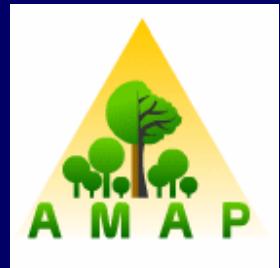


# Analyse de séquences

Partie I : Banques de données

Sèverine Bérard



AMAP - Université Montpellier 2



- Introduction
- Banques de données de séquences nucléiques
- Banques de données de séquences protéiques
- Encore des banques de données
- Interrogation des banques de données
- Formats de fichiers en bioinformatique
- Références

# 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

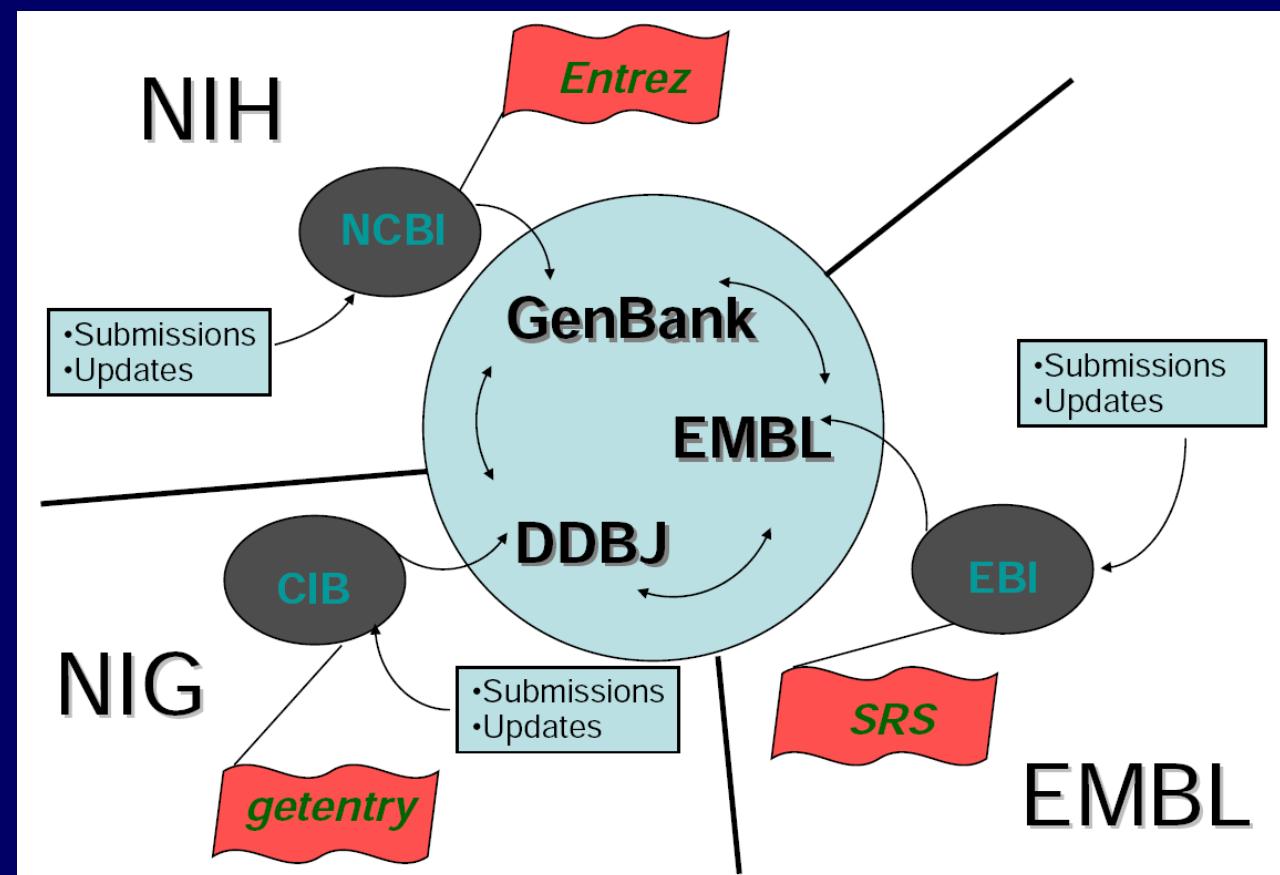
## Bref historique (1/3)

---

- ~ 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
  - 50 entrées
  - 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
  - Véritable explosion de la quantité de séquences disponibles
  - Rapidement, la DDBJ (DNA DataBank of Japan) se joint à la collaboration pour la collecte de données

## Bref historique (2/3)

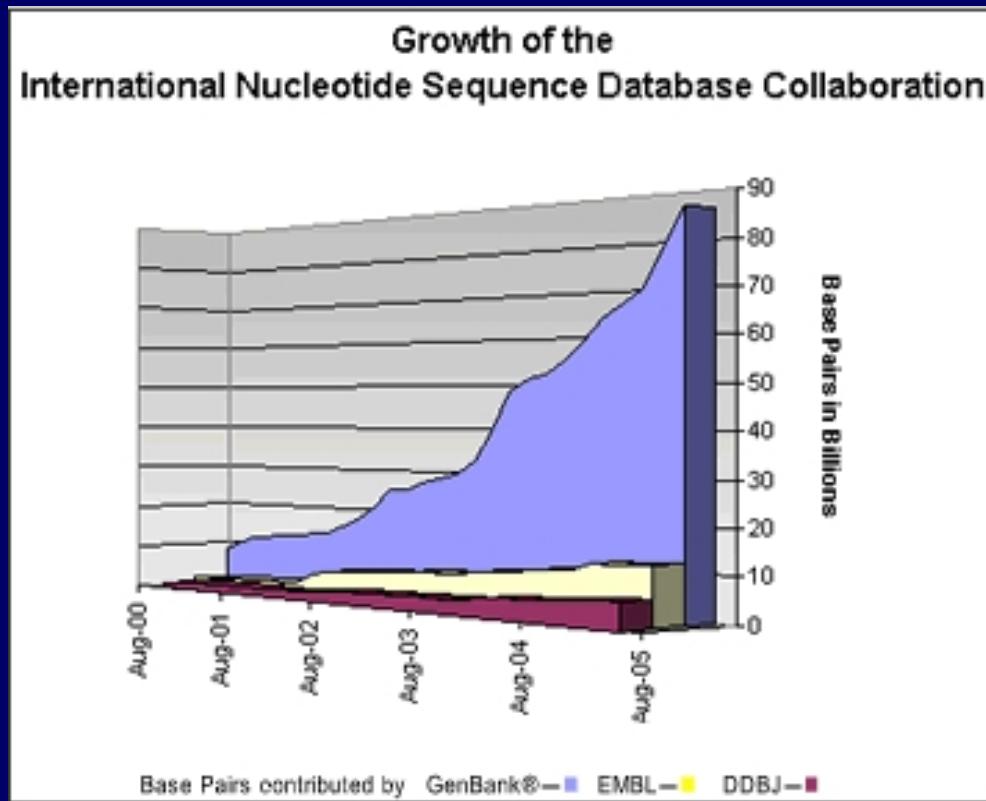
- 1988 Meeting de ces 3 groupes (appelés maintenant *International Nucleotide Sequence Database Collaboration*)
  - Accord pour un format commun
  - Échange des données journalier
  - Chaque groupe gère les mises à jour des séquences qu'il a créées



## 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  
⇒ Crédit à la création de TrEMBL (Translation of TrEMBL nucleotide sequences)
- Fin 2003 Crédit à la 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 identifiants

# 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
- Différenciations possibles des banques de données
  - Banque primaire ou secondaire
  - Banque généraliste ou spécialisée

