Cartographie génétique & physique

INRA, INSERM

Septembre 2012

Plan

Les cartes: s'orienter dans le génome

Types

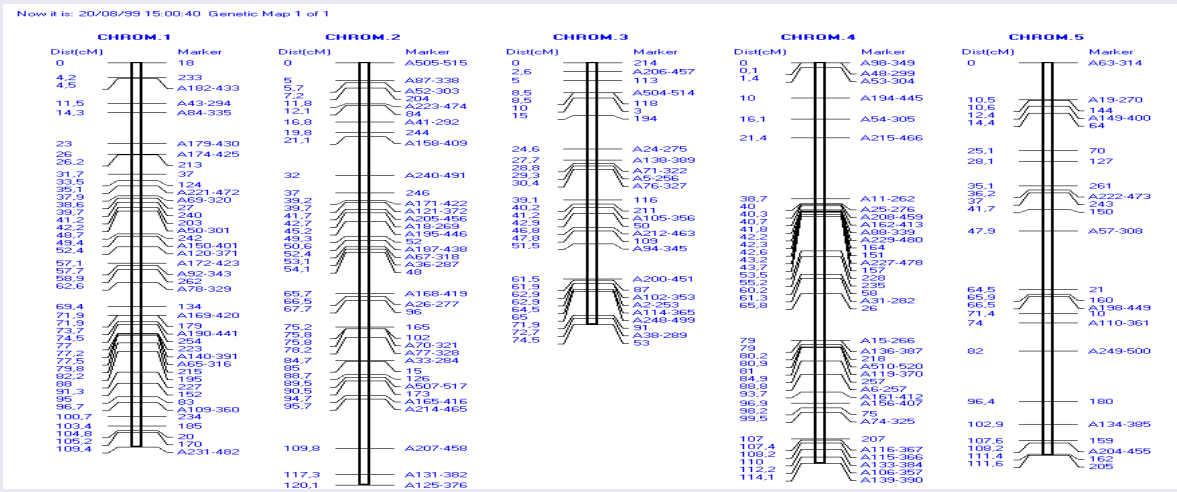
- **Cartes physiques** : distance réelle (Kb, Mb), à partir de fragments d'ADN. Résolution habituellement élevée.
- **Cartes d'hybrides irradiés** : Distance "statistique" liée à la cassure par irradiation, résolution intermédiaire.
- **Cartes génétiques** : s'appuie sur la recombinaison durant la méiose. Distance "statistique" (pas physique, pas neutre).

Carte génétique/hybrides irradiés

Représentation d'un génome positionnant un ensemble de repères (marqueurs) dont on connaît les positions sur des groupes de liaison (chromosomes idéalement).

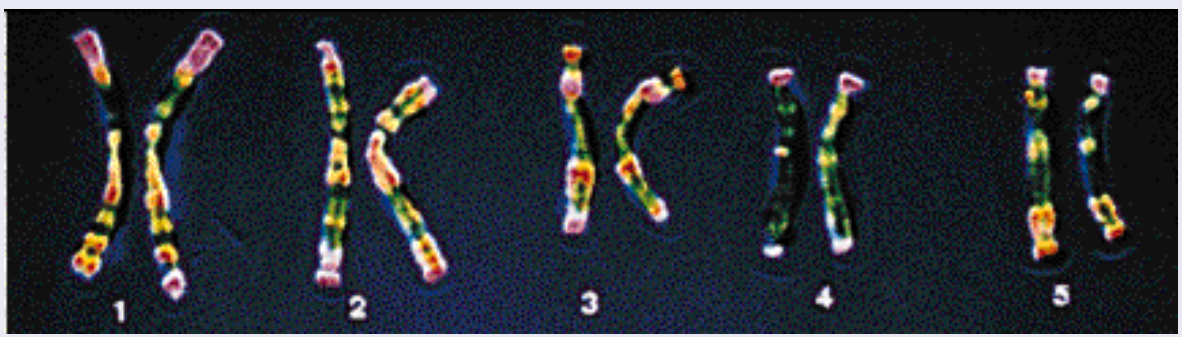
Exemple

Carte génétique



Groupes de liaison génétique

Génome



Chromosomes

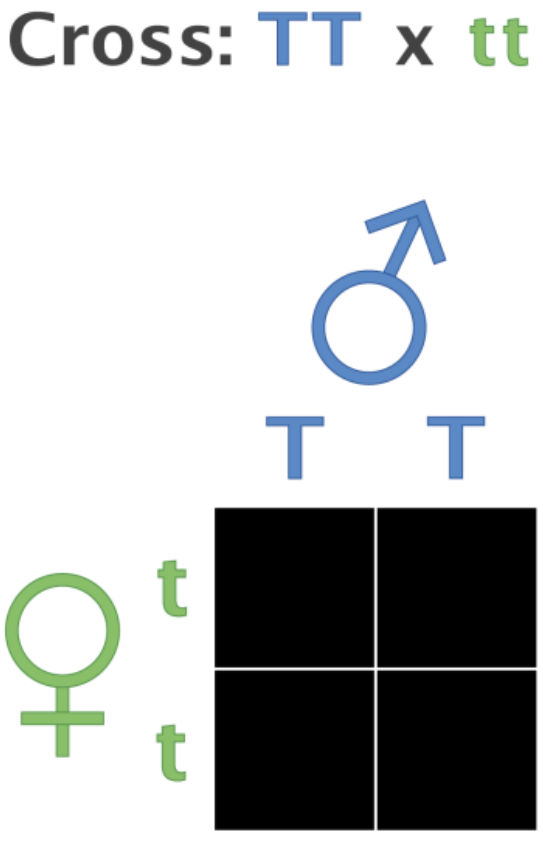
Pourquoi

- Identifier les régions du génome influençant un caractère d'intérêt (maladie ou caractère quantitatif plus complexe)
- Positionner et identifier un gène (clonage positionnel)
- Comparer les génomes (étude de la synténie, évolution, transfert d'information)
- Faciliter la construction de cartes physiques, assemblage
- Étudier la méiose

Les bases: lois de Mendel (modernes, diploïdes)

