Travaux dirigés marqueurs et cartographie

T. Schiex, INRA

October 4, 2012

1 Cartes haute densité

Le but de ce premier exercice est d'explorer les limites des jeux de données haute densité (avec un grand nombre de marqueurs, comme le permettant les technologies de typage, 160K marqueurs...).

On suppose que sur un chromosome de 1 Morgan, on a n marqueurs typés (n grand) sur p individus, dans une population de type Backcross, sans interférence. On suppose que les n marqueurs sont répartis de façon uniforme sur le chromosome et qu'il n'y a pas de manquants.

- Quelle est la probabilité pour 2 marqueurs *adjacents* d'avoir exactement le même vecteur de génotypes ? Calculer cette valeur pour 1000 marqueurs et 100 individus.
- En ne tenant compte que des possibles identités de génotypes entre paires de marqueurs adjacents, quel est le nombre moyen de marqueurs typés inutilement (avec une série de génotype identique à un marqueur adjacent) sur un jeux de n marqueurs et p individus ? Valeur pour 1000 marqueurs et 100 individus ?
- On souhaiterait typer suffisamment d'individus pour limiter la perte de marqueurs par identité de typage. Pour n marqueurs ($n-1$ paires de marqueurs adjacents), quelle nombre d'individus est-il nécessaire de typer pour que la fraction de marqueurs perdus soit plus faible qu'une valeur ε fixée. Les marqueurs restants seront les marqueurs "effectifs" et leur nombre sera noté $n*$.

  Calculez p pour $n = 100, 1000, 10000$ et une fraction de 10%.

Ces résultats montre que typer un grand nombre de marqueurs sur peu d'individus (par rapport au nombre de marqueurs) est peu efficace. On cherche maintenant à fournir des conditions absolument nécessaires sur p pour pouvoir identifier un ordre correct des marqueurs.

Pour qu'un ordre quelconque (inconnu) des marqueurs soit identifiable, il est nécessaire que l'on dispose au moins d'assez d'information pour discriminer parmi tous les ordres possibles. L'ensemble des jeux de données possibles doit

donc avoir un cardinal au moins supérieur au nombre d'ordres possibles. Cette condition nécessaire est très faible bien sûr.

Si l'on type n marqueurs et que l'on souhaite en ordonner n' parmi ces n, on se retrouve avec $O = \frac{n'!}{2}$ ordres possibles et $J = 2^{pn*}$ jeux de données effectifs différents (les marqueurs de génotypage identique ne fournisse pas d'information). Il faudrait donc que J soit au moins égal à O.

En déduire une contrainte sur le nombre d'individus minimum à typer pour ordonner n marqueurs. On utilisera plusieurs approximations liées au fait que n est grand (et donc $1/n$ petit) :

- $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$ (formule de Stirling)
- $(1+x)^p \approx 1 + px$ pour x petit

Quelle contrainte obtient t'on sur p pour $n = 1000, 10000$ marqueurs ? Cette contrainte est elle informative ?

2 Distorsion et LODs

On considère une population de type BC. On observe le jeu de données suivant de 5 marqueurs sur 50 individus. (0 est utilisé pour représenter le génotype A et 1 le génotype H:

```
*M1     01111000010110001111011100110001001000111011111011
*M2     01111000010010001010011100110001001000111011111010
*M3     10101000111001001011001100110001001000111010111010
*M4     10111011111001001010101101110111010101111001111111
*M5     10100001111001001110011100010011010110111001111011
```

- Il y a un marqueur distordu dans ce jeu de données. Lequel, vérifier le par un test avec $\alpha = 5\%$.
- Quel est le taux de recombinaison estimé entre M1 et M2 et entre M1 et M5 ?
- quelles sont les valeurs des distances estimées (en cM) pour ces deux paires en utilisant la distance de Haldane ? la distance de Kosambi ?
- calculer le LOD entre M1 et M2, entre M1 et M5. Peut-on en conclure que M1 et M5 sont ou ne sont pas dans un même groupe de liaison ?

3 Vraisemblance et hybrides irradiés

On s'intéresse à la cartographie par hybrides d'irradiation. Un hybride d'irradiation est formé par une cellule de rongeur qui contient des fragments de chromosomes issus d'un autre organisme.

