protein_id="AAA98665.1"
FT              /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEAA
FT              EVLLRVDNIIRARPRTANRQHM"
[...]
FT    CDS       complement(3300..4037)
FT              /codon_start=1
FT              /gene="REV7"
FT              /product="Rev7p"
FT              /db_xref="GOA :P38927"
FT              /db_xref="InterPro :IPR003511"
FT              /db_xref="SGD :S000001401"
FT              /db_xref="UniProtKB/Swiss-Prot :P38927"
FT              /protein_id="AAA98667.1"
FT              /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQF
FT              VPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVDKD
FT              DQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNRRVD
FT              SLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEKLISG
FT              DDKILNGVYSQYEEGESIFGSLF"
```

Les *flatfiles* en détails (7/8)


```
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon :4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI :1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     [...]
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI :1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
```

3. La séquence de nucléotides

→ EMBL

```
SQ   Sequence 5028 BP; 1510 A; 1074 C; 835 G; 1609 T; 0 other;
     gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg        60
     ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct       120
     ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa       180
     gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg       240
     ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa       300
     agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa       360
     attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat       420
     aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga       480
     gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc       540
```

→ GenBank

```
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
```

À vous de jouer

- Répondez aux questions suivantes en vous référant aux 2 fiches distribuées

→ Une des fiches vient d'EMBL, l'autre de GenBank, identifiez-les

→ Trouvez le nom des organismes d'où proviennent les séquences

→ Que sont ces séquences ?

→ À quelle division d'organisme appartiennent-elles ?

→ Quelle séquence a été entrée le + récemment ? Ont-elles été modifiées ?

→ Qui a soumis ces séquences ?

→ De quel type de molécule s'agit-il ?

→ Peut-on localiser ces séquences sur un chromosome ?

Origine des informations sur les protéines

- Origine des séquences
  - → Traduction automatique de séquences d'ADN (majoritairement)
  - → Séquençage de protéines (rare car long et coûteux)
  - → Protéines dont la structure 3D est connue

- Origine des annotations
  - → Spectrométrie de masse : régulation, rythme et localisation de l'expression des protéines ; mais aussi identification et modification post-transcriptionnelle
  - → Études d'interactions : comment les protéines s'assemblent entre elles ou avec d'autres molécules pour former des complexes moléculaires
  - → Cristallographie et résonance magnétique nucléaire : pour déterminer la forme 3D finale de la protéine

- Les données stockées : séquences + annotations (protéines entières ou fragment de protéines)
  - → Banques généralistes : protéines de toutes les espèces
  - → Banques spécialisées : familles de protéines particulières, groupes de protéines ou d'un organisme particulier

- GenPept : entrées = traductions des séquences de DDBJ/EMBL/GenBank (champ CDS) ; les annotations sont les mêmes ; la banque n'est pas vérifiée $\sim$ *archive basique*

- RefSeq : projet NCBI ; but : fournir une vue d'ensemble, intégrée et non-redondante de séquences d'ADN, d'ARN et de protéines ; lien explicite entre les séq. nucléiques et protéiques ; numéro d'accession particulier format $2+6$, ex : NT_123456

