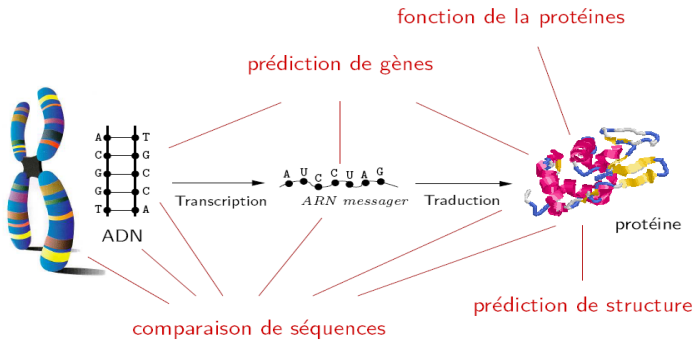
1- Définition

L'utilisation de l'outil informatique pour **traiter** les données de la biologie moléculaire.
La bioinformatique est l'application des **techniques de l'information** à la gestion des **données biologiques**.
La bioinformatique est la discipline de **l'analyse de l'information biologique**, essentiellement sous la forme de séquences nucléotidiques, de séquences d'acides aminés et de structures de protéines.
La bioinformatique n'est pas une dérivée de l'informatique : elle n'est qu'utilisatrice des ordinateurs et de leurs langages. Le suffixe "informatique" doit donc être compris comme renvoyant à l'interprétation de l'information biologique, et non pas à l'utilisation de l'ordinateur.



Comment Utiliser la bioinformatique ?
La Bioinformatique ne va pas remplacer le travail de laboratoire. La preuve expérimentale est toujours la "**règle d'or**". La bioinformatique doit être employée pour aider "à focaliser" les expériences du biologiste.
Un but est d'éliminer les fausses pistes qui conduisent à gaspiller du temps et de l'argent.

2- Description

- C'est une discipline **récente** (quelques dizaines d'années).
- C'est une discipline **"hybride"**: elle est fondée sur des concepts et des formalismes issus de la biologie, de l'informatique, des mathématiques et de la physique.

- C'est une discipline qui utilise toutes les potentialités de traitement de l'informatique: modèles théoriques, algorithmes et programmes, ordinateurs, réseau Internet, bases de données ...

3- Les bases de données

3.1-Définition

Ensemble structuré de données accessibles au moyen d'un logiciel.

- Une collection de données, structurée, indexée (table de matières), périodiquement mise à jour et contenant des références croisées avec d'autres banques
- Comporte des outils associés (logiciels) pour: accéder à la banque, mettre à jour la banque,….

-Rôle des banques/bases de données
-Collecter les informations (séquences, cartographie physique, génétique…, données structurales, relationnelles…, - *auprès de*: biologistes, littératures, autres bases de données)
-Stocker et organiser
-Distribuer l'information
-Faciliter l'exploitation
Il existe un grand nombre de bases de données d'intérêt biologique. D'une façon générale, on distingue:

- les bases de données généralistes
- les bases de données spécialisées.

3.2-Bases généralistes

Les bases de données généralistes de séquences nucléotidiques et protéiques couvrent tous les secteurs de la biologie et toutes les espèces.
Les grandes bases de séquences généralistes: Genbank, l'EMBL et DDBJ.Elles sont maintenant devenues indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique. Leur principale mission est de rendre

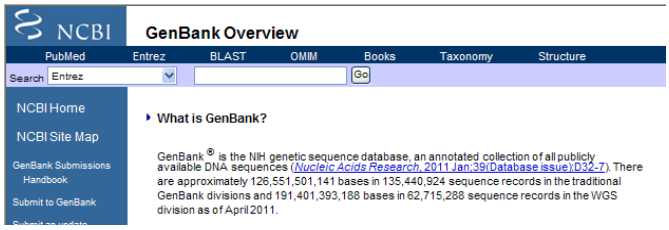
publiques les séquences qui ont été déterminées, ainsi un des premiers intérêts de ces banques est la masse de séquences qu'elles contiennent.

