

La génomique

UE Algorithmes pour la bioinformatique,
Master recherche Informatique
Maude Pupin

La génomique (source : Infobiogen)

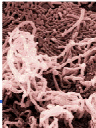
- La génomique est l'étude exhaustive des génomes et en particulier de l'ensemble des gènes, de leur disposition sur les chromosomes, de leur séquence, de leur fonction et de leur rôle.
- Les génomes des organismes vivants ont des tailles considérables allant d'une centaine de millions à des milliards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

La génomique

Présentation

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



La génomique

Présentation

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

La génomique

Présentation

Limites de la technique de séquençage

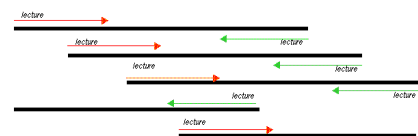
- Une manip de séquençage ne peut pas déterminer plus de 500 à 1.000 nucléotides à la suite
 - C'est très peu par rapport à un chromosome entier !
 - Besoin de couper les chromosomes en fragments puis de les reconstituer
- Il peut y avoir des erreurs de lecture
 - Oubli d'un nucléotide
 - Inversion de l'ordre des nucléotides, ...
 - Besoin de séquencer plusieurs fois un fragment pour obtenir une séquence fiable

La génomique

Présentation

Principe du séquençage d'un chromosome

- Amplification du chromosome
 - Besoin de séquencer 8 à 10 fois le chromosome
- Les copies du chr sont cassées aléatoirement en fragments de quelques milliers de nucléotides
- Séquençage des extrémités de certains des fragments obtenus
 - Certaines séquences se chevauchent en partie



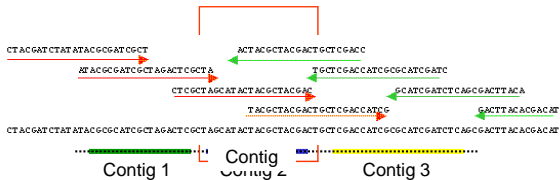
Source :
Genoscope

La génomique

Présentation

Reconstruction des contigs

- Comparaison des séquences obtenues pour aligner les parties séquencées plusieurs fois
 - Reconstitution d'enchaînements plus grands, appelés contigs
 - Traitement informatique indispensable



La génomique

Présentation

Finition (assemblage final)

- Il reste à ordonner et orienter les contigs
 - Difficulté : présence de répétitions dans les génomes qui peuvent conduire à assembler des contigs provenant de régions distantes du chr
 - Présence de « trous » qui sont comblés par un séquençage ciblé
- Correction des erreurs
 - Renouvellement du séquençage pour améliorer la qualité de la séquence
- Etape pas toujours réalisée
 - Le but est alors d'avoir rapidement l'ébauche d'un génome pour, par exemple, le comparer à une espèce proche.

La génomique

Présentation

Le séquençage du génome humain

- Le projet a commencé dans les années 1990
 - Etape préalable : cartographie des chromosomes (localisation de marqueurs) pour faciliter l'assemblage
 - 1998 : début du séquençage
 - Le travail a été réparti entre 20 institutions internationales réunies dans un consortium public
- Ebauche préliminaire célébrée en juin 2000
 - Génome séquencé seulement 5 fois
 - Entre 400.000 et 600.000 fragments de 5.000 à 6.000 nt pas toujours orientés et ordonnés
 - 90% du génome couvert, avec une erreur tous les 1.000 nt

La génomique

Présentation

La finition du génome humain

- Avril 2003 : achèvement du génome humain
 - 2 ans d'avance sur le calendrier initial grâce aux progrès des techniques de séquençage
 - Séquençage de 5 autres « équivalents génomiques »
 - Taux d'erreurs d'un nucléotide tous les 10.000
 - Travail focalisé sur les trous résiduels
 - Il en reste moins de 400
 - 2,9 milliards de nucléotides, soit 90% des 3,2 milliards de nucléotides de l'ensemble du génome humain
 - Le reste du génome est constitué de séquences répétées (notamment au niveau des centromères et télomères)



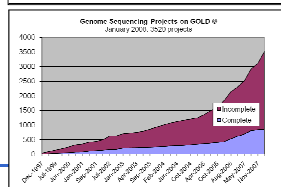
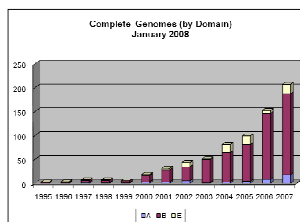
La génomique

Présentation

Bilan des projets « génomes » en 2008

- Genome Online Database <http://www.genomesonline.org/>
- **840** génomes complets
 - 693 eubactéries
 - 53 archaebactéries
 - 94 eucaryotes
- **1883** génomes procaryotes en cours de séquençage
- **97** génomes d'archae en cours
- **950** génomes eucaryotes en cours de séquençage
- **130** métagénomes

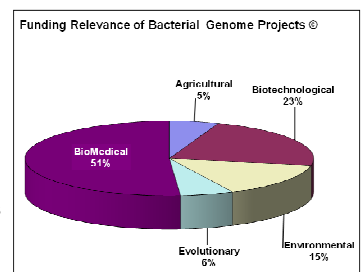
Soit un total de **3.900** projets !!



