

Séquençage de génomes

Séquençage

Comment déterminer la séquence d'une molécule ADN (ou ARN) ?

Notre répertoire d'outils de biotechnologie :

- hybridation
- PCR - *polymerase chain reaction*
- enzymes de restriction
- clonage
- séquençage : méthode de Sanger

Jouer avec les brins

- double → simple : casser les liaisons hydrogène
dénaturation (par chaleur ou agents chimiques)
- simple → double
 1. **hybridation** appariement de deux brins complémentaires
 2. enzymes (**polymérase**) pour compléter un brin (dans la présence de nucléotides à incorporer)

Brins 2

hybridation : température de fusion (*melting temperature*) — 50% de ADN dénaturé

T_m dépend de la séquence [G-C plus fort que A-T], et conditions expérimentales (agents chimiques)

approximation par formule de Wallace $T_m = 2\{A, T\} + 4\{C, G\}$ [en Celsius]

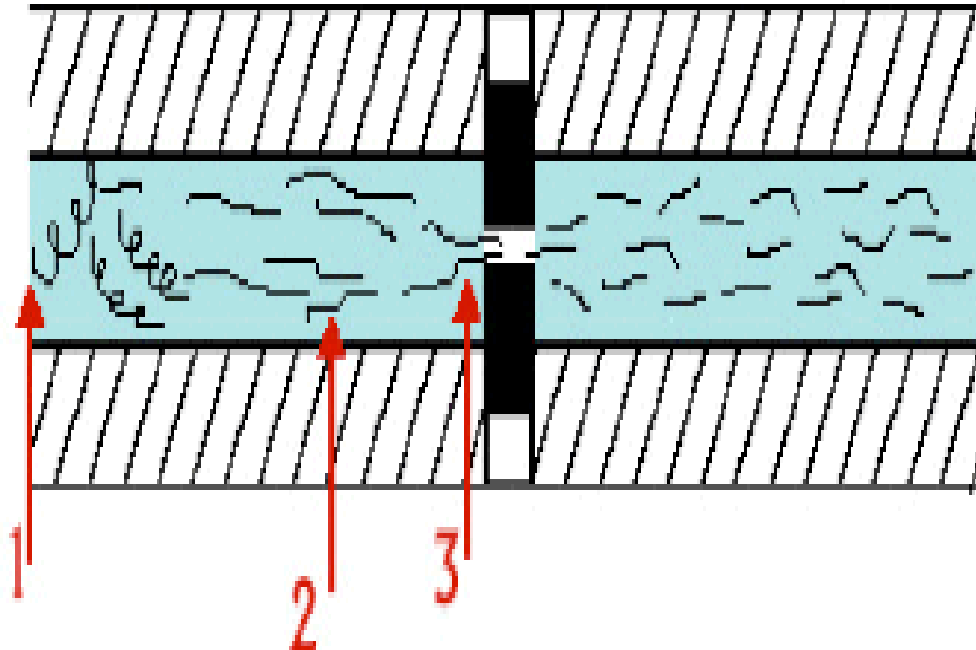
polymérase : construit le brin $5' \rightarrow 3'$ à partir d'une région déjà complétementé (**primer**)

primer aléatoire (tous 6mers) ou sélectionné

Comment couper l'ADN en petits morceaux ?

1. au hasard (p.e., **sonication** — par ondes de haute fréquence, **cisaillement**/*shearing*)
2. moins hasard (répérable) : **enzymes de restriction**

Cisaillement



cisaillement hydrodynamique : les molécules s'accélèrent en approchant la région étroite : les forces d'allongation brisent les molécules

GeneMachines

Enzymes de restriction

endonucléase — coupe l'ADN à un site spécifique (4–8 pb typiquement)

Nom	Site
AluI	AG.CT
EcoRI	G.AATTC
HindIII	A.AGCTT
centaines d'autres...	

ER 2

site d'un ER est un palindrôme : le complément inverse a la même séquence

-V--->
5' GAATTC
CTTAAG 3'
<---^ -

sticky ends

⇒ hybridation d'amorces compatibles (d'origines différentes)
+ enzyme de **ligase** pour les liaisons covalentes

ER 3

Site reconnu et point de clivage

Nom	Site
KasI	G.GCGCC
NarI	GG.CGCC
EheI	GGC.GCC
BbeI	GGCGC.C
HaeII	RGCGC.Y

$(R = \{A, G\}, Y = \{C, T\})$

KasI – BbeI sont *isoschisomeres*)

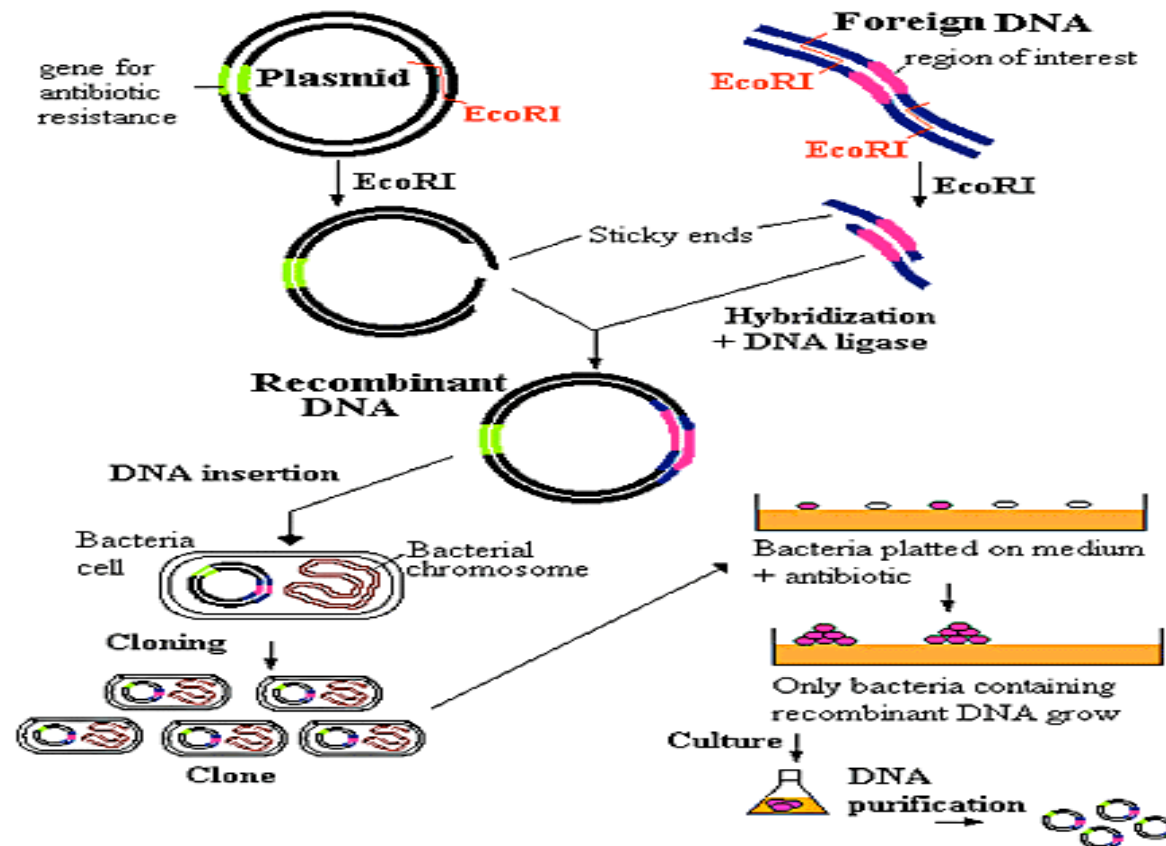
HaeII et BbeI sont *compatibles* : v. sticky end GGCGC

