UE11 – Génomique, génétique formelle et épidémiologie - n° 3 27/02/2019 Pr. Frédéric DARDEL president@parisdescartes.fr	RT: Eloise Noll, Iris Nolland RL: Timothée Moyret

Génomique, Bio-informatique, Génétique et Evolution

Plan :

I. Biologie et évolution

Commençons par une citation de T. Dobzhansky (1900-1975) :

Rien en biologie ne fait sens si ce n'est à la lumière d l'Evolution

Lorsque l'on travaille en génétique ou en génomique, ne pas se placer dans une perspective évolutive, c'est rater une partie de la problématique dont traite ce cours.

A- Biodiversité et adaptabilité du vivant

Il existe de nombreux exemples d'animaux qui ont des particularités tout à fait exceptionnelles en termes d'évolution. En voici quelques-uns…

- **Alvinella Pompeiana** : un cousin du vers de terre qui vit au fond des océans dans les « fumeurs noires » (sources hydrothermales d'où sort de l'eau chauffée à 110°C par des sources volcaniques) et qui présente une bouche (T=110°C) et un rectum (T=4°C)

=> Face à ces conditions de vie extrêmes, son génome s'est adapté

- **Hypsibius dujardini** : un petit animal dont la taille est comprise entre 100microns-1mm, qui vit dans des gouttelettes d'eau, sur de la mousse…etc. (= microenvironnement aquatique), et qui est capable de résister à des doses astronomiques de rayonnements ionisants, à la vie intersidérale, aux acides forts….

=> Toutes ces agressions extrêmes ont fragmenté son ADN en petits morceaux et ont contribué à son adaptation aux conditions thermiques et physico-chimiques de son environnement.

- **Euprymna scolopes** : un céphalopode qui vit en symbiose avec les bactéries Vibrio fischeri (qui est un cousin de Vibrio cholerae et qui est non pathogène), capable de photoluminescence (utilisée pour éloigner ses prédateurs) : le céphalopode lui fournit de la nourriture en échange de la lumière produite par le les bactéries. C'est un exemple de système de coévolution très sophistiqué.

B- La classification des espèces vivantes

L'Histoire a montré que la compréhension de l'évolution a été initiée par plusieurs scientifiques.

- **Carl von Linné (1735)** a proposé une **classification binaire des espèces** (nom de genre + nom d'espèce). Il a ensuite regroupé les espèces animales et végétales par ressemblance, par caractéristiques (présence de colonne vertébrale, de mâchoires…etc pour les animaux, de fleurs pour les plantes…etc)
- **Charles Darwin (1862)** ajoute la **dimension temporelle** à la classification de Carl von Linné : cette organisation des espèces vivantes est le résultat d'un processus temporel.

C- L'Homme est un animal

Ernst Haeckel un physiologiste allemand et un grand vulgarisateur de la science en Allemagne, propose un **arbre de l'évolution des espèces animales (1866)** : plus on monte dans l'arbre, plus on trouve des espèces complexes. Mais de manière assez surprenante, l'homme est absent de cette présentation des espèces animales.

En réponse à cette observation, il publie une 2e œuvre, **The descent of man (1874)**, dans laquelle il propose un nouvel arbre appelé « **Stammbaum des Menschen** » ***avec l'Homme au sommet.***

En janvier 1925, le congrès de l'Etat du Tennessee publie une loi, **le Butler Act** qui s'énonce ainsi: *Il est illégal pour tout professeur dans n'importe lequel des universités et établissements d'enseignement financés en totalité ou en partie par des fonds de l'Etat, d'enseigner toute théorie qui contredise l'histoire de la Création divine de l'Homme telle est enseignée dans la Bible et d'enseigner à la place que l'Homme descend d'un ordre inférieure des animaux*.

John T. Scopes, professeur de sciences naturelles et président de l'association ACLU (Association de défense des libertés civiles) proteste contre cette loi qu'il perçoit comme un retour en arrière sur la **séparation de l'Eglise et de l'Etat**. L'action de contestation de l'ACLU devant la cour suprême fédérale échoue en 1926 et le Butler Act reste en vigueur dans le Tennessee jusqu'en 1968. Mais cette vision est encore dans de nombreuses mentalités, comme le montre la persistance du créationnisme. En 2005, une école publique de la ville de Dover, en Pennsylvanie, a tenté de rendre obligatoire l'enseignement de l'Intelligent design.

II. L'ère de la génomique

Depuis un peu moins de 20 ans , nous sommes entrés dans l'ère de la génomique humaine. Le **26 juin 2000**, Nature publie un article annonçant le **premier séquençage du génome humain** par 2 scientifiques : **Francis Collins** (un des dirigeants du National Institute of Health (NIH) et qui pilotait le projet de séquençage du génome humain) et **Craig Venter** (issu du milieu académique et qui sera à la tête d'une entreprise privée par la suite). Appellation de **programme « Apollo » de la biologie moléculaire** par le président américain Bush (père) car gigantesque sur le plan scientifique et aussi financier (grands investissements).

Plusieurs milliers de génomes sont aujourd'hui connus ou en cours de séquençage : Chlamydia pneumoniae (infection chalmydia), Yersinia pestis (peste), Salmonella enterica (salmonellose), H. pylori (ulcère de l'estomac), Mycobacterium tuberculosis (tuberculose), Neisseria meningitidis (méningite)…etc Soit une grande proportion d'**organismes pathogènes**. L'intérêt est de trouver des cibles, des vulnérabilités dans le mécanisme de fonctionnement de ces organismes pour apprendre à les connaître.
On a aussi séquencé la souris et le rat parce que ce sont des **animaux modèles** et des sujets de génétique « inverse » (c'est-à-dire des KO de gènes pour observer le phénotype).

La connaissance d'informations génomiques exhaustives aura et a déjà une influence profonde sur la recherche biologique et thérapeutique. Cette révolution nécessite la mise en œuvre des outils de la Bio-informatique.

A- Des performances exceptionnelles

En **1965**, Gordon Moore (Fondateur d'Intel) a énoncé la **loi de Moore** : *La puissance des microprocesseurs double tous les 2 ans*. Cette loi représente un espoir pour la communauté scientifique qui voit les microprocesseurs comme un outil suffisamment puissant pour entreprendre le séquençage génomique. La réalité est que les biologistes vont plus vite que les ordinateurs, et ce de manière également exponentielle : **le contenu des bases de données double tous les 15-18 mois** (aujourd'hui on dépasse les 4,5Bpb).

Problèmes :

- **Le volume des données et la complexité** des problèmes à traiter requièrent des algorithmes performants et la mise en œuvre de concepts sophistiqués :
  « Big Data »
- **La connaissance détaillée des questions biologiques** posées est essentielle pour « savoir ce que l'on fait » et permettre une **analyse critique des résultats**

= Une véritable discipline **transversale**

B- Petite histoire du séquençage génomique

- Premiers exploits : les génomes viraux
  - Bactériophage ϕX174 — 5600 b — 1978
  - Bactériophage λ — 48502 pb — 1982
- L'ère génomique
  - *Haemophilus influenzae* — 1 830 137 pb — 1995 — 1ère bactérie
  - *Methanococcus jannaschi* — 1 739 933 pb — 1996 — 1ère archaebactérie
  - *Synechocystis sp.* — 3 573 470 pb — 1996 — 1er organisme photosynthétique
  - Saccharomyces cerevisiae — 12 068 000 pb — 1996 — 1er eucaryote
  - Nématode — 100 Mpb — 1998 — 1er organisme pluricellulaire
  - Drosophile — 110 Mpb — 2000 — 1er arthropode
  - Arabette — 130 Mpb — 2000 — 1ère plante supérieure
  - Homme — 3 000 Mpb — 2001

Les premiers exploits correspondent aux **génomes viraux** de bactériophage ϕX174 (1978) et du bactériophage λ (1982). Les exploits se sont ensuite rapidement succédé à partir des années 1990.

L'invention du séquençage

Le principe du séquençage classique (et du NGS) est de fragmenter l'ADN en fragments.

