Traitement informatique des données

Présentation du module:

Cette matière contribue au développement des acquis de l'étudiant en ce
Qui concerne les nouvelles bases de la bioinformatique

Connaissances préalables recommandées
Génétique
Biologie moléculaire
Microbiologie
Biochimie

Cours 05: Introduction à la génomique

Chargée de cours : ALIANE S

Introduction

En 1995, pour la première fois, la séquence complète du génome d'une cellule vivante a été déterminée. Il s'agissait d'Haemophilus influenzae, une bactérie responsable d'infections bronco-pulmonaires chez les jeunes enfants, dont le génome est composé d'un seul chromosome circulaire long de 1 830 140 paires de bases et comportant 1738 gènes. Au cours des 5 dernières années, les progrès ont été tout à fait spectaculaires : aujourd'hui (printemps 2002), les séquences d'une centaine de génomes complets sont connues, provenant de domaines très différents du Vivant : bactéries, archaebactéries, champignons, invertébrés, insectes, plantes. Cet effort a trouvé son point culminant au début de l'année 2001 avec la publication de la séquence « brute » du génome humain. À l'heure actuelle, cet effort se poursuit encore avec un grand nombre d'autres génomes en cours d'étude : bactéries, souris, poissons, plantes cultivées...

La biologie moléculaire est donc entrée depuis 1995 dans l'ère de la génomique : on dispose maintenant de l'information génétique exhaustive sur un nombre croissant d'organismes vivants et il est aujourd'hui possible d'aborder de manière globale un certain nombre de problèmes complexes dont on n'avait jusqu'à présent qu'une connaissance fragmentaire : voies métaboliques, interaction de la cellule avec l'extérieur, mécanismes globaux de régulation et de contrôle. Une nouvelle discipline est également née de la connaissance de ces séquences complètes de chromosomes : la génomique comparée. Il est maintenant possible de comparer deux organismes vivants à l'échelle de leur génome, de déterminer les gènes qu'ils ont en commun ou qui leur sont propres. Ce type d'analyse est très prometteur dans le contexte de l'identification sélective de gènes correspondant à des cibles thérapeutiques : en comparant par exemple une bactérie pathogène et une proche cousine non-pathogène, on peut essayer 9 10 Introduction de repérer les gènes impliqués dans la virulence de la souche infectieuse.

L'accélération du séquençage, permise en particulier par la robotisation et la parallélisation des méthodes d'analyse, nécessite un soutien de plus en plus important de l'outil informatique. Dans un premier stade, celui-ci est indispensable pour permettre l'assemblage du gigantesque « puzzle » que constituent les milliers ou millions de fragments de génome issus des automates de séquençage.

Ensuite l'informatique est un outil incontournable pour extraire et analyser l'information contenue dans ces gigabases (1 Gbase = 109 nucléotides) de séquence.

Le volume des données à traiter est considérable, aujourd'hui (printemps 2002) les banques de séquences rassemblent plus de 1011 nucléotides et leur taille augmente exponentiellement avec un temps de doublement de l'ordre de 15 à 18 mois. Il est clairement impossible de caractériser expérimentalement tous les gènes contenus dans ces séquences et c'est pourquoi l'analyse in silicio (grâce au silicium des microprocesseurs) doit venir au secours des biologistes pour compléter et guider les approches in vitro et in vivo.

a. Deux types de molécules support de la bioinformation : les acides nucléiques et les protéines

Le "matériaux de base" de la génomique et de la protéomique est la séquence : l'enchaînement ordonné et orienté de nucléotides (acides nucléiques) ou d'acides aminés (protéines).

ADN : Acide DésoxyriboNucléique

- macromolécule : chaîne nucléotidique
- constituée par un enchaînement d'unités élémentaires : les désoxyribonucléotides
- forme de stockage de l'information génétique. Cette information est représentée par une suite linéaire de gènes
- formée de deux brins complémentaires enroulés en double hélice ce qui lui permet de se dupliquer en deux molécules identiques entre elles et identiques à la molécule mère

On distingue :

- l'ADN du génome du noyau
- l'ADN du génome mitochondrial
- l'ADN du génome chloroplastique

ARN : Acide RiboNucléiques

- macromolécule : chaîne nucléotidique
- constitués par un enchaînement d'unités élémentaires : les ribonucléotides
- forme qui permet de transférer et de traiter l'information dans la cellule
- le plus souvent formé d'un simple brin

On distingue :

- les ARN messagers ou ARNm : ils sont transcrits à partir d'un gène (ADN). Ils sont ensuite traduits en protéines.
- les ARN de transfert
- les ARN ribosomaux
- les ARN nucléaires
- les divers "petits" ARN non codants

Protéines

- macromolécule : chaîne polypeptidique
- constituées par un enchaînement d'unités élémentaires : les acides aminés
- l'ensemble des protéines assurent les principales fonctions cellulaires
- se replient sur elles-mêmes et adoptent une conformation ou structure particulière dans l'espace. Cette structure tridimensionnelle est à l'origine de la fonction des protéines et de leur spécificité de cette fonction.

Les chaînes nucléotidiques possèdent 2 extrémités distinctes : on peut donc les représenter de manière orientées de l'extrémité dite 5' vers l'extrémité dite 3'.

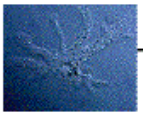
En conséquences, les chaînes nucléotidiques sont écrites sous forme d'une succession ordonnée et orientée de lettres qui représentent les unités élémentaires (les nucléotides) :

- ADN : 4 nucléotides = C, G, A et T
- ARN : 4 nucléotides = C, G, A et U

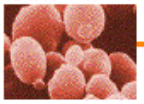
La taille des génomes nucléaires varie énormément au sein des Eucaryotes :

- de 1 à 1000 chez les plantes
- de 1 à 3300 chez les animaux
- de 1 à 300.000 chez les protistes (algues, amibes, euglènes, …)

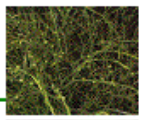
L'un des plus célèbres génome séquencé est celui de l'homme de Neandertal (Green et al., 2010).





