La génomique

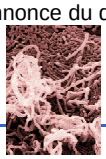
Maude Pupin
(et extraits de cours de Laurent noé)

La génomique (source : Infobiogen)

- La génomique est l'étude exhaustive des génomes et en particulier de l'ensemble des gènes, de leur disposition sur les chromosomes, de leur séquence, de leur fonction et de leur rôle.
- Les génomes des organismes vivants ont des tailles considérables allant d'une centaine de millions à des milliards de nucléotides (3 milliards pour le génome humain).

Les débuts de la génomique

- **1953** : Watson et Crick découvrent la structure de l'ADN
- **1956** : F Sanger établit la séquence en aa de l'insuline
- **1977** : F Sanger met au point le séquençage de l'ADN
- **1987** : Premier séquenceur automatisé
- **1995** : Séquençage du 1er génome bactérien
  - *Haemophilus influenzae* (1,83 Mb)
- **1996** : Séquençage du 1er génome eucaryote
  - *Saccharomyces cerevisiae* (12 Mb)
- **2001** : annonce du décryptage du génome humain



Qu'est-ce que le séquençage ?

- Déterminer l'ordre linéaire des composants d'une macromolécule (aa d'une protéine, nt de l'ADN, …)
- Le séquençage des protéines
  - Nécessite un matériel dédié qui est cher
  - Technique délicate à mettre en œuvre
  - La séquence des protéines peut-être déduite de l'ADN
- Le séquençage de l'ADN
  - Plus simple à mettre en œuvre
  - Technique très répandue, beaucoup de laboratoires possèdent un petit séquenceur automatique

Les techniques de séquençage

- Méthode Sanger (1975)
- Méthode de Maxam–Gilbert (1977)
- Automatisation de Sanger (de ~1980 à 2005)
  - Commercialisée en 1987 : premier séquenceur *Applied Biosystems 370A*
- Nouvelles Générations de Séquenceurs (depuis 2005)
  - *NGS : Next Generation Sequencing (désormais largement utilisés)*
  - *HTS : High-Throughput Sequencing*
- *NNGS : Next-Next Generation Sequencing (en cours)*
  - en particulier technologie *SMS (Single Molecule Sequencing)*

Voir cours de Laurent Noé (http://www.lifl.fr/~noe/enseignement/m1-genpro/Cours/bioinfo_bio1.pdf)

Bilan des projets « génomes » en 2011

Voir
Genome Online Database
http://www.genomesonline.org/

ARCHAEA TOTAL: 329
BACTERIA TOTAL: 8473
EUKARYA TOTAL: 2204

soit 11006 séquences de génomes ou transcriptomes !

Le séquençage ponctuel

- L'explosion du nombre de génomes séquencés est récente
- Les scientifiques séquencent depuis longtemps des fragments de génomes, selon leurs besoins :
  - Séquençage de régions d'intérêts si le génome complet n'est (n'était) pas encore connu
  - Séquençage dans le but d'étudier les variations alléliques (la même région dans des individus différents d'une même espèce)
  - Séquençage d'un ou plusieurs ARN pour localiser des gènes sur un génome et étudier leur régulation transcriptionnelle
  - …

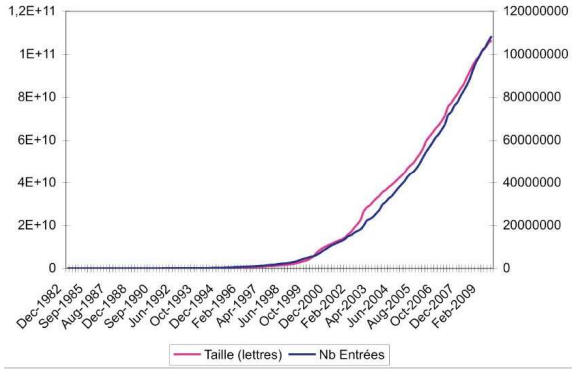
Mise à disposition des séquences

- Les séquences obtenues dans des laboratoires publics sont mises à disposition de l'ensemble de la communauté scientifique
  - Collecte des séquences par des organismes spécialisés
  - Stockage des séquences dans des banques de données, sous la forme de fichiers texte formatés
  - Les séquences sont annotées (localisation des gènes, …) et leur provenance est précisée (nom de l'espèce, laboratoire, …)
  - Les banques de données sont maintenant accessibles via Internet

