

Bioinformatique et données biologiques

Cours d'introduction à la bioinformatique et de
présentation des banques de données biologiques.

2^{ème} partie

Equipe Bonsai (2014)

COMMENT INTERROGER UNE BANQUE ?



Rechercher des données à partir d'annotations

- Recherche de mots ou expressions dans le texte des entrées via une interface d'interrogation
- Ce que souhaite les utilisateurs
 - Obtenir des données pertinentes
 - Pas trop de résultats, mais tous ceux relatifs à leur problématique
 - Prendre rapidement en main l'interface
 - Obtenir rapidement les résultats
 - Pouvoir manipuler les données obtenues
 - Changer de format, lancer des calculs, ...
- Exemple d'un système d'interrogation
 - Gquery (Entrez), le système développé par le NCBI
 - <http://www.ncbi.nlm.nih.gov/gquery/>

Gquery (Entrez), le système d'interrogation du NCBI

GQuery

Global Cross-database NCBI Search

Search NCBI databases

Literature

PubMed : scientific & medical abstracts/citations

PubMed Central : full-text journal articles

NLM Catalog : books, journals and more in the NLM Collections

MeSH : ontology used for PubMed indexing

Books : books and reports

Site Search : NCBI web and FTP site index

Health

PubMed Health : clinical effectiveness, disease and drug reports

MedGen : medical genetics literature and links

GTR : genetic testing registry

dbGaP : genotype/phenotype interaction studies

ClinVar : human variations of clinical significance

OMIM : online mendelian inheritance in man

OMIA : online mendelian inheritance in animals

Organisms

Taxonomy : taxonomic classification and nomenclature catalog

Nucleotide Sequences

Nucleotide : DNA and RNA sequences

GSS : genome survey sequences

FST : expressed sequence tag sequences

SRA : high-throughput DNA and RNA sequence read archive

PopSet : sequence sets from phylogenetic and population studies

Probe : sequence-based probes and primers

NCBI: recherche d'un terme

Quelles entrées de la banque nucléique contiennent le gène MAX ?

- Saisie de **max** dans la zone de requêtes

- Recherche du mot **max** dans tout le texte des entrées
- Pas spécifique au nom de gène : 953 619 entrées

- Saisie de **max [gene]**

- Recherche du mot **max** dans les champs correspondant au **nom de gène**
- Recherche ciblée : 64 entrées

The screenshot shows the NCBI Nucleotide search interface. The search term 'max' is entered in the search bar. The results show 262,437 nucleotide sequences. The 'Display' dropdown is set to 'Summary', and the 'Show' dropdown is set to '20'. The 'Sort by' dropdown is set to 'Relevance'. The 'Send to' dropdown is set to 'Clipboard'. The results are displayed in a table with columns for 'Accession', 'Organism', 'Length', and 'Description'. The first two results are:

- 1: [NM_001039545](#) Reports [Links](#)
Mus musculus myosin, heavy polypeptide 2, skeletal muscle, adult (Myh2), mRNA
[gi205830427|ref|NM_001039545.2|](#)[205830427]
- 2: [NM_001135145](#) Reports [Links](#)
Danio rerio hypothetical protein LOC100189619 (LOC100189619), mRNA
[gi205830390|ref|NM_001135145.1|](#)[205830390]

The screenshot shows the NCBI Nucleotide search interface. The search term 'max [gene]' is entered in the search bar. The results show 67 nucleotide sequences. The 'Display' dropdown is set to 'Summary', and the 'Show' dropdown is set to '20'. The 'Sort by' dropdown is set to 'Relevance'. The 'Send to' dropdown is set to 'Clipboard'. The results are displayed in a table with columns for 'Accession', 'Organism', 'Length', and 'Description'. The first two results are:

- 1: [NM_022210](#) Reports [Links](#)
Rattus norvegicus Max protein (Max), mRNA
[gi11559987|ref|NM_022210.1|](#)[11559987]
- 2: [XM_002119897](#) Reports [Links](#)
PREDICTED: Ciona intestinalis transcription factor protein (max), mRNA
[gi198432239|ref|XM_002119897.1|](#)[198432239]

NCBI: utilisation des champs

- Champ d'une entrée : information répertoriée dans une partie précise de l'entrée
 - Toutes les valeurs d'un champ sont répertoriées dans un index
- Permet de mieux cibler ses requêtes
 - Les termes ne sont pas cherchés dans tout le texte de l'entrée
- Syntaxe dans Entrez : **Mot_recherché [champ]**
- Exemples de champs disponibles (selon banques) :
 - **[gene]** : nom du gène
 - A utiliser avec précaution car les auteurs d'une entrée ne mettent pas tous le nom du gène au bon endroit
 - **[protein]** : nom de la protéine
 - Même remarque que pour « gene »; peuvent être associés
 - **[organism]** : nom de l'espèce ou tout autre niveau taxonomique
 - ...

NCBI: association de termes

Trois opérateurs booléens possibles : AND, OR, NOT

rattus norvegicus [organism] AND mus musculus [organism]

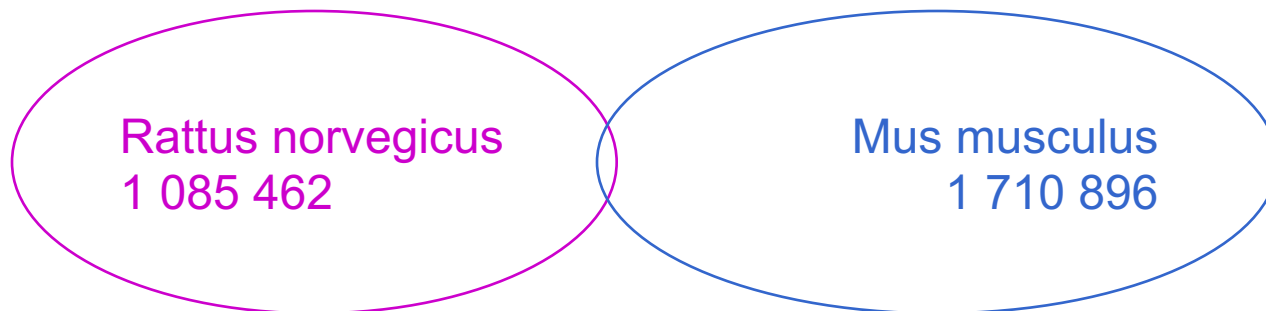
- 1 entrée : “Synthetic construct chimeric tyrosine hydroxylase”

rattus norvegicus [organism] OR mus musculus [organism]

- 2 796 357 entrées
- La séquence provient soit du rat, soit de la souris

rattus norvegicus [organism] NOT mus musculus [organism]

