Notions Générales de Bioinformatique

Qu’est-ce que la Bio-Informatique?

Champs multi-disciplinaire qui utilise des méthodes informatiques (mathématiques, statistiques, combinatoires…) pour résoudre un problème biologique :

- Formaliser des problèmes de biologie moléculaire;
- Développer des outils formels;
- Analyser les données;
- Prédire des résultats biologiques;
- Organiser les données.

Discipline relativement nouvelle, qui évolue en fonction des nouveaux problèmes posés par la biologie moléculaire.

La bioinformatique. C'est quoi ?

- Ensemble de méthodes, de logiciels et d'applications en ligne qui permettent de gérer, manipuler, et analyser des données biologiques.

- La bioinformatique met en jeu plusieurs champs disciplinaires :

Informatique **Mathématiques formelles** **Statistiques**

Biologie

La bioinformatique. Pourquoi ?

La bioinformatique a différents objectifs et différentes applications :

1-Collecter et stocker des informations dans des bases de données, accessibles en ligne.

MEDLINE PubMed Production Statistics

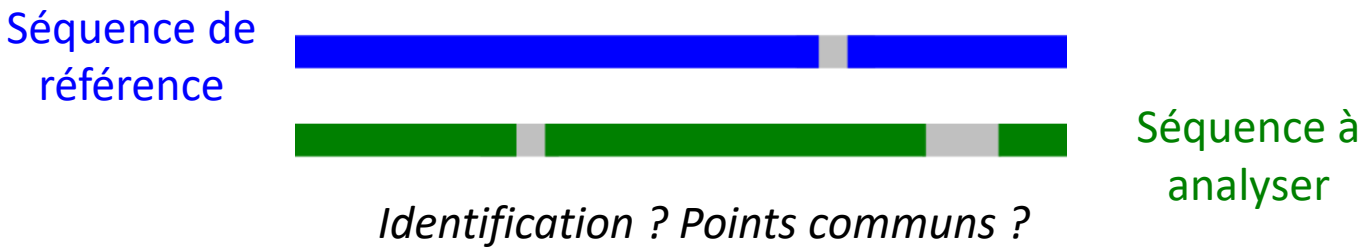
	FY2020	FY2019	FY2018	FY2017	FY2016	FY2015
MEDLINE Citations Indexed (Annual)	952,919	956,390	904,636	813,598	869,666	806,326
MEDLINE Citations Cumulative Total	27,149,277	26,196,358	25,239,968	24,335,332	23,531,948	22,391,870
MEDLINE Journal Titles	5,274	5,243	5,251	5,617	5,623	5,618
PubMed Citations (Annual)	1,514,199	1,366,447	1,329,148	1,150,125	1,165,957	1,091,693
PubMed Citations Cumulative Total	31,563,992	30,178,674	28,934,389	27,605,241	26,456,014	25,290,733
PubMed Searches	3.3 Billion	3.1 Billion	3.3 Billion	3.3 Billion	3.1 Billion	2.8 Billion
Web/Interactive	1,076 Million	896 Million	831 Million	846 Million	853 Million	910 Million
Script/E-Utilities	2.2 Billion	2.2 Billion	2.5 Billion	2.5 Billion	2.2 Billion	1.9 Billion

PubMed.gov
Go to PubMed

La bioinformatique. Pourquoi ?

La bioinformatique a différents objectifs et différentes applications :

2-Fournir des outils de comparaison de séquences (protéiques ou nucléotidiques).



<u>Objectifs</u> :

-identifier une séquence par rapport à une base de données

-déterminer le degré de similitudes entre deux séquences (intérêt en taxonomie)

-repérer des motifs structuraux :

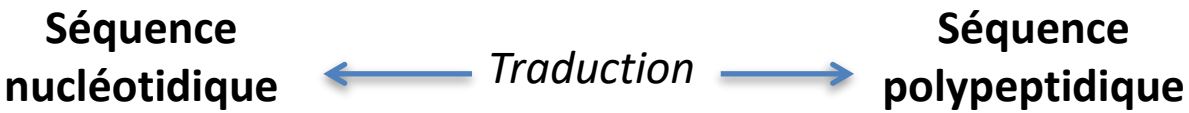
-gènes, promoteurs, etc. pour un nucléotide.

-zone de repliement, site actif, etc. pour un polypeptide.

La bioinformatique. Pourquoi ?

La bioinformatique a différents objectifs et différentes applications :

3-Fournir des outils de traduction de séquences.



Objectifs :
-simplifier les taches de traduction
-proposer plusieurs possibilités de protéines pour une même séquence
-repérer exons / introns

La bioinformatique. Pourquoi ?

La bioinformatique a différents objectifs et différentes applications :

4-Fournir des outils de prédiction

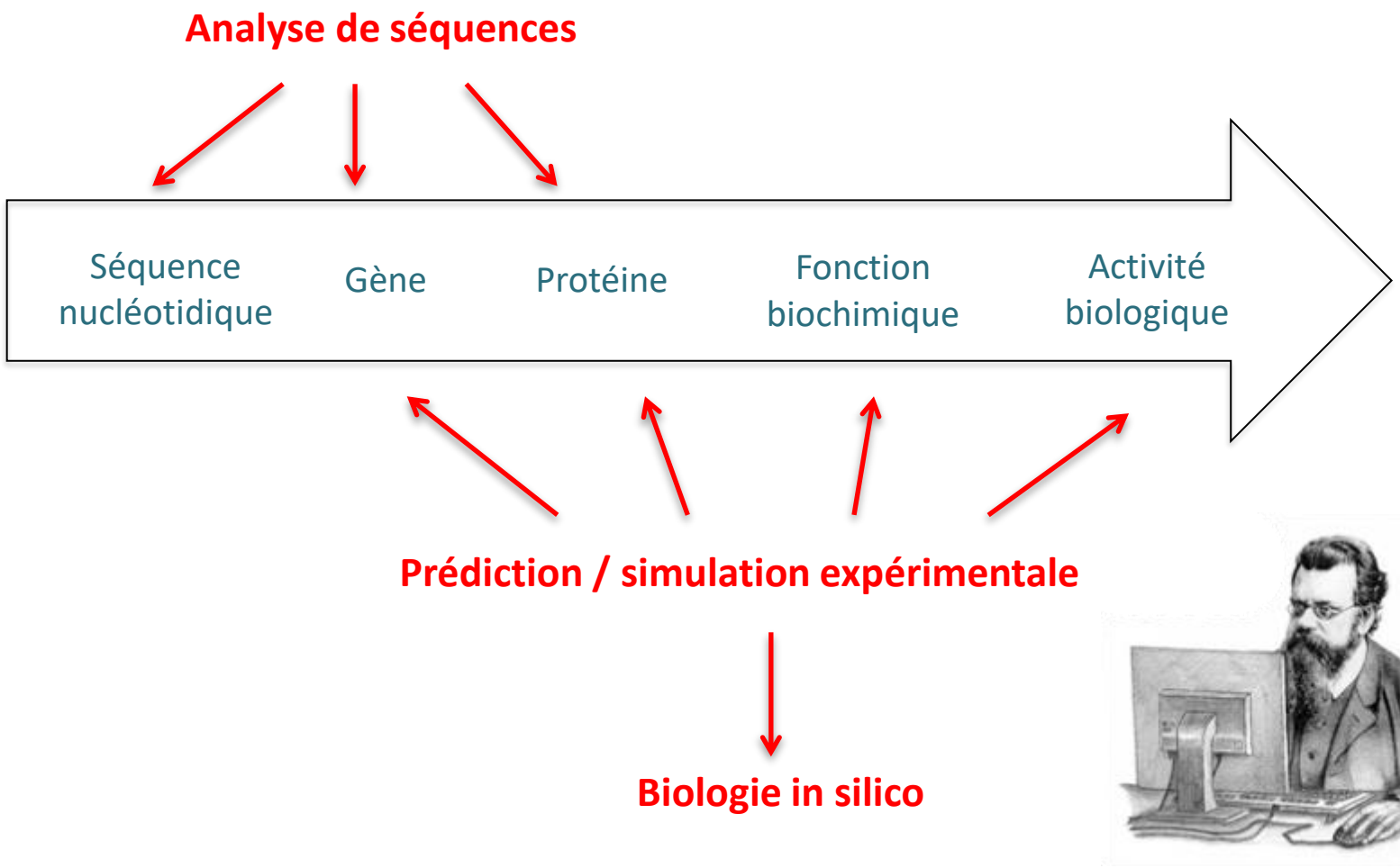
Prédiction physiologique et fonctionnelle