Loi de ségrégation

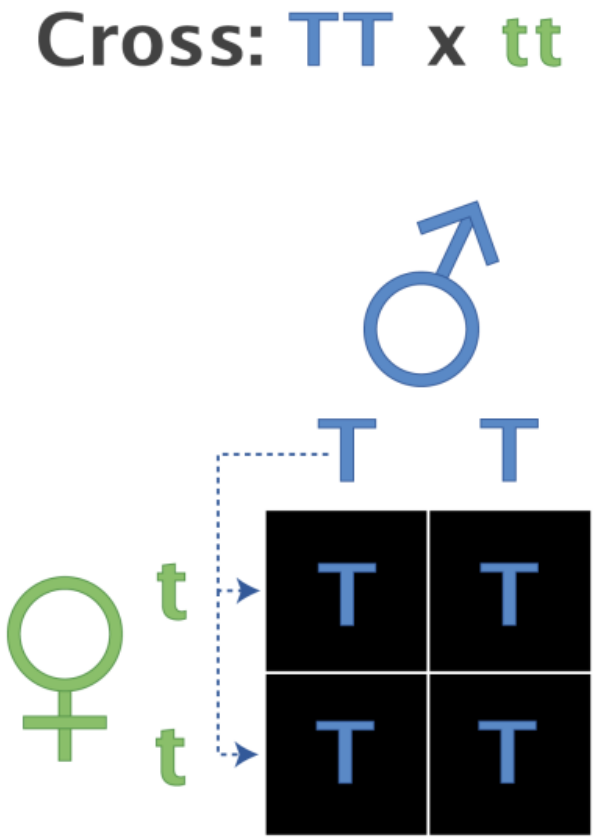
Un génome contient un ensemble de paires de gènes. Les paires ségrègent (se séparent) dans les gamètes, la moitié des gamètes portant un gène, l'autre moitié portant l'autre gène. Taille de plante (allèles *Tt*).



Les bases: lois de Mendel (modernes, diploïdes)

Loi de ségrégation

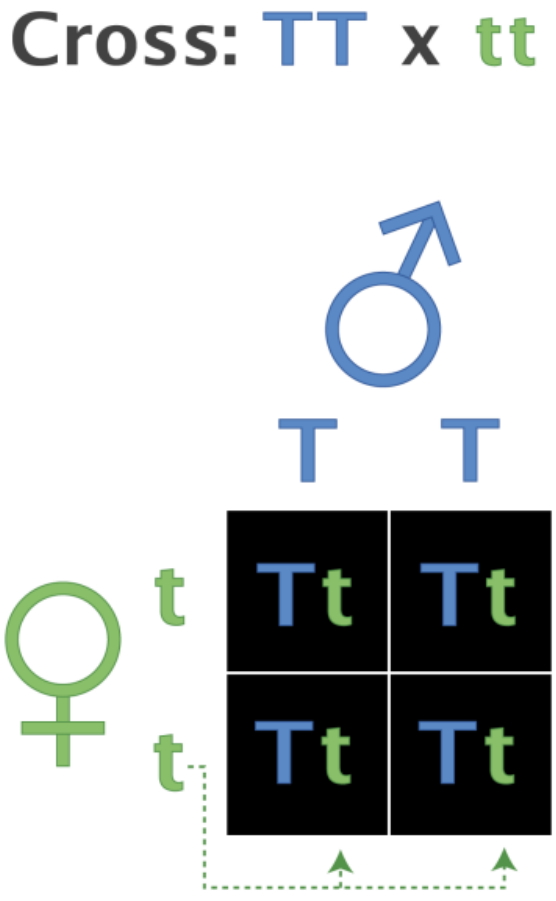
Un génome contient un ensemble de paires de gènes. Les paires ségrègent (se séparent) dans les gamètes, la moitié des gamètes portant un gène, l'autre moitié portant l'autre gène. Taille de plante (allèles *Tt*).



Les bases: lois de Mendel (modernes, diploïdes)

Loi de ségrégation

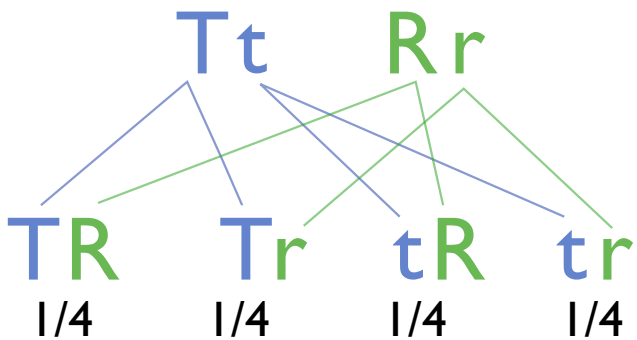
Un génome contient un ensemble de paires de gènes. Les paires ségrègent (se séparent) dans les gamètes, la moitié des gamètes portant un gène, l'autre moitié portant l'autre gène. Taille de plante (allèles *Tt*).



Les bases: lois de Mendel (modernisées, diploïdes)

Loi de ségrégation indépendante

L'assortiment de plusieurs gènes dans une cellule sexuelle se fait de façon indépendante entre les différents gènes. Taille *Tt* et forme *Rr* (ridé).

