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

Enzymes de restriction

coupent l'ADN à un site spécifique

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

Enzymes de restriction 2

souvent le site reconnu par un ER est un palindrôme : le complément inverse a la même séquence

```
   ----->
5' GAATTC
   CTTAAG 3'
   <-----
```

sticky ends

$\Rightarrow$ à l'aide d'une enzyme de ligase on peut joindre des amorces de ADN d'origines différentes coupées par le même ER

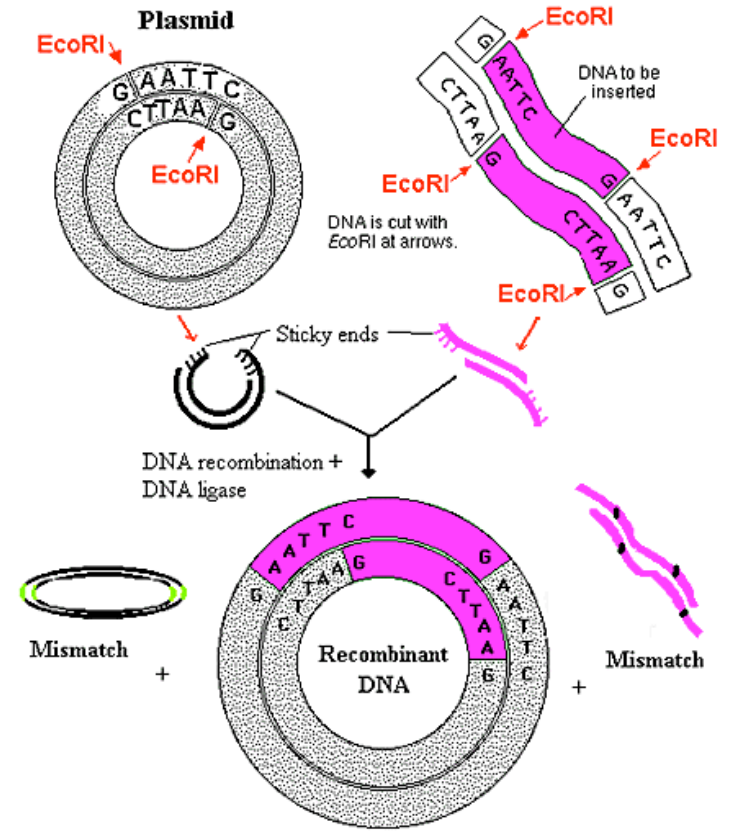
Clonage

Idée :

1. amorces de ADN après coupure par un ER
2. insertion dans vecteur dans une cellule hôte (bactérienne ou virale)
3. culture et purification

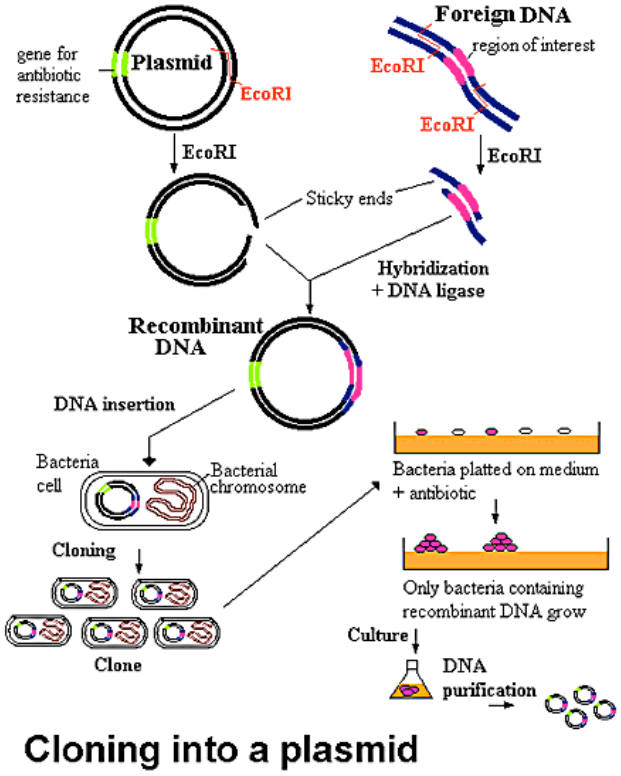
vecteurs : plasmids, phages et cosmids ; BAC (*Bacterial Artificial Chromosome*) et PAC

Clonage : insertion d'une amorce



Inserting a DNA Sample into a Plasmid

Clonage : culture



Cloning into a plasmid

[animation]