- **Méthode de Maxam et Gilbert :** dégradation chimique sélective (surtout avant)
- **Méthode de Frederick Sanger :** synthèse enzymatique sélective (encore utilisée à l'heure actuelle même si assez remplacée par les méthodes NGS)

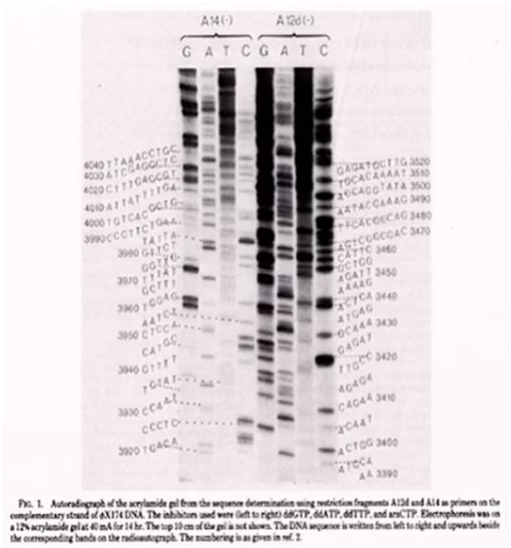
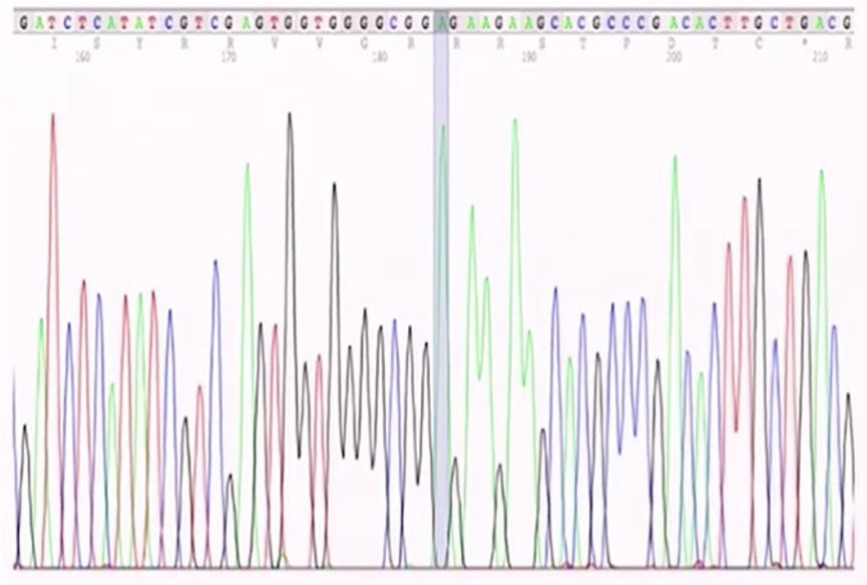


FIG. 1. Autoradiograph of the acrylamide gel from the sequence determination using restriction fragments A12d and A14 as primers on the complementary strand of ϕX174 DNA. The inhibitors used were (left to right) ddGTP, ddATP, ddTTP, and araCTP. Electrophoresis was on a 12% acrylamide gel at 40 mA for 14 hr. The top 10 cm of the gel is not shown. The DNA sequence is written from left to right and upwards beside the corresponding bands on the radioautograph. The numbering is as given in ref. 2.

Les fragments les plus courts sont en bas, la séquence se lit de bas en haut. A l'époque pour séquencer 1kb il fallait le travail d'une personne pendant 1 an.

1990 marque le **début de l'automatisation** avec le premier séquenceur automatique ABI 373A basé sur une migration sur gel avec un système de lecteur optique avec des lasers. (Débit maximal théorique = 20kb/J)

En **2000**, on a recouru à un système utilisant non pas un gel mais une électrophorèse capillaire (= tuyaux très fins dans lesquels on sépare les fragments d'ADN) avec un système de lecteur optique selon un code couleur pour chaque base. On peut ainsi séparer chaque fragment de taille croissante. L'application d'une tension électrique fait avancer les fragments dans le capillaire et leur passage devant un laser est détecté et permet d'exciter une photodiode. (Débit max théorique : 300kb/jour)



Le signal intégré ressemble à une succession de pics de différentes couleurs pour chaque base, on peut également déterminer la position de la base considérée en fonction de la taille totale du fragment. **On ne peut lire que 400 à 1000 bases par expérience, ce qui signifie qu'il faut beaucoup de fragments pour séquencer le génome**

Contraintes du séquençage de Sanger

- On ne peut séquencer que **400 (début) à 1000 (fin) bases à la fois**
- Il faut une **amorce complémentaire** de la séquence traitée pour démarrer la polymérisation.
    - Il faut découper le génome en **fragments** de **1000 nucléotides**
    - Il faut trouver une astuce pour s'affranchir du problème de **l'amorce**

Le problème majeur est ensuite de **reconstituer la séquence complète à partir de ces fragments. C'est à cette étape que la bio-informatique va être nécessaire pour assembler les fragments, les analyser (où sont les gènes, les signaux d'épissage, les signaux d'expression…etc ?) et les annoter (quelle est la fonction du gène identifié ?). La bio-informatique est incontournable :** l'analyse des séquences et dépendante de la puissance de calcul des ordinateurs. La génomique et l'informatique sont 2 technologies qui ont émergées en même temps (début des années 1980-1990) : le 1er PC IBM apparaît en 1982 (il existait déjà des ordinateurs avant 1982 mais beaucoup plus volumineux dans l'espace).

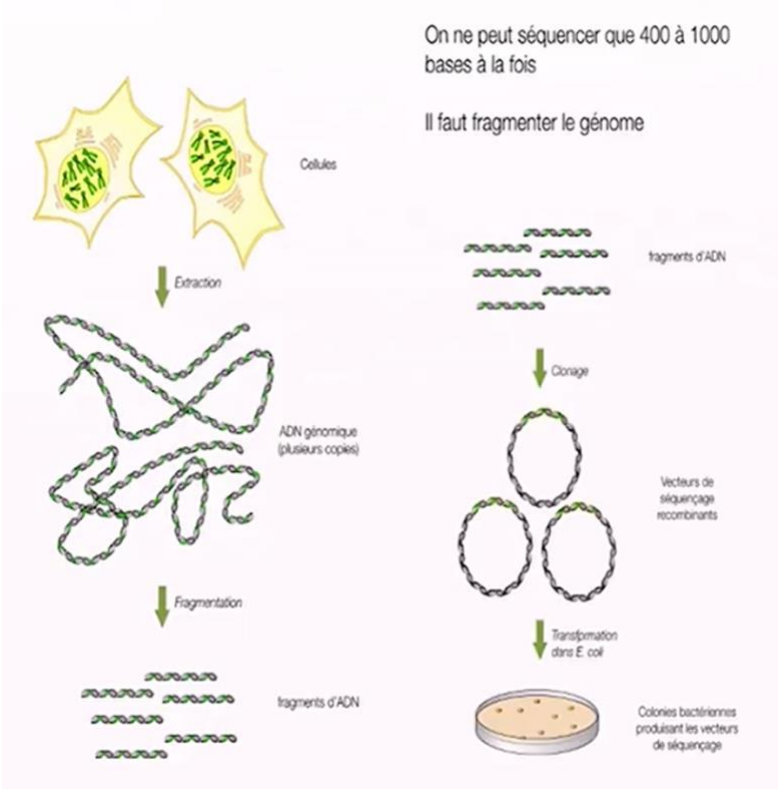
Taille des génomes

Organisme	Taille du génome (pb)
Virus du SIDA	9 750
Mycoplasma genitalium	580 000
Helicobacter pylori (ulcère stomacal)	1 667 867
Escherichia coli	4 639 221
Levure de bière	12 067 280
Plasmodium falciparum (paludisme)	25 000 000
Trypanosome	35 000 000
Nématode	110 000 000
Drosophile	150 000 000
Tétraodon (poisson-zèbre)	350 000 000
Tomate	655 000 000
Soja	1 115 000 000
Poulet	1 200 000 000
Boa constrictor	2 100 000 000
Homme	3 400 000 000

De manière assez évidente, il est nécessaire de connaître la taille du génome de l'espèce à séquencer avant de le séquencer.

Rq : Chez l'Homme le génome fait 3 000Mb ou **3 400Mb** (la vraie longueur du génome) selon que l'on considère l'euchromatine seule ou **l'euchromatine et l'hétérochromatine** (télomères, centromères…etc, qui ne sont jamais séquencés). Le record actuel du plus grand génome est tenu par **Paris Japonica** (« La Parisette », plante à fleurs blanches) qui a un génome de **150 000 Mb.**

Contraintes du séquençage de Sanger (suite)



1) Culture de cellules (10^6-10^9), chacune ayant le même nombre de copies d'ADN dans chaque noyau. Elles subissent ensuite un traitement pour extraire l'ADN génomique.

2) Fragmentation de l'ADN (1kb), par différents agents ou contraintes hydrodynamiques

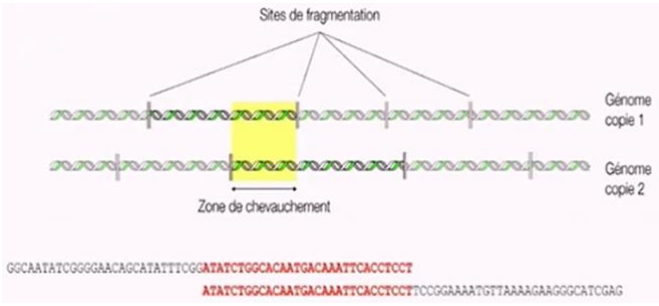
3) Isolement des fragments d'ADN et insertion dans des vecteurs de séquençage

4) Clonage des vecteurs dans des bactéries

Couverture du génome par la banque

On part d'une culture de cellules **identiques** (10^6-10^9).