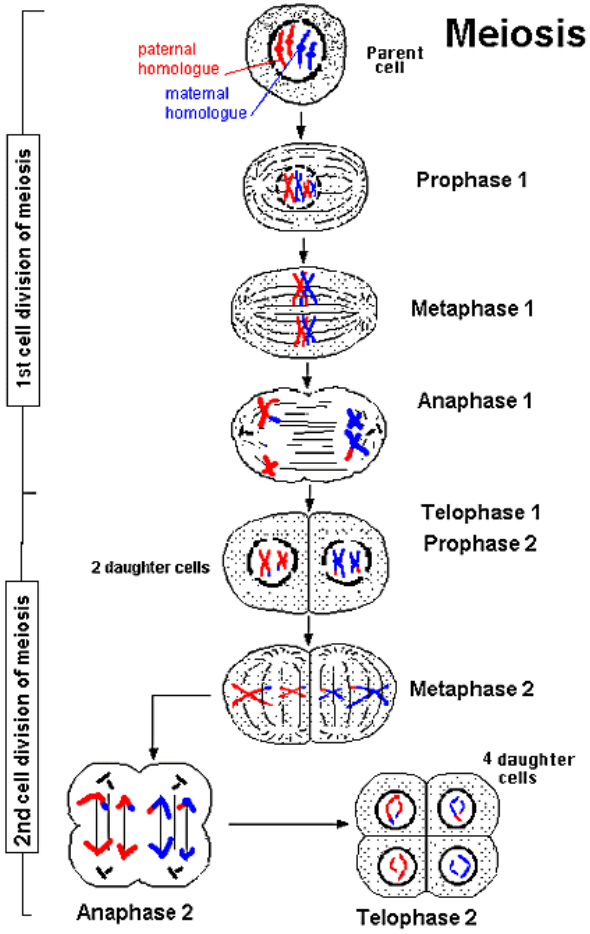


Le principe historique de la cartographie

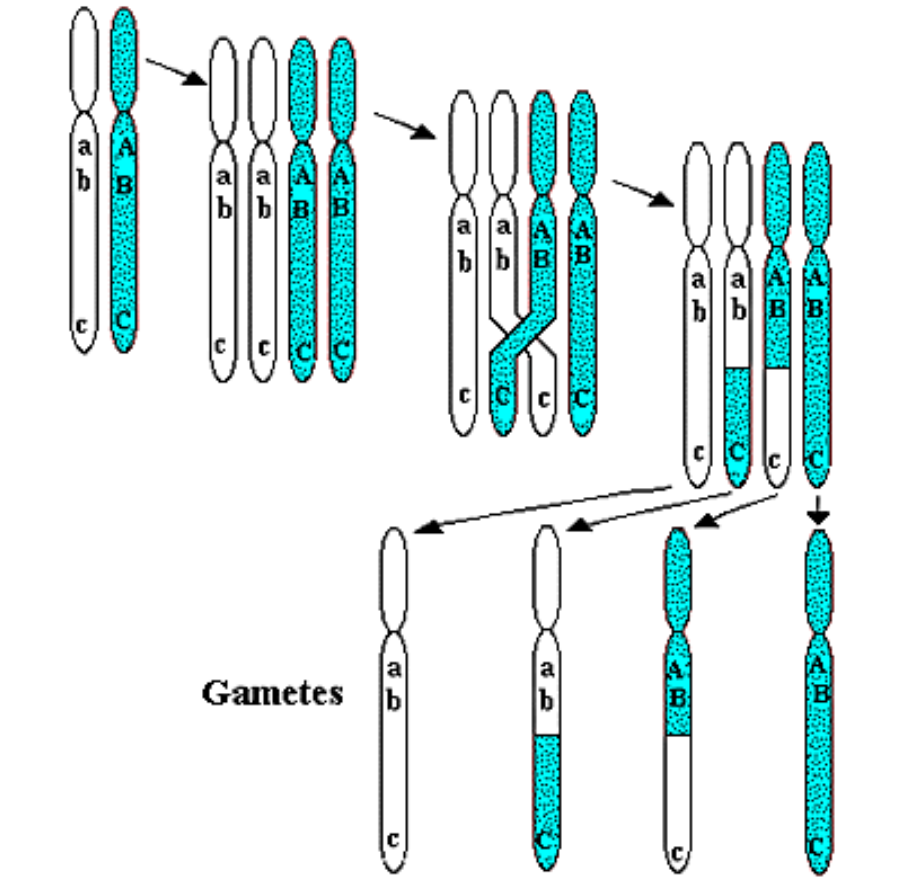
Liaison génétique (Bateson 1905)

Pour certaines paires de gènes, la fréquence des combinaisons parentales dans les gamètes est supérieure à ce que l'on attend. On parle de **liaison génétique**.

Expliqué par Morgan (1911) par l'appartenance à un même chromosome et un éventuel chiasma durant la méiose (crossing-over).



Un modèle de la méiose



Crossing-over and recombination during meiosis

50% de recombinants au plus.

Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome

Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles *Aa*).

Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles *Aa*).
- **Homozygote** : paire d'allèles identiques ($\frac{A}{A}$ ou $\frac{a}{a}$), sinon hétérozygote ($\frac{A}{a}$).

Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles *Aa*).
- **Homozygote** : paire d'allèles identiques ($\frac{A}{A}$ ou $\frac{a}{a}$), sinon hétérozygote ($\frac{A}{a}$).
- **Haplotype** : séquence des allèles portées par chacun des chromosomes ($\frac{Ab}{aB}$ par exemple).

Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles *Aa*).
- **Homozygote** : paire d'allèles identiques ($\frac{A}{A}$ ou $\frac{a}{a}$), sinon hétérozygote ($\frac{A}{a}$).
- **Haplotype** : séquence des allèles portées par chacun des chromosomes ($\frac{Ab}{aB}$ par exemple).
- **Génotype** : séquence des paires d'allèles (non ordonnées) portée par les chromosomes homologues ($\frac{A}{a}\ \frac{B}{b}$ par exemple).

Bases

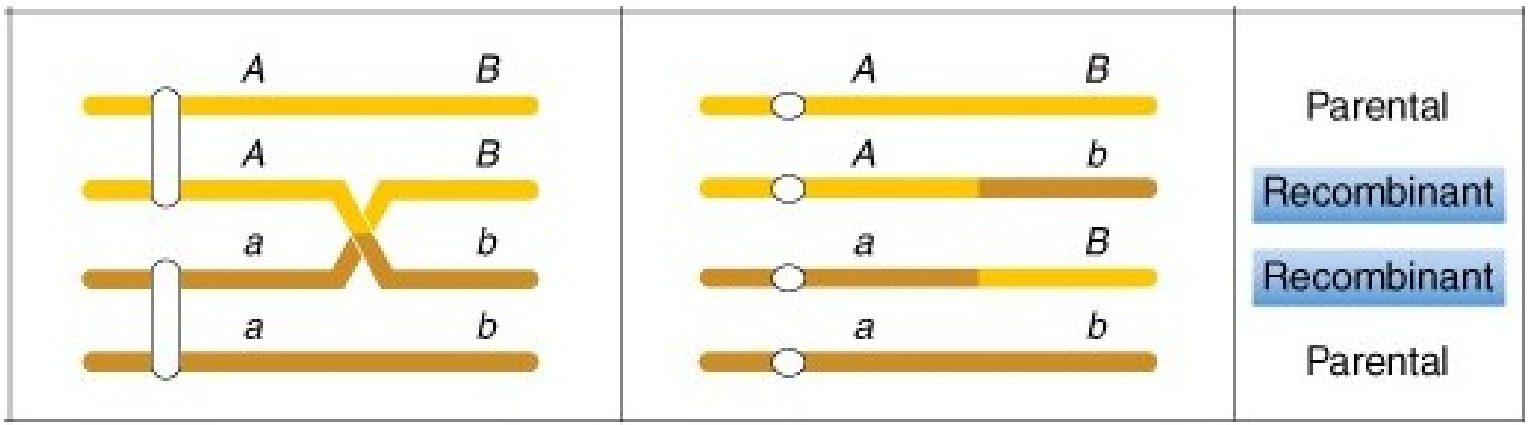
- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles *Aa*).
- **Homozygote** : paire d'allèles identiques ($\frac{A}{A}$ ou $\frac{a}{a}$), sinon hétérozygote ($\frac{A}{a}$).
- **Haplotype** : séquence des allèles portées par chacun des chromosomes ($\frac{Ab}{aB}$ par exemple).
- **Génotype** : séquence des paires d'allèles (non ordonnées) portée par les chromosomes homologues ($\frac{A}{a}\ \frac{B}{b}$ par exemple).
- **Phase**: information suffisante pour déterminer les deux haplotypes à partir du génotype.

Recombinants et Non recombinants

Marqueurs $\mathcal{A}, \mathcal{B}$

- une cellule diploïde portant les haplotypes AB/ab,
- on peut avoir les gamètes porteuses des haplotypes AB, ab, Ab, aB

Les deux premiers sont *parentaux* ou *non recombinants*. Les deux autres *recombinants* (nombre impair de cross-overs).



