遺伝的アルゴリズムにおける適応度関数のエピスタシスについて

古　谷　博　史†

遺伝的アルゴリズム（GA）において問題の困難さと密接に関連した概念にエピスタシスがある．エピスタシスは，選択の染色体全体における効果がそれを構成する各遺伝子座への効果の足し算になっていない場合をいう．一般にはエピスタシスをもたない適応度関数は，GA にとって解くことが容易であると考えられている．そして適応度関数の中でも OneMax 問題に用いられているような線形関数は，エピスタシスをもたらさない代表的な例とされている．しかし実際の GA 計算においてこのことは必ずしも正しくなく，線形適応度関数もエピスタシスをもつ．本論文では，エピスタシスを連鎖不均衡と関連させて調べ，交叉の効果との関係について述べる．

On the Epistasis of Fitness Functions in Genetic Algorithms

Hiroshi Furutani†

In genetic algorithms (GAs), epistasis is considered to be closely related to the hardness of a problem. Epistasis means a condition of chromosome in which total effect of selection is not the sum of its effect on each allele. It is generally believed that fitness functions without epistasis are easy to solve in GAs. A linear fitness function such as the one used in One-Max problem is considered as a representative example of a function without epistasis. However in actual GA calculations, this is not necessarily true, and the linear function shows epistasis. The paper describes an investigation of epistasis in the viewpoint of linkage disequilibrium and its relation to crossover.

1. はじめに

遺伝的アルゴリズム（GA）において，与えられた問題の難しさを予測することは非常に重要である．すべての問題に対し GA が必ずしも万能というわけではなく，場合により問題の表現を変えた方がよいこともあれば，他の手法を採用しなければならないこともある．そこで何らかの方法で問題の困難さ（少なくとも GA にとっての）を予測したい，という要求がでてくる．問題の困難性に関係していると考えられる概念の一つにエピスタシス（epistasis）がある [1]．エピスタシスは遺伝学における用語であり，非相加的相互作用と訳され，選択の染色体全体における効果がそれを構成する各遺伝子座への効果の単純な積み上げからずれている状況をいう．

GA においては，適応度関数の線形性と関連させて議論することが多く，線形関数をエピスタシスのない適応度関数とし，エピスタシス解析の基準にしている．このことは一見自明なことのように思われるが（そして多くの研究者がそう考えているようだが），実は必ずしも正しくない．われわれは以前，代表的な線形適応度関数をもつ One-Max 問題の解析 [2),3)] を行ったときこのことに注目した．そして，すでにこのことは多くの文献で指摘されていた．本報告の目的の一つは，この点について研究者の注意を促すことにある．

この報告のもう一つの目的は，エピスタシスを通じた選択と交叉の相互関係について考察することにある．交叉のもっとも重要な役割は，エピスタシスを解消して各遺伝子座が独立に進化できるようにすることである．そしてそのエピスタシスを生成しているのは選択過程であるから，選択と交叉はエピスタシスを通じて互いに関連しあっているといってよい．どのようなエピスタシスがどの程度生成されるかという問題は，適応度関数の形に依存する．そこで本報告では，適応度関数の形とエピスタシスの関連について考察し，それに対する交叉の役割について述べる．

2. エピスタシス

2.1 エピスタシスと連鎖不平衡

エピスタシスの程度を定量的に示す方法として連鎖不平衡 (linkage disequilibrium)[4] がある．これは遺

† 京都教育大学教育学部
Faculty of Education, Kyoto University of Education

伝子の頻度分布に注目して定義される．

いま 2 つの遺伝子座をもつ生物を考え，各遺伝子座は 2 種類の遺伝子をもつものとする．これは，GA では $\ell = 2$ ビット系のシステムに対応している．したがって $n = 2^\ell = 4$ の遺伝子型をもつ．遺伝子型の種類を $i =< i_1, i_0 >= 0, 1, 2, 3$ とし，i_1，i_0 で各遺伝子座の遺伝子を表すことにする．また各遺伝子座（ビット位置）を $b = 0, 1$ で表す．集団内での遺伝子型 i の相対頻度を x_i とすると $0 \leq x_i \leq 1$ であり，規格化の条件を満たす．

$$\sum_{i=0}^{n-1} x_i \ = \ 1.$$

集団が，遺伝子座 b においてビット値 m をとる相対頻度を $P_b(m)$ と表すことにする．また，遺伝子座 b，b' においてビット値 m，m' をとる相対頻度を $P_{bb'}(m, m')$ とする．連鎖不平衡を定量的に表すために 2 種類の量がよく用いられる．もっともよく使われているのは D 係数で

$$\begin{aligned} D &= P_{bb'}(0,0)P_{bb'}(1,1) \ - \ P_{bb'}(0,1)P_{bb'}(1,0) \\ &= x_0 x_3 - x_1 x_2, \end{aligned}$$