Depuis 1987, une convention est établie entre les trois bases nucléotidiques (EMBL, Genbank et DDBJ) permettant les échanges de séquences soumises à l'une ou à l'autre. Aujourd'hui, quelque soit la banque utilisée pour ses requêtes, les résultats obtenus doivent être les mêmes.

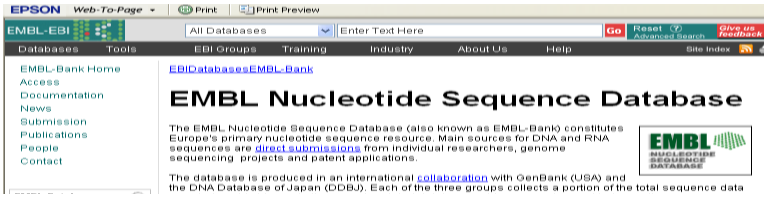
A l'exception des quelques faiblesses présentées par ces banques généralistes (manque de vérification des données soumises, retard dans l'insertion de nouvelles séquences, forte redondance «un même fragment de séquence présent dans plusieurs entrées»…), les qualités sont remarquables:

- Enorme richesse de séquences en un seul ensemble ;
- Grande diversité d'organismes ;
- Nombreuses informations qui accompagnent les séquences (annotations, expertise, bibliographie) ;
- Présence de lien vers d'autres bases de données (spécialisées), soit nucléique, soit (encore mieux) protéique.

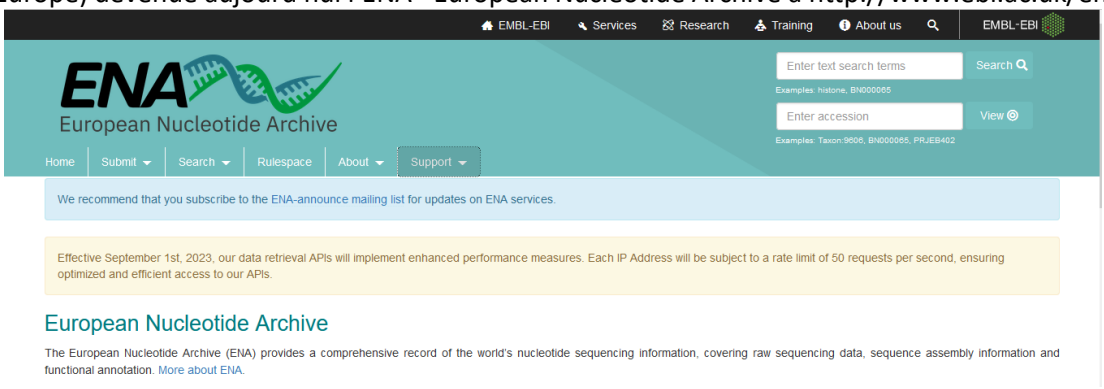
GenBank créée en 1982 au Los Alamos National Laboratory, est maintenue au NCBI (National Center for Biotechnology information), qui dépend du NIH (*National Institute of Health*) américain.



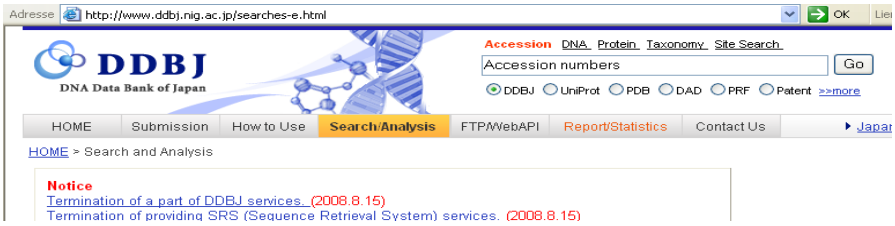
EMBL créée en 1980 par l'EMBL (European Molecular Biology Laboratory), est maintenue à l'EBI (European Bioinformatics Institute);



