5　自然選択、組み換えを含む遺伝子系図

これまで紹介してきた遺伝子系図のモデルは自然選択に対して中立な１つの遺伝子座の系図を扱ってきたが、Krone and Neuhauser(1997)および Neuhauser and Crone(1997)によって導入された自然選択を考慮に入れた祖先選択グラフ（Ancestral Selection Graph）と呼ばれる系図モデルについて第５．１節で紹介する。また連鎖した２つの遺伝子座の系図を扱う、Hudson(1983), Griffiths and Marjoram(1996,1997)等による祖先組み換えグラフ（Ancestral Recombination Graph）を第５．２節で紹介する。どちらのモデルも分枝(branching)と合祖(coalescence)を含む出生死滅過程として表現される。

５．１　自然選択を含む遺伝子系図（Ancestral Selection Graph）

５．１．１．　連続時間モランモデル

N個体からなる半数体生物集団を考えよう。二つのアレル A_1, A_2 を仮定し、$X(t)$ を集団内でのアレル A_1 の個数とする。2.1 節で紹介した離散時間モランモデルの連続時間バージョンは以下のように表現される。時間 $(t,t+h)$ の間に $X(t)$ が変化する確率は $\lambda h+o(h)$ とする。モランモデルに基づいて、N個体の中からランダムに１個体選ばれ、ある確率で子供を生む。それと同時に親の個体がランダムに１個体除かれ、先の子供によって置き換わる。親のタイプと生まれる子供のタイプの関係の確率を以下のように定義する。

親　　子　　　　　　　　親　　子

$$A_1 \to \begin{cases} A_1 & \text{確率 } 1-\gamma_1 \\ A_2 & \text{確率 } \gamma_1 \end{cases} \quad 、 \quad A_2 \to \begin{cases} A_1 & \text{確率 } \gamma_2 \\ A_2 & \text{確率 } 1-\gamma_2 \end{cases}$$

γ_1,γ_2 は突然変異率である。$X(t)$ の変化があったとき、ジャンプ過程の遷移確率は以下の様になる。

$$\begin{aligned} P(X(t+)=j+1 \mid X(t)=j) &= \left(1-\frac{j}{N}\right)\left\{\frac{j}{N}(1-\gamma_1)+(1-\frac{j}{N})\gamma_2\right\} = \lambda_j/\lambda \\ P(X(t+)=j-1 \mid X(t)=j) &= \frac{j}{N}\left\{\frac{j}{N}\gamma_1+(1-\frac{j}{N})(1-\gamma_2)\right\} = \mu_j/\lambda \qquad (5.1) \\ P(X(t+)=j \mid X(t)=j) &= 1-(\lambda_j+\mu_j)/\lambda \end{aligned}$$

これより、$\lambda_j = \lambda\left(1-\frac{j}{N}\right)\left\{\frac{j}{N}(1-\gamma_1)+\left(1-\frac{j}{N}\right)\gamma_2\right\}$, $\mu_j = \lambda\frac{j}{N}\left\{\frac{j}{N}\gamma_1+\left(1-\frac{j}{N}\right)(1-\gamma_2)\right\}$

で与えられる。　マルコフ過程 $X(t)$ の推移確率を $P_t(j,k)=P(X(t)=k \mid X(0)=j)$ とすると、次のコルモゴロフ後退方程式で表現される出生死滅過程が導かれる。

$$\frac{d}{dt}P_t(j,k) = \lambda_j P_t(j+1,k)+\mu_j P_t(j-1,k)-(\lambda_j+\mu_j)P_t(j,k) \qquad P_0(k,j)=\delta_{k,j} \qquad (5.2)$$

ここでλ_j, μ_jは(5.1)で定義される定数である。

5．1．2　自然選択を含むモランモデル

中立なモランモデルではアレルタイプによらず、すべて同じ出生率λで子供を生むと仮定した。そこでこの出生率にアレルタイプによる違いを導入することにより、自然選択を含むモデルを定義しよう。自然選択を含むモランモデルは、偏りのある投票者モデル(biased voter model)に含まれるもので、Harris(1976),Schwartz(1977), Bramson&Griffeath (!980,1981)などによって導入、研究されたモデルである。出生率を以下の様に定義する。

A_1タイプの出生率=$\lambda_1 = \dfrac{N}{2}$　、（1個体当たり$\dfrac{1}{2}$の率のポアソン過程で出生する。）

A_2タイプの出生率＝$\lambda_2 = \lambda_1(1+s_N)$,$s_N \geq 0$　とする。

$\lambda_2 \geq \lambda_1$であり、アレルタイプA_2の方が子供の出生率が高く自然選択に対して有利とする。子供が生まれるとき、親と同じタイプか異なるタイプか、すなわち突然変異の確率を次の様に定義する。子供が親と同じタイプである確率＝$1-u_N$、 異なるタイプである確率＝u_Nとする。中立遺伝子の場合の(5.2)式と同様に、自然選択を考慮に入れた連続時間モランモデルは次のコルモゴロフ後退方程式で表現される出生死滅過程となる。

