Analyse de séquences

Partie I : Banques de données

Sèverine Bérard

AMAP - Université Montpellier 2

Plan

Pourquoi des banques de séquences ?

- Comprendre ce qui constitue le “livre de la vie” : comment les millions ou milliards de bases du génome d'un organisme contiennent toute l'information nécessaire à la cellule pour la production des nombreux processus métaboliques essentiels à la survie de l'organisme ; information qui est propagée de génération en génération ?

- Pour comprendre comment cette collection de simple nucléotides est à la base de la vie, on doit collecter d'immense quantité de données de séquences et les stocker de manière à pouvoir les analyser et les retrouver facilement.

Création et maintenance des banques de données biologiques

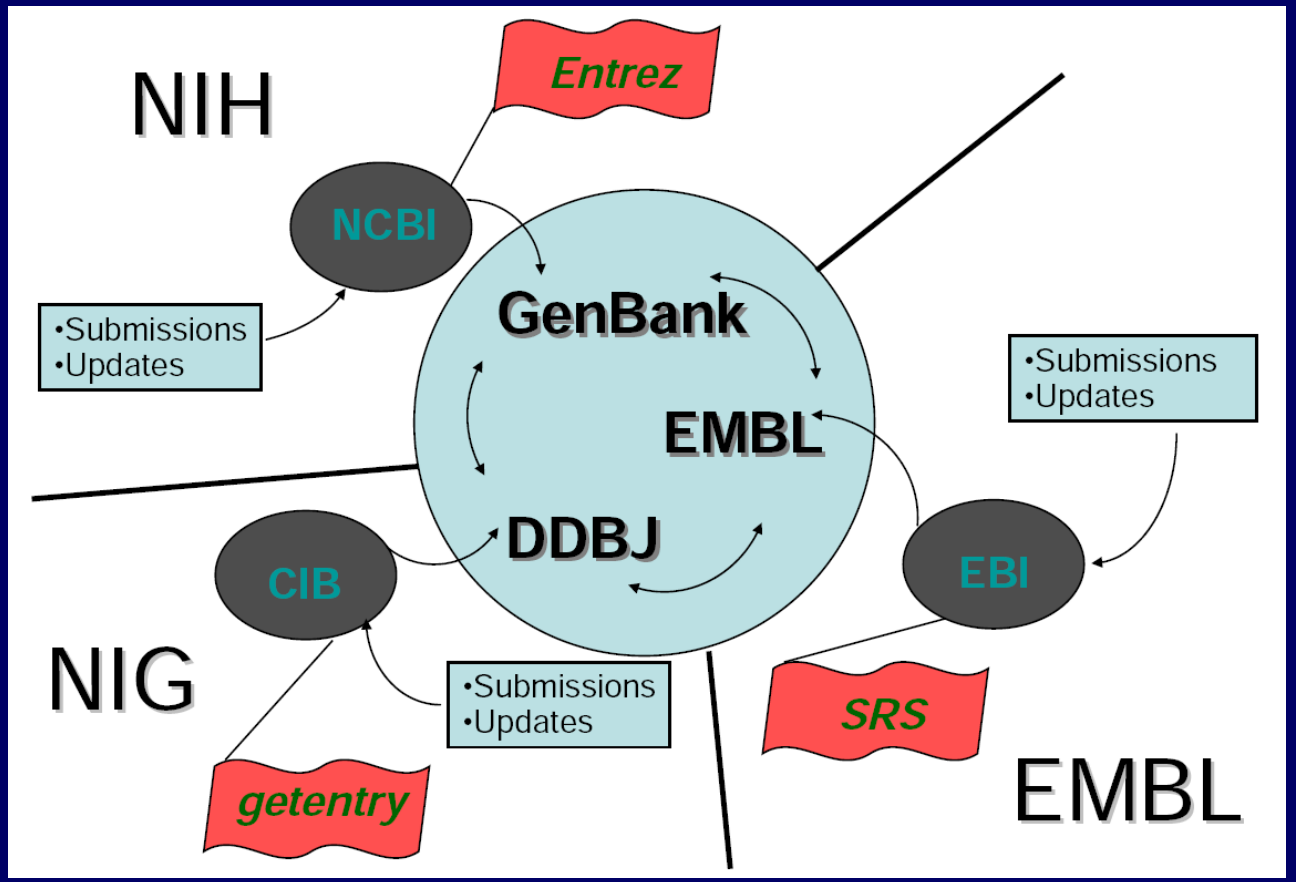
Bref historique (1/3)

- $\sim$ 1960s Margaret Dayhoff et ses collaborateurs du PIR (Protein Information Ressource) 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

- 1982 - EMBL (European Molecular Biology Laboratory) et le NCBI (National Center for Biotechnology Information) transcrivent et interprètent les 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 Meeting de ces 3 groupes (appelés maintenant *International Nucleotide Sequence Database Collaboration*)
  - → Accord pour un format commun
  - → Échange des données journalier
  - → Chaque groupe gère les mises à jour des séquences qu'il a créées



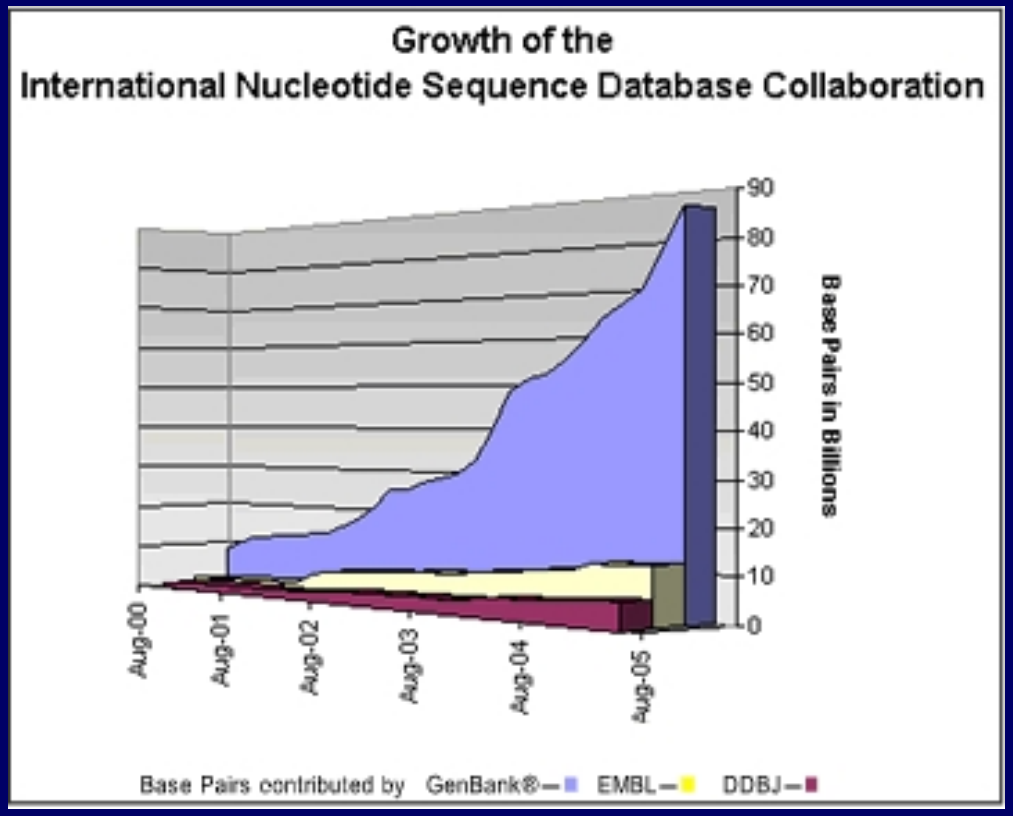
Bref historique (3/3)

- 1986 Création de SWISSPROT (3900 prot.) : conversion de l'Atlas 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
  ⇒ Création de TrEMBL (Translation of TrEMBL nucleotide sequences)

- Fin 2003 Création de UNIPROT (UNIversal PROTein Ressource) en joignant les informations de SWISSPROT, TrEMBL et PIR.
  UNIPROT à 3 composants :
  → UniProtKB (*UniProt KnowledgeBase*) = SWISSPROT + TrEMBL
  → UniRef (*UniProt Reference Clusters*) regroupe en 1 entrée les 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



International sequence databases exceed 100 gigabases (août 2005)

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) ?
  - → . . .