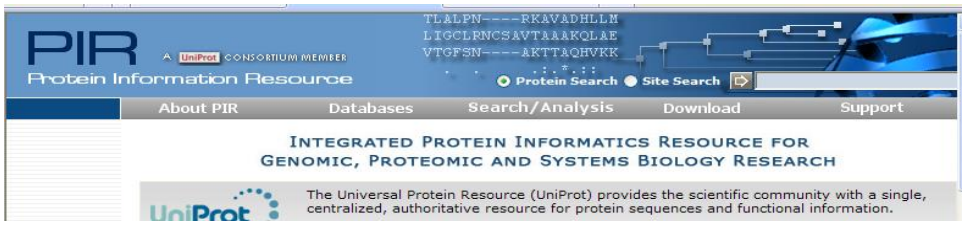
EMBL (Europe) devenue aujourd'hui l'ENA= European Nucleotide Archive à http://www.ebi.ac.uk/ena



DDBJ (DNA Data Bank of Japan) maintenue par le Centre d'Information Biologique de l'Institut National de Génétique, créée en 1986.

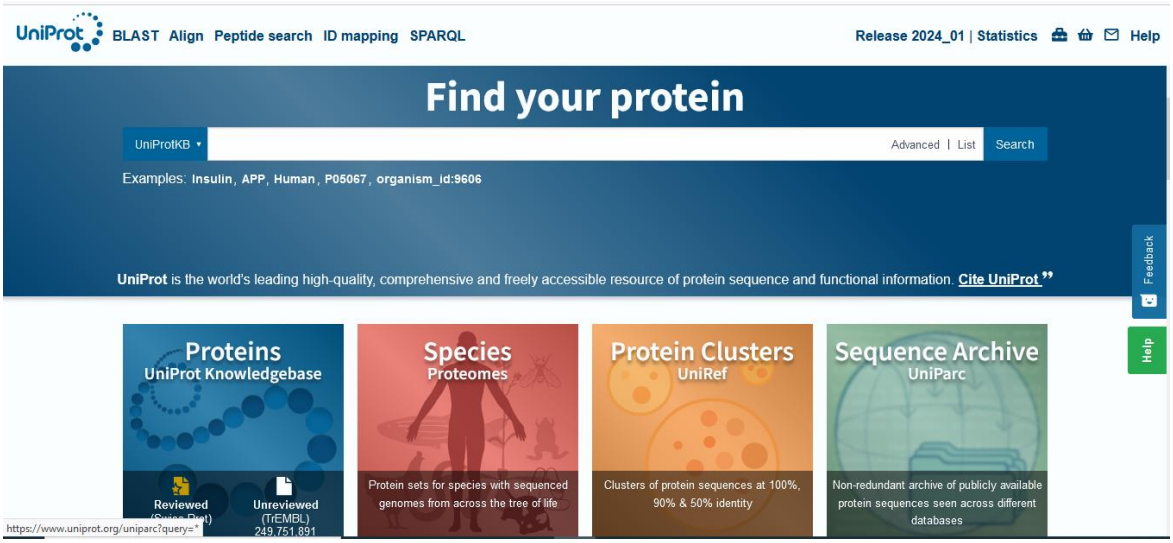


PIR-NBRF : D'abord, elle fut créée par la NBRF (National Biomedical Researche Foundation) en 1984. Actuellement, elle constitut un ensemble du à la fusion de MIPS (Martinsried Institute for Protein sequences, Munich Allemagne) et de JIPID Japan International Protein Information Database).



SwissProt : Créée par le biochimiste Amos BAIROCH en 1986 à l'Université de Genève. Actuellement développée en collaboration entre l'Institut Suisse de BioInformatique (ISB-SIB) et l'EBI. Elle contient la séquence de quasiment toutes les protéines découvertes jusque là.