- UniProt : assemblage de SWISSPROT, TrEMBL et PIR

- Objectif : proposer les séquences les plus représentatives et décrites par les experts

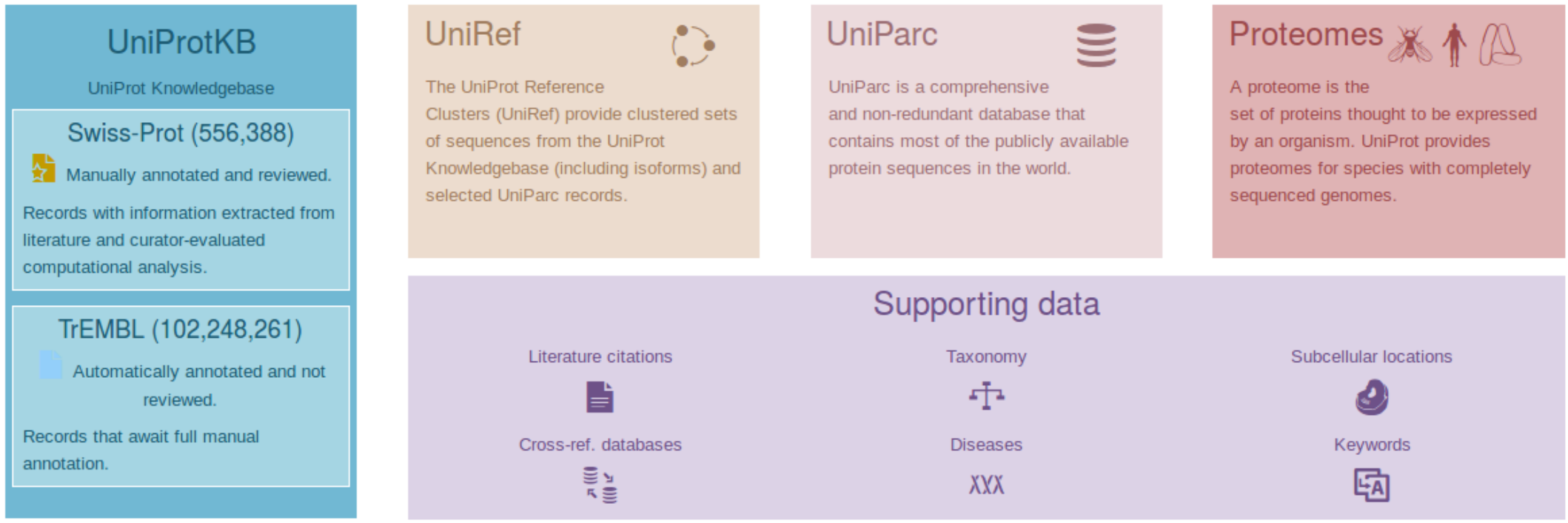
- Avantages
  - → Entrées non redondante
  - → Liens explicites entre les séquences nucléiques et protéiques
  - → Mise à jour régulière par le personnel du NCBI (statut indiqué)
  - → Validation des données et consistance des formats
  - → Synthèse des informations issues de plusieurs entrées nucl. ou prot.

- Différents niveaux de correction des données (champ COMMENT)
  - → Reviewed : revu par un membre du NCBI (ajout d'informations)
  - → Validated : une 1ère révision a été effectuée par un membre du NCBI (annotation en cours)
  - → Provisional : entrée non lue par un annotateur (contient surement un vrai transcrit ou une vrai protéine)
  - → Predicted : transcrit ou protéine issu d'une prédiction (prog info)

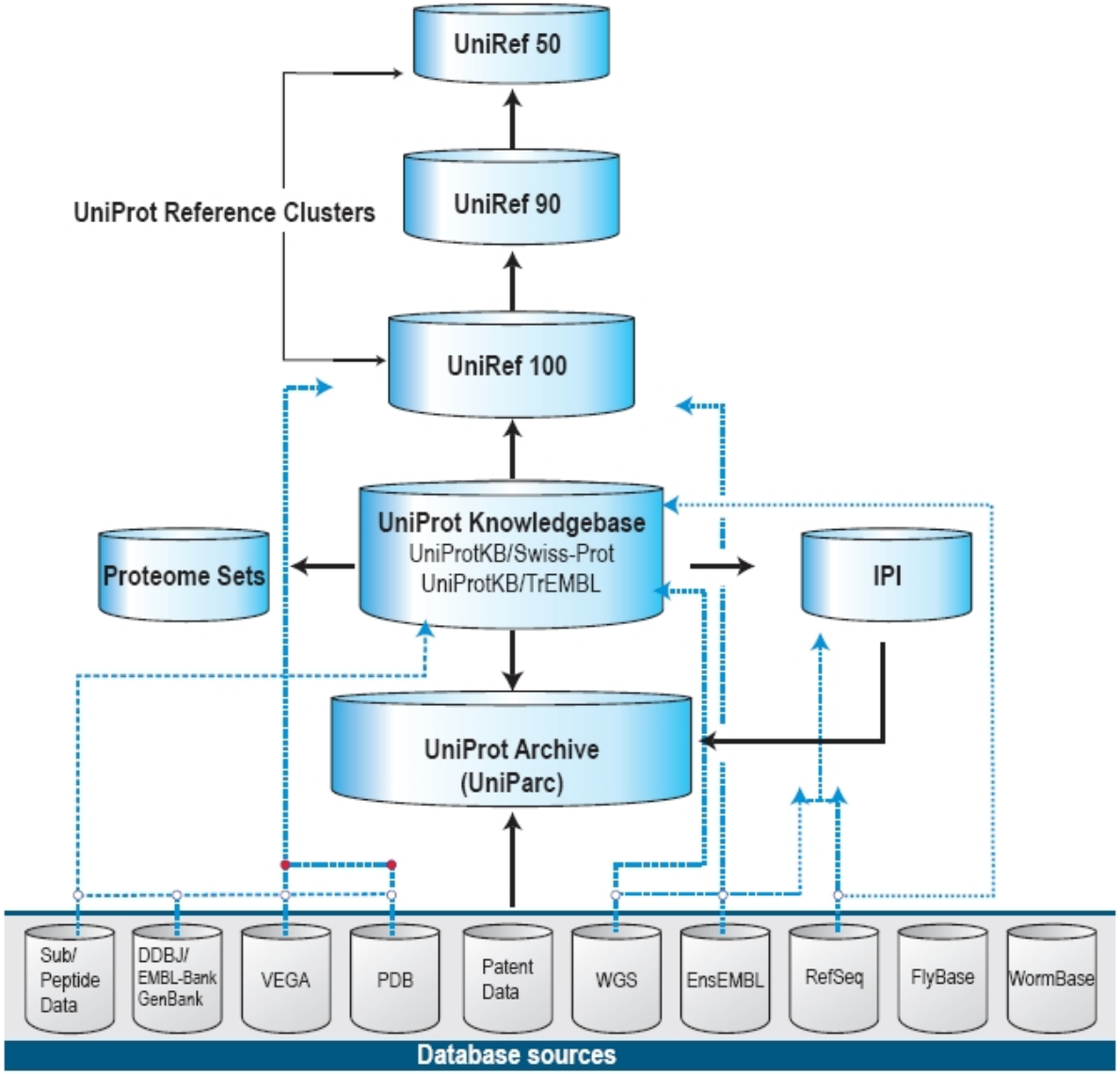
- Un numéro d'accession au format spécifique : un préfixe à 2 lettres + un muméro à 6 chiffres (convention de nommage)
- Liste des préfixes des numéros d'accession de RefSeq

→ NC_ : Complete genomic molecules
→ NG_ : Incomplete genomic region
→ NM_ : mRNA
→ NR_ : ncRNA
→ NP_ : Protein
→ XM_ : predicted mRNA model
→ XR_ : predicted ncRNA model
→ XP_ : predicted Protein model (eukaryotic sequences)
→ WP_ : predicted Protein model (prokaryotic sequences)

"The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information"



UNIPROT : les différentes banques



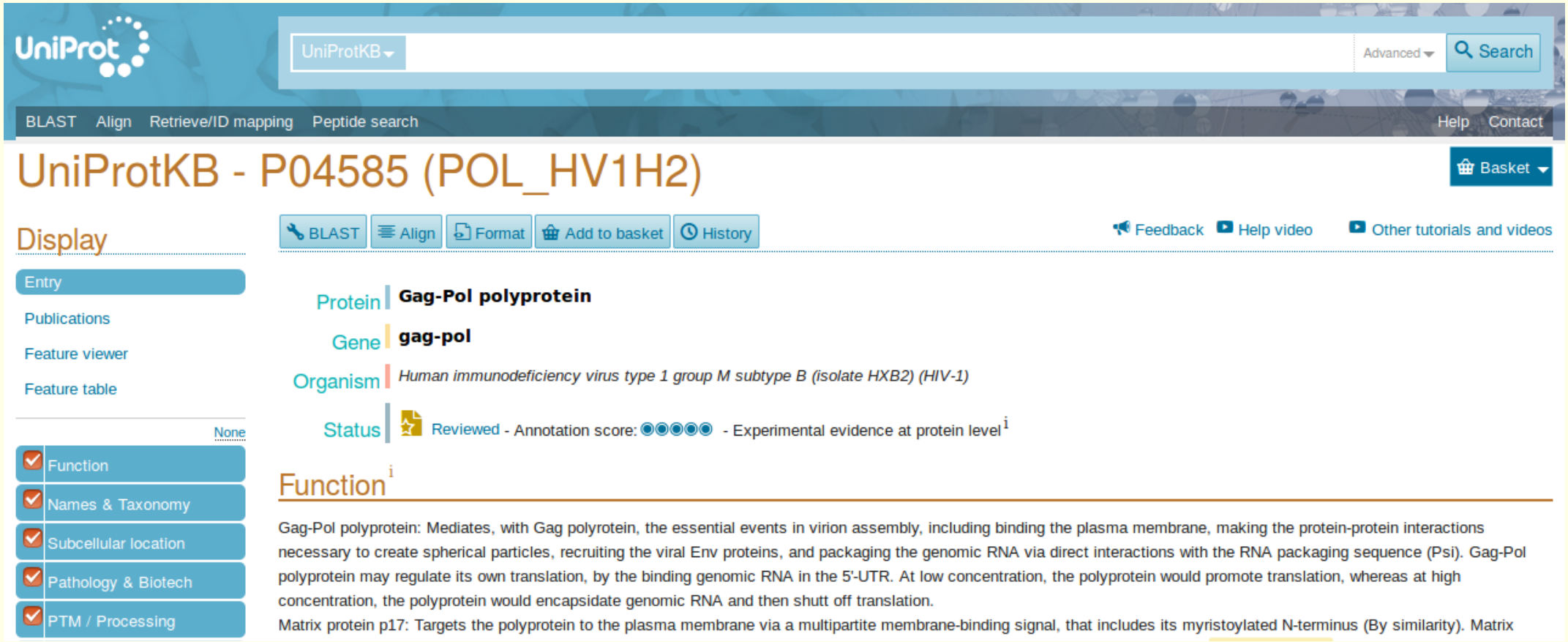
Sources and flow of data for UniProt's component databases

UniProtKB, ses deux banques

- SWISSPROT
  - → Données corrigées et validées par des experts
  - → Haut niveau d'annotation
  - → Redondance minimale
  - → Nombreux liens vers d'autres banques ($\sim$ 60)

- TrEMBL
  - → Entrées supplémentaires à SWISSPROT (pas encore annotées)
  - → Traduction automatique des CDS d'EMBL et soumissions spontanées
  - → Annotation automatique des protéines

Exemple d'entrée UniProtKB



Indication de la fiabilité des annotations - "Annotation score"
(SWISSPROT / TrEMBL)

- Localisation génomique :

  1. "Cartographie" : carte physique, carte cytogénétique, liaison génétique (*genetic linkage*), . . .

  2. "Identification" : trouver où sont les objets d'intérêt biologiques comme les gènes, les variations génétiques ou les locus de prédisposition à des maladies $\Rightarrow$ association d'une signature moléculaire à un résultat biologique

- Ex : localiser des nouveaux gènes ou affiner des régions d'intérêt

- Genome Database (GDB), eGenome, LBD2000, GeneCards et GeneLoc, GeneLynx, EuGenes, AceView, . . .

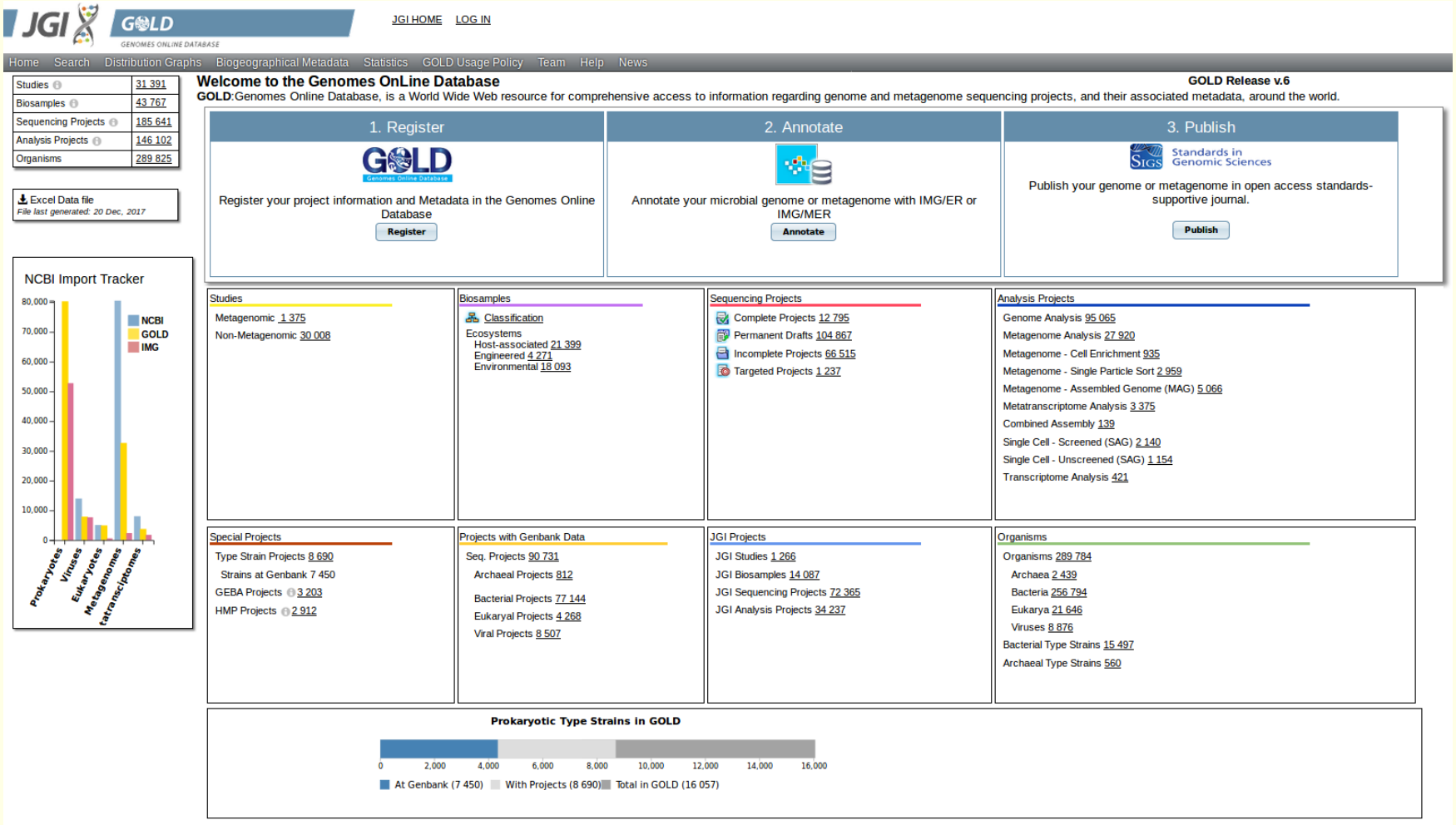
- Cartes comparatives entre plusieurs génomes (Mouse Genome Informatics Database (MGI), . . .)

- 1996 1re séquence complète d'un génome eucaroyte *Saccharomyces cerevisiae*, chromosomes entre 270 et 1500 Kb ; limite d'une entrée dans GenBank à cette époque : 350 Kb
  $\Rightarrow$ Mise en place d'une section spécifique pour les génomes, 1re vue graphique des séquences génomiques

- 2001 première ébauche du génome humain ; chromosomes entre 46 et 246 Mb

- "Navigateur" de génomes : NCBI Map Viewer, UCSC Genome Browser, Ensembl (EBI et Sanger Institute)

- Attention au version d'assemblage qui peuvent changer sans avertissement



GOLD : Genome Online Database (http://gold.jgi.doe.gov/)

Informations concernant les projets de séquençage de génomes et métagénomiques (et les métadonnées associées)

- Regroupement des protéines ayant des fonctions identiques ou proches

- Construites souvent par comparaison de toutes les protéines entre elles puis constitution de groupes

- Nombreuses banques :
  → HomoloGene, COG (NCBI)
  → KEGG SSDB (Sequence Similarity DataBase)
  → Clustr (EBI)
  → HOGENOM, OrthoMaM (LBBE, ISEM)

- Autres banques basées sur la structure 3D des protéines
  → SSF, SCOP, CATH, ...