Taux de recombinaison

Taux de recombinaison $\leq \frac{1}{2}$

Le taux de recombinaison $r_{\mathcal{AB}}$ entre les deux marqueurs $\mathcal{A}$ et $\mathcal{B}$ est la proportion de *recombinants*.

Example

Entre 3 gènes $\mathcal{Y}$ (yellow), $\mathcal{W}$ (white), $\mathcal{M}$ (miniature) de la drosophile, on observe $r_{\mathcal{Y},\mathcal{W}} = 1,3\%$, $r_{\mathcal{W},\mathcal{M}} = 32,6\%$ et $r_{\mathcal{Y},\mathcal{M}} = 33,8\%$. On peut penser que les marqueurs sont dans l'ordre $\mathcal{Y} - \mathcal{W} - \mathcal{M}$

Du fait des doubles crossing-overs, pour un ordre $\mathcal{Y} - \mathcal{W} - \mathcal{M}$:

$$r_{\mathcal{Y},\mathcal{M}} < r_{\mathcal{Y},\mathcal{W}} + r_{\mathcal{W},\mathcal{M}} \quad \text{(non additif)}$$

Distance génétique

Définition

La distance génétique $d_{\mathcal{AB}}$ entre deux marqueurs $\mathcal{A}$ et $\mathcal{B}$ est le nombre moyen de *crossing-overs* entre les deux marqueurs par méiose.

Propriétés

- Additif
- 1cM (centiMorgan) correspond à un crossover sur un haplotype pour 100 méioses.
- Les cross-overs ne sont pas facilement observables.

Distance génétique et recombinaison

Taux de recombinaison : estimable à partir de données sur la descendance de parents bien choisis.

Distance génétique : s'appuie sur un modèle de la recombinaison.

Interférence

Le taux de double recombinaison est habituellement inférieur à celui attendu sous hypothèse d'indépendance.

- $1,3\% \times 32,6\% = 0,43\%$ attendu pour la double recombinaison $\mathcal{Y} - \mathcal{W} - \mathcal{M}$.
- $0,045\%$ observé.

Fonction de distance - map functions

Entre deux marqueurs. La première fonction de distance s'appuie sur un modèle de recombinaison simplifié (sans interférence, deux chromatides).

Fonction de Haldane - sans interférence (1919)

$$r = \frac{1}{2}(1 - e^{-2d}) \qquad d = -\frac{1}{2}\log(1 - 2r)$$

Beaucoup d'autres fonctions pour l'interférence:

Fonction de Kosambi - interférence (1944)

$$r = \frac{1}{2}\tanh(2d) \qquad d = \frac{1}{2}\tanh^{-1}(2r)$$

Pour de faibles distances/taux de recombinaison, $d \approx r$.

Cartographie génétique

Comment ?

1. Accumulation d'observations du génotype sur un ensemble de marqueurs et sur un bon nombre de méioses
   - Parents bien caractérisés (phase), hétérozygotie.
   - Observation sur la descendance
2. Reconstruire les distances et l'ordre des marqueurs.

Des situations variées

Taille de l'échantillon, temps de génération, mortalité des lignées, pénétrance, finesse des observations, souplesse des croisements (plantes, animaux, humain).
Observation de certains marqueurs/allèles parfois impossible (manquants).

Les hybrides d'irradiation

Principe

1. Irradiation de cellules menant à la fragmentation des chromosomes.
2. Cellule sauvée par hybridation à une cellule hôte qui retient une partie des fragments seulement (marqueur présent (1/Here) ou absent (0/Absent)).

Deux marqueurs proches sont souvent "co-retenus/co-exclus".

- Pas besoin de polymorphisme.
- Dose de rayonnement (résolution) ajustable.

Les hybrides d'irradiation

Taux de rétention, de cassure

- r : taux de rétention, proportion des fragments retenus. Peut être défini par marqueur ou globalement.
- $c_{\mathcal{AB}}$: taux de cassure entre les marqueurs $\mathcal{AB}$.

Cassures pas directement observables.

Motifs de rétention

- 01 ou 10 : cassure obligatoire (OCB: Obligate Chromosome Break).
- 11 ou 00: cassure ou pas ?

Distance : le Ray

Distance associée

Nombre moyen de cassures entre deux marqueurs par hybride.

$$c = 1 - e^{-d} \qquad d = -\log(1 - c)$$

F2, RIL, BC, BC avancé...

Pedigrees végétaux surtout

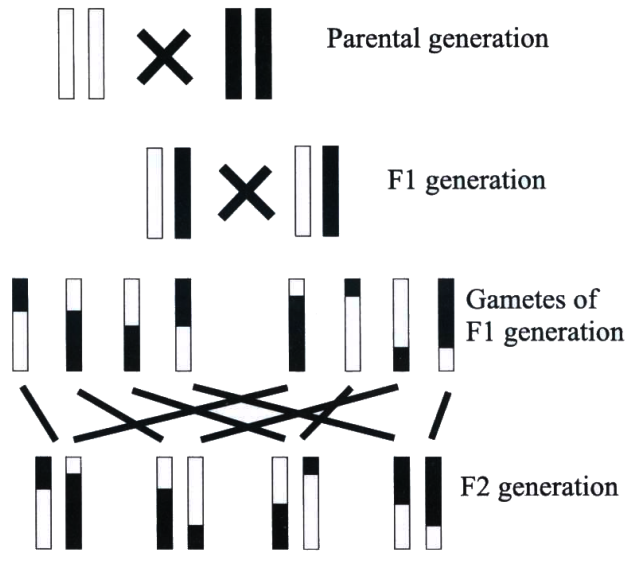
Croisements uniquement utilisés pour les organismes supportant une homozygotie quasi complète (lignées pures - inbred lines).

Inbreeding (selfing/sib)

Autofécondation répétée. Chaque croisement augmente le caractère homozygote. (20 gen., 98% d'identité). Impossible sur certaines plantes (PdT).

F2 - intercross

- lignées pures connues, différentes (*A* et *a*)
- on type la descendance en F2 : observation indirecte de deux méioses par individu.