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (**UniProtKB**), the UniProt Reference Clusters (**UniRef**), and the UniProt Archive (**UniParc**). The UniProt consortium and host institutions EMBL-EBI, SIB and PIR are committed to the long-term preservation of the UniProt databases. In 2002 the three institutes decided to pool their resources and expertise and formed the UniProt consortium.

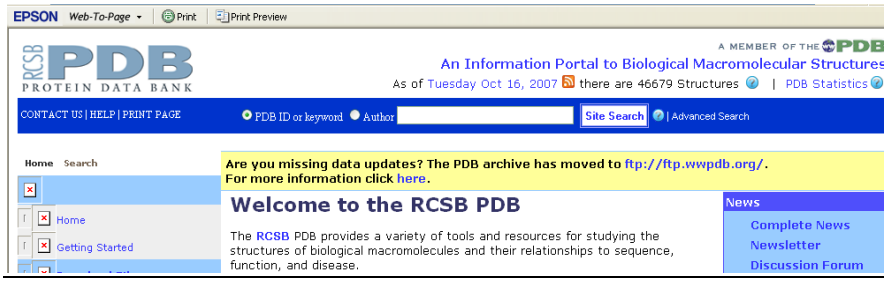


3.3-Bases spécialisées

Pour des **besoins spécifiques** liés à l'activité d'un groupe de personnes, ou encore par compilations bibliographiques, de nombreuses bases de données spécifiques ont été créées au sein des laboratoires.
Elles ont pour but:
- de recenser des familles de séquences autour de caractéristiques biologiques précises comme les signaux de régulation, les promoteurs de gènes, les signatures peptidiques ou les gènes identiques issus d'espèces différentes.
- de regrouper des classes spécifiques de séquences comme les vecteurs de clonage, les enzymes de restriction, ou toutes les séquences d'un même génome.

Quelques exemples de bases de données spécialisées:

PDB, protein DataBank (PDB): Exemple: regroupe les molécules pour lesquelles on a obtenu les coordonnées 3D par résonance magnétique ou diffraction aux rayons X. Ces structures peuvent être facilement visualisées à l'aide de logiciels de visualisation 3D.



ECD, base sur les séquences nucléiques d'*Escherichia coli*.
NRL3D, base de séquences protéiques dont la structure tridimensionnelle a été déterminée.
TFD, base de facteurs de transcription.
Prosite, bases de motifs protéiques. Elle peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique.
CATH, base sur les classifications hiérarchiques (ordonnées) des structures protéiques.
IMGT, base de séquences des immunoglobulines et des récepteurs T.
GENATLAS, base d'informations issues de la cartographie des gènes humains.
KEGG, bases de voies métaboliques.

Les bases de motifs

On sait que certains segments d'ADN ou de protéines sont déterminants dans l'analyse des séquences car ils correspondent à des sites précis d'activité biologique comme par exemple les éléments de régulation des gènes ou les signatures peptidiques. C'est pourquoi des bases spécialisées se sont naturellement constituées autour de ces séquences.
L'utilisation des bases spécialisées comme les bases de motifs, est devenue un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéines inconnues ou savoir à quelle famille appartient une séquence non encore caractérisée.

- la base **TFD** est employée pour des séquences promotrices des gènes
- **Prosite** ou **BLOCKS** sont utilisées pour des protéines inconnues ou bien des séquences protéiques traduites à partir de cDNA ou de séquences génomiques.

Pour détecter une fonctionnalité sur une séquence, il suffit d'exécuter un programme qui s'appliquera à repérer la présence de certains motifs recensés dans ces bases et ainsi à prédire l'appartenance de la séquence testée à un groupe de séquences ayant une signature commune.

La base sur les facteurs de transcription TFD

Facteur de transcription: Protéine qui régule la transcription d'un gène en se fixant sur son promoteur au niveau d'un motif nucléique ou sites de liaison (binding sites).
TFD est une base dédiée aux facteurs de transcription eucaryotes. Une partie des données a été extraite de GenBank et une autre partie provient de synthèses bibliographiques réalisées à partir de publications traitant de différents aspects de la transcription.