Clonage : idée

Pour créer des copies de fragments d'ADN (amplification *in vivo*)

1. fragments d'ADN (après coupure par ER)
2. insertion dans **vecteur** d'une cellule hôte (bactérienne ou virale)
3. culture et purification

Clonage



Cloning into a plasmid

Lodish et al. *Molecular Cell Biology*, 4th ed., W. H. Freeman 1999 ; Access Excellence

Plasmid

Caractéristiques nécessaires :

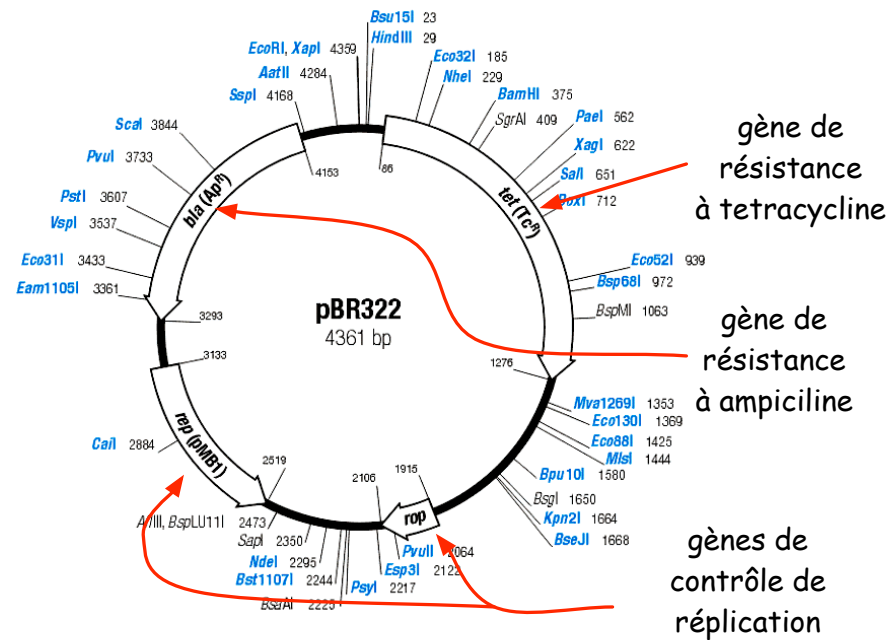
(Ins) site de restriction (avec *sticky ends*) pour insertion

(Pla) criblage pour cellules avec plasmids

(Frg) criblage pour plasmids avec fragments insérés

Thème général de **(Frg)** : fonction d'un gène interrompue par l'insertion

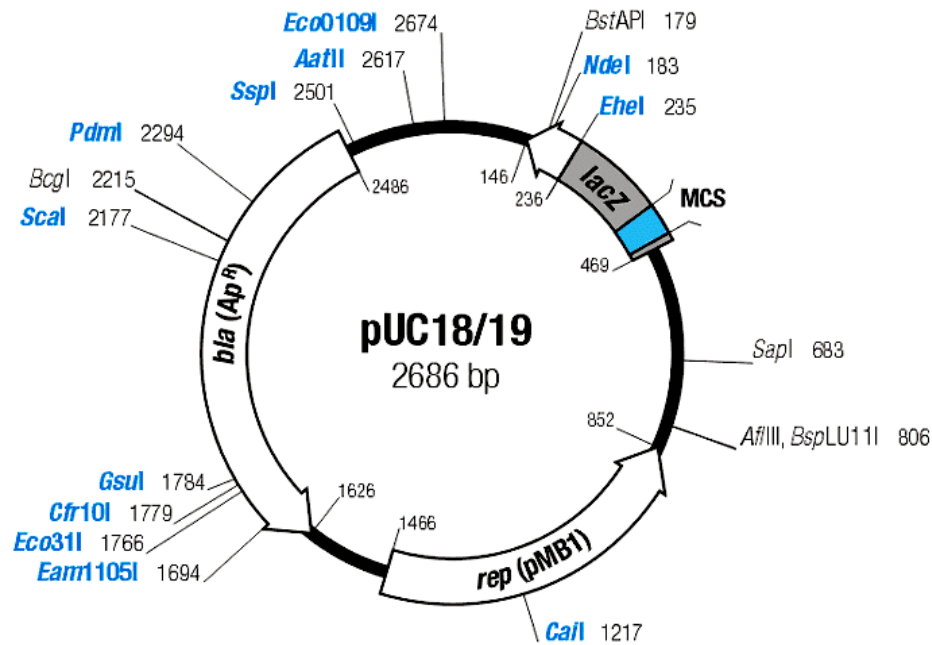
pBR322



(Ins) : site de *BamHI* (375), (Pla) : *bla*, (Frg) : copie + *tet*

Fermentas Life Sciences

pUC18



(Ins) : Multiple Cloning Site, **(Pla)** : bla, **(Frg)** : lacZ (colonies bleues en présence de X-gal sans insertion, blanches avec)