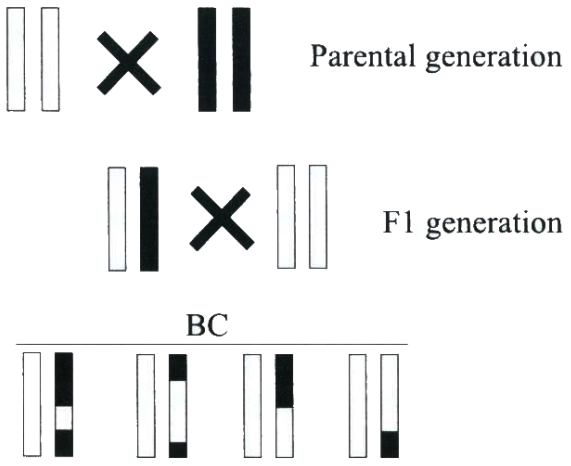


Typage F2: ségrégation 1:2:1

- homozygote *AA* $(\mathtt{A}, \frac{1}{4})$ ou *aa* $(\mathtt{B}, \frac{1}{4})$
- hétérozygote *Aa* $(\mathtt{H}, \frac{1}{2})$

	A	*a*
A	$\frac{1}{4}$	$\frac{1}{4}$
a	$\frac{1}{4}$	$\frac{1}{4}$

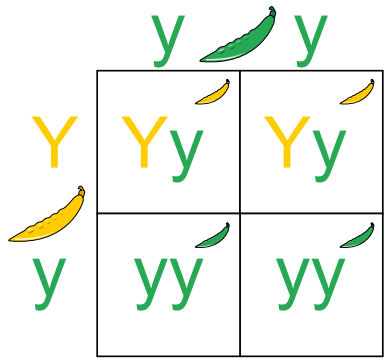
Backcross F2 - rétro-croisement



Typage F2: ségrégation 1:1

- homozygote *AA* $(\mathtt{A}, \frac{1}{2})$
- hétérozygote *Aa* $(\mathtt{H}, \frac{1}{2})$

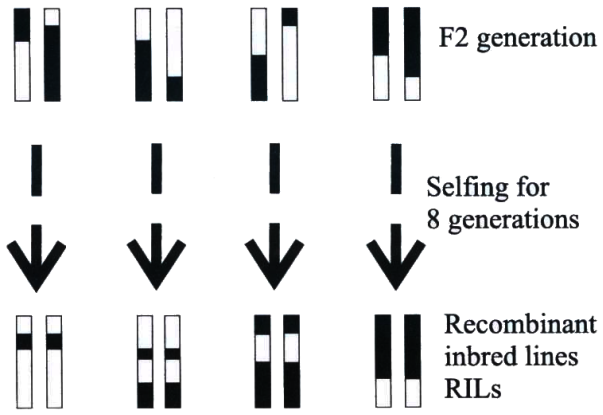
	A	*a*
A	$\frac{1}{4}$	$\frac{1}{4}$
A	$\frac{1}{4}$	$\frac{1}{4}$



Observation d'une seule méiose par individu.

Lignées recombinantes

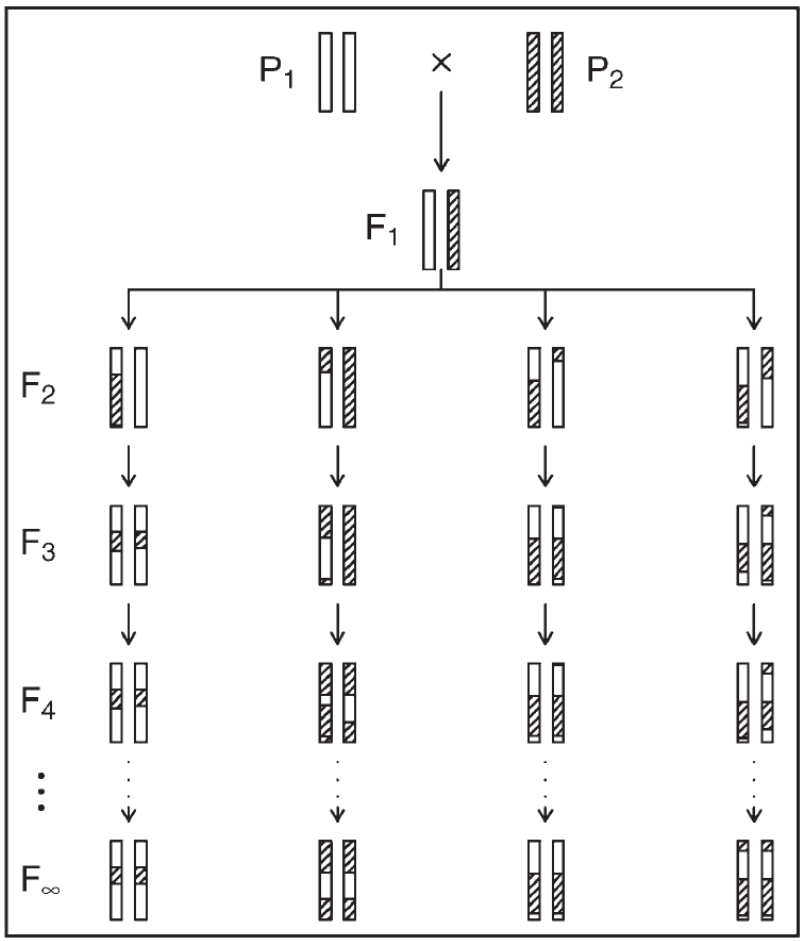
Autofécondation (self) ou croisement entre descendants d'une même génération (sib) répétée a partir d'une génération F2.