On irradie des chromosomes de l'orgaisme d'intérêt avec des rayons X qui fragmentent aléatoirement ces chromosomes. Ces fragments sont ensuite être

conservés et propagés si les cellules irradiées sont fusionnées avec des cellules non irradiées de hamster (ou autre rongeur). On suppose, par souci de simplification, que l'organisme étudié est haploïde.

En répétant ce processus sur un nombre important de cellules (et en sélectionnant celles qui ont effectivement retenu au moins un fragment avec un marqueur précis), on peut observer pour chaque hybride et marqueur si le marqueur apparaît (est retenu) ou non dans l'hybride. Chaque hybride défini ainsi un "motif de rétention" (*retention pattern*).

Les fondements de la cartographie RH est que plus la distance physique entre deux marqueurs est faible, moins il est probable qu'une cassure ait eu lieu entre eux et donc ces deux marqueurs seront probablement "co-retenus" : soit ils sont tous les deux retenus, soit ils sont tous les deux absents.

3.1 Cartographie non paramétrique

1. on considère deux marqueurs $\mathcal{M}_1$ et $\mathcal{M}_2$ et on note M_i^k le booléen qui indique si le marqueur $\mathcal{M}_i$ correspondant est retenu (1) ou non (0) sur l'hybride k.

   Pour chaque configuration possible de M_1 et M_2, indiquer si une cassure entre les deux marqueurs est possible, impossible ou obligatoire.

2. Étant donné un ensemble d'hybrides $H = \{1, \ldots, m\}$, calculer le nombre de cassures obligatoires O_{ij} entre deux marqueurs $\mathcal{M}_i$ et $\mathcal{M}_j$ en fonction des M_i^k.

3. On souhaite ordonner un ensemble de marqueurs $M_i, i \in \{1, \ldots, n\}$. Proposer, en le justifiant, un critère utilisant le nombre de cassure obligatoires entre marqueurs pour ordonner les marqueurs (le ou les ordres qui optimise(nt) ce critère seront sélectionnés).

   En exploitant sa similarité avec un problème classique en optimisation, proposer une approche algorithmique inspirée des algorithmes utilisés pour résoudre le problème d'ordonnancement en cartographie génétique pour résoudre le problème d'ordonnancement.

3.2 Approche probabiliste 2 points

On note c_{ij} la probabilité qu'une cassure au moins ait lieu entre les marqueurs $\mathcal{M}_i$ et $\mathcal{M}_j$. On suppose de plus qu'un fragment donné a une probabilité p d'être retenu.

1. Pour les deux marqueurs $\mathcal{M}_i$ et $\mathcal{M}_j$, donner la probabilité d'observer chacun des motifs de rétention observables sur un hybride k donné en fonction de p et de c_{ij}.

2. Etant donné un ensemble d'hybrides $H = \{1, \ldots, m\}$ et deux marqueurs $\mathcal{M}_i$ et $\mathcal{M}_j$, donner la vraisemblance des observations sur ces deux marqueurs sur le panel, d'hybrides en fonction de p, c_{ij} et des observations.

3. On estime p comme le ratio entre le nombre de marqueurs retenus et non retenus. Montrer, sans aller jusqu'à un calcul explicite, que l'estimateur de maximum de vraisemblance pour c_{ij} peut se calculer en résolvant une équation du second degré. On rappelle que la dérivée, pour une fonction $f(x)$ positive quelconque, de $\log(f(x))$ est $\frac{f'(x)}{f(x)}$.

4. Une telle équation admet deux racines. Ici, ces deux racines sont :

   - $\theta_{ij} = \frac{(n-n_{11}p-n_{00}(1-p))-\sqrt{(n-n_{11}p-n_{00}(1-p))^2-4np(1-p)(n_{10}+n_{01})}}{2np(1-p)}$
   - $\theta_{ij} = \frac{(n-n_{11}p-n_{00}(1-p))+\sqrt{(n-n_{11}p-n_{00}(1-p))^2-4np(1-p)(n_{10}+n_{01})}}{2np(1-p)}$

   Quelle racine choisiriez vous? Pourquoi (on examinera en particulier le cas ou le jeu de données ne contient aucune cassure obligatoire en tre i et j).