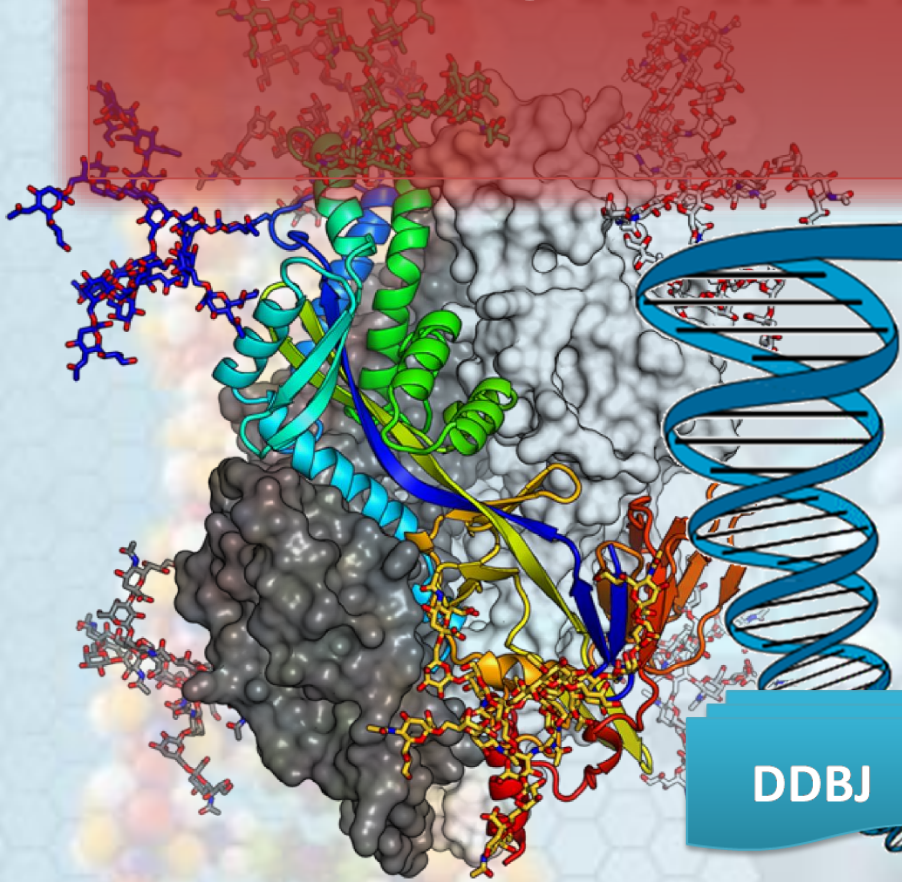
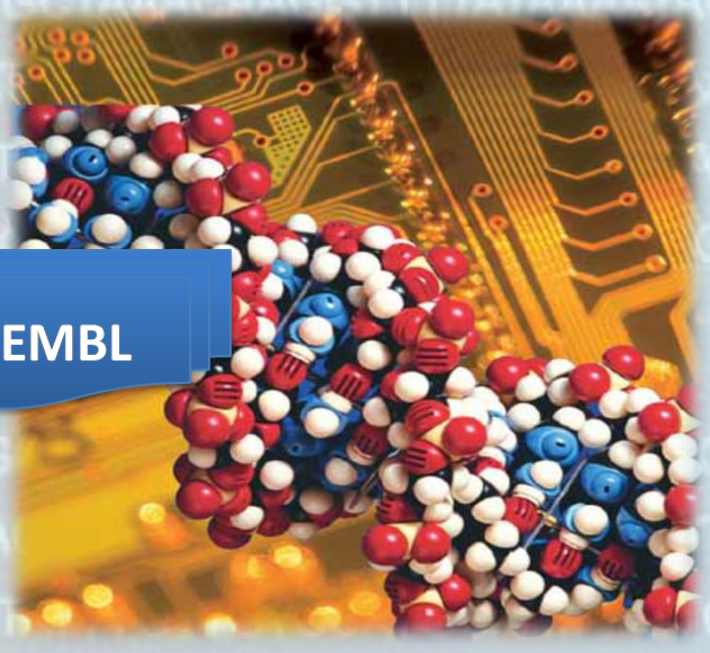
CHAPITRE I : INTRODUCTION A LA BIOINFORMATIQUE ET BANQUES DE DONNEES



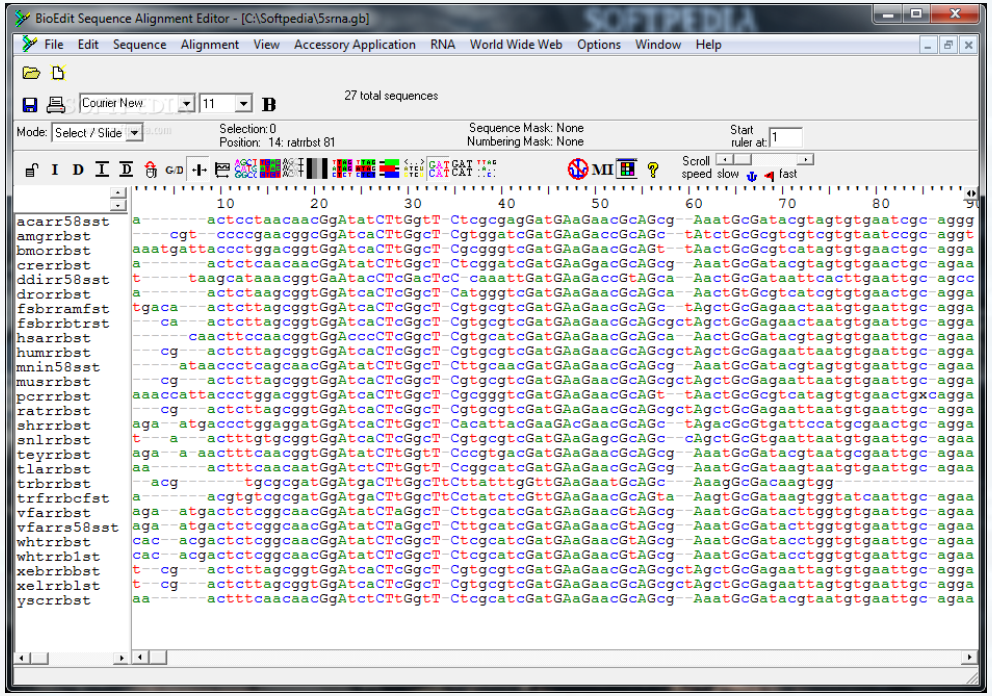
GenBank

DDBJ

EMBL



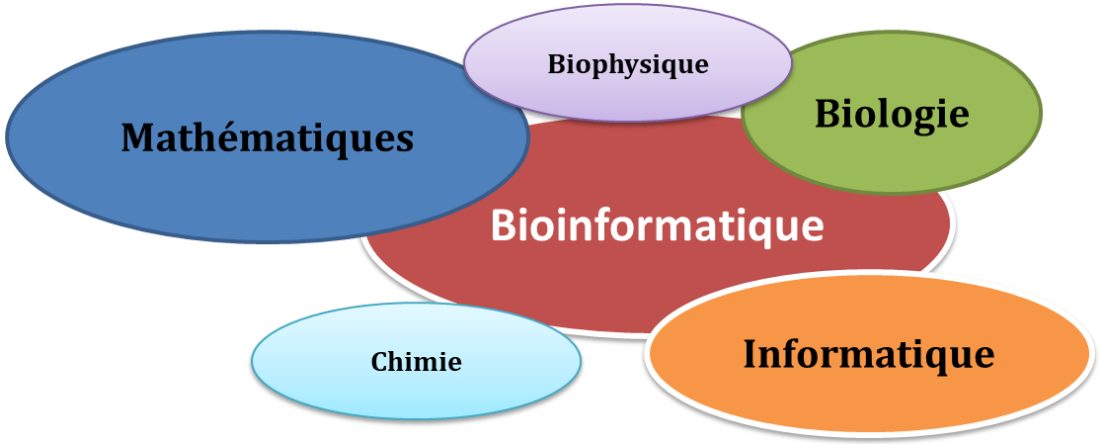
```
>EPI_ISL_15686 | A/Cygnus olor/Czech Republic/5170/06 | A / H5N1 | 2006-03-20 | MF
gtcgaaacgtacgttctctctatcatcccgtcaggcccctcaaagccgagatcgcgcagaaacttgaggatgtctttgc
aggaaagaacaccgatctcgaggctctcatggagtggctaaagacaagaccaatcctgtcacctctgactaaagggatgt
tgggatttgtattcacgctcaccgtgcccagtgagcgaggactgcagcgtagacgctttgtccagaatgccctaaatgga
>EPI_I
gtcgaa
aggaaa
tgggat
>EPI_aatgga
gtcgaggaggtc
aggaacggaagt
tgggaactatc
aatgggggatc
ggagcgggactcatcctaactctagtgctggtctgagagataatcttcttgaaaatttgcaggcttaccagaaacgaatgggagtg
cggaacagatgcagcgattcaagtgatcctcttgttgttgccgcaagtatcattgggatcttgcacttgatgttgtggattcttg
actatatcgtcttttcttcaaatgcatttatcgtcgccttaaatacggtttgaaaagagggccttctacggaaggagtacctgag
gggattctatgagggaagagtatcggcaggaacagcagaatgctg
ggactcatcctaactctagtgctggtctgagagataatcttcttgaaaatttgcaggcttaccagaaacgaatgggagtg
cagatgcagcgattcaagtgatcctcttgttgttgccgcaagtatcattgggatcttgcacttgatgttgtggattcttg
atcgtcttttcttcaaatgcatttatcgtcgccttaaatacggtttgaaaagagggccttctacggaaggagtacctgag
tctatgagggaagagtatcggcaggaacagcagaatgctg
```



INTRODUCTION

La bioinformatique est une ‘’interdiscipline’’ a la frontière de la biologie, de l’informatique , des mathématiques.

- Émergé dans les années 1980



- Le spécialiste qui travaille à mi-chemin entre ces sciences est appelé *bio-informaticien* ou *bionaute*.
- L'utilisation du terme bio-informatique est documentée pour la première fois en **1970** dans une publication de Paulien Hogeweg et Ben Hesper (université d'Utrecht, Pays-Bas)

But de la bio-informatique

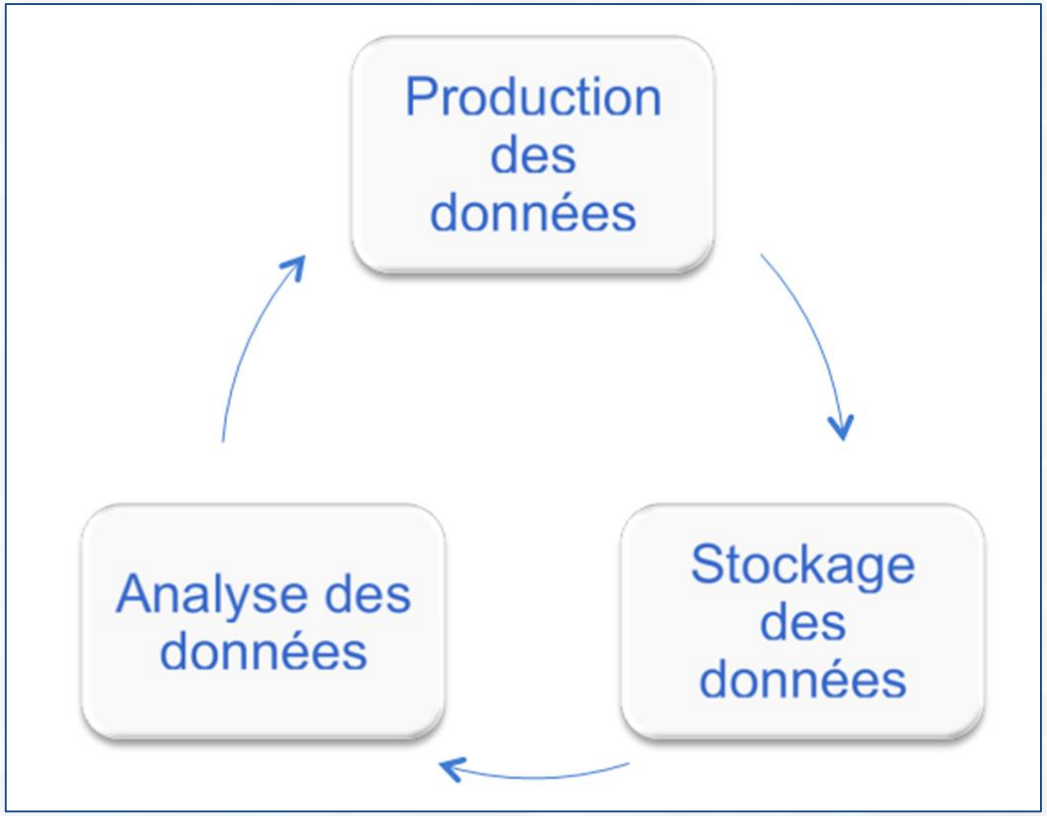
Les système biologique sont très complexes et les techniques modernes d'investigation du monde biologique fournissent une **vaste quantité** de **données expérimentales**

Donc le but ultime de la bio-informatique

« Est d'intégrer ces données d'origines très diverse pour modéliser les systèmes vivants afin de comprendre et prédire leurs comportements dans des conditions de fonctionnement normales ou pathologiques ».

LA BIOINFORMATIQUE DÉFINITION :

La bioinformatique est la discipline de l'analyse « *in silico* » de l'information biologique renfermée dans les séquences nucléotidiques (séquences de nucléotides) et protéiques (séquence des acides aminés).



- Son apparition, dans les années 1980. Coïncide avec la création des premières banques de données (**EMBL** et **GenBank**).

- À partir des années 1990, la bioinformatique devient <u>indispensable</u> avec l'accumulation des données de séquençage notamment les génomes complets.

- Fondée sur les acquis de la biologie, elle permet de produire de nouvelles connaissances et des suggestions pour de nouvelles expériences.

La bioinformatique propose des méthodes et des logiciels qui permettent :

- La collection, le stockage et la gestion des données biologiques et leur distribution à travers les réseaux.

- Le développement des outils (logiciels/algorithmes) pour analyser les problèmes de biologie moléculaire .

- L'analyse, la comparaison et la prédiction de la structure des gènes.

- La modélisation et la prédiction de la structure et de la fonction des protéines.

-Les études phylogénétiques et l'évolution moléculaire des êtres vivants.

LES BANQUE DE DONNÉES BIOLOGIQUES

Les bases de données contenant des informations biologiques et des données largement diffusées par le réseau Internet.

Elles sont généralement reliées entre elles par des liens « links » .

Il existe un grand nombre de bases de données d'intérêt biologique .