Les données

- Quelle est la fiabilité des données ?
  - → Erreurs de frappe
  - → Erreurs d'annotation
  - → Redondance

- Où sont les données ?
  - → Dans les laboratoires
  - → Dans des 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 secondaire
  - → Banque généraliste ou spécialisée

Banques de données primaires et secondaires

- Banques primaires $\sim$ archives
- Banques secondaires $\sim$ 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

Les banques de données généralistes

- 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 de localisation (*mapping databases* GeneLoc)
- Banques de structures 3D de macromolécules (PDB)
- Banques génomiques (UCSC, Ensembl)
- 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
- Ex : 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

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.
```

Division en organisme

Utilisée dans les 3 principales banques de séquences nucléiques

Division		DDBJ	EMBL	GenBank
BCT	Bacterial	√		√
FUN	Fungal		√	
HUM	Homo sapiens	√	√	
INV	Invertebrate	√	√	√
MAM	Other mammalian	√	√	√
ORG	Organelle		√	
PHG	Phage	√	√	√
PLN	Plant (also see FUN)	√	√	√
PRI	Primate (also see HUM)	√	√	√
PRO	Prokaryotic		√	
ROD	Rodent	√	√	√
SYN	Synthetic and chimeric	√	√	√
VRL	Viral	√	√	√
VRT	Other vertebrate	√	√	√

Les *flatfiles* en détails (2/8)

- Le numéro d'accession : un identificateur unique

```
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 *flatfiles* en détails (3/8)

- 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
```

Les *flatfiles* en détails (4/8)


```
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*)

```
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 (6/8)


```
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 (8/8)

- 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

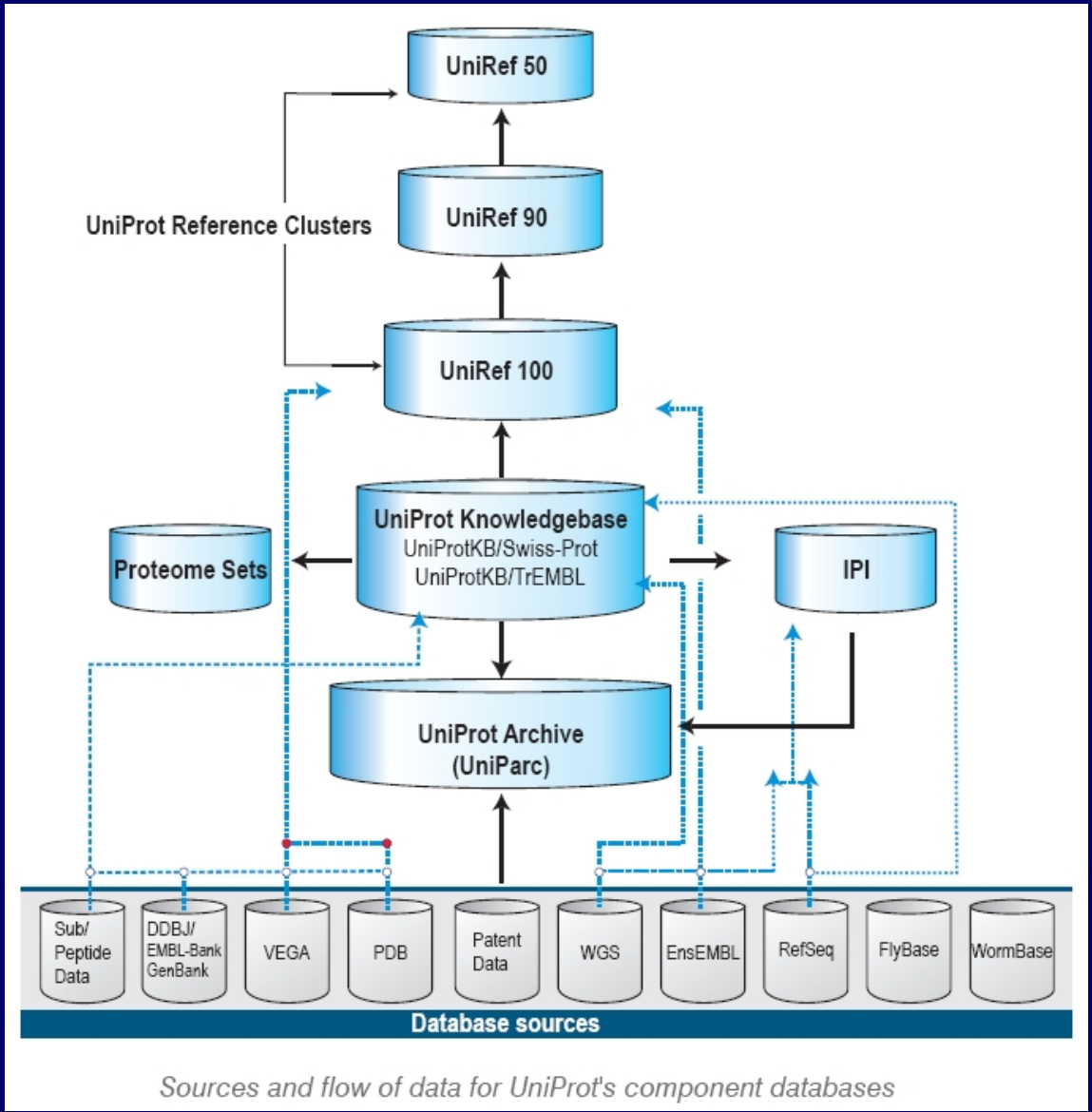
Les banques de séquences protéiques

- 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



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 SWISSPROT

Search Swiss-Prot/TrEMBL for U49845 Go Clear

UniProtKB/Swiss-Prot entry P39076

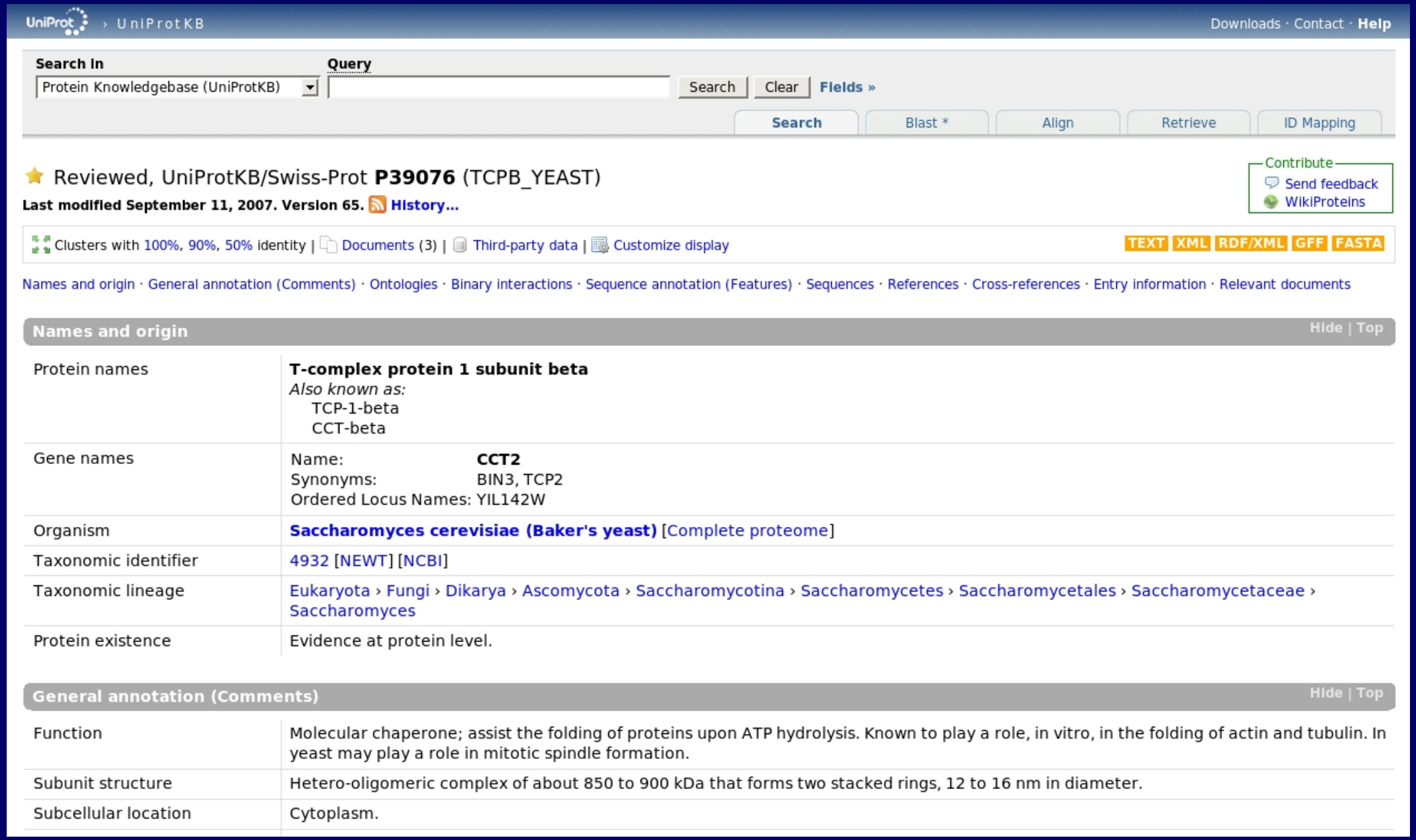
Printer-friendly view
Submit update
Quick BlastP search
Entry history

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**TCPB_YEAST**
Primary accession number	**P39076**
Secondary accession numbers	None
Integrated into Swiss-Prot on	February 1, 1995
Sequence was last modified on	February 1, 1995 (Sequence version 1)
Annotations were last modified on	October 31, 2006 (Entry version 55)
Name and origin of the protein	
Protein name	**T-complex protein 1 subunit beta**
Synonyms	**TCP-1-beta** **CCT-beta**
Gene name	**Name: CCT2** Synonyms: BIN3, TCP2 OrderedLocusNames: YIL142W
From	Saccharomyces cerevisiae (Baker's yeast) [TaxID: 4932]
Taxonomy	Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
References	
[1]	NUCLEOTIDE SEQUENCE [GENOMIC DNA]. **STRAIN**=ATCC 204511 / S288c / AB972; PubMed=7908441 [NCBI, ExPASy, EBI, Israel, Japan] Miklos D., Caplan S., Mertens D., Hynes G., Pitluk Z., Kashi Y., Harrison-Lavoie K., Stevenson S., Brown C., Barrell B.G., Horwich A.L., Willison K.; "Primary structure and function of a second essential member of the heterooligomeric TCP1 chaperonin complex of yeast, TCP1 beta."; Proc. Natl. Acad. Sci. U.S.A. 91:2743-2747(1994).
[2]	NUCLEOTIDE SEQUENCE [GENOMIC DNA].