- Banques primaires ~ archives
- Banques secondaires ~ 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
  - *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 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

- 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

- Introduction
- Banques de données de séquences nucléiques
- Banques de données de séquences protéiques
- Encore des banques de données
- Interrogation des banques de données
- Formats de fichiers en bioinformatique
- Références

- 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 versions disponibles plusieurs 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

- 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](#)"  
FT /db\_xref="[EMBL:U49845](#)"

# 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="SSIYNGISTGDLNNNTIADMRQLGIVESYKLKRAVSSASEA  
AEVLLRVDNIIRARPRTANRQHM"  
gene 687..3158  
/gene="AXL2"  
CDS 687..3158  
/gene="AXL2"

## 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

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	✓	✓	✓

- 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 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 <i>Saccharomyces cerevisiae</i>
JOURNAL	<i>Yeast</i> 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	<i>Genes Dev.</i> 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="SSIYNGISTSGLDLNNGTIADMRLQLGIVESYKLKRAVVSSASEAA
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="MNRWVEKWLRVYLKYINLIFYRNVYPPQSFDYTTYQSFNLPQF
VPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKNDLCIEKYVLDSELQHVDKD
DQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNRRVD
SLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEKLISG
DDKILNGVYSQYEEGESIFGSLF"
```

# Les flatfiles en détails (6/8)

23