Nous nous limiterons ici à une présentation des principales banques de données publiques

DEUX TYPES DE BANQUES

-Celles qui correspondent à une collecte des données l **plus exhaustive** possible et qui offrent finalement un ensemble plutôt **hétérogène** d'informations.
-Traitent des thématiques générales

"Banques de données"

OU

Banques de données ou bases de données **GÉNÉRALISTES**

-Celles qui correspondent à des données **plus homogènes et spécifiques** .
-Traitent des thématiques particulières

"Bases de données",

OU

Banques de données ou bases de données **SPÉCIALISÉES**

LES BANQUES GÉNÉRALISTES

On appelle banques généralistes, ou banques primaires, les ressources qui collectent, gèrent, archivent et mettent à disposition de la communauté scientifique un ensemble de données primaires.

Classiquement, on considère comme banques primaires les banques généralistes qui contiennent des **séquences nucléiques** et **protéique** obtenus par des méthodes expérimentales.

bien que actuellement la plupart des **séquences protéiques** ne soient pas obtenues **expérimentalement**, mais à partir des données de **séquence nucléiques.**

ainsi que les banques qui gèrent **les structure tridimensionnelles** des protéines.

Banque nucléiques

Il existe trois banques nucléique internationales

(1) GenBank

la banque américaine gérée par le National Center for Biotechnology Information (NCBI)

(2) EMBL (European Molecular Biology Laboratory)

La banque européenne maintenue à l'European Bioinformatic Institute (EBI)

(3) DDBJ

La banque japonaise ou DNA DataBase of Japan

Ces trois banques gèrent l'ensemble des séquences nucléique et leurs annotations : elles coopèrent et échange quotidiennement leurs données afin de garantir une cohérence maximale dans la mise à disposition des séquences de la communauté scientifique.

Ces séquences sont organisées dans les banque sous forme des entrées .

Les entrés nucléiques

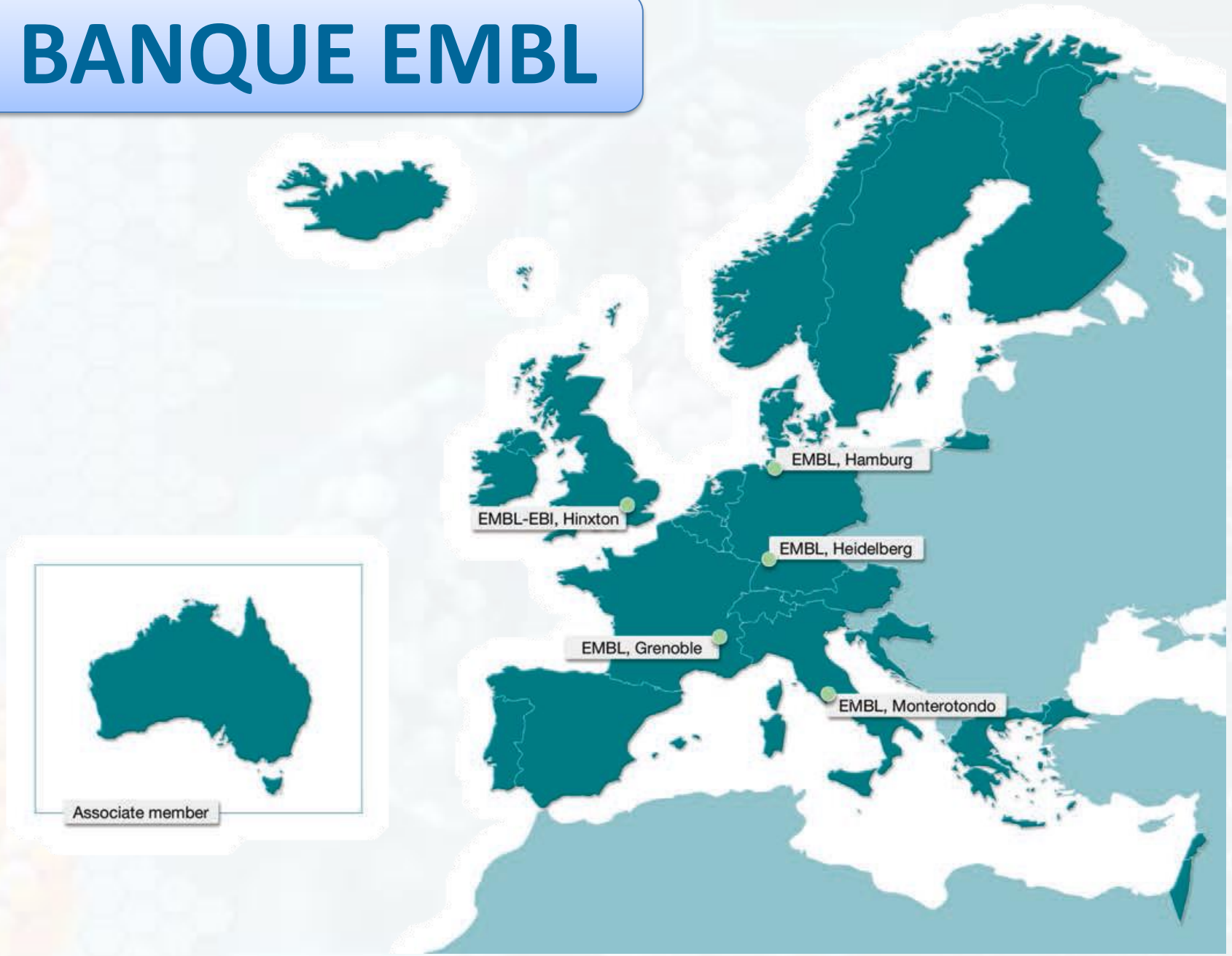
les entrées nucléiques sont organisées dans les trois banques se forme de « division », selon deux types de critères :

•**Le groupe taxonomique d'origine de la séquence** : Humans, bactéries, vertébrés, plantes, virus etc.

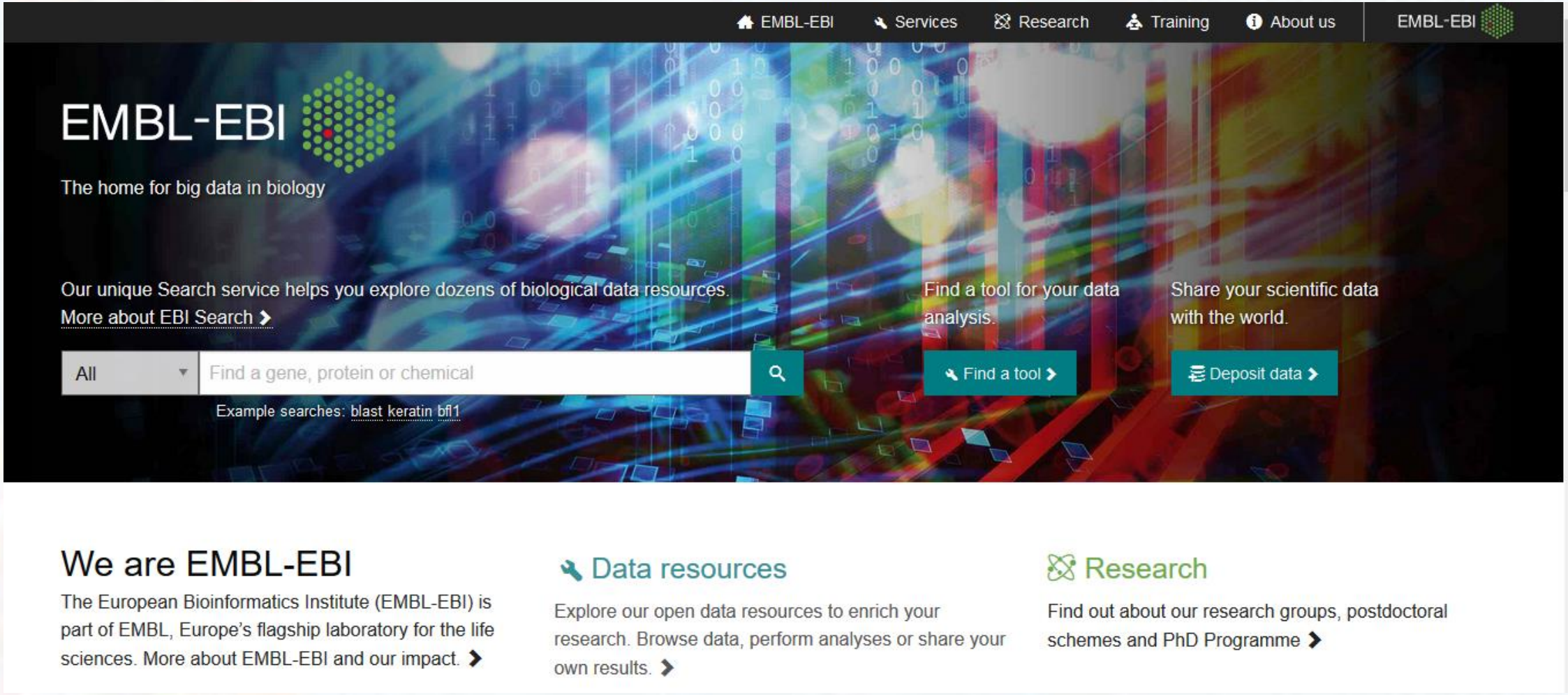
•**Le type de molécule séquencée** : **E**xpressed **S**equence **T**ag (**EST**)-et **G**enome **S**urvey **S**equence (**GSS**) -etc.

Division	Code pour les entrées
ESTs	EST
Bacteriophages	PHG
Fungi	FUN
Genome Survey	GSS
High Through Genome, HTGs	HTG
Humans	HUM
Invertebrates	INV

LA BANQUE EMBL



EMBL contient plus de 2.4 millions entrées (une entrée, entry en anglais, contient la séquence et des informations sur cette séquence).



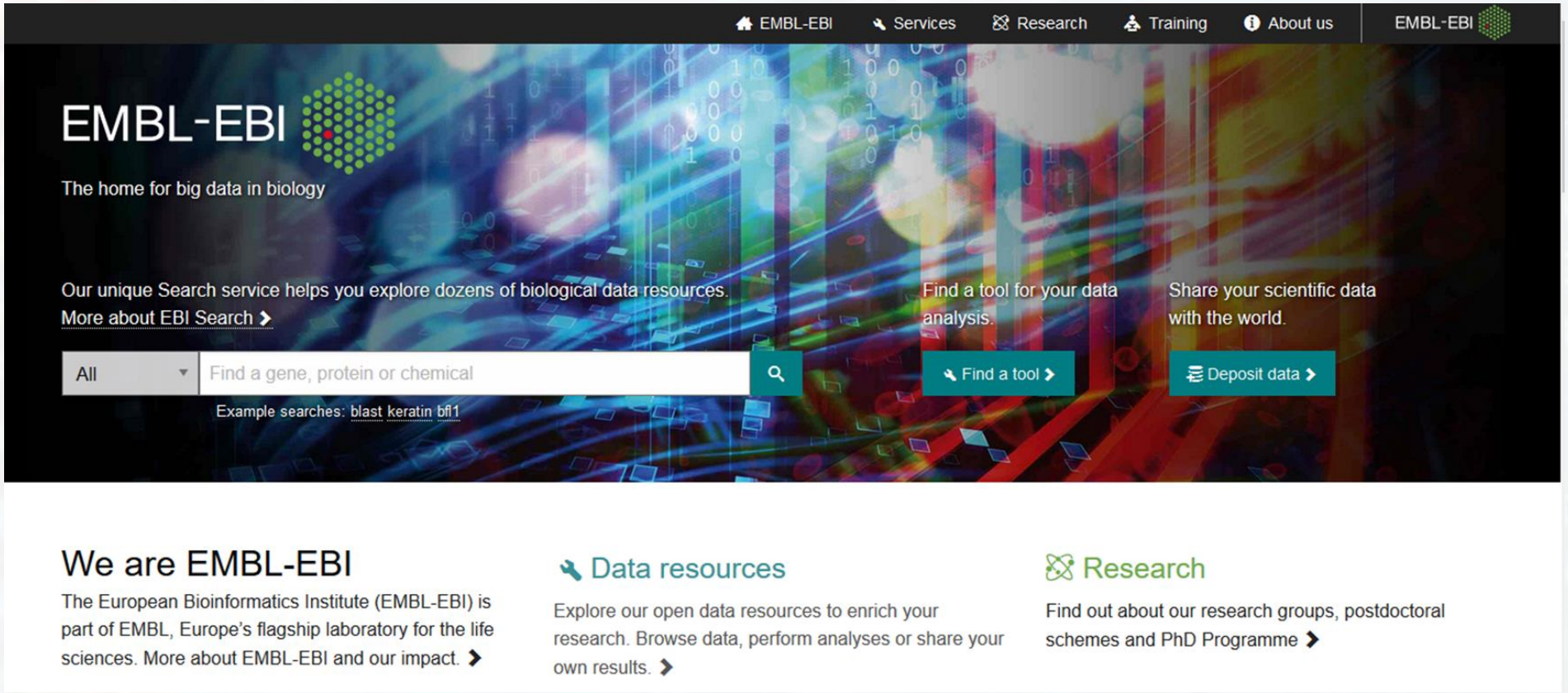
La recherche de sources de données biologiques en utilisant EMBL

Les ressources de données représentées dans EMBL comprennent:

-Séquences nucléotidiques et protéiques aux niveaux génomiques et protéomiques,

-Structures allant de produits chimiques aux complexes macromoléculaires……,

-Classifications fonctionnelles, les bibliothèques de la littérature globales qui couvrent les sciences biomédicales et la propriété intellectuelle connexe.



EMBL-EBI Services Research Training About us

EBI Search

INSULIN Search

Advanced

Examples: VAV_HUMAN, tpi1, Sulston ...

Help & Documentation | About EBI Search | Share | Feedback

Search results for *INSULIN*

Showing **19** results out of **410 211** in All results

Filter your results

Source

All results (410 211)
Genomes (1 986)
Nucleotide sequences (31 243)
Protein sequences (35 974)
Macromolecular structures (778)
Small molecules (1 718)
Gene expression (641)
Molecular interactions (418)
Reactions, pathways & diseases (1 399)
Protein families (229)
Enzymes (50)
Literature (331 467)
Samples & ontologies (4 230)
EBI web (78)

Gene & protein summaries (includes expression, structures, literature...) (4 results found)

Insulin-like receptor

InR (l(3)er10, CG18402, insulin receptor homolog, DIRbeta, DIRH, Dir-b, 18402, Inr-beta, DmelCG18402, InsR, ***Insulin***, insulin-like receptor, insulin/insulin-like growth factor receptor, DInR, DIHR, DILR, l(3)05545, FBgn0002393, insulin-receptor, lethal(3)93Dj, DIR, Inr-alpha, FBgn0010868, Drosophila insulin receptor, IR, er10, insulin receptor, insulin receptor homologue, dInsR, dInR, l(3)93Dj, FBgn0000456, INR, Dir-a, FBgn0000457, Inr, INS, FBgn0013984)

Fruit Fly (Drosophila melanogaster)

Insulin-like receptor Insulin-like receptor subunit alpha Insulin-like receptor subunit beta

daf-2 (Y55D5A.5)

Roundworm (Caenorhabditis elegans)

More...