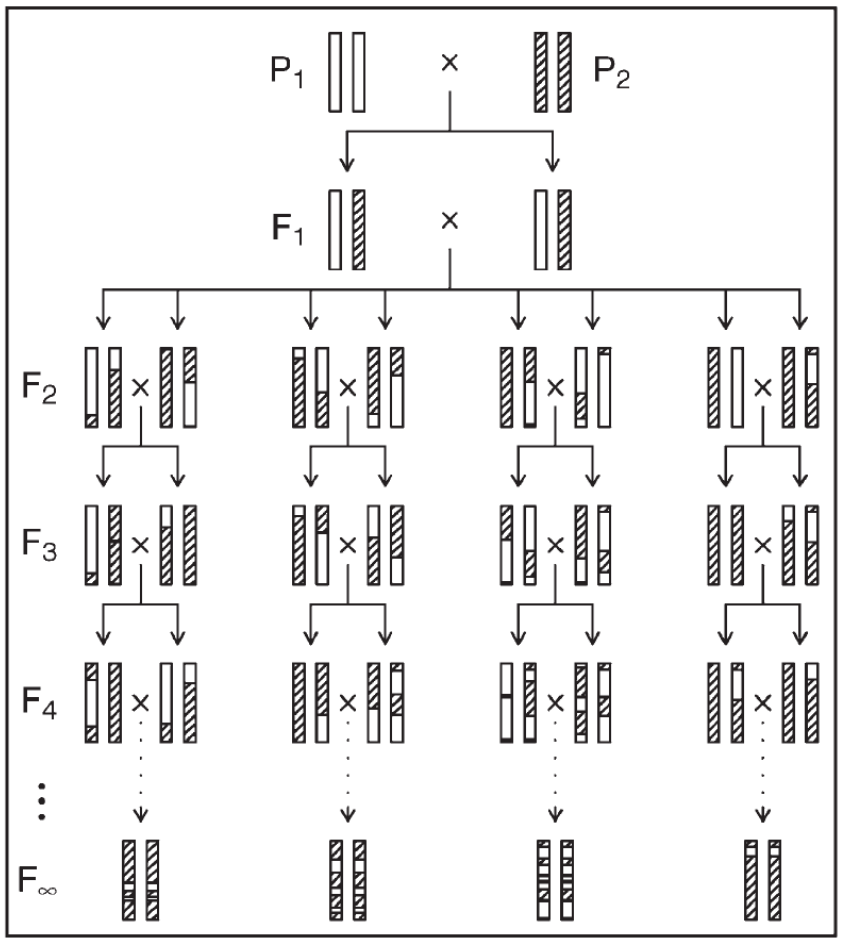


Avantages

- les individus finaux cumulent beaucoup de méioses (résolution accrue)
- Durée de vie infinie
- Se traitent comme des backcross, fonction de distance adaptée.

RIL selfing, full sib





Haploïdes doublés

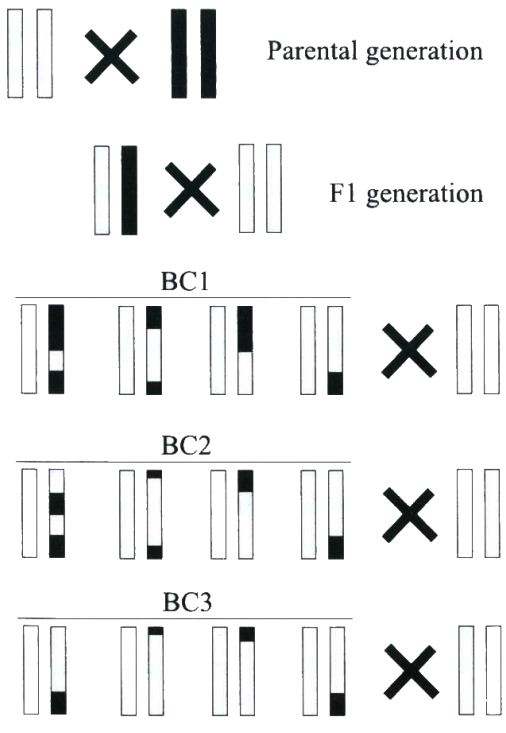
RILs couteux.

Haploides doublés - Doubled haploid - DH

Lignées totalement homozygotes obtenues en utilisant une cellule haploïde répliquée naturellement par la cellule ou par culture cellulaire de gamètes puis doublés via l'utilisation de colchicine.

Backcross avancé

On peut combiner librement les opération de type backcross et de type autofécondation. Souvent effectué avec des visées de sélection.



Example

Construction de NIL (Near Isogenic) puis autofécondations pour ne conserver qu'une partie minimale récupérée sur l'autre génome.

Outbreds - F1

Situation sans lignées pures - arbres

Croisement entre deux individus différents (maximiser l'hétérozygotie). Parents et descendance génotypés. Problème de détermination de la phase.

Example

- trois allèles A_1, A_2, A_3
- parents $A_1|A_2$ et $A_1|A_3$.
- Haplotypes A_1A_1, A_1A_2, A_1A_3, A_2A_3.

	A_1	A_2
A_1	$\frac{1}{4}$	$\frac{1}{4}$
A_3	$\frac{1}{4}$	$\frac{1}{4}$

Autres croisements: Demi-frères, CEPH

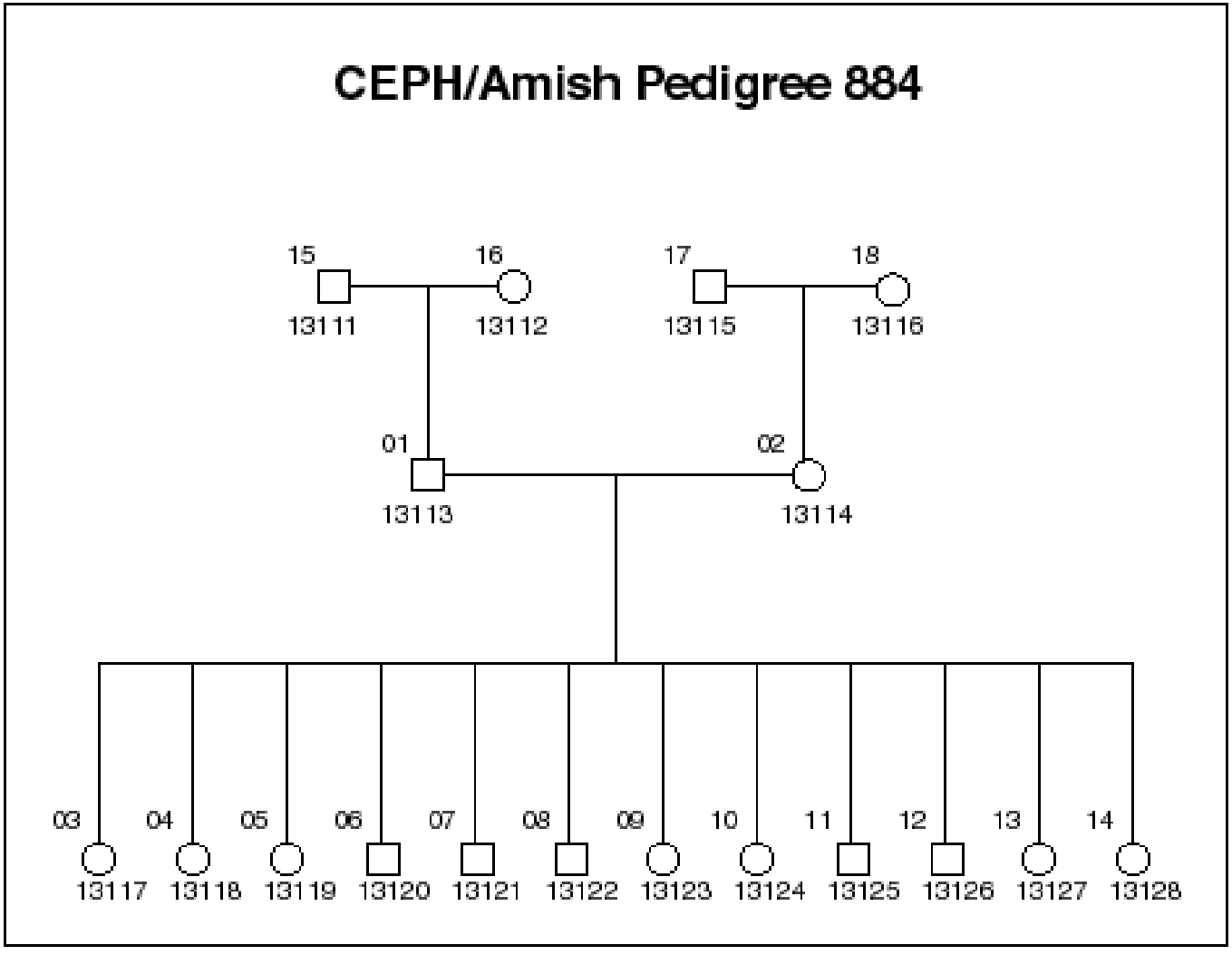
Demi-frères

Croisement utilisé pour les animaux avec utilisation d'insémination artificielle. Le père et les enfants sont typés mais pas les mères.

