Calcul de seuil pour la détection de QTL

Céline Delmas

INRA - Toulouse

23-24 Septembre 2010

Modélisation de Haldane (1919)(1/3)

- Le chromosome est modélisé par le segment $[0, T]$
- Notons X_t le génotype à la position $t \in [0, T]$
- X_t est un processus markovien de sauts de loi initiale $P[X_0 = 1] = P[X_0 = -1] = 1/2$ et de matrices de transition (P_t) :

$$P_t = \begin{bmatrix} 1 - r_t & r_t \\ r_t & 1 - r_t \end{bmatrix}$$

où r_t est la probabilité d'avoir un nombre impair de recombinations entre deux loci distants de t

Modélisation de Haldane (1919)(2/3)

- Si l'on suppose que ce nombre de recombinaisons suit une loi de Poisson de paramètre λt alors

$$r_t = \frac{1}{2}(1 - \exp(-2\lambda t)).$$

λ est classiquement choisi égal à 1 car on suppose qu'il y a en moyenne une recombinaison sur un chromosome de longueur 1M.

- La matrice génératrice du processus (X_t) définie par $A = \frac{d}{dt} P_t|_{t=0}$ est alors égale à

$$A = \left[\begin{array}{cc} -\lambda & \lambda \\ \lambda & -\lambda \end{array} \right].$$

Modélisation de Haldane (1919)(3/3)

- La matrice de transition P_t et la matrice génératrice A sont liées par l'équation $P_t = \exp(tA)$. Avec la loi initiale, elles suffisent à caractériser le processus markovien de sauts (X_t).

- Cette modélisation de l'information génomique est à la base de tous les résultats connus pour le calcul des seuils et la détection de QTL en analyse de liaison.

Extension de la modélisation de Haldane à l'analyse d'association (1/2)

- Trois génotypes sont possibles à la position t.
- La loi initiale pourra être $P[X_0 = 0] = P[X_0 = 1] = P[X_0 = 2] = 1/3$.
- L'extension naturelle la plus simple de la modélisation précédente semble être :

$$A = \begin{bmatrix} -2\lambda & \lambda & \lambda \\ \lambda & -2\lambda & \lambda \\ \lambda & \lambda & -2\lambda \end{bmatrix}.$$

Extension de la modélisation de Haldane à l'analyse d'association (2/2)

Ceci signifie que :

- Les évènements de sauts apparaissent le long du chromosome selon un processus de Poisson de paramètre 2λ.
- A chaque évènement de saut le génotype change de manière "'équiprobable"'.

Le paramètre λ pourra :

- soit être pris égal à $1/2$ (pour être en harmonie avec la modélisation classique de Haldane)
- soit être estimé sur les données.

Calcul du seuil - Contexte

- On s'interesse à un caractère quantitatif Y qui dépend de la valeur de X_t en $t^\star$, position inconnue du QTL :

$$Y_i = \mu + X(t^\star)q + \sigma\epsilon$$

  où ϵ est un bruit blanc gaussien et q est l'effet du QTL.
- En tout point $t \in [0, T]$ on effectue un test de rapport de vraisemblance de l'hypothèse nulle $q = 0$ versus l'hypothèse alternative $q \neq 0$.
- Notons $S(\cdot)$ le profil de vraisemblance obtenu.

Calcul du seuil - Contexte

- L'approche classique pour le calcul du seuil consiste à utiliser comme statistique de test :

$$\sup_{t \in [0,T]} S(t).$$

- N'est-il pas plus puissant et plus efficace d'utiliser

$$\int_0^T S(t)dt$$

Calcul du seuil - Deux marqueurs

- On suppose qu'il n'y a que deux marqueurs localisés en 0 et T.
- En notant $\theta = (q, \mu, \sigma)$, la vraisemblance pour une observation s'écrit :

$$L(\theta, t) = [p(t) f_{(\mu+q,\sigma)}(y) + \{1 - p(t)\} f_{(\mu-q,\sigma)}(y)] g(t)$$

où $f_{(m,s)}(y)$ désigne la densité d'une loi gaussienne de moyenne m et de variance s et $g(t) =$

$$\begin{aligned} & \frac{1}{2}\{\bar{r}(t_1,t_2) 1_{X(t_1)=1} 1_{X(t_2)=1} + r(t_1,t_2) 1_{X(t_1)=1} 1_{X(t_2)=-1}\} \\ + & \frac{1}{2}\{r(t_1,t_2) 1_{X(t_1)=-1} 1_{X(t_2)=1} + \bar{r}(t_1,t_2) 1_{X(t_1)=-1} 1_{X(t_2)=-1}\} \end{aligned}$$