$Z(t) = (Z_1(t), Z_2(t))$； $Z_i(t)$はアレルA_iの個数$(i = 1,2)$、その推移確率を

$P_t(j,k) = P(Z_1(t) = k | Z_1(0) = j)$　とする。

$$\frac{d}{dt}P_t(j,k) = \Lambda_j P_t(j+1,k) + \mathrm{M}_j P_t(j-1,k) - (\Lambda_j + M_j)P_t(j,k) \qquad (5.3)$$

$$\Lambda_j = \lambda_1 j(\frac{N-j}{N})(1-u_N) + \lambda_2(N-j)\frac{N-j}{N}u_N$$

$$\mathrm{M}_j = \lambda_2(N-j)\frac{j}{N}(1-u_N) + \lambda_1 j(\frac{j}{N})u_N$$

最後に集団サイズNを無限大（$N \to \infty$）とし、かつ$Nu_N \to \theta$, $Ns_N \to \sigma(>0)$とする。拡散過程近似$X(t) = \lim_{N\to\infty} Z_1(t)/N$を行うと

遺伝子頻度に関する拡散モデルが得られる。その定常分布$\phi(x)$は次の方程式を満たす。

$$\frac{1}{2}\frac{\partial^2}{\partial x^2}[b(x)\phi(x)] - \frac{\partial}{\partial x}[a(x)\phi(x)] = 0$$

ここで、$a(x) = -\dfrac{\sigma}{2}x(1-x) + \dfrac{\theta}{2}(1-2x)$,　$b(x) = x(1-x)$

これより、定常分布はWrightの公式(Wright 1949, p383)によって得られ

$\phi(x) = Kx^{\theta-1}(1-x)^{\theta-1}\exp(-\sigma x)$となる (Moran(1962), Crow&Kimura(1970))。 (5.4)

Kは規格化定数である。

5．1．3 祖先選択グラフ（Ancestral Selection Graph）

自然選択を含むモランモデルは独立なポアソン過程のシステムを用いたパーコレーションモデルを用いて表現できる。ポアソン過程について第1章で証明した定理1.18に注意しよう。再度述べると、

補題5．1

$N_t(t \geq 0)$をパラメーターλのポアソン過程とする、$P(N_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$ $(k = 0,1,2,..)$。この過程を確率pで間引き(thinning)すると、パラメーターλpのポアソン過程になる。

（証明）

パラメーターλのポアソン過程N_tで、ジャンプが起こるごとに、そのジャンプを採用する確率がpであり、そのジャンプを無視する確率を$1-p$とする。このように間引きによって定義された過程をM_tとすると、$q = 1-p$として

$$P(M_t = k) = \sum_{n=k}^{\infty} P(N_t = n)_n C_k p^k q^{n-k} = \sum_{n=k}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \times \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

$$= \frac{e^{-\lambda t} p^k}{k!} \sum_{n=k}^{\infty} \frac{(\lambda t)^n q^{n-k}}{(n-k)!} = \frac{e^{-\lambda t}(\lambda p t)^k}{k!} \sum_{j=0}^{\infty} \frac{(\lambda q t)^j}{j!} = \frac{e^{-\lambda t}(\lambda p t)^k}{k!} \times e^{\lambda q t}$$

$$= \frac{(\lambda p t)^k}{k!} \exp\left(-\lambda t(1-q)\right) = \frac{(\lambda p t)^k}{k!} \exp\left(-\lambda p t\right)$$

M_tはパラメーターλpのポアソン過程である。

ポアソン過程のこの性質を利用して、自然選択を含むモランモデルは、自然選択に対して有利な個体については、速い出生率のポアソン過程で子供を出生させ、自然選択に対して不利な個体の出生は、ある確率で間引くことによって、より低い出生率のポアソン過程による子供の出生として実現できる。さらに突然変異を考慮に入れると以下のように定義される。$I = \{1,2,...,N\}$を個体の集合、$\eta_t(x) = 1$ or 2 $(x \in I)$を個体$x \in I$のアレル状態とする。各 $(x,y) \in I \times I$に対して$W_n^{(x,y)} (n \geq 1)$を率λ_2 / Nで生じるポアソン過程の発生時刻とする。すなわちペア(x,y)に対してxからyへλ_2 / Nの発生率で矢印(→)が生じる。これは有利な個体の出生を意味する。矢印の出現と同時に突然変異と間引き(thinning)のための二つの乱数$U_n^{(x,y)}, V_n^{(x,y)}$を発生させ、次の様に突然変異及び自然選択を導入する。$U_n^{(x,y)}$を[0,1]上の一様乱数とし、ある定数u_Nに対して$U_n^{(x,y)} < u_N$のとき 突然変異が生じる