Exemple d'entrée SWISSPROT, nouvelle version



UniProt › UniProtKB — Downloads · Contact · Help

Search In: Protein Knowledgebase (UniProtKB) — Query: Search | Clear | Fields »

Search | Blast * | Align | Retrieve | ID Mapping

Reviewed, UniProtKB/Swiss-Prot **P39076** (TCPB_YEAST)

Last modified September 11, 2007. Version 65. History...

Contribute: Send feedback · WikiProteins

Clusters with 100%, 90%, 50% identity | Documents (3) | Third-party data | Customize display — TEXT XML RDF/XML GFF FASTA

Names and origin · General annotation (Comments) · Ontologies · Binary interactions · Sequence annotation (Features) · Sequences · References · Cross-references · Entry information · Relevant documents

Names and origin — Hide | Top

Protein names	**T-complex protein 1 subunit beta** *Also known as:* TCP-1-beta CCT-beta
Gene names	Name: **CCT2** Synonyms: BIN3, TCP2 Ordered Locus Names: YIL142W
Organism	**Saccharomyces cerevisiae (Baker's yeast)** [Complete proteome]
Taxonomic identifier	4932 [NEWT] [NCBI]
Taxonomic lineage	Eukaryota › Fungi › Dikarya › Ascomycota › Saccharomycotina › Saccharomycetes › Saccharomycetales › Saccharomycetaceae › Saccharomyces
Protein existence	Evidence at protein level.

General annotation (Comments) — Hide | Top

Function	Molecular chaperone; assist the folding of proteins upon ATP hydrolysis. Known to play a role, in vitro, in the folding of actin and tubulin. In yeast may play a role in mitotic spindle formation.
Subunit structure	Hetero-oligomeric complex of about 850 to 900 kDa that forms two stacked rings, 12 to 16 nm in diameter.
Subcellular location	Cytoplasm.

Banques de données de localisation

- 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 d'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), . . .)

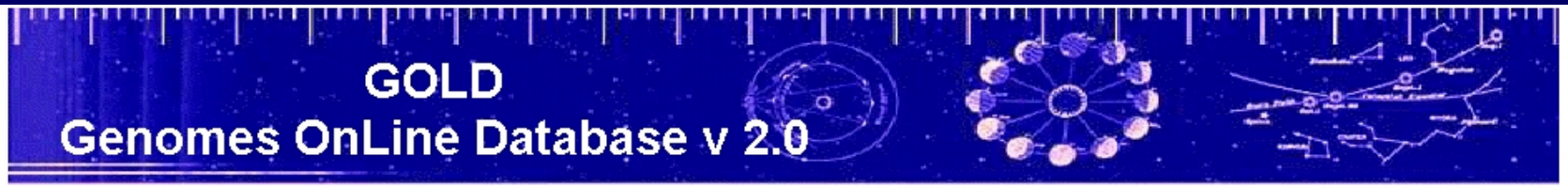
Banques de données génomiques

- 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



Contact: Genomesonline	Last Update: September 22, 2007	Location www.genomesonline.org
655 Published Complete Genomes	Search GOLD: 2934 genome projects	108 Metagenomes
59 Archaeal Ongoing Genomes	1337 Bacterial Ongoing Genomes	775 Eukaryotic Ongoing Genomes

http://www.genomesonline.org

Banques de familles de protéines

- 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)

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

Banques de motifs et domaines protéiques

- Une famille de protéine 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 : google, . . .

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

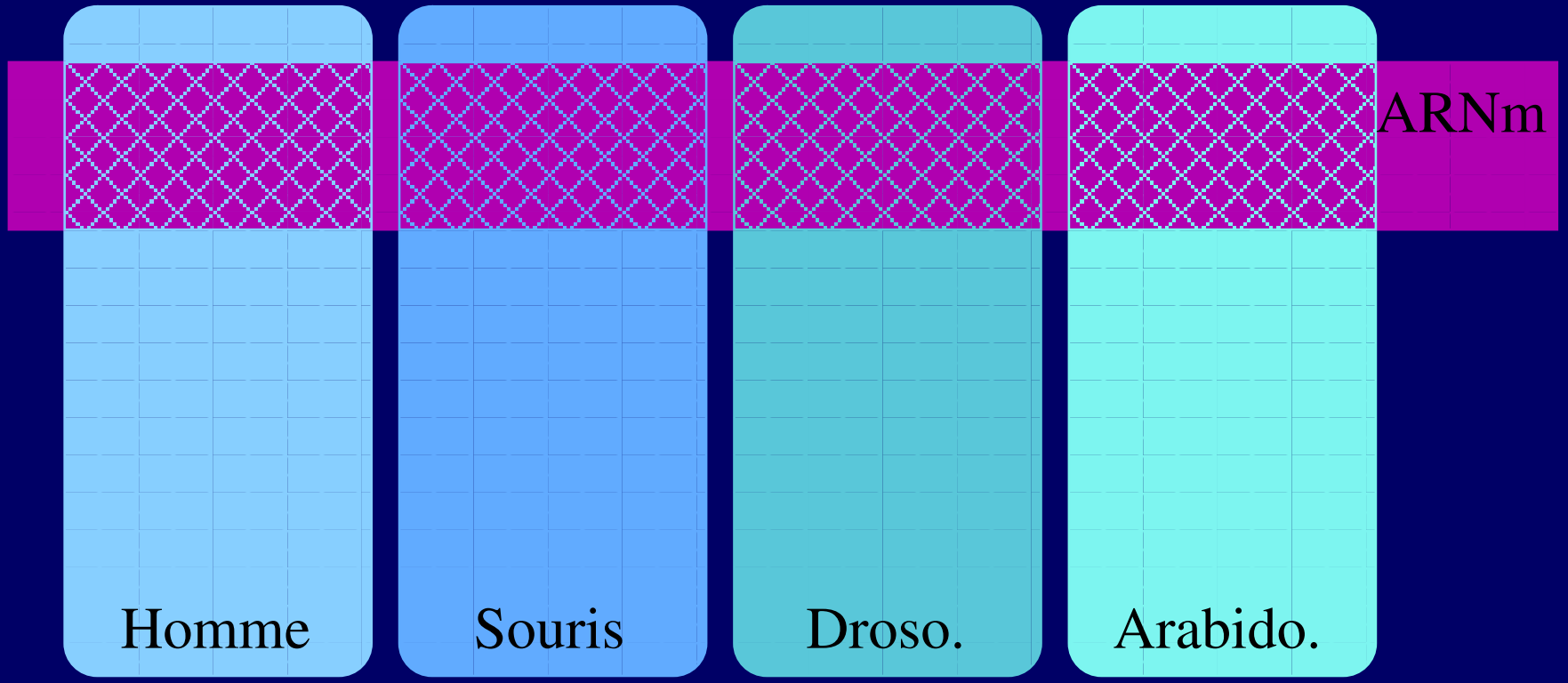
- Des catalogues d'outils
  → EMBOSS : http://emboss.sourceforge.net
  → Institut Pasteur : http://bionet.pasteur.fr/
  → . . .

- Des systèmes d'interrogation
  → Entrez (NCBI)
  → SRS (EBI)