Objectifs :
-repérer un opéron
-repérer un gène ou une protéine anormale
-prévoir la structure 3D d'une protéine
-repérer des mutations
-prédire une pathologie…

Prédiction expérimentale

Objectifs :
-repérer des sites de restriction
-prévoir la digestion d'un nucléotide
-prévoir / simuler la migration de fragments nucléotidiques ou protéiques lors d'une électrophorèse…

La bioinformatique. Pourquoi ?



- La bioinformatique recouvre des champs de recherches très différents regroupant à la fois une forte utilisation de l'informatique et des mathématiques pour répondre à une question biologique.

  - Stockage et recherche d'informations (base de données de séquences, d'annotations, outils globaux de recherche…)
  - Analyse de séquences (alignements, phylogénie, recherche de motifs, analyse de génomes…)
  - Analyses omiques (génome, transcriptome, protéome…)
  - Modélisations d'interactions géniques
  - Modélisations moléculaires (Structure des protéines, des ARN)
  - Modélisations physiologiques (cellules, organes, organismes, populations…)
  - Analyse en imagerie médicale et microscopique
  - …

- Apport de l'informatique

Stockage et organisation des données

Permet de stocker par exemple les séquences des protéines et d'y associer différentes annotations : positions des domaines, des sites actifs, d'un pro-peptide, spécificité d'expression, rôle fonctionnel, associations à des pathologies….

Automatisation de tâches manuelles

Certaines tâches simples ne peuvent pas être réalisées à la main pour de nombreuses séquences (manque de temps, d'intérêt et risque d'erreurs) et sont donc automatisées (traduction, recherche de sites d'enzymes de restriction…)

Algorithme

Un algorithme est une suite finie et non-ambiguë d'instructions permettant de donner la réponse à un problème.

Cas particulier : les heuristiques

Puisque toutes les combinaisons ne peuvent être essayées dans un temps raisonnable, certains choix stratégiques doivent être faits (cf Blast)

- Apport des mathématiques

Statistiques

Permet d'évaluer des résultats entre eux en proposant des calculs de scores et de probabilités (p-value)
=> Aide l'interprétation

Modélisation

Permet de faire des prédictions à partir d'une mise en équation d'un système et des données biologiques

Séquençage d'ADN

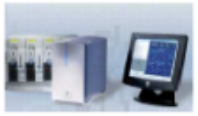
- 1977, Frédérick Sanger: Premier génome séquencé: Virus bactérien.
- 1995, J. Craig Venter: Premier génome bactérien: *H. Influenzae*
- 1996: Pemier génome eukaryote (levure *S. cerevisiae*).
- 1997: Bactérie E. coli, modèle important en microbiologie.
- 1998: Premier génome animal: le ver plat *C. elegans*
- 2000: Premier génome végétal: *A. Thaliana*; 1ère plante alimentaire: le riz.
- 2001: Génome humain …

Comment on obtient ces séquences?

Historique (rapide) des technologies….

Rappel

Découverte de la structure de l'ADN 1953 (Watson et Crick)



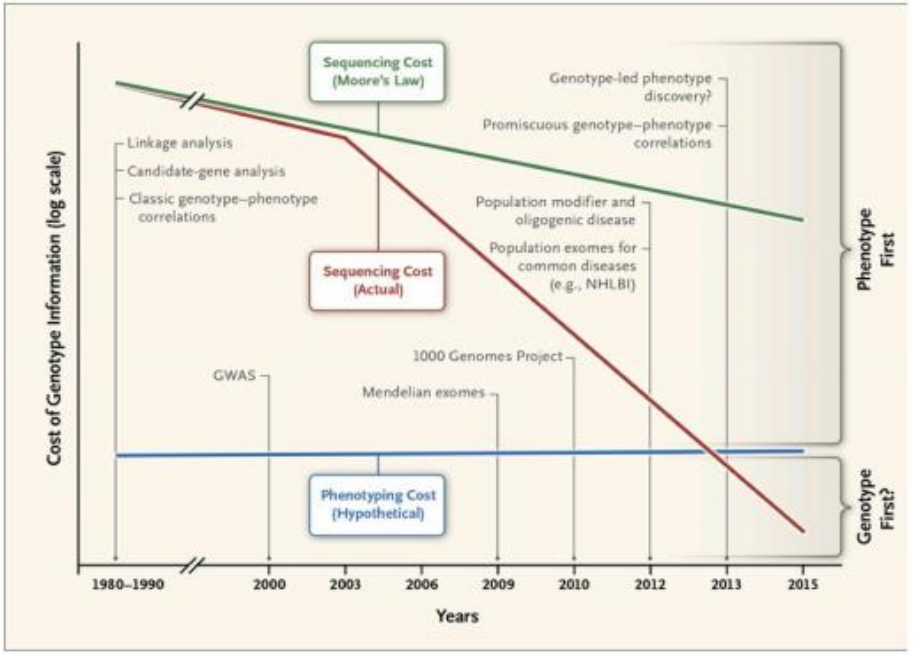
1975	1977	1990	1995	1999	2000	2007	2012
Southern Blot	Séquençage - Sanger - Gilbert	Séquençage par mesure de la fluorescence	Puce à ADN (microarray)		Séquenceur à capillaire	Séquençage à haut débit (NGS)	N-NGS

1ère GENERATION SEQUENCEUR	2ème GENERATION SEQUENCEUR	3ème GENERATION SEQUENCEUR

La révolution NGS (*next-generation sequencing*)