```
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="MNRWVEKWLRVYLKYINLILFYRNVYPPQSFDYTTYQSFNLPQ
FVPINRHPALIDYIEELLILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
KDDQIITEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
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 acagtttaggt atcgctcgaga gttacaagct aaaacgagca gtagtcagct    120
ctgcacatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa     180
gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaaccg     240
ccacacgtc attattataa ttagaaaacag aacgcaaaaa ttatccacta tataattcaa     300
agacgcgaaa aaaaaagaac aacgcgtcat agaacttttgc gcaattcgcg tcacaaataa     360
attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat   420
aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga     480
gagtgcgcct cctttgtcga gtaattttca cttttcatat gagaacttat ttcttattc      540
```

→ GenBank

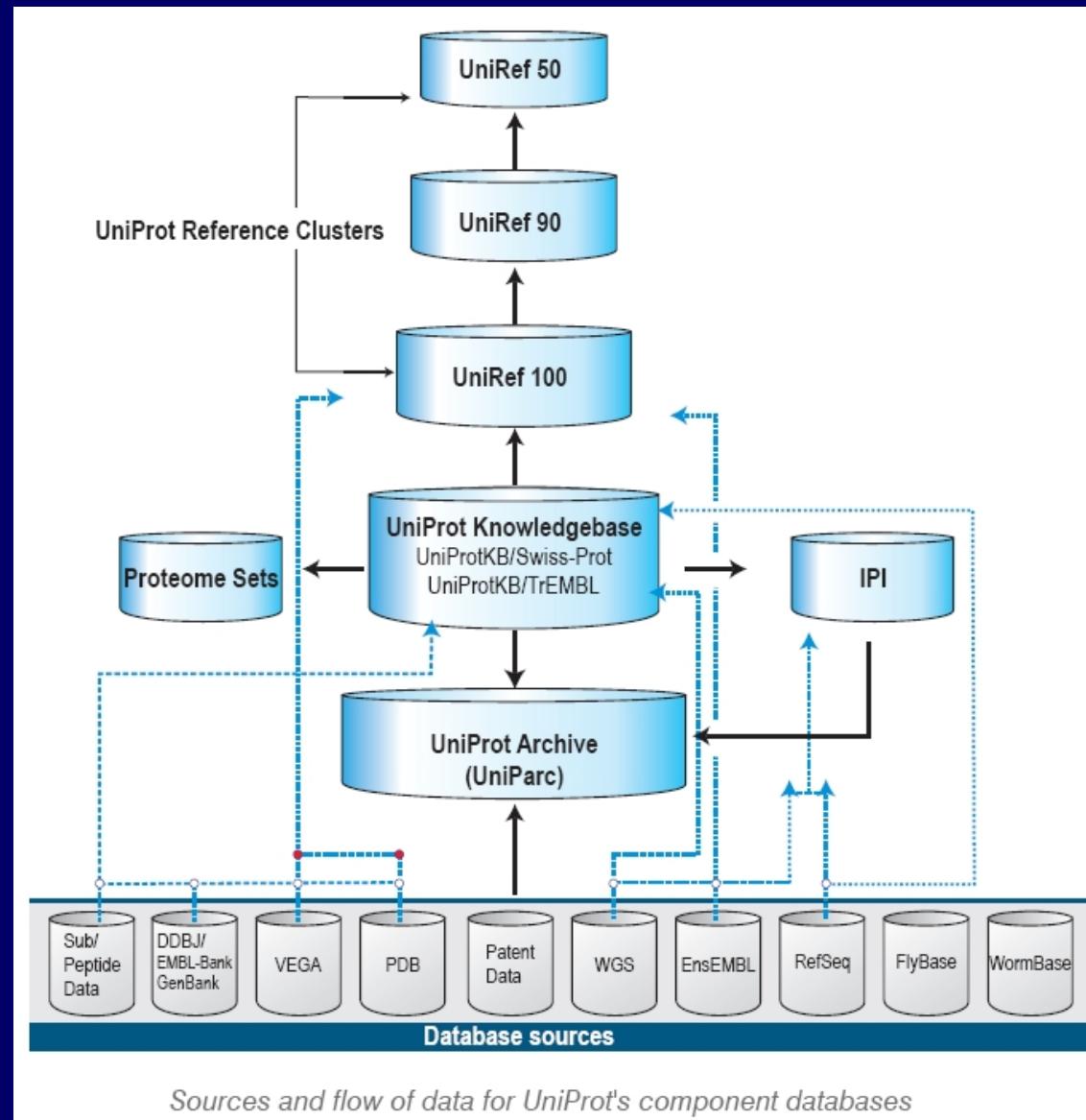
```
ORIGIN
  1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
  61 ccgacatgag acagtttaggt atcgctcgaga gttacaagct aaaacgagca gtagtcagct
 121 ctgcacatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
 181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaaccg
 241 ccacacgtc attattataa ttagaaaacag aacgcaaaaa ttatccacta tataattcaa
 301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttgc gcaattcgcg tcacaaataa
 361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
 421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
 481 gagtgcgcct cctttgtcga gtaattttca cttttcatat gagaacttat ttcttattc
 541 ttactctca catcctgttag tgattgacac tgcaacagcc accatcacta gaagaacaga
```

- 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 ?

- Introduction
- Banques de données de séquences nucléiques
- Banques de données de séquences protéiques
- Encore des banques de données
- Interrogation des banques de données
- Formats de fichiers en bioinformatique
- Références

- 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 ~ *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 seq. nucléiques et protéiques ; numéro d'accession particulier format 2 + 6, ex : NT\_123456
- UniProt : assemblage de SWISSPROT, TrEMBL et PIR



- 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

31

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	<b>Name: CCT2</b> Synonyms: BIN3, TCP2 OrderedLocusNames: YIL142W
From	<a href="#">Saccharomyces cerevisiae (Baker's yeast) [TaxID: 4932]</a>
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 32

UniProt > UniProt KB

Downloads · Contact · Help

**Search In** **Query**  
Protein Knowledgebase (UniProtKB)

★ Reviewed, UniProtKB/Swiss-Prot **P39076** (TCPB\_YEAST)  
Last modified September 11, 2007. Version 65. [History...](#)

[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](#)

[Hide](#) | [Top](#)

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

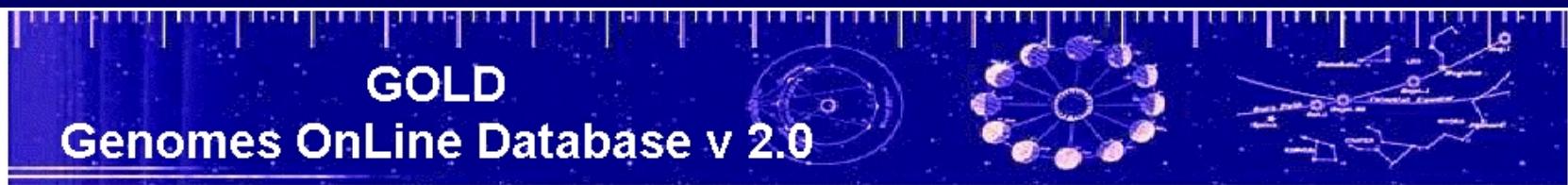
[General annotation \(Comments\)](#) [Hide](#) | [Top](#)

Function	Molecular chaperone; assist the folding of proteins upon ATP hydrolysis. Known to play a role, <i>in vitro</i> , 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.

- Introduction
- Banques de données de séquences nucléiques
- Banques de données de séquences protéiques
- Encore des banques de données
- Interrogation des banques de données
- Formats de fichiers en bioinformatique
- Références

- 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 ⇒ 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), ...)

- 1996 1re séquence complète d'un génome eucaryote *Saccharomyces cerevisiae*, chromosomes entre 270 et 1500 Kb ; limite d'une entrée dans GenBank à cette époque : 350 Kb  
⇒ 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**  
