Bases de Données en Génomique

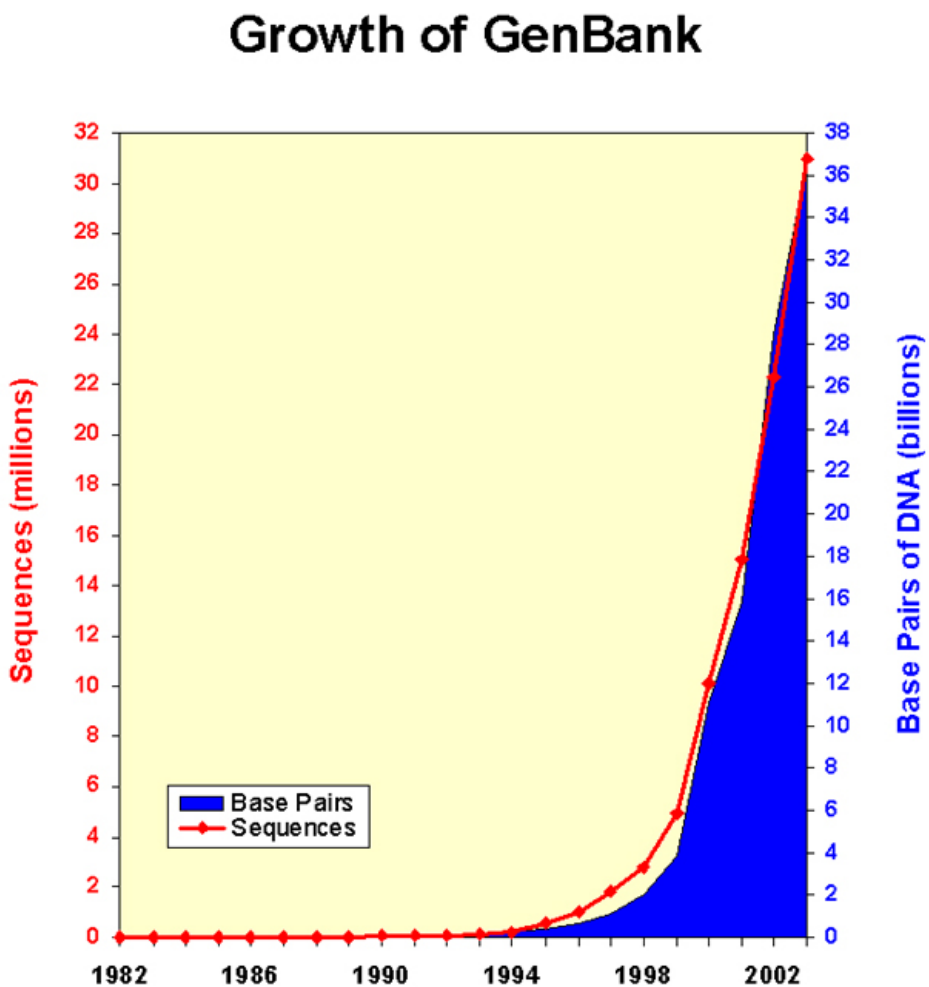
Daniel Gautheret, 2004
ESIL, Université de la Méditerranée

V.2004.1 http://www.esil.univ-mrs.fr/~dgaut/Cours

Genbank: La banque d'ADN du NIH

Etat au 2-2004

- ★ $38x10^9$ bases
- ★ $32x10^6$ séquences
- ★ Genbank double environ tous les 14 mois depuis ses débuts en 1982.
- ★ Nouvelle version tous les 2 mois



Divisions de Genbank

- **ESTs** (Expressed sequence tags): Principale division ed Genbank. 18 10^6 sequences, 580 organismes différents
- **GSS** (Genome Sequence Survey): résultats de séquençages aléatoire de BAC, dans le cadre de projets Génome
- **HTGS** (High Throughput Genomic Sequences): séquences génomiques en cours d'assemblage. Une fois assemblées, les séquences passent dans les divisions « organisme ».
- Bacteries (BCT), virus (VRL), primates (PRI), rongeurs (ROD) etc: divisions « organismes ».
- 17 divisions en tout.

