統計数理(2012)
第 60 巻 第 2 号 341–352

特集「多様性と進化の統計解析」
[研究詳解]

小分散漸近理論を用いた集団遺伝学的研究

三浦 千明†

(受付 2012 年 1 月 5 日；改訂 10 月 18 日；採択 10 月 22 日)

要　　旨

本稿では，集団遺伝学の概要について解説すると共に，集団内に生じた突然変異体の頻度の推移的な変化を解析的に表示する為に，小分散漸近理論を応用した著者自身による研究を紹介する．これによって今まで解析的な表示がなかった複雑なモデルに対しても近似的な公式が与えられるようになった．集団遺伝のモデルは台が常に有界なのに対して，漸近展開は正規分布に沿った形で与えられる．それゆえ突然変異率が低い場合に，近似公式はうまく働かない．しかし時間があまり経過していない場合や突然変異率が高い場合には，近似公式はよく機能する事が確認できた．

キーワード： 小分散漸近理論，頻度分布，正規近似．

1. はじめに

1.1 集団遺伝学とは

集団遺伝学とは，メンデル集団という互いに交配によって結ばれた単一の生物の集団の中に生まれた突然変異の挙動をモデル化し，解析及びデータの解釈を行う学問である．集団内に生じた突然変異の本体は生殖細胞のゲノム内に起こる塩基の変異である．1 塩基のみの変異からゲノム自体の倍加変異までさまざまな規模で起こりうるが，変異がゲノムに刻まれているという事実は変わらない．突然変異はセレクションや集団構造の変化(分集団化，また分集団間の移住，急激な人口増加やボトルネックなど)，またゲノム内の他の領域との組み換えなど，さまざま圧力を受ける．これら圧力の影響を表す指標として，この突然変異体の集団内での割合(頻度)がしばしば用いられる．もう 1 つ重要で理解しやすい指標として，集団からサンプルを取ってきた時の突然変異体の個数があるが，今回は話題にしない．集団に突然変異が生じたとき，勿論それは集団内に 1 つだけ存在する．ヒトのように 2 倍体であれば，その生き物はゲノムをペアで持っているので，集団の大きさが N だとすると当然変異体の頻度は $1/2N$ となる．ここからスタートし，突然変異は上述の影響の結果，その頻度を増したりまたは集団内から消失したりする．

ここで集団遺伝学上重要な仮定は，このメンデル集団が任意交配によって世代を重ねていくという仮定である．勿論実際には集団は有限であるので近親婚は起きうるし，またペア同士の選り好みなどで同類婚なども起こりうる．その場合は頻度変化の分散で規準化するなどして，任意交配をしている有効数をその集団の数とするのでやはり仮定は成り立つ．この任意交配の仮定の下で，N 個体いる現在の世代から遺伝子プールが形成され，そこから次の世代のゲノム

† 統計数理研究所：〒190–8562 東京都立川市緑町 10–3 (現 明治大学 先端数理科学インスティテュート：〒214–8571 神奈川県川崎市多摩区東三田 1–1–1)

がランダムな 2 項選択によって再び $2N$ 個抽出され，N 個体の集団となる．ここで遺伝子プールとは，現在の世代の生殖細胞から放たれた配偶子たちが作る巨大なゲノムのプールである．実際例えばヒトであれば，男性の配偶子である精子は 1 人当たり生涯に 10 の 12 乗のオーダーでプールに供給され，女性の配偶子である卵子も(精子よりは大分少ないが)1 人当たり 10 の 2 乗のオーダーで供給される．このように巨大なプールの為，その中に含まれるゲノムの数は実質無限大とみなしてよく，任意交配の仮定からランダムな 2 項選択が行われるのである．(ここではヒトで考えると精子の数と卵子の数で差がありすぎるように感じるかもしれない，しかし実際モデリングはほとんどの場合雌雄同体と仮定し，雌性配偶子も雄性配偶子も同じように混ぜて考える．またデータの解析にあたってこのように仮定しても特に問題はないと考えられている．) このように，集団に生じた突然変異体の運命にはランダムネスが生じる事になる．このことを遺伝的浮動(genetic drift)という．

1.2 モデリング

任意交配のランダムネスによって生じる頻度の浮動は，上述の説明から 1 世代前の頻度にのみ依存している．即ち遺伝子プール内の突然変異型のゲノムの頻度のその祖先型(野生型)のゲノム頻度は，そのプールを生んだ世代の中のそれらの割合をそのまま反映するからである．よってこのモデル(Wright-Fisher モデルという)はマルコフ連鎖によって記述される．(再びヒトについて考えると，ヒトは Wright-Fisher モデルの様に一斉に遺伝子プールを作ったりせず，齢構造が存在する．しかし頻度の変化のタイムスパンはヒトの個人の人生の長さに較べて非常に大きいので，このようにマルコフ性を仮定してもうまく機能することが分かっている(Charlesworth, 1994)．)

