論 文 の 内 容 の 要 旨

生産・環境生物学専攻
平成12年度博士課程進学
氏名　　中道 礼一郎
指導教官　　岸野 洋久

論文題目　不完全マーカーにもとづく自殖性・他殖性生物のQTL解析手法の開発

生物の形質は、質的形質と量的形質に大別される。一般に、質的形質とは表現型が不連続で定性的に記述できる形質のことであり、量的形質とは表現型が連続的に変化し計数値や計量値によって記述される形質のことである。量的形質には、作物の収量（重量）や草丈（長さ）、開花期（時間）など、農学研究において重要な形質が多く、量的形質に関与する遺伝子座（quantitative trait locus, QTL）の解析は重要な課題となっている。量的形質は小さな遺伝子が多数関与していることが多いため、各遺伝子型間の差異は微小である。さらに、環境による連続的な変動が加わるため、表現型値から遺伝子型を推測することは通常の方法では不可能である。そのため、解析には統計遺伝学的手法を用いる必要がある。

一般的なQTL解析では、形質値の異なる2つの親系統P_1、P_2を交配して得られた後代を、ある遺伝子座について、P_1由来のホモ接合、P_2由来のホモ接合、ヘテロ接合の3つにグループ分けする。そして、グループ間差で遺伝子座と目的形質の連鎖および遺伝効果をはかる。そのために最も簡単な方法は純系の親系統を用いることである。この場合マーカーの遺伝子型がそのまま親系統の由来を表す。いま、M個のQTLの存在を仮定すると2つの純系親P_1、P_2間の交配から得られた個体の表現型値Yは以下の遺伝モデルで表される。

$$Y = \mu + \Sigma_{x=1}^{M} g_x + e \qquad (1)$$

ここで、μは遺伝子型によらない定数であり、eは環境効果である。環境効果は中心極限定理によって正規分布で近似される。g_xはx番目（$x = 1, 2, \cdots, M$）のQTLの遺伝効果であり、その値はQTLの遺伝子型に応じて相加効果a_xと優性効果d_xで定義される。このようなサンプル個体をN個体得たとき、式1のM個のQTLの仮定は以下の尤度で評価される。

$$\log_e L = \sum_{i=1}^{N} \log_e \left\{ \sum_{k_1=1}^{3} \sum_{k_2=1}^{3} \cdots \sum_{k_M=1}^{3} \left(\phi_{i,k_1 \cdots k_M} \cdot \prod_{x=1}^{M} P_{i,k_x} \right) \right\} \tag{2}$$

ここでk_x（$x = 1, 2, \cdots M$）はx番目のQTLの遺伝子型で、P_1由来の対立遺伝子をQ_x、P_2由来の対立遺伝子をq_xとするとQTLの遺伝子型Q_xQ_x、Q_xq_x、q_xq_xに対しk_x=1, 2, 3となる。P_{i,k_x}は、i番目（$i = 1, 2, \cdots N$）の個体のマーカー遺伝子型に対するx番目のQTL遺伝子型k_xの条件付き確率である。これはQTLと隣接マーカーの組換価から算出される。$\phi_{i,k_1 \cdots k_M}$は、i番目の個体の表現型値y_iのQTL遺伝子型に対する条件付き分布である。環境効果に正規分布を仮定していることから以下のように表される。

$$\phi_{i,k_1 \cdots k_M} = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{ -\left(y_i - \mu - \Sigma_{x=1}^{M} g_{x,k_x}\right)^2 / 2\sigma^2 \right\} \tag{3}$$

ここで、σ^2は環境分散である。g_{x,k_x}はx番目のQTLの遺伝効果であり、遺伝子型k_x=1, 2, 3に対してa_x、d_x、$-a_x$である。QTLの遺伝子型は直接観察できないので、マーカーの遺伝子型からQTL遺伝子型の条件付き確率P_{i,k_x}を求め、表現型値の条件付き分布$\phi_{i,k_1 \cdots k_M}$をP_{i,k_x}で重み付けして相加平均をとることで、式2の尤度としている。

このようにQTLの数と位置を仮定すれば、その仮定の下での遺伝効果を最尤法によって推定し、その仮定の善し悪しを尤度によって検定することができる。しかし、QTLの数と位置は事前に知ることはできないため、全てのQTLの可能性について検定するのは現実的でない。そこで、このような組み合わせ最適化問題に優れた手法として、遺伝的アルゴリズム(genetic algorithm, GA)による解析法を提案する。GAは、最適解を求めるのが困難だが、最適解の候補の良し悪しを評価することは可能である問題を、生物の進化のアナロジーによって解く手法である。GAにおいて、求めたい最適解の候補は仮想的な生物個体（GA個体）の「遺伝子型」としてコード化され、その最適解候補の評価はGA個体の「適応度」として表される。まず、はじめにランダムな「遺伝子型」をもつGA個体の集団を生成する。次に「適応度」に応じてGA個体を「淘汰」し、選ばれた「適応度」の高いGA個体の「遺伝子型」から次の世代のGA集団の「遺伝子型」を生成する。これをくり返すこと

でGA個体を「進化」させると、世代が進むに連れGA集団全体に良い「適応度」を持つGA個体が増えていく。最後に生き残った最も「適応度」の高いGA個体の「遺伝子型」、すなわち最も評価の高い解候補を最適解として採用する。QTL解析におけるGA遺伝子型はQTLの数と位置であり、GA適応度はQTLの数と位置の仮定に対する尤度である。