Fermentas Life Sciences

Multiple Cloning Site

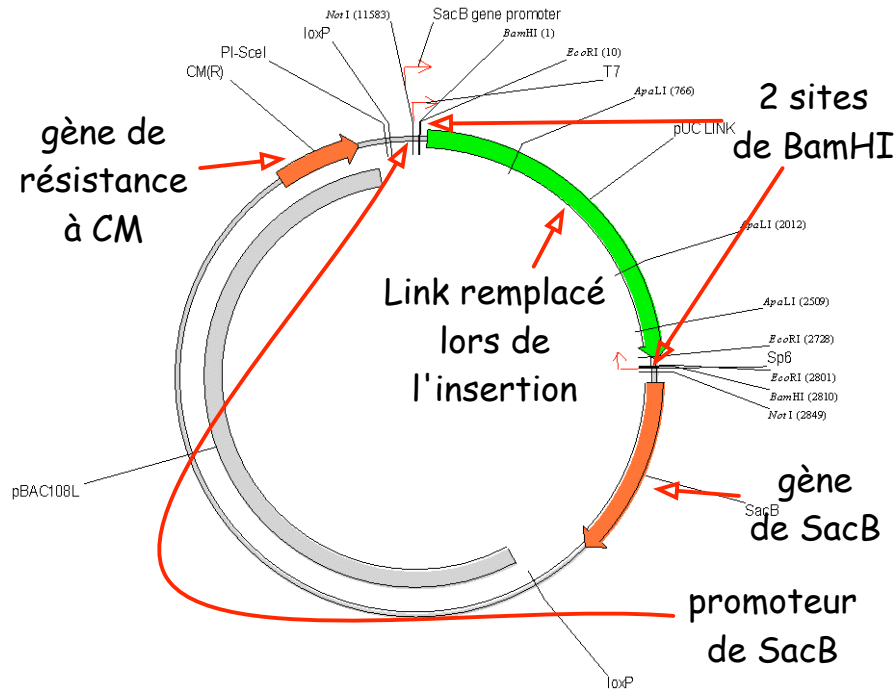


Fermentas Life Sciences

Vecteurs

Vecteur	Hôte	Taille d'insert ($\times 1000$ pb)
M13	<i>E. coli</i>	1–4
Plasmid	<i>E. coli</i>	1–5
Phage λ	<i>E. coli</i>	2–25
Cosmid	<i>E. coli</i>	35–45
BAC	<i>E. coli</i>	50–300
YAC	<i>S. cerevisiae</i>	100–2000

BAC

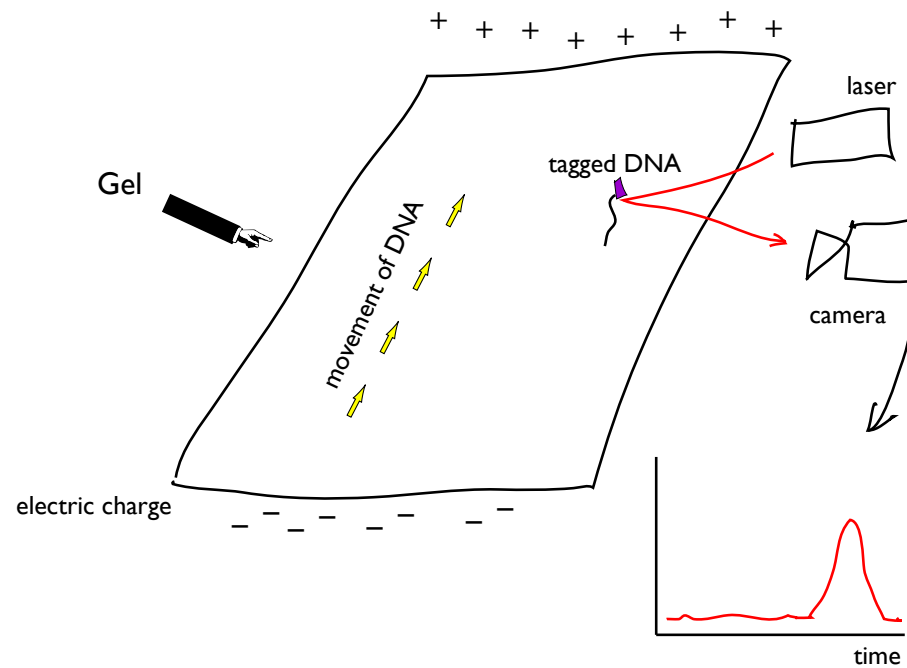


(Ins) : LINK remplacé (paires de sites de restriction), **(Pla)** : CM^R, **(Frg)** : SacB (produit une toxine)