CEPH

Utilisé par le Centre d'Étude du Polymorphisme Humain pour cartographier (génétiquement) le génome humain.
Défini par un ensemble de familles nucléaires (père, mère, descendance) et les grands parents (utilisés pour déterminer les phases), tous génotypés.

Pedigree CEPH



Multi-population

Sur une espèce, on peut accumuler des données issues de croisements variés, partageant ou pas des individus, et avec un ensemble de marqueurs typés en commun plus ou moins important.

Construction de cartes consensus (fusion, marqueurs ponts/charnières).

Backcross F2: estimer r entre 2 marqueurs

Un individu F2

- deux marqueurs homozygotes (AA, BB) ou hétérozygotes (Aa, Bb) : **non recombinants** (NR).
- un hétérozygote, un homozygote (Aa, BB ou AA, Bb) : **recombinant** (R).

Vraisemblance - probabilité des données

Si r est le taux de recombinaison (à estimer), la probabilité d'observer les données de typage D (la vraisemblance) est :

$$P(D|r) = r^R.(1-r)^{NR}$$

Backcross F2: estimer r entre 2 marqueurs

Maximum de vraisemblance

La valeur estimée $\hat{r}$ de r choisie est celle qui maximise la probabilité d'observer les données (estimateur convergent).

- Conditions nécessaires: la dérivée première est nulle au maximum. La dérivée seconde est négative.
- La vraisemblance s'écrivant comme un produit, on travaille sur son logarithme (même maxima).

$$\log(P(D|r)) = R\log(r) + NR\log(1-r)$$

$$\hat{r} = \frac{R}{R+NR}$$

En pratique

- Individus non typés sur un marqueur. Données manquantes.
- On n'observe pas toujours les génotypes. Si un allèle *A* est "dominant", *A* est compatible avec *AA*, *Aa* en BC.
- Erreurs de typages

Vraisemblance de données incomplètes : somme des vraisemblances de toutes complétions possibles des données.

- nombre exponentiel de complétions
- le logarithme d'une somme ne se simplifie pas.

Utilisation d'algorithmes d'optimisation dédiés (EM - Expectation Maximisation, Dempster et al.,1977).

Nettoyage des données: distorsion

Marqueur distordu

Allèle sur-représentée dans la descendance / à la fréquence attendue (gène lié à la reproduction/croissance, réarrangements ou problème d'échantillonage)

Test de χ^2 de Pearson sur l'hypothèse nulle: les données observées sont tirées de la distribution théorique attendue. Pour un risque $\alpha = 0.05$, $\chi^2_{1ddl} = 3,84$

Exemple

Avec 100 descendants en Backcross

44 hétérozygotes, 56 homozygotes AA. 50/50 attendus.

$$X^2 = \frac{(O_1 - A_1)^2}{A_1} + \frac{(O_0 - A_0)^2}{A_0} = \frac{(O_1 - O_0)^2}{N}$$

$$X^2 = \frac{(44 - 50)^2}{50} + \frac{(56 - 50)^2}{50} = 1,44$$

On ne peut pas rejeter H_0 (non distordu).

Nettoyage des données: “marqueurs confondus”

Jeux de données modernes

- typage de plusieurs dizaines de milliers de marqueurs
- distance minimale inter-marqueurs très faible
- pas de recombinaison/cassure obervée : même génotypes (ou génotypes compatibles avec les données manquantes).
- Supprimer ou fusionner des marqueurs

CarthaGène: `mrkdouble/mrkmerge`. Quadratique.

Construction des groupes de liaison

Groupes de liaison

Groupes de marqueurs qui appartiennent à un même chromosome (liés).

- Hypothèse 0: 2 marqueurs $\mathcal{A}$ et $\mathcal{B}$ ont un assortiment indépendant (non liés, taux de recombinaison de $\frac{1}{2}$).
- Hypothèse 1: les 2 marqueurs $\mathcal{A}$ et $\mathcal{B}$ sont liés (taux de recombinaison $< \frac{1}{2}$).

LOD score - 2 marqueurs

$$LOD = \log_{10} \frac{P(D|r=\hat{r})}{P(D|r=\frac{1}{2})}$$

Tradition: $LOD > 3$ utilisé pour conclure à la liaison.

Construction des groupes de liaison

Example

En Backcross.

M1 AaAAAaAAAaaaAaaAAaAAAaaaAAAaaAaAaAaAaAaa

M2 AaaAAaAaAaaAAaAAAAaAAaaAAAaaaAAAaAaAaAAa

$R = 10; NR = 30; \hat{r} = \frac{1}{4}; LOD =$

$R.\log_{10}(\frac{1}{4}) + NR\log_{10}(1 - \frac{1}{4}) - (R + NR).\log_{10}(\frac{1}{2}) = 2,27$

Détermination des groupes de liaison

Méthode

On fixe un seuil minimal pour le LOD (3) et un seuil maximal pour la distance définie par $\hat{r}$ (20cM).

- Si 2 marqueurs satisfont ces conditions, ils sont placés dans le même groupe de liaison.
- Dans un graphe dont les sommets sont les marqueurs et où deux marqueurs sont reliés par une arête s'il respectent les seuils LOD/$\hat{r}$, les groupes sont donnés par les composantes connexes du graphe.

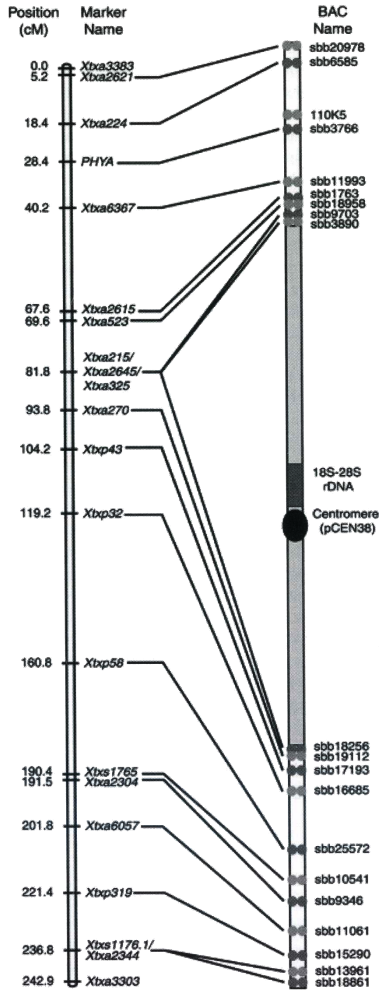
Qu'est-ce qu'une carte ?