	Taille du génome (nucléotides)	Nbre de gènes (*protein-coding*)
Amoeba dubia	~ 670 000 000 000	?
Psilotum nudum	~ 250 000 000 000	?
Fritillaria assyriaca	~ 100 000 000 000	?
Necturus lewisi	~100 000 000 000	?
Homo sapiens	2 900 000 000	23 000
Vitis vinifera	487 000 000	30 400
Drosophila melanogaster	160 000 000	14 000
Arabidopsis thaliana	115 000 000	28 000
Caenorhabditis elegans	98 000 000	19 400
Saccharomyces cerevisiae	12 500 000	5 800
Escherichia coli	4 600 000	4 300





Source : B. Dujon (2008)

GOLD ("*Genomes OnLine Database*") : base de données des génomes séquencés et en cours de séquençage.

b. "Préhistoire" du séquençage des acides nucléiques et séquençage dans l'espace

Un énorme effort humain, financier, technologique, a été fait dans les années 90 pour obtenir des outils pour les premiers pas du séquençage, de plus en plus performants et surtout automatisés.

Pour le séquençage des premiers génomes "historiques" (entre autre le génome humain), l'automatisation a requis dans les années 1990 / 2000 le développement :

- de système d'électrophorèse capillaire piloté par ordinateur qui ont remplacé les gel à plat
- de robot passeur d'échantillon qui permet d'enchaîner les échantillons
- de marqueurs fluorescents dont la lumière réfléchie après excitation par un laser est captée par une cellule CCD (*Charge-Coupled Device*)

•de suites logicielles permettant l'analyse des signaux sortant des séquenceurs et leur mise en forme sous forme de fichiers analysables (électrophorègramme et séquence)

en 2016, un séquenceur Illumina

Capacité de séquençage : 5 milliards de lectures x [300 paires de bases] = 1500 miliards de nucléotides en quelques heures à 1 jour.

le séquenceur ultra-portable MinION (*Oxford nanopore technology*) - dit de 3è génération - a été utilisé en temps réel sur le terrain lors de la crise Ebola de 2015 et de la crise Zika en 2016 (Quick et al., 2016).

Les dernières avancées

Les technologies avec des nanopores sont de plus en plus performantes. En 2018, le séquençage et l'assemblage *de novo* d'un génome humain s'est appuyé sur un protocole :

- Qui a généré des lectures ultra-longues : N50 > 100 kb avec des longueurs de lecture jusqu'à 882 kb.
- La précision de l'assemblage (après incorporation des données de séquençage à lecture courte complémentaires) a dépassé 99,8%.
- Des lectures ultra-longues ont permis l'assemblage du locus du complexe majeur d'histocompatibilité de 4 Mo dans son intégralité.
- Voir Jain *et al*. (2018)

En juillet 2016, la technologie Minion a été envoyée par la NASA dans la station spatiale internationale ("*International Space Station*", ISS) pour les premiers séquençages effectués dans l'espace d'ADN extra-terrestre potentiel.

Elle a été testée avec succès dans des conditions de gravité comparables à celles qui règnent sur Mars (G = 0,378), sur la lune (G = 0,166) et sur Europa (sattelite de Jupiter - G = 0,134). Voir : Carr *et al*. (2020).

Le séquençage en routine du génome humain est devenu possible avec le séquenceur PromethION (*Oxford Nanopore Technologies*) qui possède 3000 capteurs et 12.000 pores : ils génèrent en moyenne 70 Go de données permettant une couverture 20 X du génome humain.

Enfin, la précision des logiciel d'appel de base est sensiblement améliorée par des algorithmes basés sur des modèles de Markov cachés ou des réseaux de neurones.

Historique à signaler (Shendure et al. 2017)

1953: séquençage de l'insuline (Frederick Sanger)
1965: séquençage de l'ARNt alanine
1968: séquençage des extrémités cohésives de l'ADN du phage lambda
1977: technique de séquençage de l'ADN de Allan Maxam & Walter Gilbert
1977: technique de séquençage de l'ADN de Frederick Sanger
1981: vecteur du phage M13 de Messing
1986: Détection des bases par fluorescence au cours du séquençage par électrophorèse
1987: sequenase
1988: premier séquençage par incorporation progressive de dNTP
1990: séquençage par extrémités appariées
1992: colorants Bodipy
1993: colonies d'ARN in vitro
1996: pyroséquençage
1999: colonies d'ADN dans des gels 2000: séquençage massivement parallèles de signatures par ligation

2003: PCR en émulsion pour générer des colonies d'ADN sur des billes

2003: séquençage massivement parallèle par synthèse sur molécule unique ("*single-molecule*")

2003: guides d'ondes en mode zéro pour l'analyse de molécules uniques

2003: séquençage de colonies d'ADN par synthèse dans des gels

2005: fluorophores de terminaison réversible à quatre couleurs

2005: séquençage par ligation de colonies d'ADN sur billes

2007: capture de séquences cibles à grande échelle

2010: détection directe de la méthylation de l'ADN au cours du séquençage d'une seule molécule

2010: tunnel à électrons à résolution de base unique par détecteur à semi-conducteurs

2011: séquençage avec des semiconducteurs par détection de protons

2012: séquençage par nanopore

2012: préparation de bibliothèque simple brin d'ADN ancestral

2018: séquençage et assemblage *de novo* d'un génome humain avec des lectures ultra-longues (N50 > 100 kb avec des longueurs de lecture jusqu'à 882 kb)