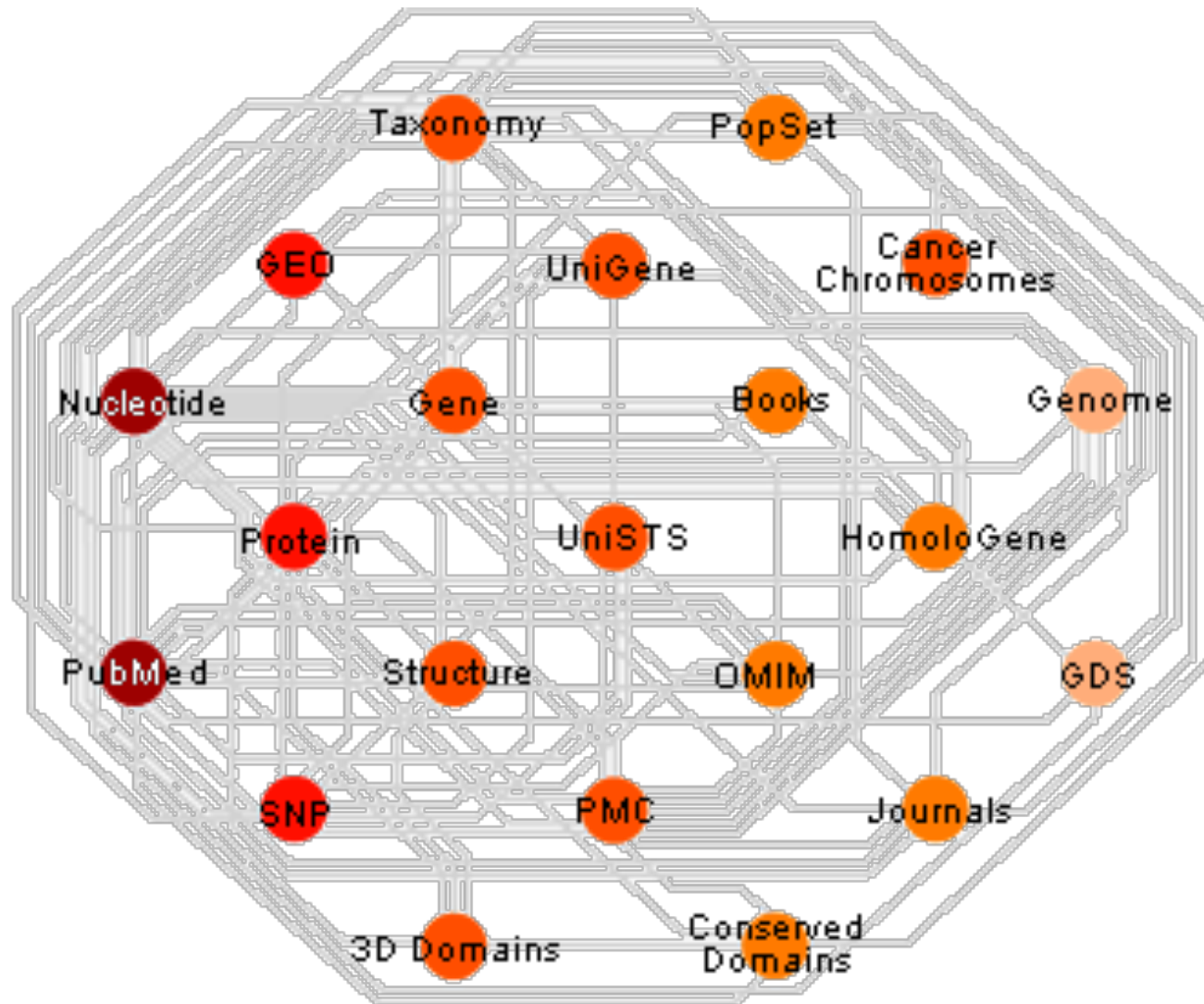
- 1 085 461 entrées
- Tous les séquences du rat, sauf la séquence chimérique



NCBI: comment construire une requête ?

- Déterminer les critères de recherche
 - Ne pas oublier les déclinaisons du mot (singulier, pluriel, ...)
 - Ne pas oublier les synonymes
- Combiner ces critères avec les bons opérateurs
 - Si différents critères qui se complètent : ET
 - Souvent, interrogation de plusieurs champs
 - Si alternative entre plusieurs termes : OU
 - Souvent, différents termes pour un même champ
 - Attention aux parenthèses :
 - ET en général *prioritaire* sur OU \neq ordre *linéaire* au NCBI :
`Gallus gallus [organism] AND (connectin OR titin)`
< `Gallus gallus [organism] AND connectin OR titin`
= `((Gallus gallus [organism]) AND connectin) OR titin)`
- Limiter la recherche des critères à un champ particulier ...

NCBI: liens entre banques



Les banques de séquences protéiques

- Origine des données
 - Traduction de séquences d'ADN
 - Nombreuses données disponibles dans les banques nucléiques
 - Séquençage de protéines
 - Rare car long et coûteux
 - Protéines dont la structure 3D est connue
- Les données stockées : séquences + annotations
 - Protéines entières
 - Fragments de protéines

Banques de séquences protéiques, les débuts

Version papier
jusqu'en 78,
puis version
électronique

NBRF (USA)
avec MIPS
(Allemagne) et
JIPID (Japon)

SIB (Swiss
Institute of
Bioinformatics)
et EBI



Mise en
commun des
informations

UniProt, ses deux banques

■ SwissProt

- Données corrigées et validées par des experts
- Haut niveau d'annotation
- Redondance minimale
- Nombreux liens vers d'autres banques

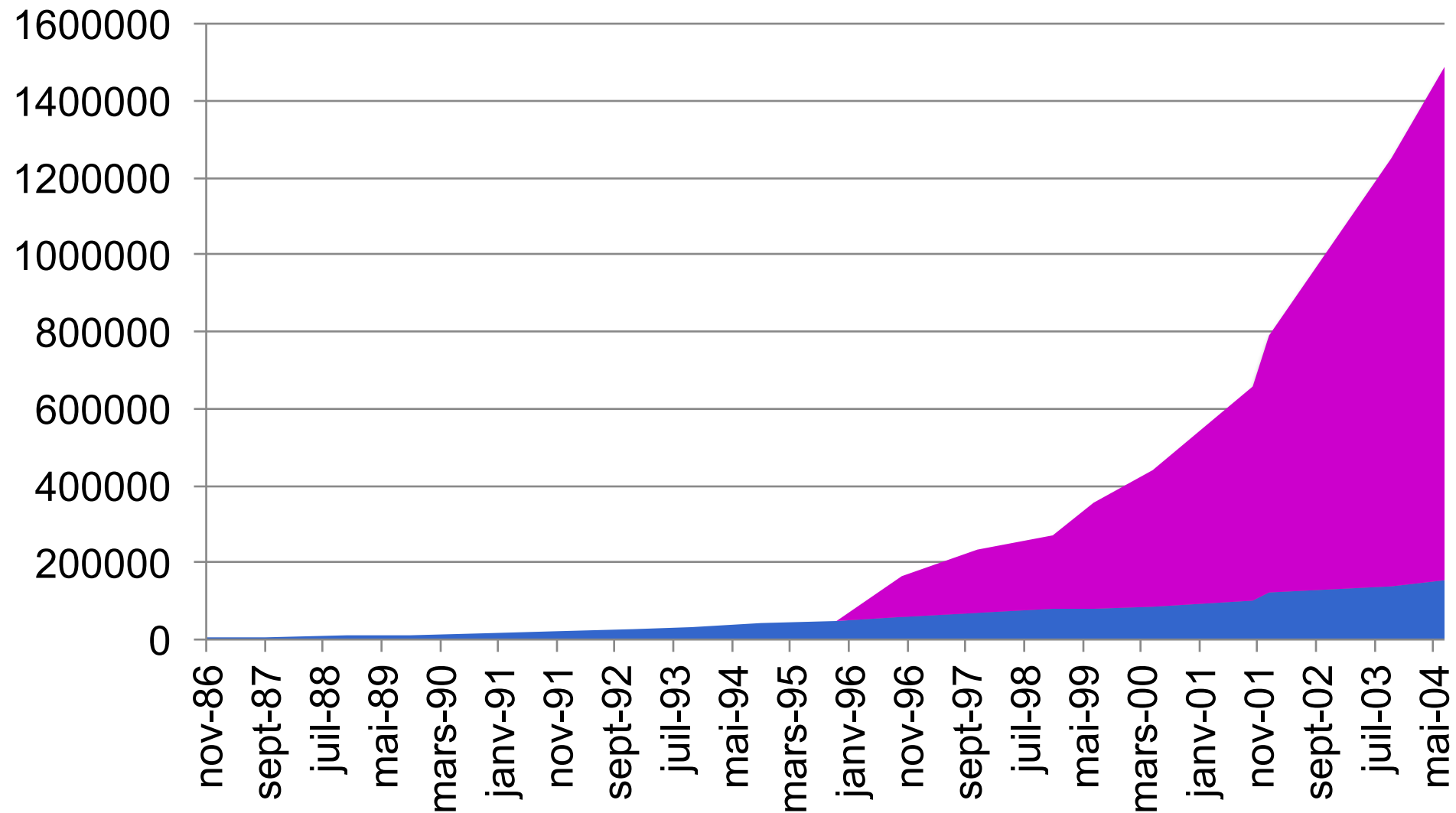
■ TrEMBL

- Entrées supplémentaires à SwissProt (pas encore annotées)
- Traduction automatique des CDS de l'EMBL et soumissions spontanées.
- Annotation automatique des protéines