La base est organisée en plusieurs fichiers permettant de regrouper différentes classes d'information que l'on connaît au niveau de la transcription. Ainsi cette base renferme non seulement des données nucléiques mais aussi des informations sur les séquences protéiques directement impliquées dans la transcription comme les domaines protéiques interagissant avec l'ADN ou les cofacteurs de transcription.

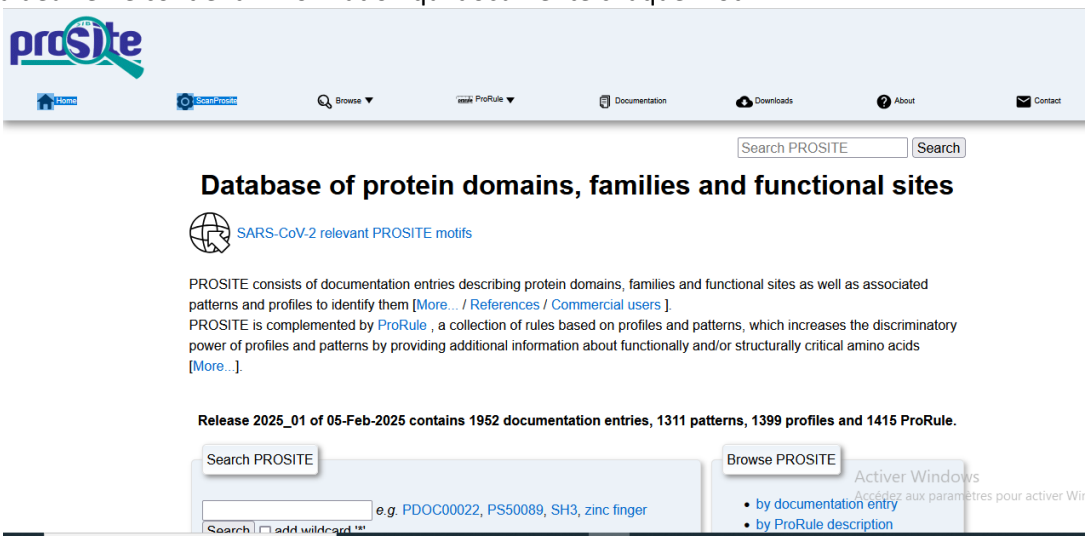
La base de motifs protéiques Prosite

La base PROSITE peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique. Elle est établie en regroupant, quand cela est possible, les protéines contenues dans Swissprot par famille comme par exemple les kinases ou les protéases. On recherche ensuite, au sein de ces groupes, des motifs consensus susceptibles de les caractériser spécifiquement.

La conception de la base repose sur quatre critères essentiels:

- collecter le plus possible de motifs significatifs,
- avoir des motifs hautement spécifiques pour caractériser au mieux une famille de protéines,
- donner une documentation complète sur chacun des motifs répertoriés, et
- faire une révision périodique des motifs pour s'assurer de leur validité par rapport aux dernières expérimentations.

La base est organisée en deux parties: la première contient l'identification et la description de chaque motif. La deuxième contient l'information qui documente chaque motif.