CHORI BACPAC

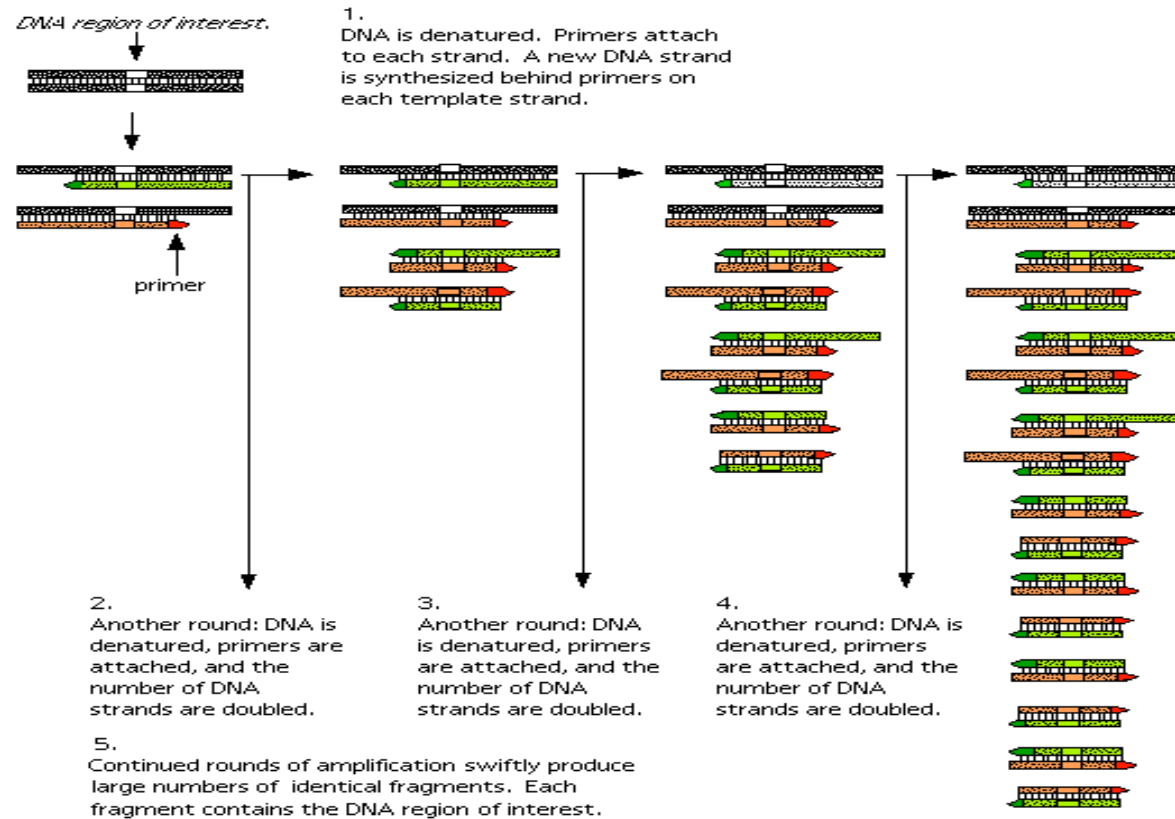
Electrophorèse

mesure la taille d'un fragment ADN



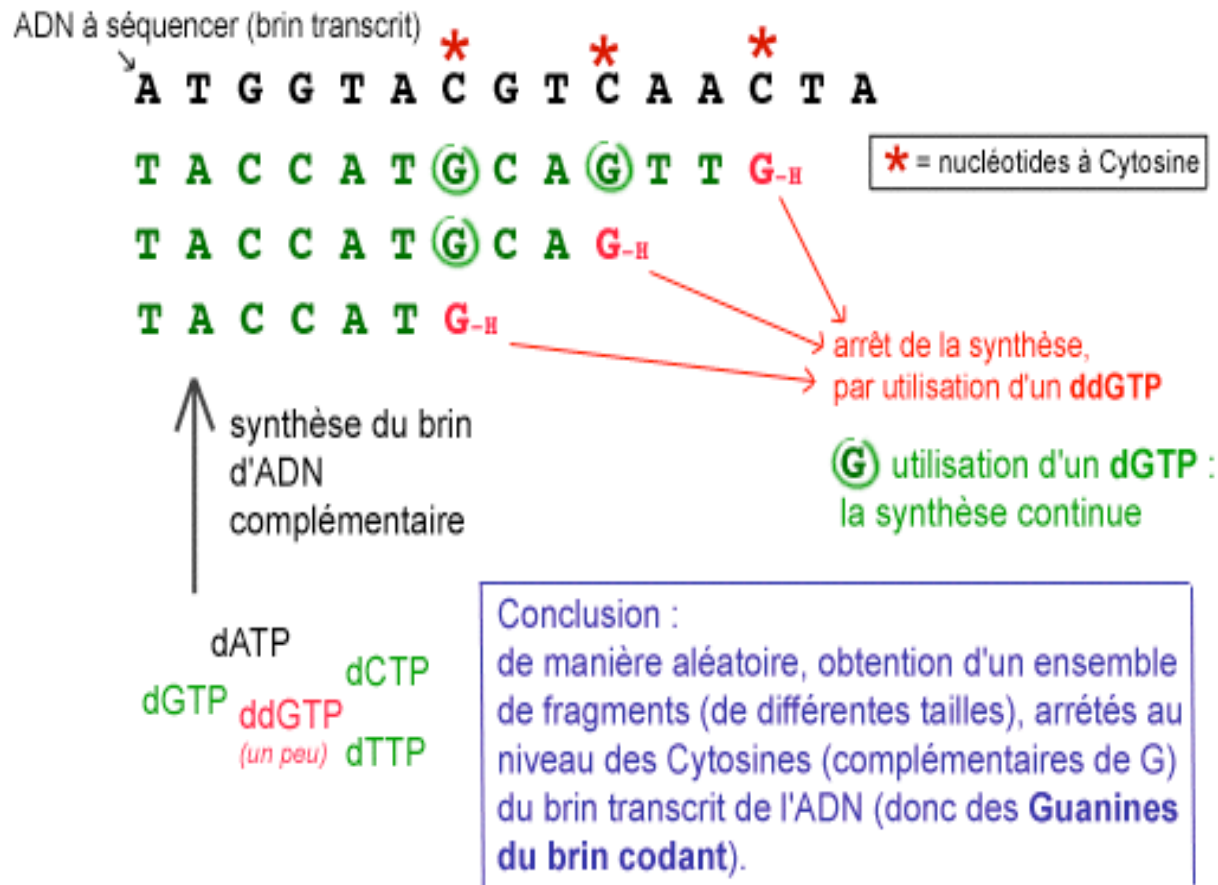
Polymerase Chain Reaction

POLYMERASE CHAIN REACTION



Access Excellence

Séquençage — méthode Sanger



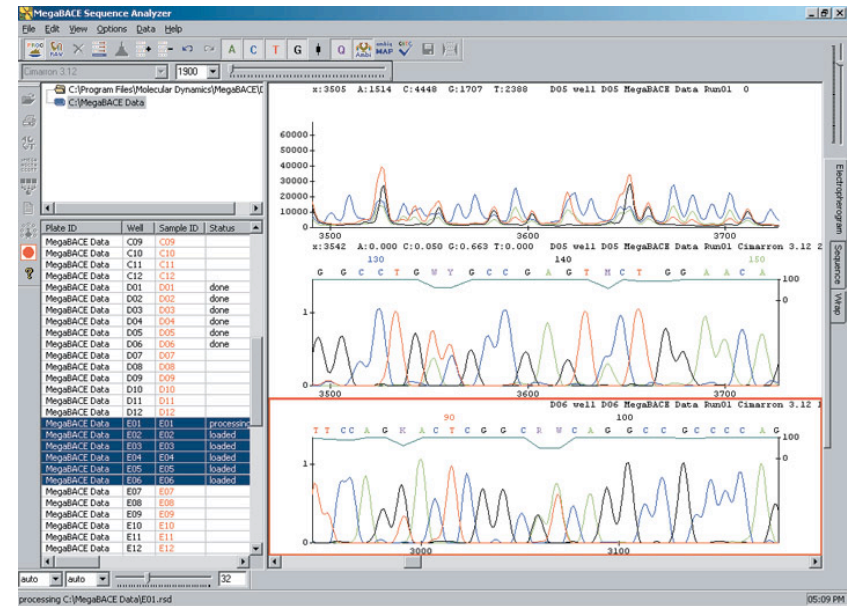
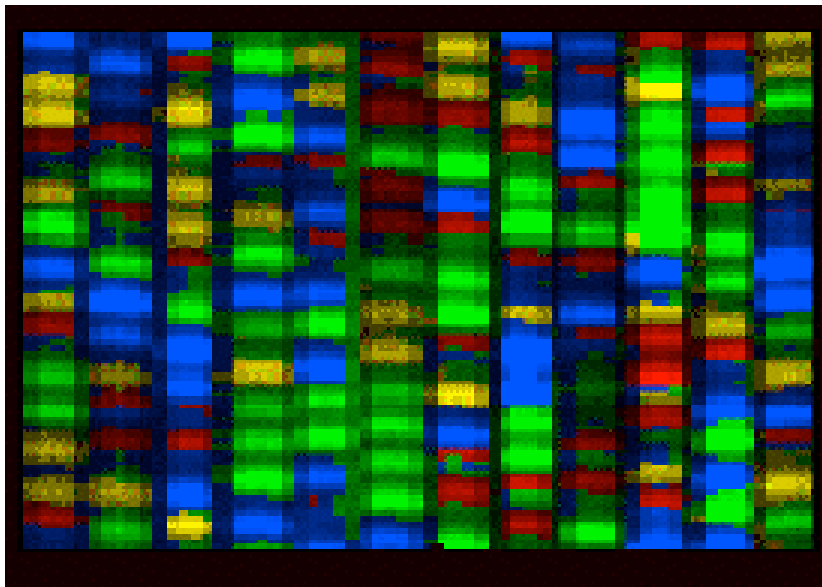
Lodish et al. *Molecular Cell Biology*, 4th ed., W. H. Freeman 1999 ; Delarue et Furelaud, Jussieu

