



Bioinformatique, cours 1

Module de Bioinformatique
Master Recherche Biologie-Santé
Maude Pupin (maude.pupin@lifl.fr)



Planning

- 13/09 : cours
 - Matin (9h-12h) :
 - Introduction à la bioinformatique
 - Recherche bibliographique
 - Banques de données
 - Comparaison de séquences
 - Après-midi (13h30-16h30) :
 - Prédiction de gènes
 - Annotation de protéines
- 14/09 : TP (mise en pratique des cours)
 - 1 groupe le matin de 9h à 13h
 - 1 groupe l'après-midi de 14h à 18h
- Page web : <http://www.fil.univ-lille1.fr/~pupin/MRBS/>

2



Qu'est-ce que la bioinformatique ?

- L'approche *in silico* de la biologie
- Un outil indispensable aux biologistes
- Trois activités principales :
 - Acquisition et organisation des données biologiques
 - Conception de logiciels pour l'analyse, la comparaison et la modélisation des données
 - Analyse des résultats produits par les logiciels
- Un nouveau domaine de recherche
 - Logiciels souvent accessibles gratuitement via Internet

3



Quelques liens utiles en bioinformatique

-  La Société Française de BioInformatique (SFBI)
<http://sfbi.impg.prd.fr/>
-  Logiciels pour la biologie de l'Institut Pasteur
<http://bioweb.pasteur.fr/>
-  Le Pôle Bioinformatique Lyonnais (PBIL)
<http://pbil.univ-lyon1.fr/pbil.html> <http://npsa-pbil.ibcp.fr/>
-  European Bioinformatics Institute (EBI)
<http://www.ebi.ac.uk/>
-  Les outils de protéomique d'ExPASy
<http://www.expasy.org/tools/>
-  National Center for Biotechnology Information (NCBI)
<http://www.ncbi.nlm.nih.gov/>

4



Mes principales sources d'inspiration

- Sites généralistes
 - Un dictionnaire : <http://fr.wiktionary.org/>
 - Une encyclopédie : <http://fr.wikipedia.org/>
- Sites français sur la Bioinformatique
 - Infobiogen (fermé maintenant) : <http://www.infobiogen.fr/>
 - Autoformation (Paris V) : <http://www.dsi.univ-paris5.fr/bio2/autof2/>
- Sites en anglais sur la bioinformatique
 - 2can (tutoriels de l'EBI) : <http://www.ebi.ac.uk/2can/tutorials/>
 - Les aides fournis par les logiciels (et les articles scientifiques)
- Les cours de mes collègues du LIFL
 - Jean-Stéphane Varré, Hélène Touzet et Laurent Noé

5



Quelques conseils

- Méfiez-vous des résultats donnés par les logiciels :
 - La qualité des résultats est parfois diminuée au profit de la rapidité
 - Certains problèmes admettent un ensemble infini de possibilités
 - Ce n'est pas toujours la solution la meilleure qui est trouvée
 - Beaucoup de logiciels ne font que de la prédiction
 - Dire ce qu'on prévoit, par raisonnement, devoir arriver. (wiktionnaire)
 - Méfiez-vous des banques de données :
 - Les données se sont pas toujours fiables
 - La mise à jour n'est pas toujours récente
- La réalité mathématique n'est pas la réalité biologique :**
Les ordinateurs ne font pas de biologie, ils calculent ... vite !

6

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>


8

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
 - C'est une grande valeur ajoutée de cette banque
- 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

9

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'informations sur :
 <http://www.nlm.nih.gov/mesh/>

10

Indexation des articles à l'aide de MeSH

- Lecture 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

11

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)

12

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

13

Différentes hiérarchies contenant ce terme

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

14

Entrez, LE système d'interrogation de PubMed

- <http://www.ncbi.nlm.nih.gov/Entrez/>
- Développé par le NCBI (National Center for Biotechnology Information)
- Interrogation de PubMed et de nombreuses autres banques
 - Banques de séquences, structures, familles de protéines, ...
 - **Bookshelf** : collection de livres dans le domaine du biomédical
 - Affichage d'extraits relatifs à une requête
 - **OMIM** (Online Mendelian Inheritance in Man) : gènes et troubles génétiques humains recensés par Dr. Victor A. McKusick *et al.*
 - **OMIA** (Online Mendelian Inheritance in Animals) : banque similaire pour les animaux (autres que Homme et souris) gérée par Pr Frank Nicholas.

15

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 selon autorisation)
 - Lien vers PMC (PubMed Central), archive gratuite



16

Entrez, un exemple de recherche simple

Quels sont les articles écrits par des « Dupont » ?

- Saisie de **Dupont** dans la boîte de requêtes
 - Recherche du mot **dupont** dans tout le texte des articles
 - Pas spécifique aux auteurs : **9382** citations trouvées
- Saisie de **Dupont [author]**
 - Recherche du mot **dupont** dans la liste des **auteurs**
 - Recherche ciblée : **5426** citations trouvées

17

Association de critères de recherche

Utilisation d'opérateurs booléens : AND, OR, NOT

- **Dupont [author] AND Martin [author]** → 15 citations
 - Un Dupont et un Martin sont auteurs de la même citation
- **Dupont [author] OR Martin [author]** → 61.644 citations
 - Soit un Martin, soit un Dupont sont un des auteurs de la citation
- **Dupont [author] NOT Martin [author]** → 4.513 citations
 - Un Dupont est auteur de la citation, mais pas un Martin



18

Comment construire une requête ?