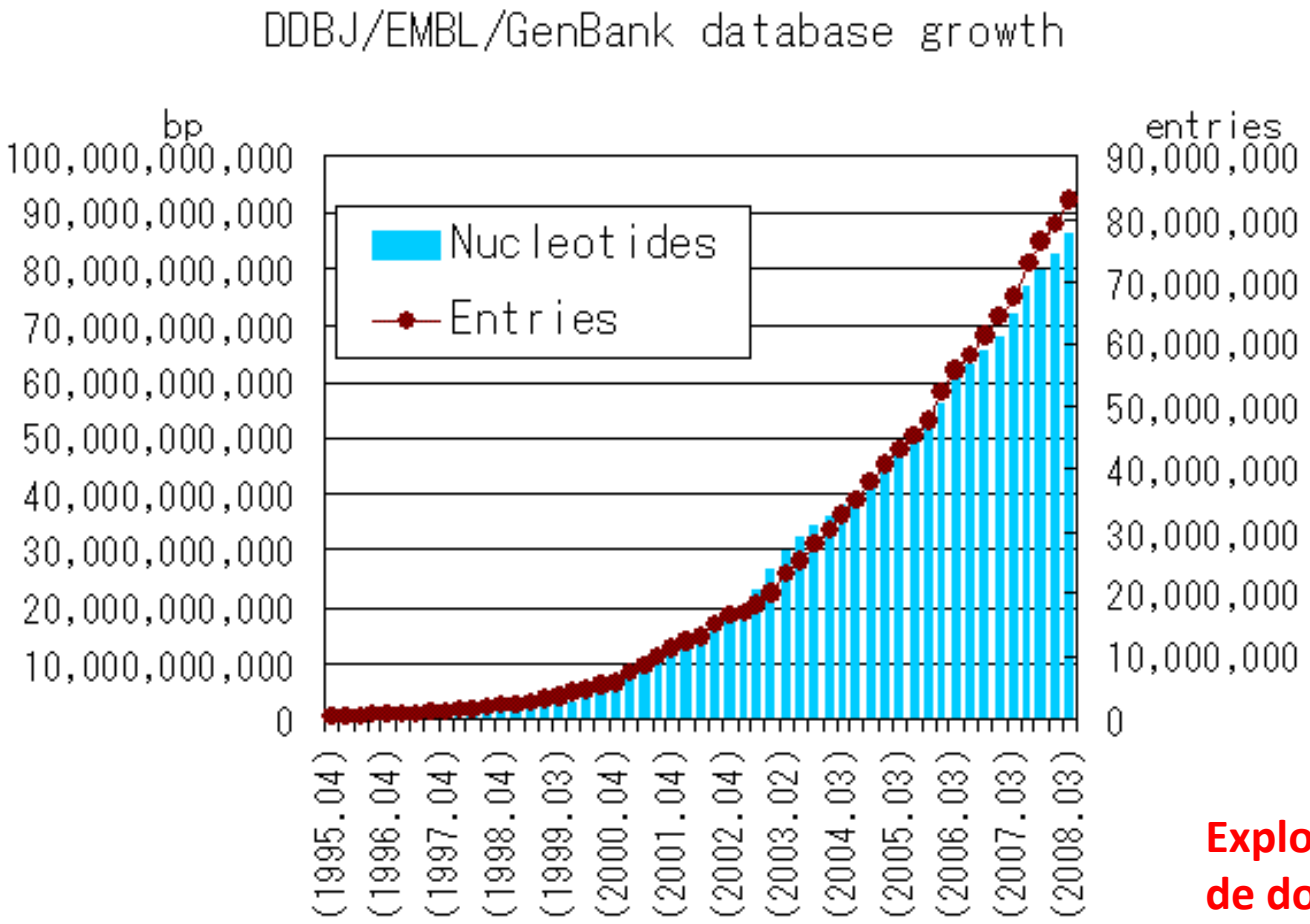
Le séquençage haut débit (HTS pour *high-throughput sequencing*) aussi appelé NGS pour *next-generation sequencing* désigne un ensemble de méthodes apparues à partir de 2005 produisant des millions de séquences en un *run* et à faibles coûts

Elles se caractérisent par l'utilisation d'approches massivement parallèles, permettant de séquencer des millions de fragments simultanément



Lu JT *et al. N Engl J Med* 2014,371:593-6

Défis de la biologie moléculaire



Explosion de la quantité de données biologiques nécessitant des outils de stockage adaptés

La Bio-Informatique s'applique à tout type de données biologiques, en particulier moléculaires :

- Les séquences d'ADN et de protéines
- Les structures d'ARN et de protéines
- Les contenus en gènes des génomes
- Les puces à ADN (microarrays)
- Les réseaux d'interactions entre protéines
- Les réseaux métaboliques
- Les arbres de phylogénie

Utilités :

- Faire avancer les connaissances en biologie, en génétique humaine, en théorie de l'évolution…
- Aider à la conception de médicaments
- Comprendre les maladies complexes..

Défis de la biologie moléculaire

- Décoder l'information contenue dans les séquences d'ADN et de protéines
  - Trouver les gènes
  - Différencier entre introns et exons
  - Analyser les répétitions dans l'ADN
  - Identifier les sites des facteurs de transcription
  - Étudier l'évolution des génomes
- Génomique structurale:
  - Modéliser les structures 3D des protéines et des ARN structurels
  - Déterminer la relation entre structure et fonction
- Génomique fonctionnelle
  - Étudier la régulation des gènes
  - Déterminer les réseaux d'interaction entre les protéines