このマルコフ連鎖は，時間と頻度に対する適当な尺度の変更を施すことで拡散過程によって近似できる．これを拡散近似という．具体的には，集団の大きさ N を無限大にしつつも，$2N$ 世代でスケーリングした単位時間間隔は有限に保たれ，なおかつ突然変異体の頻度は(消失する時以外は)ある値 $x \in (0,1]$ を取るように操作された極限での近似である．この時，時刻 0 での頻度の初期値を x_0 とすると，母集団の中の突然変異体の頻度は

$$x_t = x_0 + \int_0^t \mu(x_s)ds + \int_0^t \sigma(x_s)dB_s \tag{1.1}$$

と表される．ここで B_s はブラウン運動を表す．またやや混乱する言葉使いだが，集団遺伝学におけるランダムネスの原動力である遺伝的浮動を genetic drift といったが，これは拡散項に関する現象で，ドリフト項には関係がない．

具体的な例として，毎単位時間に一定の割合で同じ突然変異が起き続けるモデルを考える．即ち祖先型と，そこから突然変異によって生じた派生型のどちらかの状態しか取らないとする．またいったん派生型になったものからも，一定の割合で野生型に戻るという再起突然変異も起きているとする．即ちこのモデルでの突然変異とは，祖先型から派生型への突然変異と，派生型から祖先型への突然変異の相互方向への圧力の事を指す．派生型の頻度を x_t とする．祖先型から派生型への単位時間当たりの突然変異率を v とし，逆に派生型から祖先型への単位時間当たりの再起突然変異率を u とする．この拡散過程は

$$x_t = x_0 + \int_0^t \{v - (u+v)x_s\}ds + \int_0^t \sqrt{x_s(1-x_s)}dB_s \tag{1.2}$$

と表される．この拡散過程に対応する生成作用素は

$$L = \{v - (u+v)x\}\frac{d}{dx} + \frac{\sqrt{x(1-x)}}{2}\frac{d^2}{dx^2} \tag{1.3}$$

であり，定常状態における reversible な測度は，Beta 分布

$$\frac{\Gamma(2u+2v)}{\Gamma(2u)\Gamma(2v)}x^{2v-1}(1-x)^{2u-1}dx \tag{1.4}$$

となる(Wright, 1931; Feng, 2010)．このように定常状態における頻度の分布は比較的に簡単に求める事ができる．特に今の例のように変異がある座位がゲノムの他の領域と組み換えなどのインタラクションがない，すなわち独立した座位である場合，その座位の定常状態における頻度分布は，セレクションなどのさまざまな圧力を受けている場合でも解析的な解を求める統一的な方法がかなり昔から知られている(Wright, 1931, 1937)．また組み換えなどのインタラクションが存在する場合は問題が急に複雑になるが，それでも定常状態においては，セレクションがない場合のモーメントを求める方法として，Kimura-Ohta 法という一般的な方法が知られている(Ohta and Kimura, 1969a)．またセレクションがある場合の幾らかの性質なども同じく Ohta and Kimura (1969b)などで解析されている．加えて，もし連鎖している座位同士の組み換え率が非常に低いような場合には，定常状態に至るとそれぞれ完全に独立とみなしてよいことも数学的に証明されている(Ethier and Nagylaki, 1989)．

一方で頻度が時間に依存する推移的な状態の場合には，一般的にその分布を解析的に表現するのは定常状態に較べてかなり難しい問題になる．それでも拡散過程に対応する偏微分方程式が変数分離法などで解ける場合は，分布の厳密な解析解は求められている(Kimura, 1955a, 1955b; Crow and Kimura, 1956)．しかし推移的な状態の分布を解析的に表現する統一的な方法は今まで開発されてこなかった．特に組み換えのある場合については，推移分布の密度関数の解析的表現はほどんど手つかずのままであった．

そこで本稿では，Miura (2011)によって提案された小分散漸近理論を利用した遺伝子頻度の推移分布の解析的表現を得る方法について紹介する．また実際に得られる解析的な表現は近似的であるので，何らかの形で真の分布との比較が必要である．Miura (2011)では組み換えがある場合について近似表現とシミュレーションとの比較を実行しているが，厳密な解析解が存在する場合の比較は扱っていない．本稿では厳密な解析解が存在する式(1.2)で表されるモデルの場合における比較についても例を示した．

2. 小分散漸近理論について

小分散漸近理論の応用の先行研究は，数理ファイナンスにおける派生証券価格の漸近展開における評価などである(Yoshida, 1992; Takahashi, 1999; Kunitomo and Takahashi, 2001; 国友・高橋, 2003)．今回の(集団遺伝学での)応用では，分布の形のみを念頭に置いているので，それらファイナンスでの応用よりもやや単純な議論で済む．

さて，小分散漸近理論では以下の ϵ についての微小拡散モデルを考える．

$$x_t^{(\epsilon)} = x_0 + \int_0^t \mu(x_s^{(\epsilon)})ds + \epsilon\int_0^t \sigma(x_s^{(\epsilon)})dB_s \quad (0<\epsilon\leq 1) \tag{2.1}$$