Nb entrées	Nb. bases	Espèce
1355113	854232260	Homo sapiens
378892	179249409	Mus musculus
76471	139699685	Caenorhabditis elegans
66177	69663817	Arabidopsis thaliana
48963	53428355	Drosophila melanogaster
10571	28658828	Saccharomyces cerevisiae
39568	25816686	Rattus norvegicus
4923	17859484	Escherichia coli
32221	16490243	Fugu rubripes
31480	13072925	Oryza sativa
28406	11746328	Rattus sp.
9540	10912762	Schizosaccharomyces pombe
24125	10712174	Human immunodeficiency virus type 1
1086	9893044	Bacillus subtilis
15370	5794059	Brugia malayi
661	5701954	Mycobacterium tuberculosis
4852	5585160	Gallus gallus
4680	5400457	Plasmodium falciparum
5063	4559072	Bos taurus
10845	4409926	Toxoplasma gondii

Organismes dans Genbank (en 2002)

Enregistrement Genbank

- Chaque enregistrement se voit attribuer un numéro d'accession, stable et unique, et chaque séquence un numéro GI.
- Quand un changement est effectué dans un enregistrement Genbank, le num. d'accession reste, le GI change.

Enregistrement Genbank avec annotation

```
LOCUS       L10986                 47233 bp    DNA     linear   INV 21-SEP-2004
DEFINITION  Caenorhabditis elegans cosmid F10E9, complete sequence.
ACCESSION   L10986
VERSION     L10986.2  GI:38638818
KEYWORDS    HTG.
SOURCE      Caenorhabditis elegans
  ORGANISM  Caenorhabditis elegans
            Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida;
            Rhabditoidea; Rhabditidae; Peloderinae; Caenorhabditis.
REFERENCE   1  (bases 1 to 47233)
  AUTHORS   .
  CONSRTM   WormBase Consortium
  TITLE     Genome sequence of the nematode C. elegans: a platform f
            investigating biology. The C. elegans Sequencing Consort
  JOURNAL   Science 282 (5396), 2012-2018 (1998)
  MEDLINE   99069613
   PUBMED   9851916
FEATURES             Location/Qualifiers
     source          1..47233
                     /organism="Caenorhabditis elegans"
                     /mol_type="genomic DNA"
                     /strain="Bristol N2"
                     /db_xref="taxon:6239"
                     /chromosome="III"
                     /clone="F10E9"
     gene            265..26728
                     /gene="mig-10"
                     /locus_tag="F10E9.6"
     CDS             join(265..338,3266..3515,15194..15317,21507..21
                     21727..21887,23171..23335,24302..24472,24524..24608,
                     25012..25827,26284..26430,26478..26728)
                     /gene="mig-10"
 /translation="MDSCEEECDLEVDSDEEDQLFGEKCISLLSSLLPLSSSTLLSNA
                     INLELDEVERPPPLLNVLEEQQFPKVCANIEEENELEADTEEDIAETADDEESKDPVE
                     KTENFEPSVTMDTYDFPDPYPVQIRARPQVPPKPPIDTVRYSMNNIKESADWQLDELL
                     EELEALETQLNSSNGGDQLLLGVSGIPASSSRENVKSISTLPPPPPALSYHQTPQQPQ
. . .
                     QVYTGIGWEKKYKSPTPWCISIKLTALQMKRSQFIKYICAEDEMTFKKWLVALRIAKN
                     GAELLENYERACQIRRETLGPASSMSAASSSTAISEVPHSLSHHQRTPSVASSIQLSS
                     HMMNNPTHPLSVNVRNQSPASFSVNSCQQSHPSRTSAKLEIQYDEQPTGTIKRAPLDV
                     LRRVSRASTSSPTIPQEESDSDEEFPAPPPVASVMRMPPPVTPPKPCTPLTSKKAPPP
                     PPKRSDTTKLQSASPMAPAKNDLEAALARRREKMATMEC"
. . .
```

```
BASE COUNT     2598 a   2024 c   1888 g   2449 t
ORIGIN
        1 ttctaaaagt cgaaaaacga gcaatttttg atgctagatt ttttgatttg acgaattttt
       61 tcagtttttt ttctttaaaa aaggtttttg acccctaaa gttttccttt cccttccaat
      121 ttttccttc tttcttatac gacttctcaa gtttcaactc taaaacaaag ctacatgtac
      181 atttccggta aactttgtgt ctcagaagat ccattttctt tttgttacat ttattcaaga
      241 ttgaattcca aaatttcagc caatatggac agttgcgaag aggaatgcga tctggaagtt
      301 gacagtgacg aagaagatca actttttggt gaaaagtggt gagttcttat tgtggtaacc
      361 aaagaaatgt cagtggtccg taaacacttg actcccaaat ggtttctcgt aattacctta
      421 tgcacacttt tcaagtgttt gccgtttgat cttagccaat ttgaaacgtt tagatgttaa
      481 atggaaaatg ggtaaagttt tttattttat agaaaaaagg tttggaaaaa aatcgagtca
      541 ctgaatagtt tgaagaacgg aaaaataaaa ctttccaaaa atcataaaac atttagtgtt
      601 tcgaaaatta tagtgttttt tttgttggta tgttttgaca aaagctaaac catctttatt
      661 gtagttttgt aaaatgttca caaagatgcg ttttttttc aaatttggca ggctatcttt
      721 acattcacat ttggataatt caaatttttc ttatcgctaa caaattttcc tatttttcca
      781 attattcgtt tttataaagc tttggtagta tgttgtgtct atctttagtg gtcatcagtt
. . .
//
```