($-\bullet\rightarrow$)。また$V_n^{(x,y)}$を$U_n^{(x,y)}$と独立な[0,1]上の一様乱数とする。この乱数の値によって、

補題５．１より、次の手順に従って矢印のタイプを決定する。

$$V_n^{(x,y)} < \frac{\lambda_1}{\lambda_2} \Rightarrow "\rightarrow \delta" \quad : \quad \text{率}\lambda_1\text{で出現する}$$

$$V_n^{(x,y)} \geq \frac{\lambda_1}{\lambda_2} \Rightarrow "\overset{2}{\rightarrow}" \quad : \quad \text{率}\lambda_2 - \lambda_1\text{で出現する。}$$

この操作は確率$p = \frac{\lambda_1}{\lambda_2}$による間引きを表しており、$V_n^{(x,y)} < p$のとき生じる"$\rightarrow \delta$"は

２つのどちらのアレルタイプでも出生可能な矢印であり、$V_n^{(x,y)} \geq p$で生じる"$\overset{2}{\rightarrow}$"はアレルタイプ１は間引かれタイプ２の個体のみ子供を出生できる矢印を表している。矢印の先の個体は死亡し、出生した子供と置き換わる。ポアソン過程$W_n^{(x,y)}(n \geq 1)$と、乱数$U_n^{(x,y)}$と

$V_n^{(x,y)}$によって、時間を$t = 0$から t まで走らせると、グラフ上に上記の矢印の付いた図ができる。$t = 0$での各個体の状態を指定すると、時刻 t における状態も決まる。例えば図 5-1 で時刻$t = 0$に 4, 8 の個体が状態A_1とすると、A_1個体は矢印"$\rightarrow \delta$"に沿ってのみ子供を出生できるので、時刻 t にその子孫として 2,3,5,8,9 の個体が状態A_1であることが分かる。A_2個体は２種の矢印を使って子供を出生できる。

図 5-1

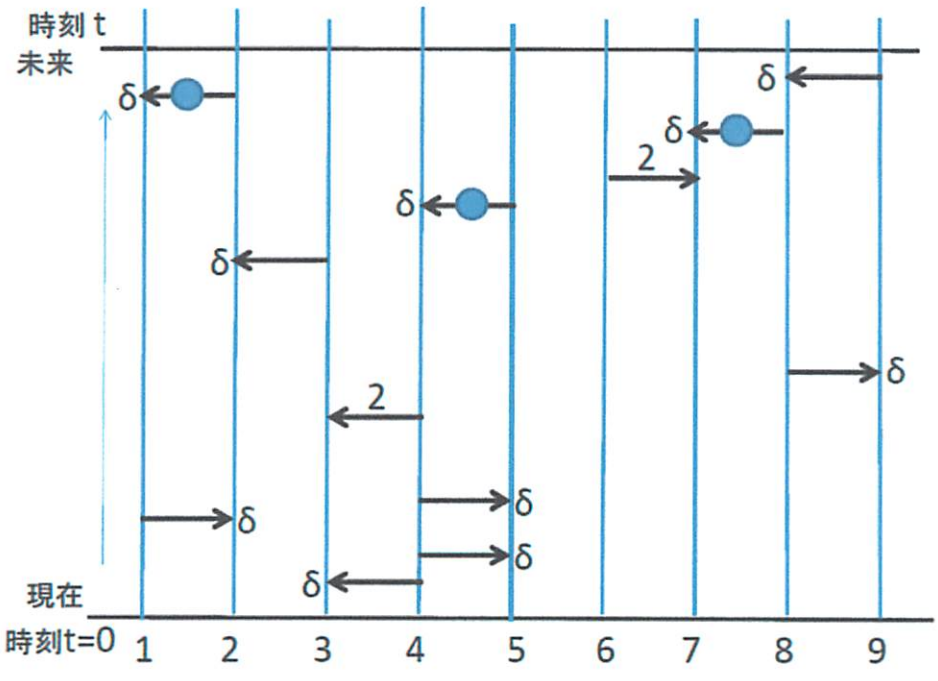


図 5-1 を時間の向きを逆向きにして、全ての矢印も逆向きにしたものが図 5-2 である。ただし、図 5-2 では図 5-1 の時刻 t を現在の時刻 0 と表示し、過去に遡った方向を正の向きとし、サンプル遺伝子の時間 t だけ遡った祖先の状態を考える。この図 5-2 を図 5-1 に双対(dual)な図と呼ぶことにする。