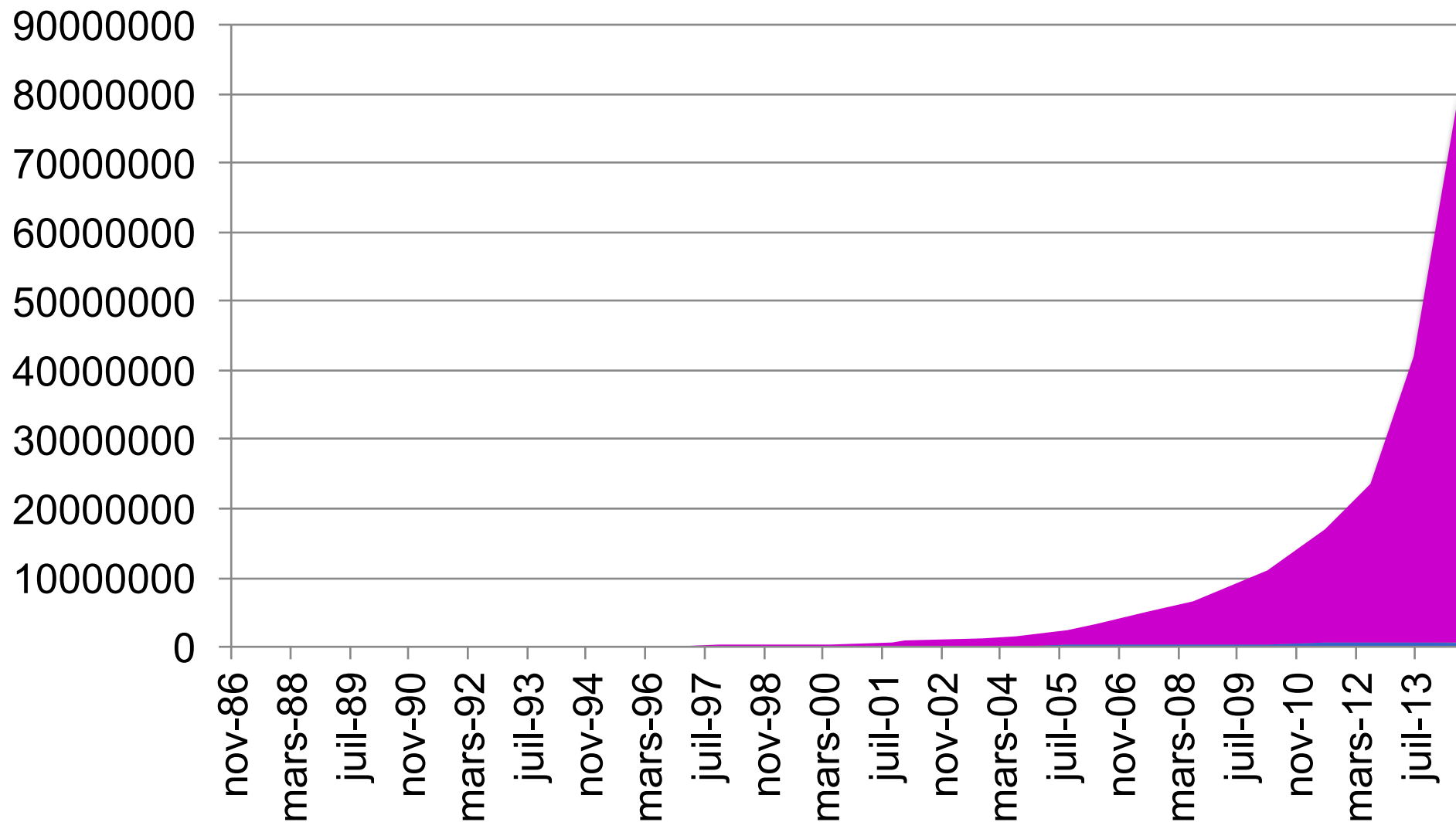
SwissProt/TrEMBL, nombre d'entrées

■ SwissProt ■ TrEMBL



SwissProt/TrEMBL, nombre d'entrées

■ SwissProt ■ TrEMBL



Les annotations de SwissProt (1/2)

- Fonction(s) de la protéine
- Modifications post-traductionnelles
- Domaines et sites fonctionnels
- Structure secondaire (hélices alpha, feuillets bêta, ...)
- Structure quaternaire (participation à un complexe)
- Ressemblance à d'autres protéines
- Maladies associées aux déficiences de la protéine
- Conflits de séquences
- Variants d'épissage, ...

Les annotations de SwissProt (2/2)

- Origine des annotations :
 - Articles spécifiques à une protéine
 - Articles de revue concernant une famille de protéines
 - Informations venant d'experts extérieurs
 - Prédiction à l'aide de programmes (vérifiée par un expert)
- Les sources des annotations sont toujours mentionnées
- Où trouver les annotations ?
 - Lignes CC (commentaires)
 - Lignes FT (caractéristiques biologiques localisées sur la séquence)

SwissProt/TrEMBL, format d'une entrée

- Format basé sur celui de l'EMBL
 - Mot-clé de 2 lettres au début de chaque ligne
 - Les mêmes mots-clés sont utilisés
 - Format différent pour les Features
- Mots-clés supplémentaires :
 - GN : les différents noms du gène qui code pour la protéine (OR)
les différents gènes qui codent pour la même protéine (AND)
 - OX : références croisées vers les banques taxonomiques
 - CC : commentaires, lignes très documentées dans SwissProt
 - KW : mots-clés issus d'un dictionnaire

SwissProt/TrEMBL, lignes CC

- Informations découpées en blocs pour plus de lisibilité

CC -!- TOPIC: First line of a comment block;

CC second and subsequent lines of a comment block.

- De nombreux sujets sont abordés

- FUNCTION : description générale de la fonction de la protéine
- CATALYTIC ACTIVITY : description des réactions catalysées par les enzymes
- DEVELOPMENTAL STAGE : description du stade spécifique auquel la protéine est exprimée
- SUBUNIT : complexes dont fait partie la protéine (+ partenaires)
- ...