**Genomes OnLine Database v 2.0**

Contact: <a href="#">Genomesonline</a>	Last Update: <b>September 22, 2007</b>	Location <a href="#">www.genomesonline.org</a>
<b>655</b> Published Complete Genomes	<b>Search GOLD:</b> 2934 genome projects	<b>108</b> Metagenomes
<b>59</b> Archaeal Ongoing Genomes	<b>1337</b> Bacterial Ongoing Genomes	<b>775</b> Eukaryotic Ongoing Genomes

<http://www.genomesonline.org>

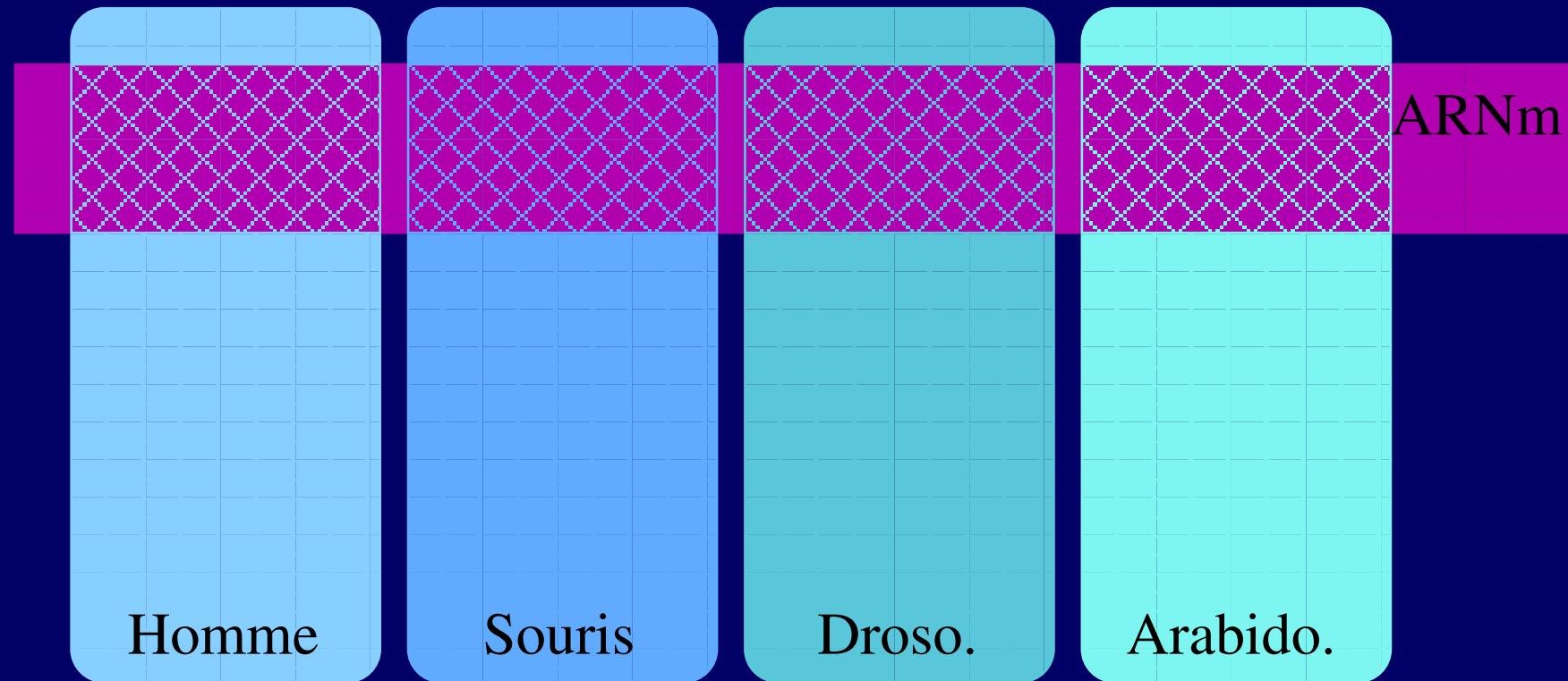
- Regroupement des protéines ayant des fonctions identiques ou proches
- Construites souvent par comparaison de toutes les protéines entre elles puis constitution de groupes
- Nombreuses banques :
  - HomoloGene, COG (NCBI)
  - KEGG SSDB (Sequence Similarity DataBase)
  - Clustr (EBI)
- Autres banques basées sur la structure 3D des protéines
  - SSF, SCOP, CATH, ...

- 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, ...

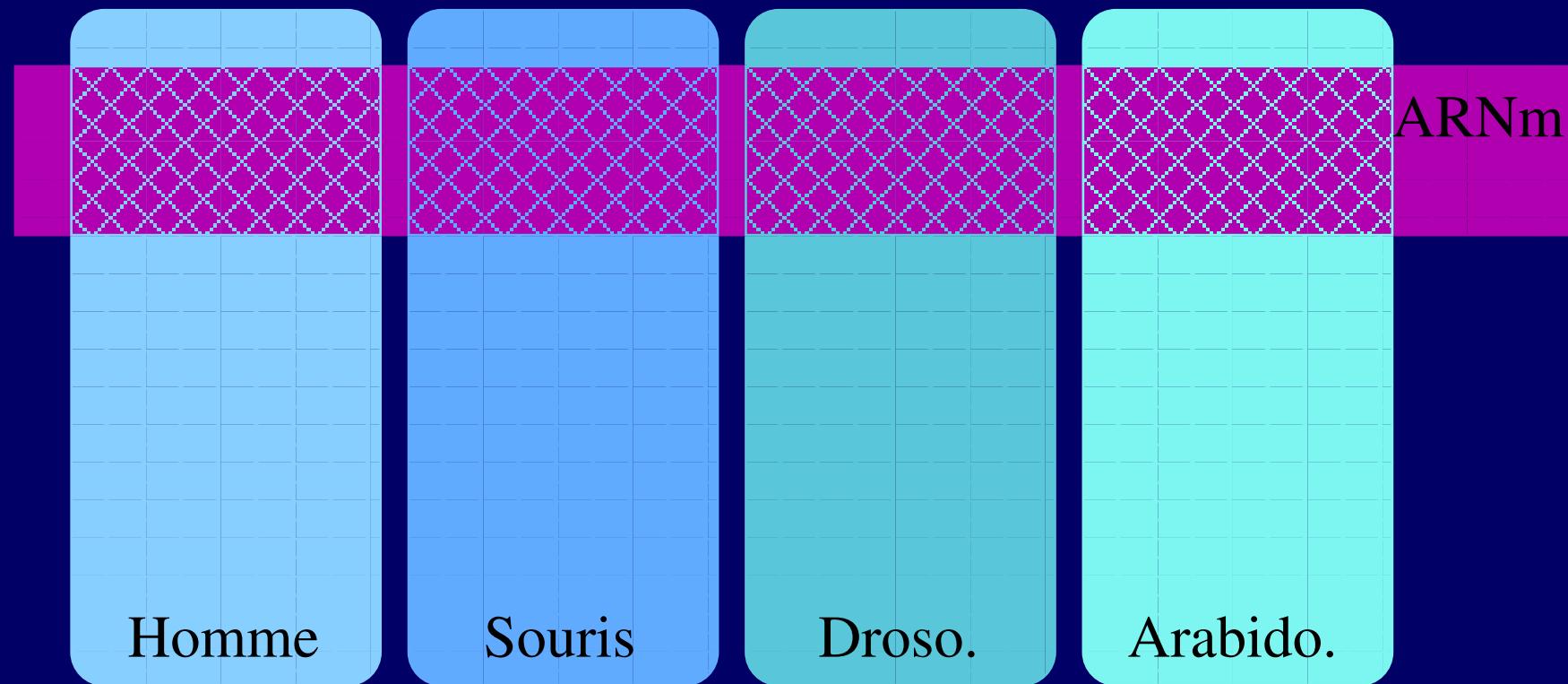
- Introduction
- Banques de données de séquences nucléiques
- Banques de données de séquences protéiques
- Encore des banques de données
- Interrogation des banques de données
- Formats de fichiers en bioinformatique
- Références

- 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)

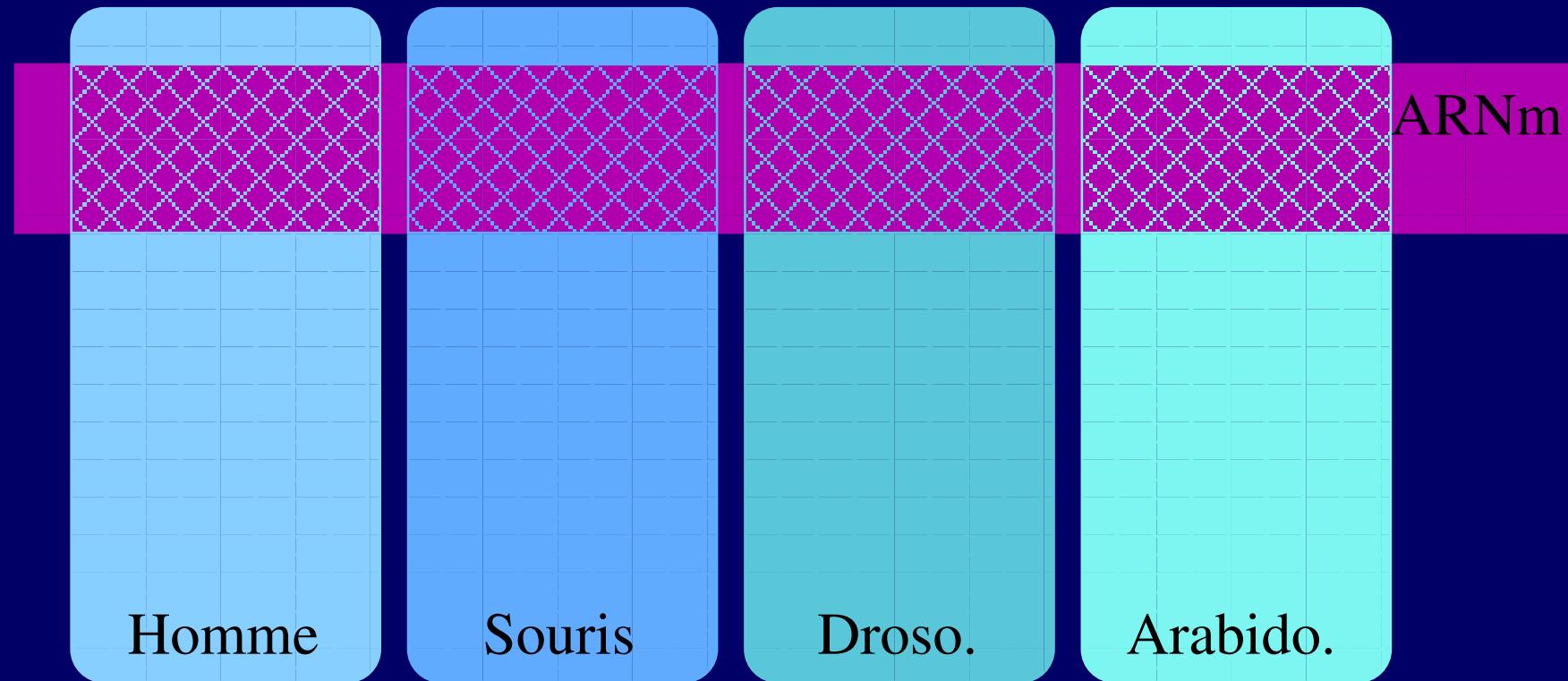
- But
  - Obtenir des informations nouvelles et pertinentes
  - Aide à la mise au point d'expérience
  - Validation des résultats d'une expérience
- Contraintes pour un système d'interrogation
  - Obtention de données pertinentes (pas trop de résultats mais tous ceux relatifs à notre problématique)
  - Simplicité d'utilisation (syntaxe d'interrogation intuitive)
  - Réponse rapide
  - Possibilité d'analyse des résultats (couplage à des outils)



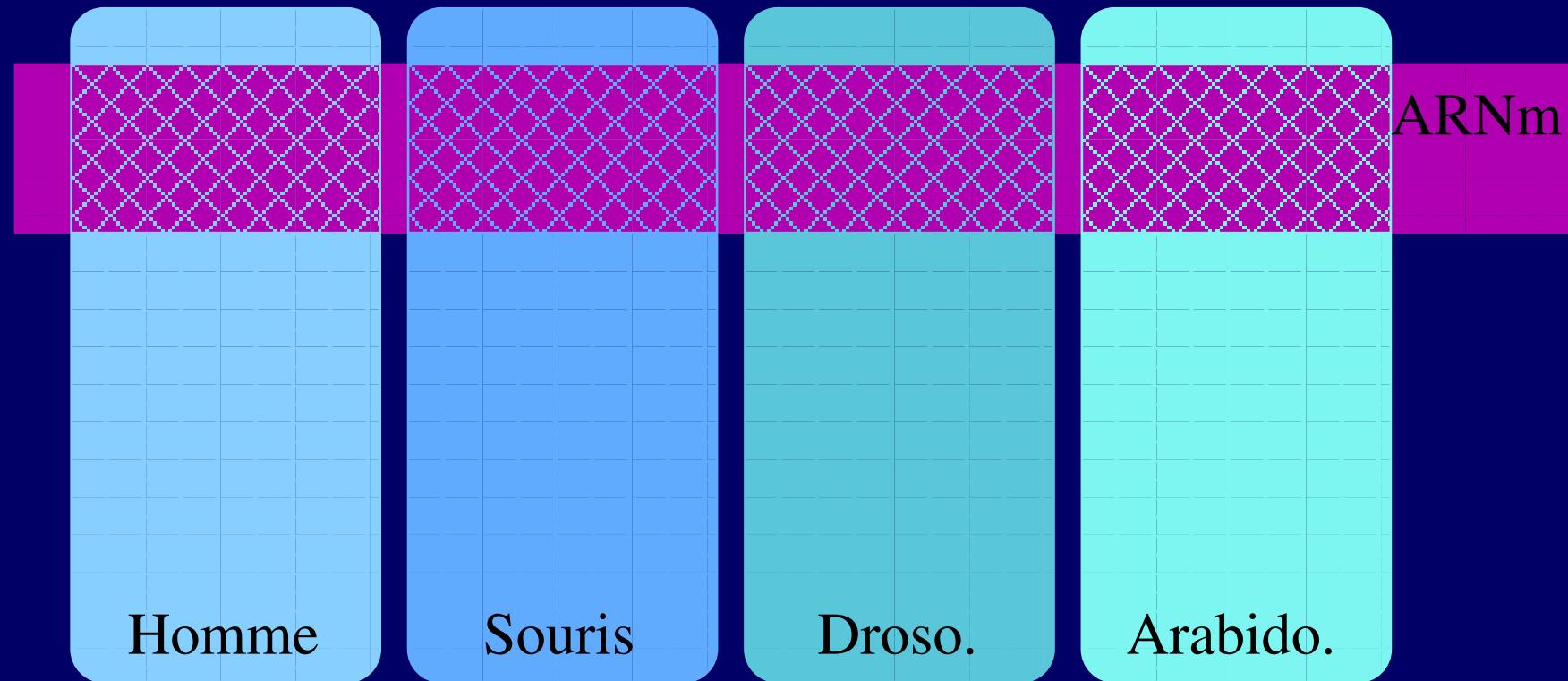
- Toutes séquences contenant 'Homme' :
- Séquences contenant 'Homme' ou 'Souris' :
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



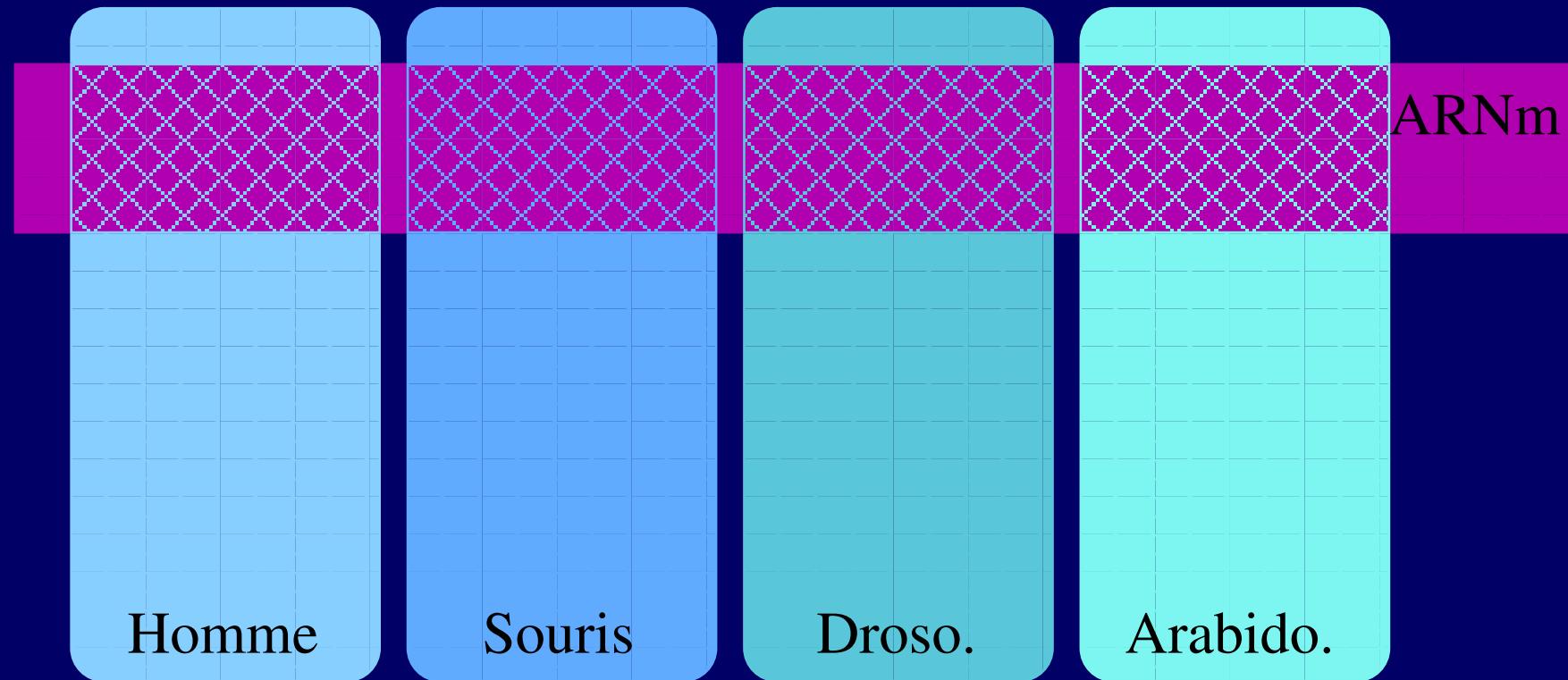
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' :
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



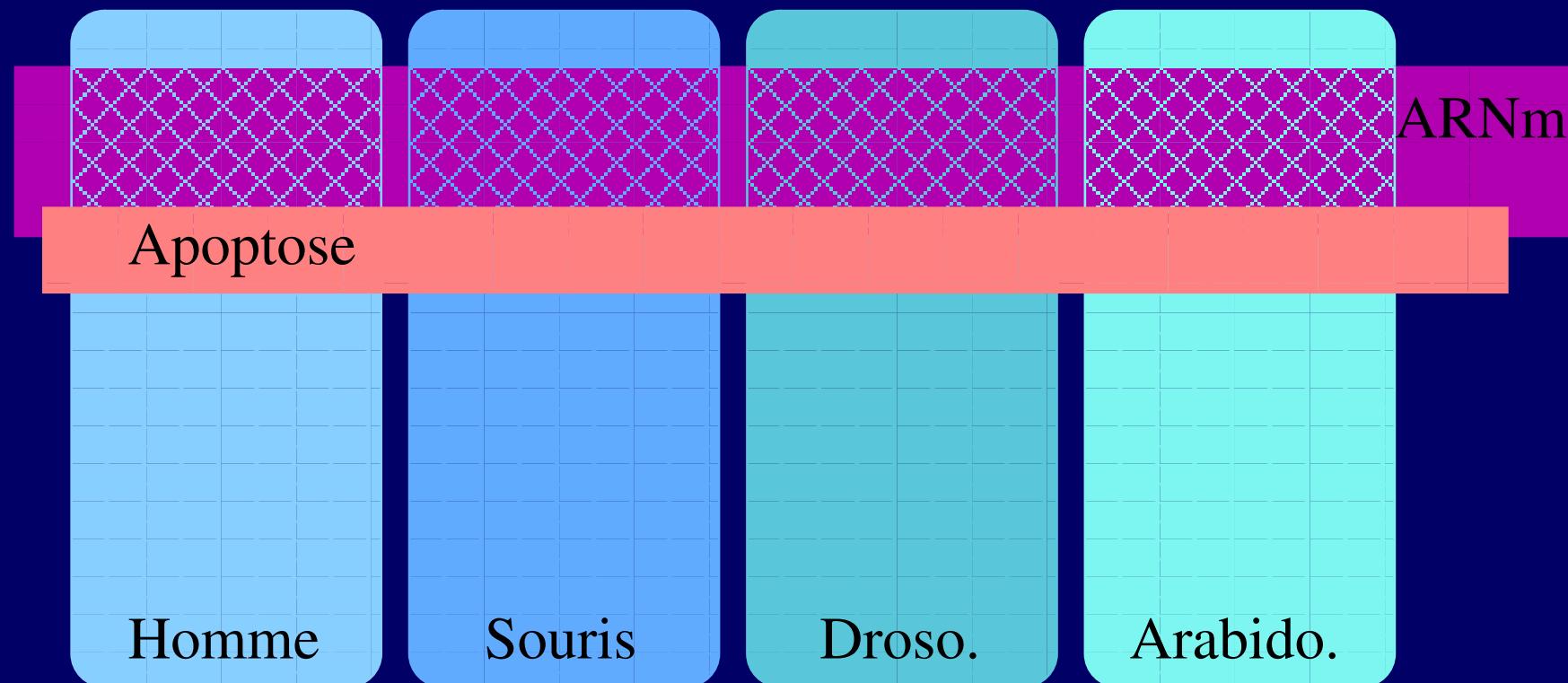
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile :
- Tous les ARNm sauf ceux d'arabidopsis :



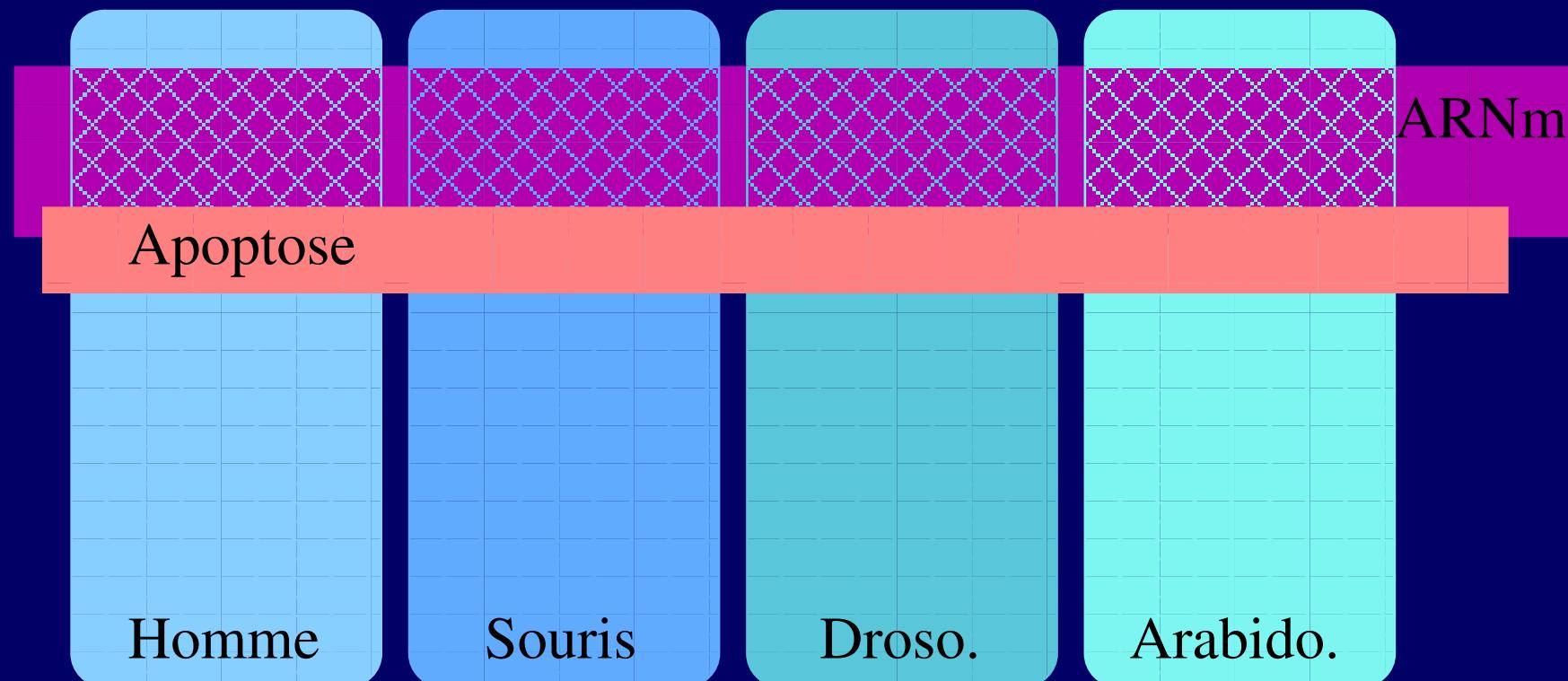
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile : Droso. ET ARNm
- Tous les ARNm sauf ceux d'arabidopsis :



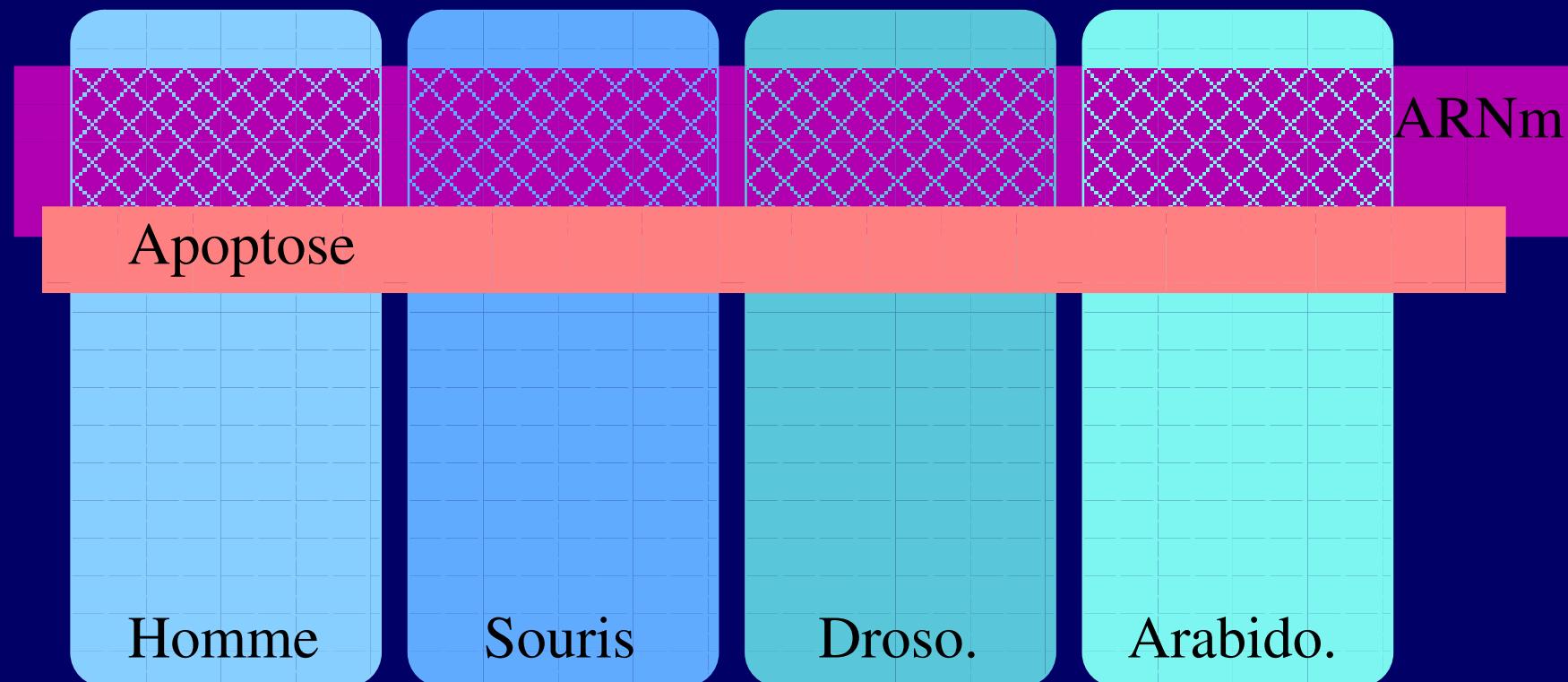
- Toutes séquences contenant 'Homme' : Homme
- Séquences contenant 'Homme' ou 'Souris' : Homme OU Souris
- Tous les ARNm de la drosophile : Droso. ET ARNm
- Tous les ARNm sauf ceux d'arabidopsis : ARNm NON Arabido.



- 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

- 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 cross-database search - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils Aide

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

NCBI Entrez, The Life Sciences Search Engine

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases  GO CLEAR Help

Welcome to the Entrez cross-database search page

 <b>PubMed:</b> biomedical literature citations and abstracts	 <b>Books:</b> online books
 <b>PubMed Central:</b> free, full text journal articles	 <b>OMIM:</b> online Mendelian Inheritance in Man
 <b>Site Search:</b> NCBI web and FTP sites	 <b>OMIA:</b> online Mendelian Inheritance in Animals

 <b>Nucleotide:</b> sequence database (includes GenBank)	 <b>UniGene:</b> gene-oriented clusters of transcript sequences
 <b>Protein:</b> sequence database	 <b>CDD:</b> conserved protein domain database
 <b>Genome:</b> whole genome sequences	 <b>3D Domains:</b> domains from Entrez Structure
 <b>Structure:</b> three-dimensional macromolecular structures	 <b>UniSTS:</b> markers and mapping data