- Plusieurs copies de chaque chromosome (une par cellule initiale)
- La fragmentation se produit **aléatoirement** donc à des endroits différentes sur chaque copie
- Une région donnée du génome est présente à de **multiples exemplaires** dans la banque, mais dans des fragments coupés différemment
- On séquence tous ces fragments
- Ces segments de séquence issus de fragments provenant de cellules différentes sont **chevauchants**



Toute la reconstruction du génome est basée sur la recherche de segments chevauchants.
Avec une fragmentation aléatoire, il faut séquencer un grand nombre de fragments pour couvrir la totalité du génome. Même dans ces conditions, il est pratiquement inévitable qu'il reste des trous. Cette observation a été étudiée statistiquement par **l'équation de Lander**, $p = e^{(-kn/L)}$, qui obéit à une loi de Poisson. Cette équation fait apparaître un terme kn/L correspondant au taux de couverture (= probabilité qu'une base donnée ne soit pas couverte par la banque).

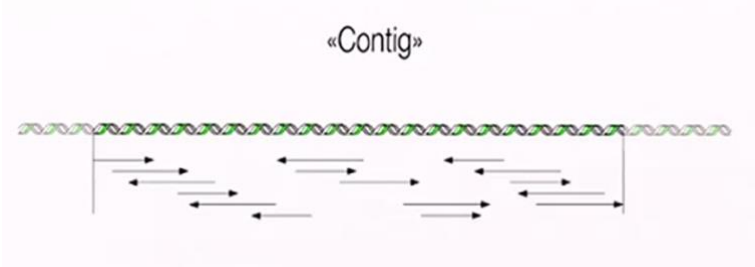
C- Stratégies de séquençage

- **Petits génomes (Limite supérieure < 10^7)**
  - Fragmentation aléatoire globale
  - Reconstruction directe

= Méthode du **« Whole Genome Shotgun » (WGS)** : tout couper, tout aligner, tout séquencer, tout remettre dans l'ordre

- **Grands génomes (Limite inférieure> 10^7)** **?**

Il faut procéder à un assemblage par chevauchements (à l'image des logiciels de reconstruction d'images qui regardent les chevauchements entre les photos). Grâce à ces chevauchements, on est capable de générer des séquences, des contigs, qui sont un ensemble de séquences qui se chevauchent mutuellement pour couvrir toute une région.



Rq : Il y a des flèches dans les 2 sens car l'ADN a 2 brins. Il faut donc chercher les chevauchements dans les 2 sens, en prenant en compte le brin complémentaire (faisable sur ordinateur).
La méthode de reconstruction de la séquence repose sur un **ajout itératif des contigs**, un **assemblage** puis, si les 2 segments sont effectivement contigus, une **fusion des contigs** correspondant.
Rq : Chaque contig est en réalité lu plusieurs fois, il y a une **redondance de l'information**.
On va utiliser cette redondance pour résoudre des **ambigüités** à une position.
ex : succession d'une même base

Obstacles à la reconstruction

La persistance de trous est inévitable et ce quel que soit la qualité de notre travail et quel que soit le taux de couverture. Ces trous peuvent être liés à l'existence de :

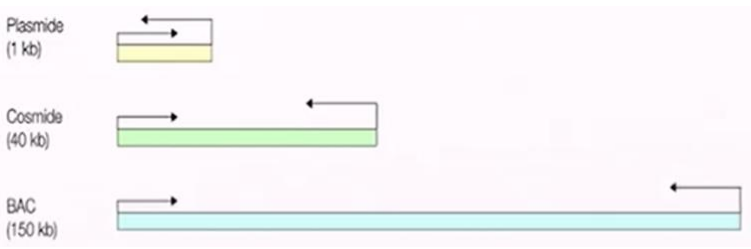
- **Séquences toxiques**
  - Pour l'hôte
  - Pour le vecteur
- **Séquences répétées** (ex : séquences Alu), qui vont créer des ambigüités d'assemblage

Comblement des trous et vérifications

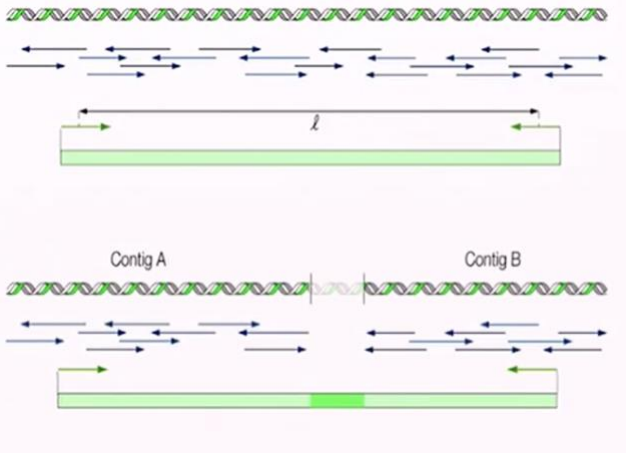
Il existe de nombreuses méthodes pour contourner cet obstacle comme :

- Intégration des données de cartographie : à partir de séquences déjà connues ou par comparaison de la carte obtenue avec des cartes génétiques
- PCR « par-dessus les trous »
- Intégration des données des grands clones : la plus généraliste et la plus utilisée
  ex : séquençage NGS pour reconstituer la continuité des chromosomes et des séquences répétées
- Re clonage sélectif

Vérification au moyen des « grands clones »



On clone des fragments plus grands dans des vecteurs adaptés.
Ex : **BAC ou Bacterial Artifical Chromosome** (150kb), **Cosmide** (40kb), **Plasmide** (1kb)
On séquence les 2 extrémités du grand fragment ou du grand clone (500-1k bases).



Le positionnement des extrémités du grand clone dans les contigs permet :

- **De vérifier la cohérence de l'assemblage** (longueur *L* compatible avec la taille de l'insert) (haut)
- **De positionner les contigs et de combler les trous** (bas)

III. Les grands génomes

A- Stratégie EST

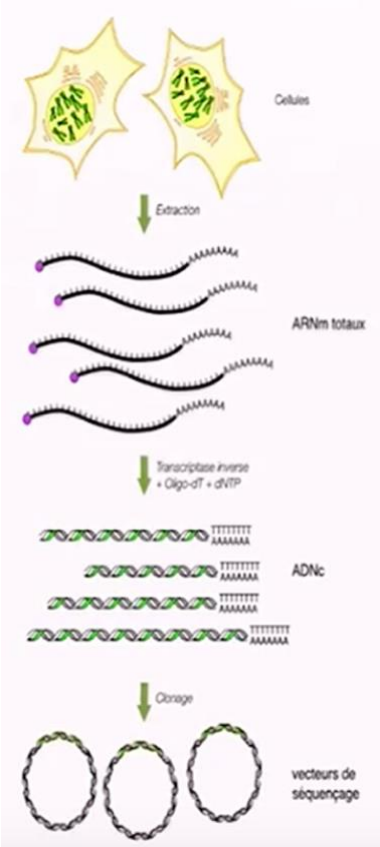
Au moment de séquencer le génome humain, la communauté scientifique a hésité à entreprendre des recherches colossales sur un génome très grand dont seulement 1,6% code pour des protéines (information d'intérêt).

Craig Venter a proposé de se restreindre au séquençage des 1,6% en proposant la méthode des EST (Expressed Sequence Tags ou Etiquettes de séquences exprimées) : le principe est de ne séquencer que les régions transcrites en ARNm. Cela nécessite la construction des banques d'ADNc, obtenues par transcription inverse des ARNm issus de différents tissus.
Les intérêts sont multiples :

- Identification des régions transcrites du génome puisque l'ADNc est obtenu à partir des ARN
- Détermination des profils d'expression tissulaire puisque les ARNm sont issus de différents tissus
- S'affranchir des problèmes d'assemblage puisque 1 ARNm = 1 gène

NB : Pourquoi différents tissus ? Pourquoi des profils d'expression tissulaire ?

Les cellules différenciées n'expriment pas les mêmes gènes, le profil d'expression sur une population différenciée de cellules identiques serait donc très biaisé par rapport au profil d'expression du génome. (ex : Hb très exprimée dans les globules rouges)
Pour avoir une représentation complète des différents profils d'expression, il faudrait séquencer les différents ARNm issus de toutes les cellules de l'organisme. Et encore, puisqu'il existe aussi une variation du profil d'expression selon le stade du développement considéré (ex : HbF uniquement au moment de la vie fœtale).



1) Extraction des ARNm totaux à partir des différents tissus
2) Synthèse des ADNc
3) Clonage dans des vecteurs de séquençage
4) **Séquençage systématique*** : centaines de milliers d'ADNc séquencés

*Séquençage systématique : constitue le gros avantage de la méthode EST

B- Le *Human Genome Project*

Séquençage exhaustif du génome humain

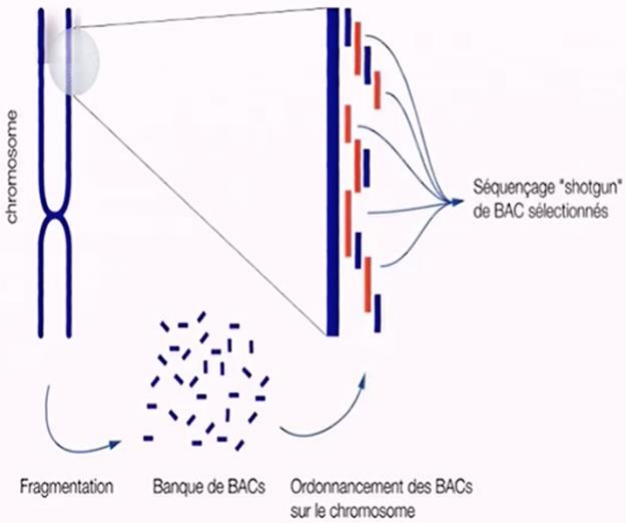
C'était un défi technologique énorme parce qu'il fallait plusieurs dizaines de millions de clones à séquencer, ce qui lui a valu l'appellation de « plus grand puzzle du monde ».
Le contenu est « diaboliquement » complexe à reconstituer compte tenu du taux très élevé de séquences répétées (>45% de notre génome) et du fait que certaines sont répétées plusieurs centaines de milliers de fois. (ex : Seq Alu)

Le *Human Genome Project* (1989-2004) était un projet vraiment fou pour plusieurs raisons.