- Une famille de protéines peut-être caractérisée par un motif ou un domaine protéique

  → Séquence plus ou moins conservée importante pour la fonction des protéines de la famille

  → Déterminée à partir d'un alignement multiple

  → Plusieurs représentations possibles : consensus, expression régulière, alignement, matrices, HMM, ...

- Nombreuses banques

  → Prosite, PFAM, BLOCKS, Prodom, CDD, ...

Accessibilité des ressources

- Le réseau internet : interface web, API, programme client, . . .

- Des centres de ressources : NCBI (http://www.ncbi.nlm.nih.gov/), EBI (http://www.ebi.ac.uk/), . . .

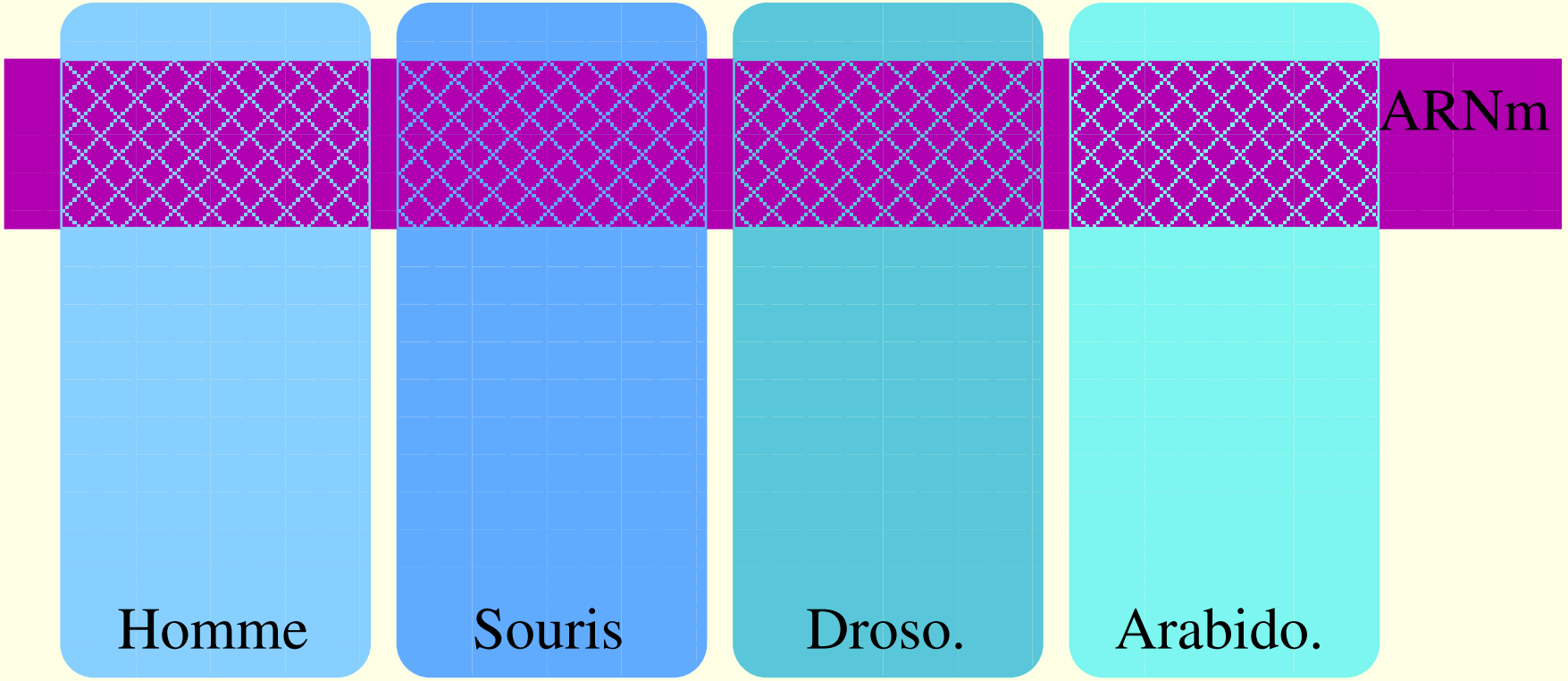
- Des catalogues d'outils : EMBOSS (http://emboss.sourceforge.net), Institut Pasteur (http://galaxy.pasteur.fr/), . . .

- Des systèmes d'interrogation
  - → GQuery (anciennement Entrez) : développé et utilisé par le NCBI