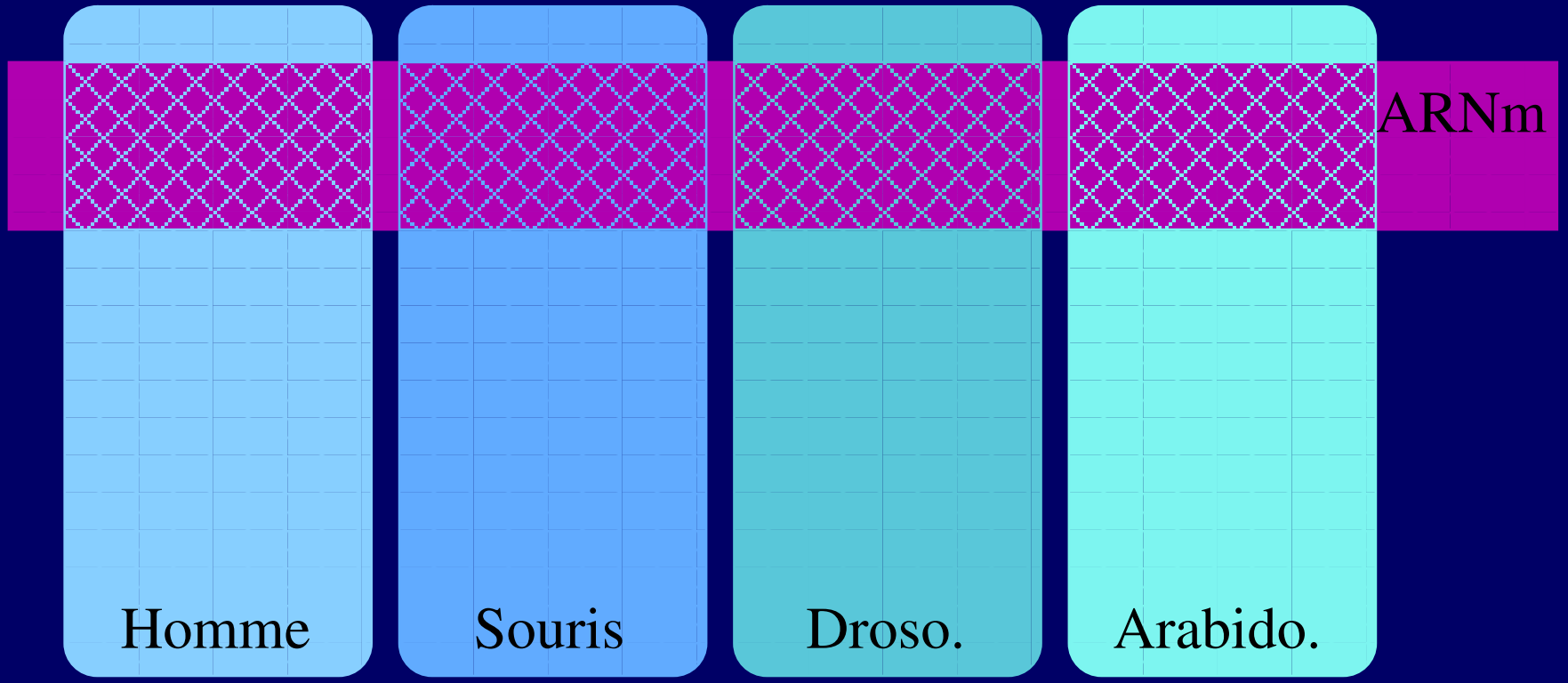
Interrogation d'une banque

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

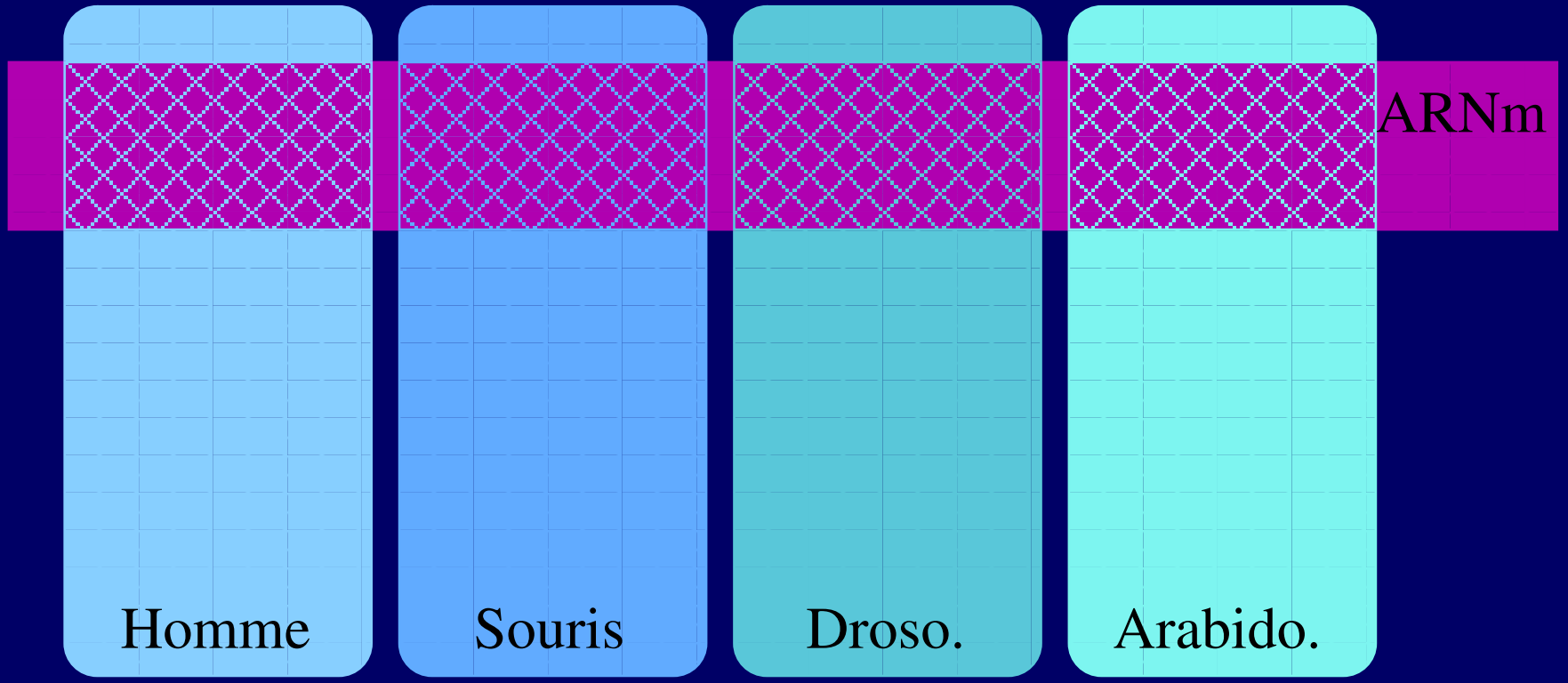
- Contraintes pour un système d'interrogation
  - → Obtention de données pertinentes (pas trop de résultats mais tous ceux relatifs à notre problématique)
  - → Simplicité d'utilisation (syntaxe d'interrogation intuitive)
  - → Réponse rapide
  - → Possibilité d'analyse des résultats (couplage à des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



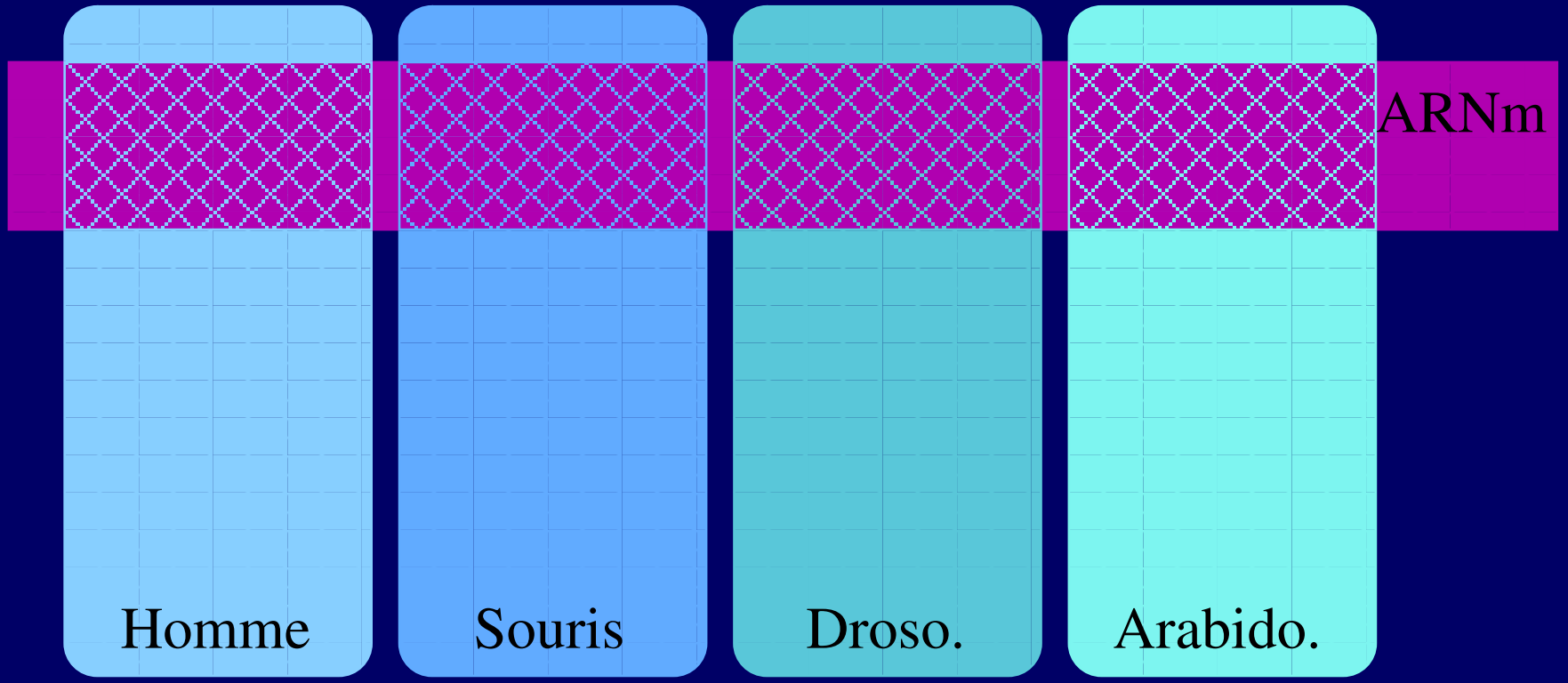
- Toutes séquences contenant 'Homme' :
- Séquences contenant 'Homme' ou 'Souris' :
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