Séquençage automatique

→ marqueurs fluorescents, multiplexage, capillaires, cycleurs thermiques

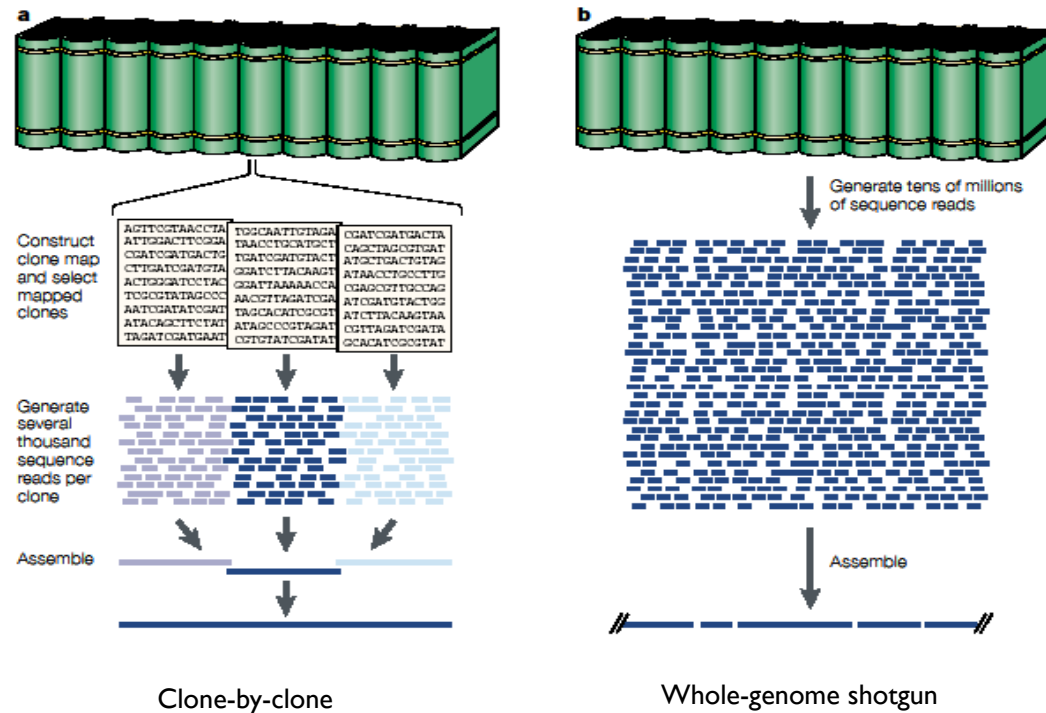
tailles jusqu'à 1000 pb (typiquement 600 pb)

Electrophorèse en séquençage



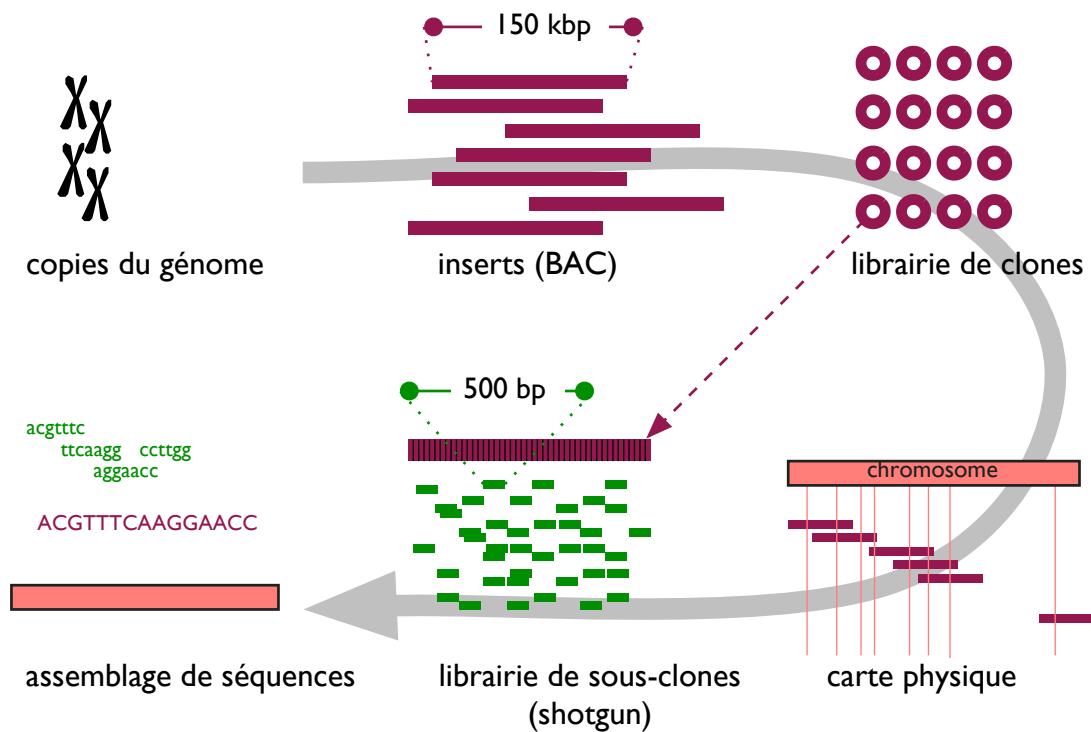
<http://www.megabace.com/>

Séquençage d'un génome



E. Green. Nature Reviews Genetics 2:573 (2001)

Approche hiérarchique



Séquençage d'un BAC

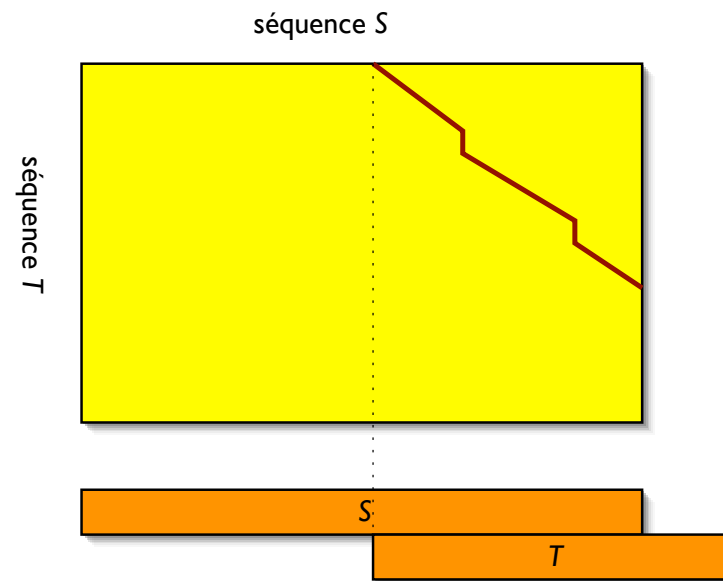
longueur de l'insert : cca. 150000 pb
(cca 50 pages dans le Livre)

fragments shotgun :
les joindre à partir des chevauchements

```
s1: AATGCC
s2:   GCCTTACAC
s3:     ACACTG
s4:       ACTGAAGG
s5:         GAAGGTTTA
-----
B : AATGCCTTACACTGAAGGTTTA
```

Alignement semiglobal

Trouver le meilleur chevauchement entre un suffixe de S et un préfixe de T .