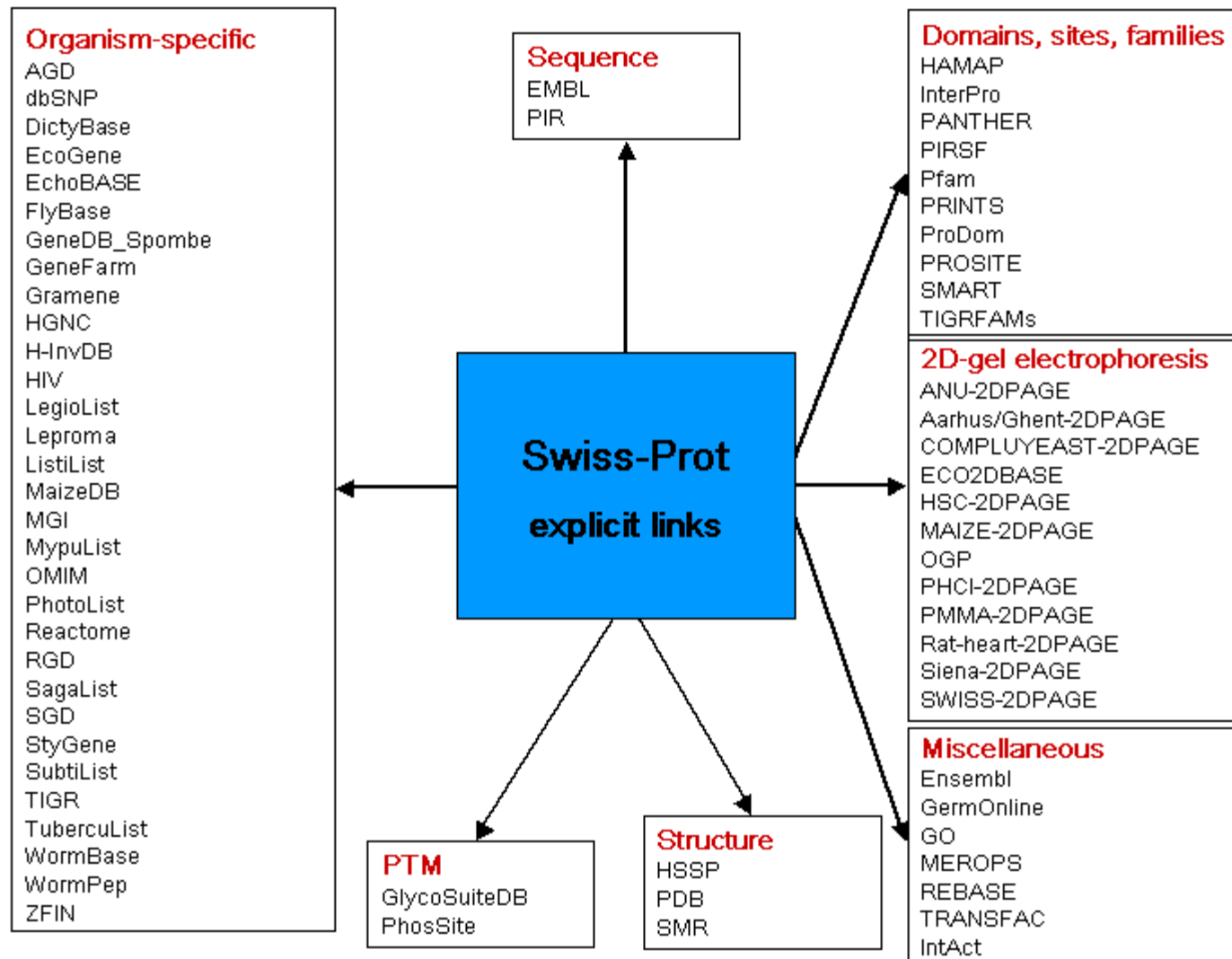
SwissProt/TrEMBL, lignes FT

- Régions ou sites d'intérêt dans la séquence
 - Modifications post-traductionnelles
 - Sites de fixation
 - Sites actifs d'enzymes
 - Structures secondaires
 - Changements de séquence (y compris les variants)
- Format en colonne (nb caractères)
 - 1-2 : FT
 - 6-13 : Key (mot-clé, vocabulaire contrôlé)
 - 15-20 22-27 : début et fin de l'objet
 - 35-75 : description (éventuellement sur plusieurs lignes)

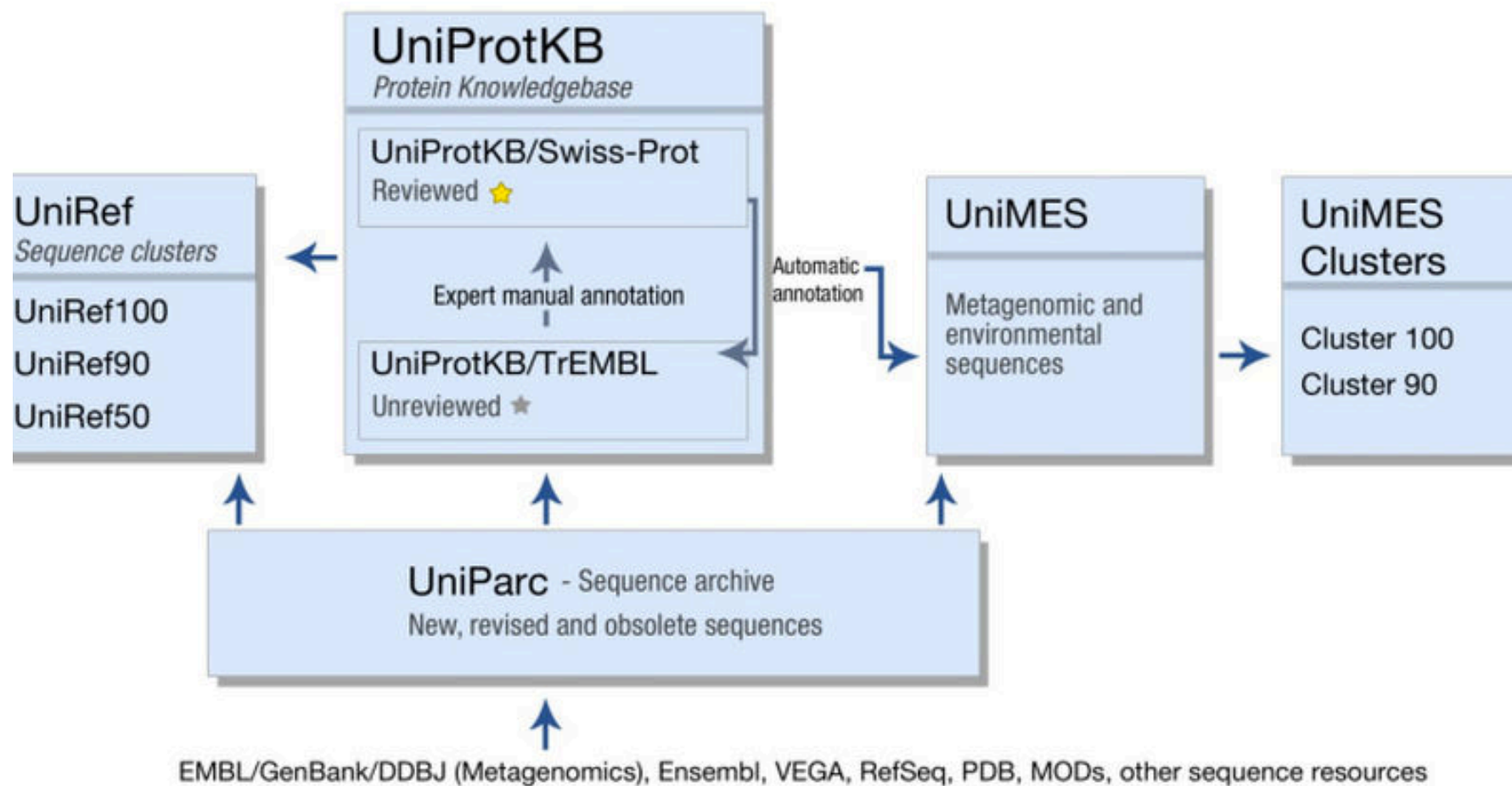
Fiabilité de l'information

- Certaines informations ne sont pas issues de données expérimentales
- Trois mots-clés sont utilisés pour les qualifier :
 - Potential : des preuves amènent à penser que l'annotation est valable (logiciel de prédiction + cohérence avec le contexte)
 - Probable : meilleure fiabilité, début de preuve expérimentale
 - By similarity : par ressemblance de séquence avec une protéine annotée expérimentalement

DR : références vers d'autres banques



UniProt, vue générale



UniProt, les différentes banques

- UniProt : UniProt Knowledgebase
 - « entrepôt » central de séquences et fonctions protéiques
 - Deux parties : SwissProt et TrEMBL
 - Plus d'informations que dans les banques d'origine
- UniParc : UniProt Archive
 - UniProt + d'autres banques (PDB, RefSeq, FlyBase, brevets, ...)
- UniRef : UniProt Non-redundant Reference database
 - UniRef100 : regroupement des séquences identiques et de leurs fragments provenant d'un même organisme
 - UniRef90 : entrées de UniRef100 avec plus de 90% d'identité
 - UniRef50 : idem pour 50% d'identité



Les banques protéiques « de deuxième niveau »

- Points de départ
 - Séquences protéiques
 - Connaissances biologiques
- Analyse des séquences pour construire de nouvelles données
- Ex : banques de familles de protéines
 - Regroupement des protéines ayant des fonctions identiques ou proches
 - Construites souvent par comparaison de toutes les protéines entre elles puis constitution de groupes
 - HomoloGene, COG, KEGG, SSDB

Banques de motifs et domaines protéiques

- Une famille de protéines peut être caractérisée par un motif ou domaine protéique
 - Séquence plus ou moins conservée importante pour la fonction des protéines de la famille
 - Déterminée à partir d'un alignement multiple
 - Plusieurs représentation possibles : séquence consensus, expression régulière, alignement multiple, matrice poids position, chaînes de Markov cachées (HMM), ...
- Nombreuses banques
 - Prosite, PFAM, Blocks, Prodom, CDD, ...