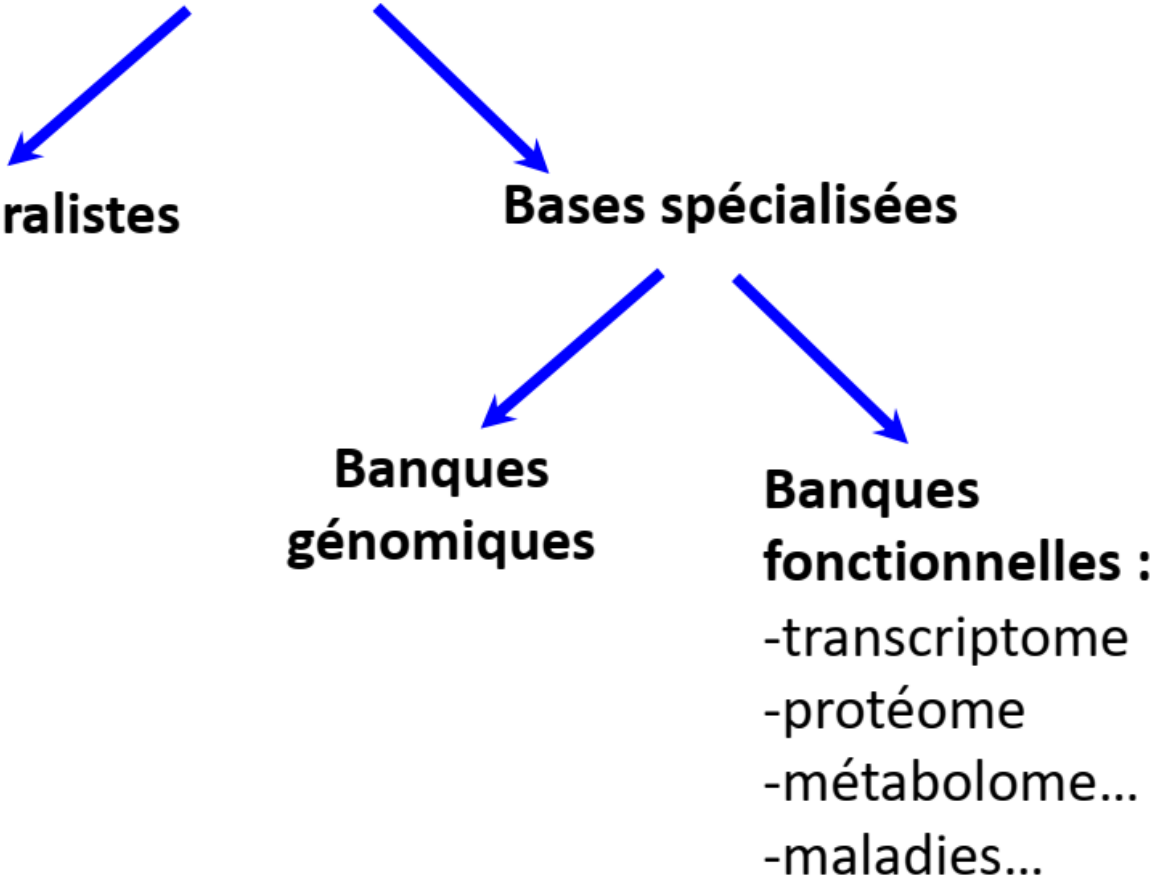
Les banques de données

Qu’est-ce qu’une banque de données?

- Ensemble de données relatives à un domaine, organisées par traitement informatique, accessibles en ligne et à distance
- Souvent, les données sont stockées sous la forme d’un fichier texte formaté (respectant une disposition particulière)
- Besoin de développer des logiciels spécifiques pour interroger les données contenues dans ces banques

Bases de données

- **Différentes catégories de bases de données :**



Les banques de données généralistes

- Ces banques contiennent des données hétérogènes
  - Collecte la plus exhaustive possible
  - Banques de séquences nucléiques
  - Banques de séquences protéiques
  - Banques de structure 3D de macromolécules
  - Banques d'articles scientifiques
- **Avantage :** tout est consultable en une fois
- **Inconvénients :** difficiles à maintenir, difficiles à interroger

Les banques de données spécialisées

- Ces banques contiennent des données homogènes
  - Collecte établie autour d'une thématique particulière
- **Avantages :** facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée, …
- **Inconvénients :** ne cible pas toujours ce que l'on veut; toutes les banques possibles n'existent pas
- **Exemples :** banques spécialisées pour un génome, banques de séquences d'immunologies, banques sur des séquences validées, …

Les bases de données bioinformatiques les plus utilisées:

- NCBI, *National Center for Biotechnology Information*
  - GenBank: Séquences d'ADN (3 billion de paires de bases)
  - Site officiel de BLAST
  - PubMed: Permet la recherche de références
  - COGs: Familles de gènes orthologues …
- EMBL, *The European Molecular Biology Laboratory*
- ExPASy, *Expert Protein Analysis System,* Protéomique
  - Swiss-Prot: Séquences de protéines
  - PROSITE: Domaines et familles de protéines
  - SWISS-MODEL: Outil de prédiction 3D de protéines
  - Différents outils de recherche
- PDB, *Protein Data Bank*
  - Base de données de structures 3D de protéines
  - Visualisation et manipulation de structures
- SCOP, *Structural Classification of Proteins*

Bases de données généralistes: publique

Pour collecter l'ensemble de séquences nucléiques:

- EMBL (Europe) devenue aujourd'hui l'ENA= European Nucleotide Archive
  → http://www.ebi.ac.uk/ena
- GenBank (USA) → http://www.ncbi.nlm.nih.gov/genbank/
- DDBJ (Japon) DNA DATA BANQUE OF JAPAN→ http://www.ddbj.nig.ac.jp

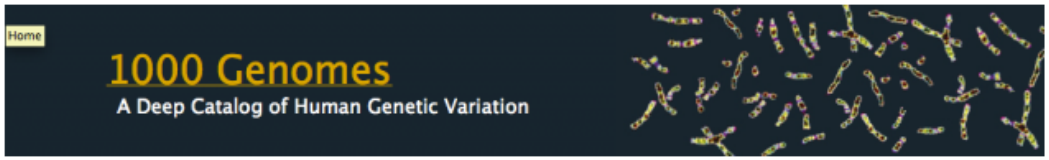
Permet d'accéder à des nombreuses données et aucun d'entre eux génère ces données

Soumettre des séquences → numéro d'accession

Bases de données généralistes: privé

- 1000 genomes project

http://www.1000genomes.org/



- http://www.jcvi.org/

J. Craig Venter®
INSTITUTE

Bases de données généralistes

Bases de données de séquences protéiques:

Les deux plus importantes :

- **SwissProt (1986) : banque manuellement annotée et «nettoyée »**
- **PIR/NBRF (1984) : banque américaine fournissant une classification des protéines basée sur la similarité entre les séquences.**

CONSORTIUM *Universal Protein Resource* → UniProt

http://www.uniprot.org

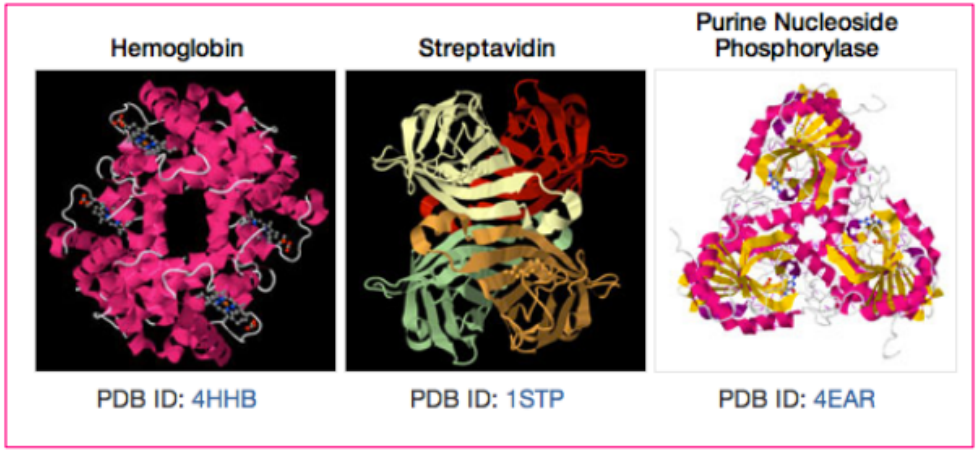
http://www.expasy.org

Bases de données généralistes

Bases de données de structure:

- La Protein Database (PDB) stockent les structures protéiques obtenues par RMN ou cristallographie
- Une entrée contient donc les coordonnées de tous les atomes de la structure

http://mm.rcsb.org



Bases de données spécialisées

- **Chaque année, en janvier, le journal Nucleic Acids Research publie un numéro spécial dédié aux bases de données « database »**

› Nucleic Acids Res. 2023 Jan 6;51(D1):D1-D8. doi: 10.1093/nar/gkac1186.

