区間マッピング法における影響分析

数理・推論研究系
特任研究員　　Dou Xiaoling

1　はじめに

遺伝学において, 個体の形質を規定する遺伝子 QTL (Quantitative trait locus) を探索するために QTL 解析という統計的手法が用いられる. この解析では, ロッドスコアとよばれる尤度関数の極大点を探索することによって QTL の位置を推測する (栗木 (2008), Siegmund and Yakir(2007)). QTL 解析に最も基本的なモデルの一つである区間マッピング法は観測されている遺伝子座 (マーカー) の情報を用いて観測されていない座のロッドスコアを推定することによりロッドスコア曲線を補間する手法である. 本稿では区間マッピング法におけるロッドスコア曲線の最大値及びその場所の推定値について影響分析を行い, 影響度の大きい個体 (グループ) を特定する.

2　区間マッピング法

簡単のために F_2 集団を対象とするモデルのみについて区間マッピング法の統計モデルを考える.

QTL がある位置に一つ存在して, それが形質に影響を与えるというモデルを考える. このような QTL のことを仮想 QTL という. 個体 t の仮想 QTL の遺伝子型を $z_*^{(t)}$ とおく. F_2 集団の遺伝子型は 3 値 (ホモ, ヘテロ, ホモ) をとり, $-1, 0, 1$ とおく.

$z_*^{(t)} \in \{-1, 0, 1\}$ は観測されない変数 (潜在変数) であるが, 同じ染色体上の座の遺伝子型は連鎖 (linkage) によって, 正の相関を持つ. このことから仮想 QTL の遺伝子型は, 近隣のマーカー遺伝子座の遺伝子型より推測することができる. いま仮想 QTL が γ に位置するとして, マーカーの遺伝子型 $z^{(t)} = (z_1^{(t)}, \ldots, z_m^{(t)})$ が与えられたときの $z_*^{(t)}$ の条件付確率を $P(z_*^{(t)}|z^{(t)}; \gamma)$ とおく.

区間マッピング法では, 次の統計モデルを仮定する (Lander and Botstein(1989))：

$$
\begin{aligned}
& z_*^{(t)} \sim P(z_*^{(t)}|z^{(t)}; \gamma), \\
& y^{(t)} = \alpha z_*^{(t)} + \beta w_*^{(t)} + \mu + \nu u^{(t)} + \varepsilon^{(t)}, \qquad \varepsilon^{(t)} \sim N(0, \sigma^2), \qquad t = 1, \ldots, n.
\end{aligned} \tag{2.1}
$$

ただし, $z_*^{(t)}$ に対して

$$
w_*^{(t)} = \begin{cases} 1 & (z_*^{(t)} = \pm 1), \\ -1 & (z_*^{(t)} = 0) \end{cases}
$$

のように定義する. $u^{(t)}$ は共変量である.

区間マッピング法のロッドスコアは, γ に位置する仮想 QTL の効果がないという帰無仮説 H_0 : $\alpha = \beta = 0$ に対する尤度比検定統計量

$$
\mathrm{LOD}(\gamma) = 2\left(\sum_{t=1}^{n} \log \sum_{k=-1}^{1} P(k|z^{(t)}; \gamma) f^{(k)}(y^{(t)}, \hat{\theta}(\gamma)) - \sum_{t=1}^{n} \log f^{(\dagger)}(y^{(t)}, \tilde{\theta}) \right) \tag{2.2}
$$

として定義される. ここで, 記号の簡単のため, $z_*^{(t)}$ の代わりに k を用いる. $P(k|z^{(t)}; \gamma)$ はホールデンの地図関数を仮定することによって求められる. $\hat{\theta}(\gamma)$ は γ が与えられたときの MLE, $f^{(k)}(\cdot, \hat{\theta})$ はパラメータ $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\nu}, \hat{\sigma}^2)$ を持つ正規分布 $N\left(\hat{\mu} + \hat{\nu}u + \hat{\alpha}k + \hat{\beta}w(k), \hat{\sigma}^2\right)$ の密度関数である.$\tilde{\theta}$ は H_0 の下での MLE, $f^{(\dagger)}(\cdot, \tilde{\theta})$ は正規分布 $N(\tilde{\mu} + \tilde{\nu}u, \tilde{\sigma}^2)$ の密度関数である.