微小拡散モデルとは，式(2.1)の拡散過程における観測時間を有限の区間 $[0, \tau]$ に固定するかわりに，拡散係数が 0 に収束する($\epsilon\to 0$)仮定の下での漸近理論である(西山, 2011)．ここでは Takahashi (1999)や国友・高橋(2003)などを参考に，特に数学的な正当化にこだわらず，具体的な計算の手順を示す．

式(2.1)は ϵ に関して以下のように形式的な展開が可能である．

$$x_t^{(\epsilon)} = x_t^{(0)} + \epsilon g_{1t} + \epsilon^2 g_{2t} + \epsilon^3 g_{3t} + \cdots \tag{2.2}$$

ここで

$$g_{it} = \frac{\epsilon}{i!}\frac{\partial^i}{\partial \epsilon^i} x_t^{(\epsilon)}\Big|_{\epsilon=0} \quad (i=1,2,3,\dots).$$

今規準化のために

$$y_t^{(\epsilon)} \stackrel{\text{def}}{=} \frac{x_t^{(\epsilon)} - x_t^{(0)}}{\epsilon} \tag{2.3}$$

と置くと

$$y_t^{(\epsilon)} = g_{1t} + \epsilon g_{2t} + \epsilon^2 g_{3t} + \cdots \tag{2.4}$$

となる．式(2.2)は式(2.1)について単純に形式的な展開を施したものであったが，実際に然るべき条件の下では，漸近展開(2.4)が適当な意味で正当化される(Watanabe, 1987, Theorem 2.2)．この時 g_{1t} は定義から Wiener 積分(非確率関数のブラウン運動に沿った積分)になる事がわかり，従って正規分布に従う．また集団遺伝学のモデルに関してもその条件が満たされ，数学的な正しさは保証されると考えられる．ここで $y_t^{(\epsilon)}$ の特性関数を考えると

$$\begin{aligned}
\psi(\xi) &= E[\exp(i\xi y_t^{(\epsilon)})] \\
&= E[\exp\{i\xi(g_{1t} + \epsilon g_{2t} + \epsilon^2 g_{3t} + \cdots)\}] \\
&= E[\exp(i\xi g_{1t})\exp(i\xi\epsilon g_{2t})\exp(i\xi\epsilon^2 g_{3t})\cdots] \\
&= E\left[\exp(i\xi g_{1t})\left\{1 + \epsilon i\xi g_{2t} + \epsilon^2 i\xi g_{3t} + \frac{\epsilon^2 i^2}{2}(\xi g_{2t})^2 + \cdots\right\}\right] \\
&= E[\exp(i\xi g_{1t})] + \epsilon i E[\exp(i\xi g_{1t})(\xi g_{2t})] + \epsilon^2 i E[\exp(i\xi g_{1t})(\xi g_{3t})] \\
&\quad + \frac{\epsilon^2 i^2}{2} E[\exp(i\xi g_{1t})(\xi g_{2t})^2] + \cdots \\
&= \exp\left\{\frac{i^2\xi \mathrm{cov}(g_{1t})\xi}{2}\right\} + \epsilon i E[\exp(i\xi g_{1t})E[\xi g_{2t}|g_{1t}]] + \epsilon^2 i E[\exp(i\xi g_{1t})E[\xi g_{3t}|g_{1t}]] \\
&\quad + \frac{\epsilon^2 i^2}{2} E[\exp(i\xi g_{1t})E[(\xi g_{2t})^2|g_{1t}]] + \cdots .
\end{aligned} \tag{2.5}$$

ここで反転公式を用いれば $y_t^{(\epsilon)}$ の密度関数は以下の様な表現を持つ．

$$\begin{aligned}
\phi_{y_t^{(\epsilon)}}(x) &= Norm(x; 0, \mathrm{cov}(g_{1t})) \\
&\quad + \epsilon\left[-\sum_{j=1}^{d}\frac{\partial}{\partial x_j}\left\{E[g_{2t}|g_{1t}=x]Norm(x;0,\mathrm{cov}(g_{1t}))\right\}\right] \\
&\quad + \epsilon^2\left[-\sum_{j=1}^{d}\frac{\partial}{\partial x_j}\left\{E[g_{3t}|g_{1t}=x]Norm(x;0,\mathrm{cov}(g_{1t}))\right\}\right. \\
&\qquad \left. + \frac{1}{2}\sum_{j,k=1}^{d}\frac{\partial^2}{\partial x_j \partial x_k}\left\{E[g_{2t}^2|g_{1t}=x]Norm(x;0,\mathrm{cov}(g_{1t}))\right\}\right] + \cdots
\end{aligned} \tag{2.6}$$

ここで $Norm(x;\ ,\ )$ は正規分布の密度関数，d は x の次元を表す．最後に式(2.3)における $y_t^{(\epsilon)}$ の定義から，非確率的な関数 $x_t^{(0)}$ によって平行移動し，また次元 d によるスケール変換を施すことで，$x_t^{(\epsilon)}$ の密度関数は

$$\phi_{x_t^{(\epsilon)}}(x) = \frac{1}{\epsilon^d}\phi_{y_t^{(\epsilon)}}\left(\frac{x - x_t^{(0)}}{\epsilon}\right) \tag{2.7}$$