- Déterminer les critères de recherche
 - Ne pas oublier les déclinaisons du mot (singulier, pluriel, ...)
 - Ne pas oublier les synonymes (si possible, passer par MeSH)
- Combiner les critères avec le ou les bons opérateurs
 - Si différents critères se complètent : ET (AND)
 - Souvent, interrogation de plusieurs champs
 - Si alternative entre plusieurs termes : OU (OR)
 - Souvent, différents termes pour un même champ
 - AND est facultatif
- Limiter la recherche des critères à un champ particulier
 - Les systèmes d'interrogation peuvent limiter la recherche à une ligne particulière du fichier.
 - Entrez : saisir le nom du champ entre crochets, après le terme

19

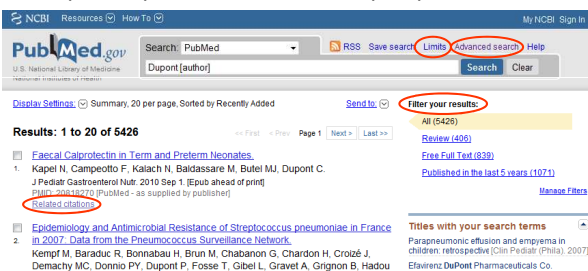
Recherche via les termes MeSH

- 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

20

Mieux cibler sa requête

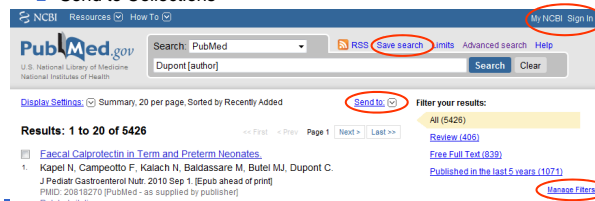
- Réduire le nombre d'articles selon certains critères
- Faire un lien vers d'autres articles sur le même thème
- Requêtes spécialisées : Clinical/Topic-specific Queries



21

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
- Mémorisation des requêtes
- Mémorisation des citations
 - Send to Collections



22

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

23

Les banques de données biologiques

❧ Pourquoi des banques de données en biologie ?

- Banques de données (selon wiktionnaire) :
 - Ensemble de données relatif à un domaine défini de connaissances et organisé pour être offert aux consultations d'utilisateurs
- Recherche publique donc données publiques
 - Partage international des données
 - Evite qu'une expérience soit refaite par différents laboratoires
 - Permet la comparaison des résultats
- Beaucoup de données générées par l'expérimentation
 - Besoin d'outils adaptés à de grands volumes de données
 - Besoin d'une gestion par des organismes spécialisés
- Accès fréquents à ces données
 - Besoin d'interrogation via Internet

25

❧ Les banques de séquences nucléiques

- Origine des données
 - Séquençage de molécules d'ADN ou d'ARN
- Les données stockées :
 - 1 séquence + ses annotations = 1 entrée
 - Fragments de génomes
 - Un ou plusieurs gènes, un bout de gène, séquence intergénique, ...
 - Génomes complets
 - ARNm, ARNt, ARNr, ... (fragments ou entiers)
- Note 1 : toutes les séquences (ADN ou ARN) sont écrites avec des T
- Note 2 : le brin donné dans la banque est appelé brin + ou brin direct

26

❧ Banques nucléiques, les débuts

- **1977** : F. Sanger met au point la méthode de Sanger pour établir le séquençage de l'ADN.
- Publication systématique des séquences dans un article
- **Début 80** : premières banques de séquences nucléiques
 - Recueil des séquences publiées dans les articles
- Puis, croissance du nombre de séquences :
 - Pas de publication systématique pour une séquence
 - Beaucoup de données à collecter
 - Gestion des données par des organismes spécialisés
 - Les séquences et leurs annotations sont soumises aux banques par les laboratoires qui ont fait le séquençage

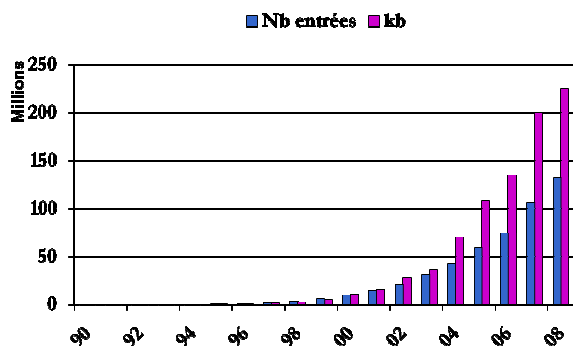
27

❧ Banques nucléiques, l'arrivée des génomes

- **1987** : Réalisation et commercialisation du 1er séquenceur automatisé (Applied Biosystems)
- **1995** : Séquençage de la 1ère bactérie, *Haemophilus influenzae* (1,83 Mb)
- **1996** : Séquençage du 1er génome eucaryote, *Saccharomyces cerevisiae* (12 Mb)
- **1998** : Séquençage du 1er organisme pluricellulaire, *Caenorhabditis elegans* (100 Mb)
- **2001** : Annonce du décryptage du génome humain
- **2010** : 8.255 projets « génome » (génomes complets ou en cours de séquençage, métagénomes)
 - cf. GOLD : <http://www.genomesonline.org/>