図 5-2

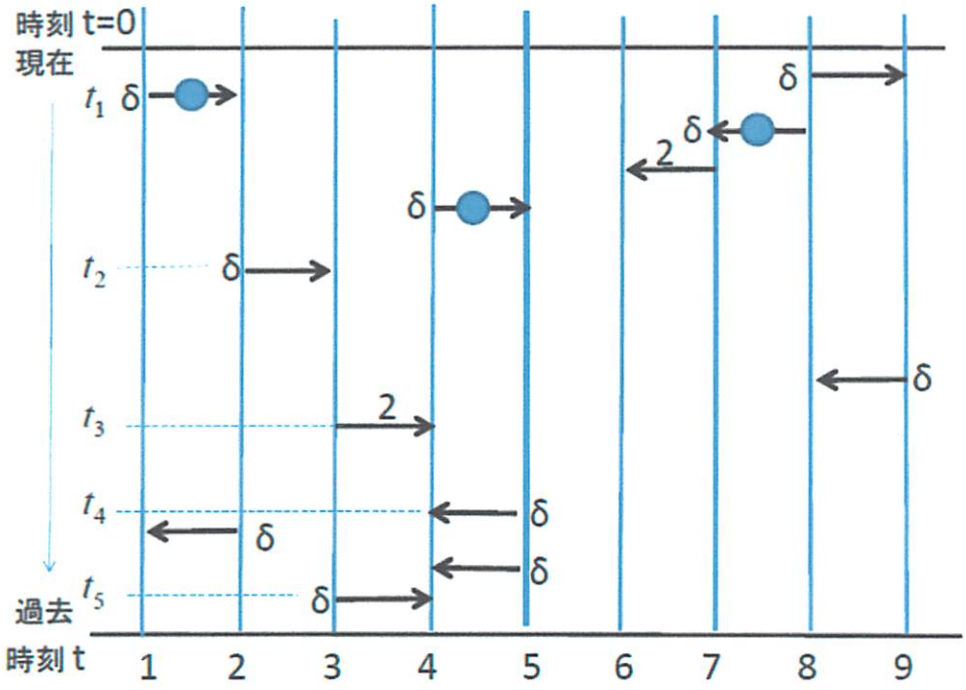


図 5-2 で、時刻$t=0$でのサンプルを{1,3,5}としよう。$t=0$の各個体の状態は未知で、祖先（時刻 t）の各個体のタイプは既知と仮定しよう。双対図 5-2 を$t=0$から下へ降りてゆく。サンプル遺伝子のタイプが未知なのでその個体のアレルタイプによって利用できる矢印に違いが生じ、下降するに従って、合祖（coalescence）と分枝（branching）が生じる。例えば個体 1 は時刻t_1で矢印に沿って個体 2 に移るが、そのとき突然変異を生じているので、この時点で個体 2 の状態は個体 1 とは別のアレル状態である。さらに下降すると時刻t_2で個体 3 と合祖(coalesce)が起こる。さらに時間を下ると、時刻t_3で矢印"$\overset{2}{\rightarrow}$"が現れるが、個体 3 がA_2タイプならば。この矢印にそって右に移るが、もしA_1タイプならば、この矢印は利用できず、そのまま下降することになる。すなわち、ここで分枝(branching)が起こる。矢印"$\overset{2}{\rightarrow}$"を通る場合は時刻t_4で個体 5 との間で合祖(coalesce)が起こり、この時点で 1 つの共通祖先に到達する。他方、矢印"$\overset{2}{\rightarrow}$"を通らずに下方へ行く場合は、時刻t_4に個体 5 が 4 に

移り、時刻t_5で個体３との間で合祖が起こり共通祖先に到達する。時刻ｔの祖先アレルタイプが既知とすると、それに基づき現れた個体のアレルタイプが確定し利用できる矢印と系図も確定する。個体間の移動（空間構造）及び突然変異は無視し、ｎ個のサンプル遺伝子の祖先について合祖と分枝のみに着目した祖先系図を表現する双対(dual)なマルコフ過程を$G_n^N(t)$で表す。その生成作用素はポアソン過程$W_n^{(x,y)}(n \geq 1)$と、乱数$U_n^{(x,y)}$と$V_n^{(x,y)}$より以下のようになる。突然変異率$u_N = 0$とする。祖先の数がk個の状態の時点において

○ 合祖(coalesce)：矢印"$\rightarrow \delta$"がサンプル遺伝子の祖先がいるサイトに着地したときに合祖が起きるので、１つのサイトでの矢印"$\rightarrow \delta$"の発生率がλ_1より

$$\text{合祖の率（Coalescence rate）} = \lambda_1 k \frac{k-1}{N} = \frac{N}{2} k \frac{k-1}{N} = \frac{k(k-1)}{2}$$

このモデルでは空間構造、この場合は親の位置は無視しているので、矢印"$\rightarrow \delta$"が系図過程で粒子がいないサイトに着地するときは、プロセスとしての変化は発生しない。

○分枝(branching)：分枝が起こるのは、矢印 "$\overset{2}{\rightarrow}$"が生じたときなので

$$\text{分枝の率（Branching rate）} = (\lambda_2 - \lambda_1)k = \frac{N}{2} s_N k \rightarrow \frac{\sigma}{2} k \quad (N \rightarrow \infty)$$