のように得られる．

式(2.1)から分かるように $x_t^{(\epsilon)}$ は $\epsilon \to 0$ につれて拡散の影響がなくなっていき，非確率的な関数 $x_t^{(0)}$ に近づいていくが，実際のモデルは式(2.1)の定義から $\epsilon = 1$ のとき $x_t^{(1)} = x_t$ である．従って次の章では x_t の真の分布に対する近似公式として，式(2.7)に $\epsilon = 1$ を代入し，展開を適当な所で打ち切った関数を採用して，その有効性を検討する．

3. 例

3.1 式(1.2)の場合

問題のセッティングは 1.2 節で述べた通りである．この問題の場合，式(1.2)に対応する偏微分方程式は変数分離法で解け(Crow and Kimura, 1956)，x_t の分布の密度関数は

$$\begin{aligned}(3.1)\qquad &\sum_{i=0}^{\infty} x^{2v-1}(1-x)^{2u-1} \\ &\times F(2(u+v)-1+i,-i,2u,1-x)F(2(u+v)-1+i,-i,2u,1-x_0) \\ &\times \frac{\Gamma(2u+i)\Gamma(2(u+v+i))\Gamma(2(u+v)+i-1)}{i!\Gamma(2u)^2\Gamma(2v+i)\Gamma(2(u+v+i)-1)} \exp\left\{-i\left(u+v\frac{i-1}{2}\right)t\right\}\end{aligned}$$

と表される．ここで $F(\ ,\ ,\ ,\ )$ はガウスの超幾何関数．$t \to \infty$ とすれば $i = 0$ 以外の項は消え，また初期値によらず Beta 分布(1.4)に収束する(Gale, 1990)．

一方で漸近展開による近似公式として，今第 2 項目までを用いた物を採用する．その為には $x_t^{(0)}$，$\mathrm{cov}(g_{1t})$ 及び $E[g_{2t}|g_{1t}=x]$ を求める必要がある．まず $x_t^{(0)}$ は $\epsilon = 0$ とおいた時の非確率的な関数であり，今の場合は非常に単純に

$$(3.2)\qquad x_t^{(0)} = \left(x_0 - \frac{v}{u+v}\right)e^{-(u+v)t} + \frac{v}{u+v}$$

と求まる．$\mathrm{cov}(g_{1t})$ はこの場合 1 次元なので単なる分散であり，g_{1t} も展開(2.2)から

$$(3.3)\qquad g_{1t} = \int_0^t \exp\{-(u+v)(t-s)\}\sqrt{x_s^{(0)}(1-x_s^{(0)})}dB_s$$

となる．よって

$$\begin{aligned}(3.4)\qquad \mathrm{var}(g_{1t}) &= E\left[\left(\int_0^t \exp\{-(u+v)(t-s)\}\sqrt{x_s^{(0)}(1-x_s^{(0)})}dB_s\right)^2\right] \\ &= \int_0^t \exp\{-2(u+v)(t-s)\}x_s^{(0)}(1-x_s^{(0)})ds.\end{aligned}$$

最後に $E[g_{2t}|g_{1t}]$ であるが，これは Takahashi (1999) の式(2.14)から

$$(3.5)\qquad E[g_{2t}|g_{1t}=x] = \alpha(x^2 - \mathrm{var}(g_{1t}))$$

$$\alpha = (\mathrm{var}(g_{1t}))^{-2}\int_0^t \left(\frac{1}{2} - x_s^{(0)}\right)\exp\{-(u+v)(t-s)\}\int_0^s \exp\{-2(u+v)(t-w)\}x_w^{(0)}(1-x_w^{(0)})dwds$$

と導かれる．これらを用いると近似公式は Takahashi (1999) の式(2.19)から以下のようになる．

$$(3.6)\qquad \left\{1 - 3\alpha(x - x_t^{(0)}) + \frac{\alpha}{\mathrm{var}(g_{1t})}(x - x_t^{(0)})^3\right\}Norm(x; x_t^{(0)}, \mathrm{var}(g_{1t}))$$

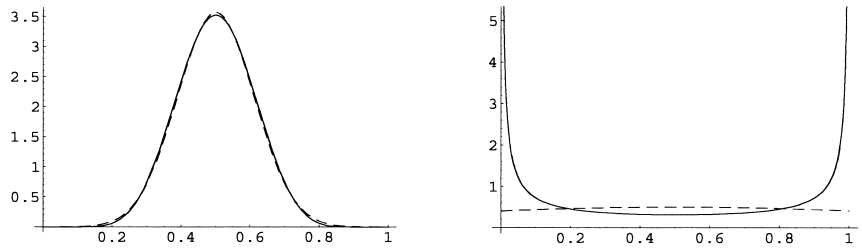

図 1. 定常状態付近での比較.