Sélection de banques de données

1-Banques de données moléculaires :

Séquences primaires:

EMBL = **GENBANK** = **DDBJ** (toutes séquences nucléiques expérimentales) :
http://www.ebi.ac.uk/embl/Contact/collaboration.html
TREMBL (traduction automatique de EMBL en séquences protéiques) :
http://www.expasy.ch/sprot/sprot-top.html
dbEST (EST - *Expressed Sequence Tags* - marqueurs d'expression) :

Séquences secondaires (dérivées des archives primaires):
SWISSPROT (séquences protéiques annotées manuellement par des biologistes) :
http://www.expasy.ch/sprot/sprot-top.html
EPD (*Eukaryotic Promoter Database*) : http://www.epd.isb-sib.ch/
UNIGENE (clusters d'EST, basés sur dbEST):
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene

Structures 3D des protéines:
PDB (Protein Data Bank, coordonnées 3D atomes de cristallographie)

2-Banques de connaissances biologiques :

Bibliographiques :

PubMed (Medline, tous les abstracts d'articles en biologie) :
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
OMIM (*Online Mendelian Inheritance in Man*, maladies génétiques de l'Homme) :
http://www3.ncbi.nlm.nih.gov/Omim/

Relationnelles et fonctionnelles (classification, structures, taxonomie etc.):
Taxonomy (phylogénie des espèces) :
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy
CATH (*Class, Architecture, Topology and Homologous superfamily*) :
http://www.biochem.ucl.ac.uk/bsm/cath_new/
SCOP (*Structural Classification of Proteins*) : http://scop.mrc-lmb.cam.ac.uk/scop
KEGG (*Kyoto Encyclopedia of Genes and Genomes*, voies et interactions) :
http://www.genome.ad.jp/kegg/

Organismes:
GDB (*Genome Data Base*, chez l'Homme) **:** http://www.gdb.org/gdb/
MGI (*Mouse Genome Informatics*, chez la souris) **:** http://www.informatics.jax.org/
Flybase (chez la drosophile) **:** http://fly.ebi.ac.uk:7081/

Cliniques:
Orphanet (Base de données sur les maladies rares et sur les médicaments orphelins)

3.4-Organisation de l'information

Le contenu d'une base

Ces bases offrent des fiches descriptives de séquences nucléiques ou protéiques (AND, ADNc, ARN, protéines) ; ces fiches sont appelées des entrées.
Une "entrée" (entry en anglais) contient principalement :

- Une séquence.
- Qui l'a déposée dans la base
- La date de dépot, voire de mise à jour
- Des informations sur l'organisme qui contient cette séquence.
- Des informations sur le génome (par exemple nucléaire ou mitochondrial).

Eventuellement (souvent) :

- une ou des références à des articles scientifiques
- la description de la composition de la séquences (les annotations : ou se situe le gène, la CDS, les introns, exons...)
- des hyperliens vers d'autres bases de données ou séquences.

Pour **identifier** ces séquences, les différentes banques de données leur assignent des **Numéros d'Accession** uniques au sein de leurs collections respectives.
Le numéro d'accession est la façon la plus directe, la plus sure et la plus simple de retrouver une séquence dans une base de données. Lors de la naissance d'une séquence (sa soumission à une base de données primaire) on lui attribue un numéro d'accession nouveau et unique. Cet identifiant sera utilisé pour citer la séquence dans les articles scientifiques ou sert à retrouver rapidement cette information car cet identifiant ne changera jamais.
Les numéros d'accession sont de la forme :
anciennes séquences : U43555, J47888, ...
nouvelles séquences: AP554922, AY11102, ...

Le numéro de version: chaque entrée possède un numéro de version. Par exemple lors de sa première soumission une séquence sera attribuée le numéro d'accession AB334763 et le numéro de version 1. Si bien que son identifiant complet sera: AB334763.1
Si on change quelque chose dans l'entrée, le numéro de version deviendra 2, et donc l'identifiant complet sera AB334763.2 etc...

3.5- Utilisation des bases de données

La diffusion

Avant 1990 : envoi postal de bandes magnétiques puis de CD-Rom aux laboratoires abonnés. A partir de 1990: Utilisation des réseaux informatiques (Internet) avec mise à disposition gratuite et mise à jour quotidienne des données. Depuis le début des années 90, avec l'installation massive des réseaux informatiques (Internet) à hauts débits, beaucoup de laboratoires consultent les bases de données via ces réseaux à partir de serveurs publics. Ces réseaux informatiques rapides et les services qui en découlent permettent une large diffusion des bases. Beaucoup de serveurs mettent gratuitement à disposition de nombreuses bases (pour la plupart des banques), dont les grandes banques de séquences généralistes comme l'EMBL ou Swissprot avec une mise à jour quotidienne des données, mais également un grand nombre d'autres bases dont la diffusion était auparavant plus restreinte.