○ 衝突(Collision)：矢印 "$\overset{2}{\rightarrow}$"がサンプルの祖先がいるサイトに着地したとき、これを衝突と呼ぶことにする。

$$\text{衝突率（Collision rate）} = \frac{N}{2} s_N k \times \frac{k-1}{N} = \frac{\sigma}{2N} k(k-1) = O(\frac{1}{N})$$

衝突によって生じた祖先（粒子）を仮想祖先（粒子）（fictious particle）と呼ぶと、上の計算より衝突の発生率は$N \rightarrow \infty$の極限で無視できることが分かる（証明は後述）。仮想祖先でない通常の祖先を実祖先(non-fictious particle)と呼ぶことにする。

仮想祖先が他の仮想祖先と合祖したときは、その合祖粒子は仮想祖先と呼ぶ、また実祖先と合祖したときは、実祖先となる。以上より$N \rightarrow \infty$において、自然選択を含むモランモデルの双対なプロセスとして現れる遺伝子系図のプロセスは合祖と分枝を含んだモデルとして表現される。

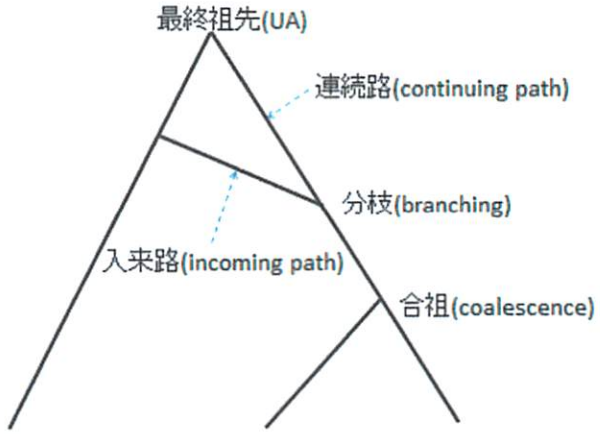
時刻ｔにおいてｎ個のサンプル遺伝子の実祖先の集合を$\mathfrak{R}_n^{N,1}(t)$、仮想祖先の集合を$\mathfrak{R}_n^{N,2}(t)$とし、$\mathfrak{R}_n^N(t) = \mathfrak{R}_n^{N,1}(t) \cup \mathfrak{R}_n^{N,2}$とすると、初期条件は$\mathfrak{R}_n^{N,1}(0) = \{1, 2, \cdots, n\}$, $\mathfrak{R}_n^{N,2} = \phi$である。

最初に合祖または分枝が起こる時刻をT_1^Nとする。時刻T_1^Nにおいて、サンプルｉとｊが共通祖先に到達し合祖が起こったとき、$\beta_1^N = 0, \gamma_1^N = (i, j), i < j$と記して、合祖した祖先は小さ

い方の数字 i で表示する。従って、$\Re_n^{N,1}(\mathrm{T}_1^{\mathrm{N}})=\{1,2,\cdots,n\}\setminus\{j\}$, $\Re_n^{N,2}(\mathrm{T}_1^{\mathrm{N}})=\phi$ となる。時刻 $\mathrm{T}_1^{\mathrm{N}}$ において、サンプル j に分枝が起こったときは、$\beta_1^N=j,\gamma_1^N=(0,0)$ と記して、分枝によって生じた祖先は使われていない最小の数字 n+1 を充てることとする。実祖先の場合は、$\Re_n^{N,1}(\mathrm{T}_1^{\mathrm{N}})=\{1,2,\cdots,n,n+1\}$, $\Re_n^{N,2}(\mathrm{T}_1^{\mathrm{N}})=\phi$ と表され、仮想祖先の場合は $\Re_n^{N,1}(\mathrm{T}_1^{\mathrm{N}})=\{1,2,\cdots,n\}$, $\Re_n^{N,2}(\mathrm{T}_1^{\mathrm{N}})=\{n+1\}$ と表される。以下同様に、m 回目の合祖または分枝の事象が生じた時点 $\mathrm{T}_{\mathrm{m}}^{\mathrm{N}}$ で、分枝ならば β_m^N、合祖ならば γ_m^N を用いて表示する。祖先系図を表すマルコフ過程 $G_n^N(t)$ は $G_n^N(t)=\left[\Re_n^N(t),\{\mathrm{T}_{\mathrm{m}}^{\mathrm{N}};m\geq 1\},\{(\beta_m^N,\gamma_m^N),m\geq 1\}\right]$ で表現される。このマルコフ過程 $G_n^N(t)$ の上にこの系図と独立に突然変異を重ねる。$G_n^N(t)$ に突然変異を重ねた遺伝子系図のプロセスを $F_n^N(t)$ で表す。マルコフ過程 $G_n^N(t)$ のサイズ過程（時刻 t において含まれる祖先の数）を $\left|G_n^N(t)\right|=\left|\Re_n^N(t)\right|=\left|\Re_n^{N,1}(t)\right|+\left|\Re_n^{N,2}(t)\right|=\mathrm{A}_{\mathrm{n}}^{\mathrm{N}}(t)$ と表し、このサイズが 1 になる時刻を $\mathrm{T}_{\mathrm{UA}}^{\mathrm{N}}=Inf\left\{t;\mathrm{A}_{\mathrm{n}}^{\mathrm{N}}(t)=\left|G_n^N(t)\right|=1\right\}$ として、マルコフ過程 $G_n^N(t)$ を時刻 $\mathrm{T}_{\mathrm{UA}}^{\mathrm{N}}$ で停止させるマルコフ過程を $\left\{G_n^N(t):0\leq t\leq \mathrm{T}_{\mathrm{UA}}^{\mathrm{N}}\right\}$ とする。