図 1 (a) (b) は式(1.4)による定常状態での真の頻度分布と，ほぼ対応する式(3.6)における 2 項目までの近似公式を重ねて表示し，それぞれパラメータ u，v を変えたものである．初期値は $x_0 = 0.5$ とした．

図 1 (a) は突然変異率が高い場合 ($u = v = 5$) である．漸近展開は微小拡散モデルを仮定しているので，理論的には $t \to \infty$ での正確さを保証していない．しかし形式的には十分時間の経過した近似公式と定常状態の式(1.4)を比較する事はできる．近似公式では $t = 10^7$ とした．図からわかる通り定常状態に至ってもなお近似が非常にうまくいっている事がわかる．しかし図 1 (b) の場合のように突然変異率が低い ($u = v = 0.1$) と，定常状態に近付くにつれて真の分布は固定(頻度が 1 になる)や消失(頻度が 0 になる)が増え，台の両端によっていくので，近似公式ではうまく状況を表現できない事が分かった．

3.2 ゲノム上の他の領域とインタラクションがある場合

3.1 節で扱ったような，式(1.2)で表現される様な領域がゲノム上に 2 つある場合を考える．この時 2 つの領域が独立であるなら，当然頻度の同時分布は単純に式(3.1)の積の形で表すことができるだろう．しかし独立でない場合，すなわちインタラクションが存在する場合，問題は非常に複雑になる．ここでインタラクションとは，既に述べている通り 2 つの領域間の組み換えの事を指す．組み換えとは 2 倍体が配偶子を形成するにあたって，生殖細胞内のペアのゲノム同士が互いに多数の領域を交換しあう現象である．ゲノムは染色体という複数本の断片に分かれて存在している．今問題にしている 2 つの領域が別の染色体に分かれて乗っているなら，組み換えが起きても領域の交換の影響は受けず，独立したままである．しかし 2 つの領域が同じ染色体に乗っている場合，組み換えの影響を受け，個々の領域(領域 A と領域 B とする)のそれぞれの祖先型と派生型の頻度の積が，2 つの領域を合わせて考えたペアの頻度とずれてくることが考えられる．

もう少し明確にモデルのセッティングをしよう．領域 A の派生型を A，野生型を a とし，その頻度をそれぞれ p_t，$1 - p_t$ とする．同様に領域 B の派生型を B，野生型を b とし，その頻度をそれぞれ q_t，$1 - q_t$ とする．A から a への(再起)突然変異率を u_1，a から A への突然変異率を v_1 とする．再び同様に，B から b への(再起)突然変異率を u_2，b から B への突然変異率を v_2 とする．また領域 A と領域 B の間の組み換え率を c とする．ここで 2 つの領域をペアにして，つなぎ合わせて考えた時，考えうる型の組み合わせ(ハプロタイプという)は，AB, Ab, aB, ab の 4 つある．これらの頻度をそれぞれ $x_{1t}, x_{2t}, x_{3t}, x_{4t} (= 1 - x_{1t} - x_{2t} - x_{3t})$ とする．今ある量 D_t を $D_t \equiv x_{1t}x_{4t} - x_{3t}x_{2t}$ で定義する．もし領域 A と領域 B が独立であれば，この量は常に 0 になるはずである．すなわち D_t はインタラクションの程度を表す変数である．これを連鎖

不平衡(LD)という．勿論作り方から D_t も確率過程であり，分布をもつ．またより直接的に独立からのずれとして $D_t \equiv x_{1t} - p_t q_t$ などと定義して上の式を導いてもよい．このようにセッティングすると，変数の組 (x_{1t}, x_{2t}, x_{3t}) と (p_t, q_t, D_t) の間には 1 対 1 の関係があることがわかる(Mano, 2005)．よって本稿では組 (p_t, q_t, D_t) について考える．

(p_t, q_t, D_t) に関する生成作用素は

$$
\begin{aligned}
L = {} & \{v_1 - (u_1+v_1)p\}\frac{\partial}{\partial p} + \{v_2 - (u_2+v_2)q\}\frac{\partial}{\partial q} - (1+k+c)D\frac{\partial}{\partial D} \\
& + D\frac{\partial^2}{\partial p \partial q} + D(1-2p)\frac{\partial^2}{\partial p \partial D} + D(1-2q)\frac{\partial^2}{\partial q \partial D} + \frac{p(1-p)}{2}\frac{\partial^2}{\partial p^2} \\
& + \frac{q(1-q)}{2}\frac{\partial^2}{\partial q^2} + \frac{pq(1-p)(1-q) + D(1-2p)(1-2q) - D^2}{2}\frac{\partial^2}{\partial D^2},
\end{aligned} \tag{3.7}
$$