  - → EB-Eye ou EBI Search : développé et utilisé par l'EBI, utilsé par d'autres projets comme EnsemblGenome, Emboss, Interpro,...
  - → ACNUC développé et utilisé par le PRABI-Doua : un des 1er système d'interrogation pour les données biologiques
  - → Systèmes spécifique comme UniProt search engine ($\sim$ EB-Eye)
  - → Base de données relationnelles / NoSQL
  - → . . .

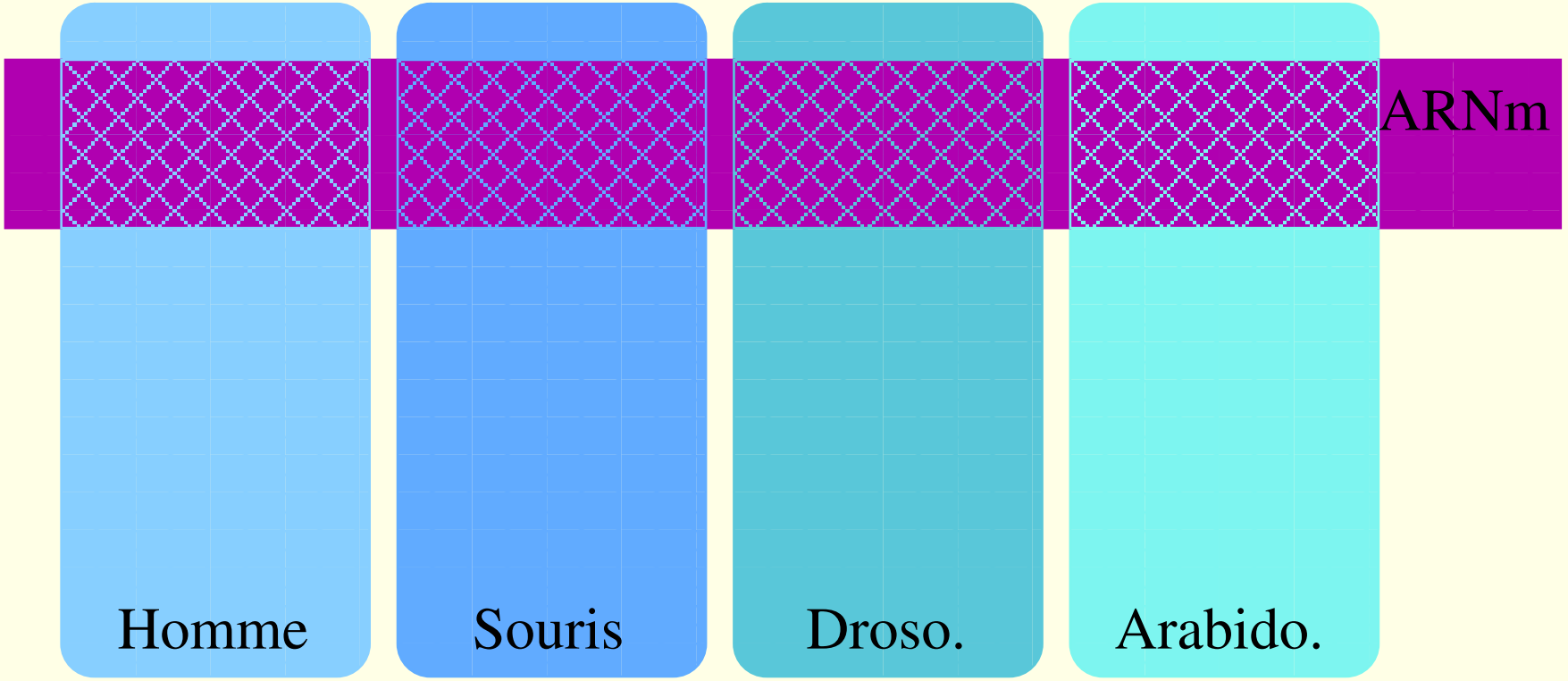
Interrogation d'une banque

- But
  - → Obtenir des information nouvelles et pertinentes
  - → Aide à la mise au point d'expérience
  - → Validation des résultats d'une expérience

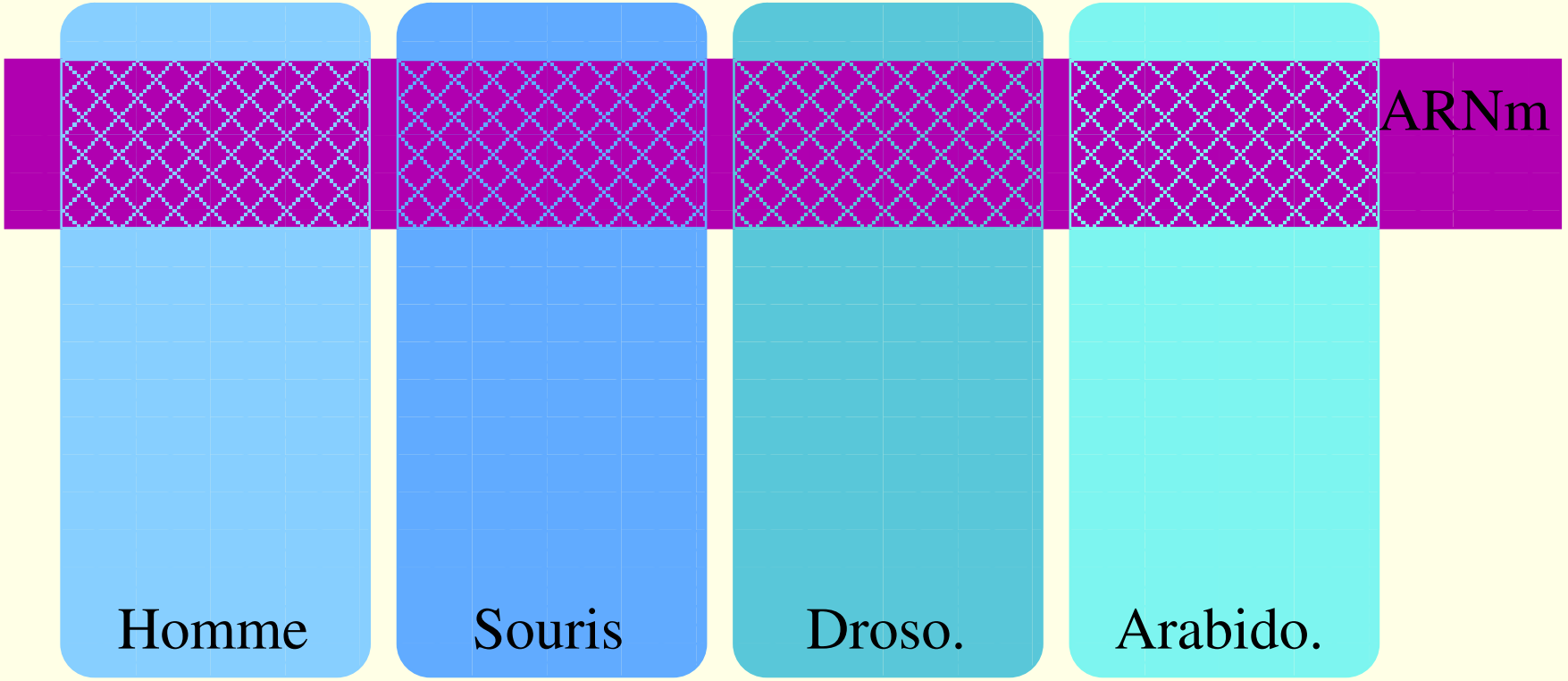
- Contraintes pour un système d'interrogation
  - → Obtention de données pertinentes (pas trop de résultats mais tous ceux relatifs à notre problématique)
  - → Simplicité d'utilisation (syntaxe d'interrogation intuitive)