- **Très très cher** : environ 3 milliards de dollars
- **Pharaonique** : en 1988-1989, la production mondiale est de 10 à 20 millions de nucléotides par an. A ce rythme, il faut plus d'un siècle…
- **Très difficile** : le génome est très redondant, près de la moitié du génome est composé de séquences répétées
- **L'utilité fait débat** : les gènes codant des protéines représentent moins de 2% du génome

Après avoir acheté la paix sociale, la communauté scientifique s'est donnée **15 ans** pour réaliser ce projet et fournir une **séquence de haute qualité (99,99% de précision).** Elle s'est accordée pour que les travaux soient faits dans le cadre d'un **consortium international public** et pour que les données soient **immédiatement déposées**. Elle a même fait un **pari sur une accélération des techniques de séquençage par un facteur 100.**

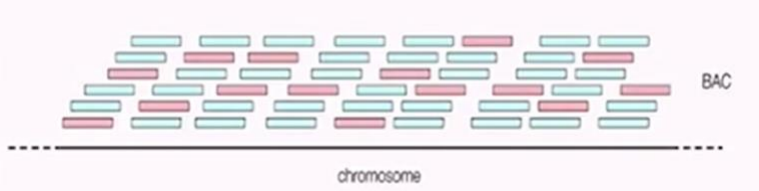
Pour acheter la paix sociale, la communauté scientifique s'est engagée à proposer une analyse simultanée d'organismes modèles (E. Coli, la levure, le nématode et la drosophile) avant d'entreprendre un travail de complexité plus importante : celui du séquençage du génome humain, en utilisant une **approche hiérarchique**.



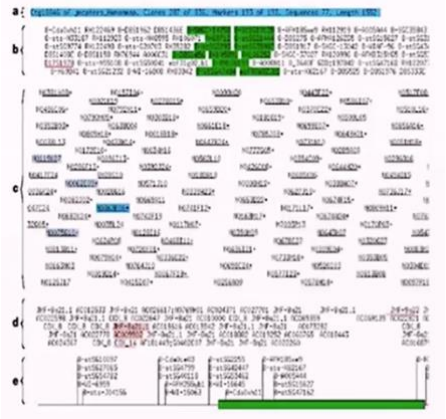
<u>Approche hiérarchique du séquençage : « Map first, Sequence later »</u>

Le principe est de découper le génome en « grands morceaux » puis de repositionner les morceaux sur le génome complet sans en faire le séquençage. Pour ça, on utilise des vecteurs adaptés à la propagation de très grands fragments d'ADN (BAC : 100-300kb, Cosmides : 30-50kb). Ces grands clones sont ensuite cartographiés (« map first ») puis séquencés (« sequence later »).

Cette approche permet de répartir le travail sur différents groupes scientifiques qui entreprennent parallèlement le séquençage WGS (shotgun) d'un BAC sélectionné et ainsi d'éviter une redondance dans le travail entre les équipes du monde entier.



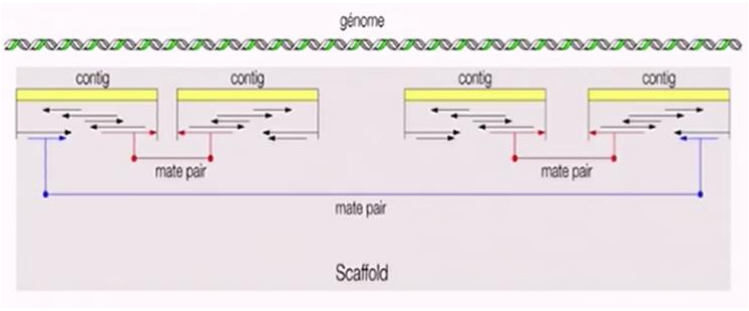
Pour faire la cartographie, on construit une banque de BAC très redondante (avec un taux de couverture > 10), on coupe ces BAC par des enzymes de restriction, on analyse les fragments par électrophorèse et on identifie des fragments communs à 2 BAC, des chevauchements, qui nous permettent de reconstituer la séquence de proche en proche.



Au total, le consortium international a cartographié 350 000 BAC mais seulement 30 000 ont été séquencés par l'étape de WGS. L'ébauche de séquence (draft sequence) de février 2001 possédait des contigs de 1000kb, ce qui équivaut à la persistance d'encore 3000 trous.

En 1998, **Celera (Craig Venter)** entre dans la course avec une **approche WGS** (ce qui signifie $>10^7$ fragments à assembler) et en lançant un système de robotisation extrême grâce à 300M$ d'investissements (**Perkin-Elmer**, une entreprise d'instrumentation scientifique tenue par Tony White ; **Applied biosystems** fabrique >100 séquenceurs). Celera a fait un appel à candidature pour trouver un super-ordinateur assez puissant pour ses recherches. Avec l'aide de **Compaq**, Celera a monté le super-ordinateur privé le plus puissant du monde. Celera s'installe à Rockville avec son plus gros calculateur privé du monde, ses 200 séquenceurs et son mégawatt de consommation électrique.

Comment résoudre le problème des séquences répétées ?



Les « mate pairs » de Gene Myers est une idée de séquençage qui utilise non pas les chevauchements mais l'information d'appariement des fragments (sachant que la taille des fragments est connue parce qu'imposée par les grands clones de séquençage type Plasmide, Cosmide et BAC). Cette méthode des « mate pairs » est utilisée à différentes échelles de fragments (petits, moyens et grands fragments) et aboutit à la construction de « scaffolds » (=échafaudages).

Le séquençage du génome humain est le seul projet de « big science » qui a mis en concurrence un consortium public et une entreprise privée (Celera genomics), qui vendait les informations aux laboratoires pharmaceutiques.

Consortium public	**Celera genomics (Craig Venter)**
Stratégie hiérarchique « divide and conquer » : Cartographie puis séquençage	« Shotgun » total Petits clones (2kb) Cosmides (50kb) BAC (150kb)
Séquençage distribué des BAC	Reconstruction globale Supercalculateur parallèle
Reconstruction hiérarchique	Intégration des données « publiques »

Aujourd'hui, l'ensemble du génome humain est accessible en ligne sur plusieurs portails

- Portail européen Ensembl : www.ensembl.org
- http://genome.usc.org

Avec le NGS, le séquençage de nouvelle génération (2005-2008), on est capable de séquencer 1 200 000 000 bases/jour/5000euros.

IV. La traduction du génome

Le **NGS** coupe des morceaux plus petits, on a donc la même problématique de recouvrement, mais le taux de couverture doit être plus élevé. Aujourd'hui, il est facile de séquencer le génome du chimpanzé grâce à celui de l'humain déjà reconstitué. Il est utilisé comme patron pour aligner les séquences de chimpanzé sur celui de l'humain. En revanche, pour une espèce exotique dont on ne connait pas l'organisation du génome, c'est plus compliqué, on utilise alors les techniques **de paires appariées**.

A. La difficulté d'analyse du génome et les différentes techniques

Il y a différents niveaux dans l'analyse génétique :

- L'analyse du génome : composition, variabilité, quelles modifications ?
- L'analyse du transcriptome : qu'est-ce qui est transcrit et dans quel tissu ?

Où sont les gènes ?
Le génome est peu lisible car il est codé par un alphabet à 4 lettres. Seul 2% du génome code pour des protéines, il est très morcelé. La taille des exons (150-200 nucléotides) varie peu, c'est la taille des introns qui varie énormément.
Ainsi lorsqu'on analyse le génome on est à la recherche d'informations codantes au sein de l'ensemble des informations. *Le prof donne l'exemple d'un film sur TF1 : pub-film-pub-film-pub pour décrire le génome.* Il faut réussir à décoder les gènes à partir de toutes les informations obtenues par le séquençage du génome → *en gros on est en quête des gènes, mais c'est difficile car ils sont cachés dans toute l'information du génome.*

Pour ce faire on utilise différentes techniques :

- On cherche des **signaux au niveau des promoteurs** et des liaisons intron-exon (signaux d'épissage) qui sont des motifs. On va chercher des enchainements de nucléotides qui correspondent à « des structures spéciales » (comme le traitement de texte).
- On cherche aussi des choses par **le contenu,** par le biais des statistiques sur la composition des nucléotides. Il y a des **contraintes** appliquées sur le code génétique, des irrégularités périodiques. *Avec une répartition de 3 nucléotides ; il y a sûrement une région codante.*
- On compare aussi les **bandes EST** du même organisme dans le langage génomique.
- On peut aussi **comparer avec d'autres espèces** voisines : on regarde si le singe a une séquence qui ressemble. Si le gène est conservé dans l'évolution alors il a surement une fonction donc c'est un gène codant. La fonction du gène retrouvé dans le singe, devrait avoir une fonction semblable à celle du gène que l'on cherche à analyser.