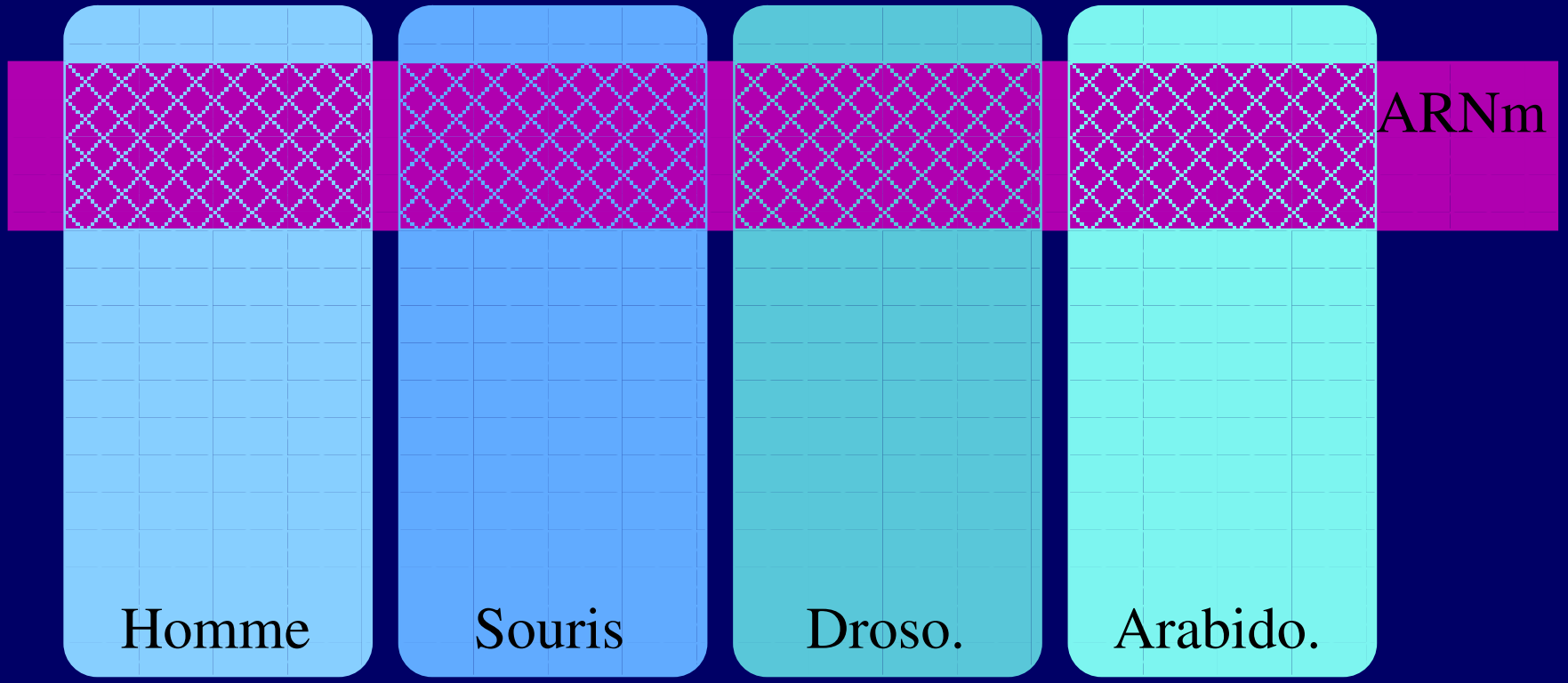
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' :
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



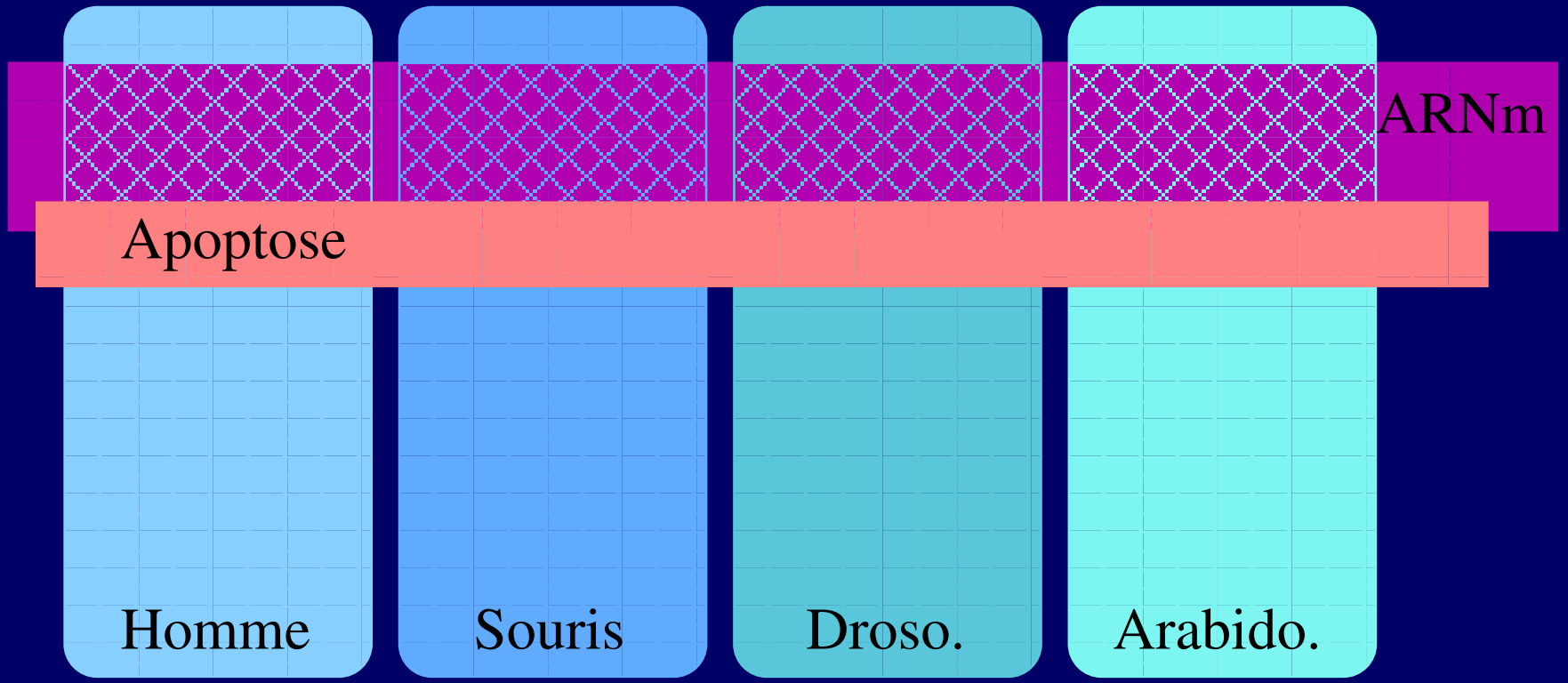
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



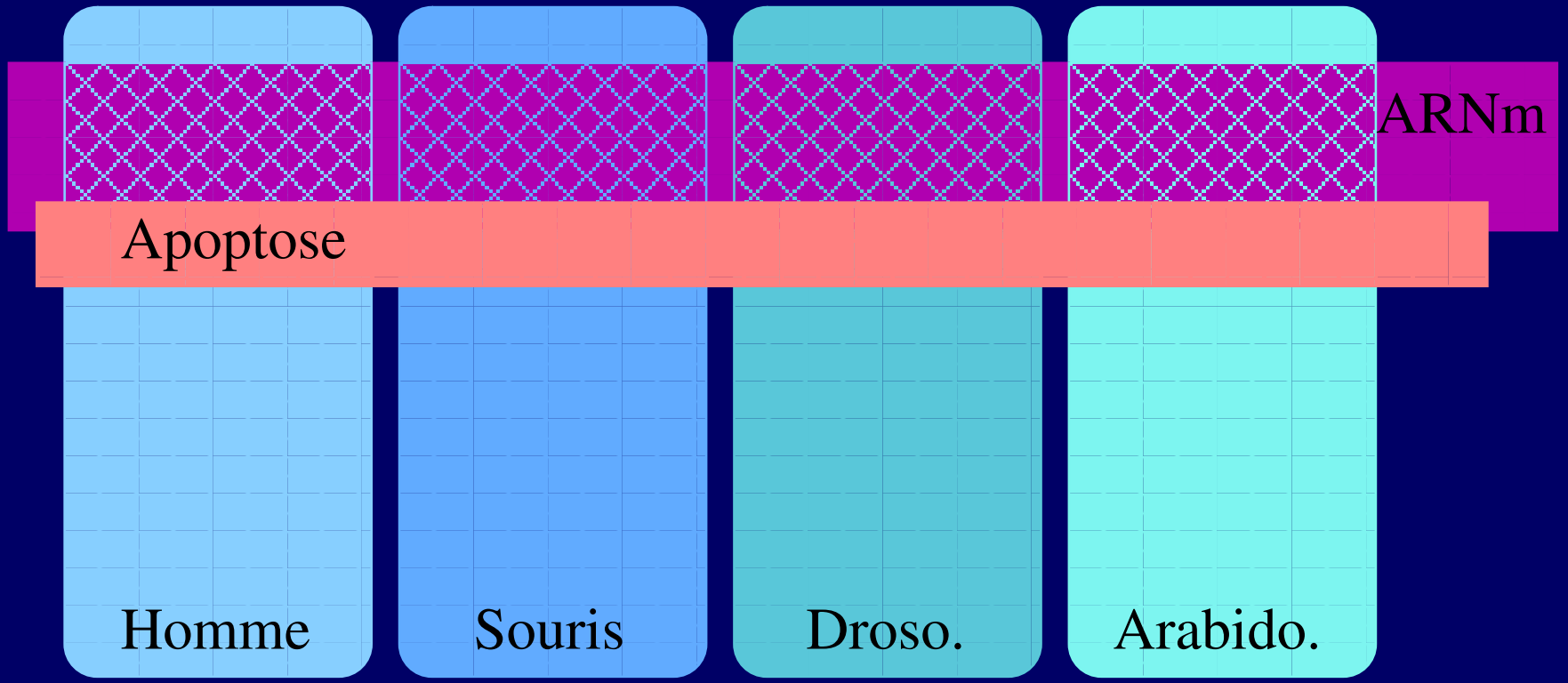
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile : Droso. ET ARNm
- Tous les ARNm sauf ceux d'arabidopsis :