Les banques de gènes

Nom	Type	Organisme	Description	Nb. Enrgst.
Refseq	gene+ mRNA+ prot	H, M, R	Itération à partir de Blast seed mRNA/EST vs contigs de Genbank. +Annotation manuelle: publications, UTR prolongées, etc.	11 405(H) 5 749 (M)
HGI (Human Gene Index)	mRNA	H, M, R, D, Arabid., Fugu, riz, etc.	Genbank mRNA+EST contigués. Toutes les solutions alternatives sont conservées.	388 000
Ensembl	gènes + transcrits	H + eucaryotes?	Genscan sur contig, puis Blast vs prot, mRNA, EST, PFAM	35 500 genes, 44 860 transcrits
Unigene	clusters	H, M, R, bovin, zebrafish, blé, riz, mais, orge	Clustering itératif à partir de mRNA+CDS génomique+ESTs. Pas de contigage.	89 371 cl

★Unigene: banque d'ESTs classifiés ("clusterisés"). Dans chaque cluster Unigene sont regroupés des EST ayant une similarité de séquence significative. On peut donc trouver des transcripts différents et des artefacts (chimères, etc.). Unigene ne propose pas de mRNA reconstruits (contigs) à partir des séquences d'un cluster.

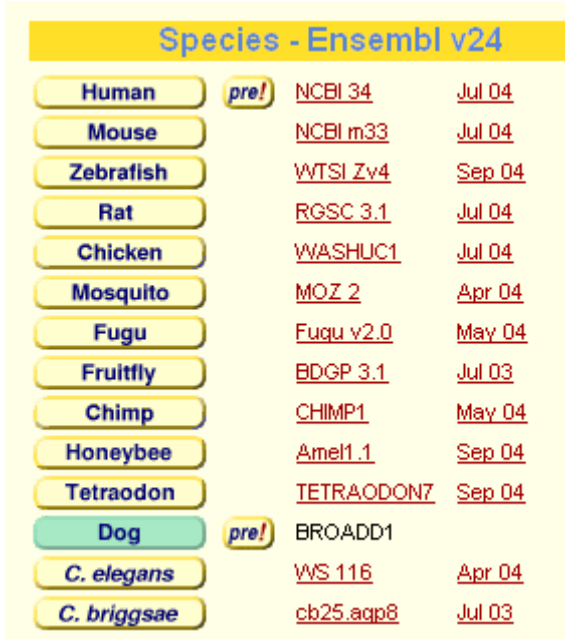
★TIGR Human Gene Index (HGI). Ici encore on a clusterisé les EST, mais HGI est une banque de "contigs", c.a.d. de séquences de mRNA reconstruites à partir des EST d'un même cluster. Les clusters étant souvent hétérogènes, ils produisent souvent plusieurs contigs. Ces contigs doivent théoriquement correspondre à des mRNA alternatifs.

Autres banques de séquences

Nucléotidiques

- ★ gbEST / dbEST: Division EST dans Genbank
- ★ EMBL: Equivalent européen de Genbank. Format différent, contenu presque identique.
- ★ Banques spécialisées Certaines collections de séquences, bien que généralement présentes dans Genbank, sont beaucoup plus utiles lorsqu'elles sont rassemblées dans des banques spécialisées, par ex:
  - Récepteurs des lymphocytes T (Réarrangements de l'ADN)
  - Génomes HIV, etc.
- ★ NR nucléique (Non-redundant). Banque combinée: Genbank+refseq (20x10e9 nt / oct. 2002)