The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection

Daniel J Rigden [1], Xosé M Fernández [2]

Affiliations + expand

PMID: 36624667 PMCID: PMC9825711 DOI: 10.1093/nar/gkac1186

Free PMC article

Abstract

The 2023 Nucleic Acids Research Database Issue contains 178 papers ranging across biology and related fields. There are 90 papers reporting on new databases and 82 updates from resources previously published in the Issue. Six more papers are updates from databases most recently published elsewhere. Major nucleic acid databases reporting updates include Genbank, ENA, ChIPBase, JASPAR, mirDIP and the Issue's first Breakthrough Article, NACDDB for Circular Dichroism data. Updates from BMRB and RCSB cover experimental protein structural data while AlphaFold 2 computational structure predictions feature widely. STRING and REBASE are stand-out updates in the signalling and enzymes section. Immunology-related databases include CEDAR, the second Breakthrough Article, for cancer epitopes and receptors alongside returning IPD-IMGT/HLA and the new PGG.MHC. Genomics-related resources include Ensembl, GWAS Central and UCSC Genome Browser. Major returning databases for drugs and their targets include Open Targets, DrugCentral, CTD and Pubchem. The EMPIAR image archive appears in the Issue for the first time. The entire database Issue is freely available online on the Nucleic Acids Research website (https://academic.oup.com/nar). The NAR online Molecular Biology Database Collection has been updated, revisiting 463 entries, adding 92 new resources and eliminating 96 discontinued URLs so bringing the current total to 1764 databases. It is available at http://www.oxfordjournals.org/nar/database/c/.

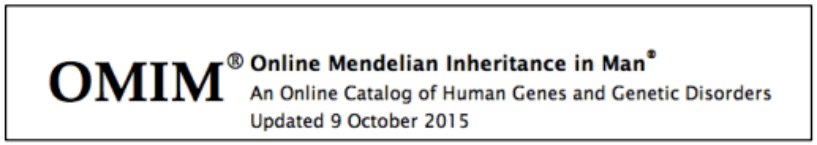
Bases de données spécialisée

Dédiées à un organisme :

- Flybase : Drosophile http://flybase.org
- HIV database : www.hiv.lanl.gov/
- Porteco: Escherichia coli http://www.porteco.org
- *Arabidopsis thaliana:* TAIR https://www.arabidopsis.org

Bases de données spécialisée

Dédiées à un type de séquences particulier :



- IMGT : données d'immunologie http://www.imgt.org
- EPD : Eukaryotic Promoter Database http://epd.vital-it.ch
- The European ribosomal RNA database http://bioinformatics.psb.ugent.be/webtools/rRNA
- Online Mendelian Inheritance in Man http://www.omim.org
- GOLD: Genome Online Database https://gold.jgi.doe.gov

Les banques de séquences

Les banques nucléiques

Les 3 banques nucléotidiques principales coexistent et coopèrent:

- **Elles collectent des informations de séquences (associées ou non à une publication) par soumission directe des auteurs (95% de l'ensemble des données) mais également par balayage systématique de la littérature scientifique (principalement les brevets).**

EMBL: Banque européenne créée en 1980 (Heidelberg, DE) et financée par l'EMBO (European Moleculary Biology Organisation), elle est aujourd'hui diffusée par l'EBI (European Bioinformatics Institute, Cambridge, GB)

Genbank: Créée en 1982 par la société IntelliGenetics (Los Alamos, US) et diffusée maintenant par le NCBI (National Center for Biotechnology Information, Bethesda, US)

DDBJ (DNA Data Bank of Japan) : Créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon).

Depuis 1987, ces banques échangent quotidiennement leurs fichiers afin de garantir dans chacune d'elles un ensemble de données le plus complet possible.

Chaque enregistrement ou « entrée » correspond à une séquence nucléique.

Organisation de l'information

Format général

- « flat file» ou fichier plat
- les banques sont distribuées sous forme de fichiers texte
- les données sont organisées séquentiellement

2 parties dans une fiche

- des informations relatives à la séquence (annotation)
- la séquence elle-même

Les champs

- ils facilitent l'accès à l'information
- chaque champ regroupe des informations de même type

Les séquences biologiques sont souvent:

-redondantes

-dispersées dans différentes banques de données

-ont des nomenclatures diverses et variées (synonymes)

Pour **identifier** ces séquences, les différentes banques de données leur assignent des **Numéros d'Accession** uniques au sein de leurs collections respectives. Pour pointer sans ambiguité sur un tel objet, on utilise la notation:

« Défauts» des banques nucléiques

Aucun contrôle des banques:

- *les auteurs sont responsables de la qualité des séquences soumises.*

Hétérogénéité:

- *ADN nucléaire, mitochondrial, chloroplastique, ARNm, ARNt, ARNs, ARNr, chromosomes entiers ...*
- *gènes, fragments … (10 bp à 350000 bp)*

Variabilité de l'état des connaissances sur les séquences:

- *Annotation effectuée ou non*
- *Annotation hétérogènes: automatique ou expérimentale*

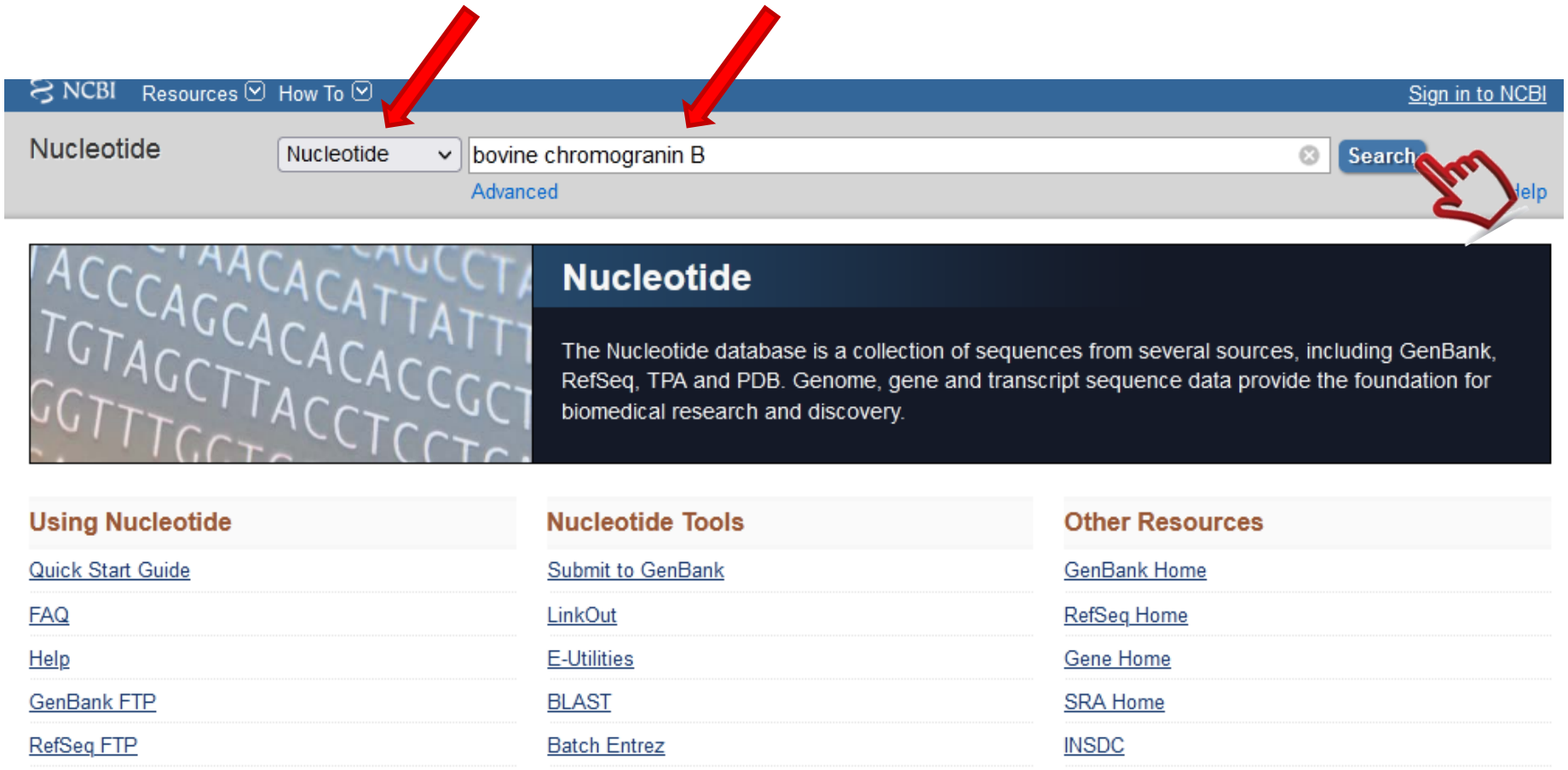
Erreurs dans les séquences (qualité inégale):