28

❧ Banques nucléiques, l'explosion des données



29

❧ Banques nucléiques, collaboration

International Nucleotide Sequence Database Collaboration

- Association des 3 banques nucléiques :
 - EMBL-Bank (European Molecular Biology Laboratory) – EBI
<http://www.ebi.ac.uk/embl/>
 - GenBank (banque des Etats-Unis d'Amérique) – NCBI
<http://www.ddbj.nig.ac.jp/>
 - DDJ (DNA Databank of Japon) – CIB
<http://www.ncbi.nlm.nih.gov/Genbank/>
- Echange quotidien des données
- Répartition de la collecte des données
 - Chaque banque collecte les données de son continent



30

Banques nucléiques, format d'une entrée

- 3 parties :

Description générale de la séquence

« Features »

Description des objets biologiques présents sur la séquence

La séquence

```
ctcgggcagc cccaggtcat cctgctagac tcagacctgg atgaacccat agaattgcgc 60
tcggtcaaga gccgcagcga gccgcgggag ccgccagct cctccaggt gaagcccgag 120
acacccggct cggcggcggt gccggtgag gccgcagcgg caccaccac gacggcgagag 180
```

31

EMBL, description générale de la séquence

- ID : toujours la 1ère ligne d'une entrée

Accession	Version	Topologie	Molécule	Classe	Taxonomie	Taille seq
M71283	SV 1	linear	genomic DNA	STD	BCT	1322 BP

- AC : numéros d'accension
 - Un n°acc principal pour chaque entrée, unique
 - Une liste de n°acc secondaires (historique de l'entrée)
- DT : dates de création et de dernière version
- DE : description du contenu de l'entrée
- KW : mots-clés ; peu renseigné
- OS, OC : organisme contenant la seq. et sa taxonomie
- RN, RC, RX, RP, RA, RT, RL : réf. bibliographiques
 - Uniquement les références données par les auteurs de l'entrée

32

GenBank et DDBJ, description générale

- LOCUS : toujours la première ligne d'une entrée

Locus name	Taille seq	Molécule	Topologie	Division	Date
BACCOMQP	1322 bp	DNA	linear	BCT	26-APR-1993

- DEFINITION = DE
- ACCESSION = AC
- VERSION ~ DT
- KEYWORDS = KW
- SOURCE, ORGANISM = OS, OC
- REFERENCE, AUTHORS, TITLE, JOURNAL, ... = R...

33

Banques nucléiques, lignes FT (Features)

Format (partagé par toutes les banques) :

- Key** : un seul mot indiquant un groupe fonctionnel
 - Vocabulaire contrôlé, hiérarchique
 - gene : séquence complète du gène (y compris les introns)
 - CDS : séquence codante (sans les introns, entre ATG et Stop)
- Location** : instructions pour trouver l'objet sur la séquence de l'entrée
 - Format décrit dans le transparent suivant
- Qualifiers** : description précise du groupe fonctionnel
 - Format : /qualifier="commentaires libres"
 - /gene="comQ" : nom du gène concerné
 - /note="competence regulation" : information concernant la fonction

34

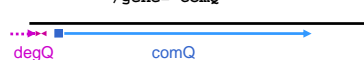
Banques nucléiques, localisation des « keys »

- 467** : l'annotation ne concerne qu'une seule base
- 109..1105** : entre les positions 109 et 1105 (incluses)
 - Toujours la position la plus petite en premier
- <1..21 ou 1275..>1322** : « Keys » tronqués
 - Commence avant le premier nt de l'entrée
 - Se termine après le dernier nt de l'entrée (taille seq = 1322)
- <234..888** : début réel inconnu, mais avant 234
- 234..>888** : fin réelle inconnue, mais après 888
- complement(340..565)** : séquence complémentaire inversée à celle de l'entrée (brin -)
- join(12..78,134..202)** : fragments indiqués mis bout à bout (concaténés) ; nombre de fragments illimité

35

Exemple de « Feature » d'une séquence ADN

```
FT CDS <1..21
FT /codon_start=1
FT /db_xref="SWISS-PROT:Q99039"
FT /transl_table=11
FT /gene="degQ"
FT /protein_id="AAA2322.1"
FT /translation="YAMKIS"
FT terminator 21..47
FT /gene="degQ"
FT promoter 109..140
FT /gene="comQ"
FT mRNA 146..1105
FT /partial
FT /gene="comQ"
```

...>>>  séquence de l'entrée

36

§ Le format FASTA

- Utilisé par les logiciels d'analyse de séquence
- Une ligne de commentaires précédée de « > »
- La séquence brute (pas d'espace, ni de nombre)

```
>Human Polycomb 2 homolog (hPc2) mRNA, partial cds
ctccggcagcccgaggtcatcctgctagactcagacctggatgaacccat
agacttgcgctcgggtcaagagccgcagcgagggcgaggagccgccagct
ccctccaggtgaagcccgagacacggcgctcgccggcggtggcggtggcg
Ggggcagcggcaccaccacgacggcgaggagaagcct
>hPc2 gene
ggacgaacctgcagagtcgctgagcaggttcaagcccttctttggaata
taattatcacccagctcaccgcgaactgcctcaccgttactttcaaggag
tacgtgacggtg
```