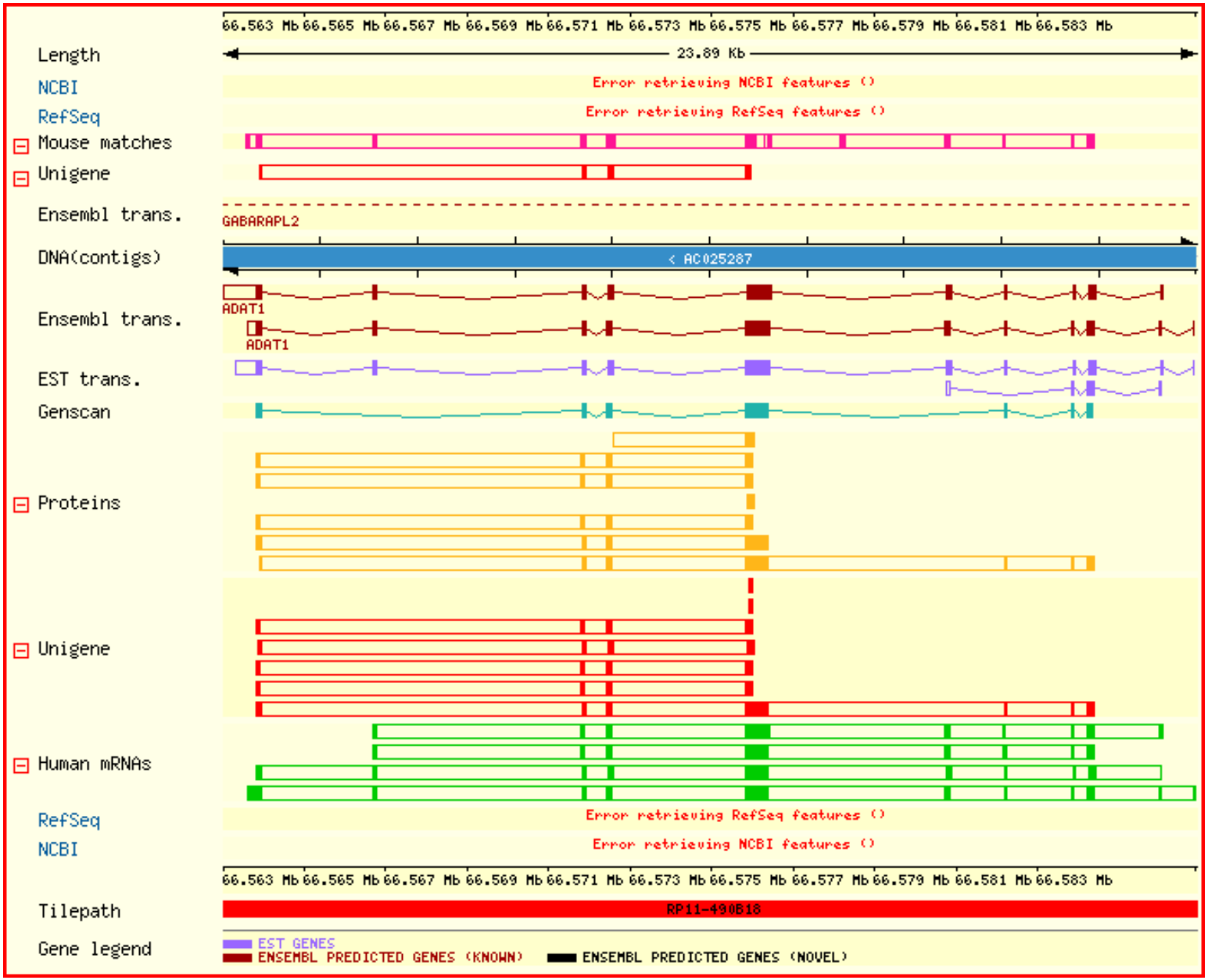
Ensembl (www.ensembl.org)



- ★ Plusieurs banques en une:
  - Peptides confirmés
  - Transcrits confirmés
  - peptides prédits
  - Transcrits prédits
  - Génome assemblé (golden path)