  - → Réponse rapide
  - → Possibilité d'analyse des résultats (couplage à des outils)



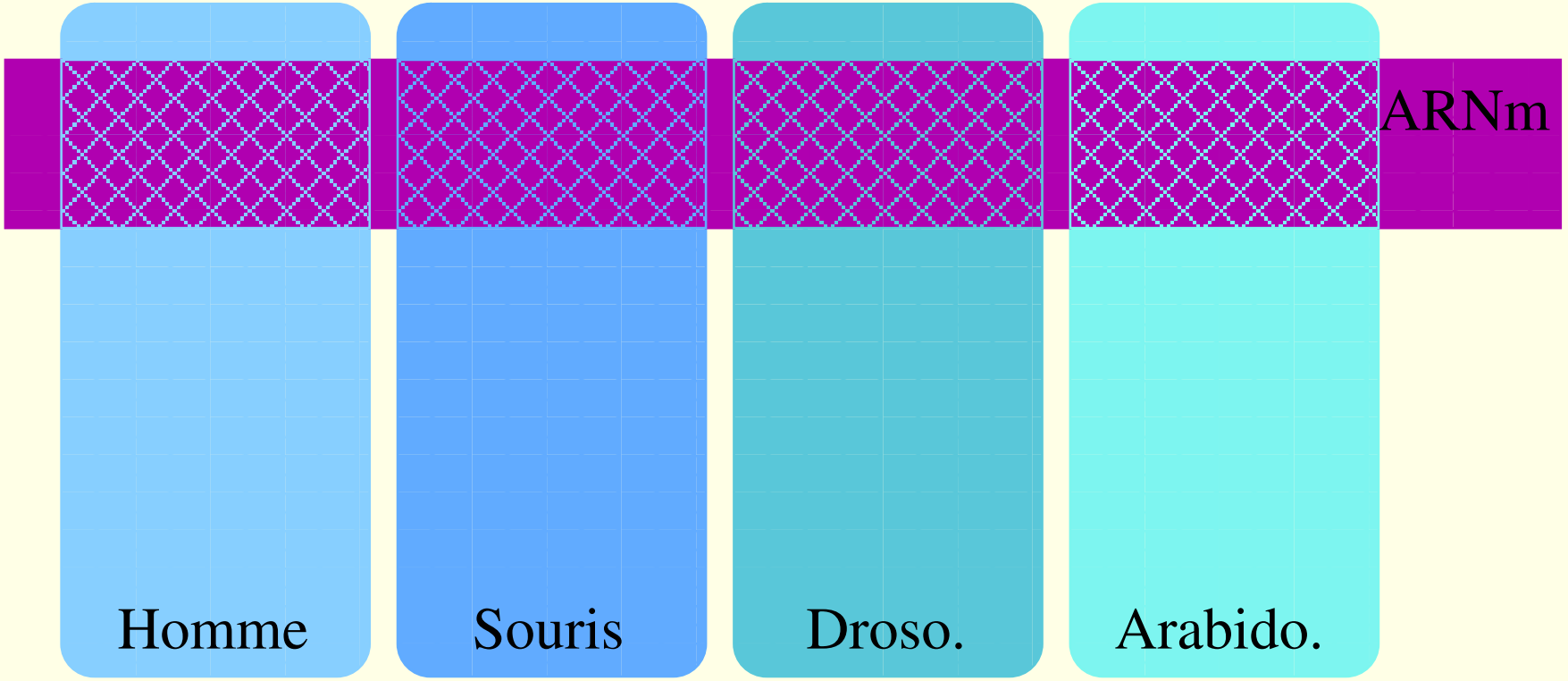
- Toutes séquences contenant 'Homme' :
- Séquences contenant 'Homme' ou 'Souris' :
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



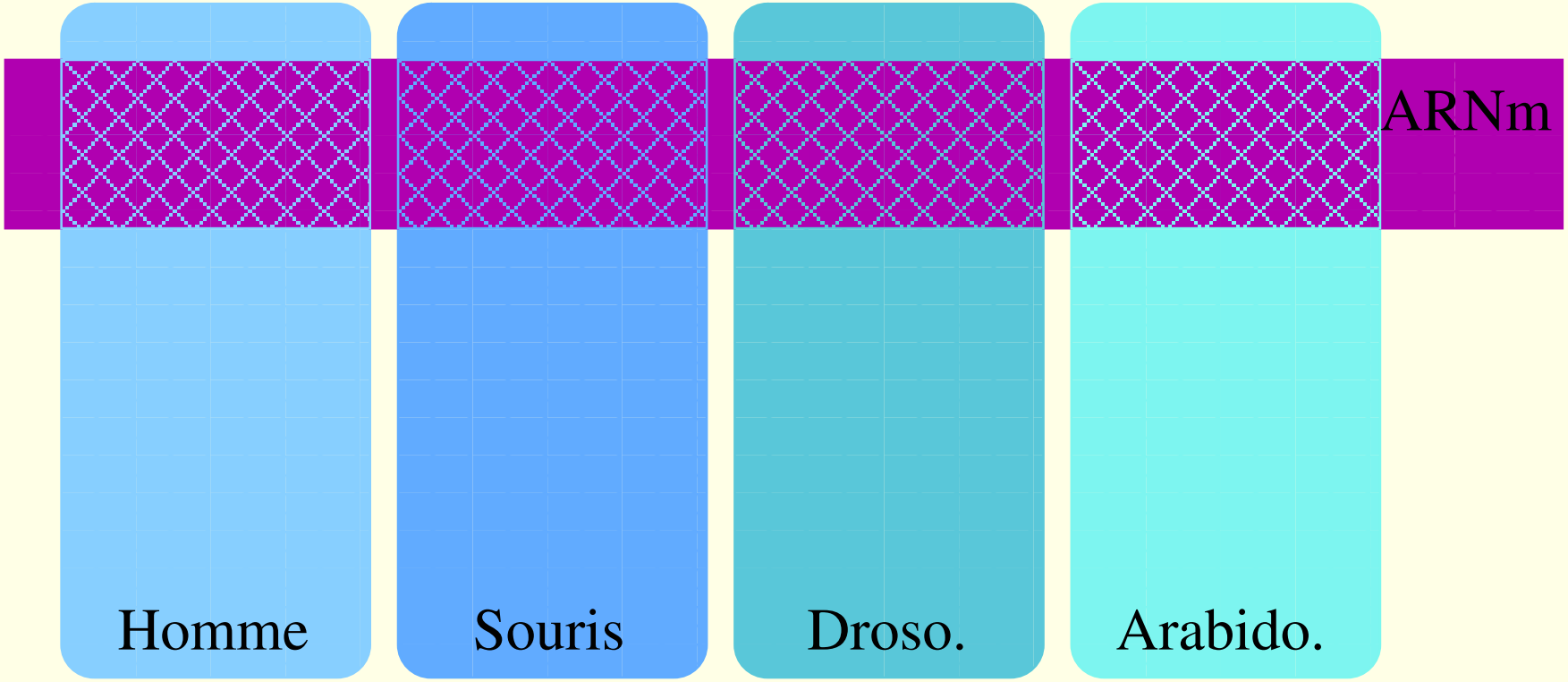
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' :
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



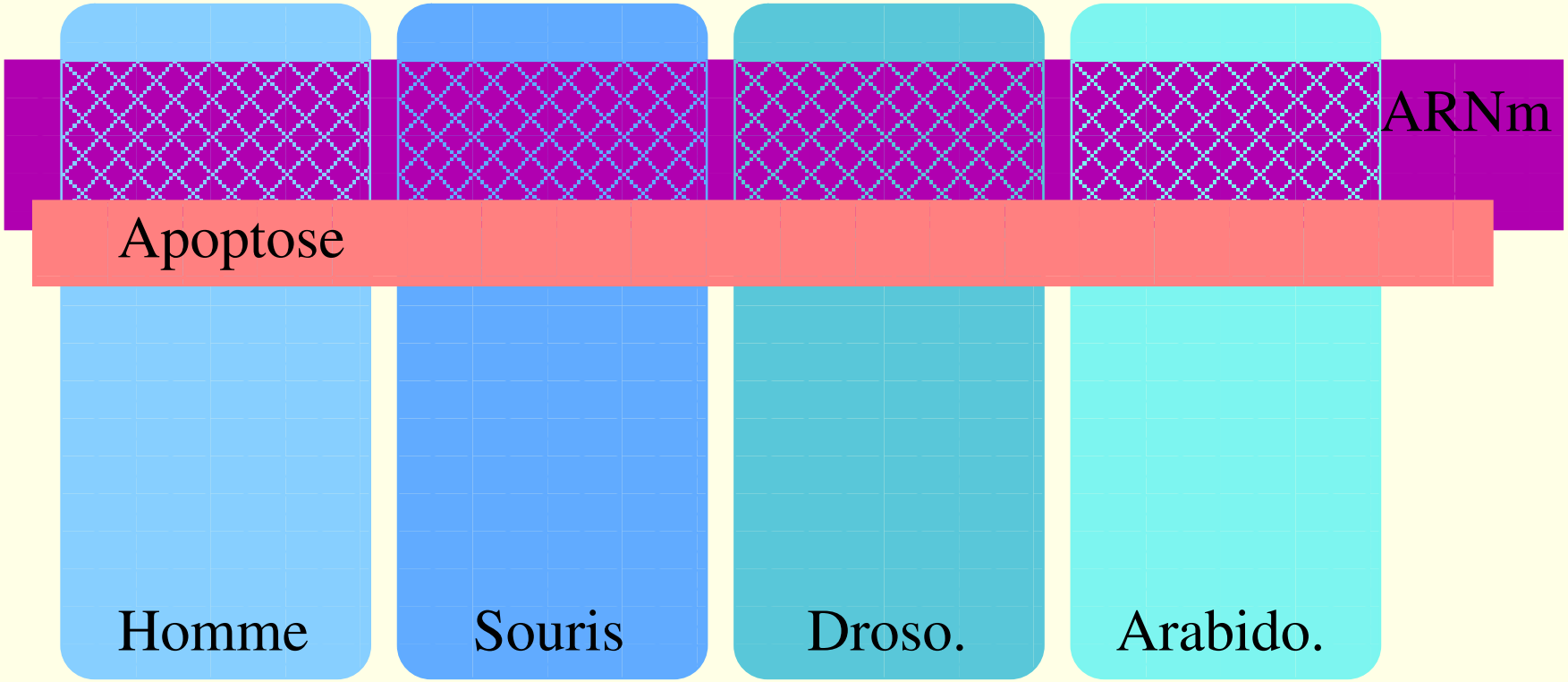
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



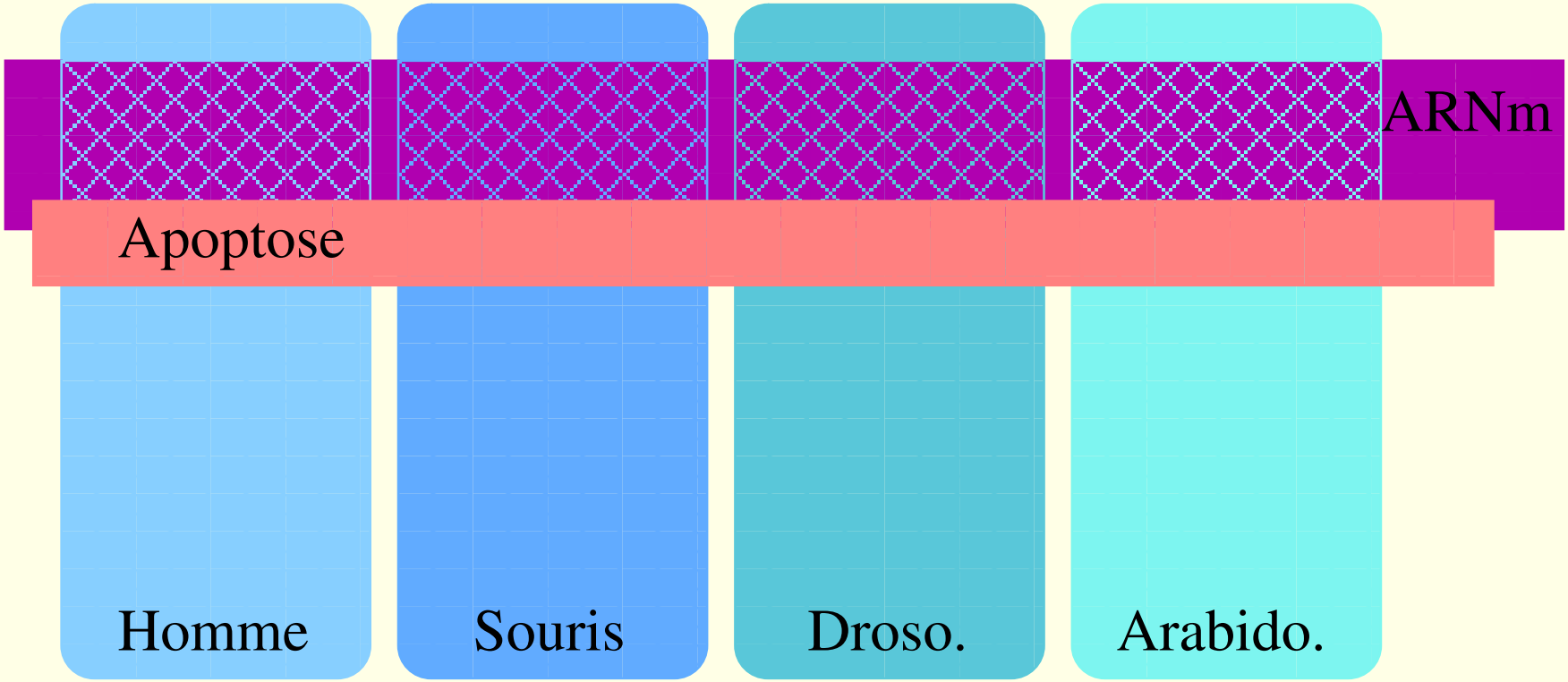
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile : Droso. ET ARNm
- Tous les ARNm sauf ceux d'arabidopsis :



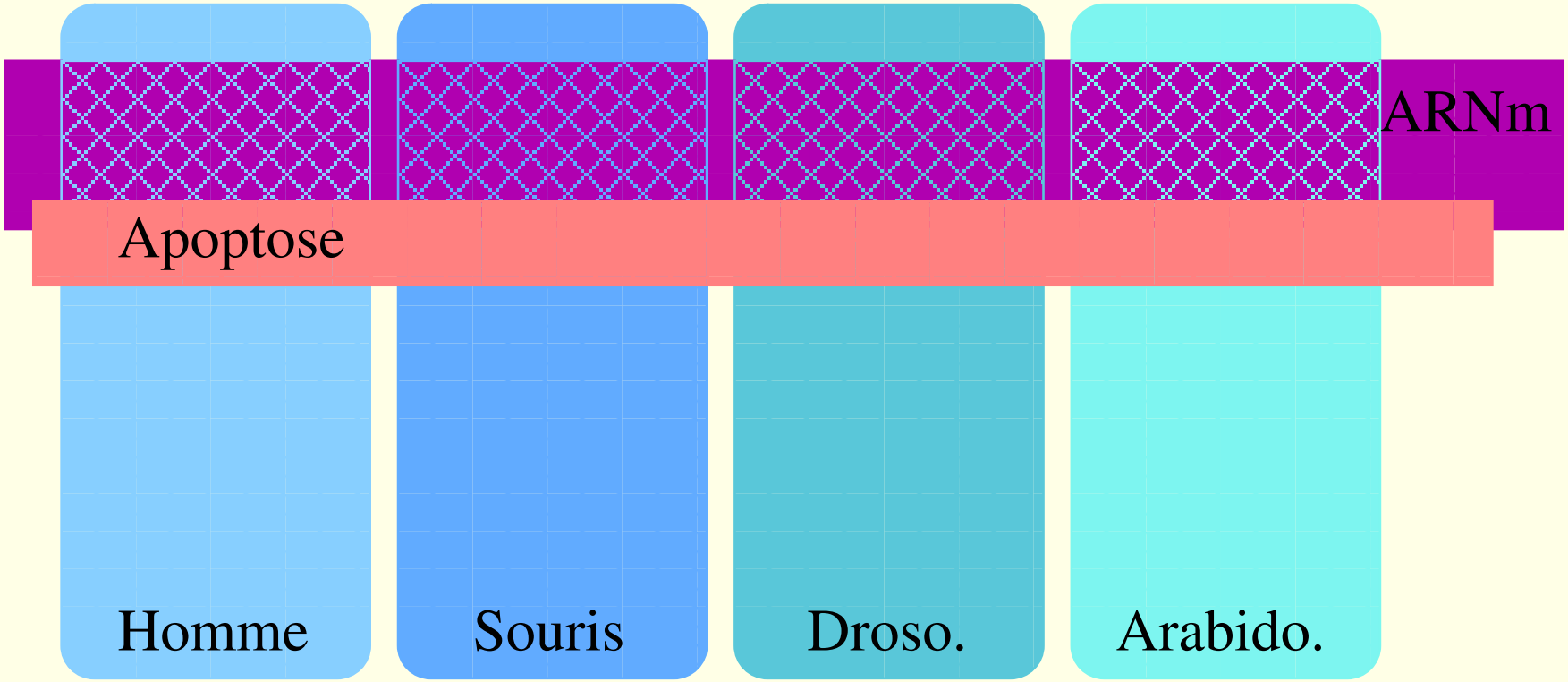
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile : Droso. ET ARNm
- Tous les ARNm sauf ceux d'arabidopsis : ARNm NON Arabido.