以上のモデルを 3 個のサンプル遺伝子について図示すると図 5-3 の様になる。

図 5-3



分枝が起きたとき、親個体の系図の経路を連続路(continuing path)、矢印 "$\overset{2}{\rightarrow}$" に沿って遡る経路を入来路(incoming path)と呼ぶ。

$N\rightarrow\infty$ の極限を取ると、$G_n^N(t)=\left[\Re_n^N(t),\{\mathrm{T}_{\mathrm{m}}^{\mathrm{N}};m\geq 1\},\{(\beta_m^N,\gamma_m^N),m\geq 1\}\right]$ は明確な連続時間集合値マルコフ過程 $G_n(t)=\left[\Re_n(t),\{\mathrm{T}_{\mathrm{m}};m\geq 1\},\{(\beta_m,\gamma_m),m\geq 1\}\right]$ に収束する。ここで、T_{m} はジャンプ時刻、(β_m,γ_m) はラベルプロセスであり、$\Re_n(t)=\Re_n^1(t)\cup\Re_n^2(t)$（$\Re_n^1(t)$ は実祖先の集

合、$\Re_n^2(t)$は仮想祖先の集合）であるが、任意の時刻tにおいて$\Re_n^2(t)=\phi$であることが示される。従って、極限$N\to\infty$においては、実祖先のみ考えれば良いことになり、実祖先の集合$\Re_n^1(t)$を単に$\Re_n(t)$と表す。$\Re_n(t)$のサイズ過程を$A_n(t)=|\Re_n(t)|$とすると、これまでの議論より、次の定理が成り立つ。

定理５．２

突然変異率$u_N=0$とする。$\Re_n^N(t)$をn個のサンプル遺伝子の時刻tの実祖先の集合、それに含まれる祖先（粒子）の数を$A_n^N(t)=|\Re_n^N(t)|$とすると$N\to\infty$のときそれぞれ、連続時間マルコフ連鎖$\Re_n(t)$, $A_n(t)$に収束する。$\Re_n(t)$は集合値(set-valued)マルコフ連鎖であり、$A_n(t)$はそのサイズを表すマルコフ連鎖である。この出生死滅過程が１個の粒子になる最初の時刻を$T_{UA}=\inf\{t\geq 0;\ A_n(t)=1\}$とするとマルコフ連鎖$\{A_n(t)\,;\,0\leq t\leq T_{UA}\}$は次のコルモゴロフ後退方程式で表される出生死滅過程である。

$P_t(n,j)=P(A_n(t)=j|A_n(0)=n)$とすると

$$\frac{d}{dt}P_t(n,j)=\frac{\sigma n}{2}P_t(n+1,j)+\frac{n(n-1)}{2}P_t(n-1,j)-\{\frac{\sigma n}{2}+\frac{n(n-1)}{2})P_t(n,j) \qquad (5.5)$$

マルコフ連鎖$\Re_n(t)$を祖先選択グラフ(Ancestral Selection Graph)と呼び、略してASGと表示される。$\sigma=0$のときは、第３章で述べた中立な合祖モデル(Kingmanの coalescentモデル）に一致する。$N\to\infty$における収束の証明は、Krone and Neuhauser(1997)のAppendix Aを参照されたい。

祖先遺伝子サイズを表す$A_n(t)$はCoalescenceまたはbranchingによって推移する出生死滅過程であり、$A_n(T_m)=k$のとき、その滞在時間$T_{m+1}-T_m$はパラメーター
$\frac{\sigma k}{2}+\frac{k(k-1)}{2}=\frac{k(\sigma+k-1)}{2}$の指数分布に従い、
時刻T_{m+1}に確率$\frac{k(k-1)/2}{k(\sigma+k-1)/2}=\frac{k-1}{\sigma+k-1}$で合祖(coalescence)が起こり、確率$\frac{\sigma}{\sigma+k-1}$で分枝(branching)が起こる。