View all available Gene & protein summaries

Filter your results

Source

All results (410 211)

Nucleotide sequences (31 243)

- Study (626)
- Sequence (Release) (23 227)
- Non-coding (Release) (8)
- Assembly contig (Release) (205)
- Assembly scaffold (Release) (466)
- Transcriptome assembly contig (Release) (248)
- Sequence (Update) (22)
- Assembly contig (Update) (6)
- Assembly scaffold (Update) (1)
- Study (Read/Analysis) (42)
- Sample (75)
- Read (Run) (1)
- Read (Experiment) (2)
- Coding (Release) (6 280)
- Coding (Update) (34)

Organisms

- Homo sapiens (9 866)
- unidentified (3 916)
- Mus musculus (3 138)
- synthetic construct (2 710)
- Acetobacter pomorum DM001 (2 474)
- Rattus norvegicus (736)
- Macrobrachium rosenbergii (309)
- Sus scrofa (285)
- Drosophila melanogaster (244)
- Bos taurus (241)

Nucleotide sequences (31 243 results found)

Transcriptome profile of peripheral blood mononuclear cells in patients with type I diabetes and their first grade relatives

Transcriptome profile of peripheral blood mononuclear cells in patients with type I diabetes and their first grade relatives

Related data | Views

Source: Study
ID: PRJNA140345

Function-based discovery of significant transcriptional temporal patterns in insulin-stimulated muscle cells

Function-based discovery of significant transcriptional temporal patterns in **insulin**-stimulated muscle cells

Related data | Views

Source: Study
ID: PRJNA140401

Co-activator function of RIP140 for NF?B/p65-dependent cytokine gene

Co-activator function of RIP140 for NF?B/p65-dependent cytokine gene

Related data | Views

Source: Study
ID: PRJNA108129

Deletion of the Mammalian INDY Homologue in Mice Mimics Aspects of Dietary Restriction and Protects Against Diet and Age-Induced Adiposity and Insulin Resistance

Deletion of the Mammalian INDY Homologue in Mice Mimics Aspects of Dietary Restriction and Protects Against Diet and Age-Induced Adiposity and **Insulin** Resistance

Related data | Views

Source: Study
ID: PRJNA140701

Histone Demethylase UTX-1 Regulates the Caenorhabditis elegans Lifespan by Targeting the Insulin/IGF-1 Signaling Pathway

Histone Demethylase UTX-1 Regulates the Caenorhabditis elegans Lifespan by Targeting the **Insulin**/IGF-1 Signaling Pathway

Related data | Views

Source: Study
ID: PRJNA140803

BCR/ABL1-Dependent Transcriptional Response Reveals Enrichment for Genes Involved in Negative Feedback Regulation

BCR/ABL1-Dependent Transcriptional Response Reveals Enrichment for Genes Involved in

Related data | Views

Source: Study
ID: PRJNA108487

EBI Search

insulin

Search

Advanced

Examples: VAV_HUMAN, tpi1, Sulston ...

Search results for *insulin*

Showing **15** results out of **9 866** in All results → Nucleotide sequences filtered by Organisms [X]

Filter your results

Source

All results (388 834)
Nucleotide sequences (9 866)
Sequence (Release) (9 445)
Non-coding (Release) (1)
Assembly scaffold (Release) (26)
Sample (38)
Coding (Release) (356)

Organisms

- [x] **Homo sapiens** (9 866)
- [ ] unidentified (3 916)
- [ ] Mus musculus (3 138)
- [ ] synthetic construct (2 710)
- [ ] Acetobacter pomorum DM001 (2 474)
- [ ] Rattus norvegicus (736)
- [ ] Macrobrachium rosenbergii (309)
- [ ] Sus scrofa (285)

Nucleotide sequences (9 866 results found)

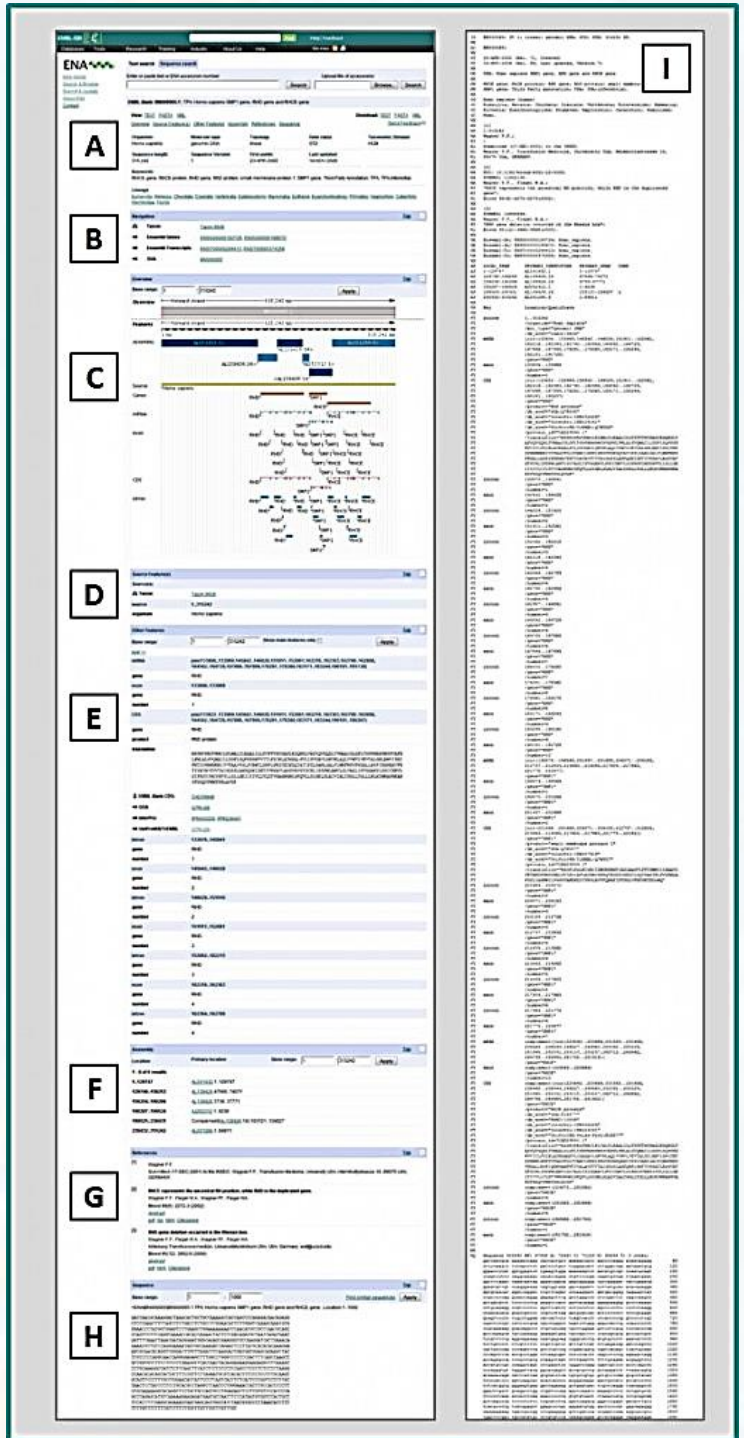
M10039
Human alpha-type **insulin** gene and 5' flanking polymorphic region.
Related data Views
Source: Sequence (Release)
ID: M10039

S99616
insulin {promoter} [human, Genomic, 450 nt].
Related data Views
Source: Sequence (Release)
ID: S99616

S99617
insulin {promoter} [human, Genomic Mutant, 21 nt].
Related data Views
Source: Sequence (Release)
ID: S99617

J05043
Human **insulin** receptor (IR) gene, exon 1.
Related data Views
Source: Sequence (Release)
ID: J05043

VUE D'ENSEMBLE D'UNE ENTRÉE EMBL-BANK



EMBL-Bank offre une vue facile à lire des données, où des informations telles que la taxonomie et des annotation sont regroupés en sections distinctes.

En outre, il a une représentation graphique de l'assemblage et des fonctions d'annotation.
EMBL-Bank a également une vue de texte brut qui est utile pour l'accès programmatique

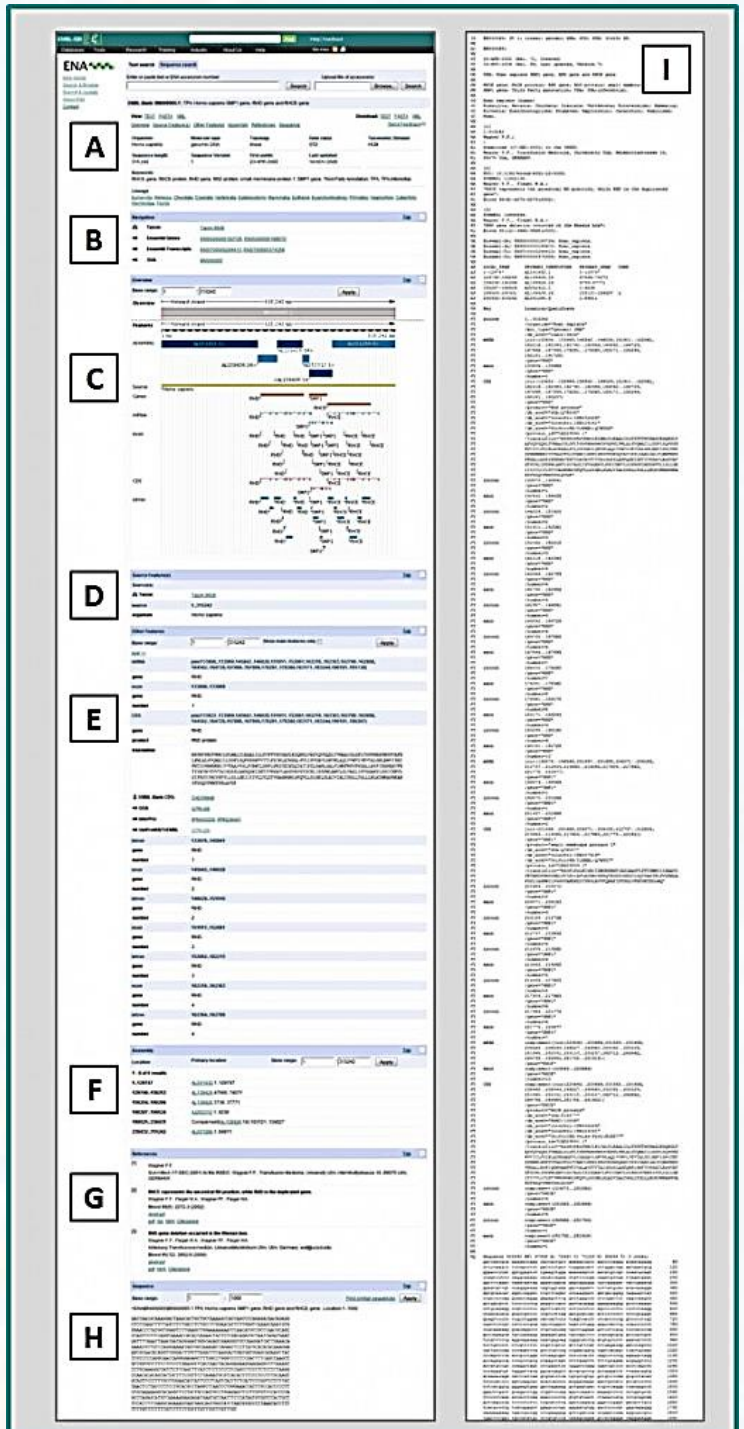
[A] Dans la vue par défaut, le sommaire d'entrée fournit des informations sur l'organisme, la division et le groupe taxonomique; vous pouvez également télécharger séquence et changer le point de vue de l'entrée.

[B] Partie fournit des liens vers d'autres ressources, y compris le portail de la taxonomie, Ensemble et la séquence Version Archive (de vieilles versions de l'entrée).

[C] Aperçu fournit une représentation graphique des données d'assemblage et d'annotation.

[D] Source entité (s) donne des informations sur la source de la séquence, telle que l'organisme, organite ou pays…

[E] Autres caractéristiques fournit des informations détaillées sur la fonction de différentes régions de la séquence.



[F] Assemblée fournit des informations détaillées sur la façon dont la séquence a été construit à partir de séquences de niveau inférieur.

[G] Références vous permettent de visualiser le document (s) citant la séquence et son annotation.

[H] Séquence peut être utilisée pour rechercher des séquences similaires dans la base de données.

[I] Le point de vue de texte de la même entrée; ce peut être consulté en cliquant sur 'TEXT' dans la section [A].

Ce point de vue est utile de vous écrivez des programmes car il fournit tous les codes de ligne qui identifie le type de ligne; par exemple «DE» identifie la ligne 'Description'.