- ARNm impliqués dans l'apoptose chez l'homme :
- Séquences de la souris ou de la droso. impliquées dans l'apoptose mais qui ne sont pas des ARNm :

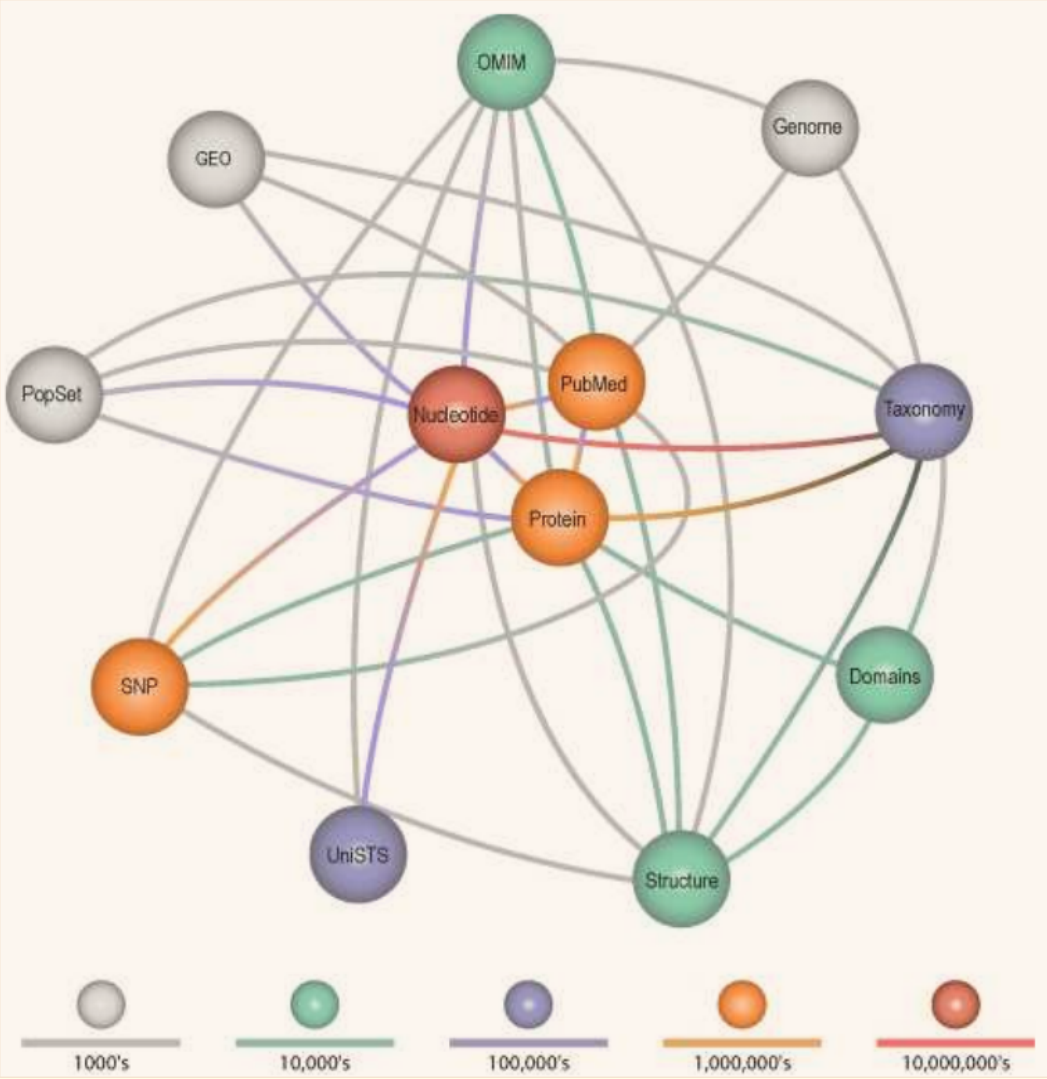


- ARNm impliqués dans l'apoptose chez l'homme :
  ARNm ET Apoptose ET Homme
- Séquences de la souris ou de la droso. impliquées dans l'apoptose mais qui ne sont pas des ARNm :



- ARNm impliqués dans l'apoptose chez l'homme :
  ARNm ET Apoptose ET Homme
- Séquences de la souris ou de la droso. impliquées dans l'apoptose mais qui ne sont pas des ARNm :
  ((Souris OU Droso.) ET Apoptose) NON ARNm

- Intégration "légère" de diverses banques de données au travers de la notion de référence croisée
  $\rightarrow$ choix de rubriques



- Interface propriétaire (ne peut être installée par autrui)

- Requête simple/manuelle (possibilité de mettre le nom du champs entre crochets)
  Exemple : `homo sapiens [organism]`

- Requête avancée (choix des champs dans une liste déroulante)

- Différentes options
  → Aide dans "Preview/Index"
  → Historique (lien "History")
  → Ajout de limites (lien "Limits")
  → Sauvegarde, format : "Display", "send to", menus déroulants associés

- Terme
- Locution (groupe de terme) : "mot1 mot2 . . ."
- Connecteurs logiques : ET (AND), OU (OR), NON (NOT)
- Caractère joker : *
- Expression complexe : parenthèse () (ET prioritaire)
- Couple descripteur (champ) / valeur en fonction des rubriques (ex : organism, gene, slen, author, title, . . .) : valeur [descripteur]
- Exemples de requêtes
  - → Human hemoglobin
  - → Human [organism] hemoglobin* [protein name]
  - → Human [organism] OR Mouse [organism]
  - → 1000 :2000 [SLEN] (Human [organism] OR Mouse [organism])
  - → 1000 :2000 [SLEN] NOT Human [organism]

- Syntaxe : valeur [descripteur]

- Descripteur Properties [prop]
  → type de molecule : biomol_genomic [prop], biomol_mrna [prop]
  → rubrique : srcdb_refseq [prop], srcdb_ddbj/embl/genbank [prop], srcdb_genbank [prop]
  → groupe d'organismes : gbdiv_pri [prop], gbdiv_rod [prop], gbdiv_pln [prop]

- Descripteur annotation [fkey] : exon [fkey] cds [fkey] 5'UTR [fkey] mRNA [fkey]

- Autres descripteurs en fonction de la rubrique : [organism], [gene], ...

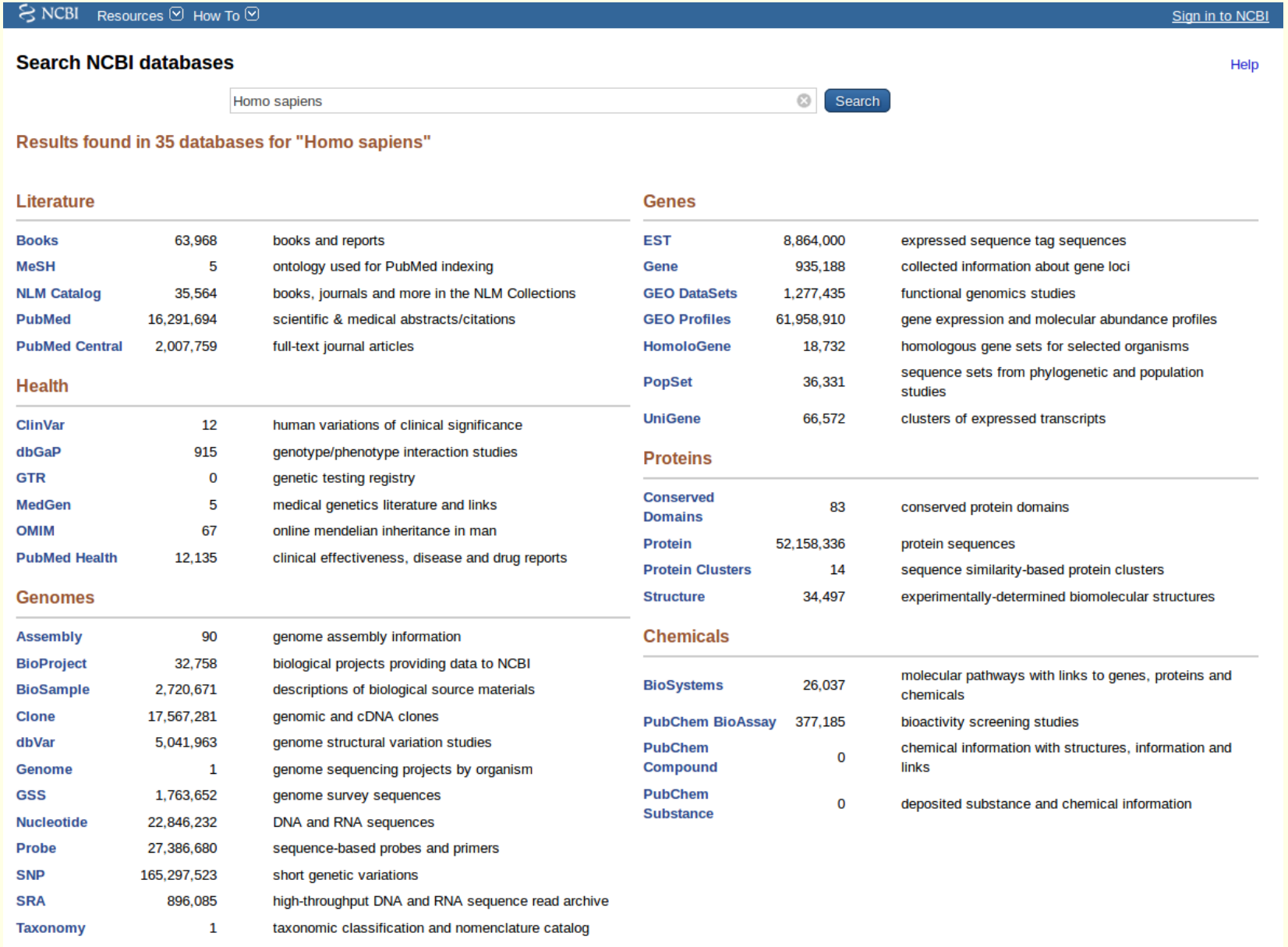
GQuery, les rubriques

- Nucleotide : séquences nucléiques
- Gene : synthèse sur les gènes
- Protein : séquences protéiques (Genpept, Swissprot)
- CDD : domaines protéiques fonctionnels (Conserv Domain protein database)
- PubMed : publications dans les journaux, revues, conférences (Medline)
- Taxonomy : classification du vivant
- OMIM : maladies génétiques humaines
- RefSeq : séquences de la banque RefSeq
- . . .