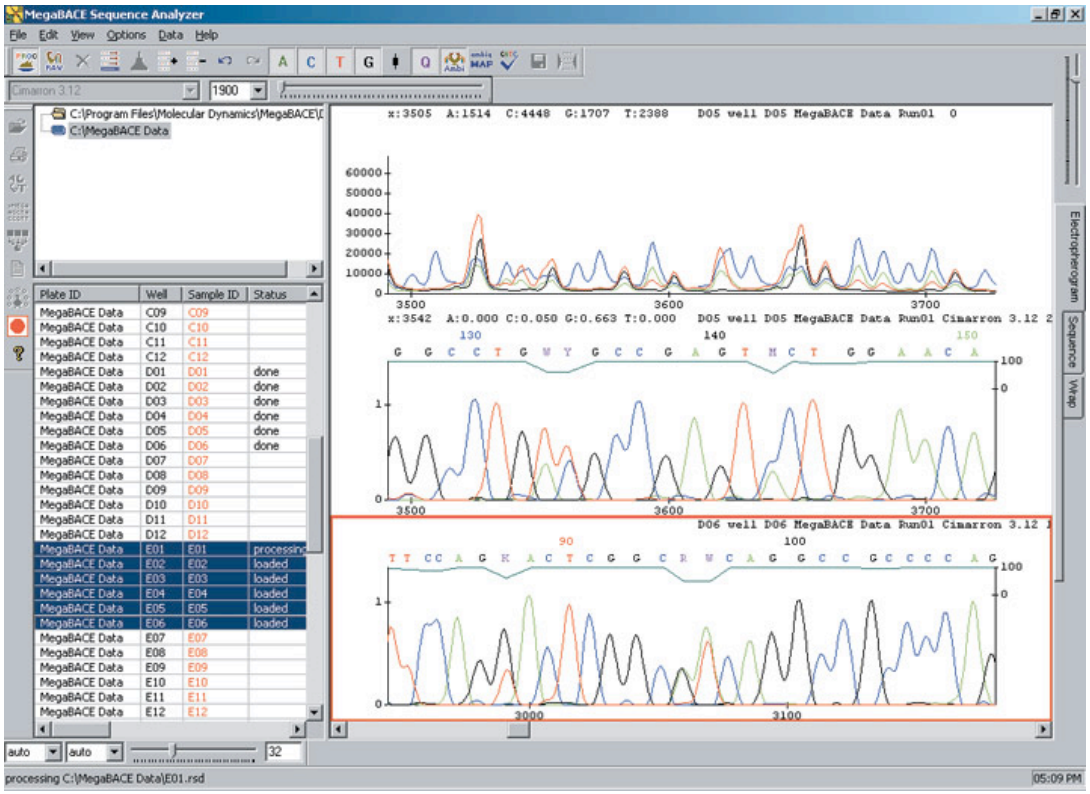
Séquençage d’un fragment

séquençage par la méthode de Sanger

[animation]

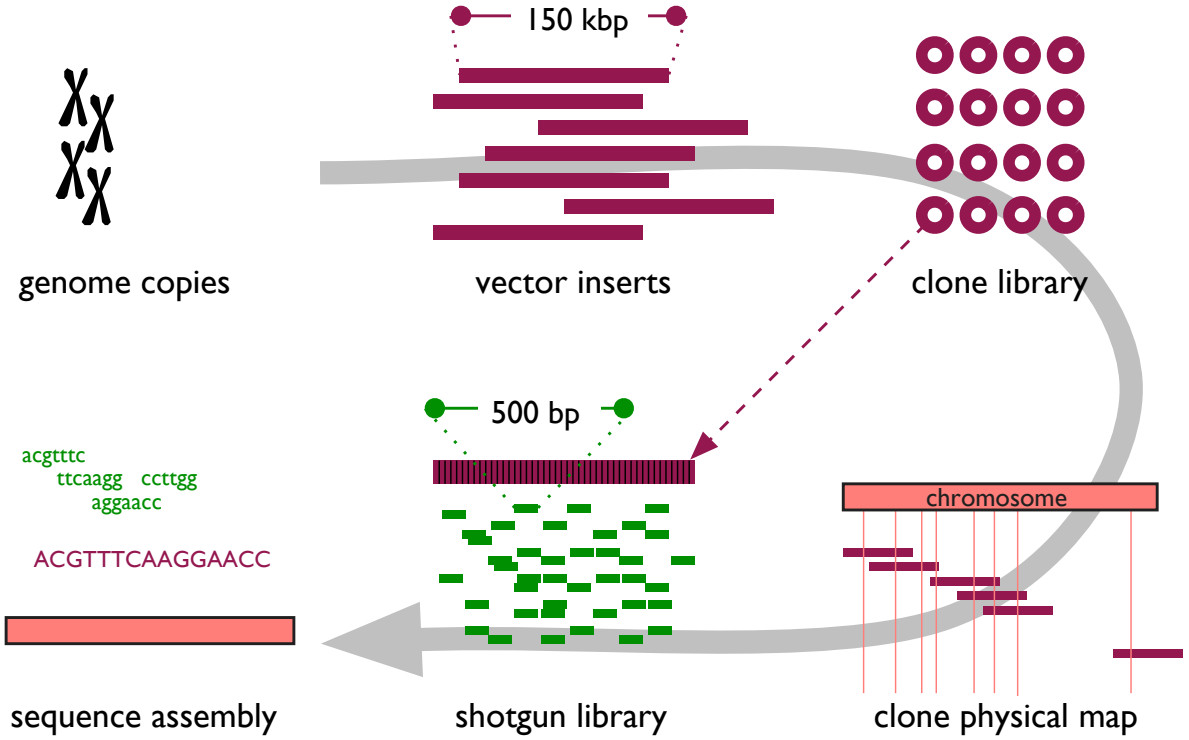
produit des séquences de longueur entre 50-600