Exemple d'entrée de la base EMBL :

```
ID   AI436639; SV 1; linear; mRNA; EST; HUM; 281 BP.
XX
AC   AI436639;
XX
DT   16-MAR-1999 (Rel. 59, Created)
DT   28-JAN-2011 (Rel. 107, Last updated, Version 5)
XX
DE   th61h01.x1 NCI_CGAP_Ov23 Homo sapiens cDNA clone IMAGE:2122801 3' similar
DE   to gb:X07868_rna1 PUTATIVE INSULIN-LIKE GROWTH FACTOR II ASSOCIATED
DE   (HUMAN), mRNA sequence.
XX
KW   EST.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-281
RA   NCI-CGAP;
RT   "National Cancer Institute, Cancer Genome Anatomy Project (CGAP), Tumor
RT   Gene Index http://www.ncbi.nlm.nih.gov/ncicgap";
RL   Unpublished.
XX
DR   UNILIB; 130; 1454.
XX
CC   Contact: Robert Strausberg, Ph.D.
CC   Email: cgapbs-r@mail.nih.gov
CC   Tissue Procurement: Christopher Moskaluk, M.D., Ph.D., Michael R.
CC   Emmert-Buck, M.D., Ph.D.
CC    cDNA Library Preparation: Life Technologies, Inc.
CC    cDNA Library Arrayed by: Greg Lennon, Ph.D.
CC    DNA Sequencing by: Washington University Genome Sequencing Center
CC    Clone distribution: NCI-CGAP clone distribution information can be
CC   found through the I.M.A.G.E. Consortium/LLNL at:
CC   www-bio.llnl.gov/bbrp/image/image.html
CC   Trace considered overall poor quality
CC   Insert Length: 638    Std Error: 0.00
CC   Seq primer: -40UP from Gibco
CC   Sequence considered overall poor quality.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..281
FT                   /organism="Homo sapiens"
FT                   /lab_host="DH10B"
FT                   /mol_type="mRNA"
FT                   /clone_lib="LIBEST_001454 NCI_CGAP_Ov23"
FT                   /clone="IMAGE:2122801"
FT                   /tissue_type="tumor, 5 pooled (see description)"
FT                   /note="Organ: ovary; Vector: pCMV-SPORT6; Site_1: SalI;
FT                   Site_2: NotI; Cloned unidirectionally. Primer: Oligo dT.
FT                   Average insert size 1.35 kb. Tumor types include: mixed
FT                   Mullerian tumor, papillary serous, clear cell, spindle
FT                   cell. All are primary tumors, metastasis positive. Life
FT                   Technologies catalog #: 11534-013"
FT                   /db_xref="taxon:9606"
FT                   /db_xref="UNILIB:130"
XX
SQ   Sequence 281 BP; 83 A; 62 C; 88 G; 47 T; 1 other;
     tgctttattg ggattgcaag cgttacaagg ttaaagacaa aacccaagca tgggattttg        60
     ccggaaatat tatcgttaaa ggagctgagt tgagtcaaac acgggcccca agggggaccg       120
     aggcggcagg cacaggtgac attcaatgtt tggcgtgggg gtcttcaagt gatggcaaaa       180
     gaggggaccc aaaagggggc cccccactga agacattggg gacaccggga ggagacaaaa       240
     tggaaagcca ccacttggcc ccgaggtcaa acaggcancc c                           281
//
```

tableau des codes et leurs significations :

Code	Signification, contenu de la ligne	Nombre /entrée
ID	C'est l'identificateur de l'entrée contenant la séquence. Cette ligne a la structure suivante : nom de l'entrée classe de la donnée ; molécule (DNA, RNA, RNAm, XXX si l'entrée n'a pas été annotée) ; division ; longueur de la séquence en paire de bases (BP).	1
XX	Cette ligne est une ligne vide qui sert à limiter les différents champs de l'entrée et à clarifier sa lecture.	Plusieurs
NI	Indique l'identificateur de l'acide nucléique.	1
AC	Donne le numéro d'accession de l'entrée.	>=1
DT	Donne la date d'incorporation dans la base (1ère ligne) et la date de la dernière mise à jour de l'entrée (2ème ligne).	>=1
DE	Contient des informations descriptives sur la séquence : comme la région du génome dont elle est issue …	>=1
KW	Donne le(s) mot(s)-clé(s) qui peuvent être utilisés pour retrouver l'entrée dans la base. Les mots-clés, séparés par des ; , sont rangés par ordre alphabétique.	>=1

```
ID   AI436639; SV 1; linear; mRNA; EST; HUM; 281 BP.
XX
AC   AI436639;
XX
DT   16-MAR-1999 (Rel. 59, Created)
DT   28-JAN-2011 (Rel. 107, Last updated, Version 5)
XX
DE   th61h01.x1 NCI_CGAP_Ov23 Homo sapiens cDNA clone IMAGE:2122801 3' similar
DE   to gb:X07868_rna1 PUTATIVE INSULIN-LIKE GROWTH FACTOR II ASSOCIATED
DE   (HUMAN), mRNA sequence.
XX
KW   EST.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-281
RA   NCI-CGAP;
RT   "National Cancer Institute, Cancer Genome Anatomy Project (CGAP), Tumor
RT   Gene Index http://www.ncbi.nlm.nih.gov/ncicgap";
RL   Unpublished.
XX
DR   UNILIB; 130; 1454.
XX
CC   Contact: Robert Strausberg, Ph.D.
CC   Email: cgapbs-r@mail.nih.gov
CC   Tissue Procurement: Christopher Moskaluk, M.D., Ph.D., Michael R.
CC   Emmert-Buck, M.D., Ph.D.
CC    cDNA Library Preparation: Life Technologies, Inc.
CC    cDNA Library Arrayed by: Greg Lennon, Ph.D.
CC    DNA Sequencing by: Washington University Genome Sequencing Center
CC    Clone distribution: NCI-CGAP clone distribution information can be
CC   found through the I.M.A.G.E. Consortium/LLNL at:
CC    www-bio.llnl.gov/bbrp/image/image.html
CC   Trace considered overall poor quality
CC   Insert Length: 638    Std Error: 0.00
CC   Seq primer: -40UP from Gibco
CC   Sequence considered overall poor quality.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..281
FT                   /organism="Homo sapiens"
FT                   /lab_host="DH10B"
FT                   /mol_type="mRNA"
FT                   /clone_lib="LIBEST_001454 NCI_CGAP_Ov23"
FT                   /clone="IMAGE:2122801"
FT                   /tissue_type="tumor, 5 pooled (see description)"
FT                   /note="Organ: ovary; Vector: pCMV-SPORT6; Site_1: SalI;
FT                   Site_2: NotI; Cloned unidirectionally. Primer: Oligo dT.
FT                   Average insert size 1.35 kb. Tumor types include: mixed
FT                   Mullerian tumor, papillary serous, clear cell, spindle
FT                   cell. All are primary tumors, metastasis positive. Life
FT                   Technologies catalog #: 11534-013"
FT                   /db_xref="taxon:9606"
FT                   /db_xref="UNILIB:130"
XX
SQ   Sequence 281 BP; 83 A; 62 C; 88 G; 47 T; 1 other;
     tgctttattg ggattgcaag cgttacaagg ttaaagacaa aacccaagca tgggattttg        60
     ccggaaatat tatcgttaaa ggagctgagt tgagtcaaac acgggcccca agggggaccg       120
     aggcggcagg cacaggtgac attcaatgtt tggcgtgggg gtcttcaagt gatggcaaaa       180
     gaggggaccc aaaaggggcc cccccactga agacattggg gacaccggga ggagacaaaa       240
     tggaaagcca ccacttggcc ccgaggtcaa acaggcancc c                           281
//
```

Code	Signification, contenu de la ligne	Nombre/ entrée
OS	Spécifie l'organisme d'où provient la séquence ; le plus souvent on donne le nom latin suivi du nom anglais entre parenthèses. Dans le cas d'hybrides, les lignes OC/OS sont spécifiées pour chaque organisme de l'hybride.	>=1
OC	1ère ligne : Donne le nom scientifique de l'organisme. 2ème ligne :Donne la classification taxonomique de l'organisme avec le groupe le plus général en premier, chaque groupe est séparé par un ;. Cette classification peut s'étendre sur plusieurs lignes OC.	>=1
OG	Indique la localisation sub-cellulaire des séquences non nucléaires.	0 ou 1
RN	Donne le numéro unique attribué à chaque référence bibliographique de l'entrée. Ce numéro est utilisé pour désigner la référence dans les commentaires (CC) et dans la table des caractéristiques (FT).	>=1
RC	Donne des commentaires sur la référence.	>=0
RX	Donne la région pour laquelle la référence bibliographique est associée.	>=0

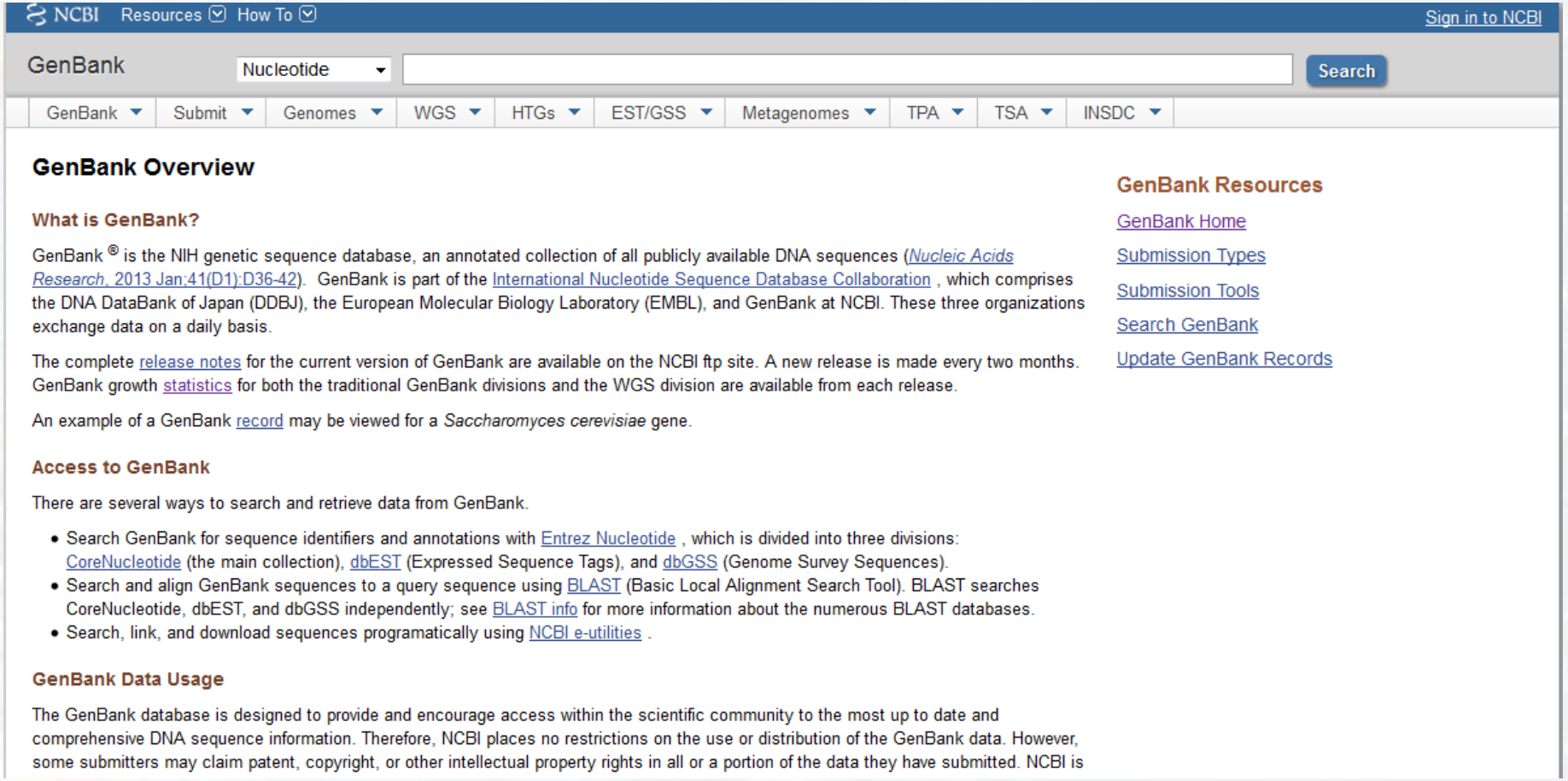
```
ID   AI436639; SV 1; linear; mRNA; EST; HUM; 281 BP.
XX
AC   AI436639;
XX
DT   16-MAR-1999 (Rel. 59, Created)
DT   28-JAN-2011 (Rel. 107, Last updated, Version 5)
XX
DE   th61h01.x1 NCI_CGAP_Ov23 Homo sapiens cDNA clone IMAGE:2122801 3' similar
DE   to gb:X07868_rna1 PUTATIVE INSULIN-LIKE GROWTH FACTOR II ASSOCIATED
DE   (HUMAN), mRNA sequence.
XX
KW   EST.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-281
RA   NCI-CGAP;
RT   "National Cancer Institute, Cancer Genome Anatomy Project (CGAP), Tumor
RT   Gene Index http://www.ncbi.nlm.nih.gov/ncicgap";
RL   Unpublished.
XX
DR   UNILIB; 130; 1454.
XX
CC   Contact: Robert Strausberg, Ph.D.
CC   Email: cgapbs-r@mail.nih.gov
CC   Tissue Procurement: Christopher Moskaluk, M.D., Ph.D., Michael R.
CC   Emmert-Buck, M.D., Ph.D.
CC    cDNA Library Preparation: Life Technologies, Inc.
CC    cDNA Library Arrayed by: Greg Lennon, Ph.D.
CC    DNA Sequencing by: Washington University Genome Sequencing Center
CC    Clone distribution: NCI-CGAP clone distribution information can be
CC   found through the I.M.A.G.E. Consortium/LLNL at:
CC    www-bio.llnl.gov/bbrp/image/image.html
CC   Trace considered overall poor quality
CC   Insert Length: 638    Std Error: 0.00
CC   Seq primer: -40UP from Gibco
CC   Sequence considered overall poor quality.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..281
FT                   /organism="Homo sapiens"
FT                   /lab_host="DH10B"
FT                   /mol_type="mRNA"
FT                   /clone_lib="LIBEST_001454 NCI_CGAP_Ov23"
FT                   /clone="IMAGE:2122801"
FT                   /tissue_type="tumor, 5 pooled (see description)"
FT                   /note="Organ: ovary; Vector: pCMV-SPORT6; Site_1: SalI;
FT                   Site_2: NotI; Cloned unidirectionally. Primer: Oligo dT.
FT                   Average insert size 1.35 kb. Tumor types include: mixed
FT                   Mullerian tumor, papillary serous, clear cell, spindle
FT                   cell. All are primary tumors, metastasis positive. Life
FT                   Technologies catalog #: 11534-013"
FT                   /db_xref="taxon:9606"
FT                   /db_xref="UNILIB:130"
XX
SQ   Sequence 281 BP; 83 A; 62 C; 88 G; 47 T; 1 other;
     tgctttattg ggattgcaag cgttacaagg ttaaagacaa aacccaagca tgggattttg        60
     ccggaaatat tatcgttaaa ggagctgagt tgagtcaaac acgggcccca agggggaccg       120
     aggcggcagg cacaggtgac attcaatgtt tggcgtgggg gtcttcaagt gatggcaaaa       180
     gaggggaccc aaaagggggc cccccactga agacattggg gacaccggga ggagacaaaa       240
     tggaaagcca ccacttggcc ccgaggtcaa acaggcancc c                           281
//
```

RP	Donne les références associées aux différentes régions de la séquence.	>=1
RA	Indique les auteurs de l'article ou du travail cité, ils sont inscrits dans l'ordre donné dans la publication.	>=1
RT	Indique le titre de l'article, si la séquence a été soumise à la base et non publiée, la ligne ne contiendra qu'un point virgule.	>=1
RL	Donne d'une manière abrégée, les références du journal.	>=1
DR	Etablit des liaisons avec d'autres bases de données qui contiennent une information en relation avec cette entrée. Par exemple, si la traduction protéique d'une séquence existe dans la banques de données Swiss-Prot, la ligne DR pointera sur l'entrée correspondante dans Swiss-Prot.	>=0
FH	Sert à améliorer la lecture d'une entrée : c'est l'en-tête du champ FT	0 ou 2
FT	Enumère les caractéristiques de la séquence, elle répond aux abréviations utilisées dans " the feature table ".	>=0
SQ	Donne la longueur de la séquence en paire de bases (bp) ainsi que le résumé de son contenu.	1
CC	Donne les commentaires sur la séquence.	>=0
	Des blancs pour introduire la séquence.	>=1
//	Indique la fin de l'entrée.	1

GenBank

GenBank, est une base de données de séquences nucléiques de NCBI, est une collection annotée de toutes les séquences nucléotidiques et protéiques disponibles publiquement.