Bases de connaissances protéiques

- Rassemble divers données
 - Données primaires : séquences et annotations
 - Données secondaires : issues de calculs (familles, motifs, ...)
- Inférence de nouvelles connaissances
 - Familles construites par similarité de séquences
 - Mise en commun des annotations
 - Prédiction de la fonction de protéines inconnues
- Evite la consultation de plusieurs banques pour étudier une séquence

InterPro

- <http://www.ebi.ac.uk/interpro/>
- Contenu
 - Superfamilles, familles, domaines, motifs, sites fonctionnels, modifications post-traductionnelles, structures 3D
- Regroupe plusieurs banques existantes
 - Prosite, PFAM, Blocks, Prodom, Smart, Prints, TIGRFams, Superfamily, SCOP, CATH, MSD
- Une entrée
 - Description biologique détaillée
 - Représentation de l'objet par les différentes banques



Banques d'interactions protéiques

- Différents niveaux d'interactions possible
 - Physique (formation de complexes)
 - Collaboratif (intervention dans un même processus)
- Données issues d'expériences
 - Doubles-hybrides chez la Levure, co-immunoprécipitation
 - Biomolecular Interaction Network Data (BIND) database, Database of Interacting Proteins (DIP)
- Données extraites d'articles scientifiques
 - Analyse informatique des articles, lecture par des « curateurs »
 - Molecular Interaction database (MINT), IntAct database
- Mise au point d'un format de stockage des données
 - Standards Initiative (PSI) Molecular Interaction (PSI-MI)

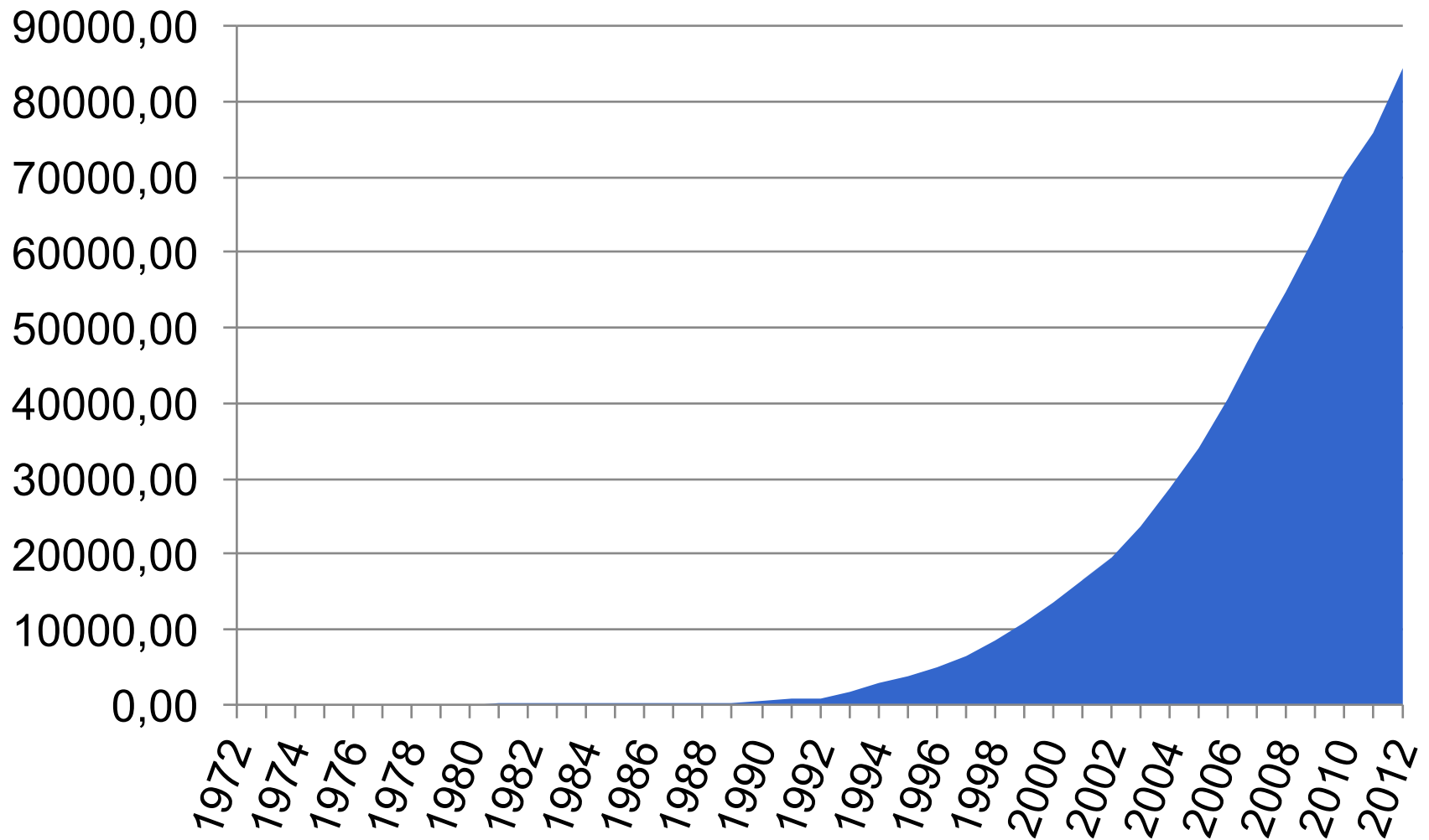
Structures 3D de protéines

- 1958 : détermination de la première structure 3D de protéine par Kendrew et Perutz
 - Découverte de la complexité de la structure 3D d'une protéine
- Hypothèses de l'époque :
 - Deux protéines avec des séquences proches se replient de façon semblable
 - Deux protéines ayant des structures 3D proches ont des séquences proches
- La structure 3D des protéines est déterminante pour leur fonction