GAの最大の利点は実装の単純さである。そのため、複雑な問題を複雑なまま自由にモデル化して解析できる。より複雑なケースでのQTL解析として欠測値の問題を考察する。QTL解析はマーカー遺伝子型情報に依存しているが、実験の不手際でデータが欠落することもあれば、優性マーカーによって遺伝子型の一部が観察できないこともある。また、純系親の作成に際し、ホモ接合性が完全でなく、不完全分離・非分離マーカーを生じることもある。不完全なマーカー情報が得られた場合でも、それ以外のマーカーを同時に使用することで失われた情報を補うことができる。不完全マーカーの遺伝子型の条件付き確率は周辺のマーカーとの組換価から求められる。失われた遺伝子型情報はこの条件付き確率に従うランダムサンプリングによって決定される。不完全マーカーの遺伝子型は、他のマーカーから得た条件付き確率に従っているとはいえ、遺伝子型を決め打ちすることで、推定に多少のずれが生じているはずである。そこで、不完全マーカーの遺伝子型のランダムサンプリングは、GA個体ごと、GA世代ごとに全てやり直し推定値を更新する。通常、GA集団の個体数は数百で、数十世代の世代交代を繰り返すことから、不完全マーカーの遺伝子型は数千から数万のランダムサンプリングがなされることになる。これによって、GAによる推定全体では不完全マーカーの遺伝子型の偏りは解消され適切な推定がなされる。

これまでのQTL解析手法は遺伝モデルを単純化するため純系親由来の交配に依存してきた。一方、多くの他殖性生物では純系親の作成は困難である。そこで、GAによる柔軟なモデル構築を生かして純系親を用いない手法を提案する。いま、任意の自然集団から無作為にN_p個体の生物個体をとりだし、それら無作為交配して一交配あたりのN_o個体の子個体が得られたとき、M個のQTLを仮定した遺伝モデルとその尤度は以下のようなる。

$$y_i = \mu + \sum_{x=1}^{M}\left(a_{x,i,1} + a_{x,i,2}\right) + e \qquad y_{i_1,i_2,j} = \mu + \sum_{x=1}^{M}\left(a_{x,i_1,h_1} + a_{x,i_2,h_2}\right) + e \qquad (4)$$

$$\log_e L = \sum_{i=1}^{N_p}\log_e\left(\phi_i^p\right) \times \sum_{i_1=1}^{N_p-1}\sum_{i_2=i_1+1}^{N_p}\sum_{j=1}^{N_o}\log_e\left\{\sum_{k_1=1}^{4}\cdots\sum_{k_M=1}^{4}\left(\phi^o_{i_1,i_2,j,k_1\cdots k_M}\cdot\prod_{x=1}^{M}P_{i_1,i_2,j,k_x}\right)\right\} \qquad (5)$$

ここで、y_iはi番目（$i = 1, 2, \cdots N_p$）の親個体の表現型値で、$y_{i_1,i_2,j}$はi_1番目とi_2番目（i_1, i_2

$= 1, 2, \cdots N_p; i_1 \neq i_2$) の親個体の交配の j 番目 ($j = 1, 2, \cdots N_o$) の子個体の表現型値である。μ は遺伝子型によらない定数であり、e は正規分布に従う環境効果である。$a_{x,i,h}$ は i 番目の親個体の x 番目 ($x = 1, 2, \cdots, M$) の QTL の h 番目 ($h = 1, 2$) の対立遺伝子の遺伝効果である。h_1 と h_2 はそれぞれ、i_1 番目と i_2 番目の親個体から受け継いだ QTL 対立遺伝子の由来である。ϕ_i^p は、i 番目の親個体の表現型値 y_i の、QTL 対立遺伝子の遺伝効果に対する分布である。$\phi^o_{i_1,i_2,j,k_1 \cdots k_M}$ は、i_1 番目と i_2 番目の親個体の交配の j 番目の子個体の表現型値 $y_{i_1,i_2,j}$ の QTL 対立遺伝子の由来型 k_x に対する条件付き分布である。これらは、環境効果の正規性の仮定から式 3 同様、平均が定数 μ と遺伝効果の和で分散が環境分散に等しい正規分布となる。P_{i_1,i_2,j,k_x} は、i_1 番目と i_2 番目の親個体の交配の j 番目の子個体のマーカー遺伝子型に対する x 番目の QTL の対立遺伝子由来型 k_x の条件付き確率である。これは QTL と隣接マーカーの組換価から算出されるが、純系親を用いない交配ではマーカー遺伝子型情報は不完全である。他殖性自然集団ではマーカーがきれいに分離している保証はない。たとえきれいに分離しても、linkage phase、つまり対立遺伝子の組み合わせは観察できない。すなわち、親個体のマーカー遺伝子型が *AaBb* であっても、その半数体型が *AB* と *ab* であるのか *Ab* と *aB* であるのかは観察できない。つまり非純系親交配ではマーカーは全て不完全分離状態で、その由来情報は観察できない。しかしマーカー由来情報の条件付き確率は算出可能である。親個体の linkage phase の事後確率は子集団内の遺伝子型の分布から求められ、親の linkage phase が決定されれば子のマーカー由来情報の条件付き確率は周辺マーカーとの組換価から求められる。前述の純系交配での欠測値の問題と同様に、これらの事後確率と条件付き確率に従うランダムサンプリングでマーカー由来情報が決定される。ランダムサンプリングは GA 個体ごと、GA 世代ごとに全てやり直し、推定値を更新する。

シミュレーション実験でこれらの手法の有効性を確認したところ、従来の手法が対処できない状況においても効果的に QTL 検出がなされることが示された。QTL 解析では、純系親作成のコスト、サンプル飼育のコスト、マーカー開発のコスト、マーカーの共優性・優性による解析の容易さなどで実験の負荷が決まる。これらの要因は生物種によって異なるが、GA ではこれを考慮してその生物に最適な実験を設計できる。つまり、従来のように解析手法の制約にあわせて実験を設計するのではなく、実験者の都合にあわせて柔軟に解析手法を構築することが可能となった。