５．１．４　祖先選択グラフ（ASG）の性質および解釈

n個のサンプルについて、最終祖先（ultimate ancestor (UA)）に到達する最初の時刻を$T_{UA}=\inf\{t\geq 0;\ A_n(t)=1\}$とする。このとき、出生死滅過程の一般論より、次の定理が

成り立つ。

定理５．３

全ての自然数 n に対して、$P_n(T_{UA}<\infty)=1$ であり、次の期待値を得る。

$$E_n[T_{UA}]=2(1-\frac{1}{n})+2\sum_{r=1}^{n-1}\frac{1}{r(r+1)}\frac{e^{\sigma}}{\sigma^{r+1}}\int_0^{\sigma} t^{r+1}e^{-t}dt \qquad (5.6)$$

（証明）(5.1)で定義される出生死滅過程は状態１が吸収状態となっている。従って $P_n(T_{UA}<\infty)=1$ は吸収確率が１であることを示せばよい。この出生死滅過程の死亡率、出生率が $\mu_i=\frac{i(i-1)}{2}$, $\lambda_i=\frac{\sigma i}{2}$ より、ジャンプ過程の推移確率は次式で与えられる。

$$\begin{cases} i\to i+1 : \lambda_i/(\lambda_i+\mu_i)=\sigma/(\sigma+i-1) \\ i\to i-1 : \mu_i/(\lambda_i+\mu_i)=(i-1)/(\sigma+i-1) \end{cases} \qquad (5.7)$$

第１章の定理１．９において、$p_j=\sigma/(\sigma+j-1)$, $q_j=(j-1)/(\sigma+j-1)$ と置いて

状態１が吸収状態であることに注意すると、

$\sum_{i=2}^{\infty}\left(\prod_{j=2}^{i}\frac{q_j}{p_j}\right)=\sum_{i=2}^{\infty}\left(\prod_{j=2}^{i}\frac{j-1}{\sigma}\right)=\sum_{i=2}^{\infty}\left(\frac{1}{\sigma}\right)^{i-1}(i-1)!=+\infty$。よって定理１．９より

吸収確率 $P_n(T_{UA}<\infty)=1$ である。

次に吸収までの平均時間 $E_n[T_{UA}]$ を求める。この出生死滅過程の状態 j (≥ 2) での滞在時間 τ_j はパラメーター $(\lambda_j+\mu_j)$ の指数分布に従う。よって滞在時間の平均は $\frac{1}{\lambda_j+\mu_j}$ であり、滞在時間に関する強マルコフ性と、その後のジャンプに着目すると次式を得る。

$h(j)=E_j[T_{UA}]$ とすると $h(1)=0$ であり、$j\geq 2$ のとき

$h(j)=\frac{1}{\lambda_j+\mu_j}+\frac{\lambda_j}{\lambda_j+\mu_j}h(j+1)+\frac{\mu_j}{\lambda_j+\mu_j}h(j-1)$ が成り立つ。解は定理１．１０と同様にして次のように与えられる（Karlin&Tailor(1975), Chap4, Theorem7.1 参照)。

$\rho_2=\frac{1}{\mu_2}=1$ ，$i\geq 3$のとき $\rho_i=\frac{\lambda_2\lambda_3\cdots\lambda_{i-1}}{\mu_2\mu_3\cdots\mu_i}=\frac{2\sigma^{i-2}}{i!}$ とする。

$\sum_{i=2}^{\infty}\rho_i=\sum_{i=2}^{\infty}\frac{2\sigma^{i-2}}{i!}=\frac{2}{\sigma^2}\left\{\sum_{i=0}^{\infty}\frac{\sigma^i}{i!}-1-\sigma\right\}=\frac{2}{\sigma^2}\{\exp(\sigma)-1-\sigma\}<\infty$、よって有界な解をも

ち、$h(j)=\sum_{i=2}^{\infty}\rho_i+\sum_{r=2}^{j-1}\left(\prod_{i=2}^{r}\frac{\mu_i}{\lambda_i}\right)\left(\sum_{k=r+1}^{\infty}\rho_k\right)=\sum_{r=1}^{j-1}\left(\prod_{i=2}^{r}\frac{\mu_i}{\lambda_i}\right)\left(\sum_{k=r+1}^{\infty}\rho_k\right)=2\sum_{r=1}^{j-1}\sum_{i=0}^{\infty}\frac{(r-1)!}{(r+i+1)!}\sigma^i$

特に $h(n)=E_n[T_{UA}]=2\sum_{r=1}^{n-1}\sum_{i=0}^{\infty}\frac{(r-1)!}{(r+i+1)!}\sigma^i$

$=2\left(1-\frac{1}{n}\right)+2\sum_{r=1}^{n-1}\frac{1}{r(r-1)}\frac{e^{\sigma}}{\sigma^{r+1}}\int_0^{\sigma}t^{r+1}e^{-t}dt$