 <b>Taxonomy:</b> organisms in GenBank	 <b>PopSet:</b> population study data sets
 <b>SNP:</b> single nucleotide polymorphism	 <b>GEO Profiles:</b> expression and molecular abundance profiles
 <b>Gene:</b> gene-centered information	 <b>GEO DataSets:</b> experimental sets of GEO data
 <b>HomoloGene:</b> eukaryotic homology groups	 <b>Cancer Chromosomes:</b> cytogenetic databases
 <b>PubChem Compound:</b> unique small molecule chemical structures	 <b>PubChem BioAssay:</b> bioactivity screens of chemical substances
 <b>PubChem Substance:</b> deposited chemical substance records	 <b>GENSAT:</b> gene expression atlas of mouse central nervous system
 <b>Genome Project:</b> genome project information	 <b>Probe:</b> sequence-specific reagents

 <b>Journals:</b> detailed information about the journals indexed in PubMed and other Entrez databases	 <b>MeSH:</b> detailed information about NLM's controlled vocabulary
 <b>NLM Catalog:</b> catalog of books, journals, and audiovisuals in the NLM collections	

Enter terms and click 'GO' to run the search against ALL the databases, OR  
Click Database Name or Icon to go directly to the Search Page for that database, OR  
Click Question Mark for a short explanation of that database.

Counts in XML | Entrez Utilities | Disclaimer | Privacy statement | Accessibility |

# Entrez, résultats d'une recherche

46

Entrez Nucleotide - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils Aide

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucleotide

NCBI All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for homo sapiens [orgn] Go Clear Save Search

Limits Preview/Index History Clipboard Details

All: 11532479 bacteria: 0 mRNA: 8160382 RefSeq: 48517

Show only records from: CoreNucleotide (2665510), EST (7895578), GSS (971391). [What's this?]

Items 1 - 20 of 11532479 Page | 1 of 576624 Next

1: NM\_006387 Reports Homo sapiens calcium homeostasis endoplasmic reticulum protein (CHERP), mRNA gi|119226259|ref|NM\_006387.5|[119226259] Links

2: NM\_032207 Reports Homo sapiens chromosome 19 open reading frame 44 (C19orf44), mRNA gi|119226258|ref|NM\_032207.2|[119226258] Links

3: NM\_153688 Reports Homo sapiens zinc finger protein 1 homolog (mouse) (ZFP1), mRNA gi|119226228|ref|NM\_153688.2|[119226228] Links

4: NM\_001079691 Reports Homo sapiens hypothetical gene CG018 (CG018), transcript variant 2, mRNA gi|119226226|ref|NM\_001079691.1|[119226226] Links

5: NM\_199050 Reports Homo sapiens chromosome 21 open reading frame 25 (C21orf25), transcript variant 2, mRNA gi|119226225|ref|NM\_199050.2|[119226225] Links

6: NM\_052818 Reports Homo sapiens hypothetical gene CG018 (CG018), transcript variant 1, mRNA gi|119226223|ref|NM\_052818.2|[119226223] Links

7: NM\_015500 Reports Homo sapiens chromosome 21 open reading frame 25 (C21orf25), transcript variant 1, mRNA gi|119226221|ref|NM\_015500.1|[119226221] Links

8: NM\_001079692 Reports Homo sapiens MS4A13 protein (NYD-SP21), transcript variant 2, mRNA gi|119226219|ref|NM\_001079692.1|[119226219] Links

9: NM\_032597 Reports Homo sapiens MS4A13 protein (NYD-SP21), transcript variant 1, mRNA

- 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

Select Databanks to Search - Mozilla Firefox

Echier Édition Affichage Historique Marque-pages Outils Aide

EB-eye Search NEW http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+top

EMBL-EBI All Databases Enter Text Here Go Reset Advanced Search

Databases Tools Groups Training Industry About Us Help Site Index

Quick Search Library Page Query Form Tools Results Projects Views Databanks HELP Job Status

**Search Options**

1. Select the databanks you want to search

2. Enter your search terms in the Quick Search box, or choose a query form from below