3　ロッドスコアの最大値とその位置の影響関数

$(y^{(t)}, z^{(t)})_{t=1,\ldots,n}$ の経験分布関数を F_n とおく．$\hat{\theta}(\gamma)$ と $\tilde{\theta}$ も F_n の汎関数であるので，$\hat{\theta}(\gamma) = T(\gamma, F_n)$, $\tilde{\theta} = T^0(F_n)$ とおくと，$\mathrm{LOD}(\gamma)/2n$ が F_n の統計的汎関数として

$$\mathrm{FLOD}(\gamma) = \frac{\mathrm{LOD}(\gamma)}{2n} = \frac{1}{n}\sum_{t=1}^{n} \log \sum_{k=-1}^{1} P(k|z^{(t)};\gamma) f^{(k)}(y^{(t)}, \hat{\theta}(\gamma)) - \frac{1}{n}\sum_{t=1}^{n} \log f^{(\dagger)}(y^{(t)}, \tilde{\theta})$$

$$= L(\gamma, T(\gamma, F_n), F_n) - L^0(T^0(F_n), F_n) \tag{3.1}$$

ただし

$$L(\gamma, \theta, F) = \int \log \sum_{k=-1}^{1} P(k|z;\gamma) f^{(k)}(y, \theta) dF(y, z), \qquad L^0(\theta, F) = \int \log f^{(\dagger)}(y, \theta) dF(y)$$

のように記述できる．ロッドスコアを最大にする点 $\hat{\gamma} = G(F_n)$, $G(F) := \mathrm{argmax}_\gamma L(\gamma, T(\gamma, F), F)$ も F_n の汎関数である．y, z を固定して，$F_n^\epsilon = (1-\epsilon)F_n + \epsilon\delta_{(y,z)}$ とおく．$\left(\frac{d}{d\epsilon}\right)_0$ を $\epsilon = 0$ の点での ϵ の微分とすると，$\mathrm{FLOD}(\hat{\gamma})$ の経験影響関数 (Tanaka (1999)) は

$$\left(\frac{d}{d\epsilon}\right)_0 \mathrm{FLOD}\left(G(F_n^\epsilon)\right) = \left(\frac{d}{d\epsilon}\right)_0 L(\hat{\gamma}, T(\hat{\gamma}, F_n^\epsilon), F_n^\epsilon) - \left(\frac{d}{d\epsilon}\right)_0 L^0(T^0(F_n^\epsilon), F_n^\epsilon)$$

$$= \log \sum_k P(k|z;\hat{\gamma}) f^{(k)}(y, \hat{\theta}) - \log f^{(\dagger)}(y, \tilde{\theta}) - \mathrm{LOD}(\hat{\gamma}) \tag{3.2}$$

のように求められる．

また，ロッドスコアの最大値を与える点 $\hat{\gamma} = G(F_n)$ の経験影響関数

$$\left(\frac{d}{d\epsilon}\right)_0 G(F_n^\epsilon) = -\frac{L_\gamma(\hat{\gamma}, \hat{\theta}, \delta_{(y,z)}) - L_{\gamma\theta}(\hat{\gamma}, \hat{\theta}, F_n) L_{\theta\theta}(\hat{\gamma}, \hat{\theta}, F_n)^{-1} L_\theta(\hat{\gamma}, \hat{\theta}, \delta_{(y,z)})}{L_{\gamma\gamma}(\hat{\gamma}, \hat{\theta}, F_n) - L_{\gamma\theta}(\hat{\gamma}, \hat{\theta}, F_n) L_{\theta\theta}(\hat{\gamma}, \hat{\theta}, F_n)^{-1} L_{\theta\gamma}(\hat{\gamma}, \hat{\theta}, F_n)} \tag{3.3}$$

で与えられる．ここで，$L_\cdot$ と $L_{\cdot\cdot}$ は γ または θ に関する偏微分を表し，数値微分や微分公式で求められる．

4　データ解析

マウスの肥満の原因となる遺伝子を探索するため，肥満と関連性が深いとされる血中アディポネクチン濃度に着目し，標準的マウス近交系である B6 と，日本産亜種由来の MSM 系統の F_2 雑種 170 個体に対して解析を行う．

第 3 染色体上で観測された 6 箇所の遺伝子型について区間マッピング法を用いて計算されたロッドスコア曲線の最大値とその位置に着目する．(3.2) と (3.3) を $t = 1, \ldots, n$ について計算し，絶対値の大きな経験影響関数を持つ個体がロッドスコアの最大値またはその場所に対して大きな影響を与えると判断される．

参考文献

栗木 哲 (2008). QTL 解析の統計モデルと検定の多重性調整,「21 世紀の統計科学」, II, (小西 貞則, 国友 直人編), 東京大学出版会, 315–356.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121** (1), 185–199.

Siegmund, D. and Yakir, B. (2007). *The Statistics of Gene Mapping*, Springer, New York.

Tanaka, Y. (1999). Recent advance in sensitivity analysis in multivariate statistical methods, *Journal of the Japanese Society of Computational Statistics*, **7**, 1–25.