La version 218, Feb 2020 GenBank contient plus de 199 millions entrées (Séquences) et plus de 228 milliards nucléotides.



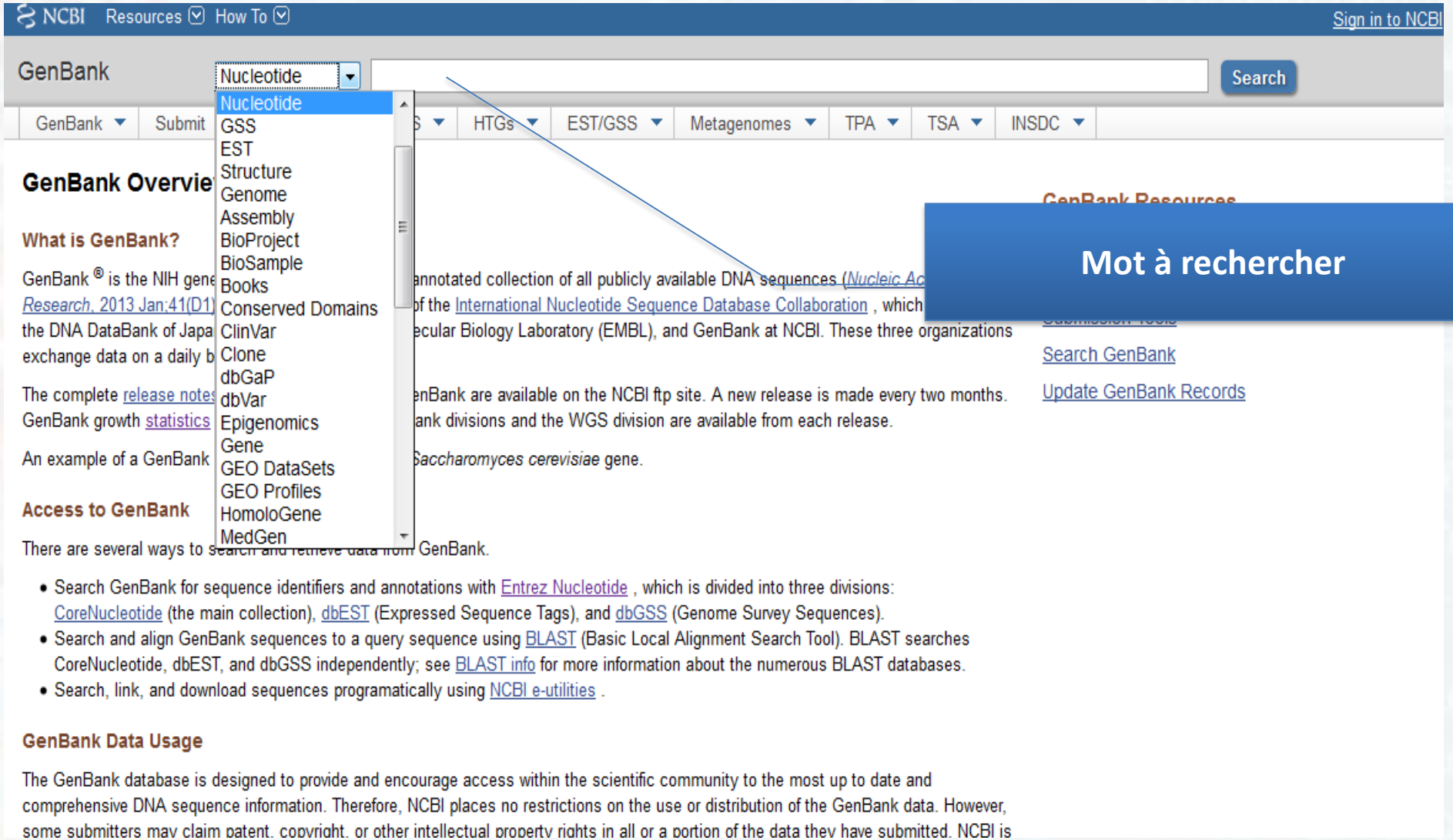
Division	Code pour le LOCUS des entrées
Primate	PRI
Rodent	ROD
Other mammalian	MAM
Other vertebrate	VRT
Invertebrate	INV
Plant	PLN
Bacterial	BCT
Structural RNA	RNA
Viral	VRL
Phage	PHG
Synthétic and chimeric	SYN
Unnanoted	UNA
Expressed Sequence Tag	EST
Patent	PAT
Sequence Tagged Site	STS
Genome Survey Sequence	GSS
High Throughput Genomic Sequencing	HTG

Actuellement, tous les enregistrements dans GenBank sont généré à partir soumissions directes aux bases de données de séquences d'ADN à partir des auteurs originaux, qui offrir leurs enregistrements pour rendre les données accessibles au public.

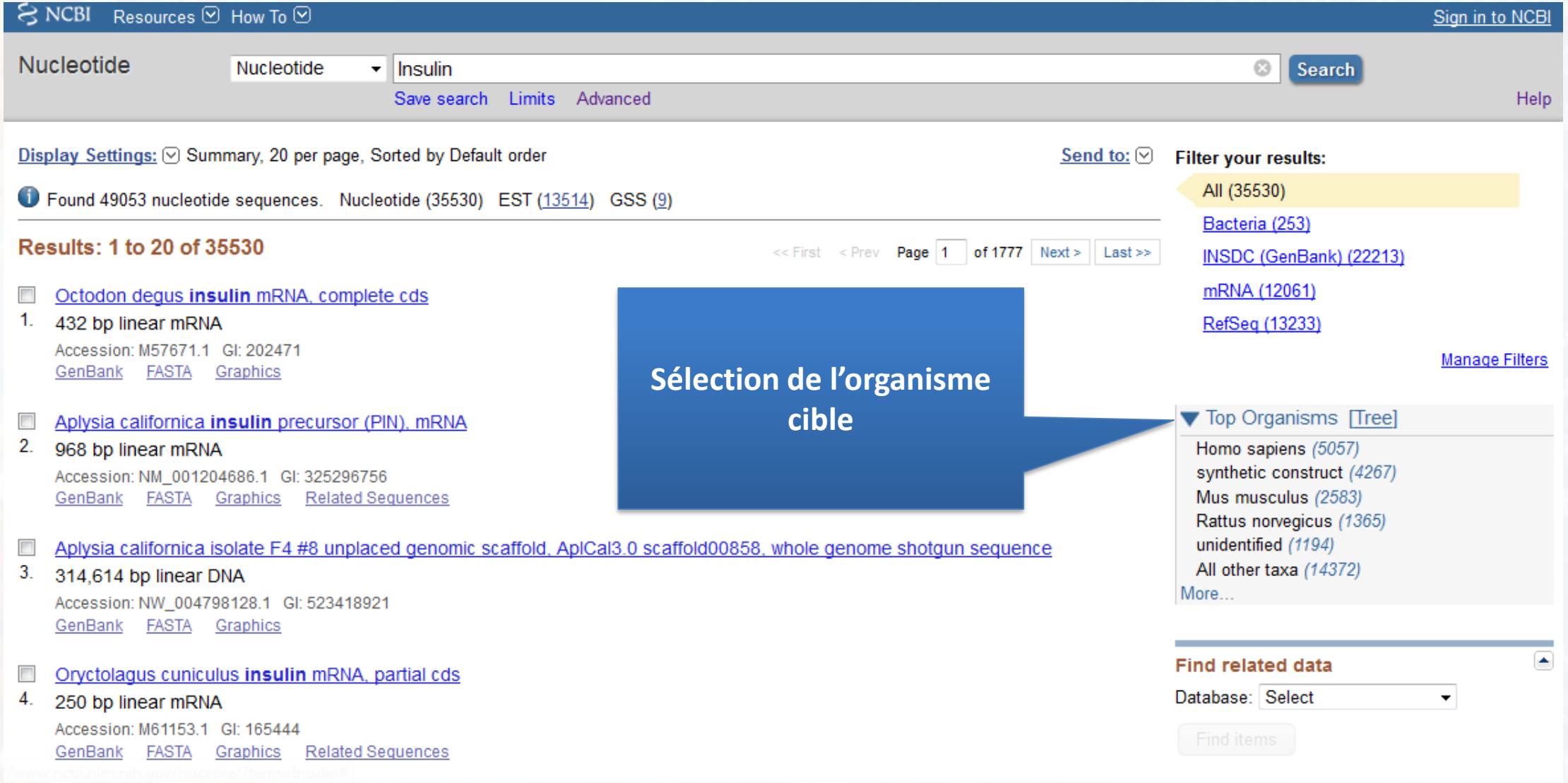
RECHERCHE DE SÉQUENCES NUCLÉOTIDIQUES DANS LE GenBank

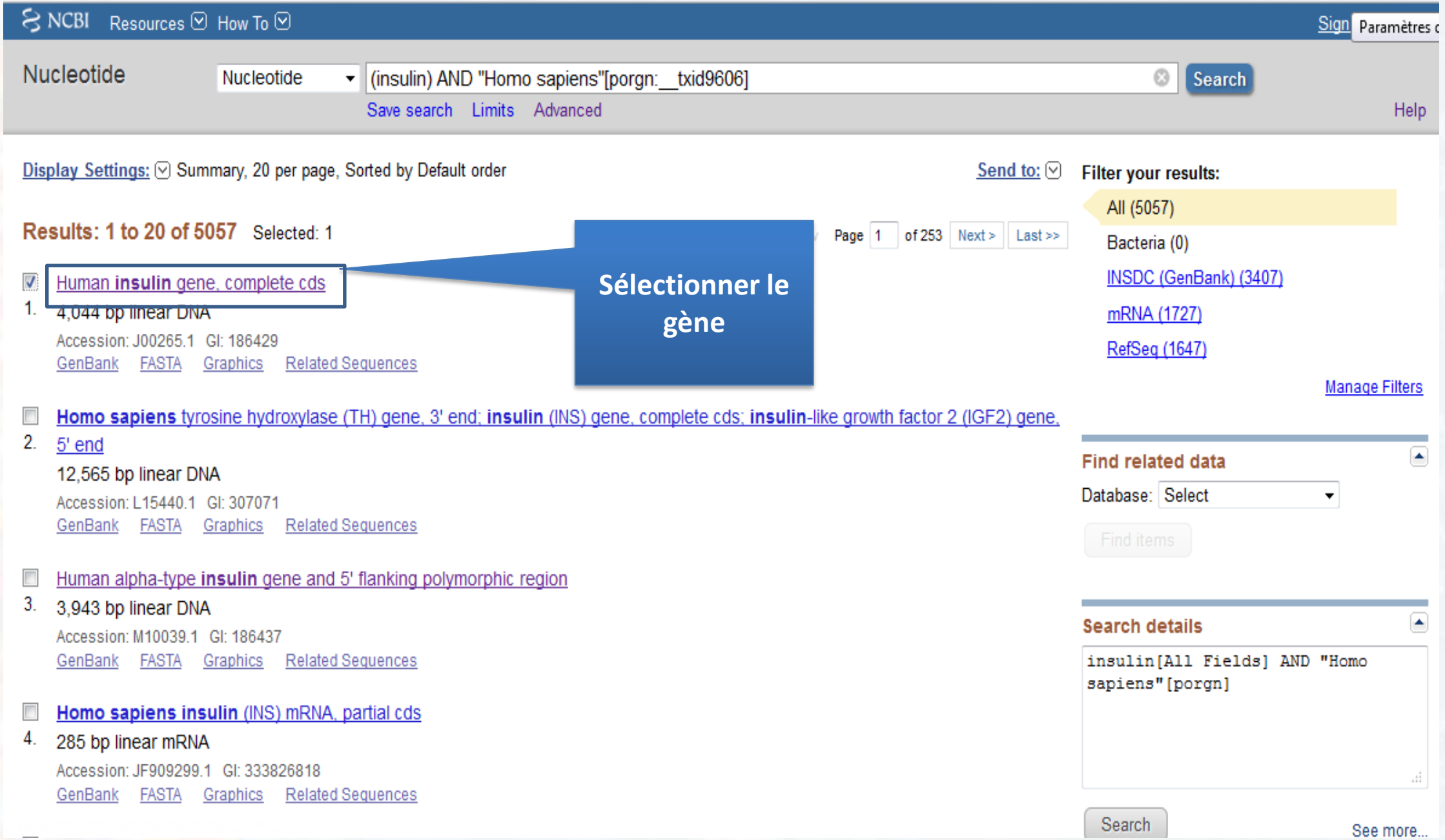
Recherche dans GenBank sur une séquence d'intérêt:

A l'aide du moteur de recherche intégré, rechercher la séquence souhaitée



Exemple Pour rechercher un gène, sélectionner la catégorie Gene dans le bandeau de recherche, pour un ARNm sélectionner Nucléotide.




NCBI Resources How To
Sign
Paramètres
Nucleotide
Nucleotide
(insulin) AND "Homo sapiens"[porgn:__txid9606]
Search
Save search Limits Advanced
Help
Display Settings: Summary, 20 per page, Sorted by Default order
Send to:
Results: 1 to 20 of 5057 Selected: 1
Page 1 of 253 Next > Last >>
Human insulin gene, complete cds
1. 4,044 bp linear DNA
Accession: J00265.1 GI: 186429
GenBank FASTA Graphics Related Sequences
Sélectionner le gène
Homo sapiens tyrosine hydroxylase (TH) gene, 3' end; insulin (INS) gene, complete cds; insulin-like growth factor 2 (IGF2) gene, 5' end
2. 12,565 bp linear DNA
Accession: L15440.1 GI: 307071
GenBank FASTA Graphics Related Sequences
Human alpha-type insulin gene and 5' flanking polymorphic region
3. 3,943 bp linear DNA
Accession: M10039.1 GI: 186437
GenBank FASTA Graphics Related Sequences
Homo sapiens insulin (INS) mRNA, partial cds
4. 285 bp linear mRNA
Accession: JF909299.1 GI: 333826818
GenBank FASTA Graphics Related Sequences
Filter your results:
All (5057)
Bacteria (0)
INSDC (GenBank) (3407)
mRNA (1727)
RefSeq (1647)
Manage Filters
Find related data
Database: Select
Find items
Search details
insulin[All Fields] AND "Homo sapiens"[porgn]
Search
See more...