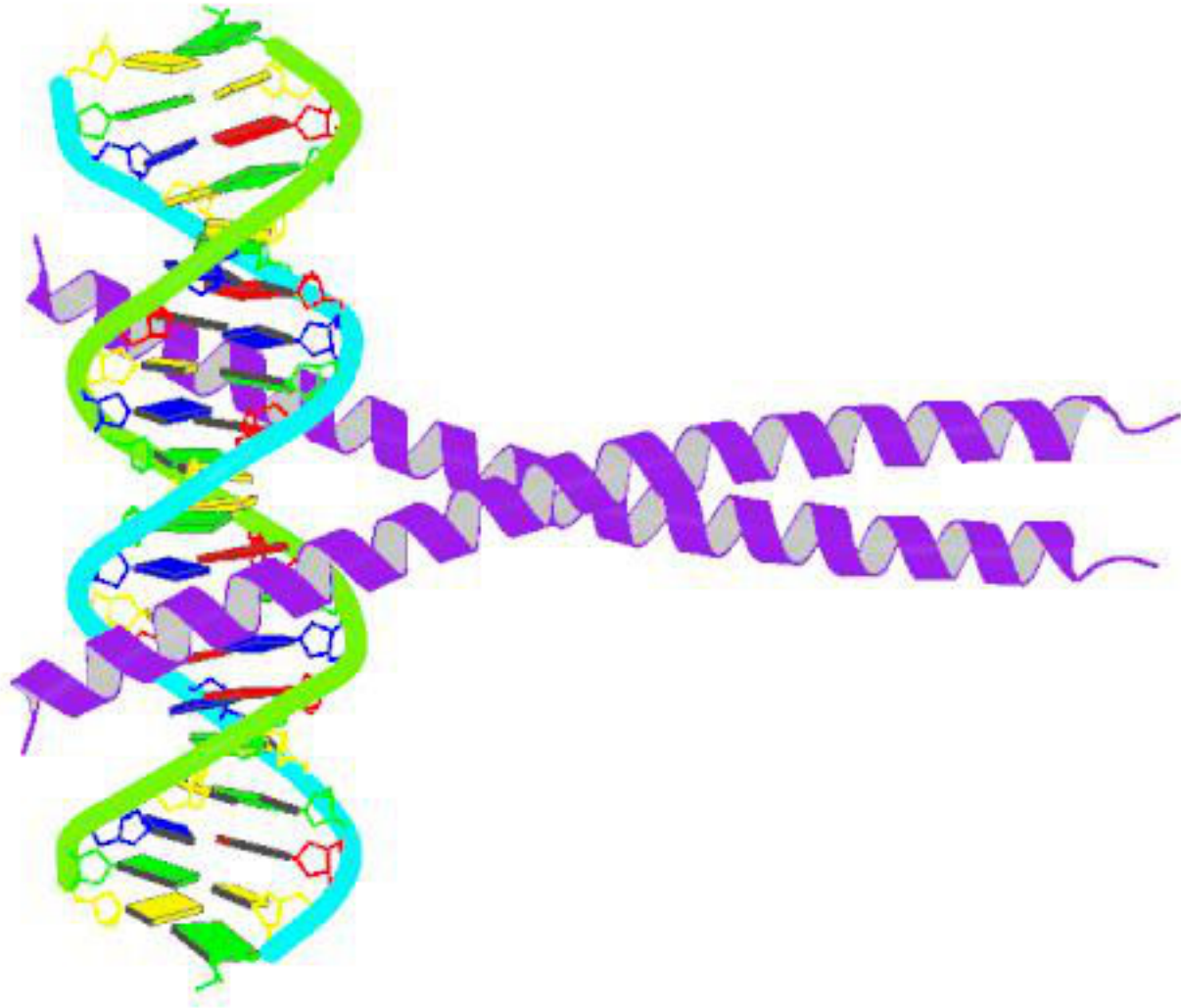
wwPDB : la banque de structures 3D

- worldwilde Protein Data Bank
- Seule banque de structures 3D de protéines, acides aminés et grosses molécules biologiques
- 1971 : le RSCB (Research Collaboratory for Structural Bioinformatics) créé la banque PDB
- 2003 : regroupement des 3 banques de structures 3D en une seule
 - RSCB (Research Collaboratory for Structural Bioinformatics)
 - MSD (Macromolecular Structure Database)
 - PDBj (Protein Data Bank Japan)

PDB, nombre d'entrées



Structure d'une partie de AP1_human



Classification structurale des protéines

- Classification des protéines basée sur leurs structures 2D, 3D et leur fonction
 - Construction manuelle aidée d'outils de comparaison de structures et de séquences
- 2 banques « concurrentes » :
 - **SCOP** : Structural Classification of Proteins
 - **CATH** : Protein Structure Classification
- 4 niveaux :
 - **Class** / **Fold** / **Superfamily** / **Family**
 - **Class** / **Architecture** / **Topology** / **Homologous superfamily** / ... / ...
- Se méfier des classifications : contiennent des aberrations
 - Des structures proches qui ne sont pas dans les mêmes familles
 - Des structures distinctes qui sont dans une même famille

Similarité structure

Homologie structurelle
+ fonctionnelle

Similarité
séquence

SCOP, hiérarchie principale

■ **Fold** (similarités structurales majeures)

Similarité structure

- Même éléments 2D, dans le même ordre et avec la même topologie

■ **Superfamily** (possibilité d'un ancêtre commun)

Homologie structurelle
+ fonctionnelle

- Faible conservation de séquence
- Mais caractéristiques structurales et fonctionnelles liées

■ **Family** (lien dans l'évolution clairement démontré)

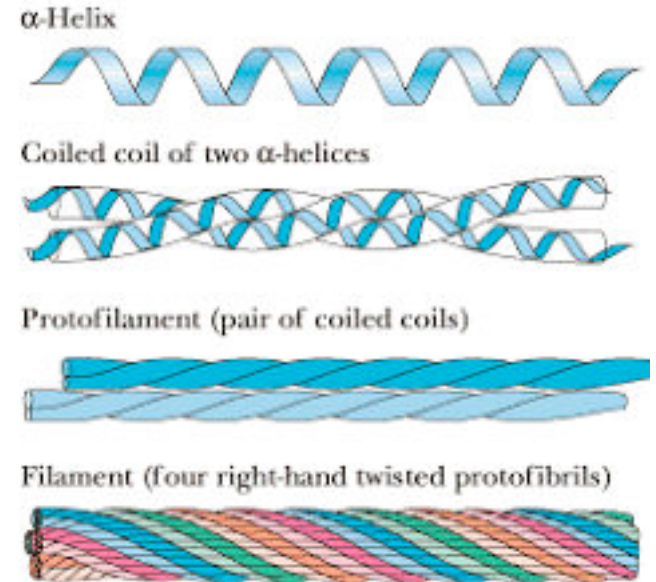
- Souvent > 30% identité

Similarité
séquence

SCOP, les classes (haut de la hiérarchie)

CLASS (avant FOLD) :

- Protéines tout α
- Protéines tout β
- Protéines α/β (éléments α et β mélangés)
- Protéines $\alpha+\beta$ (éléments α et β séparés)
- Protéines multi-domaine
- Peptides/protéines de membrane/surface
- Petites protéines
- Protéines « coiled-coil » (e.g. α -keratin)
- Peptides
- Protéines artificielles



SCOP, exemple : hiérarchie de AP1_human

1. **Root:** Scop

2. **Class:** Coiled coil proteins [57942]

Not a true class

Similarité structure

3. **Fold:** Parallel coiled-coil [57943]

Not a true fold; includes oligomers of shorter identical helices

4. **Superfamily:** Leucine zipper domain [57959]

Homologie structurelle
+ fonctionnelle

5. **Family:** Leucine zipper domain [57960]

6. **Protein:** C-jun [57975]

Similarité
séquence

7. **Species:** Human (Homo sapiens) [57976]

CATH, hiérarchie principale

■ Class [C-Level]

- 3 classes : *mainly- α* , *mainly- β* , α & β

■ Architecture [A-level] :

- *Similarités structurales majeures*

■ Topology (Fold family) [T-level]

- *Même topologies/repliement, même connectivité globale*

■ Homologous Superfamily [H-level]

- *Ancêtre commun « certain » (probabilité élevée)*

■ Sequence Family Levels [S,O,L,I,D]

- **Sequence Family (S35)** (35% d'identité, *automatique*)
- **Non-identical (S95)** (...)
- **Identical (S100)** (...)