ここで $k = u_1 + v_1 + u_2 + v_2$ (Ohta and Kimura, 1969b)．対応する拡散過程は以下のようになる(Mano, 2007)．

$$
x_t = x_0 + \int_0^t \mu(x_s)ds + \int_0^t \sigma(x_s)dB_s, \tag{3.8}
$$

ここで

$$
x_t = \begin{pmatrix} p_t \\ q_t \\ D_t \end{pmatrix}, \quad x_0 = \begin{pmatrix} p_0 \\ q_0 \\ D_0 \end{pmatrix}, \quad \mu(x_t) = \begin{pmatrix} v_1 - (u_1+v_1)p_t \\ v_2 - (u_2+v_2)q_t \\ -(1+k+c)D_t \end{pmatrix} \tag{3.9}
$$

$$
\sigma(x_t)\big(\sigma(x_t)\big)^{\mathrm{T}} = \begin{pmatrix} p_t(1-p_t) & D_t & D_t(1-2p_t) \\ D_t & q_t(1-q_t) & D_t(1-2q_t) \\ D_t(1-2p_t) & D_t(1-2q_t) & p_t q_t(1-p_t)(1-q_t) + D_t(1-2p_t)(1-2q_t) - D_t^2 \end{pmatrix}.
$$

このモデルにおいては，真の分布に対する解析的な解は求められておらず，3.1 節で扱った場合と違って小分散漸近理論による近似公式が解析的な表現として意味を持つ．ただし生成作用素を眺めてみても分かるように，3.1 節の場合とは違い近似公式を求めるといっても，かなり困難なことが予想される．実際例えば第 2 項目を求めるために必要な $E[g_{2t}|g_{1t}]$ は，勿論計算を実行することは可能であるが，非常に冗長な表現になってしまい事実上あまり有用ではないと思われる．より高次の計算に対しては，更なる煩雑さがついて回る．小分散漸近理論を用いた近似公式の構成の大きな利点の 1 つは，正規分布に沿った高次の展開にある．しかし以上のような理由から，組み換えがある 2 つの領域のモデルに関しては，Miura (2011) では第 1 項までの近似，即ち正規分布 $Norm(x; x_t^{(0)}, \mathrm{cov}(g_{1t}))$ を採用しその有効性を検討している．それを以下に示す．

3.1 節の場合と同様に $x_t^{(0)}$ と $\mathrm{cov}(g_{1t})$ を求める．$x_t^{(0)} = (p_t^{(0)}, q_t^{(0)}, D_t^{(0)})^{\mathrm{T}}$ の各成分は式(3.9)から分かるようにお互いに独立しており，それぞれ簡単に求めることができて

$$
x_t^{(0)} = \begin{pmatrix} p_t^{(0)} \\ q_t^{(0)} \\ D_t^{(0)} \end{pmatrix} = \begin{pmatrix} \left(x_0 - \frac{v_1}{u_1+v_1}\right)e^{-(u_1+v_1)t} + \frac{v_1}{u_1+v_1} \\ \left(x_0 - \frac{v_2}{u_2+v_2}\right)e^{-(u_2+v_2)t} + \frac{v_2}{u_2+v_2} \\ D_0 e^{-(1+k+c)t} \end{pmatrix} \tag{3.10}
$$

となる．つぎに $\mathrm{cov}(g_{1t})$ であるが，まず g_{1t} は形式的な展開(2.2)から

$$
g_{1t} = \int_0^t \partial\mu(x_s^{(0)})g_{1s}ds + \int_0^t \sigma(x_s^{(0)})dB_s, \tag{3.11}
$$

ここで $\partial\mu(x_s^{(0)})$ は $\partial\mu$ の各要素をそれぞれ (p,q,D) で偏微分したものを要素に持つ (3×3) の行列．この方程式の解は，解核行列 $Y_tY_s^{-1}$ を用いて以下のように解ける．

$$g_{1t}=\int_0^t (Y_tY_s^{-1})\sigma(x_s^{(0)})dB_s \tag{3.12}$$

ここで Y_t は $dY_t/dt=\partial\mu(x_s^{(0)})Y_t$, $Y_0=I_3$ を満たし，

$$Y_tY_s^{-1}=\begin{pmatrix} e^{-(u_1+v_1)(t-s)} & 0 & o \\ 0 & e^{-(u_2+v_2)(t-s)} & 0 \\ 0 & 0 & e^{-(1+k+c)(t-s)} \end{pmatrix}=\left(Y_tY_s^{-1}\right)^{\mathrm{T}}. \tag{3.13}$$

以上から $\mathrm{cov}(g_{1t})$ は

$$\mathrm{cov}(g_{1t})=\int_0^t (Y_tY_s^{-1})\sigma(x_s^{(0)})\left(\sigma(x_s^{(0)})\right)^{\mathrm{T}}\left(Y_tY_s^{-1}\right)^{\mathrm{T}}ds. \tag{3.14}$$

図 2 及び図 3 はシミュレーションによる真の分布(の周辺分布)と近似公式から得られた曲線を重ね合わせたものである(Miura, 2011)．時間は(a)から(c)へ流れている．図 2 は突然変異率が高い場合である．図から明らかな通り，突然変異率が高ければ時間が十分経過しても真の分布は固定や消失する事がなくそれ故単鋒型を保ち，正規分布による近似はかなり良く機能している．図 3 は突然変異率が低い場合である．この場合でも(a)の様に時間があまり経過していなければ，近似は機能している．しかし時間が経過するにつれ，真の分布では固定や消失するものが増え，近似公式は機能しなくなる．LD については固定や消失が増えた影響で，座位が独立に見える場合が増え，単鋒型の分布であるものの正規分布での近似はできなくなっている．また図 2，図 3 から，頻度の周辺分布についてはインタラクションが存在しても座位が 1 つの場合である 3.1 節と同じ傾向を示す事がわかった．