2. Détermination des séquences de nucléotides

a. Méthode historique de Fréderick Sanger

Fréderick Sanger est décédé en 2013. Il fût l'un des plus admirables scientifiques biochimistes (Prix Nobel de chimie 1958 et Prix Nobel de chimie 1980),

Bien qu'elle ait cédé la place aux nouvelles technologies de séquençage, la méthode de Sanger est historiquement capitale puisqu'elle a permis les premiers séquençages de génomes complets :

- *Haemophilus influenzae* 1995
- *Saccharomyces cerevisiae* 1996
- *Escherichia coli* K-12 1997
- *Caenorhabditis elegans* 1998
- *Arabidopsis thaliana* 2000

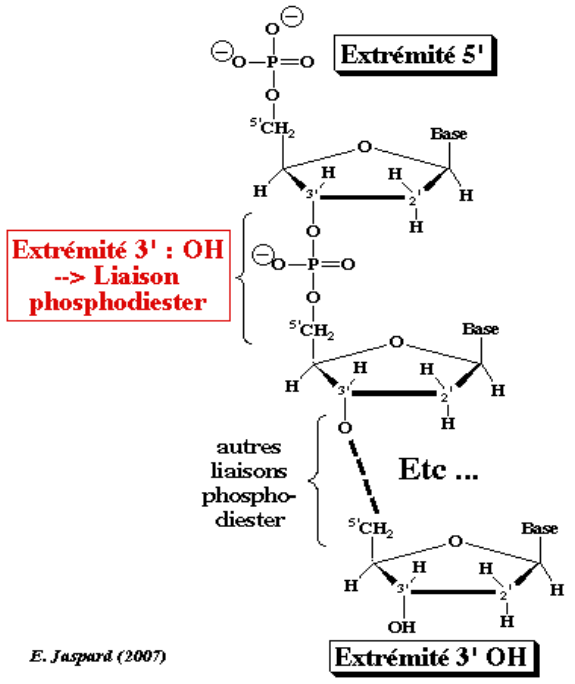
- *Drosophila melanogaster* 2000
- Homme 2001
- *Mus musculus* 2002
- Rat 2004

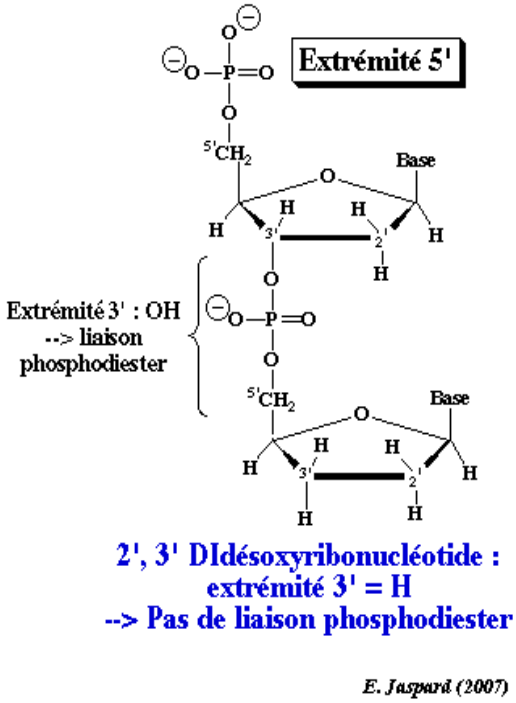
Les nucléotides au sein des acides nucléiques sont liés par une liaison phosphodiester qui s'établit entre le groupement OH sur le carbone 3' du ribose du nucléotide dit en position 5' et le phosphore du groupe phosphoryle en position α du nucléotide dit en position 3'.

La méthode de séquençage de Sanger (dite par terminaison de chaîne) utilise des nucléotides appelés didésoxyribonucléotides (ddNTP) qui ont un atome d'hydrogène à la place du groupement OH sur le carbone 3' du ribose.

Ils peuvent donc être incorporés dans un brin d'ADN en cours de synthèse, mais ils ne permettent pas qu'un autre nucléotide soit incorporé après eux : en effet, l'absence de l'atome d'oxygène en 3' empêche la formation d'une nouvelle liaison phosphodiester.

L'allongement du brin d'ADN s'arrête donc au niveau du ddNTP incorporé, d'où terminaison de la synthèse de l'ADN.





Source : Sanger et al., 1977

La méthode de séquençage de Sanger utilise une amorce marquée radioactivement ("*dye-labeled primer*") car la polymérase nécessite un court fragment complémentaire du brin à séquencer pour initier la synthèse du brin copie.

Quatre réactions de séquençage sont donc menées en parallèle dans quatre tubes distincts, contenant chacun un seul didésoxyribonucléotide (ddTTP, ddATP, ddCTP et ddGTP) :

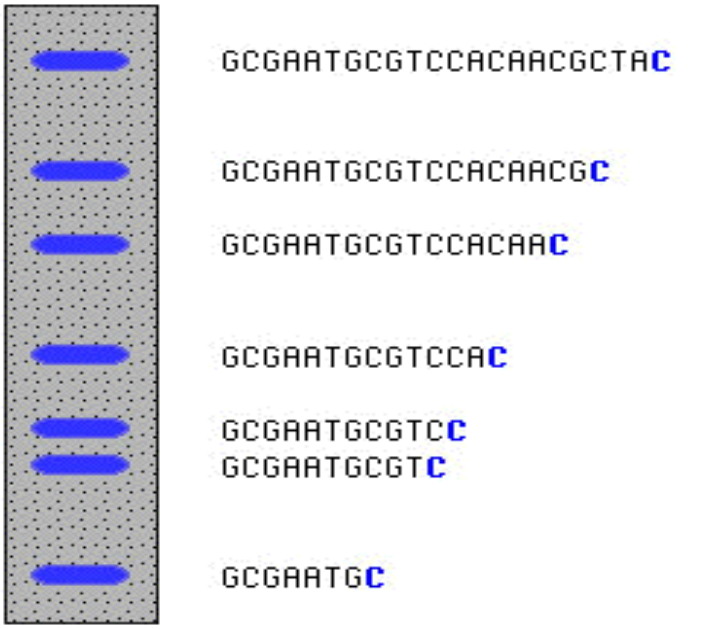
- ADN matrice + amorce marquée + dNTP + ddTTP
- ADN matrice + amorce marquée + dNTP + ddATP
- ADN matrice + amorce marquée + dNTP + ddCTP
- ADN matrice + amorce marquée + dNTP + ddGTP

Dans chaque tube, toutes les copies d'ADN synthétisé sont interrompues derrière le même nucléotide.