La consultation

L'utilisation de l'Internet pour la recherche de l'information biologique est d'actualité.
Si la méthode n'est pas structurée, le chercheur de l'information aura le sentiment d'être perdu au sein de cette gigantesque toile d'araignée qui est le web.
C'est pour cela qu'une structuration et une modélisation de la méthode de recherche s'imposent.
Cela permet, en effet de gagner énormément de temps et d'effectuer des recherches plus spécifiques.
Pour que les bases de données soient plus facilement exploitables et que les utilisateurs puissent extraire des sous-ensembles de séquences qui les intéressent, il existe des programmes de consultation.

Les étapes pour une consultation de bases de séquences biologiques

Classiquement, ces étapes sont les suivantes :

- Le choix de la ou des bases à interroger
- La formulation de la recherche
- Le choix pour la présentation des résultats

Il existe des règles spécifiques pour la formulation des requêtes d'interrogation de bases de séquences biologiques, en particulier: le nom d'espèce à laquelle appartiennent une séquence et les mots-clés à utiliser pour décrire les séquences recherchées

Interrogation des bases de données

On peut interroger une BD pour plusieurs raisons :
• Pour connaître la séquence d'un gène ou d'une portion de ce gène
• Pour connaître la structure primaire d'une protéine
• Pour comparer deux séquences, …
Le résultat de l'interrogation des BD est une fiche descriptive de la molécule.
On parlera alors d'une **entrée** (ou fiche descriptive de la séquence recherchée).
La structure d'une entrée est presque la même quelque soit la BD interrogée.

Systèmes d'interrogation des bases de données

Toutes les banques de données possèdent leurs systèmes (ou outils, ou logiciels) d'interrogation.
Chaque banque de séquences a son propre système d'interrogation, avec quelquefois des versions différentes proposées par certains serveurs. Des outils d'interrogation qui permettent des interrogations dans de nombreuses banques de séquences, généralistes ou spécialisées, ont été développés, les plus connus et utilisés sont :
SRS (Sequence Retrieval System) : il permet une interrogation simple ou croisée sur les bases en biologie moléculaire. C'est un outil d'accès privilégié aux banques de séquences généralistes et spécialisées.
ENTREZ : Ce serveur permet l'interrogation des banques de séquences Medline et PubMed, GenBank, EMBL, DDBJ, PIR, SwissProt, PRF, PDB, SNP, CDD, UniSTS, OMIM... (Medline et pubMed sont des bases de données bibliographiques biologiques). Ce serveur est au NCBI (USA)
ACNUC : Système d'interrogation (au choix) des banques EMBL, Genbank, PIR, Hovergen, NRSub, NRBact, etc .. au total une trentaine de banques. Ce serveur est accessible au Pôle Bio- Informatique Lyonnais (PBIL).
DBGET : Système d'interrogation des banques PubMed, EMBL, Genbank, SwissProt, PIR, PRF, LITDB, PDB, PDBSTR, EPD, Prosite, Ligand, PMD, AA- Index, OMIM. Ce serveur est accessible au GenomeNet (Japon)
Une "entrée" (entry en anglais) dans la base de donnée Genbank