**Standard Query Form**

**Extended Query Form**

You can browse through all the entries in any databases. First, select the databanks you want to browse, then click:

**Browse Entries**

**Tips**

- ▶ bookmark this link to return to your project
- ▶ Linking to SRS?  
- Please read our [Linking to SRS](#) guide for important information regarding linking to our SRS server.

**BookMarkLets**

[About BookMarkLets](#)

- Protein Seq
- DNA/RNA Seq
- Structures

**Available Databanks**

+ Expand all - Collapse all Show databanks tooltips:

**Literature, Bibliography and Reference Databases**

(all)  TAXONOMY  GENETICCODE  OMIM  MEDLINE  
 Patent Abstracts  Karyn's Genomes  
**Literature, Bibliography and Reference Databases - subsections**  
 (all)  MEDLINE (Updates)  MED2PUB  MEDLINE (Main Release 2006)

**Gene Dictionaries and Ontologies**

**Nucleotide sequence databases**

(all)  EMBL  Patent DNA  IMGT/LIGM-DB  IMGT/HLA  
 IPD-KIR  EMBL (Contig)  EMBL (Contigs expanded)  EMBL (Annotated Cons)  
 EMBL (Coding Sequences)  Genome Reviews  RefSeq Genome  LiveLists  
 EMBL ID/Accession Mapping  EMBL MGA  
**Nucleotide sequence databases - subsections**  
 (all)  EMBL (Updates)  EMBL (Release)  EMBL (Whole Genome Shotgun)  
 EMBL (Whole Genome Shotgun release)  EMBL (Whole Genome Shotgun updates)  EMBL (Contig release)  
 EMBL (Contig updates)  EMBL (Contigs expanded release)  EMBL (Contigs expanded updates)  
 EMBL (Annotated Cons release)  EMBL (Annotated Cons updates)  RefSeq Genome (Release)  
 RefSeq Genome (Updates)  EMBL (Whole Genome Shotgun Masters)

**Nucleotide related databases**

**UniProt Universal Protein Resource**

(all)  UniProtKB  UniProtKB/Swiss-Prot  UniProtKB/Trembl  UniRef100  UniRef90  
 UniRef50  UniParc

**Other protein sequence databases**

**Protein function, structure and interaction databases**

**Enzymes, reactions and metabolic pathway databases**

**Mutation and SNP databases**

**Biological Resources Catalogues (CABRI)**

**Mapping databases**

**Other databases**

**User owned databases**

**Application result databases**

**EMBOSS result databases**

**Eurofir Food data**

**EMBLCDs Grouped By**

Terms of Use | Feedback & Support | SRS Release 7.1.3.2 Copyright © 1997-2003 LION bioscience AG. All Rights Reserved.

- 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

50

Standard Query Form - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils Aide

EB-eye Search NEW http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz

EMBL-EBI All Databases Enter Text Here Go Reset Advanced Search

Databases Tools Groups Training Industry About Us Help Site Index

Quick Search Library Page Query Form Tools Results Projects Views Databanks HELP Job Status

Reset search EMBL

**Search Options**

Combine search terms with: & (AND) Use wildcards Get results of type: Entry

**Fields you can search**

In a single field, you can separate multiple values by &, |, !

Organism Name homo sapiens

AllText

AllText

AllText

**Result Display Options**

View results using: EMBLSeqSimpleView or Create a view Show 30 results per page

**Create a view**

Select the fields you want displayed in your view and choose the format

Choose 1 or more fields: ID Topology Molecule Data Class Division Sequence Length Accession Number

Display As: Table List Sequence Format: embl

Search

Tips

To do more advanced queries, use the [Extended Query Form](#).

Terms of Use | Feedback & Support | SRS Release 7.1.3.2 Copyright © 1997-2003 LION bioscience AG. All Rights Reserved.

# SRS, résultats d'une recherche

51

Query Results - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils Aide