Calcul du seuil - Deux marqueurs

$$p(t) = P\{X(t) = 1 | X(t_1), X(t_2)\}$$

$$\begin{aligned} p(t) &= Q_t^{1,1} 1_{X(t_1)=1} 1_{X(t_2)=1} + Q_t^{1,-1} 1_{X(t_1)=1} 1_{X(t_2)=-1} \\ &+ Q_t^{-1,1} 1_{X(t_1)=-1} 1_{X(t_2)=1} + Q_t^{-1,-1} 1_{X(t_1)=-1} 1_{X(t_2)=-1} \end{aligned}$$

où

$$Q_t^{1,1} = \frac{\bar{r}(t_1,t)\bar{r}(t,t_2)}{\bar{r}(t_1,t_2)} \quad , \quad Q_t^{1,-1} = \frac{\bar{r}(t_1,t)r(t,t_2)}{r(t_1,t_2)}$$

$$Q_t^{-1,1} = \frac{r(t_1,t)\bar{r}(t,t_2)}{r(t_1,t_2)} \quad , \quad Q_t^{-1,-1} = \frac{r(t_1,t)r(t,t_2)}{\bar{r}(t_1,t_2)}$$

Calcul du seuil - Deux marqueurs

Soit $S_n(\cdot)$ le processus de test de rapport de vraisemblance pour n observations,

$$S_n(\cdot) \Rightarrow \{Z(\cdot)\}^2$$

où $Z(\cdot)$ est un processus gaussien de variance 1, de fonction de covariance $\forall(t,t') \in [t_1, t_2]^2$:

$$\Gamma(t,t') = \frac{4E\{p(t)p(t')\} - 1}{\sqrt{E[\{2p(t) - 1)\}^2]}\sqrt{E[\{2p(t') - 1)\}^2]}}$$

de fonction moyenne, sous l'hypothèse de l'existence d'un QTL en $t^\star$ d'effet $q = a/\sqrt{n}$:

$$m_{t^\star}(t) = \frac{aE[X(t^\star)\{2p(t) - 1\}]}{\sigma\sqrt{E[\{2p(t) - 1\}^2]}}$$

Calcul du seuil - Deux marqueurs

Le processus $Z(\cdot)$ est en particulier caractérisé par $\forall t \in [t_1, t_2]$:

$$Z(t) = \{\alpha(t)Z(t_1) + \beta(t)Z(t_2)\}/\sqrt{E[\{2p(t)-1\}^2]}$$

où

$$\alpha(t) = Q_t^{1,1} + Q_t^{1,-1} - 1, \beta(t) = Q_t^{1,1} - Q_t^{1,-1}$$
$$\alpha(t_1) = 1, \beta(t_1) = 0, \alpha(t_2) = 0, \beta(t_2) = 1$$
$$\text{Cov}\{Z(t_1), Z(t_2)\} = \exp(-2(t_2 - t_1))$$

De même :

$$m_{t^\star}(t) = \{\alpha(t)m_{t^\star}(t_1) + \beta(t)m_{t^\star}(t_2)\}/\sqrt{E[\{2p(t)-1\}^2]}$$

$\Rightarrow$ Le processus $Z(\cdot)$ est facilement simulable

Calcul du seuil - Deux marqueurs

Notons :

$$\xi = \frac{\{Z(t_1)\}^2 + \{Z(t_2)\}^2 - 2e^{-2(t_2-t_1)}Z(t_1)Z(t_2)}{\{1+e^{-2(t_2-t_1)}\}\{1-e^{-2(t_2-t_1)}\}}$$

Alors

$$\sup_{t\in[t_1,t_2]} \{Z(t)\}^2 = \begin{cases} \xi \text{ si } \frac{Z(t_2)}{Z(t_1)} \in]e^{-2(t_2-t_1)}, e^{2(t_2-t_1)}[\\ \\ \max[\{Z(t_1)\}^2, \{Z(t_2)\}^2] \text{ sinon} \end{cases}$$