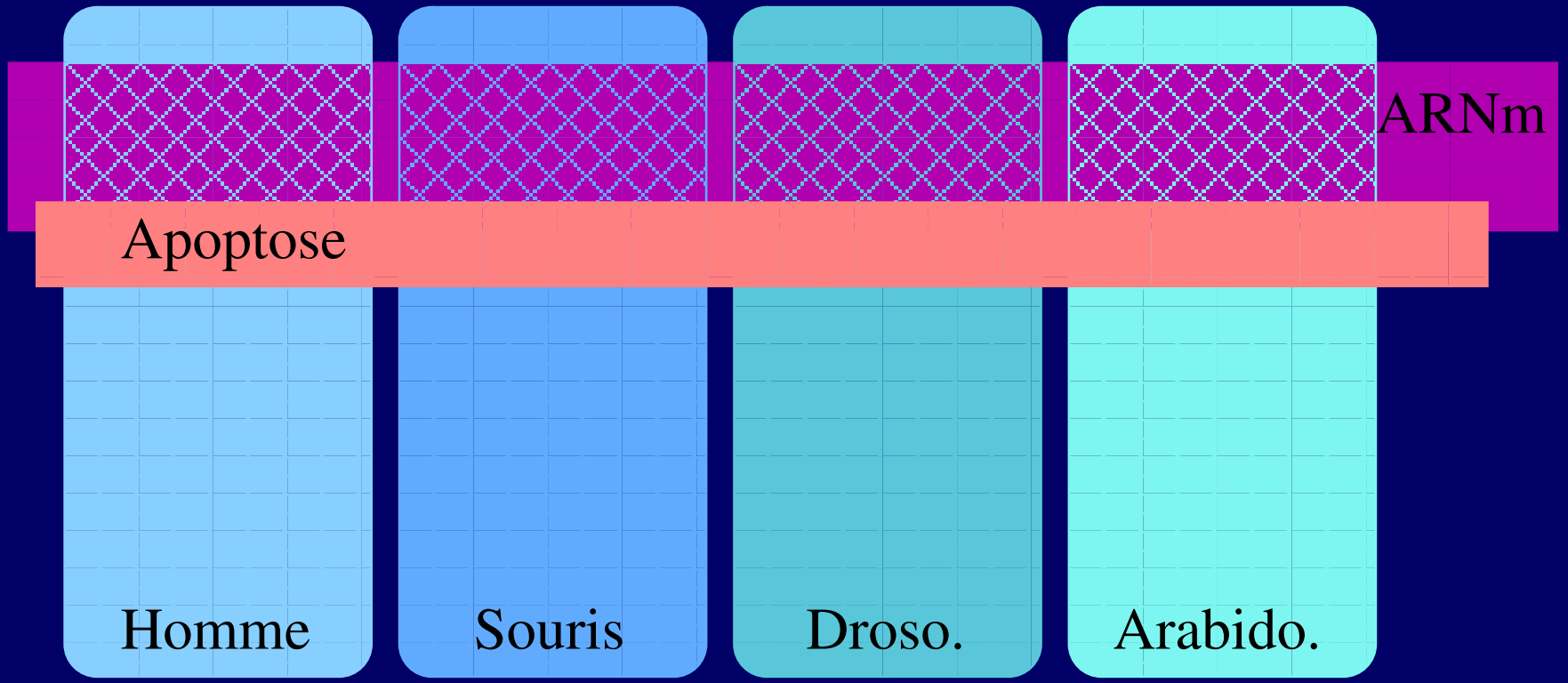
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile : Droso. ET ARNm
- Tous les ARNm sauf ceux d'arabidopsis : ARNm NON Arabido.



- 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 :

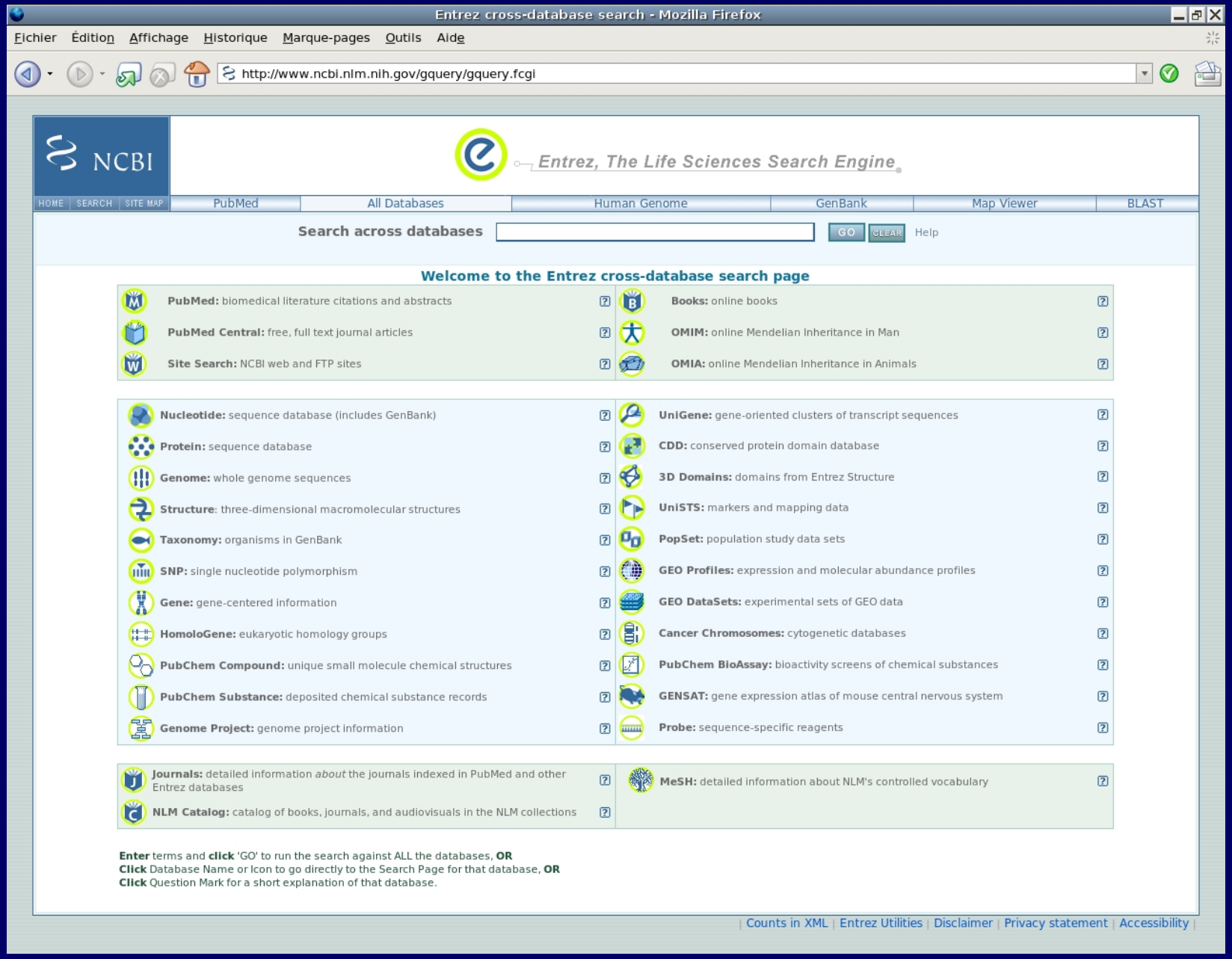


- ARNm impliqués dans l'apoptose chez l'homme :
  ARNm ET Apoptose ET Homme
- Séquences de la souris ou de la droso. impliquées dans l'apoptose mais qui ne sont pas des ARNm :
  ((Souris OU Droso.) ET Apoptose) NON ARNm

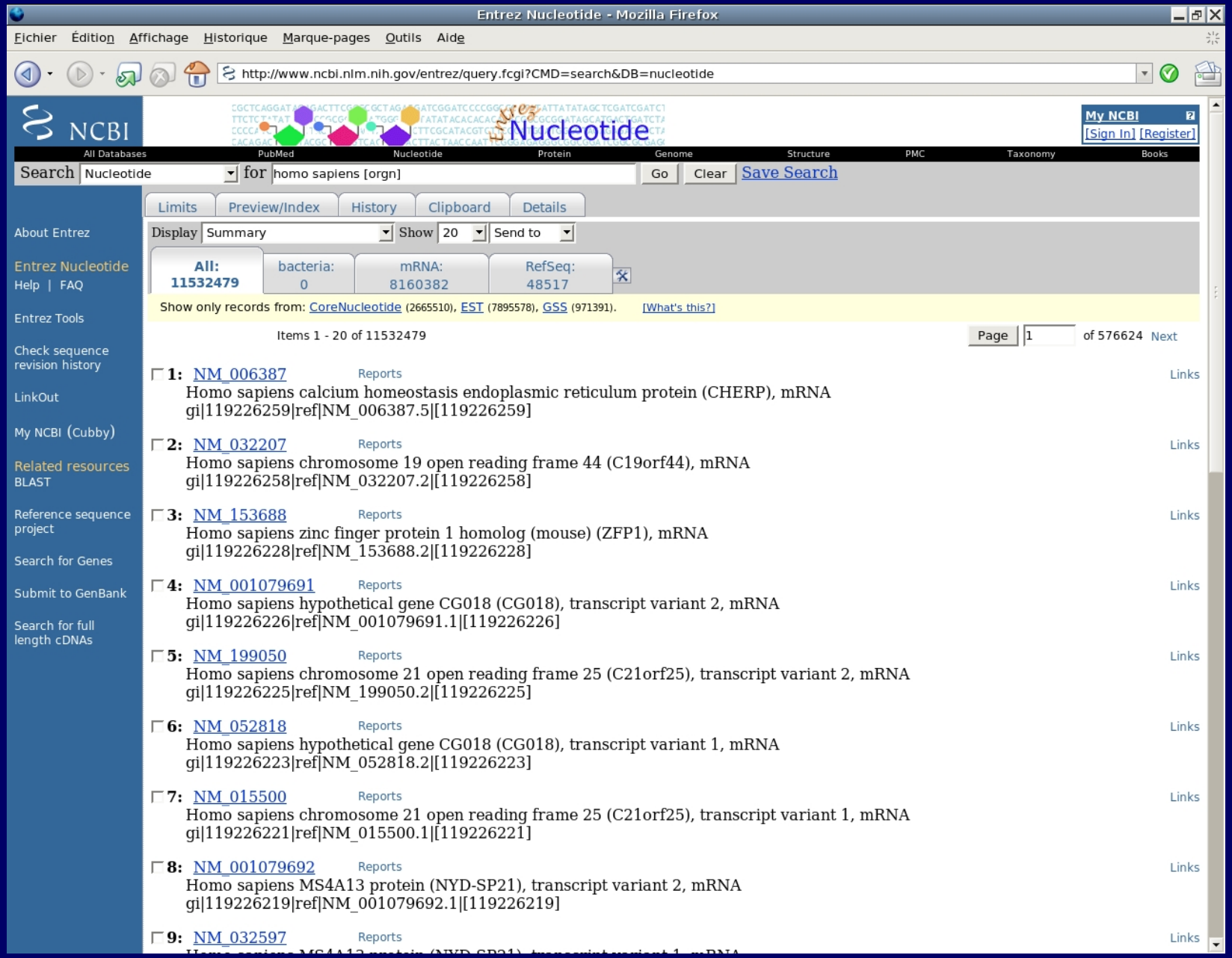
Entrez, le système d'interrogation du NCBI