Bacillus subtilis 16S ribosomal RNA gene, partial sequence

GenBank: JX403943.1

FASTA Graphics

```
LOCUS       JX403943               1447 bp    DNA     linear   BCT 03-SEP-2012
DEFINITION  Bacillus subtilis 16S ribosomal RNA gene, partial sequence.
ACCESSION   JX403943
VERSION     JX403943.1  GI:402535856
KEYWORDS    .
SOURCE      Bacillus subtilis
  ORGANISM  Bacillus subtilis
            Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus.
REFERENCE   1  (bases 1 to 1447)
  AUTHORS   Wang,X.
  TITLE     16S rDNA of antagonistic bacterium W-2 against wheat take-all
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1447)
  AUTHORS   Wang,X.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-JUL-2012) Library, Agricultural University of Hebei,
            Lingyusi Street No. 289, Baoding, Hebei 071001, China
FEATURES             Location/Qualifiers
     source          1..1447
                     /organism="Bacillus subtilis"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:1423"
                     /PCR_primers="fwd_name: 27f, fwd_seq:
                     agagtttgatcctggctcag, rev_name: 1492r, rev_seq:
                     ggttaccttgttacgactt"
     rRNA            <1..>1447
                     /product="16S ribosomal RNA"
ORIGIN
        1 aggcgggtgc tatacatgca gtcgagcgga cagatgggag cttgctccct gatgttagcg
       61 gcggacgggt gagtaacacg tgggtaacct gcctgtaaga ctgggataac tccgggaaac
      121 cggggctaat accggatgct tgtttgaacc gcatggttca gacataaaag gtggcttcgg
      181 ctaccactta cagatggacc cgcggcgcat tagctagttg gtgaggtaac ggctcaccaa
      241 ggcaacgatg cgtagccgac ctgagagggt gatcggccac actgggactg agacacggcc
      301 cagactccta cgggaggcag cagtagggaa tcttccgcaa tggacgaaag tctgacggag
      361 caacgccgcg tgagtgatga aggttttcgg atcgtaaagc tctgttgtta gggaagaaca
      421 agtgccgttc aaatagggcg gcaccttgac ggtacctaac cagaaagcca cggctaacta
      481 cgtgccagca gccgcggtaa tacgtaggtg gcaagcgttg tccggaatta ttgggcgtaa
      541 agggctcgca ggcggtttct taagtctgat gtgaaagccc ccggctcaac cggggagggt
      601 cattggaaac tggggaactt gagtgcagaa gaggagagtg gaattccacg tgtagcggtg
      661 aaatgcgtag agatgtggag gaacaccagt ggcgaaggcg actctctggt ctgtaactga
      721 cgctgaggag cgaaagcgtg gggagcgaac aggattagat accctggtag tccacgccgt
      781 aaacgatgag tgctaagtgt taggggtttt ccgccccttа gtgctgcagc taacgcatta
      841 agcactccgc ctggggagta cggtcgcaag actgaaactc aaaggaattg acgggggccc
      901 gcacaagcgg tggagcatgt ggtttaattc gaagcaacgc gaagaacctt accaggtctt
      961 gacatcctct gacaatccta gagataggac gtcccccttcg ggggcagagt gacaggtggt
     1021 gcatggttgt cgtcagctcg tgtcgtgaga tgttgggtta agtcccgcaa cgagcgcaac
     1081 ccttgatctt agttgccagc attcagttgg gcactctaag gtgactgccg gtgacaaacc
     1141 ggaggaaggt ggggatgacg tcaaatcatc atgcccctta tgacctgggc tacacacgtg
     1201 ctacaatgga cagaacaaag ggcagcgaaa ccgcgaggtt aagccaatcc cacaaatctg
     1261 ttctcagttc ggatcgcagt ctgcaactcg actgcgtgaa gctggaatcg ctagtaatcg
     1321 cggatcagca tgccgcggtg aatacgttcc cgggccttgt acacaccgcc cgtcacacca
     1381 cgagagtttg taacacccga agtcggtgag gtaacctttt aggagccagc cgccgaaggt
     1441 gacagga
//
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Related information

Taxonomy

Recent activity

Bacillus subtilis 16S ribosor

16S rDNA bacillus subtilis p

16 Sr DNA bacillus subtilis

16 Sr DNA bacillus subtilis