Le rapport des concentrations entre les dNTP et les didésoxyribonucléotides (ddNTP) et le nombre de réactions simultanées catalysées par la polymérase assure statistiquement que toutes les copies partielles intermédiaires possibles de la molécule d'ADN sont synthétisées.

On sépare alors les copies selon leur taille par une migration électrophorétique dans un gel poreux (entre 2 larges plaques de verre), le contenu de chaque tube étant déposé dans un puits distinct. Ces gels permettent de séparer deux intermédiaires consécutifs qui ont une différence de taille d'un seul nucléotide.

Exemple ci-dessous : profil d'électrophorèse du contenu du tube avec le ddCTP. Toutes les copies intermédiaires d'ADN synthétisé sont terminées par un C.

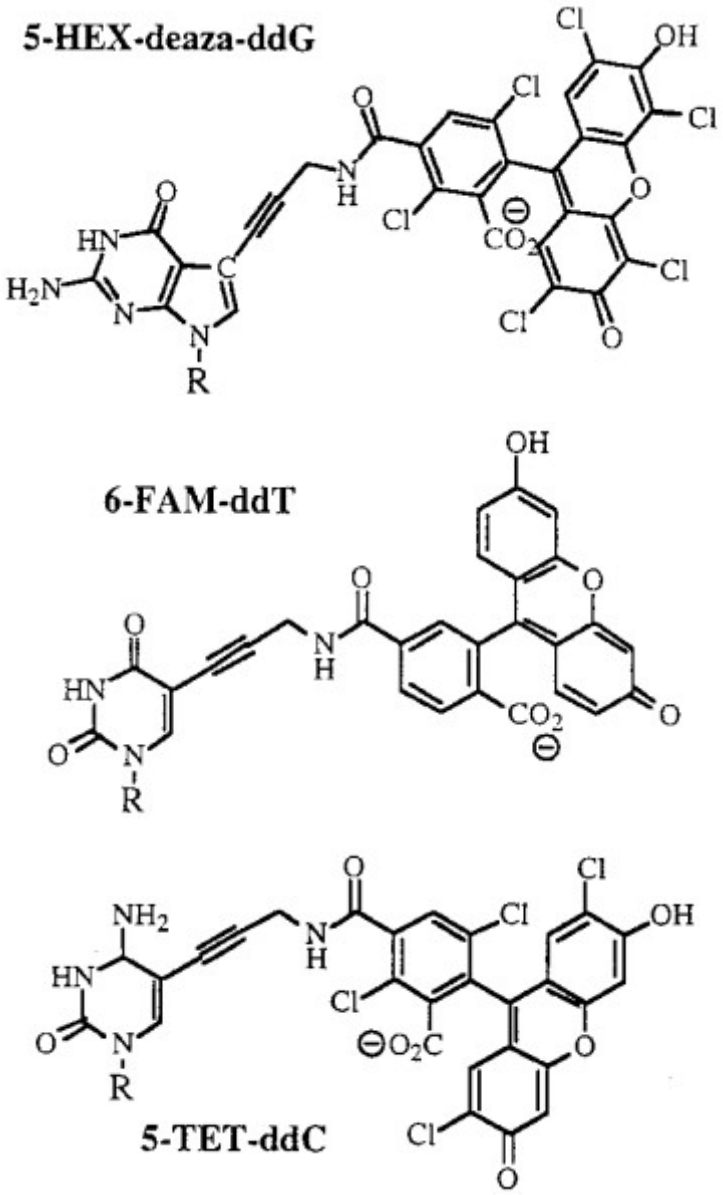


b. La technique de séquençage avec des didésoxyribonucléotides fluorescents ("dye terminator sequencing")

Smith et al. (1986) "Fluorescence detection in automated DNA sequencing" Nature 321, 674 - 679

Cette technique utilise des didésoxyribonucléotides dont chacun est marqué par un fluorophore spécifique. Les fragments d'ADN synthétisés portent ce fluorophore terminal. On les appelle des terminateurs d'élongation ou "*Big Dye Terminators*" ou "*Dye-labeled terminator*".

Ci-dessous, exemple de structures de ddNTP fluorescents :



Source : Drandis, 1999

- 6-TAMRA-ddTTP / 6-FAM-ddTTP / 5-TET-ddCTP / 5-HEX-deaza-ddGTTP
- R = 2',3'-dideoxyribose-5'-triphosphate / FAM = 6-carboxyfluorescéine
-

Améliorations apportées par la méthode des ddNTP fluorescents par rapport à la méthode de Sanger