4. まとめ

本稿では遺伝子の推移的な頻度分布の近似公式を与えるために，Miura (2011)によって提案された小分散漸近理論を応用した方法とその例をみてきた．上述のように漸近展開はブラウン運動に沿った形で表現され，実際の公式も正規分布に多項式(のベクトル)を掛けた形によって与えられる．他方で突然変異体の頻度のモデルは常にシンプレックス(か，またはそれを変換した有界な領域)によって与えられる．すなわちこれらの確率モデルはサポートが違う．この為にいくら形式的に展開が与えられ，またその収束が保証されたとしても，その漸近的な近づき方は非常に遅いことが予想される．その為収束の速さや有効性を具体的に見積もるのは今後の課題であると思われる．しかしながら 3 章の例で示された通り，時間があまり経過していない状態での近似公式はかなり良く真の分布を近似していることが見て取れる．このような状況が役に立つのは，例えばある定常状態にあった集団が，時刻 t でその集団の大きさが(すなわち時間や頻度のスケールが)急に変化したときなどが挙げられる．この場合は定常な分布は正確な分布が得られているのだから，スケール変化後の推移的な分布を漸近近似公式で表現して，これらを畳み込めば簡便にそのような状況が再現できる．その他に有用だと考えられる応用としては，例でも見たように，分布が固定や消失に寄らないような状況ではこの公式はうまく機能する．突然変異が祖先型と派生型相互に起きるという圧力以外でそのような状況が起こるケースとしては，超優性による多型の保持がある．ここで超優性とはセレクションの一形態で，2 倍体において野生型と派生型両方のゲノムを持っていた方が，ゲノムが両方とも野生型または派生型である場合よりも有利に働く場合であり，ヒトでは獲得免疫に関わる遺伝子群が代表的な超優性の働く座位(群)である．このように集団遺伝学に導入されたばかりの小分散漸近理論