37

§ Banques nucléiques, inconvénients

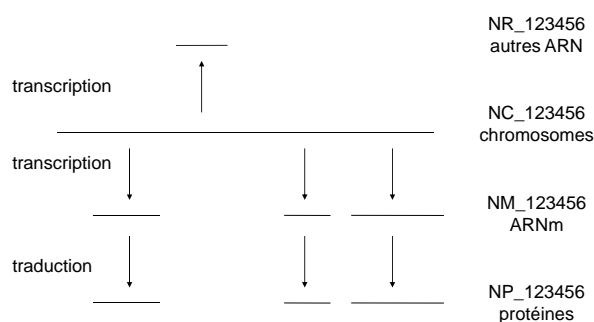
- Difficulté de mise à jour des données
 - Version plus récente d'une séquence ou d'une annotation dans d'autres banques (ex : banques dédiées à un génome complet)
- Forte redondance
 - Un même fragment de séquence présent dans plusieurs entrées
- Annotations peu normalisées
 - Difficulté de recherche d'une information précise
- Annotations peu précises
 - Peu de descriptions sur les gènes et leur produit
- Erreurs dans les annotations et dans les séquences

38

§ RefSeq : Reference Sequence collection

- Banque créée et mise à jour par le NCBI
 - <http://www.ncbi.nlm.nih.gov/RefSeq/>
- « The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms. »
- Avantages :
 - Non redondante
 - Liens explicites entre les séquences nucléiques et protéiques
 - Mise à jour régulière par le personnel du NCBI avec indication du statut de l'entrée
 - Validation des données et consistance des formats

§ Quelques numéros d'accèsion de RefSeq



§ Autres banques du NCBI

- Gene :
 - Banque centrée sur les gènes
 - Source : RefSeq
 - Localisation sur le génome, variants d'épissage, protéines codées par le gène, bibliographie, gènes homologues, ...
- UniGene :
 - Regroupement de séquences nucléiques dicté par les gènes
 - Un groupe contient toutes les séquences qui représentent un gène unique (ARNm et EST)

§ Ensembl (EBI) / UCSC Genome (USA)

- Bases de données concernant les génomes complets d'eucaryotes métazoaires
- Regroupe les annotations provenant de différentes sources
 - Gènes connus (annotations provenant d'autres banques)
 - ARNm et EST localisés sur le génome (variants d'épissage)
 - Protéines localisées sur le génome
 - Prédications statistiques
- Comparaisons entre organismes

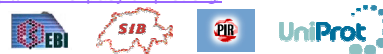
Les banques de séquences protéiques

- Origine des données
 - Traduction automatique des séquences d'ADN ou d'ARNm
 - 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

43

Banques « protéiques », les débuts

- 1956 : F. Sanger établit la séquence en aa de l'insuline
- 1965 : Atlas of Protein Sequences, M. Dayhoff
 - Version papier jusqu'en 78, puis version électronique
- 1984 : création de PIR-NBRF (Protein Information Resource - National Biomedical Research Foundation)
 - Collaboration avec MIPS (Allemagne) et JIPID (Japon)
- 1986 : création de SwissProt
 - Collaboration entre SIB (Swiss Institut of Bioinformatics) et EBI
- Fin 2003 : UniProt (Universal Protein Resource)
 - Mise en commun des informations de PIR et SwissProt/TrEMBL
 - <http://www.expasy.uniprot.org/>



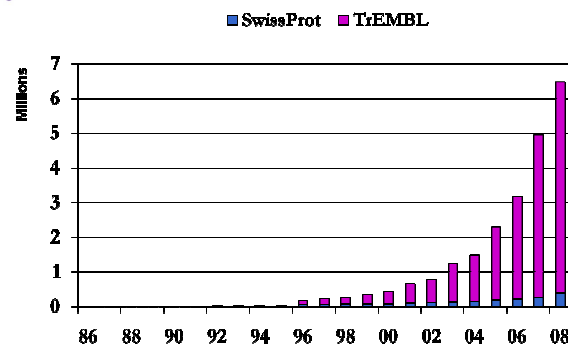
44

UniProt, ses deux parties

- SwissProt
 - Données corrigées et validées par des experts
 - Haut niveau d'annotation
 - Description de la fonction (références associées)
 - Localisation des domaines fonctionnels
 - Modifications post-traductionnelles
 - Existence de variants, ...
 - Redondance minimale
 - Nombreux liens vers d'autres banques (60 BD)
- TrEMBL
 - Entrées supplémentaires à SwissProt (pas encore annotées)
 - Traduction automatique de l'EMBL

45

SwissProt/TrEMBL, nombre d'entrées



46

SwissProt/TrEMBL, format d'une entrée

- Format basé sur celui de l'EMBL
 - Mot-clé de 2 lettres au début de chaque ligne
 - Format différent pour les Features
- Mots-clés supplémentaires :
 - GN : les différents noms du gène qui code pour la protéine (OR) et 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
 - DR : références vers d'autres banques de données
 - Vers les séquences nucléiques (EMBL/GenBank/DBJ)
 - Vers les structures 3D
 - ...