Mot-clé	Sous mot-clé	Signification et contenu de la ligne	Lignes/entrée
LOCUS		C'est l'identificateur de l'entrée contenant la séquence. Cette ligne a la structure suivante : nom de l'entrée classe de la donnée ; molécule (ADN, ARN ou ADNc, XXX si l'entrée n'a pas été annotée) ; division ; longueur de la séquence en paire de bases (BP) ; date.	1
DEFINITION		Contient des informations descriptives sur la séquence : comme la région du génome dont elle est issue …	>=1
ACCESSION		Donne le numéro d'accession de l'entrée.	>=1
NID		Indique l'identificateur de l'acide nucléique.	1
VERSION		C'est un nouveau champ, il indique la version de l'entrée.	1
KEYWORDS		Donne le(s) mot(s)-clé(s) qui peuvent être utilisés pour retrouver l'entrée dans la base. Les mots-clés, séparés par des ; , sont rangés par ordre alphabétique.	>=1
SEGMENT		Indique la localisation de la séquence dans la molécule.	0 ou 1
SOURCE		Spécifie l'organisme d'où provient la séquence ; le plus souvent on donne le nom latin suivi du nom anglais entre parenthèses. Dans le cas d'hybrides, les lignes OC/OS sont spécifiées pour chaque organisme de l'hybride.	>=1
	ORGANISM	1ère ligne : Donne le nom scientifique de l'organisme. 2ème ligne :Donne la classification taxonomique de l'organisme avec le groupe le plus général en premier, chaque groupe est séparé par un ;. Cette classification peut s'étendre sur plusieurs lignes.	>=2
REFERENCE (mot-clé qui peut apparaître plusieurs fois dans une entrée)		Donne le numéro unique attribué à chaque référence bibliographique de l'entrée. Ce numéro est utilisé pour désigner la référence dans les commentaires et dans la table des caractéristiques.	>=1
	AUTHORS	Indique les auteurs de l'article ou du travail cité, ils sont inscrits dans l'ordre donné dans la publication.	>=1
	TITLE	Indique le titre de l'article, si la séquence a été soumise à la base et non publiée, la ligne ne contiendra qu'un point virgule.	>=0
	JOURNAL	Donne d'une manière abrégée, les références du journal.	>=1
	MEDLINE	Donne l'identifiant de la citation.	0 ou 1
	REMARK	Donne des commentaires sur la référence.	>=0
COMMENT		Donne les commentaires sur la séquences.	>=0
FEATURES		Enumère les caractéristiques de la séquence, elle répond aux abréviations utilisées dans " the feature table ".	1
BASE COUNT		Donne la longueur de la séquence en paire de bases (bp) ainsi que le résumé de son contenu.	1
ORIGIN		Sert à améliorer la lecture d'une entrée, introduit la séquence.	1

QUELQUES FORMATS DE FICHIERS DANS LES BANQUES DE DONNEES

Exemples de formats liés aux logiciels de traitement des séquences

1. Format FASTA

Sans doute le plus répandu et l'un des plus pratiques car très simple. La séquence, sous forme de lignes de 80 caractères maximum, est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">".

Plusieurs séquences peuvent être mises dans un même fichier.

Avec le format FASTA, un seul fichier peut contenir plusieurs enregistrements (séquences). Chaque enregistrement commence par ">".

```
>J00265.1 HUMINS01 Human insulin gene, complete Cds
CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCGGCCAGGGT
GTCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGCTGGGACACCAGCTGGC
CTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGCTGGGATCCTGGATCTCAGCTCCCT
GGCCGACAACACTGGCAAACTCCTACTCATCCACGAAGGCCCTCCTGGGCATGGTGGTCCTTCCCAGC
CTGGCAGTCTGTTCCTCACACACCTTGTTAGTGCCCAGCCCCTGAGGTTGCAGCTGGGGGTGTCTCTG
AAGGGCTGTGAGCCCCCAGGAAGCCCTGGGGAAGTGCCTGCCTTGCCTCCCCCCGGCCCTGCCAGCGC
CTGGCTCTGCCCTCCTACCTGGGCTCCCCCCATCCAGCCTCCCTCCCTACACACTCCTCTCAAGGAGG
CACCCATGTCCTCTCCAGCTGCCGGGCCTCAGAGCACTGTGGCGTCCTGGGGCAGCCACCGCATGTCC
TGCTGTGGCATGGCTCAGGGTGGAAAGGGCGGAAGGGAGGGGTCCTGCAGATAGCTGGTGCCCACTAC
CAAACCCGCTCGGGGCAGGAGAGCCAAAGGCTGGGTGTGTGCAGAGCGGCCCCGAGAGGTTCCGAGGC
TGAGGCCAGGGTGGGACATAGGGATGCGAGGGGCCGGGGCACAGGATACTCCAACCTGCCTGCCCCCA
TGGTCTCATCCTCCTGCTTCTGGGACCTCCTGATCCTGCCCCTGGTGCTAAGAGGCAGGTAAGGGGCT
GCAGGCAGCAGGGCTCGGAGCCCATGCCCCCTCACCATGGGTCAGGCTGGACCTCCAGGTGCCTGTTC
TGGGGAGCTGGGAGGGCCGGAGGGGTGTACCCCAGGGGCTCAGCCCAGATGACACTATGGGGGTGATG
GTGTCATGGGACCTGGCCAGGAGAGGGG
```

1. Format FASTA

Sans doute le plus répandu et l'un des plus pratiques car très simple. La séquence, sous forme de lignes de 80 caractères maximum, est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">".
Plusieurs séquences peuvent être mises dans un même fichier.

```
>J00265.1|HUMINS01 Human insulin gene, complete Cds
CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCGGCCAGGGTG
TCCC TTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGCTGGGACACCAGCTGGCCT
TCA GTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGCTGGGATCCTGGATCTCAGCTCCCTGGC
CG ACACTGGCAAACTCCTACTCATCCACGAAGGCCCTCCTGGGCATGGTGGTCCTTCCCAGCCTGG
C GTTCCTCACACACCTTGTTAGTGCCCAGCCCCTGAGGTTGCAGCTGGGGGTGTCTCTGAAGGG
C GCCCCCAGGAAGCCCTGGGGAAGTGCCTGCCTTGCCTCCCCCCGGCCCTGCCAGCGCCTGGCT
 TCCTACCTGGGCTCCCCCCATCCAGCCTCCCTCCCTACACACTCCTCTCAAGGAGGCACCCAT
 TCCAGCTGCCGGGCCTCAGAGCACTGTGGCGTCCTGGGGCAGCCACCGCATGTCCTGCTGTGG
 TCAGGGTGGAAAGGGCGGAAGGGAGGGGTCCTGCAGATAGCTGGTGCCCACTACCAAACCCGC
 GTGTGTGCAGAGCGGCCCCGAGAGGTTCCGAGGCTGAGGCCAGG
 GGGCACAGGATACTCCAACCTGCCTGCCCCCATGGTCTCATCC
 GCCCCTGGTGCTAAGAGGCAGGTAAGGGGCTGCAGGCAGCAGG
 GGGTCAGGCTGGACCTCCAGGTGCCTGTTCTGGGGAGCTGGGA
 TCAGCCCAGATGACACTATGGGGGTGATGGTGTCATGGGACCT
```

J00265.1 un enregistrement pourrait exister dans différentes bases de données
le «.1» indique que la séquence a été modifiée une fois

1. Format FASTA

Sans doute le plus répandu et l'un des plus pratiques car très simple. La séquence, sous forme de lignes de 80 caractères maximum, est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">".
Plusieurs séquences peuvent être mises dans un même fichier.

```
>J00265.1|HUMINS01 Human insulin gene, complete Cds
CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCGGCCAGGGTGTCCCC
TTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGCTGGGACACCAGCTGGCCTTCAAGGTCTC
TGCCTCCCTCCAGCCACCCCACTACACGCTGCTGGATCCTGGATCTCAGCTCCCTGGCCGACAACACTGGCAA
ACTCCTACTCATCCACGAAGGCCCTCCTGGGCGGTGGTCCTTCCCAGCCTGGCAGTCTGTTCCTCACACACC
TTGTTAGTGCCCAGCCCCTGAGGTTGCAGCTGGTGTCTCTGAAGGGCTGTGAGCCCCCAGGAAGCCCTGGG
GAAGTGCCTGCCTTGCCTCCCCCCGGCCCTGCGCCTGGCTCTGCCCTCCTACCTGGGCTCCCCCCATCCA
GCCTCCCTCCCTACACACTCCTCTCAAGCCCCATGTCCTCTCCAGCTGCCGGGCCTCAGAGCACTGTGG
CGTCCTGGGGCAGCCACCGCATGTCCTCCATGGCTCAGGGTGGAAAGGGCGGAAGGGAGGGGTCCTGC
AGATAGCTGGTGCCCACTACCAAACCCGCAGGAGAGCCAAAGGCTGGGTGTGTGCAGAGCGGCCCCG
AGAGGTTGCGAGGGGCCGGGGCACAGGATACTCCAACCTGCC
TGCCCCCGATCCTGCCCCTGGTGCTAAGAGGCAGGTAAGGGG
CTGCAGGTGGGTCAGGCTGGACCTCCAGGTGCCTGTTCTGGG
GAGCTGGCCCAGATGACACTATGGGGGTGATGGTGTCATGGG
ACCTGGC
```

Description de la séquence Dans cet exemple, « Insulin" est le nom du gène et Human est l'organisme à partir duquel il a été déterminé