Exemple de champs de la rubrique Gene

Nom du champ	Abréviation	Définition	Exemples
Symbol	SYM	nom normalisé du gène	SRY [SYM]
Chromosome	CHR	numéro chromosome portant le gène	9 [CHR]
Filter	Filter	lien avec une autre banque	gene_omim [FILTER]
Gene Ontology	GO	descripteur normalisé	cell adhesion [GO]
Title	TI	Terme dans la définition	binding [TI]
Properties	PROP	différentes propriétés	gene_in_mitochondrion[PROP]
Organism	ORGN	noms pour l'ensemble des taxons	human[ORGN]
All Fields	[ALL]	tous les champs	SRY [ALL]

GQuery, formulaire de requête et résultats d'une recherche sur toutes les rubriques



NCBI Resources How To Sign in to NCBI

Search NCBI databases Help

Homo sapiens Search

Results found in 35 databases for "Homo sapiens"

Literature

Books	63,968	books and reports
MeSH	5	ontology used for PubMed indexing
NLM Catalog	35,564	books, journals and more in the NLM Collections
PubMed	16,291,694	scientific & medical abstracts/citations
PubMed Central	2,007,759	full-text journal articles

Health

ClinVar	12	human variations of clinical significance
dbGaP	915	genotype/phenotype interaction studies
GTR	0	genetic testing registry
MedGen	5	medical genetics literature and links
OMIM	67	online mendelian inheritance in man
PubMed Health	12,135	clinical effectiveness, disease and drug reports

Genomes

Assembly	90	genome assembly information
BioProject	32,758	biological projects providing data to NCBI
BioSample	2,720,671	descriptions of biological source materials
Clone	17,567,281	genomic and cDNA clones
dbVar	5,041,963	genome structural variation studies
Genome	1	genome sequencing projects by organism
GSS	1,763,652	genome survey sequences
Nucleotide	22,846,232	DNA and RNA sequences
Probe	27,386,680	sequence-based probes and primers
SNP	165,297,523	short genetic variations
SRA	896,085	high-throughput DNA and RNA sequence read archive
Taxonomy	1	taxonomic classification and nomenclature catalog

Genes

EST	8,864,000	expressed sequence tag sequences
Gene	935,188	collected information about gene loci
GEO DataSets	1,277,435	functional genomics studies
GEO Profiles	61,958,910	gene expression and molecular abundance profiles
HomoloGene	18,732	homologous gene sets for selected organisms
PopSet	36,331	sequence sets from phylogenetic and population studies
UniGene	66,572	clusters of expressed transcripts

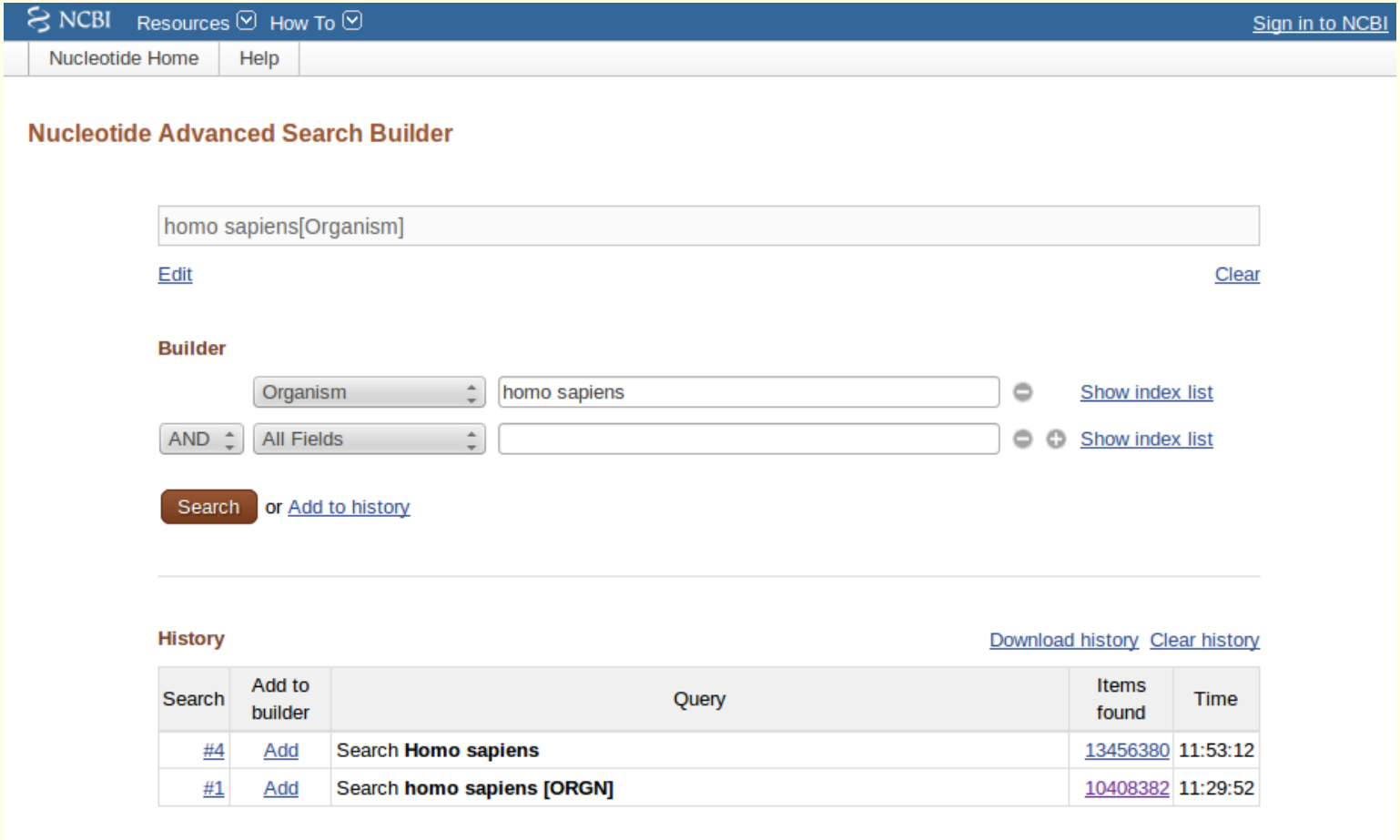
Proteins

Conserved Domains	83	conserved protein domains
Protein	52,158,336	protein sequences
Protein Clusters	14	sequence similarity-based protein clusters
Structure	34,497	experimentally-determined biomolecular structures

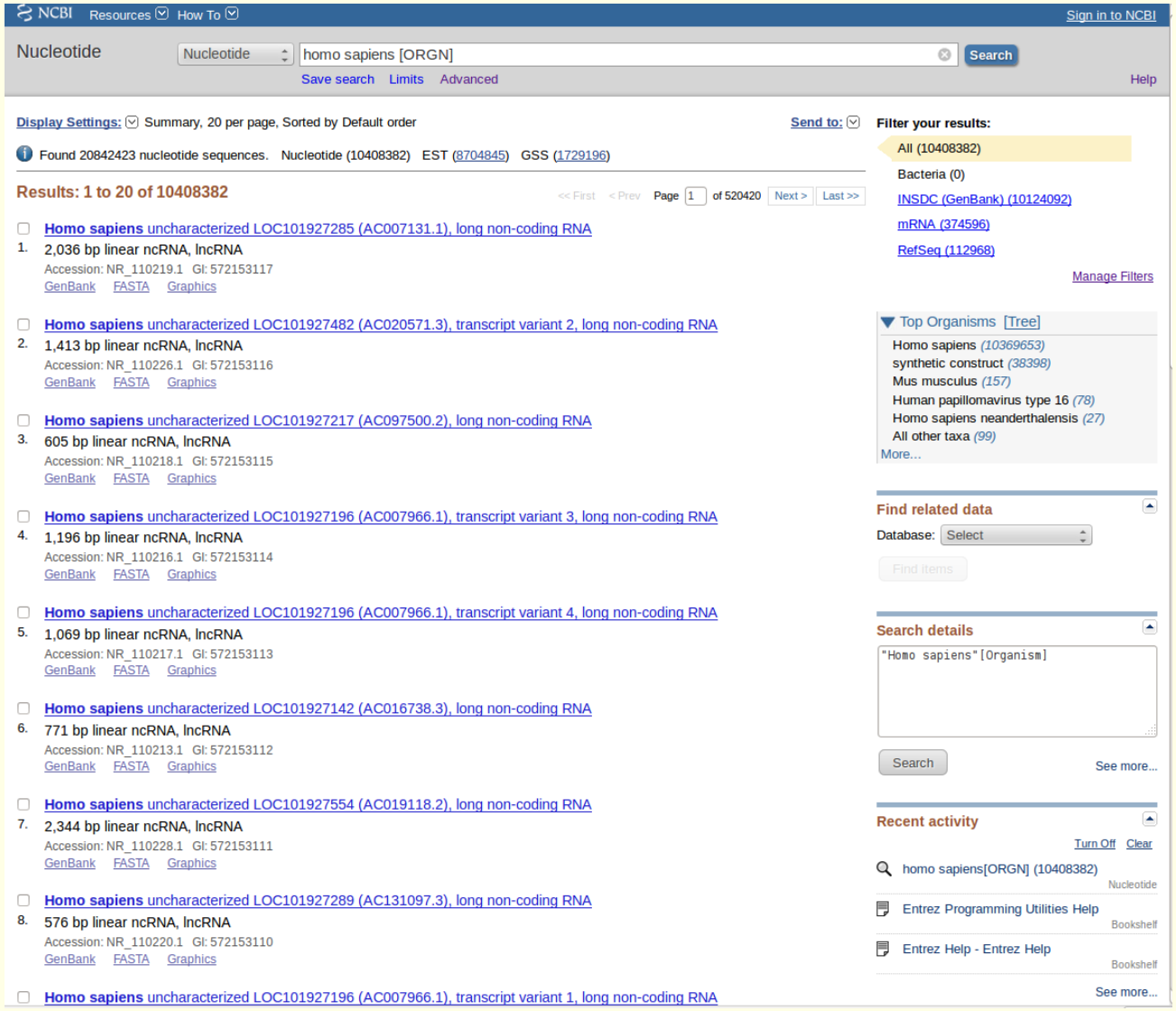
Chemicals

BioSystems	26,037	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	377,185	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	0	deposited substance and chemical information

GQuery, formulaire de requête avancé sur une rubrique (ex : Nucleotide)



GQuery, résultats d'une recherche sur une rubrique (ex : Nucleotide)



Quelques formats classiques en bioinformatique

1. Les fichiers des banques de séquences (*flatfiles*)
   → Ceux qu'on a vu mais aussi ceux d'autres banques

2. Le format FASTA : une 1re ligne de définition introduite par un >, contenant un identifiant sans espace et une description suivie de la séquence elle-même, sur plusieurs lignes (de taille 60 ou 80 classiquement)

```
>embl|U49845|U49845 Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and Rev7p (REV7) genes, complet
gatcctccatatacaacggtatctccacctcaggtttagatctcaacaacggaaccattg
ccgacatgagacagttaggtatcgtcgagagttacaagctaaaacgagcagtagtcagct
ctgcatctgaagccgctgaagttctactaagggtggataacatcatccgtgcaagaccaa
gaaccgccaatagacaacatatgtaacatatttaggatatacctcgaaaataataaaccg
ccacactgtcattattataattagaaacagaacgcaaaaattatccactatataattcaa
agacgcgaaaaaaaaagaacaacgcgtcatagaacttttggcaattcgcgtcacaaataa
attttggcaacttatgtttcctcttcgagcagtactcgagccctgtctcaagaatgtaat
aatacccatcgtaggtatggttaaagatagcatctccacaacctcaaagctccttgccga
```

3. Les formats des outils phylogénétiques

→ Le format NEXUS [Maddison *et al.*, 97] (utilisé par PAUP)

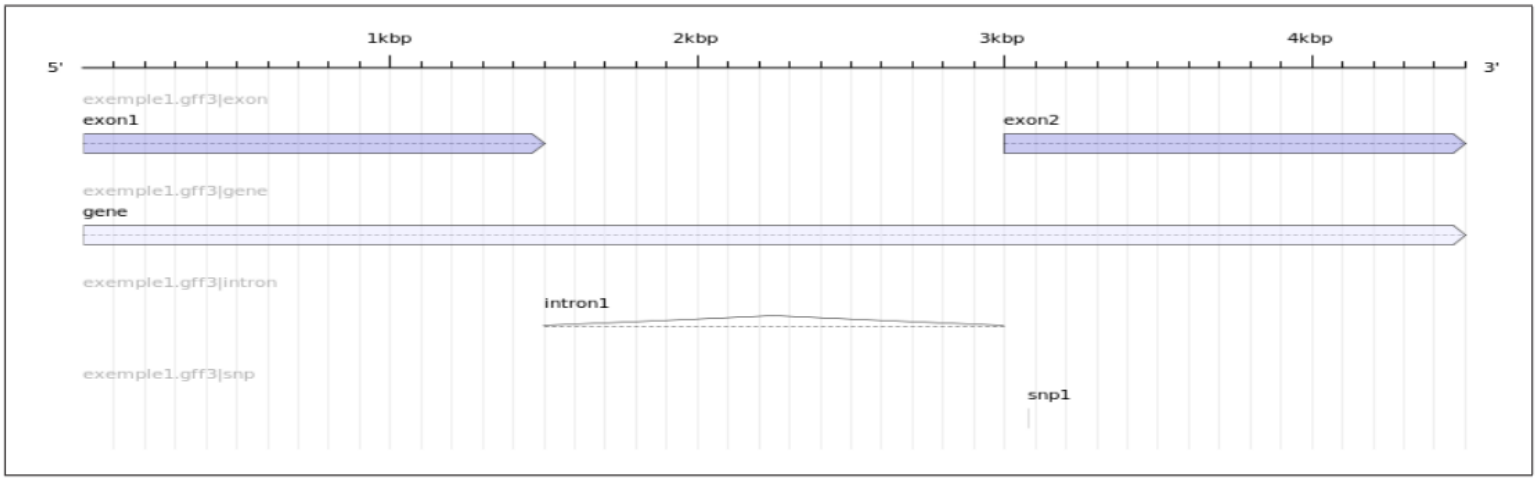
```
#nexus

BEGIN Taxa;
DIMENSIONS ntax=6;
TAXLABELS
[1] 'Europe'
[2] 'Albania'
[3] 'Andorra'
[4] 'Belarus'
[5] 'Belgium'
[6] 'BosniaHerzeg'
;
END; [Taxa]
BEGIN Distances;
DIMENSIONS ntax=6;
FORMAT labels=left diagonal triangle=both;
MATRIX
[1] 'Europe'        0 64 48 37 41 57
[2] 'Albania'       64 0 73 81 51 75
[3] 'Andorra'       48 73 0 50 70 68
[4] 'Belarus'       37 81 50 0 46 68