Similarité structure

Homologie structurelle
+ fonctionnelle

Similarité
séquence

CATH, exemple: hiérarchie de AP1_human

1.20.5.170.8.1.1

1. **Class** : Mainly Alpha (1)

Similarité structure

2. **Architecture** : Up-down Bundle (20)

3. **Topology** : Single alpha-helices involved in coiled-coils or other helix-helix interfaces (5)

Homologie structurelle
+ fonctionnelle

4. **Homologous Superfamily** : TRANSCRIPTION/DNA (170)



5. **Sequence Family (S35)** : TRANSCRIPTION/DNA

6. **Non-identical (S95)** : TRANSCRIPTION/DNA

7. **Identical (S100)** : TRANSCRIPTION/DNA

Similarité
séquence

Banques de données de puces à ADN

- Stockage des données d'expériences de puces à ADN
- Constitution de la société savante MGED (→ FGED)
Microarray Gene Expression Data (→ Functional Genomics Data Society)
 - Soutien pour que les auteurs soumettent leurs données avant publication
- Mise au point d'un « format » pour mémoriser les données
 - *MIAME : Minimum Information About a Microarray Experiment*
 - Description des conditions expérimentales
 - Possibilité de reproduire l'expérience
 - Interprétation appropriée des données
- Les banques de *microarrays* :
 - NCBI **Gene Expression Omnibus (GEO)** 
 - **ArrayExpress** (à l'EBI) 
 - **Chemical Effects in Biological Systems (CEBS)** Knowledgebase

Banques de séquences brutes : SRA/TRACE

- **Sequence Read Archive (SRA)** <http://www.ncbi.nlm.nih.gov/sra/>
 - Stocke les données brutes (séquence + qualités) NGS. (*Séquenceurs Nouvelle Génération*)
 - Support du NCBI, EBI, et DDBJ
 - Soutien pour que les auteurs soumettent leurs données :
 - *Formats : SFF (Roche 454), Illumina Native, Illumina SRF, AB SOLiD Native, AB SOLiD SRF*
 - *Hold–Until–Published (HUP)*
 - **Sequence Read Format (SRF)** : format adapté compressé.
 - Taille (2013) : 1770 téra-bases ou 1,77 péta-bases (*peta = 10^{15} ; téra = 10^{12} ; giga = 10^9*)
- **Trace Archive (TRACE)** <http://www.ncbi.nlm.nih.gov/Traces/>
 - Stocke toutes les séquences brutes Sanger (~300-800bp)
 - Taille (2013) : ~ 2,2 milliards de séquences (stable depuis 2010)



Banques de séq. préassemblés : HTGS,WGS,TSA

ADN : fragments de génomes ...

■ **High Throughput Genome Sequencing (HTGS)**

- clone-based sequencing (mapping with BAC,YAC)
- > 2kb

■ **Whole Genome Shotgun (WGS)**

- Whole genome shotgun sequencing skips the mapping step of clone-based sequencing
- > 200pb

ARN (cDNA) : ARN messagers partiels ...

■ **Transcriptome Shotgun Assembly (TSA)**

- Assemblage « informatique » incomplet de transcriptome (= mRNA partiels)
- Données initiales issus d'EST (Sanger) et surtout de NGS ...
 - dépôt préalable respectif des données brutes dans une banque EST/ la banque SRA
- > 200bp

BLAST® Basic

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Or, upload file

Job Title

☐ Align two or more

Choose Search Set

Database

Organism Optional

Program Selection

Optimize for

- ☒ Highly similar sequences (megablast)
- ☐ More dissimilar sequences (discontigu)
- ☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Genomic plus Transcript

- Human genomic plus transcript (Human)
- Mouse genomic plus transcript (Mouse G

Other Databases

- Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq_rna)
- Reference genomic sequences (refseq_ge
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HT
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats
- Sequence tagged sites (dbsts)
- Whole-genome shotgun contigs (wgs)
- Transcriptome Shotgun Assembly (TSA)**
- 16S ribosomal RNA sequences (Bacteria
- Whole-genome shotgun contigs (wgs)

Banques de seq. générales (NCBI) : NR, RefSeq, ...

■ **NR = Non Redundant**

- ... à l'origine (mais plus le cas sur la banque ADN)
- Regroupe (GenBank + EMBL + DDBJ) + PDB

■ **REFSEQ :**

- Séquences non redondantes, filtrés, issus d'organismes modèles « suffisamment » séquencés.
- Limité en taille (+rapide & + propre), mais aussi en résultats (- de résultats)

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite

blastn blastp blastx tblastn tblastx

BLASTN programs search

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Genomic plus Transcript

- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases

- Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun contigs (wgs)
- Transcriptome Shotgun Assembly (TSA)**
- 16S ribosomal RNA sequences (Bacteria and Archaea)**
- Whole-genome shotgun contigs (wgs)

Or, upload file

Job Title

☐ Align two or more

Choose Search Strategy

Database

Organism Optional

Enter organism name or id--completions will be suggested ☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top tax

Program Selection

Optimize for

- ☒ Highly similar sequences (megablast)
- ☐ More dissimilar sequences (discontiguous megablast)
- ☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Analyse du transcriptome

- Transcriptome =
 - Étude de l'expression des gènes dans un cellule
 - Techniques expérimentales globales : **les puces à ADN // les NGS**
 - Etude de plusieurs milliers de gènes en une seule fois
 - Mesure de leur taux d'expression
 - Etude d'une cellule dans différentes conditions expérimentales
 - Génère une grande quantité de données
 - Besoin d'outils issus de la fouille de données et des statistiques
- Objectif : identifier des groupes (clusters) de gènes ayant des profils d'expression similaires
 - Aide à l'annotation

Etude du fonctionnement de la cellule, objectifs

- Reconstruire l'ensemble des processus cellulaires
 - Voies métaboliques, transport des molécules
 - Voies de régulation des gènes, transduction du signal
 - Réplication et réparation de l'ADN, synthèse des protéines, ...
- Prédire des comportements cellulaires
 - Prédire la réponse à un stimulus
 - Prédire le processus à l'origine d'une maladie
 - Simulation des expériences de mutation de gènes, ...
- Mettre en évidence des modules fonctionnels conservés au cours de l'évolution

Etude du fonctionnement de la cellule, méthodes

- Constitution de bases de connaissances
 - Collecte d'informations sur les génomes, les gènes, les protéines, les interactions entre molécules, les processus cellulaires, ...
 - Établissement de liens entre les données de différentes natures
 - Inférence de nouvelles connaissances
- Représentation dynamique des réseaux cellulaires
 - Équations différentielles
 - Réseaux booléens
 - Théorie des graphes, ...

Recherche bibliographique

Recherche bibliographique

- Effectuée lors de la prise en main d'un sujet
 - Etat de l'art sur les connaissances actuelles
 - Evite de « réinventer la roue »
 - Diminue le nombre d'expériences à réaliser
 - Evaluation de la « concurrence »
 - MAIS : prend beaucoup de temps !
- Veille nécessaire
 - De nouveaux articles sont publiés régulièrement
- Recherche de [nouvelles] techniques expérimentales
- Points d'entrée :

 ■ <http://www.pubmed.gov>

 ■ <http://scholar.google.com>

PubMed et MEDLINE

- MEDLINE est la banque de citations et de résumés biomédicaux du NLM (U.S. National Library of Medicine)
 - Environ 4800 journaux recensés à partir de 1966
 - Nombreux articles indexés par des termes MeSH
- PubMed est une extension de MEDLINE
 - 1.760.000 citations parues entre 1950 et 1965
 - Articles hors sujet (tectonique des plaques, ...) publiés dans des journaux présents dans MEDLINE (Science, Nature, ...)
 - Articles non encore référencés dans MEDLINE car pas encore indexés par des termes MESH
 - Journaux dans le domaine des sciences naturelles qui n'ont pas été sélectionnés par MEDLINE

Termes MeSH

- MeSH : Medical Subject Headings (rubriques médicales)
- Vocabulaire contrôlé de termes biomédicaux et de molécules chimiques établi par le NLM
- 22.997 « descripteurs » et plus de 151.000 éléments chimiques
- Plus de 136.062 synonymes (au sens large) référencés
- Classement hiérarchique des termes
 - Des termes les plus généraux aux termes les plus précis
- Mis à jour régulièrement
- Plus d'information sur :



<http://www.nlm.nih.gov/mesh/>

Indexation des articles à l'aide de MeSH

- Information ajoutée par MedLine sur les articles
- Lectures des articles scientifiques par des experts
- Attribution d'une liste de termes MeSH associés à cet article
 - Ceux correspondant aux thèmes principaux de l'article (Major Topic)
 - Ceux évoqués dans l'article, mais non centraux
 - Recherche du niveau hiérarchique le plus approprié
- 83 qualificatifs (subheadings) permettent de préciser à quel aspect du terme il est fait référence dans l'article