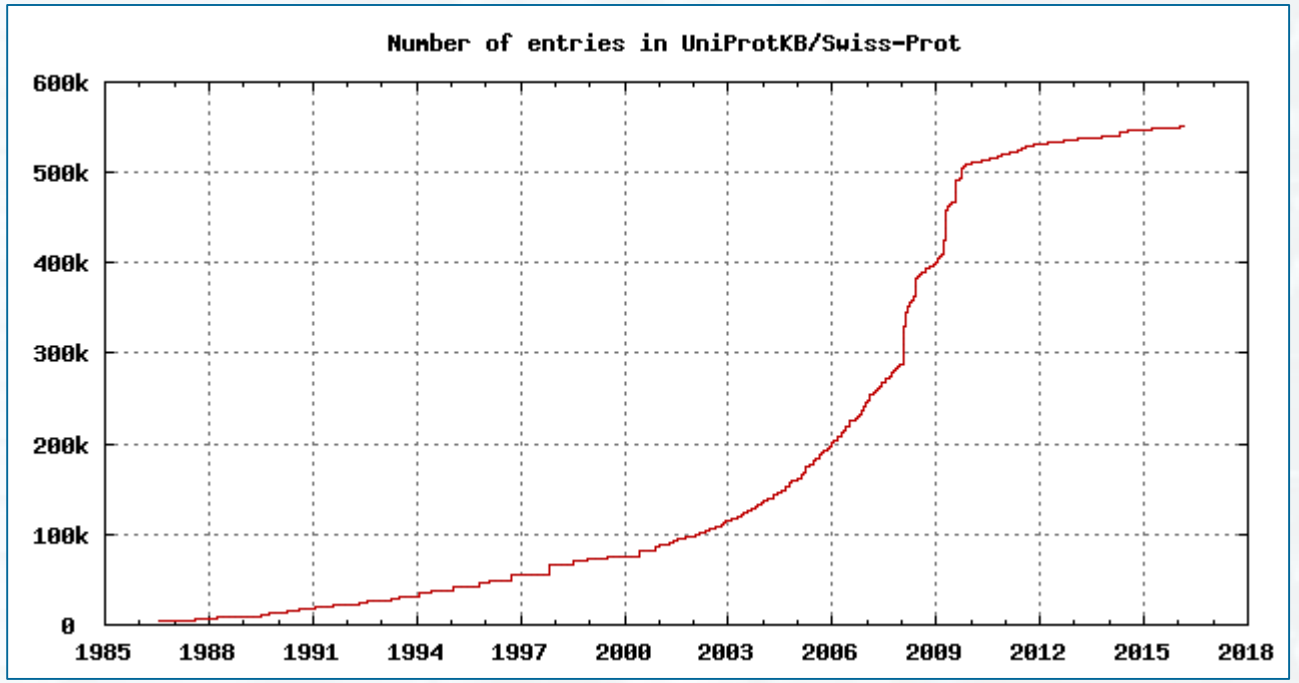
BANQUES PROTÉIQUES

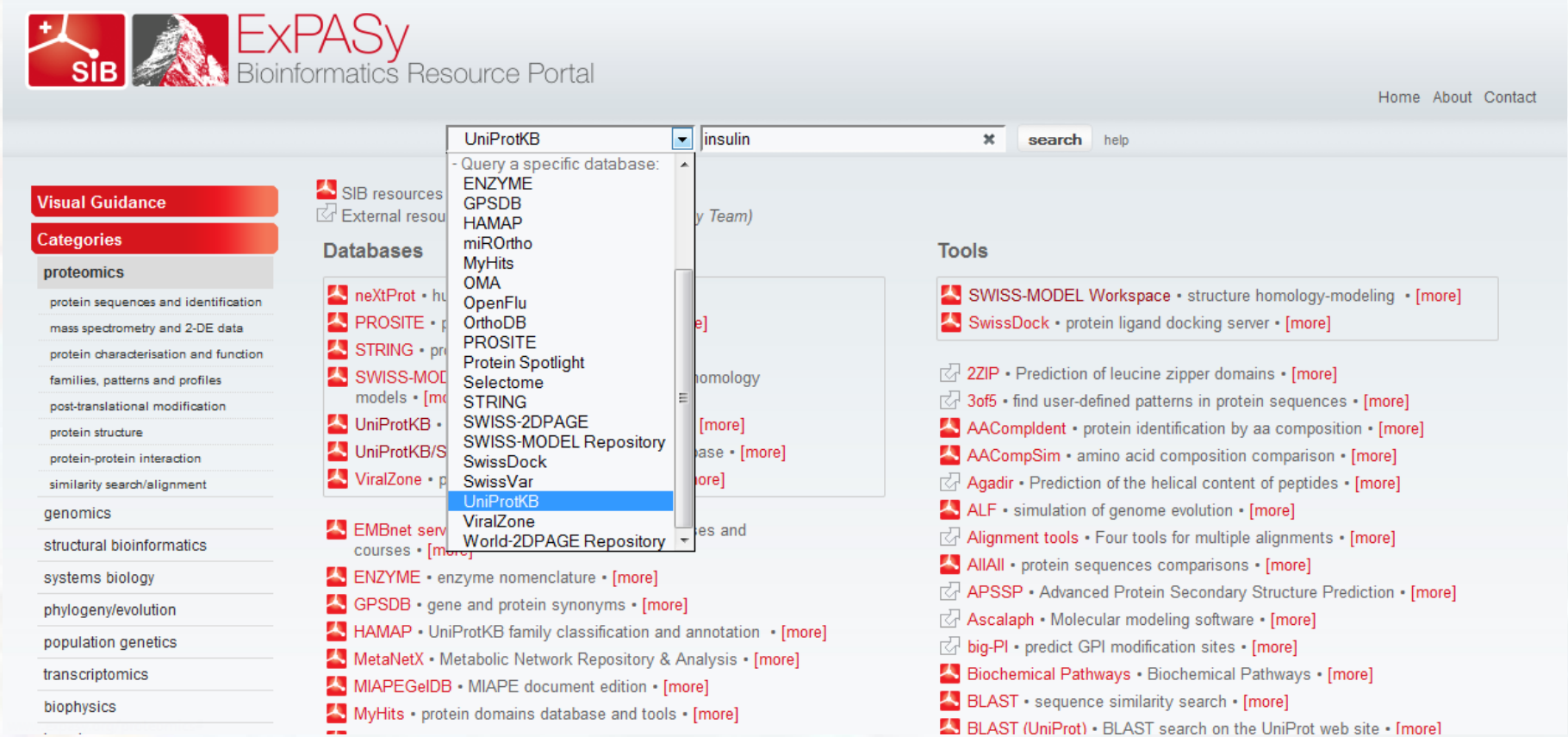
SwissProt

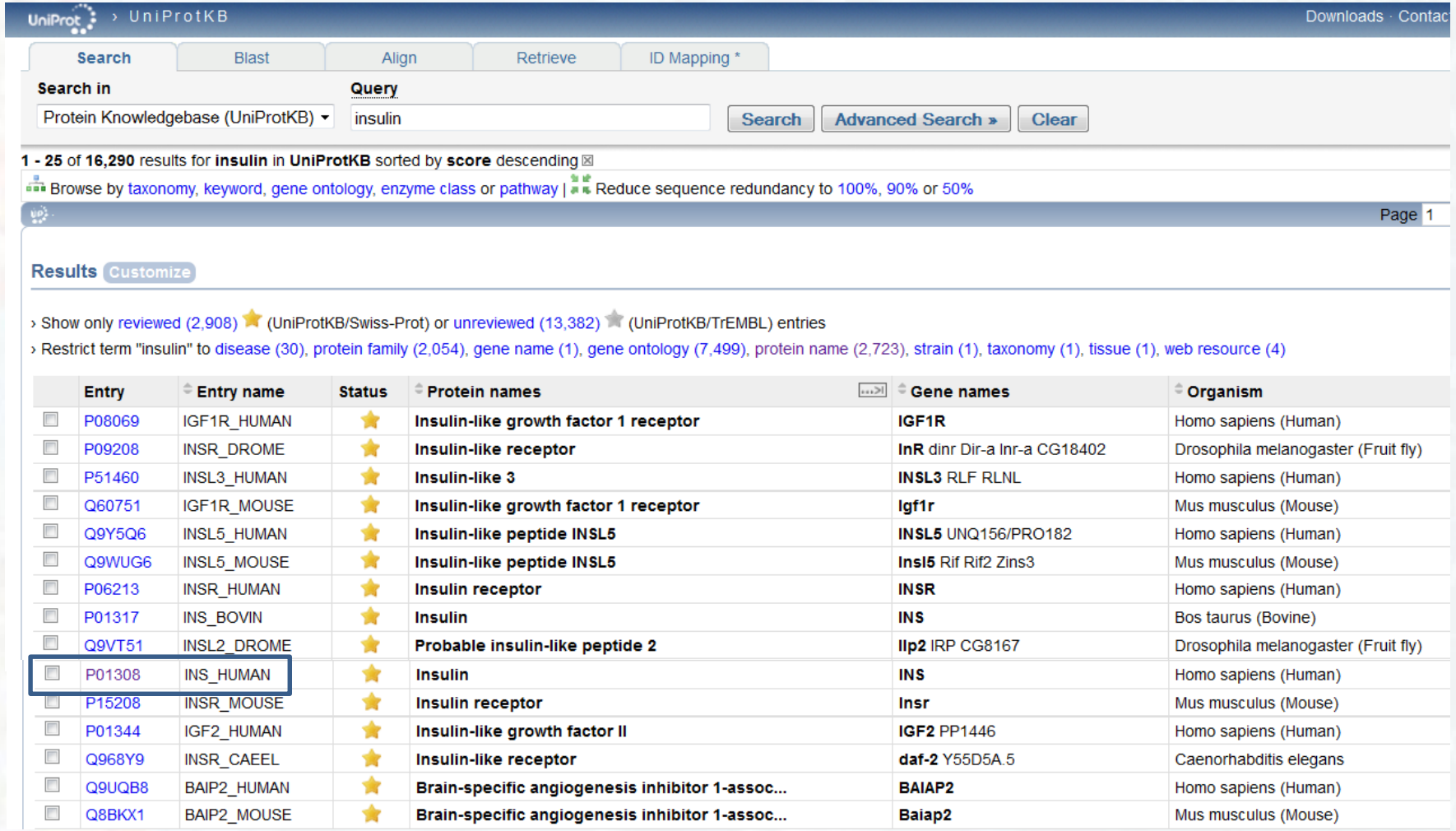
Créée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI.

SwissProt regroupe aussi des séquences annotées de la banque PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL

La version 2016_02 du 17-Feb-2016 de UniProtKB / Swiss-Prot contient 550552 entrées de séquence, comprenant 196472675 acides aminés




ExPASy
Bioinformatics Resource Portal
Home About Contact
UniProtKB
insulin
search
help
- Query a specific database:
ENZYME
GPSDB
HAMAP
miROrtho
MyHits
OMA
OpenFlu
OrthoDB
PROSITE
Protein Spotlight
Selectome
STRING
SWISS-2DPAGE
SWISS-MODEL Repository
SwissDock
SwissVar
UniProtKB
ViralZone
World-2DPAGE Repository
Visual Guidance
Categories
proteomics
protein sequences and identification
mass spectrometry and 2-DE data
protein characterisation and function
families, patterns and profiles
post-translational modification
protein structure
protein-protein interaction
similarity search/alignment
genomics
structural bioinformatics
systems biology
phylogeny/evolution
population genetics
transcriptomics
biophysics
SIB resources
External resou
Databases
neXtProt • hu
PROSITE • p
STRING • pr
SWISS-MOD
models • [mo
UniProtKB •
UniProtKB/S
ViralZone • p
EMBnet serv
courses • [more]
ENZYME • enzyme nomenclature • [more]
GPSDB • gene and protein synonyms • [more]
HAMAP • UniProtKB family classification and annotation • [more]
MetaNetX • Metabolic Network Repository & Analysis • [more]
MIAPEGelDB • MIAPE document edition • [more]
MyHits • protein domains database and tools • [more]
Tools
SWISS-MODEL Workspace • structure homology-modeling • [more]
SwissDock • protein ligand docking server • [more]
2ZIP • Prediction of leucine zipper domains • [more]
3of5 • find user-defined patterns in protein sequences • [more]
AACompIdent • protein identification by aa composition • [more]
AACompSim • amino acid composition comparison • [more]
Agadir • Prediction of the helical content of peptides • [more]
ALF • simulation of genome evolution • [more]
Alignment tools • Four tools for multiple alignments • [more]
AllAll • protein sequences comparisons • [more]
APSSP • Advanced Protein Secondary Structure Prediction • [more]
Ascalaph • Molecular modeling software • [more]
big-PI • predict GPI modification sites • [more]
Biochemical Pathways • Biochemical Pathways • [more]
BLAST • sequence similarity search • [more]
BLAST (UniProt) • BLAST search on the UniProt web site • [more]


UniProt › UniProtKB
Downloads · Contac
Search
Blast
Align
Retrieve
ID Mapping *
Search in
Protein Knowledgebase (UniProtKB)
Query
insulin
Search
Advanced Search »
Clear
1 - 25 of 16,290 results for insulin in UniProtKB sorted by score descending
Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%
Page 1
Results Customize
› Show only reviewed (2,908) (UniProtKB/Swiss-Prot) or unreviewed (13,382) (UniProtKB/TrEMBL) entries
› Restrict term "insulin" to disease (30), protein family (2,054), gene name (1), gene ontology (7,499), protein name (2,723), strain (1), taxonomy (1), tissue (1), web resource (4)
Entry
Entry name
Status
Protein names
Gene names
Organism
P08069 IGF1R_HUMAN Insulin-like growth factor 1 receptor IGF1R Homo sapiens (Human)
P09208 INSR_DROME Insulin-like receptor InR dinr Dir-a Inr-a CG18402 Drosophila melanogaster (Fruit fly)
P51460 INSL3_HUMAN Insulin-like 3 INSL3 RLF RLNL Homo sapiens (Human)
Q60751 IGF1R_MOUSE Insulin-like growth factor 1 receptor Igf1r Mus musculus (Mouse)
Q9Y5Q6 INSL5_HUMAN Insulin-like peptide INSL5 INSL5 UNQ156/PRO182 Homo sapiens (Human)
Q9WUG6 INSL5_MOUSE Insulin-like peptide INSL5 Insl5 Rif Rif2 Zins3 Mus musculus (Mouse)
P06213 INSR_HUMAN Insulin receptor INSR Homo sapiens (Human)
P01317 INS_BOVIN Insulin INS Bos taurus (Bovine)
Q9VT51 INSL2_DROME Probable insulin-like peptide 2 Ilp2 IRP CG8167 Drosophila melanogaster (Fruit fly)
P01308 INS_HUMAN Insulin INS Homo sapiens (Human)
P15208 INSR_MOUSE Insulin receptor Insr Mus musculus (Mouse)
P01344 IGF2_HUMAN Insulin-like growth factor II IGF2 PP1446 Homo sapiens (Human)
Q968Y9 INSR_CAEEL Insulin-like receptor daf-2 Y55D5A.5 Caenorhabditis elegans
Q9UQB8 BAIP2_HUMAN Brain-specific angiogenesis inhibitor 1-assoc... BAIAP2 Homo sapiens (Human)
Q8BKX1 BAIP2_MOUSE Brain-specific angiogenesis inhibitor 1-assoc... Baiap2 Mus musculus (Mouse)

LES BANQUES SPÉCIALISÉES

Les bases de données spécialisées sont d'intérêt divers et la masse des données qu'elles contiennent peut varier d'une base à une autre.

Ces bases correspondent à des améliorations ou à des regroupements par rapport aux données issues des bases généralistes
Donc on peut les appeler **banques secondaires**.

- Ces bases contiennent des données homogènes
- Collecte des données établie autour d'une thématique particulière

Exemple : bases spécialisée pour un génome spécifique, bases de séquences immunologiques, de voies métaboliques, de cartes génétiques, de motifs protéiques, d'expression de gènes, de structures, . . .

Quelques exemples

(LEAPdb)

Late Embryogenesis Abundant Proteins database (G. Hunault & E. Jaspard) : cette base de données contient un grand nombre d'informations sur les protéines LEA impliqués dans la tolérance à de nombreux stress, notamment la déshydratation et le froid. Pour l'instant, on les a mises en évidence principalement chez les plantes.

LEAPDB
1642 proteins

Home
Browse
Search
Blast
Statistical analysis
Export
Submit
Help
Contacts

Admin

Late Embryogenesis Abundant Proteins

LEA proteins have been discovered in 1981. Although, they are almost associated with abiotic stress tolerance (particularly dehydration and cold stress), their actual function remains unknown.

The LEAP database provides usefull data about the different CLASSES of LEA proteins for the analysis of their structure - function relationships. More...

Main classifications of LEAP with time – introduction of CLASS nomenclature

PFAM	Dure et al.	Bray	Turnaciffe and Wise	Battaglia et al.	Bies-Etheve et al.	Hundertmark and Hincha	LEAPdb
	1989	1993	2007	2008	2008	2008	2010
PF00257	D11	Group 2	Group 2	Group 2	Group 2	Dehydrin	Classes 1 to 4
PF00477	D19 D132	Group 1	Group 1	Group 1	Group 1	LEA_5	Class 5
PF02987	D7	Group 3	Group 3	Group 3A	Group 6	LEA_4	Class 6
	D29	Group 5	--	Group 3B	--		
PF03168	D95	--	--	Group 5C	Group 7	LEA_2	Classes 7 & 8
PF03242	D73	--	LEA5	Group 5B	Group 6	LEA_3	Class 9
PF03760	--	Group 4	Group 4	Group 4A	Group 4	LEA_1	Class 10
	D113			Group 4B			
PF04927	D34	Group 6	Group 6	Group 5A	Group 5	SMP	Class 11
PF10714	--	--	--	Group 6	Group 8	PvLEA18	Class 12
				Group 7			

Gilles Hunault and Emmanuel Jaspard,
LEAPdb: a database for the late embryogenesis abundant proteins
BMC Genomics 2010, 11:221

Emmanuel Jaspard, David Macherel and Gilles Hunault,
Computational and Statistical Analyses of Amino Acid Usage and Physico-Chemical Properties of the Twelve Late Embryogenesis Abundant Protein Classes
PLoS ONE 2012, 7(5)

Acc. # NCBI-GenPept: NP 001152689 - UniprotKB: B6UI06 protein 1001 / 1642
Class number 6
gi 226529441
Definition late embryogenesis abundant protein, group 3
organism Zea mays

Fasta seq.
MASRQQHPTSYHAGETKARAEEKTGQVMGATQEKGREAKHKASDASDRAMGMGHDAMEATREKARAAADR
TMGMGHDAGEAAKDRAYRAKDAASGAAGRARDTASDAAGAAGDRARDGAQQTGSYVAQTAEAARQKAAGA
ALYAKDTVVAGKDKTGALLQQAGEKVMSTAVGAKDTVVSTAVGAKDTVVSTAVGAKDAMMNSLGMAGEDK
DGTTTTDAGKDTSTRKPGRDY

Acc. # NCBI-GenPept: NP 001234972 - UniprotKB: Q39805 protein 1002 / 1642
Class number 4
gi 351723341
Definition dehydrin-like protein
organism Glycine max

Fasta seq.
MASYQKHYDDQGRKVDEYGNVEKQTDEYGNPVHAASVTYVATRTAAGGYSDDINKQHDTTNAYGVDTGRQ
HSSGGYDGDTNKHHGTTGGYNDDTNRHHGTTGVYGIDTDRQQHGTTGGYAGDTGRQHGNIGGPYYGTNTA
DTGTGPRSGTTGGTGYGGTGGTDYGTTGGTGYGSGTGYGVNTGGAHTEAGYRKEHRQHDQSHGDQNEKKG
IMDKIKEKLPGGHSDK

Acc. # NCBI-GenPept: NP 001235370 - UniprotKB: Q7XAW0 protein 1003 / 1642
Class number 4
gi 351727463
Definition lea protein
organism Glycine max

Fasta seq.
MASYQKHYDDQGRKVDEYGNVERQTDEYGNPVHATSVTYVATKSVGGYNDDANKQHDITGVYPEKDTGRH
HFGRGYDGVTNKQHDATGVYPGIDIGRDHGTTGVYGLNTDRHHGSTGVNPGIDTHNQQHGTTGGYAGDTG
RQHGNTGGLYYGTDTADTGAGPRSGNTGGTGYGGTGGTDYGTAGGTGYGSGTGYGINTGGAHTEAGYGKE
HRQHEQSHGGQHEKKGILDKIKEKLPGGHSDK

Acc. # NCBI-GenPept: NP 001237152 - UniprotKB: Q9XFT0 protein 1004 / 1642

LEAPDB

1642 proteins

Home

Browse

Search

Blast

Statistical analysis

Export

Submit

Help

Contacts

Returns to the Home Page

Admin

Search the LEAPDB: you can select one or more filtering parameters

ch AND OR

By LEAP class — detail of LEAP classes

With your own motif — you may use regular expressions.