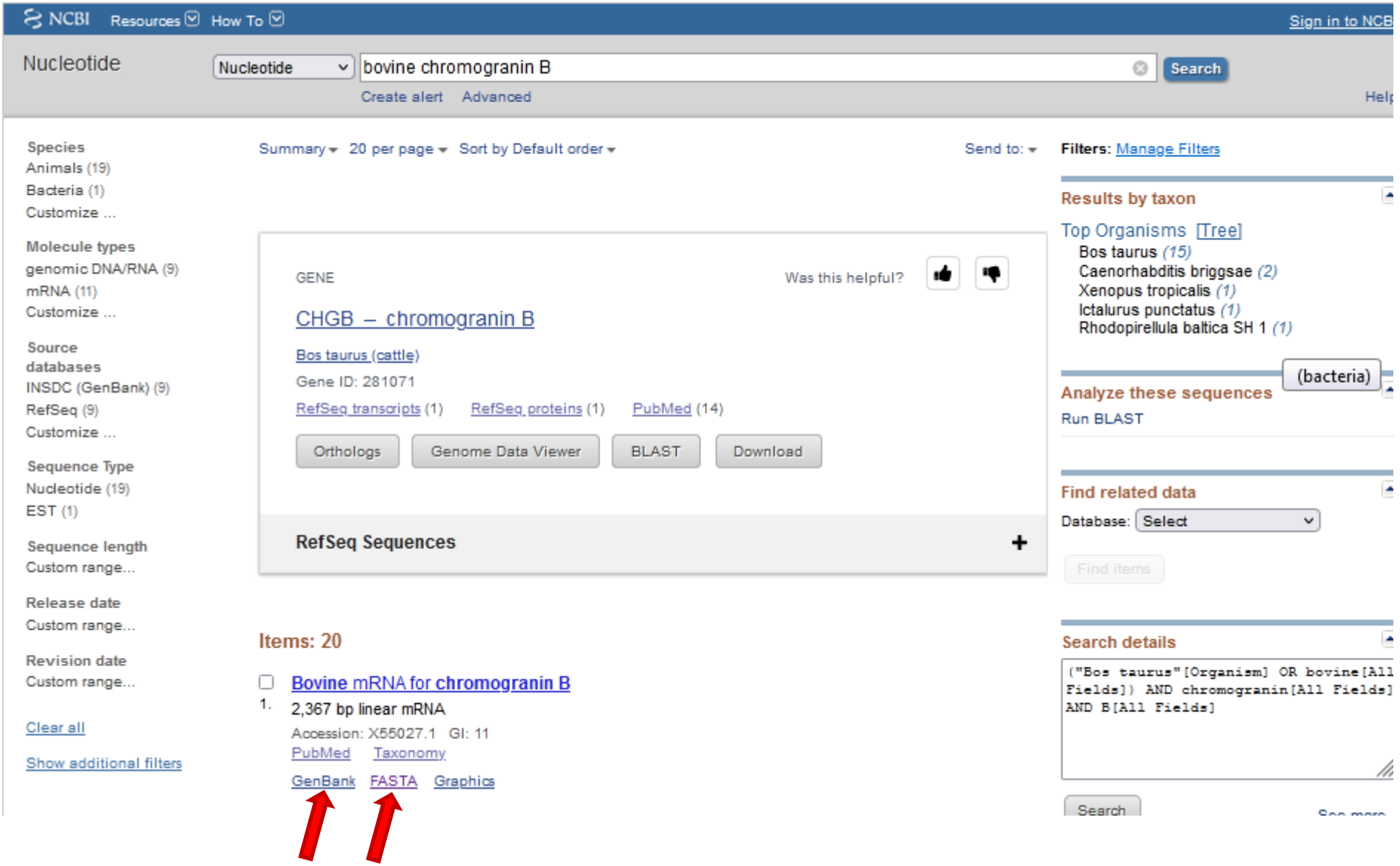
- *origine du fragment*
- *cultures infectés*
- *présence de séquences de vecteurs de clonage*
- *erreurs de saisie*

Redondance des données: plusieurs entrées pour une même séquence

- *Certains gènes sont séquencés à la fois sous forme d'ARNm et de fragments génomiques.*
- *Certaines séquences ont été saisies plusieurs fois dans la banque.*

- Allez dans «NCBI»
- Entrer la recherche suivante dans la base de donnée «Nucleotide»: **Bovine chromogranin B**


NCBI
Resources
How To
Sign in to NCBI
Nucleotide
Nucleotide
bovine chromogranin B
Search
Advanced
Help
Nucleotide
The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.
Using Nucleotide
Quick Start Guide
FAQ
Help
GenBank FTP
RefSeq FTP
Nucleotide Tools
Submit to GenBank
LinkOut
E-Utilities
BLAST
Batch Entrez
Other Resources
GenBank Home
RefSeq Home
Gene Home
SRA Home
INSDC