EB-eye Search [NEW] http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz

All Databases Enter Text Here Go Reset Advanced Search

Databases Tools Groups Training Industry About Us Help Site Index

Quick Search Library Page Query Form Tools Results Projects Views Databanks HELP Job Status

Reset

Query "(([embl-AllText:homo\*] & [embl-AllText:sapiens\*]) | [embl-AllText:homo sapiens\*]) & [embl-AllText:tetraodon\*]" found 1508 entries next

Apply Options to:  
 selected results only  
 unselected results only

Result Options  
Launch analysis tool:  
BlastN Launch  
Show tools relevant to these results:  
Tools  
Link to related information:  
Link  
Save results:  
Save

Display Options  
View results using:  
EMBLSeqSimpleView  
Show 30 results per page  
Printer friendly view  
Apply Display Options

EMBL	Primary Accession (Links to SVA)	Accession List	Description	Sequence Length
<a href="#">EMBL:DQ322650</a>	<a href="#">DQ322650</a>	<a href="#">DQ322650</a>	Streptomyces sp. 44414 plasmid pRL2, complete sequence.	20252
<a href="#">EMBL:CK829240</a>	<a href="#">CK829240</a>	<a href="#">CK829240</a>	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
<a href="#">EMBL:CK829247</a>	<a href="#">CK829247</a>	<a href="#">CK829247</a>	Fr_Fwd_12D03_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_12D03 similar to gbi AH16170.1  Similar to aldolase A, fructose-bisphosphate - Homo sapiens. Score = 40.0 bits (92), Expect = 0.009, mRNA sequence.	348
<a href="#">EMBL:CK829327</a>	<a href="#">CK829327</a>	<a href="#">CK829327</a>	Fr_Fwd_10F12_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_10F12 similar to gbi AH11837.2  DNAJC7 protein - Homo sapiens. Score = 238 bits (608), Expect = 5e-62, mRNA sequence.	647
<a href="#">EMBL:CK829363</a>	<a href="#">CK829363</a>	<a href="#">CK829363</a>	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
<a href="#">EMBL:CK829398</a>	<a href="#">CK829398</a>	<a href="#">CK829398</a>	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 983016G03 - Homo sapiens. Score = 156 bits (394), Expect = 1e-37, mRNA sequence.	455
<a href="#">EMBL:CK829423</a>	<a href="#">CK829423</a>	<a href="#">CK829423</a>	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
<a href="#">EMBL:CK829449</a>	<a href="#">CK829449</a>	<a href="#">CK829449</a>	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 MGCG509 - Homo sapiens. Score = 74.3 bits (181), Expect = 2e-12, mRNA sequence.	703
<a href="#">EMBL:CK829487</a>	<a href="#">CK829487</a>	<a href="#">CK829487</a>	Fr_Fwd_08B02_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_08B02 similar to dbj BAAT6781.1  KIAA0937 protein - Homo sapiens. Score = 31.6 bits (70), Expect = 9.4, mRNA sequence.	588
<a href="#">EMBL:CK829545</a>	<a href="#">CK829545</a>	<a href="#">CK829545</a>	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
<a href="#">EMBL:CK829575</a>	<a href="#">CK829575</a>	<a href="#">CK829575</a>	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
<a href="#">EMBL:CK829656</a>	<a href="#">CK829656</a>	<a href="#">CK829656</a>	Fr_Fwd_05G12_T3 Forward subtracted cDNA library from fast skeletal muscle Takifugu rubripes cDNA clone Fr_Fwd_05G12 similar to dbj BAAT6781.1  KIAA0937 protein - Homo sapiens. Score = 31.6 bits (70), Expect = 7.7, mRNA sequence.	543
<a href="#">EMBL:CK829703</a>	<a href="#">CK829703</a>	<a href="#">CK829703</a>	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

- Introduction
- Banques de données de séquences nucléiques
- Banques de données de séquences protéiques
- Encore des banques de données
- Interrogation des banques de données
- Formats de fichiers en bioinformatique
- Références

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)