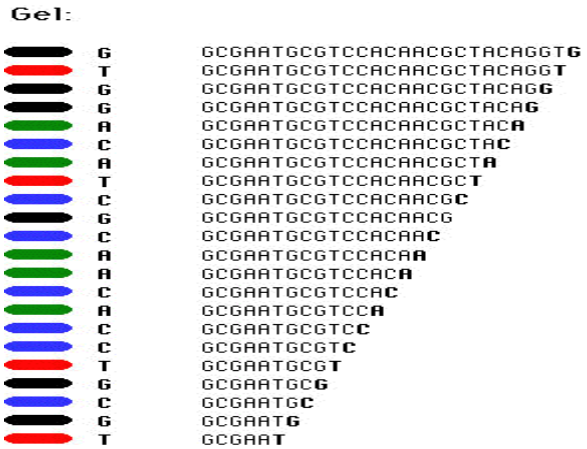
a. La méthode initiale de Sanger utilisant une amorce marquée radioactivement est plus laborieuse, coûteuse (4 réactions distinctes) et dangereuse (radioactivité) que celle des ddNTP fluorescents.

b. Par ailleurs, l'un des problème du séquençage est la formation de "faux-stop" : c'est la terminaison prématurée d'une copie qui implique un désoxyribonucléotide à la place d'un ddNTP. Avec la méthode des ddNTP fluorescents, les "faux-stop" ne sont pas détectés car ils ne fluorescent pas.

c. Avec la méthode des ddNTP fluorescents, il n'y a qu'une réaction de séquençage en présence des 4 didésoxyribonucléotides :

ADN matrice + dNTP + ddCTP fluorescent bleu + ddATP fluorescent vert + ddGTP fluorescent jaune + ddTTP fluorescent rouge

•L'excitation se fait à 2 longueurs d'onde différentes par un laser à l'argon. L'émission de fluorescence est mesurée à 4 longueurs d'onde correspondant aux 4 fluorophores.
•Chaque base a donc un signal spécifique qui permet de l'identifier lors de son passage dans le faisceau d'un photomètre situé à la sortie du capillaire.
•L'analyse des signaux reçus est réalisée par un ordinateur et permet de reconstituer la séquence avec une grande précision (figure ci-dessous).



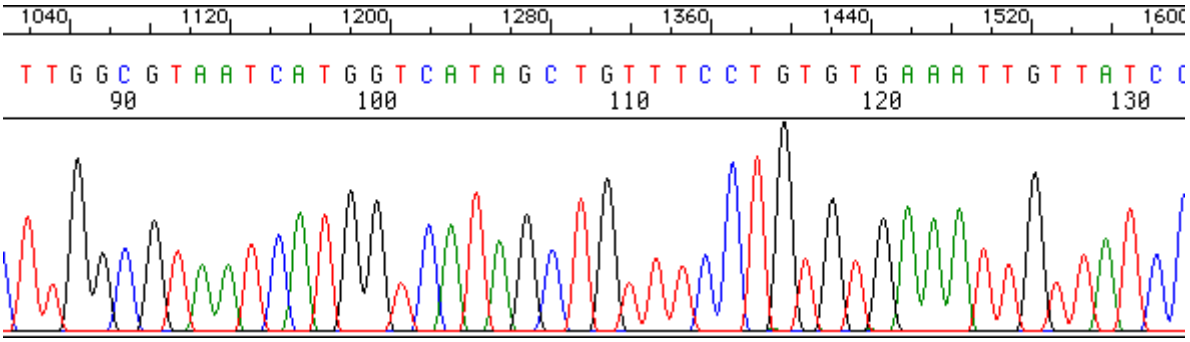
Source :University of Michigan

le séquenceur "*MegaBACE*" (société Amersham) : plateforme capillaire à haut débit pour le séquençage d'ADN.

Schématiquement, l'appareil est composé de 96 capillaires, d'un système d'électrophorèse, d'un laser et d'une caméra CCD (*Charge-Coupled Device*).

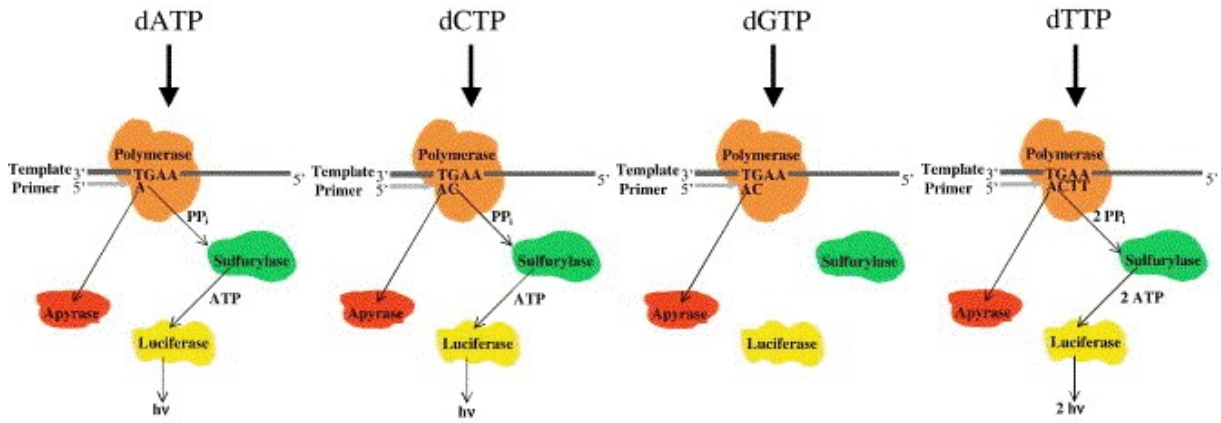
- Les capillaires (diamètre environ 250 µm), sont remplis d'un polymère qui sert de tamis moléculaire.
- Les molécules d'ADN sont introduites à une extrémité des capillaires par électro-injection et migrent ensuite tout au long de ceux-ci sous l'effet d'un très haut voltage (8500 volts) de façon à les séparer en fonction de leur longueur.
- Près de l'anode, un rayon laser traverse chaque capillaire afin d'exciter les ddNTP fluorescents incorporées à l'ADN au cours de la réaction de séquençage.
- Une caméra CCD mesure l'émission de fluorescence au fur et à mesure que les copies d'ADN passent devant le laser. Les ddNTP fluorescents sont distingués les uns des autres selon la longueur d'onde émise. Exemples : TAMRA : excitation 552 nm - émission 575 nm / FAM : excitation 490 nm - émission 520 nm.