Séquençage d'un fragment 2



http ://www.megabace.com/

Séquençage d'un génome



Séquençage d'un BAC

(méthode de shotgun : fragments aléatoires séquencés)

[exemple]

approche :

1. détection de chevauchements
2. layout
3. séquence de consensus

Chevauchements

- PD pour alignement semi-globale

- recherche rapide

$\Rightarrow$ graphe de chevauchements (*overlap graph*)

Chemins dans le graphe de chevauchements

chemin = super-séquence

pondération des arêtes par la taille du chevauchement

chemin Hamiltonien : super-séquence

la super-séquence plus courte : NP-difficile

problèmes : erreurs de lecture, répétitions, orientation inconnue. . .

Séquence de consensus

- majorité

- pondération par qualité des fragments (Phrap)

$\Rightarrow$ contigs

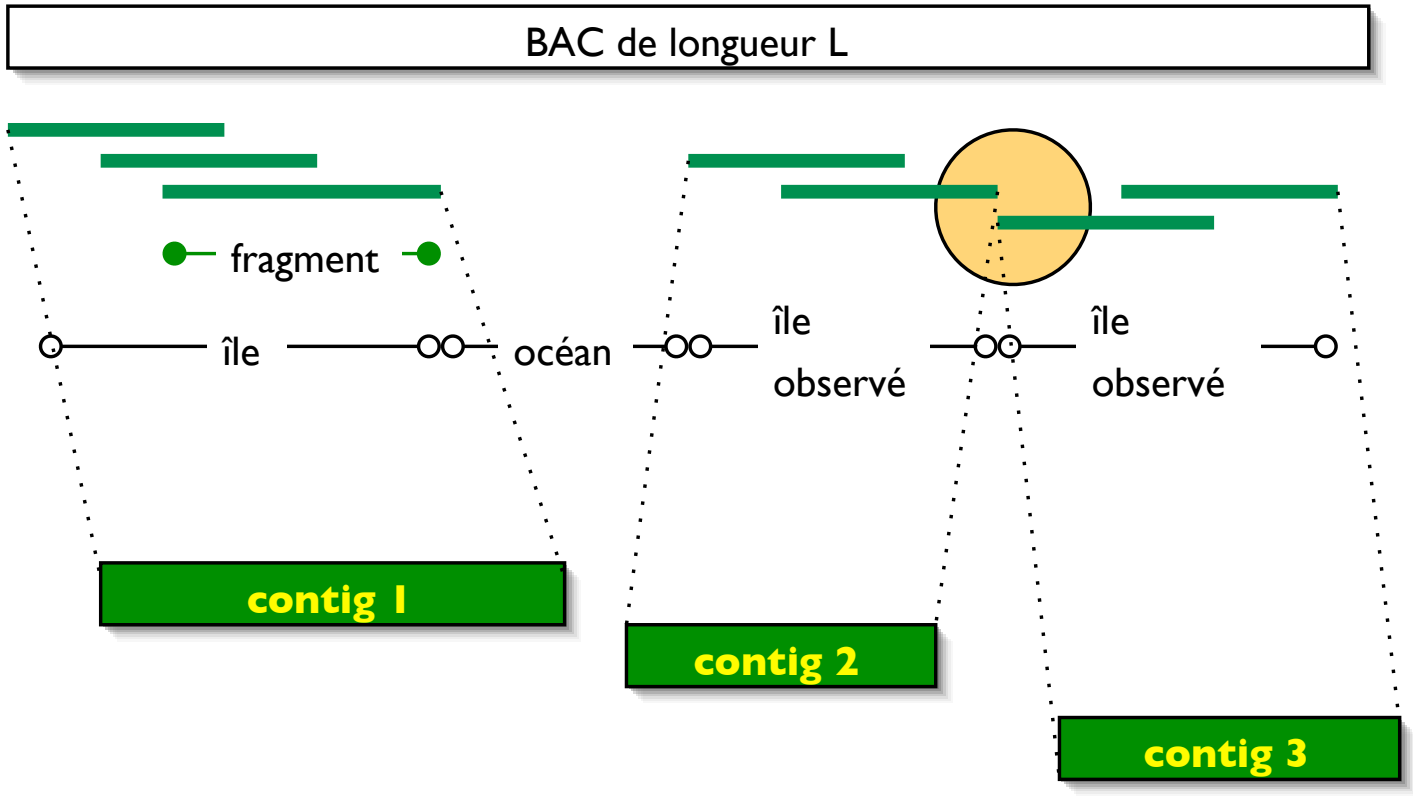
Modèle statistique pour shotgun

nombre de fragments : n
longueur d'un fragment : ℓ
longueur du BAC : L

couverture (*coverage*) : $c = n\ell/L$

chevauchement minimal : $\theta\ell$, $0 < \theta < 1$

Modèle - terminologie



Modèle - cont.

Thm. La probabilité qu'une position du BAC est couverte par au moins un fragment est cca. $(1 - e^{-c})$.

Preuve. Probabilité qu'un fragment fixé couvre la position : ℓ/L

Probabilité qu'aucun fragment ne la couvre pas : $\left(1 - \frac{\ell}{L}\right)^n$.

Approximation : $(1 - a/x)^x \approx e^{-a}$.

Modèle - cont.

Thm. Le nombre des océans est cca. $ne^{-c(1-\theta)} = \frac{\ell}{L}ce^{-c(1-\theta)}$.

Preuve. Probabilité qu'un fragment fixé est le dernier fragment d'un île observé :

$$p = \left(1 - \frac{(1-\theta)\ell}{L}\right)^{n-1}.$$

+approximation comme avant

Espérance du nombre des océans = np.

Modèle - cont.