と定義される．しかし，この定義よりむしろ直感的な理解のためにはつぎの同値な定義のほうがふさわしい．

$$D[b, b'] \ = \ P_{bb'}(0,0) \ - \ P_b(0)P_{b'}(0). \qquad (1)$$

2 つの遺伝子座が独立で互いに相関がなければ $P_{bb'}(0,0) \ = \ P_b(0)P_{b'}(0)$ となり，$D[b, b'] \ = \ 0$ が成り立つ．

もう 1 つの定義は Z 係数によるもので

$$Z \ = \ \ln \frac{P_{bb'}(0,0)P_{bb'}(1,1)}{P_{bb'}(0,1)P_{bb'}(1,0)}, \qquad (2)$$

ここですべての $P_{bb'}(m, m')$ が 0 でないことを仮定する．$D = 0$ ならば $Z = 0$ であり，それ以外の場合も両者は同じ符号をもつことに注意して欲しい．

2.2 エピスタシスと選択

エピスタシスの原因として選択（淘汰）が挙げられる．したがって，エピスタシスについては，選択の役割を中心に研究が行われている．ここでは，Felsenstein の研究 [5] を中心に紹介する．

先ほど示した 2 ビットの集団を例にとる．突然変異と交叉を無視し，選択の過程だけ考えるものとする．選択の効果を理論的に解析するためつぎの 2 つのモデルを考える．

- 連続時間モデル
- 離散時間モデル

連続時間モデルは GA の進化を連続的と考えてモデル化したもので，離散時間モデルは段階的進化を仮定した GA モデルである．GA の応用面では，世代ごとに集団の個体が入れ替わる世代的 GA が採用されることが多く，離散時間モデルによる進化の記述のほうがふさわしいと思われる．

連続時間モデルによる選択過程の表現は木村により連立微分方程式の形で与えられた．

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} \ = \ (f_i - \bar{f}(t))x_i(t), \qquad (3)$$

ここで f_i は遺伝型 i の適応度で，$f_i > 0$ とする．また $\bar{f}(t)$ は世代 t における集団の平均適応度を表す．

$$\bar{f}(t) = \sum_{i=0}^{n-1} f_i x_i(t).$$

いま 2 ビットの系を考え，その適応度関数を

$f_0 = 1$
$f_1 = 1 + \alpha$
$f_2 = 1 + \beta$
$f_3 = 1 + \alpha + \beta + E$

とし，f_3 が最大の適応度をもつものとする．線形の適応度では $E = 0$ であり，E はエピスタシスの程度を表している．この系の連鎖不平衡を表す Z 係数の世代による変化 $\mathrm{d}Z/\mathrm{d}t$ は，進化方程式 (3) から，

$$\frac{\mathrm{d}Z(t)}{\mathrm{d}t} \ = f_0 + f_3 - f_1 - f_2 \ = \ E, \qquad (4)$$

で与えられる．初期状態 $(t = 0)$ において，系が連鎖平衡状態 $(Z = 0)$ にあったとすると，$E = 0$ ならばすべての世代において系は $Z(t) = 0$ となり，常に連鎖平衡状態にある．したがって，連続時間モデルにおいてエピスタシスのなくなる条件は

$$E \ = \ 0 \qquad (5)$$

であり，われわれの常識とよく一致する．ここで注意して欲しいことは，$E > 0$ なら Z は（そして D も）正の値をとって増加していき，$E < 0$ ならその逆の結果になる．

つぎに離散時間モデルにおける選択の効果について考察する．進化方程式は連立差分方程式

$$x_i(t+1) \ = \ \frac{f_i x_i(t)}{\bar{f}(t)}, \qquad (6)$$

で与えられる．2 ビット系における適応度関数は

$f_0 = 1$
$f_1 = 1 + \alpha$
$f_2 = 1 + \beta$
$f_3 = (1 + \alpha)(1 + \beta)\gamma,$

とパラメータ化する．この系での連鎖不平衡の変化 $\Delta Z(t) \equiv Z(t+1) - Z(t)$ は $\mathcal{E} = \ln\gamma$ として

$$\Delta Z(t) = \ln f_0 + \ln f_3 - \ln f_1 - \ln f_2 = \mathcal{E}, \ (7)$$

となる．先ほどの議論と同様に，初期状態が連鎖平衡 $Z(0)=0$ で，適応度が $\mathcal{E}=0$ ならば集団はエピスタシスのない状態 $Z(t)=D(t)=0$ になる．また，$\mathcal{E}>0$ ならば $Z(t)$ は正の値をとりつつ増加していく．逆に $\mathcal{E}<0$ ならば負の値をとって減少する．このことの意味は交叉の役割と関連させて後で議論する．