[5] 'Belgium'       41 51 70 46 0 70
[6] 'BosniaHerzeg'  57 75 68 68 70 0
;
END; [Distances]
```

→ Le format Phylip (suite EMBOSS) : la 1re ligne contient le nb de séquences suivi de leur taille (les séq. ayant été alignées, elles ont nécessairement toutes la même taille). Viennent ensuite les informations concernant chaque espèce ou gène

```
        7      50
C1          GCCAACCCCA CGGTCACTCT GTTCCCGCCC TCCTGGAGCT CCAAGACAAG
C2          GCTGCCCCCT CGGTCACTCT GTTCCCGCCC TCCTGGAGCT TCAAGACAAG
C3          GCTGCCCCCT CGGTCACTCT GTTCCCACCC TCCTGGAGCT TCAAGACAAG
C4          GAGACACCTT CATCTCCTCT GACCCCAGAG GCAGGGAGCT CCAAGACAAG
C5          GCCACCCCCT TGGTCACTCT GTTC---CCC TCCTGGAGCT CCAAGACAAG
C6          GCTGCCCCAT CGGTCACTCT GTTCCCGCCC TCCTGGAGCT TCAAGACAAG
C7          GCTGCCCCCT CGGTCACTCT GTTCCCACCC TCCTGGAGCT TCAAGACAAG
```

4. Les formats d'échange
   - → GFF (*Generic Feature Format*) : exploiter la richesse informationnelle des annotations au travers d'une structure tabulaire à 9 colonnes
   - → Rendre compte de la présence de sous-régions fonctionnelles imbriquées
   - → Outils de validation

```
##gff-version 3
##gff-version 3
ctg123 GenBank  exon 1 1500 . + . ID=exon1;note=chromobox homolog 8
ctg123 GenBank  gene 1 4500 . + . ID=gene;Dbxref=GeneID:779897
ctg123 GenBank  exon 3000 4500 . + . ID=exon2
ctg123 GenBank  intron 1501 2999 . + . ID=intron1
ctg123 GenBank  snp 3080 3080 . + . ID=snp1
```



→ ASN.1 (*Abstract Syntax Notation number one*) : norme qui définit un formalisme de description de types de données abstraits

Standard international ; format semi-structuré ; format de base pour les données du NCBI

```
Seq-entry ::= set {
  level 1 ,
  class nuc-prot ,
  descr {
    title "Mus musculus Brca1 mRNA, and translated products" ,
    source {
      org {
        taxname "Mus musculus" ,
        db {
          {
            db "taxon" ,
            tag
              id 10090 } } ,
        orgname {
          name
            binomial {
              genus "Mus" ,
              species "musculus" } , ...
```

→ XML (eXtensible Markup Language) : fournit une moyen pour implémenter des ontologies (vocabulaire structuré)

Langage à balise comme HTML ; standard international ; format semi-structuré

```
<?xml version="1.0"?>
<!DOCTYPE GBSeq PUBLIC "-//NCBI//NCBI GBSeq/EN" http://www.ncbi.nlm.nih.gov/dtd/NCBI_GBSeq.dtd">
<GBSet>
<GBSeq>
  <GBSeq_locus>MMU35641</GBSeq_locus>
  <GBSeq_length>5538</GBSeq_length>
  <GBSeq_strandedness value="not-set">0</GBSeq_strandedness>
  <GBSeq_moltype value="mrna">5</GBSeq_moltype>
  <GBSeq_topology value="linear">1</GBSeq_topology>
  <GBSeq_division>ROD</GBSeq_division>
  <GBSeq_update-date>18-OCT-1996</GBSeq_update-date>
  <GBSeq_create-date>25-OCT-1995</GBSeq_create-date>
  <GBSeq_definition>Mus musculus Brca1 mRNA, complete cds</GBSeq_definition>
  <GBSeq_primary-accession>U35641</GBSeq_primary-accession>
  <GBSeq_accession-version>U35641.1</GBSeq_accession-version>
  ...
```

5. Formats correspondants aux alignements multiples
   - → MSF *Multiple Sequence Format*
   - → MAF*Multiple Alignment Format*
   - → ALN (CLUSTALW)
   - → ...

6. ...

- Cette liste de formats classiques en bioinformatique n'est bien sûr pas exhaustive

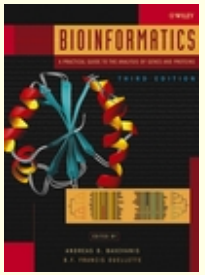
- Pour certains formats, représentation entrelacée (interleaved) ou séquentielle (sequential) possible

Certains formats tendent à devenir des standards ou sont plus génériques que d'autres, il faut encourager leur utilisation

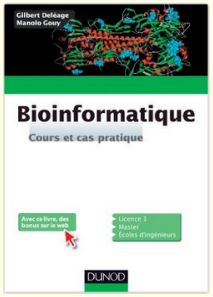
Conversion des formats et visualisation

- Les logiciels peuvent n'accepter qu'un type de format particulier
- Des outils existent pour transformer un format en un autre :
  - → ReadSeq (http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi)
    Formats supportés : IG/Stanford, GenBank/GB, NBRF, EMBL, GCG, DNAStrider, Fitch, Pearson/Fasta, Phylip3.2, Phylip, PIR/CODATA, MSF, ASN.1 et PAUP/NEXUS
  - → Outils de conversion de l'EBI (http://www.ebi.ac.uk/Tools/sfc/)
  - → Fonctions BioPython, BioPerl, ...
  - → ...
- Des outils de visualisation de données existent et acceptent, en général, plusieurs formats (AnnotationSketch, Seaview, MEGA, ...)

- Chap 1-4 dans "Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition"
  [Baxevanis & Ouellette,2005]



- "Bioinformatique - Cours et cas pratique"
  [Deléage Gouy, 2013]



- Cours en ligne de Maude Pupin et Jean-Stéphane Varré (université de Lille)