TIGR assembler (1995)

une approche glouton utilisée pour assembler le génome de *H. influenzae*

1. analyse de k -mers dans les fragments : chevauchements potentiels entre fragments avec k -mers partagés (score déterminé par nombre de k -mers en commun)
2. identification de fragments avec régions répétées
3. initialisation de la séquence assemblée (contig) par un fragment
4. répéter : ajout du meilleur fragment à la séquence assemblée

En 2 : fragment avec trop de chevauchements potentiel=repeat

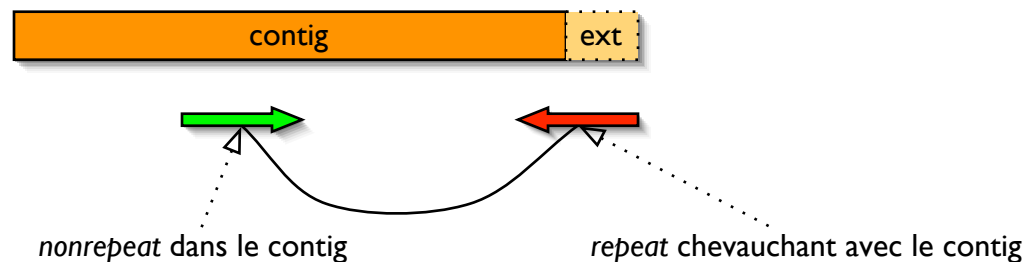
Sutton et al, *Genome Sci. Techn.* 1 : 9.

TIGR 2

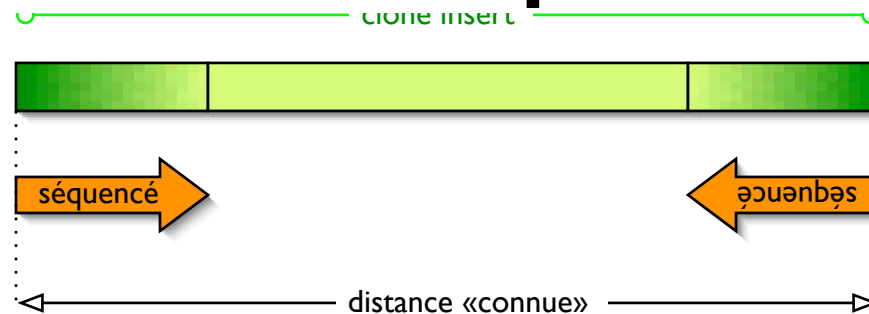
trouver meilleur fragment à ajouter : alignement local (Smith-Waterman),
pour tous les fragments avec chevauchements potentiels

contig finit s'il n'y a plus de fragments à ajouter : à la frontière d'une région
répétée ou à un vrai trou

extension dans la région répétée : utiliser des séquences shotgun ap-
pariées (*mate pairs*)



Mate pairs

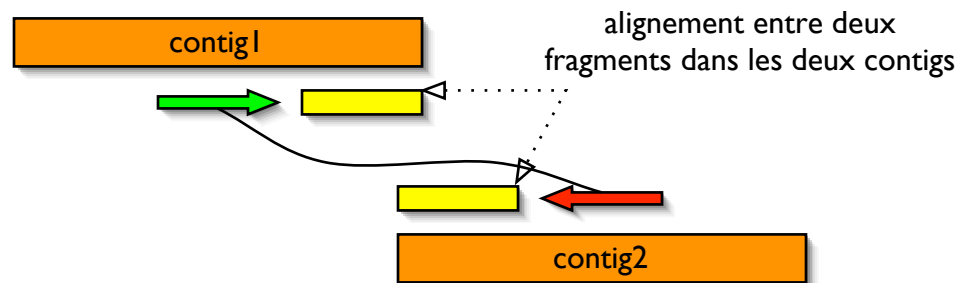


distances : 2k (M13), 10k (plasmid), 100k (BAC)

aident à orienter des contigs, à construire des ossatures (*scaffolds*), et à traverser des régions répétées

TIGR 3

joindre des contigs si évidence par chevauchements et mate pairs



Assemblage : overlap-layout-consensus

Overlap : déterminer les chevauchements parmi les séquences shotgun

Layout : déterminer l'ordre des séquences shotgun

Consensus : déterminer la séquence des contigs

Calcul de tous les chevauchements

$\mathcal{F} = \{f_1, \dots, f_n\}$ ensemble de séquences shotgun

trouver le meilleur chevauchement entre chaque paire : $O(n^2\ell^2)$ où ℓ est une borne supérieure sur la longueur des fragments.

Améliorer :

1. moins de paires comparés
2. trouver le meilleur chevauchement plus rapidement

Chevauchements

- PD pour alignement semi-globale

- recherche rapide

⇒ graphe de chevauchements (*overlap graph*)

Un exemple (CAP3)

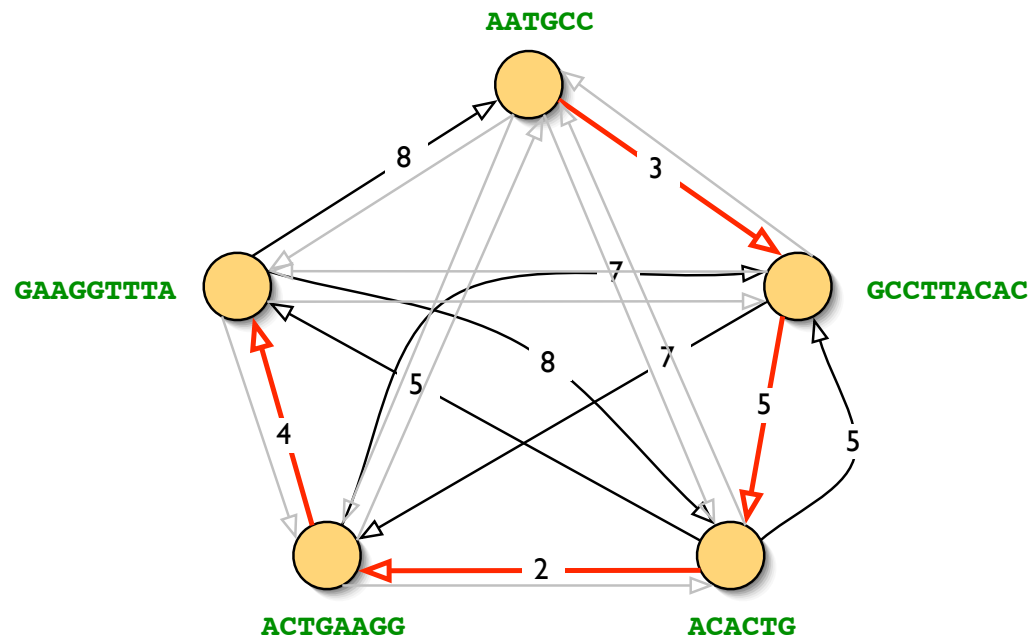
Séquence combinée de tous les fragments:



1. Tableau de hachage des k -mers de la séquence combinée.
2. Chaque fragment f ainsi que son complément sont comparés à la séquence combinée, pour trouver des HSPs (v. BLAST).
3. Les chevauchements potentiels de 2) sont évalués par alignement dans une bande.