On analyse aussi la synténie qui sera détaillée plus tard. **L'annotation génomique** combine l'ensemble de ces techniques.

Pour identifier des signaux (exemple dans les jonctions intron-exon) on utilise des outils de geeks *comme les caractères jokers*
Les expressions régulières permettent de définir les expressions compliquées. *Ex : codons stop (TGA , TAA et TAG).* Ces outils informatiques permettent de faire des recherches de haut niveau. Ils font partie de langages de programmations comme *Python, perl ou TCL*. Bref, il faut savoir que ces outils intelligents existent et qu'ils permettent de détecter des enchainements de nucléotides complexes.

L'enchaînement du contenu par les statistiques :
Il y a des statistiques effectuées sur les nucléotides, qui permettent de classifier les espèces, de dire : cette séquence est une séquence de cette espèce.
Ex de Jurassic Park : est-ce que c'est vraiment de l'ADN de dinosaures ?
Donc si le dinosaure est un vertébré, dans son ADN on doit retrouver 60% de AT et 40% de CG, ce qui fait 20% de C et de G et 30% de A et de T. Non, ce n'est pas un dinosaure car le taux de CG est de 60% et non pas de 40%.
De même les séquence de méthylation sont toujours sous représentées chez les humains et il y a de forts biais imposés par le code génétique dans les exons.
Fun fact : le poulet est un dinosaure encore vivant aujourd'hui.

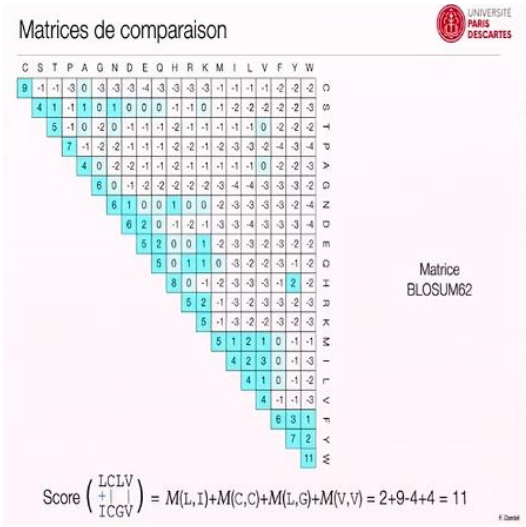
Les séquences biologiques sont-elles aléatoires ? Non il existe des biais dans la distribution des nucléotides (% AT/CG) liées aux **pressions de sélection exercées par l'évolution.**
À partir de statistiques caractéristiques d'un type de séquence, on connait tel biais qui correspond à telle catégorie. Donc on calcule la probabilité d'appartenance à une espèce.
Exemple : les séquences CPG sont sous représentées dans notre génome, on trouve peu de successions CG.
Il existe des **outils de prédictions** basés sur des méthodes statistiques déterminées en balayant une séquence. Ça permet de cibler dans un génome les séquences qui pourraient nous intéresser mais sans nous les donner précisément. Par exemple on trouve que par là il y a un début d'exon, mais pour savoir où est ce que l'exon commence précisément on cherche un **motif** (site exact)**,** alors que cela nous donne une **zone.**

La biologie se base sur la **comparaison** de séquences : quand on a deux molécules de fonctions apparentées, leurs séquences seront ressemblantes et inversement, si on a deux molécules avec une séquence semblable, alors elles auront surement une fonction apparentée. On est dans un système d'une **évolution à partir d'un ancêtre commun**.
Ex : notre récepteur de la vitamine D est ressemblant à celui du poisson zèbre.
On a 20 AA, donc si l'ensemble du génome était réparti au hasard, on aurait seulement 5% du génome en commun (100/20 = 5). Si on en a plus, c'est que les protéines ont la même fonction, ou une fonction ressemblante.

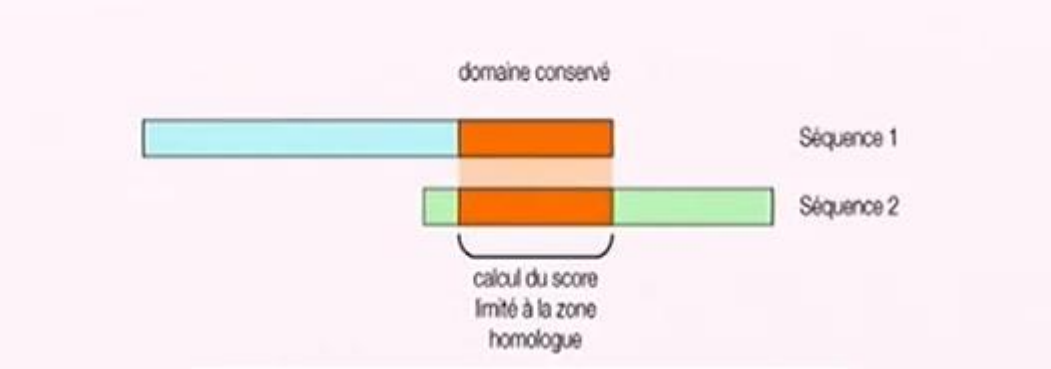
On code par un système de points la ressemblance :

- Deux AA opposés (tryptophane et isoleucine : points négatifs
- Deux AA neutres (alanine et tryptophane) : zéro
- Deux AA avec ressemblance (isoleucine et leucine): +2
- Deux AA identique : +4

= **score d'alignement**



Il existe aussi des algorithmes d'alignement dans l'ordinateur qui calculent entre deux séquences le score d'alignement et déterminent la séquence avec laquelle il y a le meilleur alignement dans les bases de données. Si on compare deux séquences majoritairement semblables, on essaye de les aligner sur toute la longueur. Ce sont des **comparaisons globales**. Mais dans d'autres cas, on va s'intéresser seulement à des briques de structures de protéines. On aura alors des nucléotides assez différents sur l'ensemble de la longueur, mais aussi quelques domaines identiques (forment les briques). C'est ce qu'on appelle un **alignement local**. Donc quand on aligne les séquences on va avoir un domaine conservé puis un non conservé.



Ex : domaine en bleu conservé, vert de chaque côté est non conservé.

BLAST (outil d'alignement de base de séquence, très rapide) est un programme informatique d'alignement qui recherche dans les bandes de séquences celles qui vont se rapprocher le plus de la séquence à analyser : *voici ma séquence, trouve-moi dans les bases de données internet tout ce qui ressemble à ma séquence*. Il donne le score d'alignement, mais aussi le % de similitude, la longueur de la séquence. Il donne aussi le **E-values** qui est une estimation que la probabilité que cet alignement soit dû au hasard (bruit de fond).
On obtient une multitude d'homologues (dans d'autre espèces ou dans la même espèce). Maintenant si 10 séquences se ressemblent, on cherche à toutes les aligner : c'est de l'**alignement multiple**. Plusieurs caractéristiques rentrent en jeu : des région 100% conservées et d'autres non. Sur le plan fonctionnel, cela se traduit par : **si on n'a pas de fonction, pas de pression de sélection, donc des mutations aléatoires.** Cependant si les régions sont 100% conservées, cela signifie que les mutations sur ces régions n'ont pas survécu par le fait de la sélection naturelle. Donc ce gène-là a une fonction car il y a eu un désavantage sélectif.

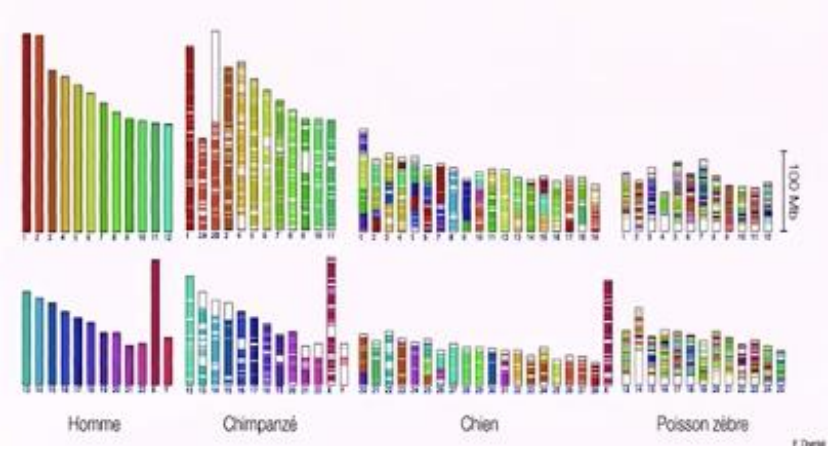
- ⇨ régions conservées = régions fonctionnelles : **prédiction fonctionnelle**.