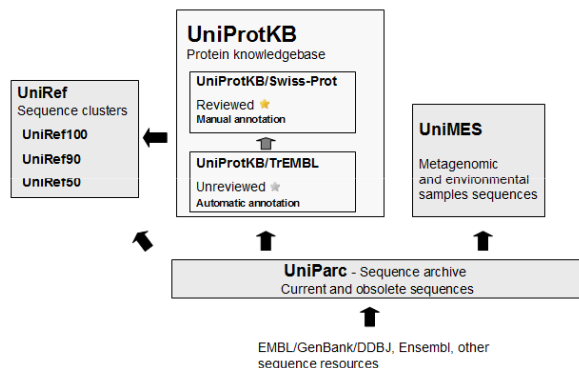
47

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)
 - ...

48

UniProt, vue générale



49

Comment et pourquoi interroger une banque ?

Comment ?

- Recherche dans les annotations
 - Systèmes d'interrogation de banques de données : Entrez, SRS
- Recherche dans les séquences
 - Comparaison d'une séquence aux séquences de la banque : BLAST, FASTA

Pourquoi ?

- Obtenir des informations nouvelles et pertinentes
- Aide à la mise au point d'expériences
- Validation des résultats d'une expérience
- ...

50

Interrogation des annotations d'une banque

- Recherche de mots ou expressions dans le texte des entrées
- Contraintes pour un système d'interrogation
 - Obtention de données pertinentes (pas trop de résultats, mais tous ceux relatifs à notre problématique)
 - Simplicité d'utilisation (syntaxe d'interrogation intuitive)
 - Réponse rapide
 - Possibilité d'analyse des résultats (couplage à d'autres outils)
- Systèmes d'interrogation
 - Entrez, permet aussi d'interroger des banques de séquences, ...
 - Même fonctionnement que pour interroger PubMed
 - SRS : un autre système d'interrogation, une autre philosophie
 - <http://srs.ebi.ac.uk/>

51

Recherche parmi les séquences d'une banque

- Pourquoi ?
 - Savoir si ma séquence ressemble à d'autres déjà connues
 - Trouver toutes les séquences d'une même famille
 - Rechercher toutes les séquences qui contiennent un motif donné
- Comment ?
 - Développement de programmes dédiés à la comparaison d'une séquence à toutes les séquences d'une banque
 - BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>, ...)
 - FASTA (<http://www.ebi.ac.uk/fasta/>)
 - YASS (<http://bioinfo.lifl.fr/yass/>)
 - ...

52

Comparaison d'une séquence à une banque

- Identifier les [fragments de] séquences de la banque ayant une forte ressemblance avec la séquence requête
- Requête : une séquence (ADN/ARN ou protéine)
- Résultat : une liste de séquences ressemblant à la séquence entrée
 - Ressemblance = bonne correspondance entre les lettres des deux séquences
 - Représentée grâce à la construction d'un alignement
 - Significativité statistique de la ressemblance estimée à l'aide du calcul d'une « e-value »
 - Tri des résultats pertinents
 - Plus la e-value est proche de 0, mieux c'est (en général, les résultats pertinents sont <1)

53

Un exemple : BLAST

- Développé en 1990 par Altschul, Gish, Miller, Myers and Lipman (NCBI)
- Différents programmes :

Prog	Requête	Banque	Utilisation type
BlastN	ADN/ARN	ADN/ARN	Localisation d'ARN sur les génomes
BlastP	Protéine	Protéines	Etude de familles de protéines
BlastX	ADN/ARN traduit	Protéines	Prédiction de gènes et de CDS
TblastN	Protéine	ADN/ARN traduits	Localisation du gène codant la protéine dans les génomes
TblastX	ADN/ARN traduit	ADN/ARN traduits	Annotation de séquences

54

Affichage des résultats (1/2)

1. Représentation graphique des résultats



2. Liste des séquences de la banque possédant au moins une région commune avec la séquence requête

- N°acc ; description de l'entrée ; Score de l'alignement ; e-value
- Classement des entrées selon leur e-value

Sequences producing significant alignments:	Score (bits)	E Value
gi120594329 gb CA54446.1 Homo sapiens insulin (Homo sapiens ce...	1.69	2e-41
gi145374727 gb U001141.1 P Proteins precursor (Homo sapien...	1.69	2e-41
gi157112977 cd U001099P2.1 proinsulin precursor (Dan. tre...	1.69	2e-41
gi12086868 cd AA747123.1 synthetic preproinsulin	1.69	4e-41
gi148418141 gb G020719.1 Insulin precursor (Canis lupus...	1.69	5e-41
gi14682888 gb U001447.1 Insulin precursor (Canis lupus...	1.69	7e-41
gi11663751 gb C034074.1 Insulin precursor (Canis lupus...	1.67	1e-40

55

Affichage des résultats (2/2)