Huang et Madan, *Genome Res.* 9 : 868.

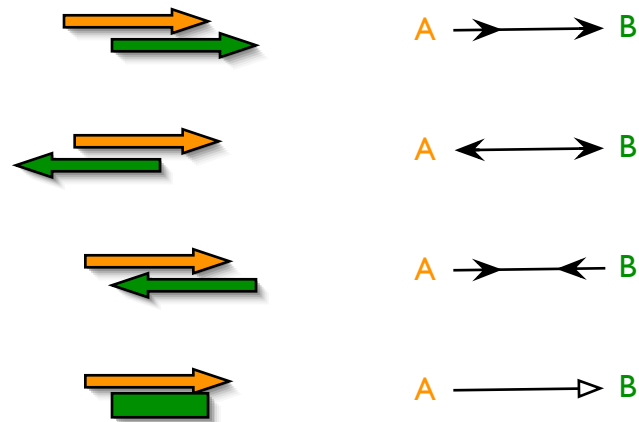
Graphe de chevauchements



[problème d'orientation ignoré]

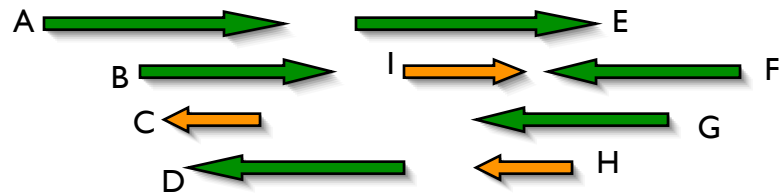
Layout

Catégories de chevauchements (orientation inconnue des fragments)



Myers, J. *Comput. Biol.* 2 : 275.

Exemple



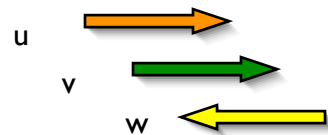
compter les flèches arrivant à un vertex : layout=chemin avec arêtes de chevauchements propres+ forêts avec arêtes de contention

Myers, J. *Comput. Biol.* 2 : 275.

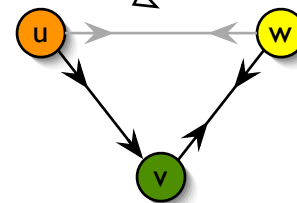
Simplification du graphe

1. enlever les arêtes de contention

2. enlever des arêtes «transitifs» : si $u \Rightarrow v$, $v \Rightarrow w$ et $u \Rightarrow w$ sont des arêtes compatibles (orientation+taille de chevauchements), alors enlever $u \Rightarrow w$



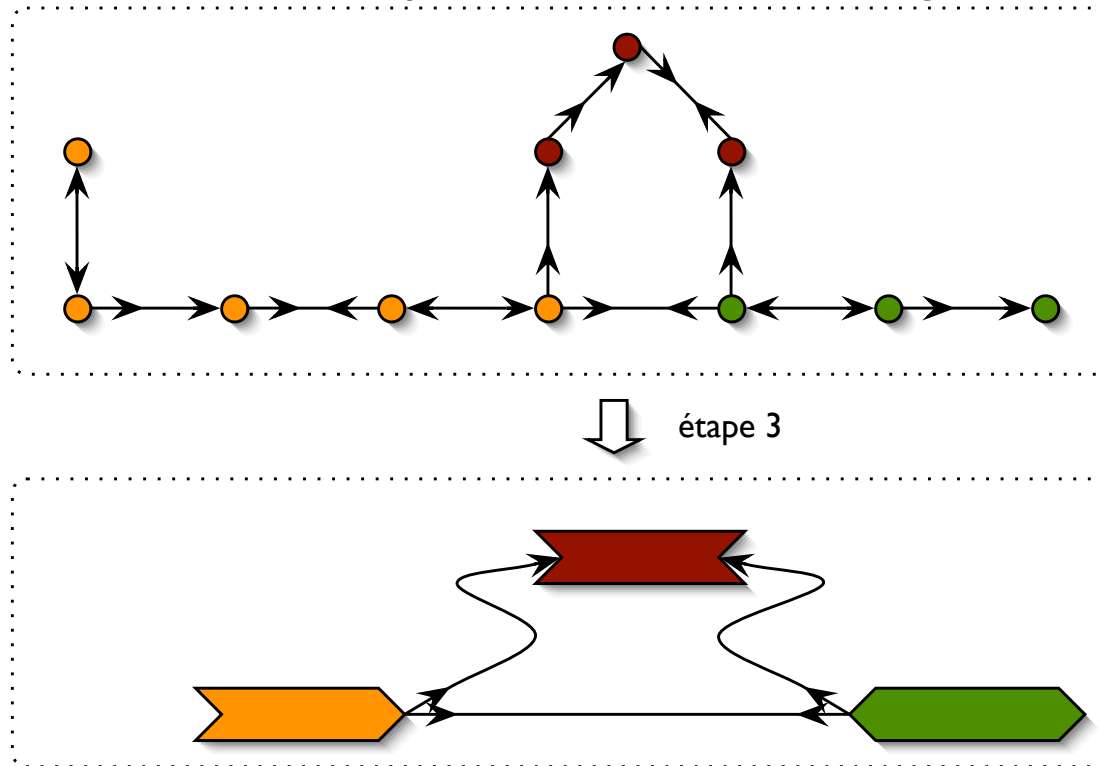
arête impliqué par
les tailles des
chevauchements



Myers, J. *Comput. Biol.* 2 : 275.

Simplification du graphe 2

3. collapser des chemins : «super-vertices» ou contigs



Myers, J. *Comput. Biol.* 2 : 275.