- Interface propriétaire (ne peut être installée par autrui)
- Opérateurs en majuscule : AND, OR, NOT
- Nom du champs entre crochets `homo sapiens [organism]`
- Aide dans "Preview/Index"
- Historique (lien "History")
- Ajout de limites (lien "Limits")
- Sauvegarde, format
  → Boutons "Display", "send to"
  → Menus déroulants associés

Entrez, page d'accueil



Entrez, résultats d'une recherche



SRS, le système d'interrogation de l'EBI

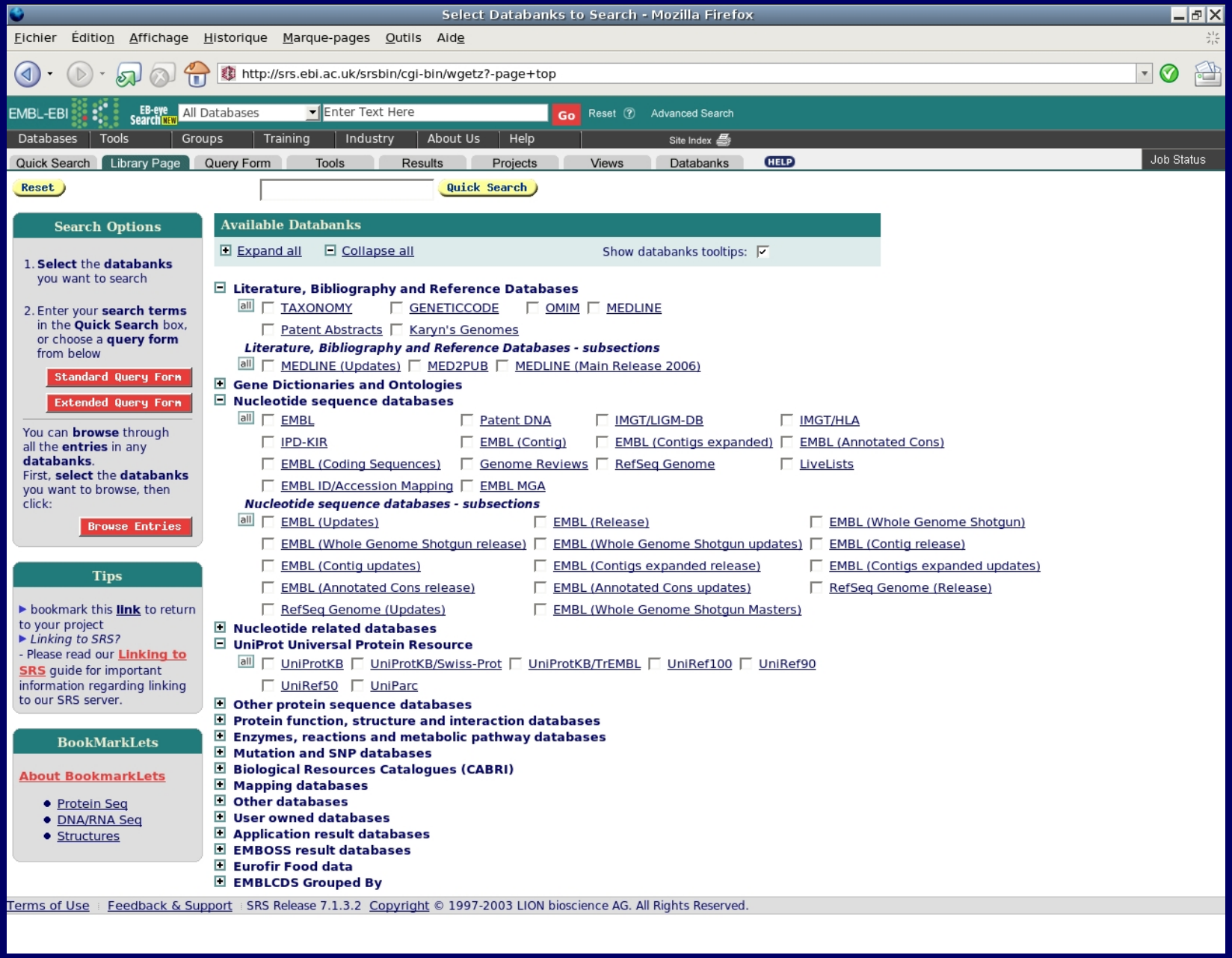
- Système libre, de nombreux miroir existent

- Onglet "Library"
  → Choix de la ou des banques interrogées
  → Lancement du formulaire

- Onglet "Results"
  → Historique des requêtes

- Onglet "View"
  → Création du format d'affichage

SRS, onglet Library



Interrogation *via* SRS

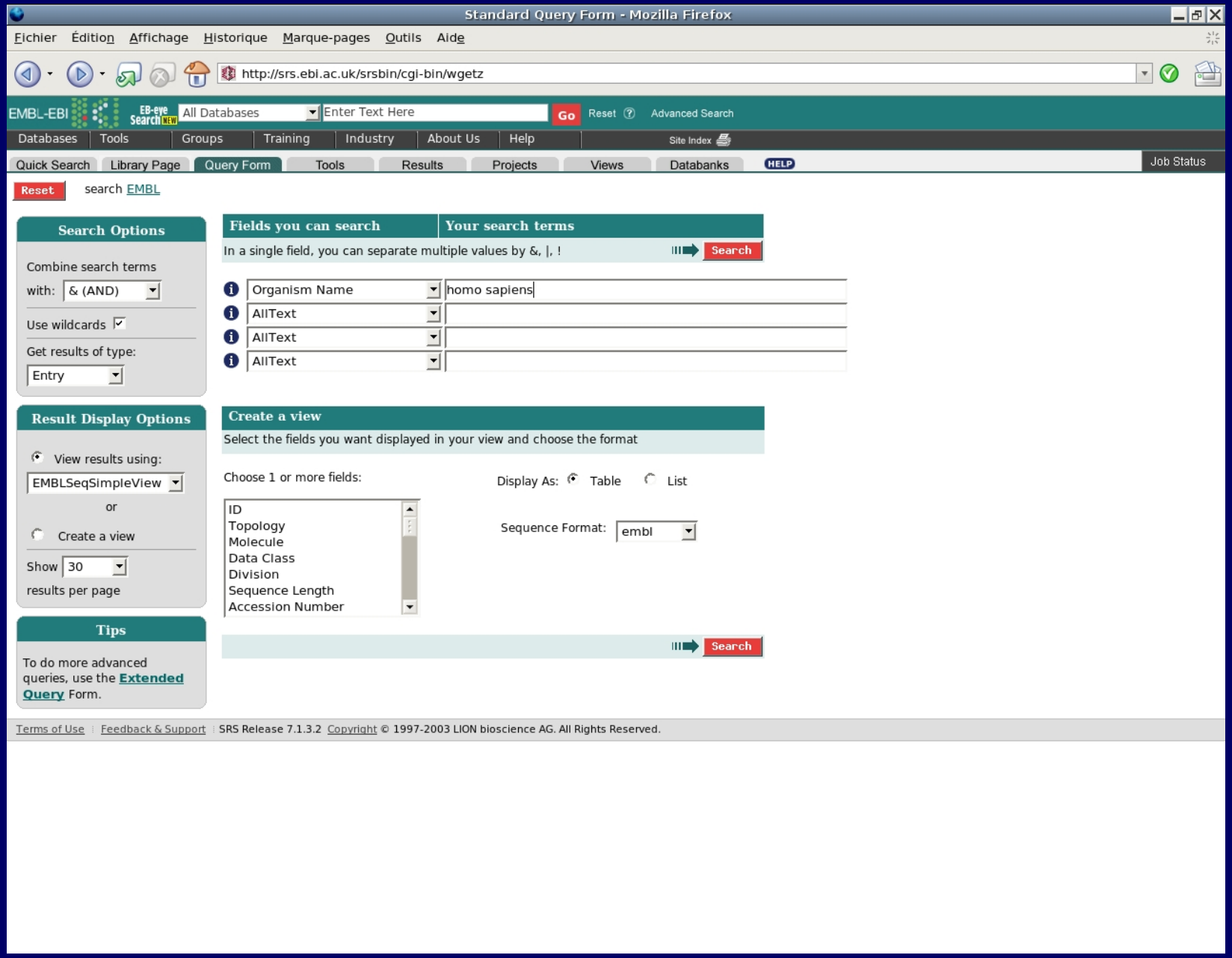
- Opérateurs : & (et), | (ou), ! (non)