On va par la suite faire un alignement des séquences en commençant par celles qui se ressemblent le plus. Cela permet de construire par la suite des arbres d'évolution de ressemblance. On peut donc introduire une dimension temporelle : probablement les protéines qui se ressemblent dérivent d'un ancêtre commun dans l'évolution puis différentes régions ont divergé et se sont spécialisées pour des hormones différentes. Plus on a de ressemblances, plus les deux espèces se sont séparées récemment et moins elles se ressemblent, plus leur séparation est lointaine dans l'évolution.
Il existe un lien entre la modification de la séquence/génétique et la modification du phénotype.
Moins il y de divergence, plus les espèces sont proches.

L'ordre des gènes dans les chromosomes : **la carte des chromosomes**

Pour ce faire on a colorié la carte des chromosomes avec une couleur pour une région.
Chez le poison zèbre, on observe une multitude de couleurs différentes sur les chromosomes. Ainsi entre le poisson zèbre et l'homme il y a eu un réel mélange des bouts chromosomes et l'ordre des gènes n'est pas le même → véritable réarrangement des chromosomes.
Alors que le chimpanzé est plus proche de nous car on observe, globalement la même répartition des couleurs. Mais le chimpanzé à 48 chromosomes et nous 46 chr. 2 explications possibles : soit le chimpanzé a divisé son chromosome, soit nous avons fusionné 2 chromosomes → 2ème proposition est la bonne car on a retrouvé un centrosome fantôme dans le chromosome 2 de l'homme.

La synténie est la propriété de conservation dans l'ordre des chromosomes.



B. Le contenu du génome :

La taille des génomes est très variable, le nombre de gène varie peu (un facteur 10), mais la densité diminue avec l'augmentation de la taille du génome. Il n'y a pas de corrélation entre nombre de gènes et taille du génome et la taille du génome n'a pas de lien avec la complexité de l'espèce.

Il y a une réelle difficulté d'analyse du génome :

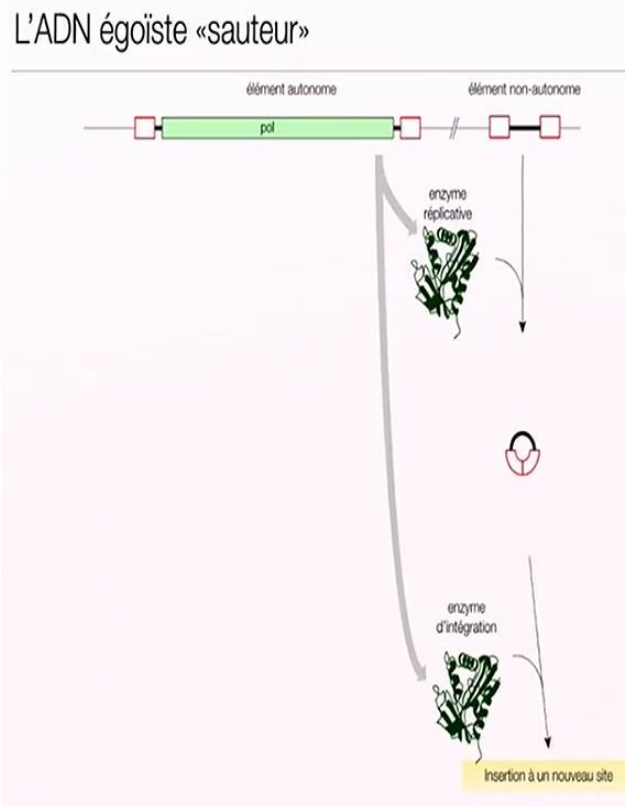
- **très dilué** : 2%
- **les pseudogènes** = gènes devenus inactifs suite à une mutation. Aussi nombreux que les gènes !
- **les séquences répétées de gènes** (45% de notre génome) : ils viennent d'un processus actif de réplication de segments qui se multiplient en permanence dans le génome. Elles envahissent le génome.

ADN égoïste

Elément du génome encadré par des séquences spécifiques souvent répétées elles même, qui code pour une/des protéines impliquées dans la réplication d'éléments autonomes.

L'enzyme réplicative permet de fabriquer un intermédiaire réplicatif linéaire ou circulaire qui peut être soit en ADN soit en ARN suivant la nature des éléments répliqués, puis une seconde enzyme d'intégration, va se mettre à un autre endroit dans le génome. On obtient donc deux, puis 3, … copies.

Il existe aussi des éléments <u>non autonomes</u> qui partagent des éléments de contrôle répétés de l'autonome mais sans rien au milieu, sans gêne, qui va pouvoir se répliquer par le biais de l'autonome et de l'enzyme d'intégration: l'enzyme réplicative de l'élément autonome diffuse dans le noyau en envoyant des infos à l'élément non autonome en pensant que c'est un élément autonome, qui vont être traduit grâce à l'élément autonome et par la suite va se mettre à un autre endroit dans le génome. Donc il peut se répliquer mais seulement avec l'aide de l'élément autonome qui fournit les protéines nécessaires à la réplication.



On a également en notre sein 450 000 copies de **rétrovirus fossiles endommagées** : traces d'infections qui ont eu lieu dans la lignée humaine au cours de l'Histoire. Souvent délétées par des éléments de recombinaisons et d'évolution. Mais certaines sont encore partiellement fonctionnelles telles que la capside, la transcriptase inverse, l'intégrase. Ce sont des virus qui ont perdu leur protéine d'enveloppe, le virus ne peut donc plus sortir. Il est prisonnier mais peut se répliquer. Dans notre génome il existe encore HERVK qui est encore actif et qui se réplique. Il faut que ça se produise dans une cellule germinale pour être transmis à la descendance. Il y a aussi des transposons. Ce sont des séquences qui passent par l'intermédiaire ADN ou ARN, des formes autonomes ou non. Il y a des formes directes ou inversées (par les transposases). Il y a des séquences longues répétées et des séquences courtes répétées = short interse nuclear.
Il existe également des éléments nucléaires dispersées courts et des éléments nucléaires dispersés longs. Ce sont des séquences transcrites avec un promoteur reconnu par les ARN polymérases cellulaires. Puis elles sont devenues des ARN intégrés à un autre endroit par transcription inverse. *ex : la séquence ALU qui est encore active car elle a encore un élément autonome pour lui permette de se répliquer dans plusieurs endroits.*
Le maïs est l'espèce vivante chez qui ce mécanisme de transposition a été découvert. Dans le maïs les éléments les plus actifs sont les transposons qui ont fait quadrupler la taille du génome au cours des 18 millions dernières années.
Ainsi tous ces éléments participent à **la variabilité génomique** : on n'est pas tous pareil.

C. L'histoire du séquençage du génome

De qui a-t-on séquencé le génome ?
Dans le cadre du consortium public, celui qui a permis de séquencer le premier génome humain, on a fait des prélèvements anonymes chez des donneurs mâles (car on devait séquencer aussi le chromosome Y). Ils ont réalisé des bacs avec pour chaque bac un seul haplotype, un seul segment d'ADN d'un seul donneur. Quand on hydrolyse les cellules d'un organisme diploïde, on a deux génomes mélangés : celui qui vient de la mère et celui qui vient du père, cela peut compliquer l'analyse. 70% des bacs, viennent d'un seul donneur qui avait des très bonnes capacités de recouvrement.

Chez Celera, (entreprise privée), ils utilisent des prélèvements de 9 donneurs anonymes tous très différents. Ils font des shotguns séparés sur chacun des donneurs car les séquences sont différentes, pour ne pas louper de chevauchements, et reconstituent les séquences de donneurs chacune séparément. Dans chaque shotgun, il y a déjà les deux haplotypes du père et de la mère.

Fun fact parce qu'il est 20h est qu'on est encore à la fac : Craig Venter s'est fait séquencer lui-même, et on a trouvé un variant d'un gène de prédisposition aux maladies cardio-vasculaires. Il prend depuis des statines et comme le séquençage de son génome a couté 300 millions de dollars, c'était le diagnostic le plus cher du monde.
´

D. Les variabilités génomiques :

- SNIPs : variation de 1 nucléotides
- Insertions et délétions : ajout ou retrait d'un ou plusieurs nucléotides entre un individu A et un B
- Répétitions de microsatellites ex : dans un génome il y a 3 copie de cet ensemble de nucléotides, alors que dans l'autre il y en a 6
- Copy number variation : duplications plus importantes
- Variation de grande partie : Alu

La variabilité génétique chez l'homme est faible : taux est de 3 4000 000, donc de 0,1% du génome. Il y a 99,9% de conservation dans les régions d'euchromatine, donc une variation tous les 1000 nucléotides. C'est moins que pour les chimpanzés qui ont 0,2% de variations.
La variabilité génétique est différente selon les populations humaines. Elle est particulièrement élevée en Afrique. Donc si on prend deux Africains, ils ont plus de différences entre eux que deux individus ailleurs dans le globe *(ex : un japonais et un américain).* Cela s'explique par le fait qu'une partie de la population africaine soit partie conquérir le reste du monde, et l'autre restée en Afrique.