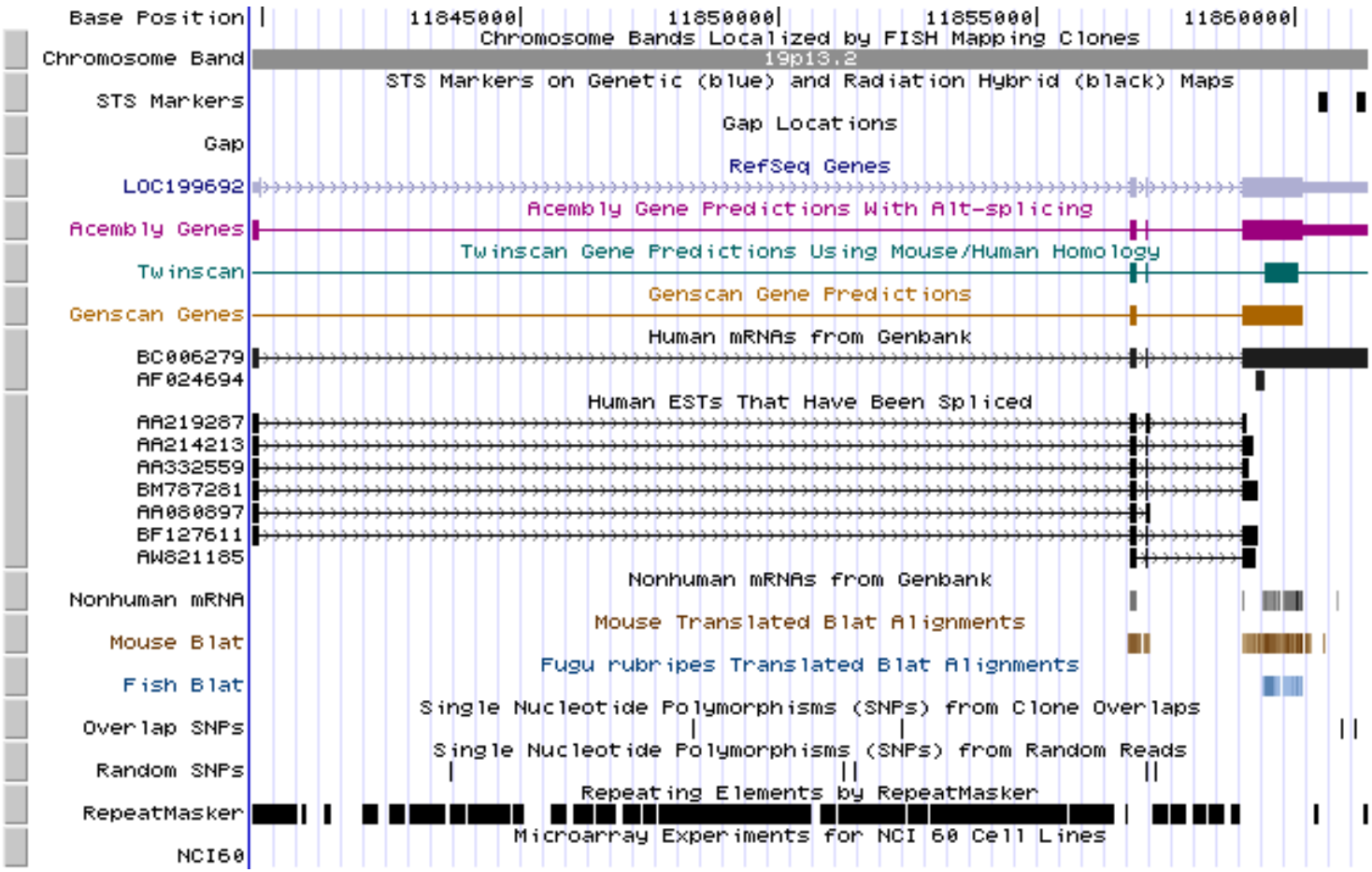
- ★ Méthode de prédiction (système Genewise): Genscan sur contig, puis Blast contre: protéines, mRNA, EST
- ★ Version Juillet 2001, humain: Confirmed genes: 21921; Predicted genes: 24636; Confirmed exons: 143479; Predicted exons: 770562; Transcripts: 23931; Contigs: 329154; Sequences: 29080; base pairs: 4318661441.

Ensembl: « contig view »