Séquence de consensus

Le layout donne la position approximative de chaque fragment. Trouver la séquence de consensus : alignement multiple

Profile : enregistrer la fréquence de symbols dans l'alignement multiple

Joindre les séquences consécutives au contig dans l'ordre spécifié dans la phase *layout*, maintenir un profile dans le contig

Problème : alignement d'une séquence à un profile (dans une bande autour de la position approximative)

⇒ contigs

Cartes

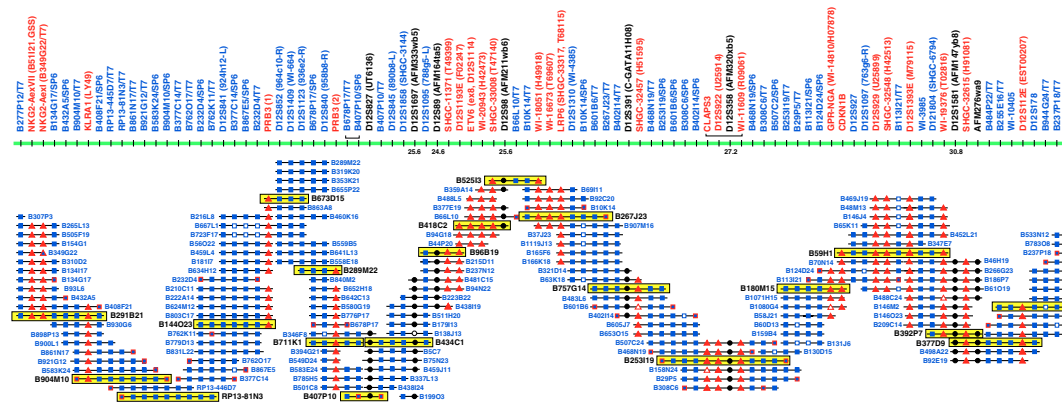
cartes physiques et génomiques

cartes physiques : par digestion ou hybridation

empreintes, sondes uniques (STS) ou non-uniques

utilité : p.e sélection de clones pour séquençage complet

Exemple



Resolution

- ordre de clones
- coordonnées physiques

Hybridation

Ensemble de sondes : $P_1, P_2, \dots, P_m$ -[non-unique]

Ensemble de clones : $C_1, C_2, \dots, C_n$

Données : $n \times m$ matrice H ou $H[i, j] = 1$ ssi clone C_i s'hybride avec sonde P_j .

But : assigner des coordonnées physiques aux C_i et P_j qui explique H

Séquençage par hybridation (SBH)

Idée : $\mathcal{C}$ ensemble de sondes (p.e. tous les k -mers) arrangé sur une puce

On teste la présence de chaque $c \in \mathcal{C}$ dans une molécule d'ADN (séquence u) par hybridation

spectrum : ensemble de k -mers dans u

$$\mathcal{S}_k(u) = \{u[i, \dots, i + k - 1] : i = 1, \dots, |u| - k + 1\}.$$

Problème : reconstruction de u à partir de $\mathcal{S}_k(u)$.

Reconstruction à partir du spectrum

Approche 1 : graphe de chevauchements où chaque k -mer de u est un vertex et des arêtes représentent des chevauchements de taille $(k - 1)$. On cherche un chemin hamiltonien.

Approche 2 : g.d.c où chaque k -mer de u est une arête entre son préfixe et suffixe de taille $(k - 1)$ [quand le spectrum inclut tous les k -mers, c'est un **graphe de Bruijn**] On cherche un chemin eulérien.

Chemin eulérien

Déf Un chemin/cycle eulérien d'un graphe visite chaque arête exactement une fois.

Thm. Il existe un cycle eulérien dans un graphe connecté orienté ssi $\text{degré}_{arr}(u) = \text{degré}_{sort}(u)$ dans chaque vertice u .

Il existe un algorithme qui trouve un cycle [ou chemin] eulérien (ou annonce qu'il n'y en a pas) en temps linéaire.

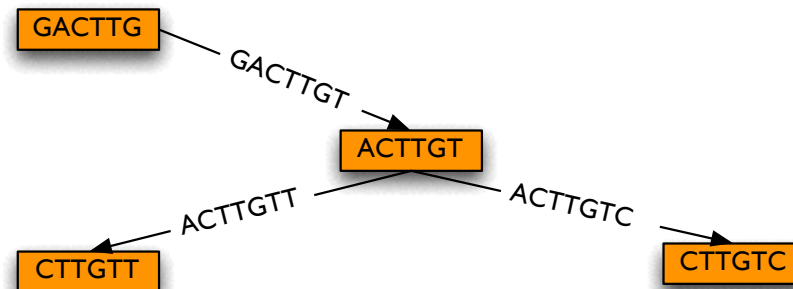
Séquençage shotgun par SBH

Il est désirable de calculer le spectrum avec un grand k (longueur d'une séquence qui peut être reconstruite est $\approx 2^k$) mais il n'est pas [encore] pratique d'utiliser des puces avec $k = 20$ p.e.

Idée : pourquoi ne pas générer le spectrum à partir de séquences shotgun ?

1. ensemble $\mathcal{F} = \{s_1, s_2, \dots, s_n\}$ de séquences shotgun
2. ensemble de k -mers (p.e. $k = 24$) $\mathcal{F}_k = \bigcup_{i=1}^n \{s_i[j..j+k-1] : j = 1, \dots, |s_i| - k + 1\}$
3. assembler la séquence à partir de $\mathcal{F}_k$ comme en SBH (chemins eulériens)

Idury-Waterman



1. Réduction du graphe par transformations pareilles à ce qu'on a vu pour layout (Myers 1995).
2. Chemins eulériens modifiés : la même arête peut être visitée plus qu'une fois (p.e. région répétée); quelques arêtes ne sont pas visitées de tout (erreurs de séquençage)

Idury et Waterman, *J. Comput. Biol.* 2 : 291.

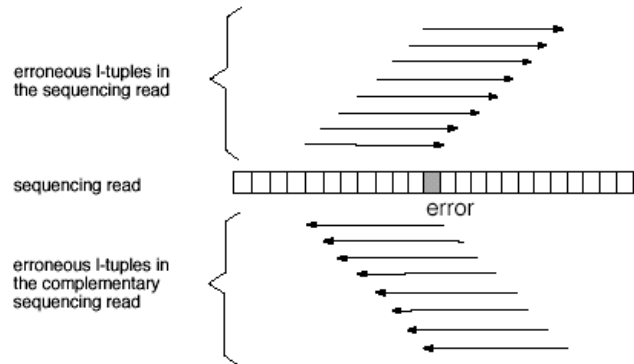
Euler

Une autre approche inspirée par SBH : Euler

- correction d'erreurs : identification de k -mers rares
- augmentation de graphe de Bruijn pour transformer le problème du **super-chemin eulérien** en celui du chemin eulérien

Pevzner, Tang, et Waterman, *RECOMB 2001* 256.

Euler : erreurs



problème avec Idury-Waterman : trop de sommets dans le graphe créés par des erreurs de séquençage

solution de Euler : **orphelin** est un rare k -mer u dans le [multi-]spectrum $\mathcal{S}$ t.q. il existe exactement un $v \in \mathcal{S}$ avec $\|u - v\| = 1$ et v est fréquent — remplacer u par v .

Euler : super-chemins

Départ : ensemble de k -mers avec info sur leurs occurrences (lecture shotgun + position dans la lecture)

construire le graphe de Bruijn à arêtes multiples (une arête pour chaque occurrence d'un k -mer)

chemins initiaux : définis par les séquences shotgun