La dernière étape est la lecture des profils bruts ou "*base-calling*" (détermination de la séquence par appel de bases).



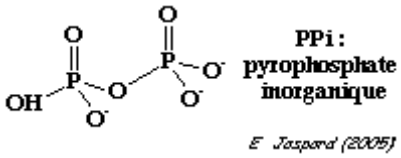
3. Méthode du pyroséquençage

Elle permet d'effectuer un séquençage moins cher et rapide qu'un séquençage par la méthode de Sanger car elle ne nécessite pas de clonage et la lecture de la séquence est directe,



Source : Ahmadian *et al*., 2006

Les désoxyribonucléotides triphosphate (dNTP) sont ajoutés l'un après l'autre (et non pas tous ensemble comme dans la méthode de Sanger). Si le désoxyribonucléotide ajouté est complémentaire du désoxyribonucléotide du brin matrice, il est incorporé dans le brin en cours de synthèse et un pyrophosphate inorganique (PPi) est libéré.



L'ATP sulfurylase transforme stoechiomètriquement le pyrophophate libéré en ATP en présence d'un substrat : l'adénosine 5' - phosphosulfate (APS).

- L'ATP formé est utilisé par une luciférase qui transforme la luciférine en oxyluciférine qui génère un signal lumineux dans le visible proportionnel à la quantité d'ATP.
- L'apyrase dégrade les nucléotides non incorporés et l'excès d'ATP.
- Remarque importante : l'ATP est le substrat de la polymérase (pour l'élongation du brin en cours de synthèse) mais il est aussi formé par l'ATP sulfurylase. Pour la polymérisation, on utilise donc un analogue de l'ATP : la désoxyadénosine α-thio triphosphate (dATPαS) qui n'est pas un substrat de la luciférase.

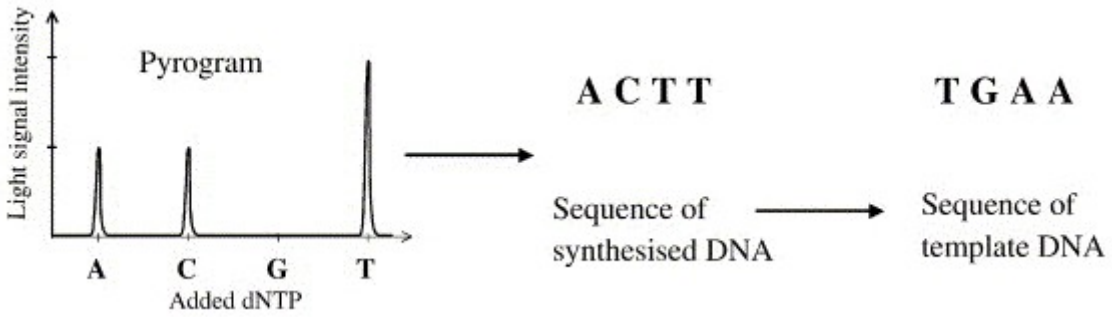
$$[ADN]_n + dNTP \xrightarrow{\text{Polymérase}} [ADN]_{n+1} + PP_i$$

$$dNTP \xrightarrow{\text{Apyrase}} dNDP + dNMP + phosphate$$

$$ATP \xrightarrow{\text{Apyrase}} ADP + AMP + phosphate$$

E. Jospard (2008)

Le capteur CCD du séquenceur capte le signal lumineux et le traduit par un pic sur le pyrogramme,



Source : Ahmadian et al., 2006

La hauteur du pic est proportionnelle à l'intensité du signal lumineux, elle même proportionnelle au nombre de nucléotides incorporés au même moment. On déduit la séquence à partir de la taille des pics obtenus.

En cas de mélange de nucléotides à une même position (polymorphisme de séquence), la taille des pics permet d'avoir une quantification de la proportion de brins porteurs de l'un ou l'autre des nucléotides.

Exemple d'application du pyroséquençage à l'étude du transcriptome de *Arabidopsis thaliana* : Weber et al. (2007) "Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing" Plant Physiol. 144, 32-42.

Méthode	longueur des lecture (nucléotides)	nombre de lectures	total par tour ("*run*") (Mpb)	coût relatif par nucléotide
Sanger	700 - 800	96	0,07	1
pyroséquençage	250	400.000	100	0,1
phase solide	25 - 35	40 à 80 millions	1000 - 2000	0,01

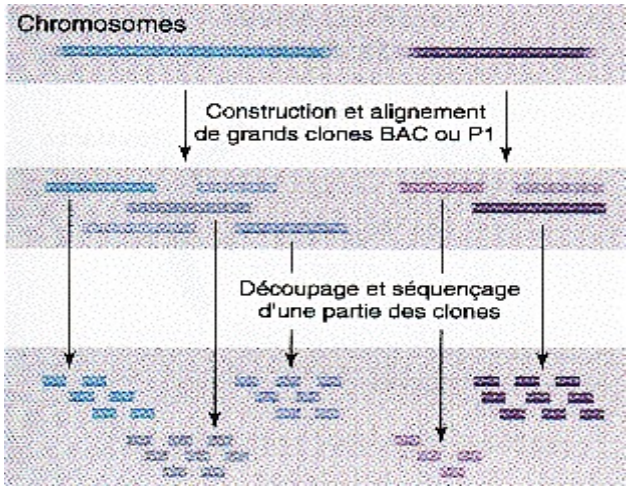
4. Stratégies initiales de séquençage des génomes

a. La méthode hiérarchique ou "clone par clone"

Le génome est découpé en un nombre "restreint" (quelques dizaines de milliers) de fragments de grande taille (50 à 200 kilo paires de base) qui couvrent l'ensemble du génome.