```
>emb1|U49845|U49845 Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and Rev7p (REV7) genes, comple  
gatcctccatataacaacggtatctccacctcaggttagatctcaacaacggaccattg  
ccgacatgagacagttaggatcgtagagttacaagactaaaacggcgcgttagtgc  
ctgcattgtgaagccgctgaagttactaagggtggataaacatcatccgtgcaagacaa  
gaaccgccaatagacaacatatgtAACATATTAGGATAACCTCGAAAATAATAACCG  
CCACACTGTCAATTATAATTAGAAACAGAACGCAAAATTATCCACTATATAATTCAA  
AGACGCGAAAAAAAAGAACACGCGTCATAGAACCTTGGCAATTCGCGTCACAAATAA  
ATTTGGCAACTTATGTTCTCTTGAGCAGTACTCGAGGCCCTGTCTCAAGAATGTAAT  
AATAACCCATCGTAGGTATGGTTAAAGATAGCATCTCCACACCTCAAAGCTCCTGCCGA
```

## 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 EMBOS) : 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		GCTGCCCTCG CGGTCACTCT GTTCCCGCCC TCCTGGAGCT TCAAGACAAG
C3		GCTGCCCTCG CGGTCACTCT GTTCCCACCC TCCTGGAGCT TCAAGACAAG
C4		GAGACACCTT CATCTCCTCT GACCCAGAG GCAGGGAGCT CCAAGACAAG
C5		GCCACCCCT TGTTCACTCT GTTC---CCC TCCTGGAGCT CCAAGACAAG
C6		GCTGCCCTCAT CGGTCACTCT GTTCCCGCCC TCCTGGAGCT TCAAGACAAG
C7		GCTGCCCTCG 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" } , ...
```

- 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>R0D</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

- 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>

- Introduction
- Banques de données de séquences nucléiques
- Banques de données de séquences protéiques
- Encore des banques de données
- Interrogation des banques de données
- Formats de fichiers en bioinformatique
- 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/>