Calcul du seuil - Deux marqueurs

Notons :

$$\tau = \frac{(t_2 - t_1)\{e^{-2(t_2-t_1)} Z(t_1) - Z(t_2)\}}{\{e^{-2(t_2-t_1)} - 1\}\{Z(t_1) + Z(t_2)\}} + t_1$$

et définissons τ' tel que

$$\frac{(t_2 - t_1)\beta(\tau')}{\alpha(\tau') + \beta(\tau')} + t_1 = \tau$$

Alors

$$\text{argsup}_{t \in [t_1, t_2]} \{Z(t)\}^2 = \tau'$$

Calcul du seuil-Plusieurs marqueurs

- On suppose que K marqueurs sont positionnés en $0 = t_1 < t_2 < \cdots < t_K = T$
- Notons $\mathcal{T}_k = \{t_1, \cdots, t_K\}$, soit $t \in [t_1, t_K]$, définissons t^l et t^r tels que :

$$t^l = \sup\{t_k \in \mathcal{T}_k : t_k < t\} \text{ , } t^r = \inf\{t_k \in \mathcal{T}_k : t < t_k\}$$

Alors

$$Z(t) = \{\alpha(t)Z(t^l) + \beta(t)Z(t^r)\}/\sqrt{E[\{2p(t)-1\}^2]}$$

avec $\forall k \; \forall k'$, $\text{cov}\{Z(t_k), Z(t_{k'})\} = e^{-2|t_k - t_{k'}|}$; de plus :

$$m_{t^\star}(t) = \{\alpha(t)m_{t^\star}(t^l) + \beta(t)m_{t^\star}(t^r)\}/\sqrt{E[\{2p(t)-1\}^2]}$$

Calcul du seuil - Plusieurs marqueurs

Soit $2 \leq k \leq K$, notons :

$$\xi_k = \frac{\{Z(t_{k-1})\}^2 + \{Z(t_k)\}^2 - 2e^{-2(t_k - t_{k-1})} Z(t_{k-1}) Z(t_k)}{\{1 + e^{-2(t_k - t_{k-1})}\}\{1 - e^{-2(t_k - t_{k-1})}\}}$$

Notons

$$H_k = \begin{cases} \xi_k \text{ si } \frac{Z(t_k)}{Z(t_{k-1})} \in]e^{-2(t_k - t_{k-1})}, e^{2(t_k - t_{k-1})}[\\ \\ \max[\{Z(t_{k-1})\}^2, \{Z(t_k)\}^2] \text{ sinon} \end{cases}$$

Alors

$$\sup_{t \in [0,T]} \{Z(t)\}^2 = \max(H_2, \cdots, H_K) = H_{t^\star}$$

Calcul du seuil - Plusieurs marqueurs

Soit $2 \leq k \leq K$, notons :

$$\tau_k = \frac{(t_k - t_{k-1})\{e^{-2(t_k - t_{k-1})} Z(t_{k-1}) - Z(t_k)\}}{\{e^{-2(t_k - t_{k-1})} - 1\}\{Z(t_{k-1}) + Z(t_k)\}} + t_{k-1}$$

et définissons τ'_k tel que

$$\frac{(t_k - t_{k-1})\beta(\tau'_k)}{\alpha(\tau'_k) + \beta(\tau'_k)} + t_{k-1} = \tau_k$$

Alors

$$\text{argsup}_{t \in [t_1, t_2]} \{Z(t)\}^2 = \tau'_{k^\star}$$

Calcul du seuil - Implémentation

Un package Matlab "'imapping"' pour le calcul du seuil a été réalisé par C.E. Rabier et A. Genz. Il est téléchargeable à l'adresse suivante :

http ://www.math.univ-toulouse.fr/ rabier/doc/articles.html

Calcul du seuil - Plusieurs familles

- On considère I familles
- C est la variable aléatoire indiquant la famille
- Un individu appartient à la famille i avec probabilité $\pi_i = P[C = i]$
- Le caractère Y dépend de la valeur de $X(t)$ en $t^\star \in [t_1, t_K]$ qui est la position du QTL.
- La valeur du caractère Y dépend aussi de la famille sur laquelle il est mesurée.

$$(Y|C = i) = \mu_i + X(t^\star)q_i + \sigma\epsilon$$

où μ_i et q_i sont respectivement l'effet polygénique et l'effet QTL dans la famille i ; ϵ est un bruit blanc gaussien.

Calcul du seuil - Plusieurs familles

- n est le nombre total d'observations j, $(Y_j, X_j(t_1), \cdots, X_j(t_K), C_j)$. Ces observations sont supposées iid.

- Existe t'il un QTL dans au moins une des familles en un point $t^\star$ inconnu ?