3. Présentation des alignements obtenus pour chaque séquence sélectionnée de la banque

```
>gi|2497407|sp|Q62587|INS_PSAOB Insulin precursor [Contains: Insulin B
chain; Insulin A chain]
gi|1370283|emb|CAA66897.1 preproinsulin [Psammomys obesus]
Length=110
```

Score = 142 bits (357), Expect = 6e-33, Method: Composition-based stats.
Identities = 70/86 (81%), Positives = 75/86 (87%), Gaps = 0/86 (0%)

Query	25	FVNQHLGSGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEG	84
		FVNQHLGSGSHLVEALYLVCGERGFFYTPK RR +D Q+ Q+ELGG PGAG L+ LALE	
Sbjct	25	FVNQHLGSGSHLVEALYLVCGERGFFYTPKFRGVDDPQMPQLLGSGPGAGDLRALALEV	84

```
Query 85 SLQKRGIVEQCCTSICSLYQLENYCN 110
+ QKRGIVEQCCT ICSLYQLENYCN
Sbjct 85 ARQKRGIVEQCCTGICSLYQLENYCN 110
```

56

Pourquoi comparer des séquences ?

- **Du point de vue biologique :**
 - Beaucoup de gènes et de protéines appartiennent à des familles possédant des fonctions similaires ou partageant une origine commune (un ancêtre commun).
- **Mise en évidence de points communs au niveau fonctionnel, structural ou de l'origine :**
 - séquences similaires ↔ fonction similaire, structure similaire.
 - Découverte de domaines similaires permettant d'attribuer une fonction putative lorsque la fonction de la séquence n'est pas connue.

58

Les familles de protéines

- Différentes protéines qui possèdent des fonction proches
 - Ex : Catalyser la polymérisation de l'ADN, réguler les gènes impliqués dans la synthèse du tryptophane, ...
- Ce sont des protéines dites homologues
 - Elles ont un ancêtre commun (un gène dans un organisme)
- Ce sont souvent des protéines similaires
 - Ressemblance au niveau de leur séquence (> 30% identité)
 - Mais des protéines avec des séquences différentes peuvent avoir des fonctions proches (ressemblance en 3D)

59

Evolution d'une famille de gènes

- **Spéciation :**
 - Naissance d'une nouvelle espèce
 - Gènes issus du même ancêtre dans **des espèces différentes**
 - **Gènes orthologues**
- **Duplication :**
 - Doublement d'un gène **au sein d'un génome**
 - Evolution indépendante des deux gènes
 - **Gènes paralogues**
 - Possibilité d'inventer de nouvelles fonctions (un des 2 gènes subit des mutations et l'autre garde la fonction d'origine)

50

§ Mutations au niveau de l'ADN

- Substitution : changement d'un nucléotide par un autre au moment de la réplication
- Insertion-délétion : ajout ou suppression d'un fragment d'ADN (plusieurs causes possibles, différentes échelles)
- Duplication : doublement d'un fragment d'ADN (duplication de gènes ou de fragments de chromosomes)
- Recombinaison : échange de fragments de séquences entre chromosomes
- Inversion : Changement de sens d'un fragment d'ADN

61

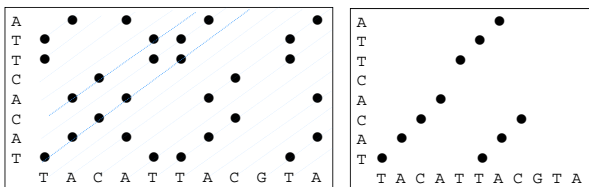
§ Des conséquences plus ou moins graves

- **Mutations neutres :**
 - Pas dans un gène
 - Pas de changement d'aa (codons synonymes)
 - Changement d'un aa par un autre équivalent
- **Mutations défavorables :**
 - Altèrent la fonction de la protéine
- **Mutations bénéfiques :**
 - Améliorent le fonctionnement d'une protéine
 - Invention d'une nouvelle fonction
- **Mutations létales :**
 - Rendent une protéine vitale non fonctionnelle

62

§ Dotplot ou graphe par matrice de points

- **Idee simple :** localiser les positions où les lettres des deux séquences sont identiques
- Outil graphique (Staden, 1982)
 - Abscisse = une séquence ; Ordonnée = autre séquence
 - Point = présence de deux lettres identiques
- Dotplots filtrés : mots exacts ou avec erreurs



63

§ Dotplots

Avantages

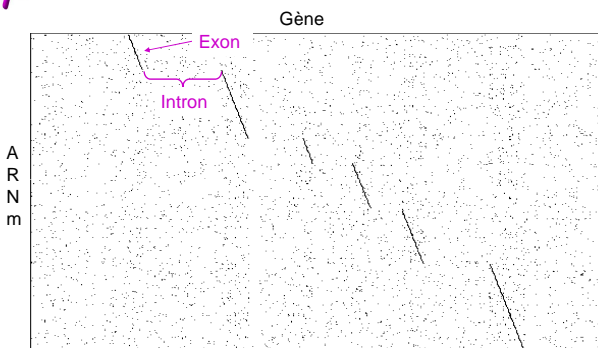
- Principe basé sur le visuel
 - L'œil est un organe très performant
- Pas ou peu de perte d'information
 - Le filtrage n'est pas obligatoire
- Aucun autre programme ne peut fournir autant d'informations