By accession_number — (% for wildcards)

By organism

By pfam

By cdd

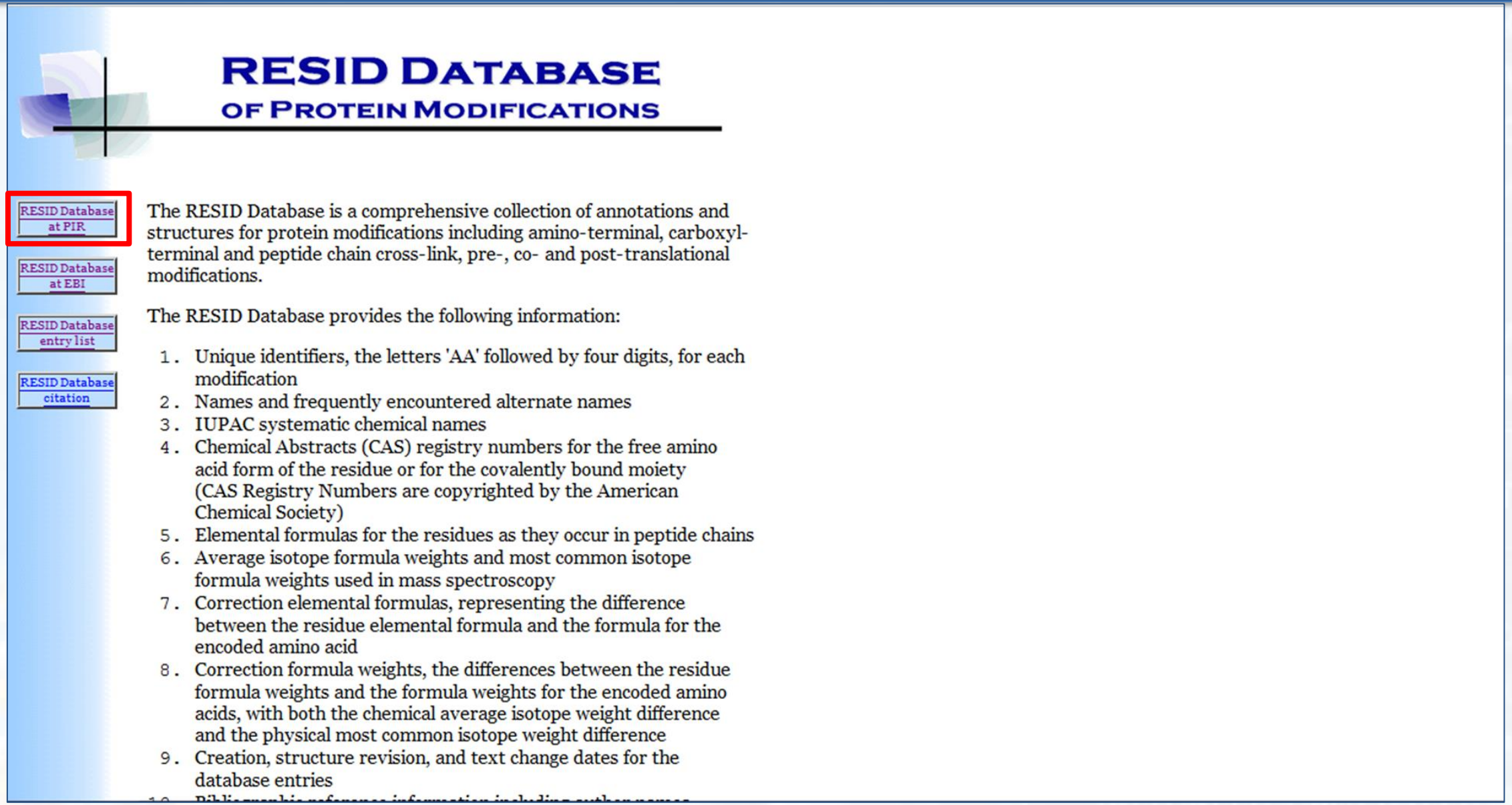
By date

By length — minimal length : maximal length :

proceed

RESID Database

Base de données sur les acides aminés peu fréquents (sous-partie de la base de données PIR)



RESID DATABASE
OF PROTEIN MODIFICATIONS

RESID Database at PIR
RESID Database at EBI
RESID Database entry list
RESID Database citation

The RESID Database is a comprehensive collection of annotations and structures for protein modifications including amino-terminal, carboxyl-terminal and peptide chain cross-link, pre-, co- and post-translational modifications.

The RESID Database provides the following information:

1. Unique identifiers, the letters 'AA' followed by four digits, for each modification
2. Names and frequently encountered alternate names
3. IUPAC systematic chemical names
4. Chemical Abstracts (CAS) registry numbers for the free amino acid form of the residue or for the covalently bound moiety (CAS Registry Numbers are copyrighted by the American Chemical Society)
5. Elemental formulas for the residues as they occur in peptide chains
6. Average isotope formula weights and most common isotope formula weights used in mass spectroscopy
7. Correction elemental formulas, representing the difference between the residue elemental formula and the formula for the encoded amino acid
8. Correction formula weights, the differences between the residue formula weights and the formula weights for the encoded amino acids, with both the chemical average isotope weight difference and the physical most common isotope weight difference
9. Creation, structure revision, and text change dates for the database entries

- Search
- Download
- Documentation and browsing tables
- FAQ
- Publications

RESID Database at PIR

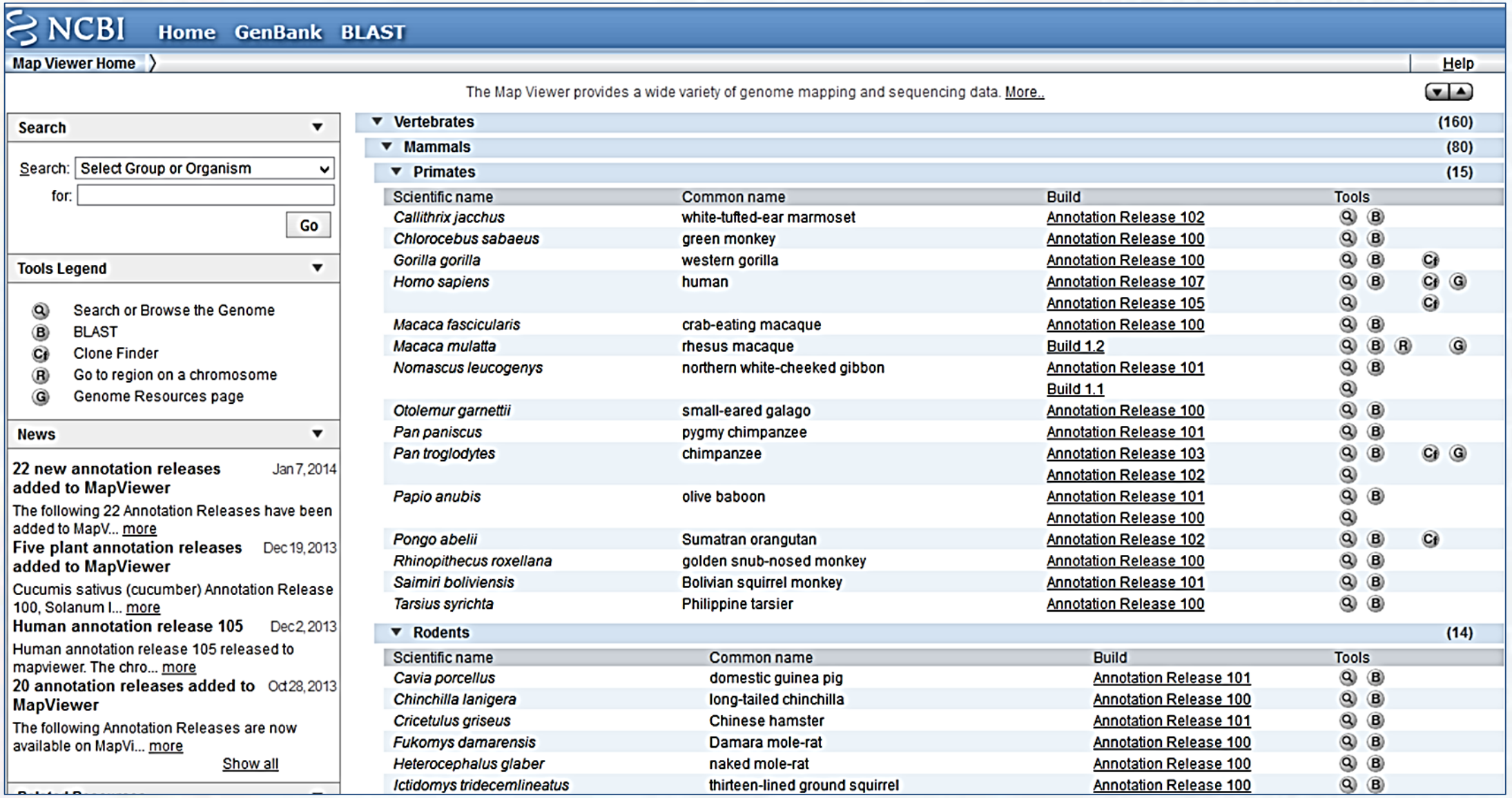
The RESID Database of Protein Modifications is a comprehensive collection of annotations and structures for protein modifications including amino-terminal, carboxyl-terminal and peptide chain cross-link post-translational modifications.

- RESID Home
- Search
- Download
- Documentation and browsing tables
- FAQ
- Publications

The RESID Database is produced by:
John S. Garavelli jsgarave@udel.edu
Center for Bioinformatics & Computational Biology
Delaware Biotechnology Institute
University of Delaware

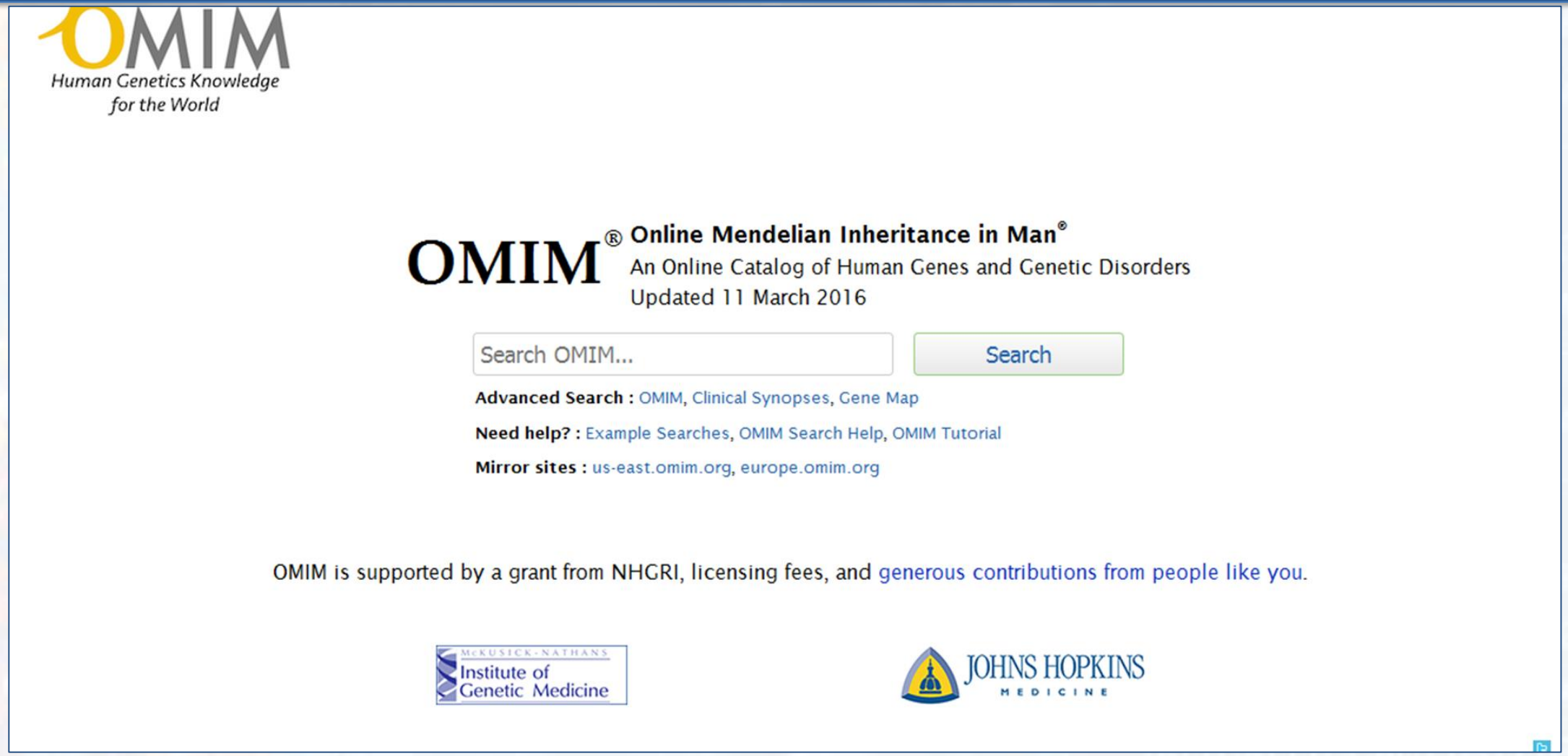
Map Viewer

Map Viewer vous permet de visualiser et de rechercher génome complet d'un organisme, affichage des cartes chromosomiques.
Le nombre et les types de cartes disponibles varient selon l'organisme



LA BASE DE DONNÉES OMIM (ONLINE MENDELIAN INHERITANCE IN MAN)

Donne de nombreuses informations sur la classification des maladies génétiques, des présentations cliniques et la cartographie génomique du localisation de la maladie.
La base de données est mise à jour continuellement et offre probablement le meilleur lors de la recherche d'information sur les maladies héréditaires.



NCBI Resources How To Sign in to NCBI

OMIM OMIM Search

Limits Advanced Help

OMIM

OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. Its official home is omim.org.

Using OMIM

Getting Started

FAQ

OMIM tools

OMIM API

Related Resources

ClinVar

Gene

GTR

MedGen

Last updated on: 11 Mar 2016