Calcul du seuil - Plusieurs familles

- On note $\theta = (q_1, \cdots, q_I, \mu_1, \cdots, \mu_I, \sigma)$ le paramètre du modèle à t fixé.
- La vraisemblance pour une observation s'écrit :

$$L(\theta, t) = \sum_{i=1}^{I} [p(t) f_{(\mu_i + q_i, \sigma)}(y) + \{1 - p(t)\} f_{(\mu_i - q_i, \sigma)}(y)] 1_{C=i} \pi_i g(t)$$

- La vraisemblance pour n observations s'obtient par le produit de n termes tels que ci-dessus.
- Sous $\mathcal{H}_0$, il n'y a pas de QTL sur $[0, T]$. Sous $\mathcal{H}_1$, il n'y a qu'un seul QTL en $t^\star$ sur $[0, T]$ dans au moins une des familles d'effet $q_i = \lambda_i / \sqrt{n}$.

Calcul du seuil - Plusieurs familles

Notons $S_n(\cdot)$ le profil de vraisemblance. Quand n tend vers l'infini, sous $\mathcal{H}_0$ et $\mathcal{H}_1$:

$$S_n(\cdot) \Rightarrow \sum_{i=1}^{I} \{Z^i(\cdot)\}^2$$

- Les $Z^i(\cdot)$ sont des processus gaussiens indépendants de variance 1.
- $\forall k \in \{1, \cdots, K-1\}$, $\forall t \in [t_k, t_{k+1}]$:

$$Z^i(t) = \{\alpha(t) Z^i(t_k) + \beta(t) Z^i(t_{k+1})\} / \sqrt{E[\{2p(t)-1\}^2]}$$

où $\text{Cov}\{Z(t_k), Z(t_{k+1})\} = e^{-2(t_{k+1}-t_k)}$.

Calcul du seuil - Plusieurs familles

- La fonction moyenne de $Z^i(\cdot)$ est nulle sous $\mathcal{H}_0$ et vérifie sous $\mathcal{H}_1$ $\forall (t, t^\star) \in [t_k, t_{k+1}] \times [t_1, t_K]$:

$$m^i_{t^\star}(t) = \{\alpha(t) m^i_{t^\star}(t_k) + \beta(t) m^i_{t^\star}(t_{k+1})\} / \sqrt{E[\{2p(t) - 1\}^2]}$$

où :

$$\alpha(t) = Q_t^{1,1} + Q_t^{1,-1} - 1 \text{ , } \beta(t) = Q_t^{1,1} - Q_t^{1,-1}$$

$$E[\{2p(t) - 1\}^2] = \{\alpha(t)\}^2 + \{\beta(t)\}^2 + 2\alpha(t)\beta(t) e^{-2(t_{k+1} - t_k)}$$

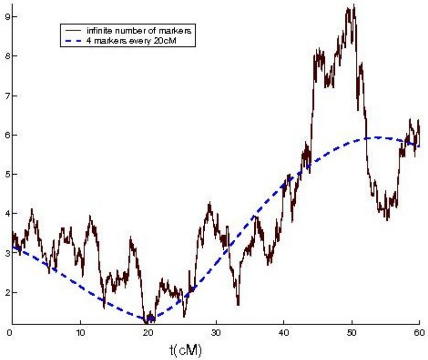
$$Q_t^{1,1} = \frac{\bar{r}(t_k, t)\bar{r}(t, t_{k+1})}{\bar{r}(t_k, t_{k+1})} \text{ , } Q_t^{1,-1} = \frac{\bar{r}(t_k, t) r(t, t_{k+1})}{r(t_k, t_{k+1})}$$

$$m^i_{t^\star}(t_k) = \lambda_i \sqrt{\pi_i} e^{-2|t^\star - t_k|} / \sigma$$

Calcul du seuil - Plusieurs familles

Remarque : Lorsque le nombre de marqueurs est infini, les processus $Z^i(\cdot)$ sont des processus d'Ornstein-Uhlenbeck ; par conséquent $\sum_{i=1}^{I}\{Z^i(\cdot)\}^2$ est un processus de χ^2 d'Ornstein-Uhlenbeck à I degrés de liberté.

Calcul du seuil - Plusieurs familles Profils pour $T = 60cM$ et $I = 3$



Implémentation

Un package Matlab "'imappingfamily"' pour le calcul du seuil a été réalisé par C.E. Rabier et A. Genz. Il est téléchargeable à l'adresse suivante :

http ://www.math.univ-toulouse.fr/ rabier/doc/articles.html

Calcul du seuil - Contexte

- On s'interesse à un caractère quantitatif Y qui dépend de la valeur de $X(t_1^\star)X(t_2^\star)$ en $(t_1^\star, t_2^\star)$, position inconnue de l'interaction :

$$Y_i = \mu + X(t_1^\star)X(t_2^\star)q + \sigma\epsilon$$

  où ϵ est un bruit blanc gaussien et q est l'effet de l'interaction.
- En tout point $(t_1, t_2) \in [0, T]^2$ on effectue un test de rapport de vraisemblance de l'hypothèse nulle $q = 0$ versus l'hypothèse alternative $q \neq 0$.
- Notons $S(\cdot, \cdot)$ la surface de vraisemblance obtenue.