On peut dater les migrations à partir de l'Afrique grâce :

- Aux mitochondries pour remonter aux infos de la mère,
- Au chromosome Y pour remonter à celui du père.

Le sud de l'Amérique du sud est la dernière zone du globe à avoir été conquise par les hommes.

Il n'y a pas d'allèle spécifique d'une population donnée, c'est les **fréquences alléliques** qui varient. Donc il n'existe pas de gène spécial japonais, mais il va y avoir plus de fréquence du gène de non résistance à l'alcool. Ce gène est aussi présent chez les Européens mais avec une plus faible fréquence. Il y a une variation appelée golden chez le poisson zèbre. Ce variant-là est déficient dans la pigmentation. On a remarqué l'existence de ce gène chez l'homme et qu'une mutation de ce gène qui impacte la mélanine chez les populations d'Europe du nord qui sont les dernières à avoir colonisé leur territoire. Cette mutation peut être datée dans les populations d'Europe du nord : entre 3000 et 8000 ans avant aujourd'hui ; c'est une mutation récente. En 4000 avant JC, certaines populations ont changé de régime alimentaire en le rendant plus abondant en Vitamine D. Comme ils étaient dans un pays froid avec peu de soleil (déterminant dans la synthèse de vitamine D), les individus à la peau claire étaient avantagés par rapport à ceux à la peau foncée en raison d'une fabrication accrue de la vitamine D. Cette explication est plausible mais pas certaine.

New fact : notre cher président a pu faire le malin dans le musée de la préhistoire en disant que les Hommes de la préhistoire avaient la peau foncée. Ce qui n'a pas plu à sa mère.

On est identique à 99,9% et les chimpanzés à 99,8% mais on a l'impression d'avoir un visage plus différent que les chimpanzés. Pour ça il y a deux explications :
1- dans notre cerveau on a appris très jeune à reconnaitre les visages, qui est un élément important de survie. Et c'est identique chez les chimpanzés, ils se reconnaissent entre eux et ils ont l'impression que les hommes sont tous pareils.
2- Certains critères sont sélectionnés dans l'évolution plus vite que d'autre dans les caractères qui nous différencient. Jusqu'à il y a un siècle les gens bougeaient peu. On se reproduit sur des critères extérieurs : la silhouette, la force, la taille et non pas sur les caractéristiques du glycéraldéhyde. Ces critères ont été soumis à une **pression de sélection non visible**. Les critères de la beauté féminine varient beaucoup d'une population à une autre et au cours du temps.

V. L'utilisation de la génomique

A. Les progrès pharmaceutiques :

Dans le premier Vidal : on soignait les maladies des bronches par des cigarettes américaines, ou bien on soignait les hémorroïdes avec une pommade à base de cocaïne car à l'époque on n'avait pas d'autres anesthésiants locaux. En un siècle, la pharmacie a fait des progrès considérables. Avant on avait que les opiacées et la digitaline pour les maladies du cœur.
Ça a progressé de la manière suivante : on caractérisait les dysfonctionnements physiologiques et les mécanismes moléculaires associés. On identifiait par la suite les cibles moléculaires et des molécules actives.

Depuis la génomique, tout a changé : on a 1/3 de gènes inconnus, 1/3 de gènes connus et 1/3 de gènes qui ressemblent à des gènes connus. C'est à cette dernière catégorie que l'on s'intéresse : on sait que ce sont des enzymes d'une famille importante et connue mais on ne connait ni leurs cibles ni leurs fonctions. Les laboratoires pharmaceutiques les veulent pour chercher une cible pour de nouveaux médicaments. On identifie des gènes candidats en cherchant informatiquement sur le génome, puis on essaye de caractériser la fonction des gènes cibles et enfin on cherche des molécules. La compréhension de la physiologie est devenue la dernière étape. Avant on partait de la physiologie vers le gène cible et maintenant on part du gène séquencé et on remonte à la physiologie. C'est ce qu'on appelle **l'approche pharmacogénomique**. Si on cherche un antibactérien, on cherche une enzyme absente chez l'homme qui tue l'ensemble des bactéries néfastes dans les **banques génomiques.**

B. Génomique et évolution :

Aujourd'hui toute la classification des espèces est basée sur l'analyse comparative des séquences qui est l'outil le plus précis. Les éléments différenciant : les poissons ont acquis une nouvelle capacité de fixer le calcium dans les os. Donc on est plus proche dans l'évolution de la sole et de la carpe, que la carpe et la sole ne sont proches de requin. Certaine fois, l'évolution n'est pas intuitive.
L'Homme n'est pas une forme améliorée du singe, car on a évolué pendant la même période. Il y a des bactéries qui savent faire des choses que l'homme ne sait pas faire. Chaque espèce s'est adaptée à son environnement. **L'évolution n'est pas une progression, mais du bidouillage permanent**. L'évolution, c'est de la récupération et du bricolage.

Ce sont trois protéines de structure voisines mais de fonctions différentes :
1- réparations de l'ADN
2- protéine qui intervient dans la germination.
3- régulation du cycle circadien.
donc elles sont impliquées dans des mécanismes différents mais elles ont toutes la capacité de détecter la lumière bleue. Leurs structures sont identiques, elles dérivent du même ancêtre commun, on a réutilisé le récepteur à la lumière bleue pour de multiples fonctions.

C. Évolution et médecine : l'histoire de la peste

541-544 peste de Justinien : en train de conquérir l'Empire romain, son armée fut décimée par la peste qui a ensuite décimée 40% de la population d'Europe. On a retrouvé les bactéries dans des squelettes. La peste est une bactérie qui se transfert à une puce, cette dernière pique soit des rats qui sont le réservoir, soit des hommes.

Pourquoi cette maladie est aussi virulente ?

Elle a acquis un gène qui lui permet de survivre dans l'intestin de la puce, le gène YMT porté par un plasmide extra chromosomique. Il y a aussi une mutation dans le gène PLA qui lui permet d'envahir les tissus et de remonter aux ganglions, et de donnait la forme lymphatique. Puis une autre mutation, dont la perte du flagelle qui la rend invisible du système immunitaire. Ces différentes mutations sont apparues à différents temps de l'histoire et ont permis d'acquérir au fur et à mesure de nouvelles fonctions. L'échappement au système immunitaire est apparu il y a 4000 ans et coïncide avec l'épidémie de Justinien qui est surement la première épidémie de peste.

Pourquoi certaines sont très virulentes et d'autres moins ? Est-ce que les maladies infectieuses nouvelles sont plus virulentes que les anciennes ? Quelle est la stratégie du pathogène ?

Pour optimiser sa dissémination il doit optimiser deux choses en même temps :

- efficacité de prolifération dans l'hôte en évitant le système immunitaire.
- entre les hôtes

Il faut s'intéresser au mode de transmission :

- transmission directe
- transmission par un vecteur.

Si le vecteur n'est pas un insecte, ça peut être l'eau qui est un vecteur très efficace. Les épidémies hydriques ont plus d'impact sur la mortalité car dans une transmission par voie directe, si on tue son hôte direct, ce dernier n'a pas le temps de le transmettre donc c'est un mauvais mécanisme. Il y a une corrélation entre le vecteur et la virulence. **Donc la transmission directe ou par vecteur détermine la virulence de l'infection.**
Les bactéries moins pathogènes ont un avantage par transmission directe et les bactéries très pathogènes ont un avantage dans la transmission par un vecteur. La meilleure stratégie c'est quand le vecteur reste résistant et n'est pas affecté par le virus.
Pourquoi les infections nosocomiales sont plus graves que celles en ville : c'est parce que le médecin est le vecteur des infections car il est vacciné donc devient un bon vecteur.

D. <u>D'autres applications de la génomique</u>

Encode : le prof donne un avis personnel sur « on peut donner une fonction à 80% du génome »
Absurde, car on peut mesurer grâce à une fonction (= quelque chose que l'on peut détruire par une mutation, que ça sert à quelque chose, quand on le détruit ça a un impact.) On est en mesure de calculer la pression de sélection sur le génome humain: il ne varie pas , il est entre 5 et 8% du génome. Les biologistes de l'évolution calculent le pourcentage de fonctionnalité du génome. Plus le génome est utile, plus il y a de risques que des mutations détruisent la viabilité des descendants. Si 3% est fonctionnelle, une mutation a moins d'impact car elle a plus de chance de se trouver dans des mutations moins importantes.

L'homme est une espèce en évolution : on estime que le développement des césariennes a un impact sur la taille des nouveaux nés. Avant, si le bébé était trop gros, il mourrait et la mère avec, aujourd'hui ce n'est plus le cas. On a fait face à une pression de sélection évolutive.

Le big DATA : dans l'astronomie il y a aussi un nombre important d'informations. Attention !
« Ce n'est pas parce qu'on a beaucoup de données qu'on a beaucoup d'informations »
L'information n'est pas la connaissance, et la connaissance n'est pas la compréhension.
En génétique il faut traduire les données en informations.