2.3 One-Max 問題のエピスタシス

従来，One-Max 問題はエピスタシスをもたない適応度の代表例として多くの研究がされてきた．One-Max 問題の適応度は

$$f_i = |i| = \sum_{b=1}^{\ell-1} i_b, \qquad (8)$$
$$i = < i_\ell, i_{\ell-1}, \ldots, i_1 > .$$

と定義され，ビット列 i の中のビット 1 の個数に等しい．選択のみの進化方程式 (6) の解から [1] 連鎖不平衡係数の具体的な形を導くことができる．初期条件として一様な分布（$x_0(0) = \ldots = x_{n-1}(0)$）を仮定すれば

- $\ell = 2$ の場合
$$D[1,2] \;=\; -1/(2+2^t)$$
- $\ell = 3$ の場合
$$D[b,b'] \;=\; -(1+2^t-3^t+4^t)/(3+3\cdot 2^t+3^t)$$

と求めることができ，常に負の値をとることがわかる．このように One-Max 問題の適応度関数はエピスタシスをもち，負の連鎖不平衡を生成する [2),3)]．

実際，世代的 GA において連鎖不平衡をもたらさないことがわかっているのは積型適応度関数であり [1]，2 ビット系の解析結果と一致する．

3. 交叉の役割

3.1 交叉と連鎖不平衡係数

生物にとって（また GA にとって）なぜ交叉が必要か，という疑問に答えるためには，エピスタシスの働きを研究することが大切である．遺伝における交叉の効果を調べるために種々の方法が提案されているが，ここでは連鎖不平衡係数（D 係数）に対する効果をみていくことにする．一点交叉や一様交叉による D 係数の変化は解析的に計算することができ，交叉率を $\chi = 1$ とした場合

- 一点交叉
$$D(t+1) \;=\; (1-\frac{b'-b}{\ell-1})D(t)$$
- 一様交叉
$$D(t+1) \;=\; \frac{1}{2}D(t)$$

で与えられる [1),2)]．いずれにしろ交叉は D 係数の絶対値を小さくするよう働く．それではこのことは集団にとってどのような意味があるのであろうか．それに対する答えは

(1)　交叉の直接的効果
(2)　交叉の間接的効果

の 2 つがある．直接的効果は，交叉により集団内の分布が直接変化することをいい，間接的効果は，選択の過程を通じて分布が変化することをいう．

3.2 交叉の直接的効果

交叉の直接的効果について 2 ビット系を例に説明する．D 係数の定義から

$$x_0 = P_b(0)P_{b'}(0) + D$$
$$x_1 = P_b(0)P_{b'}(1) - D$$
$$x_2 = P_b(1)P_{b'}(0) - D$$
$$x_3 = P_b(1)P_{b'}(1) + D$$

となる．交叉は，P_b について直接にはなんら効果を及ぼさない（後で述べるように間接的には影響を与える）．交叉の効果があるのは D 係数の部分である．

交叉による効果は D 係数の符号に依存する．いま f_3 が最大で $i=3$ が最適解になるものとする．もし $D<0$ ならば D は増加し，結果として x_3 も増加する．したがってこの場合，交叉はわれわれにとって望ましい働きをする．逆に $D>0$ ならば交叉によって D は減少し，また x_3 減少してしまう．このように，交叉が有益なオペレータになるか有害な働きをするか，ということはその時点での集団の状態（連鎖不平衡）によって決まる．

任意のビット数 ℓ をもつ一般の系について同じ問題を検討するためには，もう少し精密な定式化が必要になる．連鎖不平衡係数についても 2 つの遺伝子座に関連して定義した（2 次の係数）が，より詳しい議論のためには m 次の連鎖不均衡係数を定義し，交叉のそれらに対する効果を知る必要がある．m 次の連鎖不平衡係数に対する交叉の効果の解析的な表現を得ることは可能だが，少し複雑になるのでここでは省略する．しかしいずれにしろ，交叉の効果の効果を予測する上で連鎖不平衡係数（とくにその符号）との関係を知ることが重要なことは理解できるであろう．

3.3 交叉の間接的効果

交叉のもう一つの効果は選択を通じて間接的に現れる．そのため，遺伝学者 Fisher が導いた「自然選択の基本定理」をもとに進化速度と集団の分散の関係に注目する [6)]．