Ces fragments sont clonés dans des vecteurs spéciaux : les YAC ("*Yeast Artificial Chromosome*" - problème d'échange de fragments d'ADN), les BAC ("*Bacterial Artificial Chromosome*") ou des vecteurs dérivés du phage P1 (les PAC).

Une carte physique des clones est établie pour faciliter l'obtention de la séquence finale du génome : elle permet d'ordonner les clones dans le génome.



Source : Gibson and Muse, 2004

Les cartes de liaison disposent des marqueurs ordonnés le long des chromosomes par la mesure de leur liaison deux à deux. Ces cartes de liaison permettent de se repérer dans le génome et sont une aide essentielle dans la construction de la carte physique.

Un sous-ensemble avec un minimum de recouvrement (pour avoir une couverture la plus complète possible du génome) est ensuite choisi et séquencé en "vrac" (voir ci-dessous) : chaque clone de grande taille est découpé en un grand nombre de fragments de petite taille (environ 2000 paires de bases) et les extrémités sont séquencées individuellement.

Les problèmes d'assemblage ne se posent qu'à l'échelle des grands fragments et sont facilement résolus en multipliant le nombre de lectures dans ces zones.

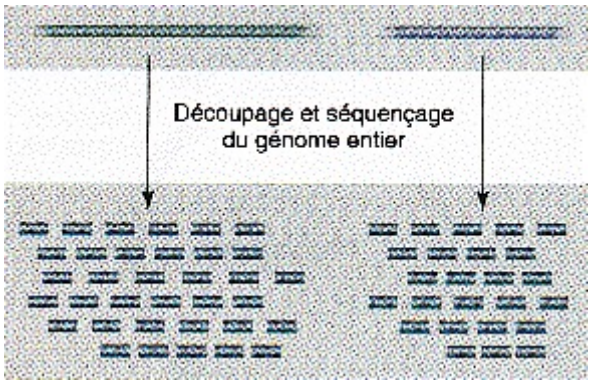
b. La méthode de séquençage aléatoire global ou "en vrac" ou "shotgun"

C'est une méthode très différente et complémentaire de la méthode hiérarchique.

Une carte de grands fragments ordonnés n'est pas établie au préalable.

Un trés grand nombre de séquences sont obtenues de façon aléatoire à l'échelle du génome entier. Les extrémités d'une partie de ces fragments sont séquencées. Puis ces séquences sont assemblées selon leurs recouvrements.

Du fait du grand nombre de fragments et du clonage, certaines séquences ne sont jamais séquencées.



Source : Gibson and Muse, 2004

La difficulté d'assemblage est beaucoup plus grande que dans la stratégie "clone par clone" et le nombre énorme de comparaisons de séquences nécessite une puissance de calcul considérable.

Il n'est pas possible, pour combler les trous entre les contigs (voir ci-dessous), de diriger le travail de séquençage supplémentaire sur un grand fragment bien identifié.

Compléments sur la méthode "shotgun"

C'est un processus aléatoire d'échantillonnage de N lectures de taille L, pour un génome de taille G :

- couverture : $a = N . L / G$
- nombre de contig obtenus (N_c) en fonction de la couverture : $N_c = (a . G / L) e^{-a}$
- taille moyenne des contigs : $L_c = (e^a - 1) . L / a$ (Lander and Waterman, 1988),

Evolution des "stratégies" de séquençage de type "shotgun" :

- Roach *et al.* (1995)
- stratégie "parking"
- les méthodes "*paired end sequencing*"
- les nouvelles technologies de séquençage à très haut débit, non limitantes (exemple : "*WGS sequences = whole genome shotgun sequences*")

Quelle que soit la stratégie adoptée, lors de l'assemblage terminal du génome, il faut éliminer :

•Les fragments d'ADN contaminants d'origine bactérienne.

•Les clones ne provenant pas, à l'origine, d'un même fragment du génome du fait d'une recombinaison à l'intérieur du BAC ou d'une mauvaise annotation lors de la construction de la collection de fragments pour la phase de séquençage en vrac.

•Les séquences répétées peuvent aussi poser un problème lors de l'assemblage des grands génomes car elles peuvent conduire à assembler 2 séquences provenant de régions distantes du génome. Lors de l'assemblage, elles sont donc "masquées" par des programmes informatiques tel que *RepeatMasker*. Ces logiciels remplaçent les nucléotides de ces régions par le symbole "N" qui décrit n'importe quel nucléotide.

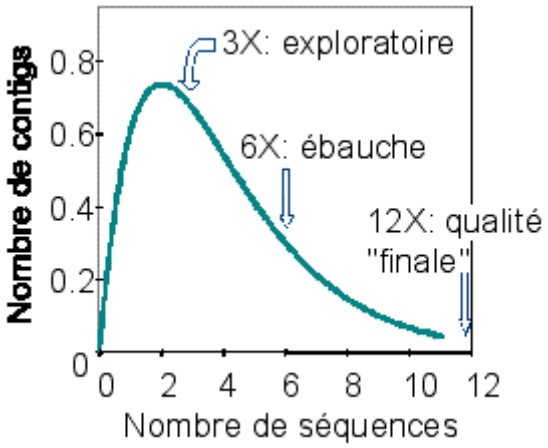
5. Les contigs et l'assemblage, les trous, l'appel de base

Avec les technologies encore courantes dans de nombreux laboratoires, chaque séquençage ne permet d'obtenir une lecture que de quelques milliers de paires de base. Il n'est donc pas possible de séquencer en une seule fois des molécules d'ADN aussi grandes que les chromosomes.