FICHE RECAPITULATIVE

I. Biologie et évolution

Lorsque l'on travaille en génétique ou en génomique, ne pas se placer dans une perspective évolutive, c'est rater une partie de la problématique. Au long de l'évolution, le vivant s'est adapté, et de nombreux exemples montrent des particularités évolutives exceptionnelles.
La compréhension de l'évolution a été initiée par plusieurs scientifiques :
- Carl von Linné (1735) propose une classification binaire des espèces (nom de genre + nom d'espèce) et a regroupé les espèces par ressemblances, par caractéristiques
- Charles Darwin (1862) ajoute la dimension temporelle à la classification de Carl von Linné : cette organisation des espèces vivantes est le résultat d'un processus temporel
- Ernst Haeckel (1866) propose un arbre de l'évolution des espèces animales, dans lequel il rajoute l'homme en 1874 « Stammbaum des Menschen »
Butler Act (1925): loi interdisant dans le Tennessee de contredire l'histoire de la création divine de l'Homme, qui restera en vigueur jusqu'en 1968 !

II. L'ère de la génomique

Depuis un peu moins de 20 ans , nous sommes entrés dans l'ère de la génomique humaine : programme « Apollo » de la biologie moléculaire avec le **26 juin 2000**, la publication dans Nature d'un article annonçant le **premier séquençage du génome humain**. Plusieurs milliers de génomes sont aujourd'hui connus ou en cours de séquençage. La connaissance d'informations génomiques exhaustives aura et a déjà une influence profonde sur la recherche biologique et thérapeutique, d'où l'importance de la mise en œuvre des outils de la Bio-informatique.

La loi de Moore : la puissance des microprocesseurs double tous les 2 ans
Mais : le contenu des bases de données double tous les 15-18 mois
Les algorithmes deviennent une **véritable discipline transversale**

Invention du séquençage par fragmentation de l'ADN
-Méthode de Maxam et Gilbert → dégradation chimique sélective
-Méthode de Frederick Sanger → synthèse enzymatique sélective

Avec des contraintes importantes : on ne peut séquencer que 400 (début) à 1000 (fin) bases à la fois, et il faut une amorce complémentaire de la séquence traitée pour démarrer la polymérisation. Mais surtout, il faut reconstituer la séquence complète à partir de ces fragments. C'est à cette étape que la bio-informatique va être nécessaire pour assembler les fragments, les analyser et les annoter. **La bio-informatique est devenue incontournable !**

Stratégies de séquençage
•Petits génomes
Fragmentation aléatoire globale et reconstruction directe Méthode du « Whole Genome Shotgun » (WGS) : tout couper, tout aligner, tout séquencer, tout remettre dans l'ordre
•Grands génomes
Il faut procéder à un assemblage par chevauchements. Grâce à ces chevauchements, on est capable de générer des séquences, des contigs, qui sont un ensemble de séquences qui se chevauchent mutuellement pour couvrir toute une région.
Pour combler les trous dont la persistance est inévitable (séquences toxiques ou répétées), on peut intégrer des données de cartographie, faire une PCR par-dessus les trous, faire du re clonage sélectif, ou alors **intégrer des données des grands clones**(méthode la plus utilisée, qui permet également de vérifier la cohérence de l'assemblage en comblant les trous)

III. Les grands génomes

Plusieurs stratégies pour séquencer le génome humain :

→**Stratégie EST** : le principe est de ne séquencer que les régions transcrites en ARNm, grâce à la construction de banques d'ADNc après transcription inverse. Cette stratégie a pour intérêt **d'identifier les régions transcrites** du génome puisque l'ADNc est obtenu à partir des ARN, de **déterminer des profils d'expression tissulaire** puisque les ARNm sont issus de différents tissus et également de **s'affranchir des problèmes d'assemblage** puisque 1 ARNm = 1 gène. Cependant attention aux cellules différenciées qui n'expriment plus forcément le même génome

→**Séquençage exhaustif du génome humain** : un projet fou, car extrêmement cher, d'une durée pharaonique, très difficile, mais surtout dont l'utilité fait débat ! Pour réussir ce pari, le consortium public adopte une **approche hiérarchique « Map first, Sequence later »** qui privilégie le positionnement des fragments à leur séquençage immédiat ; ce qui permet notamment de répartir le travail.

Le **problème des séquences répétées** sera résolu en utilisant les « **mate pairs** » de Gene Myers (utilise non plus les chevauchements mais l'information d'appariement des fragments), ce qui aboutit à la construction de « scaffolds ».

IV. La traduction du génome

Le **génome est peu lisible** car il est codé par un alphabet à 4 lettres. Seul 2% du génome code pour des protéines, il est très morcelé. Ainsi lorsqu'on analyse le génome on est à la recherche **d'informations codantes au sein de l'ensemble des informations**.

Pour ce faire on utilise différentes techniques :

-On cherche des **signaux au niveau des promoteurs et des liaisons intron-exon** (signaux d'épissage) qui sont des motifs.

-On cherche aussi des choses **par le contenu,** par le biais des statistiques sur la composition des nucléotides. Il y a des contraintes appliquées sur le code génétique, des irrégularités périodiques.

-On compare aussi les **bandes EST du même organisme** dans le langage génomique.

-On peut aussi comparer avec d'autres **espèces voisines**

L'annotation génomique combine l'ensemble de ces techniques.

De plus, il existe des **biais dans la distribution des nucléotides** (% AT/CG) liées aux pressions de sélection exercées par l'évolution et à partir de statistiques caractéristiques d'un type de séquence, on connait tel biais qui correspond à telle catégorie. Donc on peut calculer par exemple la probabilité d'appartenance à une espèce.

Il existe des outils de prédictions basés sur des méthodes statistiques déterminées en balayant une séquence. Ça permet de cibler dans un génome les séquences qui pourraient nous intéresser.

La **biologie se base sur la comparaison de séquences** : quand on a deux molécules de fonctions apparentées, leurs séquences seront ressemblantes et inversement, si on a deux molécules avec une séquence semblable, alors elles auront surement une fonction apparentée. On est dans un système d'une évolution à partir d'un ancêtre commun. **La ressemblance est codée par un système de points.**

Il existe aussi des **algorithmes d'alignement** dans l'ordinateur qui calculent entre deux séquences le score d'alignement et déterminent la séquence avec laquelle il y a le meilleur alignement dans les bases de données. On peut faire des **comparaisons globales** ou chercher des **alignements locaux**. **BLAST** est un exemple de ces outils d'alignement de base de séquence.

Régions conservées = régions fonctionnelles : prédiction fonctionnelle. Il existe un lien entre la modification de la séquence/génétique et la modification du phénotype. **Moins il y de**

divergences, plus les espèces sont proches. La **synténie** est la propriété de conservation dans l'ordre des chromosomes.

Il y a une réelle difficulté d'analyse du génome car il est **très dilué**, contient des **pseudogènes** aussi nombreux que les gènes et des **séquences répétées** de gènes
ADN égoïste : élément du génome encadré par des séquences spécifiques souvent répétées elles-mêmes, qui code pour une/des protéines impliquées dans la réplication d'éléments autonomes.
On a également en notre sein 450 000 copies de **rétrovirus fossiles** endommagées : traces d'infections qui ont eu lieu dans la lignée humaine au cours de l'Histoire. Souvent délétées par des éléments de recombinaisons et d'évolution. Mais **certaines sont encore partiellement fonctionnelles.**

-**SNIPs** : variation de 1 nucléotides
-**Insertions et délétions** : ajout ou retrait d'un ou plusieurs nucléotides entre 2 individus
-**Répétitions de microsatellites**
-**Copy number variation** : duplications plus importantes
-**Variation de grande partie** : Alu
La **variabilité génétique chez l'homme est faible** : une variation tous les 1000 nucléotides. Elle est différente selon les populations humaines étudiées (très élevée en Afrique) mais il n'y a pas d'allèle spécifique d'une population donnée, c'est les fréquences alléliques qui varient.

V. L'utilisation de la génomique

En un siècle, la pharmacie a fait des progrès considérables. Ça a progressé de la manière suivante : on caractérisait les dysfonctionnements physiologiques et les mécanismes moléculaires associés. On identifiait par la suite les cibles moléculaires et des molécules actives. Depuis la génomique, tout a changé : on a 1/3 de gènes inconnus, 1/3 de gènes connus et 1/3 de gènes qui ressemblent à des gènes connus.
Avant on partait de la physiologie vers le gène cible et maintenant on part du gène séquencé et on remonte à la physiologie. C'est ce qu'on appelle l'approche pharmacogénomique.
Aujourd'hui toute la classification des espèces est basée sur l'analyse comparative des séquences qui est l'outil le plus précis

Histoire de la peste : Pourquoi cette maladie est aussi virulente ?
Pour optimiser sa dissémination l'élément pathogène doit optimiser deux choses en même temps :
-efficacité de prolifération dans l'hôte en évitant le système immunitaire.
-entre les hôtes
Il faut donc s'intéresser au mode de transmission :
-transmission directe
-transmission par un vecteur

« Ce n'est pas parce qu'on a beaucoup de données qu'on a beaucoup d'informations »
L'information n'est pas la connaissance, et la connaissance n'est pas la compréhension.
En génétique il faut traduire les données en informations.