Inconvénients

- Difficulté pour retrouver les diagonales observées
- Pas toujours possibilité de construire un alignement linéaire
- Pas de traitement automatisé possible de l'information
- Pas de mesure de similarité possible
- Difficulté d'évaluation de la pertinence de l'information

64

§ Dotplot d'un gène d'actine avec son ARNm



65

§ Autre possibilité : l'alignement

- Mise en correspondance des lettres d'une séquence par rapport à une autre
- Ordre des lettres conservé
- Trois paires possibles :
 - Deux résidus identiques
 - ☞ Appariement (match)
 - Deux résidus différents
 - ☞ Mésappariement (mismatch)
 - Un résidu avec rien
 - ☞ Insertion ou délétion (indel, gap)

ACGGCTA-TC
| | | | |
ACTG-TAATG

66

Quel alignement construire ?

- Nombreux alignements possibles entre 2 séquences
- Besoin d'un guide pour construire le "bon" alignement
- Choix d'un critère de qualité : score de similarité
 - Somme des poids de toutes les paires de l'alignement
- Recherche de l'alignement de score optimal
 - Poids donnés aux opérations élémentaires (paires) :
 - Lettres identiques
 - Lettres différentes (matrices de substitution)
 - Insertion-délétion (pénalités d'indel)

```

ACGGCTATC   ACGGCTA-TC   AC-GGCTA-TC
||| ||| |||  ||| ||| |||  ||| ||| |||
ACTGTAATG   ACTG-TAATG   ACTG--TAATG
  
```

67

Alignement global ou local

- Alignement global :
 - Premier algorithme d'alignement
 - Needleman et Wunsch, 1970
 - Aligne les séquences sur toute leur longueur
- Alignement local :
 - Smith et Waterman, 1981
 - Modification du programme de Needleman et Wunsch
 - Recherche de la région de plus forte similarité entre les séquences
- Ce sont des algorithmes exacts qui garantissent de construire l'alignement optimal

68

Comparaison des résultats

- Alignement global


```

GGCTGACCACCTT
| | | | |
GA-TCACTTCCATG
      
```
- Alignement local


```

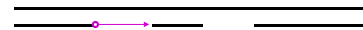
5 GACCACCTT 13
  ||| ||| |||
1 GATCAC-TT 8
      
```



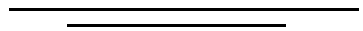
69

Problème des insertions-délétions

- **Biologie** : indel de plusieurs lettres = UNE SEULE mutation
- **Algorithmes** : plus l'indel est longue, plus la pénalité est forte, moins le programme considère cette éventualité
- **Parade** : Pénalité de création d'indel + Pénalité d'extension



- Pas de pénalités d'indel en début et en fin d'alignement
 - Glissement d'une séquence par rapport à l'autre



70

Matrices de substitution nucléiques

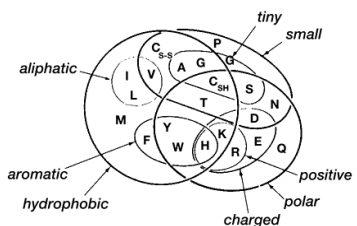
- Donnent un coût (un poids) aux paires de lettres dans les alignements.
- Pour les nucléotides, matrice simple :
 - Une valeur pour les matches et une pour les mismatches

	A	C	G	T	
A	5	-4	-4	-4	
C	-4	5	-4	-4	match = 5
G	-4	-4	5	-4	mismatch = -4
T	-4	-4	-4	5	

71

Matrices de substitution protéiques

- Certains acides aminés ont des propriétés physico-chimiques proches
- Matrices avec des coûts différents selon les paires
 - Problème : comment estimer ces coûts ?



72

§ Matrices basées sur l'évolution

- **Idée** : La nature tolère certaines mutations
 - Ces mutations peuvent être observées
- **Construction** :
 - Alignement multiple d'une famille de protéines
 - Comptage de toutes les paires d'aa observées dans chaque colonne
 - Calcul d'un poids pour le changement d'un aa i par un autre j
- **Interprétation** de « Poids (i,j) » :
 - la paire (i, j) est plus fréquente que ce qui est attendu par hasard
 - Donc : la mutation de i vers j est tolérée par la nature
 - OU la paire (i, j) est moins fréquente que ce qui est attendu par hasard
 - Donc : la mutation de i vers j est évitée par la nature

73

§ PAM (M. Dayhoff, 1970 puis 90)

- Construction à partir d'alignements globaux
- PAM-1 : 1 mutation ponctuelle pour 100 acides aminés
 - Peu de divergence entre les séquences, forte similarité
- $PAM-250 = (PAM-1)^{250}$: 250 mutations pour 100 aa
 - Calculée en élevant à la puissance 250 la matrice PAM-1
 - Long "temps" de divergence entre les séquences
- Plus la valeur de PAM est élevée, plus le nombre d'aa considérés comme équivalents est grand (plus la matrice est tolérante aux mutations)

74

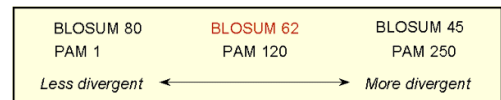
§ BLOSUM (Henikoff and Henikoff, 1990)