La génomique

Pourquoi séquençer les génomes ?

- Intérêt économique
 - Médecine
 - Biotechnologies
 - Environnement
- Intérêt scientifique
 - Evolution des espèces
 - Fonctionnement des cellules
 - Etude des êtres vivants
- Utilité publique
 - Nutrition
 - Propagation des maladies
 - Environnement



La génomique

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
 - ...

La génomique

Présentation

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

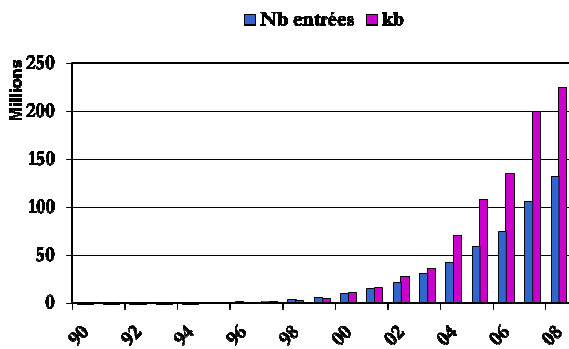
<http://www.ebi.ac.uk/embl/> <http://www.ddbj.nig.ac.jp/>
<http://www.ncbi.nlm.nih.gov/Genbank/>



La génomique

Présentation

Banques nucléiques, croissance



La génomique

Présentation

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)

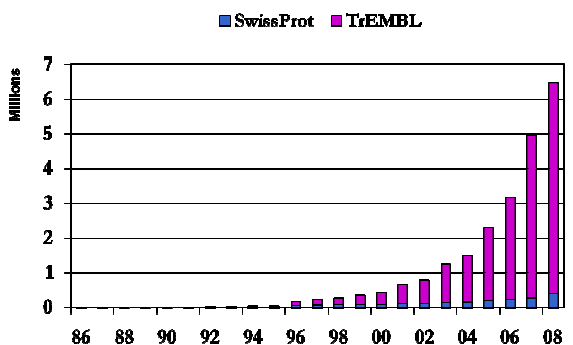
<http://www.expasy.uniprot.org/>



La génomique

Présentation

SwissProt/TrEMBL, nombre d'entrées



La génomique

Présentation

Prédiction de gènes

§ 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

La génomique

Prédiction de gènes

§ 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

La génomique

Prédiction de gènes

§ 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

La génomique

Prédiction de gènes

§ 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

La génomique

Prédiction de gènes

§ 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)

La génomique

Familles de protéines

§ 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

La génomique

Familles de protéines

§ 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

La génomique

Familles de protéines

§ 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)

La génomique

Familles de protéines

§ 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 en une fois)
 - 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

La génomique

Familles de protéines

§ Un exemple d'alignement multiple : l'insuline

	10	20	30	40	50	60
Human	MALWMRLLP	LALLALW	GGDDPAAAFV	NQHL	CGSHLVEAL	YLVCGERGFFYTPKTRREAED
Gorilla	MALWMRLLP	LALLALW	GGDDPAAAFV	NQHL	CGSHLVEAL	YLVCGERGFFYTPKTRREAED
Chimpanzee	MALWMRLLP	LALLALW	GGDDPAAAFV	NQHL	CGSHLVEAL	YLVCGERGFFYTPKTRREAED
Pig	MALWTRLLP	LALLALW	AFAPAFV	NQHL	CGSHLVEAL	YLVCGERGFFYTPKARRAEN
Chicken	MALWIRSLP	LALLALV	FGSGT	SYAANQHL	CGSHLVEAL	YLVCGERGFFYSPKARRDVEQ
	70	80	90	100	110	
Human	LQVGV	VELGGGPGAGSLQPLALEGSLQ	KRGIVEQCCTSI	CSLYQLENYCN		
Gorilla	LQVGV	VELGGGPGAGSLQPLALEGSLQ	KRGIVEQCCTSI	CSLYQLENYCN		
Chimpanzee	LQVGV	VELGGGPGAGSLQPLALEGSLQ	KRGIVEQCCTSI	CSLYQLENYCN		
Pig	PQAGAVE	LGGGLGG--LQALALEGPPQ	KRGIVEQCCTSI	CSLYQLENYCN		
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 %

La génomique

Familles de protéines