Banque génomique UCSC

- ★ http://www.genome.ucsc.edu/



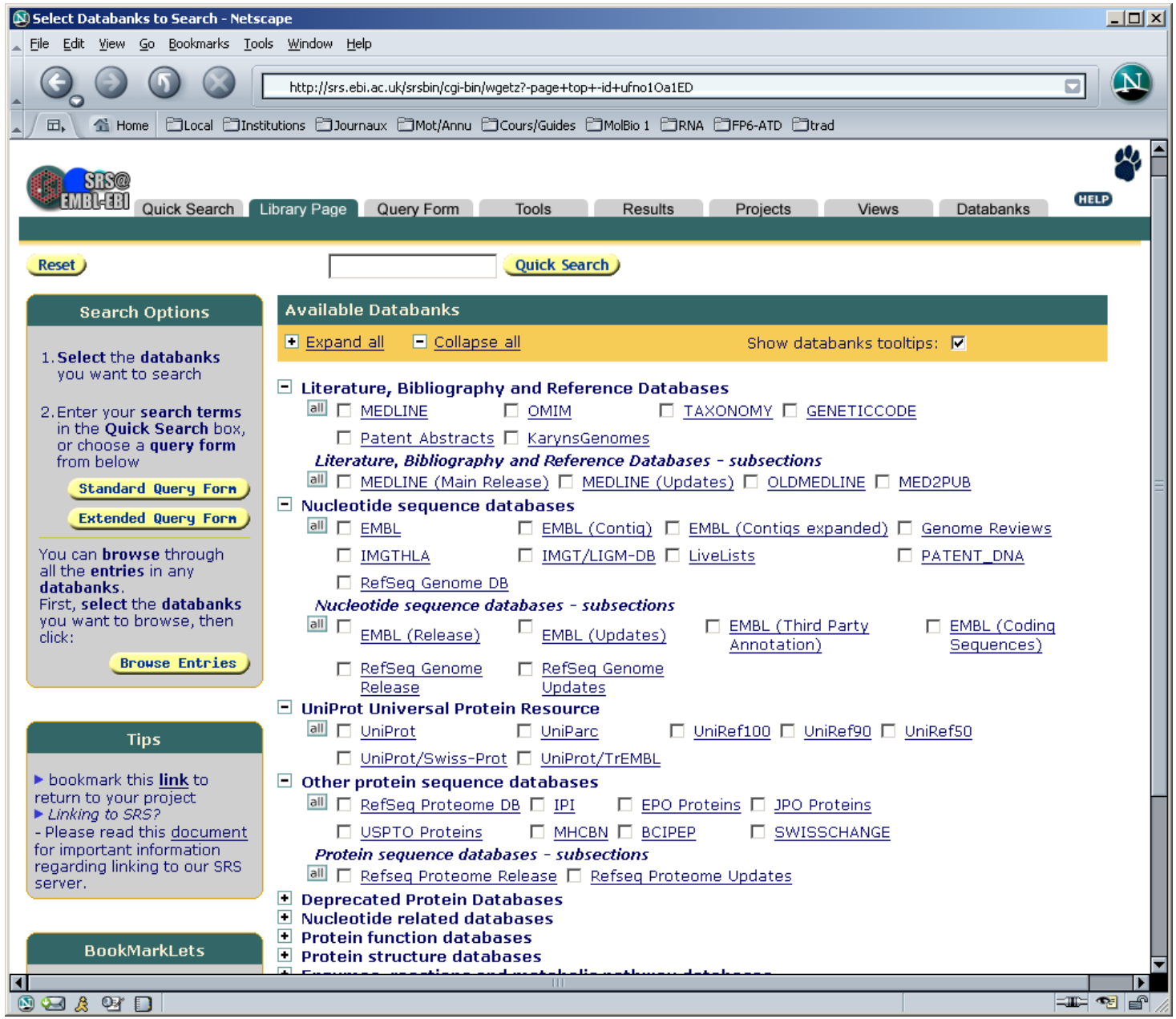
Banques protéiques

- ★ Swissprot. La mieux annotée des banques protéiques. Release 39 (2001): 101247 entrées, 37 135 523 aa. Attention: toutes les protéines connues n'y sont pas! Visiter le serveur
- ★ PIR (Protein Identification Resource), EMBL.
- ★ NR Protéique (Non-redundent): Banque protéique du NCBI = Traduction de tous les CDS de GenBank + PDB + SwissProt + PIR + PRF - redondances.
- ★ Banques spécialisées
  - Cazy (Carbohydrate Active Enzymes)
  - Etc.