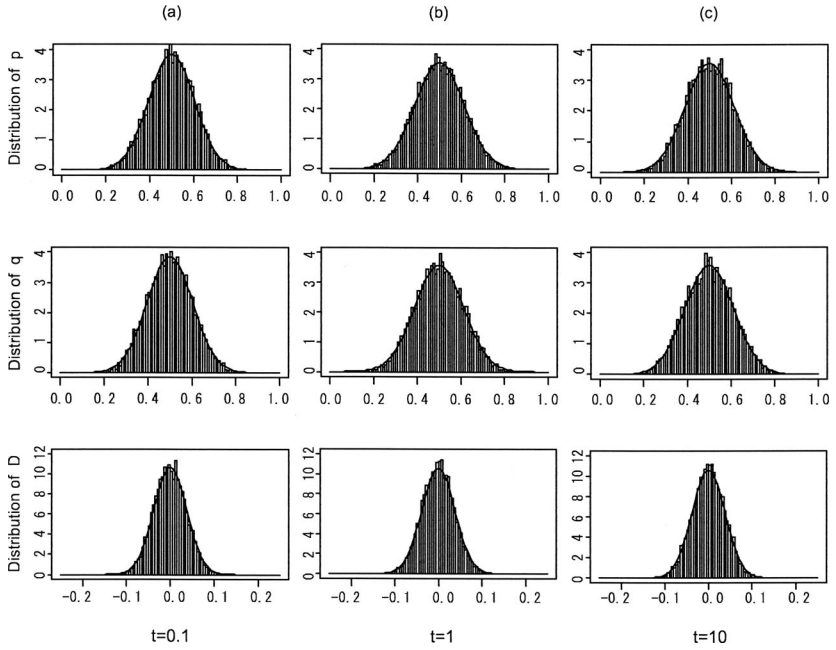

図 2. 突然変異が高い場合. $p_0 = q_0 = 0,\ D_0 = 0,\ c = 1,\ u_1 = v_1 = u_2 = v_2 = 5.$

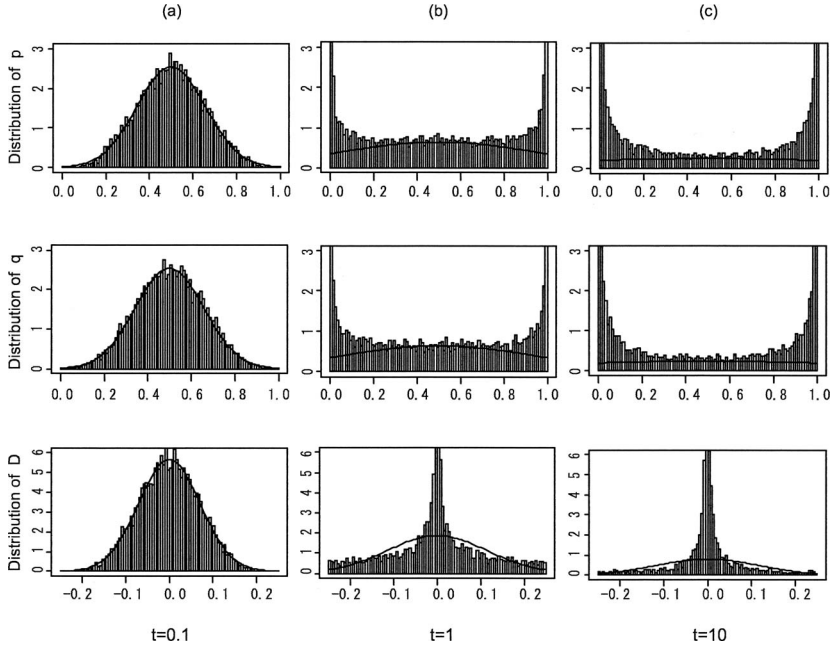

図 3. 突然変異が低い場合. $p_0 = q_0 = 0,\ D_0 = 0,\ c = 1,\ u_1 = v_1 = u_2 = v_2 = 0.1.$

であるが，問題点と応用の可能性双方とも，これから更に研究を深めていく事が重要だと考えられる．

謝　辞

査読者の方には誤り，不備をご指摘頂き，また多くの有益なご意見を頂きました．心から御礼申し上げます．

参 考 文 献

Charlesworth, Brian (1994). *Evolution in Age-structured Populations*, Cambridge Studies in Mathematical Biology, 2nd ed., Cambridge University Press, Cambridge.

Crow, J. F. and Kimura, M. (1956). Some genetic problems in natural populations, *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, 1-22, University of California Press, Berkeley, California.

Ethier, S. N. and Nagylaki, T. (1989). Diffusion approximations of the two-locus Wright-Fisher model, *Journal of Mathematical Biology*, **27**, 17–28.

Feng, Shui (2010). *The Poisson-Dirichlet Distribution and Related Topics*: *Models and Asymptotic Behaviors*, Probability and Its Applications, Springer-Verlag, Berlin, Heidelberg.

Gale, J. S. (1990). *Theoretical Population Genetics*, UNWIN HYMAN, London.

Kimura, M. (1955a). Solution of a process of random genetic drift with a continuous model, *Proceedings of National Academy of Sciences of the United States of America*, **41**, 144–150.

Kimura, M. (1955b). Stochastic process and distribution of gene frequencies under natural selection, *Cold Spring Harbor Symposia on Quantitative Biology*, **20**, 33–53.

Kunitomo, N. and Takahashi, A. (2001). The asymptotic expansion approach to the valuation of interest rate contingent claims, *Mathmatical Finance*, **11**, 117–151.

国友直人，高橋明彦(2003)．『数理ファイナンスの基礎―マリアバン解析と漸近展開の応用』，東洋経済新報社，東京．

Mano, S. (2005). Random genetic drift and gamete frequency, *Genetics*, **171**, 2043–2050.

Mano, S. (2007). Evolution of linkage disequilibrium of the founders in exponentially growing populations, *Theoretical Population Biology*, **71**, 95–108.

Miura, C. (2011). On an approximate formula for the distribution of 2-locus 2-allele model with mutual mutations, *Genes and Genetic Systems*, **86**, 207–214.

西山陽一(2011)．『マルチンゲール理論による統計解析，ISM シリーズ:進化する統計数理』，近代科学社，東京．

Ohta, T. and Kimura, M. (1969a). Linkage disequilibrium due to random genetic drift, *Genetical Research*, **13**, 47–55.

Ohta, T. and Kimura, M. (1969b). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation, *Genetics*, **63**, 229–238.

Takahashi, A. (1999). An asymptotic expansion approach to pricing financial contingent claims, *Asia-Pacific Financial Markets*, **6**, 115–151.

Watanabe, S. (1987). Analysis of Wiener functionals (Malliavin calculus) and its applications to heat kernels, *The Annals of Probability*, **15**, 1–39.

Wright, S. (1931). Evolution in Mendelian populations, *Genetics*, **16**, 97–159.

Wright, S. (1937). The distribution of gene frequencies in poplations, *Proceedings of National Academy of Sciences of the United States of America*, **23**, 307–320.

Yoshida, N. (1992). Asymptotic expansions of maximum likelihood estimators for small diffusions via the theory of Malliavin-Watanabe, *Probability Theory and Related Fields*, **92**, 275–311.

A Population Genetics Study Using the Small Disturbance Asymptotic Theory

Chiaki Miura

The Institute of Statistical Mathematics

To capture analytically the change of the transient frequency distribution of a mutant arising in a population, we apply the small disturbance asymptotic theory. This enables us to obtain an approximate formula for a model that does not have an analytical description. Model of population genetics always has a finite support. On the other hand, the asymptotic expansion is given by a form in terms of a normal distribution. Hence the formula does not work well when the mutation rate is low. However, we are able to confirm that the formula gives a good approximation when time has not passed and also when the mutation rate is high.

Key words: Small asymptotic disturbance theory, frequency distribution, normal approximation.