http://www.ebi.ac.uk/embl/ http://www.ddbj.nig.ac.jp/

http://www.ncbi.nlm.nih.gov/Genbank/

Banques nucléiques, croissance

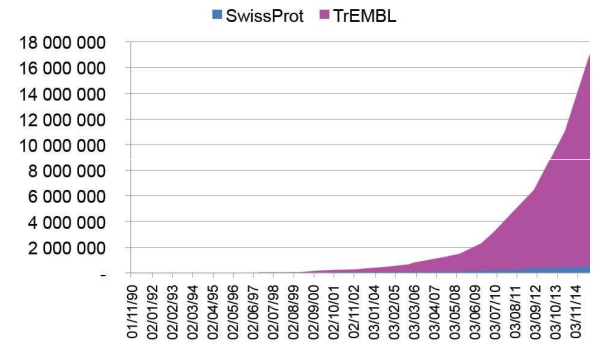


Les séquences protéiques disponibles

- Les banques produisent elles-mêmes les données
  - Traduction automatique des séquences ADN et ARNm
  - Peu de séquençage de protéines car long et coûteux
- Deux types de banques
  - Annotation « complète » et produite par des experts
  - Ex : Banque SwissProt (24/07/2007 : 276.256 entrées)
  - Annotation « légère » et produite par analyse informatique
  - Ex : Banque TrEMBL (traduction EMBL) (4.672.908 entrées)

http://www.expasy.uniprot.org/

SwissProt/TrEMBL, nombre d'entrées



La bioinformatique

Définition de la bioinformatique

Un domaine de recherche qui analyse et interprète des données biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie.

Source : article présentant la bioinformatique, sur le site d'*Interstices*
Auteur(s) :
Isabelle Quinkal (Journaliste)
François Rechenmann (Chercheur)

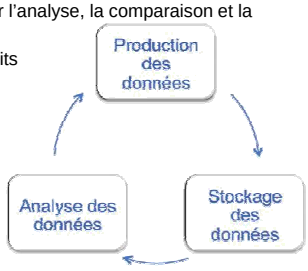
Définition de la bioinformatique

en anglais, il y a distinction entre « Bioinformatics » et « Computational Biology »

- ***« Bioinformatics »***
  - *applique des algorithmes, modèles statistiques dans l'objectif d'interpréter, classer et comprendre des données biologiques.*
- ***« Computational Biology »***
  - *développer des modèles mathématiques et outils associés pour résoudre des problèmes biologiques.*

Qu'est-ce que la bioinformatique ?

- L'approche *in silico de* la biologie
- 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



Prédiction de gènes

Pourquoi annoter les séquences ?

- La séquence d'ADN est produite brute
  - Pas d'information sur la position des gènes, …
  - Besoin de « décoder » le message du génome
- Les expériences en laboratoire fournissent de nombreuses données
  - Etude d'un gène et de son produit (fonction de la protéine)
  - Extraction d'ARNm
  - Nombreuses publications et informations dans les banques
- Possibilité de croiser les informations pour améliorer la qualité des annotations

Comment annoter les séquences ?

- Etude de la séquence ADN pour localiser les gènes
  - Traitement informatique essentiellement basé sur la comparaison de séquences
  - Traitement manuel avec l'expertise humaine qui valide ou non les résultats proposés par les logiciels et approfondit l'étude
- Puis étude de la protéine pour prédire sa fonction
  - Prédiction des structures 2D et 3D
  - Prédiction de la localisation cellulaire
  - Prédiction des domaines fonctionnels
  - Intégration dans les réseaux cellulaires

Quelles sont les méthodes de prédiction de gènes ?

- Détection des ORF (Open Reading Frame)
  - Méthode naïve
  - Localisation des régions de plus de 99 bp entre un codon d'initiation (Cinit) et un codon de terminaison (Cterm)
- Comparaison aux banques
  - Méthode exploitant les données disponibles
  - Recherche des séquences d'ARNm et de protéines qui ressemblent à la séquence étudiée
- Etude statistique (ab initio)
  - Localisation des séquences codantes et non codantes sur la base de critères discriminants

Une idée simple : les phases ouvertes de lecture

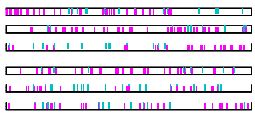
- Une séquence codante :
  - Débute par un codon d'initiation (Cinit) (ATG + autres) et se termine par un codon de terminaison (Cterm) (TAA, TAG, TGA)
  - A une taille multiple de 3 (si les introns sont enlevés)
  - Taille moyenne : 1.000 bp (bactéries)
- Une phase ouverte de lecture (ORF, Open Reading Frame)
  - Plus de 99 nt entre un Cinit et un Cterm (statistiquement rare)
  - Peut contenir un gène
- Problèmes :
  - Un gène peut être sur un brin ou sur l'autre
  - Plusieurs phases de lecture possibles