Cartographier

Pour chaque groupe de liaison (partie de chromosome), déterminer l'ordre des marqueurs et les distances (taux de recombinaisons) qui séparent deux marqueurs adjacents

Carte saturée

autant de groupes que de chromosomes, tous les marqueurs de la carte sont liés à un groupe.



Qu'est-ce qu'une "bonne" carte

Deux grands types d'approches :

- non paramétriques : on cherche un ordre des marqueurs qui minimise le nombre de recombinants (génétique) ou de cassures (hybrides irradiés) obligatoires.
- paramétriques : on va chercher un ordre qui maximise la vraisemblance (paramètres = taux de recombinaison).
  - multi point: on utilise la vraisemblance maximum pour l'ordre des marqueurs (estimation multipoint)
  - multi 2-points: on minimise la somme des vraisemblances 2-points (entre paires de marqueurs adjacents), des LODs ou des distances. LODs sensible aux manquants.

Une carte "fiable" ?

Trouver une bonne carte

Un problème combinatoire

Pour n marqueurs, il y a $n!/2$ ordres de marqueurs définissant des cartes différentes. $\frac{10!}{2} = 1,8.10^6$.

- Impossible d'énumérer les ordres.
- Problème d'optimisation difficile (même dans ses versions les plus simples).

Lien avec le “Traveling Salesman Problem”

OCB en hybrides irradiés

Soit $\mathcal{A}, \mathcal{B}$, 2 marqueurs, on peut calculer au préalable le nombre $d_{\mathcal{AB}}$ de cassures obligatoires entre $\mathcal{A}$ et $\mathcal{B}$.
On cherche à trouver un ordre des marqueurs qui minimise la somme des $d_{\mathcal{AB}}$ entre les paires de marqueurs successifs.

Wandering Salesman Problem - WSP/TSP

Soit n villes et d_{ij} la distance entre les villes i et j. Trouver un chemin qui passe une fois et une seule par chaque ville et de longueur totale minimum.

Analogie marqueur/ville, distance/OCB. Si l’on sait résoudre le problème TSP, on sait trouver une carte qui minimise OCB.

Paramétrique : BC sans données manquantes

Marqueurs $\mathcal{A}_i, i \in \{1, \ldots, n\}$, typés sur p individus
R_{ij}, NR_{ij}: nombre de (non) recombinants entre i et j.
r_{ij}: taux de recombinaison entre $\mathcal{A}_i$ et $\mathcal{A}_j$. Pas d'interférence.

$$P(D|r) = \prod_{i=1}^{n-1} r_{i,i+1}^{R_{i,i+1}} . (1 - r_{i,i+1})^{NR_{i,i+1}}$$

$$\log P(D|r) = \sum_{i=1}^{n-1} R_{i,i+1} \log(r_{i,i+1}) + NR_{i,i+1} \log(1 - r_{i,i+1})$$

$$\hat{r}_{i,i+1} = \frac{R_{i,i+1}}{R_{i,i+1} + NR_{i,i+1}}$$

Paramétrique: BC sans données manquantes

Vraisemblance multipoint

Soit $d_{ij} = R_{i,j} \log(\hat{r}_{i,j}) + NR_{i,j} \log(1 - \hat{r}_{i,j})$.

$$\log P(D|r) = \sum_{i=1}^{n-1} d_{i,i+1}$$

1. Trouver un ordre de vraisemblance maximum se ramène à résoudre un WSP avec des distances d_{ij} 2 points, précalculables.
2. Ne s'applique pas à des échantillons avec données manquantes (ou avec des observations ne permettant pas de fixer le génotype).

Méthodes heuristiques pour le TSP

- par extension: aller toujours vers le marqueur le plus "proche" (en termes de d_{ij}). Nearest Neighbor (`nicemapl/d`).
- par insertion: insérer l'arête de d_{ij} minimum ne créant pas de sommet de degré 3 ou de cycle (multi-fragment, `mfmapl/d`). Proche de la "sériation" (Buetow et al, 1987).
- test de toutes les permutations dans une fenêtre glissante de k marqueurs (`flips`).
- ...

Certaines de ces méthodes ont des garanties (si l'inégalité triangulaire est vérifiée, multi-fragment n'est pas à plus d'un facteur $\frac{1}{2}(1 + \lceil \log_2(n) \rceil)$ de l'optimum).

Méthodes méta-heuristiques pour le TSP

Deux catégories

- métaheuristiques générales: recuit simulé, algorithmes génétiques ou évolutionnaires, recherche tabou,. . .
- métaheuristiques dédiées: algorithme de Lin-Kernighan (LKH: implémentation de Lin-Kernighan parmi les plus efficaces. Utilisée dans CarthaGène).

Point commun : méthodes de recherche “locale”.

Recherche locale

Voisinage d'une carte

Ensemble des cartes que l'on peut atteindre en perturbant la carte selon un mécanisme bien précis.

exemple: 2-OPT

Inversion d'une sous-section de la carte.

1. On part d'un (ou plusieurs) ordre(s) initial(aux).
2. On considère un sous-ensemble des ordres dans le voisinage de l'ordre courant
3. on décide éventuellement de se déplacer vers un de ces voisins.

Méthodes exactes pour le TSP

Concorde

- Méthodes issues de l'optimisation combinatoire.
- Package CONCORDE disponible sur http://www.tsp.gatech.edu/concorde.html.
- CarthaGène exporte les problèmes au format TSPLIB.

Limitations pratiques

- erreurs de typage. Des modèles probabilistes dédiés existent mais utilisent la connaissance d'un ordre supposé exact.
- Forte densité: la probabilité d'observer une recombinaison entre 2 marqueurs adjacents est de plus en plus faible. Nombreux marqueurs identiques.

High-throughput

Cartes denses mais peu fiables sur tous les marqueurs.

- ajout d'information (cartographie comparative)
- diminution du nombre de marqueurs (framework mapping)

Logiciels de cartographie

- végétaux: MapMaker, CarthaGene, JoinMap
- animaux: CRIMAP,
- homme: MapMaker
- hybrides irradiés: RHMAP, RHO, CarthaGene

Voir http://linkage.rockefeller.edu/soft/