Calcul du seuil - Deux marqueurs

- On suppose qu'il n'y a que deux marqueurs localisés en 0 et T.
- En notant $\theta = (q, \mu, \sigma)$, la vraisemblance pour une observation s'écrit :

$$L(\theta, t_1^\star, t_2^\star) = [p(t_1^\star, t_2^\star) f_{(\mu+q,\sigma)}(y) + \{1 - p(t_1^\star, t_2^\star)\} f_{(\mu-q,\sigma)}(y)] g(t^\star)$$

où $f_{(m,s)}(y)$ désigne la densité d'une loi gaussienne de moyenne m et de variance s et $g(t^\star) =$

$$\begin{aligned} & \frac{1}{2}\{\bar{r}(t_1,t_2)1_{X(t_1)=1}1_{X(t_2)=1} + r(t_1,t_2)1_{X(t_1)=1}1_{X(t_2)=-1}\} \\ + \quad & \frac{1}{2}\{r(t_1,t_2)1_{X(t_1)=-1}1_{X(t_2)=1} + \bar{r}(t_1,t_2)1_{X(t_1)=-1}1_{X(t_2)=-1}\} \end{aligned}$$

Calcul du seuil - Deux marqueurs

$$p(t_1^\star, t_2^\star) = P\{X(t_1^\star) = X(t_2^\star) | X(t_1), X(t_2)\}$$

$$p(t_1^\star, t_2^\star) = Q^{1,1}_{t_1^\star,t_2^\star} 1_{X(t_1)=1} 1_{X(t_2)=1} + Q^{1,-1}_{t_1^\star,t_2^\star} 1_{X(t_1)=1} 1_{X(t_2)=-1}$$
$$+ Q^{-1,1}_{t_1^\star,t_2^\star} 1_{X(t_1)=-1} 1_{X(t_2)=1} + Q^{-1,-1}_{t_1^\star,t_2^\star} 1_{X(t_1)=-1} 1_{X(t_2)=-1}$$

où

$$Q^{1,1}_{t_1^\star,t_2^\star} = Q^{-1,-1}_{t_1^\star,t_2^\star} = \frac{\bar{r}(t_1^\star, t_2^\star)}{\bar{r}(t_1, t_2)} [\bar{r}(t_2, t_2^\star)\bar{r}(t_1, t_1^\star) + r(t_2, t_2^\star) r(t_1, t_1^\star)]$$

$$Q^{1,-1}_{t_1^\star,t_2^\star} = Q^{-1,1}_{t_1^\star,t_2^\star} = \frac{\bar{r}(t_1^\star, t_2^\star)}{r(t_1, t_2)} [\bar{r}(t_2, t_2^\star) r(t_1, t_1^\star) + r(t_2, t_2^\star)\bar{r}(t_1, t_1^\star)]$$

Calcul du seuil - Deux marqueurs

Soit $S_n(\cdot,\cdot)$ la surface de test de rapport de vraisemblance pour n observations,

$$S_n(\cdot,\cdot) \Rightarrow \{Z(\cdot,\cdot)\}^2$$

où $Z(\cdot,\cdot)$ est un champ aléatoire gaussien de variance 1, de fonction de covariance $\forall((t_1,t_2),(t_1',t_2')) \in [0,T]^2 \times [0,T]^2$:

$$\Gamma((t_1,t_2),(t_1',t_2')) =$$

$$\frac{E[\{2p(t_1,t_2)-1-E(2p(t_1,t_2)-1)\}\{2p(t_1',t_2')-1-E(2p(t_1',t_2')-1)\}]}{\sqrt{\mathrm{Var}\{2p(t_1,t_2)-1)\}}\sqrt{\mathrm{Var}\{2p(t_1',t_2')-1)\}}}$$

de fonction moyenne, sous l'hypothèse de l'existence d'une interaction en $(t_1^\star,t_2^\star)$ d'effet $q=a/\sqrt{n}$: $m_{t_1^\star,t_2^\star}(t_1,t_2) =$

$$\frac{aE[X(t_1^\star)X(t_2^\star)\{2p(t_1,t_2)-1-E(2p(t_1,t_2)-1)\}]}{\sigma\sqrt{\mathrm{Var}\{2p(t_1,t_2)-1\}}}$$