突然変異と交叉を無視し，比例選択を適用すると次式を導くことができる [1)]．

$$\Delta\bar{f}(t) = \frac{1}{\bar{f}(t)}\mathrm{VAR}(f), \tag{9}$$
$$\mathrm{VAR}(f) = \sum_i f_i^2 x_i(t) - \bar{f}(t)^2.$$

ここで $\Delta\bar{f}(t) = \bar{f}(t+1) - \bar{f}(t)$ は世代あたりの平均適応度の変化を表し，$\mathrm{VAR}(f)$ は適応度の分散である．$\Delta\bar{f}$ を進化速度の指標と考えれば，進化のためには適応度の分散を大きくすることが重要であることがわかる．

もう一つの重要な指標として，個体の間の Hamming 距離の分布がある．いま，$i = 0 =< 0, \ldots, 0 >$ を基準にとると Hamming 距離の平均 $E\{|i|\}$ は次式で与えられる．

$$E\{|i|\} = \sum_{b=1}^{\ell} P_b(1). \tag{10}$$

同様にして分散は，

$$V\{|i|\} = V_A + V_I, \tag{11}$$

で与えられる．ここで分散 $V\{|i|\}$ は

$$V_A = \sum_{b=1}^{\ell} P_b(1)\,(1 - P_b(1)), \tag{12}$$

$$V_I = 2\sum_{b<b'} D[b, b']. \tag{13}$$

と 2 つの項に分離することができる．相加的分散 V_A は常に正だが，エピスタシス分散 V_I は正負いずれの符号もとりうることに注意してほしい．Hamming 距離に関する 3 つの統計量 $E\{|i|\}$，V_A，V_I のうち交叉の影響を受けるのは V_I のみである．交叉は V_I の絶対値を減少させるので，$V_I < 0$ の場合は交叉により Hamming 距離の分散 $V\{|i|\}$ は増加する．

GA では解集団の分散は大きいほどいい，といわれている．このことは Fisher の定理から理解でき，適応度の分散が大きいほど進化速度は速くなる．そして，Hamming 距離の分散が大きければ一般に適応度の分散も大きいと期待される．One-Max 関数の場合は適応度が基準解からの Hamming 距離に等しいので，このことがそのままあてはまる．そして One-Max 問題では負の連鎖不均衡を生成するので，結果的に交叉が遺伝にとって有益な演算となる．そして，このことはかなり一般的に成り立つと考えられる．間接的効果についても直接的効果の場合と同様に，連鎖不平衡が負になることが重要である．

4. おわりに

GA の理論的基礎付けに関する研究は様々な面から進められているが，一つの中心的テーマが適応度関数の解析である．その中で，エピスタシスの概念は非常に重要な視点を与えている．直感的には線形の適応度関数

$$f_i = \sum_{b=1}^{\ell-1} \alpha_b\, i_b + \mathrm{const.}$$

が文字どおりエピスタシスのない形をしており，多くの研究者が線形適応度関数を第 0 近似として解析の出発点にしてきた．

Reeves らは実験計画法を用いた適応度関数 f_i の解析法を提案した [7]．彼らの方法は，適応度関数をエピスタシスの度合いに応じて展開し，その係数から問題の困難さを予測しようとするものである．しかし，その中でもっともエピスタシスの程度が低いとされているのは線形の項である．彼らも線形項がエピスタシスをもつことには気がついていない．この問題に対する解答は非常に簡単で

$f_i \to \ln f_i$

と適応度を変換して同じ解析を進めることである．

このような例は GA の理論的研究において非常に多く見ることができ，ここに示したものと同様な改良が可能であると考える．

参 考 文 献

1) 古谷博史: “遺伝的アルゴリズムにおける交叉の Walsh 解析,” 情報処理学会論文誌, **42** (2001) 掲載予定.
2) H. Furutani: “Study of Crossover in One Max Problem by Linkage Analysis,” L. Specter et al. (Eds.) Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-2001 Morgan Kaufmann, (2001) 320–327.
3) 古谷博史, 藤林由紀, 村田真知子: “遺伝的アルゴリズムにおける交叉の役割の連鎖解析,” 情報処理学会研究会報告, **2001-MPS-33** (2001) 53–56.
4) J. Maynard Smith: Evolutionary Genetics, Oxford University Press, Oxford (1998).
5) J. Felsenstein: “The Effect of Linkage on Directional Selection,” *Genetics*, **52** (1965) 349–363.
6) R.A. Fisher: The Genetical Theory of Natural selection, 2nd edition, Dover, New York (1958).
7) C. Reeves and C. Wright: “An Experimental Design Perspective on Genetic Algorithms,” In L.D. Whitley and M.D. Vose (Eds.) *Foundations of genetic algorithms 3*, Morgan Kaufmann (1995) 7–22.