Pour reconstituer ces immenses séquences, il faut effectuer un grand nombre de séquençages, plusieurs fois supérieur à la taille du chromosome. Ces séquençages redondants permettent :

•de raccorder les séquences les unes aux autres
•de s'assurer de la qualité du résultat de chaque lecture

Pour les premiers séquençages des génomes (avant l'avènement des nouvelles technologies de séquençage à très haut débit), la redondance était d'un facteur 8 à 10 (une profondeur de 8 à 10X).



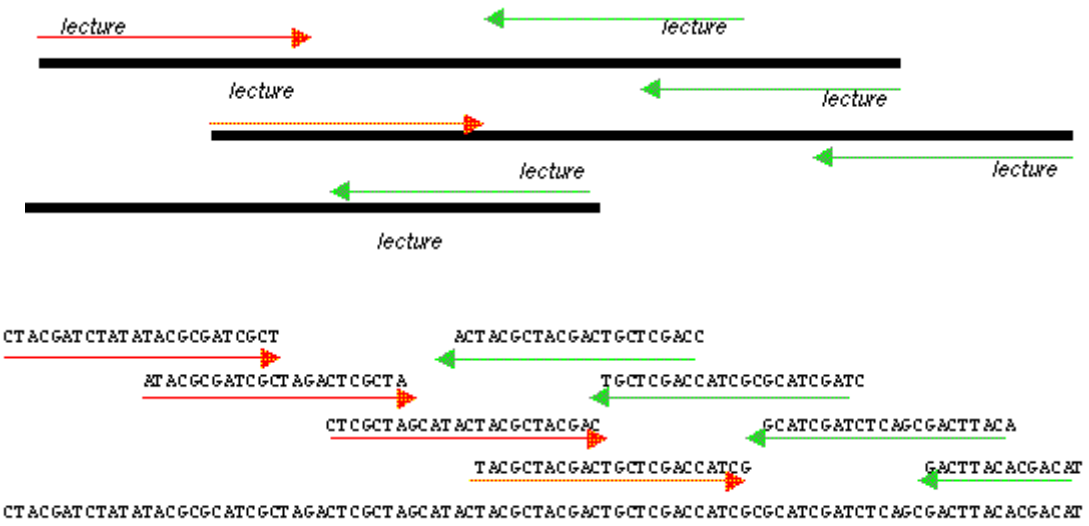
Source : Dujon, 2008

Celà signifie :

- fractionner le fragment à séquencer en sous-fragments
- effectuer un nombre de séquençage tel que l'ensemble de ces séquençages, mis bout à bout, représentent 10 fois la longueur de la séquence du fragment initial
- en d'autres termes, chaque base du fragment initial doit apparaître en moyenne dans 10 lectures

L'assemblage

La comparaison des séquences permet d'aligner les parties qui se recouvrent partiellement ou chevauchantes.



Source : Genoscope - FAQ

es séquences chevauchantes peuvent être reliées en enchaînements plus grands que l'on appelle contigs.

En reliant l'ensemble des contigs, on reconstitue des séquences de plusieurs millions à plusieurs dizaines de millions de nucléotides (les "*scaffold*").

Ces opérations sont effectuées par des programmes bioinformatiques.

Les trous ou "gap" : Comme le séquençage est effectué sur des sous-fragments pris de manière aléatoire, même avec un tel niveau de redondance, il reste des parties non assemblées : des trous ("*gap*") qui peuvent être "comblés" par un travail ciblé.

Scaffold : ensemble de contigs orientés et ordonnés. Les trous ("*gaps*" - voir ci-dessous) sont de longueur connue.

Mapped scaffold : ensemble de *scaffolds* localisés le long des chromosomes (pas forcément ordonnés ou orientés). Les trous sont de longueur inconnue.

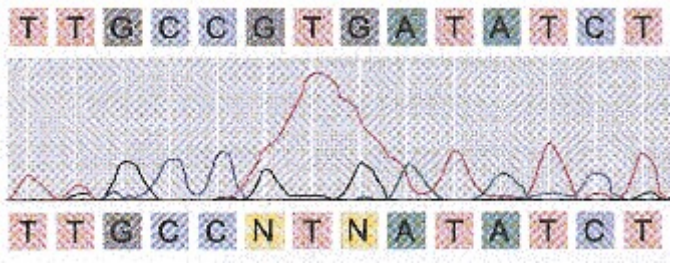
Pour déterminer les relations de voisinage des contigs, les liens clones sont considérés, c'est-à-dire les lectures obtenues aux deux extrémités d'un même fragment d'ADN. On recherche parmi ces paires celles qui s'ancrent dans deux contigs différents.

Cela permet de jeter un pont entre les deux contigs et de les orienter. De plus, le fragment d'ADN "à cheval" sur le trou entre les deux contigs peut faire l'objet d'un séquençage supplémentaire, ce qui permet de combler le trou.

La lecture des profils bruts ou "base-calling" : c'est la détermination de la séquence par appel de bases qui s'effectue en routine par des programmes informatiques qui déterminent l'identité des bases, comparent les séquences et fournissent une plate-forme intuitive de correction.

La suite logicielle publique développée à l'Université de Washington contient les programmes :

- Phred : il convertit les fichiers "traces" (chromatogramme au milieu de la figure ci-dessous) en séquences qui sont immédiatement déposées dans des banques.
- Phrap / CrossMatch / Swat : ensembles de programmes pour l'assemblage de séquences d'ADN en contigs.
- Consed : outil graphique de visualisation et d'édition des séquences assemblées par Phrap.
- La fonction "Autofinish" (Gordon *et al*. 2001) du programme Consed permet de combler les trous en proposant des amorces et en identifiant des matrices d'ADN qui permettent de franchir les discontinuités entre 2 contigs.



Source : Gibson and Muse, 2004