Détection des ORF, fonctionnement

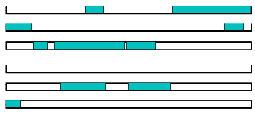
- Traduction à l'aveugle
  - → 6 phases de lecture
  - = 6 séquences protéiques possibles

Détection des ORF, résultats

- 6 phases de lectures :
  - codons d'initiation (ATG)
  - codons de terminaison (TAA, TAG, TGA)



- Sélection des phases ouvertes de lecture (ORF)
  - Régions mesurant plus de 99 nt entre un Cinit et un Cterm
  - Choix du Cinit le plus loin du Cterm
  - Peut contenir un gène
  - une ORF



ATTENTION : ORF ne veut pas dire gène !

Comparaison de séquences

- Possibilité d'isoler puis séquencer un ARNm (*in vivo*)
  - Comparaison de l'ARNm au génome pour localiser le gène
  - Détermination des positions de début et de fin du gène, ainsi que des introns (car ARNm mature)
- Nombreuses séquences de protéines dans les banques
  - Comparaison de l'ADN aux protéines pour trouver des protéines de même fonction
  - Détermination des positions de début et de fin de la séquence codante, ainsi que des introns car ARNm mature

Usage du code

- N codons codent le même aa (codons synonymes)
- Pour un aa donné, il y a un codon préféré
  - Différences entre gènes selon leur taux d'expression (classe)
  - Les gènes « de ménage » (nécessaires au fonctionnement de toutes les cellules) partagent le même usage du code
  - Différences entre organismes selon leur pourcentage en G+C
  - Choix des codons riches en GC dans les génomes riches en GC
- Les séquences codantes suivent l'usage du code de leur organisme et de leur classe
- Les séquences non codantes n'ont pas de pression de sélection pour l'usage du code

Prédiction statistique

- Apprentissage de l'usage du code pour un organisme donné à partir d'un ensemble fiable de séquences codantes
- Détermination de classes de gènes avec des usages du code différents au sein de l'organisme
- Calcul de la probabilité pour qu'une fenêtre soit codante
  - Une fenêtre est une suite de lettres dans une séquence
- Analyse des résultats obtenus en faisant coulisser la fenêtre le long de la séquence étudiée

Difficulté de prédiction des gènes avec introns

- Taille des introns non multiple de 3
  - Changement de phase d'un exon à l'autre
  - Pas de changement de brin
- Existence d'exons courts (~10nt)
  - Taille en dessous des limites de résolution des logiciels
- Existence d'introns très longs (plus longs que les exons)
  - Difficulté pour localiser les exons (ils sont noyés)
- Un intron peut couper un codon en deux

Les familles de protéines

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
- 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)

Mutations dans l'ADN ⇨ évolution des protéines

- 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

Des mutations 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

Evolution d'une famille de protéine

- Spéciation :
  - Naissance d'une nouvelle espèce
  - Gènes issus du même ancêtre dans des espèces différentes
  - Gène orthologues
- Duplication :
  - Doublement d'un gène
  - 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)

Etude bioinformatique d'une famille de protéines

- Recherche dans les banques de protéines similaires à une protéine donnée
  - Constitution de la famille
  - Nécessité de définir des critères pour accepter une séquence dans une famille (statistiques et/ou biologiques – règles –)
- Alignement des protéines (2 à 2 ou famille entière)
  - Détermination des acides aminés communs à une ou plusieurs séquences
    - Notion d'aa équivalents, c'est-à-dire pouvant être échangés sans altérer la fonction de la protéine
    - L'ordre des aa dans les protéines et maintenu, possibilité d'insérer des « blancs » (gaps) pour décaler un block de protéine
- Détermination de blocks conservés
  - Conservation => important pour la fonction de la protéine

Un exemple d'alignement multiple : l'insuline

```
                    10        20        30        40        50        60
                     |         |         |         |         |         |
Human       MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
Gorilla     MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
Chimpanzee  MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
Pig         MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAEN
Chicken     MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ
            **** * ****.**.: .* .: * .************************:**:**:.*:
                    70        80        90       100       110
                     |         |         |         |         |
Human       LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
Gorilla     LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
Chimpanzee  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
Pig         PQAGAVELGGGLGG--LQALALEGPPQKRGIVEQCCTSICSLYQLENYCN
Chicken     PLVSS-PLRGEAGVLPFQQEEYEK--VKRGIVEQCCHNTCSLYQLENYCN
              ..    * *  *   .*     *    ********* . ***********
```

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