NCBI Resources How To
Sign in to NCB
Nucleotide
Nucleotide
bovine chromogranin B
Search
Create alert Advanced
Help
Species
Animals (19)
Bacteria (1)
Customize ...
Molecule types
genomic DNA/RNA (9)
mRNA (11)
Customize ...
Source databases
INSDC (GenBank) (9)
RefSeq (9)
Customize ...
Sequence Type
Nucleotide (19)
EST (1)
Sequence length
Custom range...
Release date
Custom range...
Revision date
Custom range...
Clear all
Show additional filters
Summary
20 per page
Sort by Default order
Send to:
Filters: Manage Filters
GENE
Was this helpful?
CHGB – chromogranin B
Bos taurus (cattle)
Gene ID: 281071
RefSeq transcripts (1)
RefSeq proteins (1)
PubMed (14)
Orthologs
Genome Data Viewer
BLAST
Download
RefSeq Sequences
Items: 20
Bovine mRNA for chromogranin B
1.
2,367 bp linear mRNA
Accession: X55027.1 GI: 11
PubMed
Taxonomy
GenBank
FASTA
Graphics
Results by taxon
Top Organisms [Tree]
Bos taurus (15)
Caenorhabditis briggsae (2)
Xenopus tropicalis (1)
Ictalurus punctatus (1)
Rhodopirellula baltica SH 1 (1)
(bacteria)
Analyze these sequences
Run BLAST
Find related data
Database: Select
Find items
Search details
("Bos taurus"[Organism] OR bovine[All Fields]) AND chromogranin[All Fields] AND B[All Fields]
Search