Exemple de terme MeSH, sa définition

Encephalopathy, Bovine Spongiform

A transmissible spongiform encephalopathy of cattle associated with abnormal prion proteins in the brain. Affected animals develop excitability and salivation followed by ATAXIA. This disorder has been associated with consumption of SCRAPIE infected ruminant derived protein. This condition may be transmitted to humans, where it is referred to as variant or new variant CREUTZFELDT-JAKOB SYNDROME. (Vet Rec 1998 Jul 25;143(41):101-5)

Year introduced: 1992

Previous Indexing: Brain Diseases/veterinary (1988-1991) Cattle Diseases (1988-1991)

Ex, ses qualificatifs et synonymes

Subheadings:

Blood ; cerebrospinal fluid ; chemically induced ; classification ; complications ; diagnosis ; drug ; therapy ; economics ; enzymology ; epidemiology ; etiology ; genetics ; history ; immunology ; metabolism ; microbiology ; mortality ; nursing ; pathology ; physiopathology ; prevention and control ; psychology ; surgery ; therapy ; transmission ; virology

Entry Terms:

Bovine Spongiform Encephalopathy ; BSE (Bovine Spongiform Encephalopathy) ; BSEs (Bovine Spongiform Encephalopathy) ; Encephalitis, Bovine Spongiform ; Bovine Spongiform Encephalitis ; Mad Cow Disease ; Mad Cow Diseases ; Spongiform Encephalopathy, Bovine

Ex, les différentes hiérarchies

Diseases Category

Nervous System Diseases

Central Nervous System Diseases

Central Nervous System Infections

Prion Diseases

Encephalopathy, Bovine Spongiform

Diseases Category

Nervous System Diseases

Neurodegenerative Diseases

Prion Diseases

Encephalopathy, Bovine Spongiform

Diseases Category

Animal Diseases

Cattle Diseases

Encephalopathy, Bovine Spongiform

Entrez, consulter un article

- Un formulaire adapté : Single citation matcher
- A l'affichage d'une citation possibilité de :
 - Lien vers le site du journal (accès à l'article soumis à conditions)
 - Lien vers PMC (PubMed Central), archive gratuite

The screenshot shows the NCBI PubMed Single Citation Matcher interface. On the left is a blue sidebar with navigation links. The main content area has a title bar with 'Entrez', 'PubMed', 'Nucleotide', 'Protein', 'Genome', and 'Structure'. Below this, there are instructions and a search form. The 'Single Citation Matcher' link in the sidebar is circled in red.

NCBI PubMed Single Citation Matcher

Entrez PubMed Nucleotide Protein Genome Structure

About Entrez

Text Version

Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy

E-Utilities

PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Use this tool to find PubMed citations. You may omit any field.

Journal may be the full title or the title abbreviation.

For first and last author searching, use smith jc format.

Journal:

Date: (month and day are optional)

Volume: Issue: First page:

Author name (see [help](#))

☐ Only as first author ☐ Only as last author

Title words:

Go Clear

Recherche via les termes MeSH

- Deux moyens de trouver des termes MeSH pertinents :
 - ① Interroger directement la banque des termes MeSH
 - ② Rechercher les critères dans les « titres et résumés » des citations de PubMed
- Identifier les articles intéressants
- Etudier les termes MeSH associés à ces articles (vus avec le format MEDLINE – menu déroulant « Display » –)
- Puis interroger PubMed avec les termes MeSH trouvés
 - Définir si certains termes doivent être des « Major Topics »
 - Combiner les termes avec les opérateurs appropriés
 - Ajouter les critères qui ne correspondent pas à un terme MeSH

Mieux cibler sa requête

- Réduire le nombre d'articles selon certains critères
- Faire un lien vers d'autres articles sur le même thème
- Requêtes spécialisées : Clinical/Special Queries

The screenshot shows the PubMed website interface. At the top, there's the NCBI logo and the PubMed logo with the text "A service of the National Library of Medicine and the National Institutes of Health". Below this is a navigation bar with links to "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "OMIM", "PMC", "Journals", and "Books". The search bar contains the text "Dupont [author]" and buttons for "Go", "Clear", and "Save Search". Below the search bar are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The "Limits" button is circled in red. Below these buttons are dropdown menus for "Display" (set to "Summary"), "Show" (set to "20"), "First Author", and "Send to". Below these are the results counts: "All: 4531" and "Review: 316". Below the counts is the text "Items 1 - 20 of 4531" and a pagination bar showing "Page 1 of 227 Next". The list of results shows two articles. The first article is titled "Prevalence of mixed cryoglobulins in relation to CD4 cell count among patients coinfectd with HIV and hepatitis C virus." and has a "Related Articles" link circled in red. The second article is titled "Tuberculosis in HIV-infected patients: a comprehensive review." and has a "Related Articles, Links" link. In the left sidebar, under "PubMed Services", the link "Clinical Queries" is circled in red.

NCBI PubMed
A service of the National Library of Medicine and the National Institutes of Health
www.pubmed.gov

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for Dupont [author] Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 First Author Send to

All: 4531 Review: 316

Items 1 - 20 of 4531 Page 1 of 227 Next

1: Aaron L, Lebray P, Alyanakian MA, Roudiere L, Therby A, Chaix ML, Dupont B, Pol S, Viard JP. Prevalence of mixed cryoglobulins in relation to CD4 cell count among patients coinfectd with HIV and hepatitis C virus. Clin Infect Dis. 2005 Jan 15;40(2):306-8. Epub 2004 Dec 20. PMID: 15655752 [PubMed - indexed for MEDLINE] Related Articles, Links

2: Aaron L, Saadoun D, Calatroni I, Launay O, Memain N, Vincent V, Marchal G, Dupont B, Bouchaud O, Valeyre D, Lortholary O. Tuberculosis in HIV-infected patients: a comprehensive review. Clin Microbiol Infect. 2004 May;10(5):388-98. Review. PMID: 15113314 [PubMed - indexed for MEDLINE] Related Articles, Links

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation
Matcher
Batch Citation Matcher
Clinical Queries
Special Queries

Un outil puissant : My NCBI

- Possibilité de se créer un compte permanent
 - Création immédiate d'un login et d'un mot-de-passe ([Register])
- Filtrage personnalisé des résultats
 - Ajout d'onglets de tri des résultats
- Mémorisation des requêtes
- Mémorisation des citations
 - Send to Clipboard puis, à partir du Clipboard, send to My NCBI collections

The screenshot shows the PubMed website interface. The top navigation bar includes the NCBI logo, the PubMed logo, and the text "A service of the National Library of Medicine and the National Institutes of Health". The main search bar contains the text "Search PubMed for Dupont [author]". To the right of the search bar, the "My NCBI" button is circled in red, displaying "Welcome pupin. [Sign Out]". Below the search bar, the "Save Search" button is circled in red. The "Clipboard" button is also circled in red. The "Send to" dropdown menu is circled in red. The "Free full text: 445" and "Published in the last 5 years: 830" filters are circled in red. The "Items 1 - 20 of 4531" text is circled in red. The "Page 1 of 227 Next" text is circled in red.

NCBI PubMed A service of the National Library of Medicine and the National Institutes of Health www.pubmed.gov

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for Dupont [author] Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Send to

All: 4531 Free full text: 445 Published in the last 5 years: 830

Items 1 - 20 of 4531 Page 1 of 227 Next

Comment faire de la veille sur un sujet ?

- Faire une première recherche bibliographique
 - Construire une requête pertinente
 - Consulter et trier les articles obtenus
 - Mémoriser l'ensemble des articles pertinents
- Mémoriser la requête dans My NCBI
 - Soit mise en place d'une alerte automatique par e-mail si une nouvelle citation répondant à la requête est parue
 - Soit relance de la requête sur les citations parues depuis la dernière consultation
- Surveiller les articles parus dans les journaux thématiques
 - Inscription aux eTOC (email Table Of Contents) sur le site des journaux
- Abonnement aux flux **RSS** des journaux