SRS

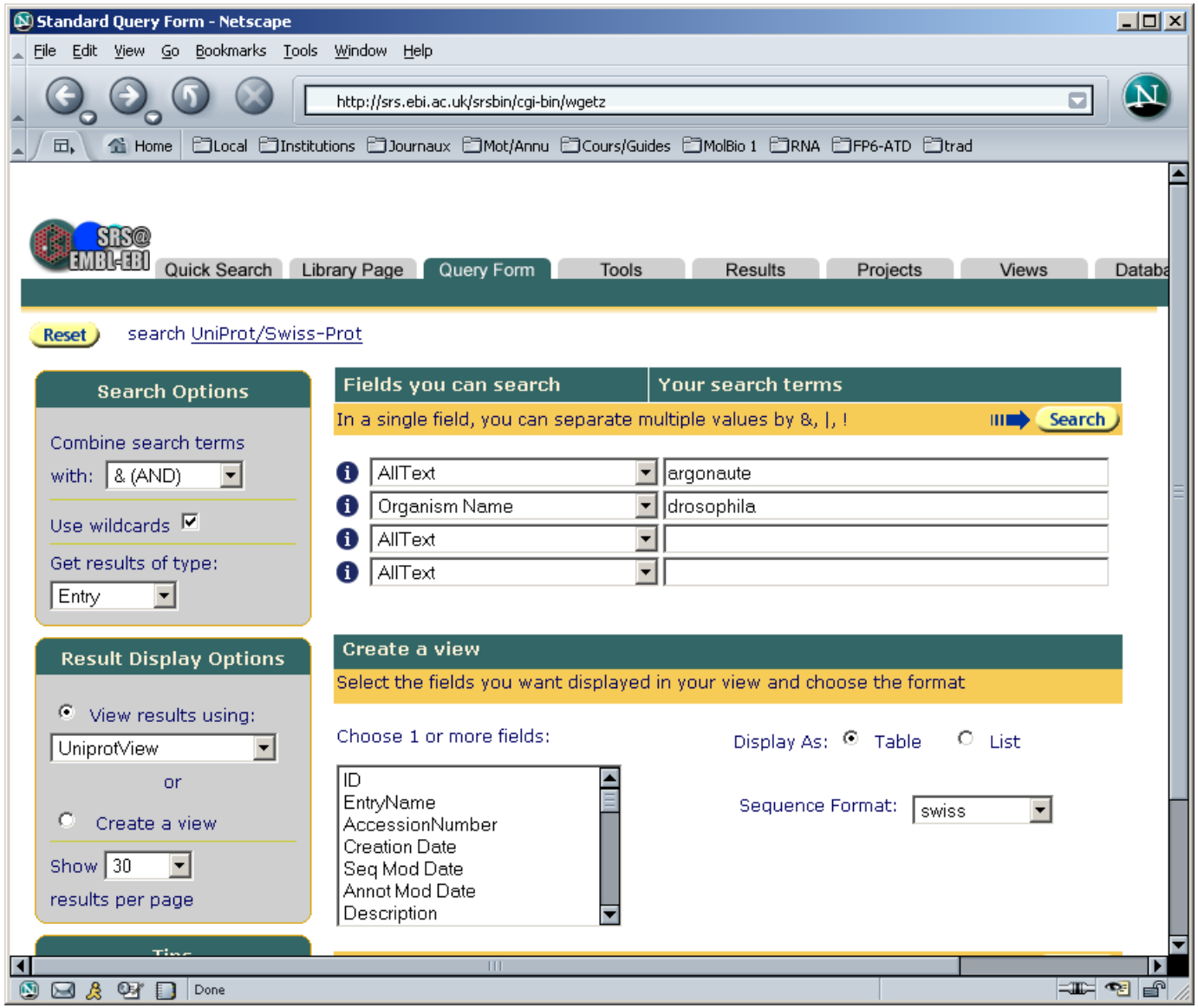
Sequence Retrieval System

Database selection page

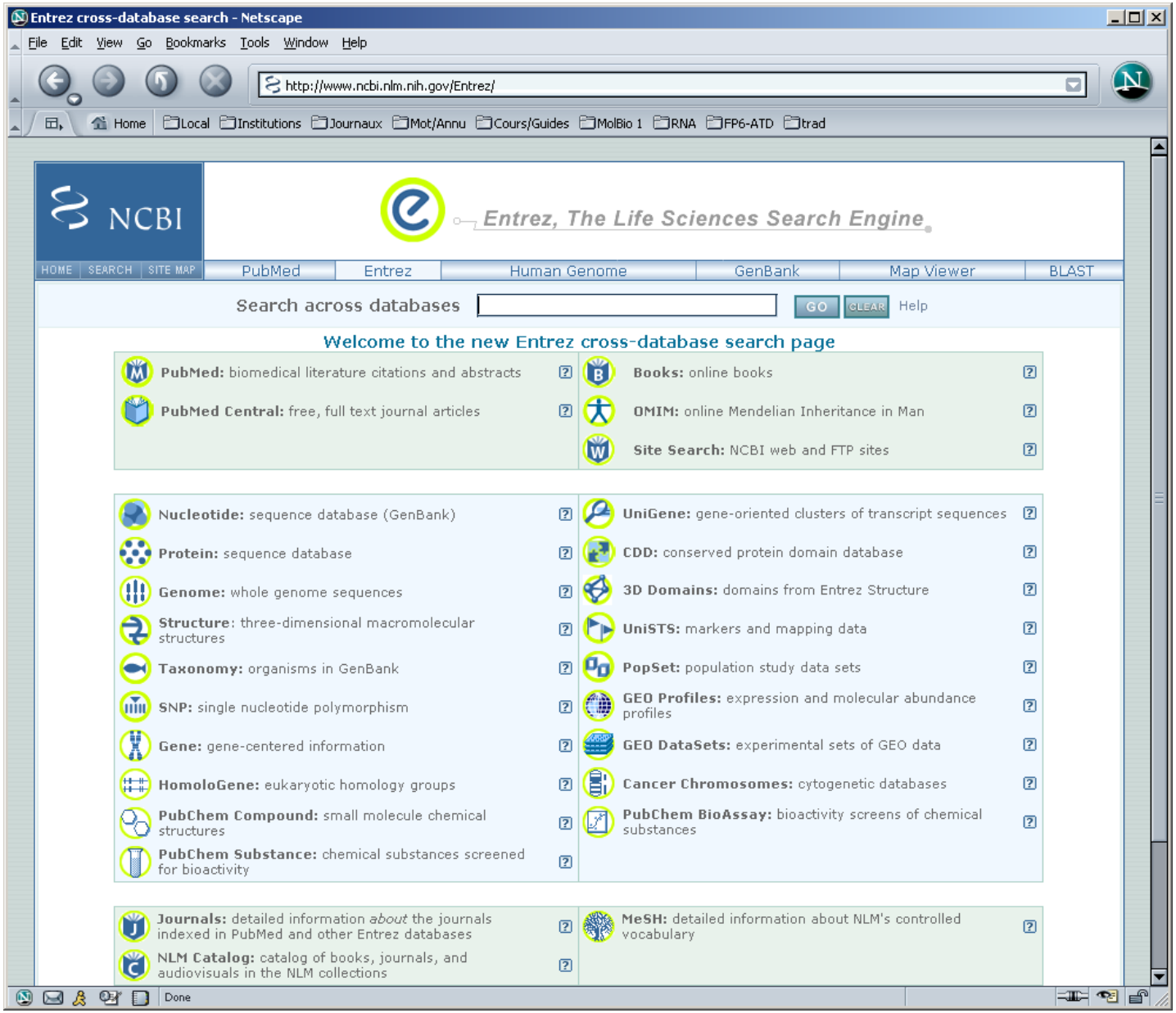


SRS

Query:


Standard Query Form - Netscape
File Edit View Go Bookmarks Tools Window Help
http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz
Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 RNA FP6-ATD trad
SRS@ EMBL-EBI
Quick Search Library Page Query Form Tools Results Projects Views
Reset
search UniProt/Swiss-Prot
Search Options
Combine search terms with: & (AND)
Use wildcards
Get results of type: Entry
Result Display Options
View results using: UniprotView
or
Create a view
Show 30 results per page
Fields you can search
Your search terms
In a single field, you can separate multiple values by &, |, !
Search
AllText argonaute
Organism Name drosophila
AllText
AllText
Create a view
Select the fields you want displayed in your view and choose the format
Choose 1 or more fields:
ID
EntryName
AccessionNumber
Creation Date
Seq Mod Date
Annot Mod Date
Description
Display As: Table List
Sequence Format: swiss
Done

Entrez


Entrez cross-database search - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/Entrez/
Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 RNA FP6-ATD trad
NCBI
Entrez, The Life Sciences Search Engine
HOME SEARCH SITE MAP
PubMed Entrez Human Genome GenBank Map Viewer BLAST
Search across databases
GO CLEAR Help
Welcome to the new Entrez cross-database search page
PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Books: online books
OMIM: online Mendelian Inheritance in Man
Site Search: NCBI web and FTP sites
Nucleotide: sequence database (GenBank)
Protein: sequence database
Genome: whole genome sequences
Structure: three-dimensional macromolecular structures
Taxonomy: organisms in GenBank
SNP: single nucleotide polymorphism
Gene: gene-centered information
HomoloGene: eukaryotic homology groups
PubChem Compound: small molecule chemical structures
PubChem Substance: chemical substances screened for bioactivity
UniGene: gene-oriented clusters of transcript sequences
CDD: conserved protein domain database
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
PopSet: population study data sets
GEO Profiles: expression and molecular abundance profiles
GEO DataSets: experimental sets of GEO data
Cancer Chromosomes: cytogenetic databases
PubChem BioAssay: bioactivity screens of chemical substances
Journals: detailed information about the journals indexed in PubMed and other Entrez databases
NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections
MeSH: detailed information about NLM's controlled vocabulary
Done

Glossaire de génomique…

http://www.sciencemag.org/cgi/content/full/291/5507/1197

(S*cience*, Vol 291, 1197)