Les champs d'une fiche: GENBANK

Identifiant

Numéro d'accession

Organisme

Classification taxonomique

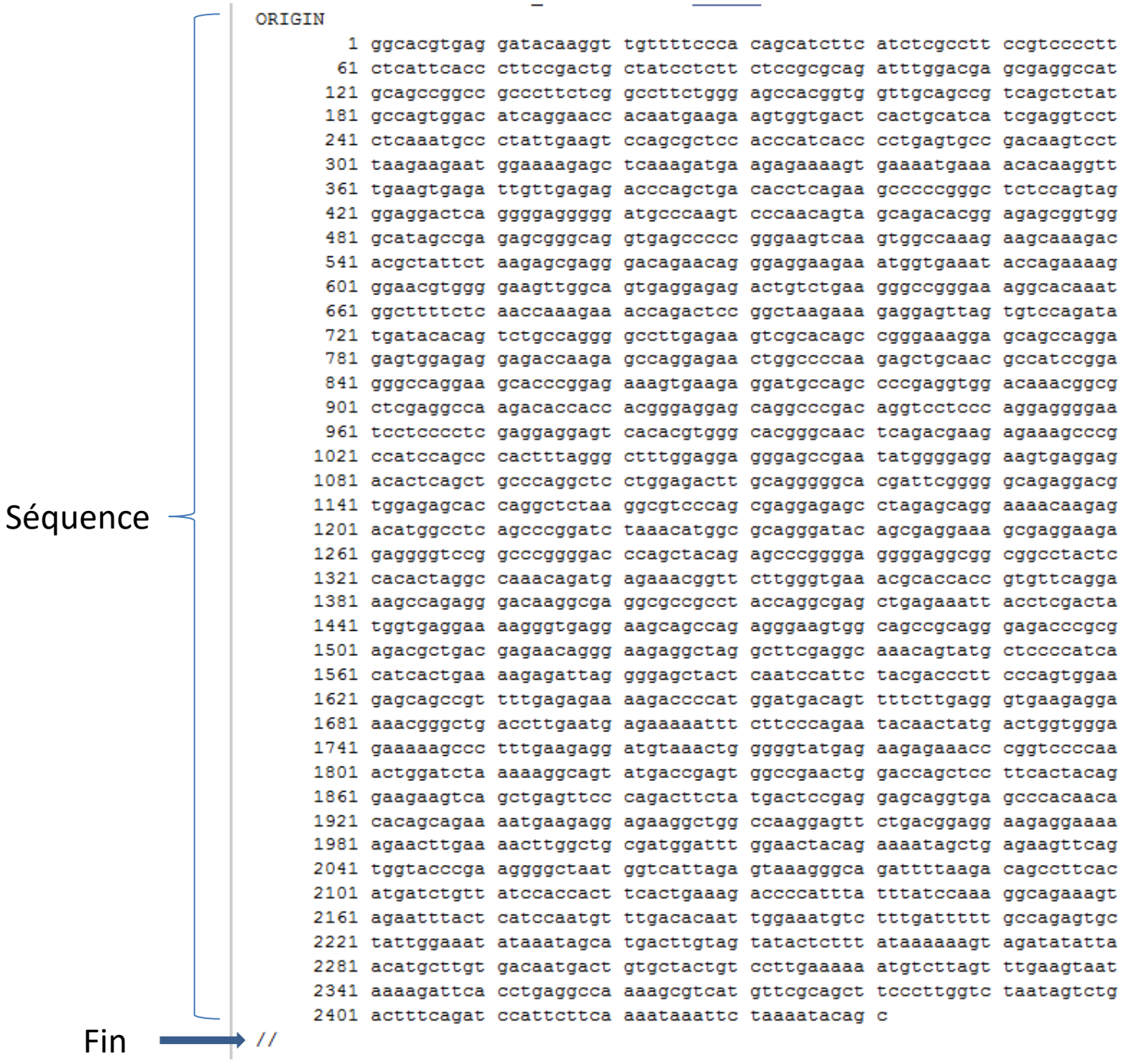
Références

Caractéristiques

Nature

```
LOCUS       NM_181006               2441 bp    mRNA    linear   MAM 11-SEP-2013
DEFINITION  Bos taurus chromogranin B (secretogranin 1) (CHGB), mRNA.
ACCESSION   NM_181006
VERSION     NM_181006.2  GI:31341071
KEYWORDS    RefSeq.
SOURCE      Bos taurus (cattle)
  ORGANISM  Bos taurus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Ruminantia;
            Pecora; Bovidae; Bovinae; Bos.
REFERENCE   1  (bases 1 to 2441)
  AUTHORS   Zimin,A.V., Delcher,A.L., Florea,L., Kelley,D.R., Schatz,M.C.,
            Puiu,D., Hanrahan,F., Pertea,G., Van Tassell,C.P., Sonstegard,T.S.,
            Marcais,G., Roberts,M., Subramanian,P., Yorke,J.A. and
            Salzberg,S.L.
  TITLE     A whole-genome assembly of the domestic cow, Bos taurus
  JOURNAL   Genome Biol. 10 (4), R42 (2009)
FEATURES             Location/Qualifiers
     source          1..2441
                     /organism="Bos taurus"
                     /mol_type="mRNA"
                     /db_xref="taxon:9913"
                     /chromosome="13"
                     /map="13"
     gene            1..2441
                     /gene="CHGB"
                     /note="chromogranin B (secretogranin 1)"
                     /db_xref="GeneID:281071"
     CDS             119..2059
                     /gene="CHGB"
                     /function="calcium binding protein"
                     /note="secretogranin-1; cgB; sgI; chromogranin-B;
                     secretogranin I"
                     /codon_start=1
                     /product="secretogranin-1 precursor"
                     /protein_id="NP_851349.1"
                     /db_xref="GI:30794308"
                     /db_xref="GeneID:281071"
```

```
                     /db_xref="GeneID:281071"
     CDS             119..2059
                     /gene="CHGB"
                     /function="calcium binding protein"
                     /note="secretogranin-1; cgB; sgI; chromogranin-B;
                     secretogranin I"
                     /codon_start=1
                     /product="secretogranin-1 precursor"
                     /protein_id="NP_851349.1"
                     /db_xref="GI:30794308"
                     /db_xref="GeneID:281071"
                     /translation="MQPAALLGLLGATVVAAVSSMPVDIRNHNEEVVTHCIIEVLSNA
                     LLKSSAPPITPECRQVLKKNGKELKDEEKSENENTRFEVRLLRDPADTSEAPGLSSRE
                     DSGEGDAQVPTVADTESGGHSRERAGEPPGSQVAKEAKTRYSKSEGQNREEEMVKYQK
                     RERGEVGSEERLSEGPGKAQMAFLNQRNQTPAKKEELVSRYDTQSARGLEKSHSRERS
                     SQESGEETKSQENWPQELQRHPEGQEAPGESEEDASPEVDKRRSRPRHHHGRSRPDRS
                     SQEGNPPLEEESHVGTGNSDEEKARHPAHFRALEEGAEYGEEVRRHSAAQAPGDLQGA
                     RFGGRGRGEHQALRRPSEESLEQENKRHGLSPDLNMAQGYSEESEEERGPARGPSYRA
                     RGGEAAAYSTLGQTDEKRFLGETHHRVQESQRDKARRRLPGELRNYLDYGEEKGEEAA
                     RGKWQPQGDPRDADENREEARLRGKQYAPHHITEKRLGELLNPFYDPSQWKSSRFERK
                     DPMDDSFLEGEEENGLTLNEKNFFPEYNYDWWEKKPFEEDVNWGYEKRNPVPKLDLKR
                     QYDRVAELDQLLHYRKKSAEFPDFYDSEEQVSPQHTAENEEEKAGQGVLTEEEEKELE
                     NLAAMDLELQKIAEKFSGTRRG"
     sig_peptide     119..178
                     /gene="CHGB"
                     /inference="COORDINATES: ab initio prediction:SignalP:4.0"
     mat_peptide     179..2056
                     /gene="CHGB"
                     /product="Secretogranin-1"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="propagated from UniProtKB/Swiss-Prot (P23389.2)"
     misc_feature    599..604
                     /gene="CHGB"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Cleavage, major site; propagated from
                     UniProtKB/Swiss-Prot (P23389.2); cleavage site"
```




```
ORIGIN
        1 ggcacgtgag gatacaaggt tgttttccca cagcatcttc atctcgcctt ccgtcccctt
       61 ctcattcacc cttccgactg ctatcctctt ctccgcgcag atttggacga gcgaggccat
      121 gcagccggcc gcccttctcg gccttctggg agccacggtg gttgcagccg tcagctctat
      181 gccagtggac atcaggaacc acaatgaaga agtggtgact cactgcatca tcgaggtcct
      241 ctcaaatgcc ctattgaagt ccagcgctcc acccatcacc cctgagtgcc gacaagtcct
      301 taagaagaat ggaaaagagc tcaaagatga agagaaaagt gaaaatgaaa acacaaggtt
      361 tgaagtgaga ttgttgagag acccagctga cacctcagaa gccccccggc tctccagtag
      421 ggaggactca ggggaggggg atgcccaagt cccaacagta gcagacacgg agagcggtgg
      481 gcatagccga gagcgggcag gtgagccccc gggaagtcaa gtggccaaag aagcaaagac
      541 acgctattct aagagcgagg gacagaacag ggaggaagaa atggtgaaat accagaaaag
      601 ggaacgtggg gaagttggca gtgaggagag actgtctgaa gggccgggaa aggcacaaat
      661 ggcttttctc aaccaaagaa accagactcc ggctaagaaa gaggagttag tgtccagata
      721 tgatacacag tctgccaggg gccttgagaa gtcgcacagc cgggaaagga gcagccagga
      781 gagtggagag gagaccaaga gccaggagaa ctggccccaa gagctgcaac gccatccgga
      841 gggccaggaa gcacccggag aaagtgaaga ggatgccagc cccgaggtgg acaaacggcg
      901 ctcgaggcca agacaccacc acgggaggag caggcccgac aggtcctccc aggaggggaa
      961 tcctcccctc gaggaggagt cacacgtggg cacgggcaac tcagacgaag agaaagcccg
     1021 ccatccagcc cactttaggg ctttggagga gggagccgaa tatggggagg aagtgaggag
     1081 acactcagct gcccaggctc ctggagactt gcaggggggca cgattcgggg gcagaggacg
     1141 tggagagcac caggctctaa ggcgtcccag cgaggagagc ctagagcagg aaaacaagag
     1201 acatggcctc agcccggatc taaacatggc gcagggatac agcgaggaaa gcgaggaaga
     1261 gaggggtccg gcccggggac ccagctacag agcccgggga ggggaggcgg cggcctactc
     1321 cacactaggc caaacagatg agaaacggtt cttgggtgaa acgcaccacc gtgttcagga
     1381 aagccagagg gacaaggcga ggcgccgcct accaggcgag ctgagaaatt acctcgacta
     1441 tggtgaggaa aagggtgagg aagcagccag agggaagtgg cagccgcagg gagacccgcg
     1501 agacgctgac gagaacaggg aagaggctag gcttcgaggc aaacagtatg ctccccatca
     1561 catcactgaa aagagattag gggagctact caatccattc tacgaccctt cccagtggaa
     1621 gagcagccgt tttgagagaa aagaccccat ggatgacagt tttcttgagg gtgaagagga
     1681 aaacgggctg accttgaatg agaaaaattt cttcccagaa tacaactatg actggtggga
     1741 gaaaaagccc tttgaagagg atgtaaactg ggggtatgag aagagaaacc cggtccccaa
     1801 actggatcta aaaaggcagt atgaccgagt ggccgaactg gaccagctcc ttcactacag
     1861 gaagaagtca gctgagttcc cagacttcta tgactccgag gagcaggtga gcccacaaca
     1921 cacagcagaa aatgaagagg agaaggctgg ccaaggagtt ctgacggagg aagaggaaaa
     1981 agaacttgaa aacttggctg cgatggattt ggaactacag aaaatagctg agaagttcag
     2041 tggtacccga aggggctaat ggtcattaga gtaaagggca gattttaaga cagccttcac
     2101 atgatctgtt atccaccact tcactgaaag accccattta tttatccaaa ggcagaaagt
     2161 agaatttact catccaatgt ttgacacaat tggaaatgtc tttgattttt gccagagtgc
     2221 tattggaaat ataaatagca tgacttgtag tatactcttt ataaaaaagt agatatatta
     2281 acatgcttgt gacaatgact gtgctactgt ccttgaaaaa atgtcttagt ttgaagtaat
     2341 aaaagattca cctgaggcca aaagcgtcat gttcgcagct tcccttggtc taatagtctg
     2401 actttcagat ccattcttca aaataaattc taaaatacag c
//
```

Amino acid codes		
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamine or Glutamic acid
X	Xaa	Any amino acid

▪ Codes UIPAC

Nucleotide ambiguity code

Nucleic acid codes	
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base