- Noms des champs dans des menus déroulants

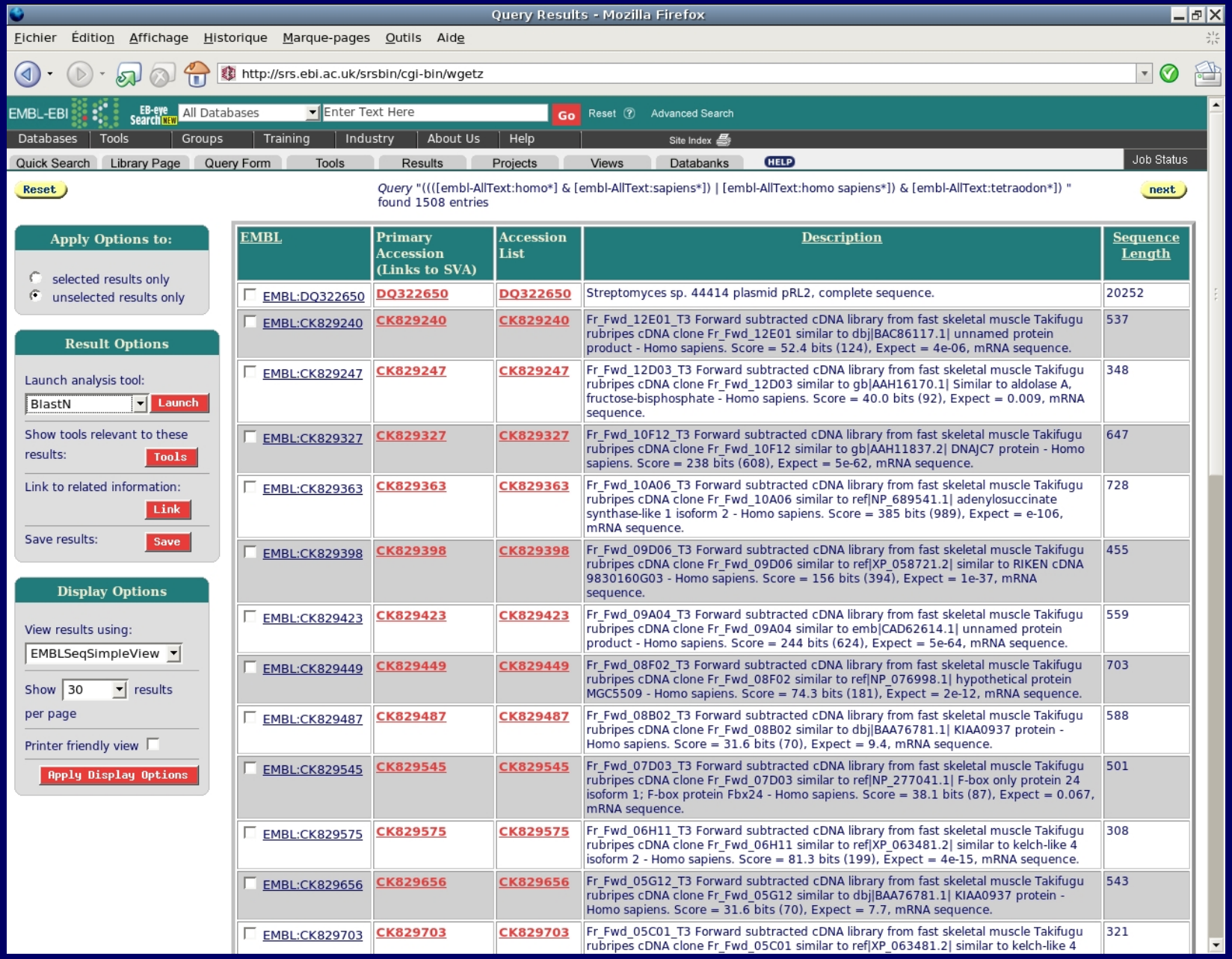
- Sauvegarde, format
  → Bouton "Save"
  → Bouton "Rerun query"
  → Options associés

- Analyse bioinformatique des entrées
  → Bouton "Launch"

SRS, page d'accueil



SRS, résultats d'une recherche



EMBL	Primary Accession (Links to SVA)	Accession List	Description	Sequence Length
EMBL:DQ322650	DQ322650	DQ322650	Streptomyces sp. 44414 plasmid pRL2, complete sequence.	20252
EMBL:CK829240	CK829240	CK829240	Fr_Fwd_12E01_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_12E01 similar to dbj|BAC86117.1| unnamed protein product - Homo sapiens. Score = 52.4 bits (124), Expect = 4e-06, mRNA sequence.	537
EMBL:CK829247	CK829247	CK829247	Fr_Fwd_12D03_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_12D03 similar to gb|AAH16170.1| Similar to aldolase A, fructose-bisphosphate - Homo sapiens. Score = 40.0 bits (92), Expect = 0.009, mRNA sequence.	348
EMBL:CK829327	CK829327	CK829327	Fr_Fwd_10F12_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_10F12 similar to gb|AAH11837.2| DNAJC7 protein - Homo sapiens. Score = 238 bits (608), Expect = 5e-62, mRNA sequence.	647
EMBL:CK829363	CK829363	CK829363	Fr_Fwd_10A06_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_10A06 similar to ref|NP_689541.1| adenylosuccinate synthase-like 1 isoform 2 - Homo sapiens. Score = 385 bits (989), Expect = e-106, mRNA sequence.	728
EMBL:CK829398	CK829398	CK829398	Fr_Fwd_09D06_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_09D06 similar to ref|XP_058721.2| similar to RIKEN cDNA 9830160G03 - Homo sapiens. Score = 156 bits (394), Expect = 1e-37, mRNA sequence.	455
EMBL:CK829423	CK829423	CK829423	Fr_Fwd_09A04_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_09A04 similar to emb|CAD62614.1| unnamed protein product - Homo sapiens. Score = 244 bits (624), Expect = 5e-64, mRNA sequence.	559
EMBL:CK829449	CK829449	CK829449	Fr_Fwd_08F02_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_08F02 similar to ref|NP_076998.1| hypothetical protein MGC5509 - Homo sapiens. Score = 74.3 bits (181), Expect = 2e-12, mRNA sequence.	703
EMBL:CK829487	CK829487	CK829487	Fr_Fwd_08B02_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_08B02 similar to dbj|BAA76781.1| KIAA0937 protein - Homo sapiens. Score = 31.6 bits (70), Expect = 9.4, mRNA sequence.	588
EMBL:CK829545	CK829545	CK829545	Fr_Fwd_07D03_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_07D03 similar to ref|NP_277041.1| F-box only protein 24 isoform 1; F-box protein Fbx24 - Homo sapiens. Score = 38.1 bits (87), Expect = 0.067, mRNA sequence.	501
EMBL:CK829575	CK829575	CK829575	Fr_Fwd_06H11_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_06H11 similar to ref|XP_063481.2| similar to kelch-like 4 isoform 2 - Homo sapiens. Score = 81.3 bits (199), Expect = 4e-15, mRNA sequence.	308
EMBL:CK829656	CK829656	CK829656	Fr_Fwd_05G12_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_05G12 similar to dbj|BAA76781.1| KIAA0937 protein - Homo sapiens. Score = 31.6 bits (70), Expect = 7.7, mRNA sequence.	543
EMBL:CK829703	CK829703	CK829703	Fr_Fwd_05C01_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_05C01 similar to ref|XP_063481.2| similar to kelch-like 4	321

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=39;
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

→ 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" } , ...
```

Quelques formats classiques en bioinformatique

→ 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

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

Conversion des formats

- Les logiciels peuvent n'accepter qu'un type de format particulier

- Des outils existent pour transformer un format en un autre :

  → ReadSeq

  http://searchlauncher.bcm.tmc.edu/seq-util/Options/readseq.html

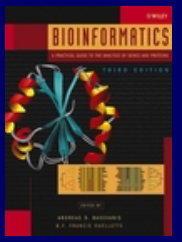
  Formats supportés : IG/Stanford, GenBank/GB, NBRF, EMBL, GCG, DNAStrider, Fitch, Pearson/Fasta, Phylip3.2, Phylip, PIR/CODATA, MSF, ASN.1 et PAUP/NEXUS

  → SEQIO (plus ancien et non maintenu)

  http: ://www.cs.ucdavis.edu/~gusfield/seqio.html

Références

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



- Cours en ligne de Maude Pupin (http://www2.lifl.fr/~pupin) et Jean-Stéphane Varré ((http://www2.lifl.fr/~varre), maîtres de conférences à l’université de Lille
- Illustrations :
  - → Transp. 4 : The Jackson Laboratory ; Presentations from 11th Annual Short Course on Genetic Approaches to Complex Heart, Lung, and Blood Diseases - 2006 ; Maglott 2 - Highlights of using NCBI’s Resources (http://pga.jax.org/hlb06coursefiles/coursemenu.html)
  - → Transp. 6 : http://www.uvm.edu/
  - → Transp. 29 : http://www.uniprot.org/