- Construction à partir d'alignements locaux
 - Les régions d'une protéine n'évoluent pas toutes à la même vitesse
- Chaque matrice est calculée à partir d'un jeu de données réelles
 - Sélection de séquences ayant un %id donné
- BLOSUM-62 = au moins 62% d'identité entre les séquences du jeu de données
- Plus la valeur de Blosum est élevée, plus le nombre d'aa considérés comme équivalents est petit (moins la matrice est tolérante aux mutations)

75

§ Comparaison PAM et BLOSUM

- Mieux vaut utiliser BLOSUM
- BLOSUM-62 est souvent proposée par défaut
- PAM est utile pour les valeurs extrêmes



76

§ Les scores d'alignement

- Calcul du score d'alignement :
 - Somme des poids de chaque paire de l'alignement
- Dépend de la matrice et des pénalités d'indel
 - Variations pour un même jeu de séquences
 - Ex : DHLPBA score : -14 avec PAM30 ; 8 avec PAM250
GRFSKG
- Score dépendant de la longueur des séquences alignées
 - Pas comparable entre deux alignements
- § N'évalue pas le degré de ressemblance des séquences
 - Pas de score seuil au delà duquel un alignement est bon

77

§ Evaluation de la ressemblance entre séquences

- Par comptage :
 - Identité : Nombre de paires identiques / nombre total de paires
 - Similarité : Nb (paires similaires et id) / nombre total de paires
 - Paires similaires calculées grâce à la matrice de substitution
 - Valeurs comparables entre alignements
- Par calcul de la probabilité d'observer un alignement aussi bon par hasard
 - Z-score : utilisé pour les alignements de deux séquences
 - Mélange d'une des deux séquences (>100 fois)
 - Calcul des scores d'alignement avec une séquence mélangée
 - Constitution d'un histogramme de scores aléatoires
 - Calcul du Z-score : (score – moyenne) / écart type
 - Comparaison du score initial avec les scores aléatoires

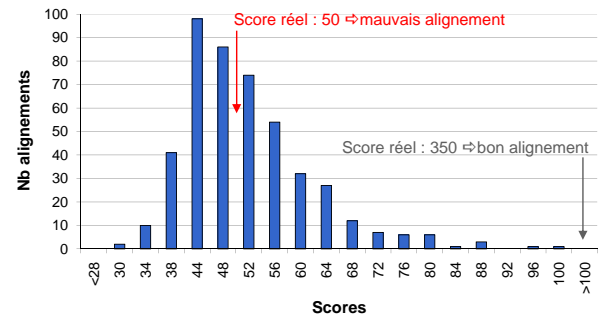
78

Homologie ?

- Séquences protéiques ≥ 100 aa, identité $\geq 25\%$
 - Ancêtre commun
- Séquences ADN ≥ 100 nt, identité $\geq 50\%$
 - Pas forcément de relation biologique
- Mauvais alignement si
 - Plus d'une insertion tous les 20 aa
 - Faibles changements des pénalités de gap \Rightarrow changements dans l'alignement

79

Histogramme des scores



80

L'alignement multiple de séquences

- Plus de 2 séquences sont alignées en même temps
- **But** : étudier une famille de séquences
- **Problème** : les algorithmes utilisés pour la comparaison de 2 séquences sont trop gourmands en calculs
 - Besoin de simplifier les calculs au détriment de la qualité des résultats
- Chaque programme est efficace sur un type de données particulier

81

Les principaux programmes d'ali mult

- Clustalw (Thompson *et al*, 1994) : mutations locales
 - Alignement de toutes les paires de séquences
 - Construction du dendrogramme (arbre)
 - Alignement progressif des séquences, dans l'ordre du dendrogramme
- Dialign (Morgensten *et al*, 1999) : blocks conservés
 - Identification des régions conservées (diagonales)
 - Gestion des conflits entre diagonales
 - Construction de l'alignement à l'aide des diagonales
- Correction manuelle de l'alignement souvent nécessaire
 - Intégration des connaissances biologiques

82

Un exemple d'alignement multiple : l'insuline

	10	20	30	40	50	60
Human	MALWMRLLPALLALLAWGPDPAFAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED					
Gorilla	MALWMRLLPALLALLAWGPDPAFAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED					
Chimpanzee	MALWMRLLPALLALLAWGPDPAFAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED					
Pig	MALWMRLLPALLALLAWGPDPAFAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED					
Chicken	MALWIRSLPALLALLVFSFGPTSYAANQHLGSHLVEALYLVCGERGFFYSPKARRDVEQ					
	70	80	90	100	110	
Human	LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN					
Gorilla	LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN					
Chimpanzee	LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN					
Pig	PQAGAVELGGLG--LQALALEGPPQKRGIVEQCCTSIICSLYQLENYCN					
Chicken	PLVSS-PLRGEAGVLPFQQBEYEK--VKGRIEVCCHNTCSLYQLENYCN					

Identity (*) : 67 is 60.91 %
 Strongly similar (:) : 7 is 6.36 %
 Weakly similar (.) : 9 is 8.18 %
 Different : 27 is 24.55 %

83