Position du fragment définie par la position de son côté droit : variables aléatoires $X_1, X_2, \ldots, X_n$.

Fixons un fragment (X_1). Quelle est la position Y_1 du premier fragment après X_1 ? Probabilité que $Y_1 > X_1 + h\ell$ est

$$J(h) = \left(1 - \frac{h\ell}{L}\right)^{n-1} \approx \left(1 - \frac{ch}{n}\right)^{n} \approx e^{-ch}.$$

Modèle - cont.

Thm. Le nombre de fragments dans un île est cca. $e^{c(1-\theta)}$.

Preuve. Soit M le nombre des fragments dans l'île.
Considérons le premier fragment de l'île. Probabilité que c'est un île singulaire ($M = 1$) : $p_1 = J(1-\theta)$.
Probabilité que $M = k$: $p_k = \Big(1 - J(1-\theta)\Big)^{k-1} J(1-\theta)$ — distribution géometrique. Espérance de M est $1/J(1-\theta)$.

Cartes physiques

1. hybridation : STS (*sequence-tagged site*)

2. empreintes (*fingerprints*)

utilité : filtrage de clônes (detection de chevauchements entre les clônes)
$\Rightarrow$ selection des clônes pour séquençage complet (*sequence-ready map*)

Cartes STS

retrouver l'ordre des marqueurs et des clônes

$\Rightarrow$ problème des '1's consécutifs

en pratique : voyageur commerçant (erreurs de hybridation)

Hybridation : pooling

10	15	20	25	5
14	19	24	4	9
18	23	3	8	13
22	2	7	12	17
1	6	11	16	21

8	11	19	22	5
15	18	21	4	7
17	25	3	6	14
24	2	10	13	16
1	9	12	20	23

Cartes d’empreintes

Problème de *Double digest*.
Digestion part deux enzymes de restriction (A et B) : taille de fragments pour A, taille de fragments pour B, taille de fragments pour A et B.

Modèle statistique : même que pour séquençage shotgun

Scaffolds

Problème de répétitions avec shotgun simple

Une solution : *double barrel shotgun*
(fragments séquencés des deux extremités)

Cartes d’empreintes