系５．４

σが小さい値のとき、σでテーラー展開すると

$$E_n[T_{UA}]=2\left(1-\frac{1}{n}\right)+\frac{(n-1)(n+2)}{2n(n+1)}\sigma+\frac{(n-1)(6+4n+n^2)}{9n(n+1)(n+2)}\sigma^2+O(\sigma^3)。 \qquad (5.8)$$

また任意の$\sigma>0$に対して、$\lim_{n\to\infty}E_n[T_{UA}]=2+\frac{2}{\sigma}\left(-\gamma-\sigma+Ei(\sigma)-\ln\sigma\right)$。

ここで、$\gamma=0.577216\cdots$(オイラー定数), $Ei(\sigma)=-\int_{-\sigma}^{\infty}\frac{e^{-t}}{t}dt$ である。

（証明）

$$h(n)=E_n[T_{UA}]=2\sum_{r=1}^{n-1}\sum_{i=0}^{\infty}\frac{(r-1)!}{(r+i+1)!}\sigma^i=2\sum_{i=0}^{\infty}\left(\sum_{r=1}^{n-1}\frac{(r-1)!}{(r+i+1)!}\right)\sigma^i$$

$$=2\left\{\sum_{r=1}^{n-1}\frac{1}{r(r+1)}+\left(\sum_{r=1}^{n-1}\frac{1}{r(r+1)(r+2)}\right)\sigma+\left(\sum_{r=1}^{n-1}\frac{1}{r(r+1)(r+2)(r+3)}\right)\sigma^2+O\left(\sigma^3\right)\right\}$$

$$=2\left(1-\frac{1}{n}\right)+\frac{(n-1)(n+2)}{2n(n+1)}\sigma+\frac{(n-1)(6+4n+n^2)}{9n(n+1)(n+2)}\sigma^2+O\left(\sigma^3\right)$$

◎ 祖先選択グラフ（Ancestral Selection Graph）の解釈

突然変異が存在しないという条件の下で合祖と分枝を含む祖先選択グラフが得られた。突然変異は祖先選択グラフが与えられた後に、系図上に過去から現在に向かってポアソン過程で発生させて重ねる。まず、突然変異は親が子供を産むときに、アレルタイプによらずu_Nの確率で発生する。出生の矢印$\to\delta$および$\overset{2}{\to}$はそれぞれλ_1, $\lambda_2-\lambda_1$の率で発生するので、出生率は合計λ_2である。よって突然変異の発生は

$\lambda_2 u_N=\frac{N}{2}(1+s_N)u_N=\frac{N}{2}\left(1+\frac{\sigma}{N}\right)\frac{\theta}{N}$であり、$N\to\infty$とすると、$\lambda_2 u_N\to\frac{\theta}{2}$より、突然変異は極限$N\to\infty$でパラメーター$\frac{\theta}{2}$のポアソン過程で系図上に発生する。ただし、

矢印 $\xrightarrow{2}$ 上に突然変異を生じる率は $(\lambda_2-\lambda_1)u_N=\frac{Ns_N}{2}\times\frac{\theta}{N}=\frac{\theta}{2}\times\frac{\theta}{N}=O\left(\frac{1}{N}\right)$ であり、

$N\to\infty$ のとき、無視できることが分かる。

現実の遺伝子系図には分枝は生じないはずである。実際の系図は共通祖先遺伝子の状態を指定することによって確定される。以下でサンプル遺伝子の系図を得る手順を説明しよう。遺伝子頻度 $X(t)$ は定常分布 $\phi(x)$ で与えられる定常確率過程と仮定する。

（1） 最終祖先（ultimate ancestor(UA)）に到達するまで合祖（coalescing）と分枝（branching） を含む出生死滅過程に従ってによって祖先選択グラフを作成する。

（2） 突然変異と自然選択存在下で上記の定常分布 $\phi(x)$ に従って UA の状態を選ぶ。（$X(t)$ が非定常な場合、遺伝子頻度分布に従ってＵＡのタイプを決める）

（3） 時間前向きにパラメーター $\frac{\theta}{2}$ のポアソン過程で突然変異を各枝に独立に発生させる。

（4） UA の状態が決まるとサンプル遺伝子のアレルタイプおよび系図が確定する。

図 5-4

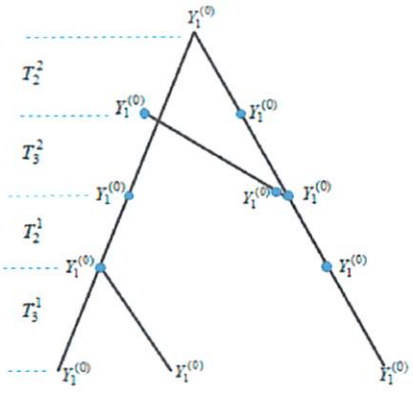
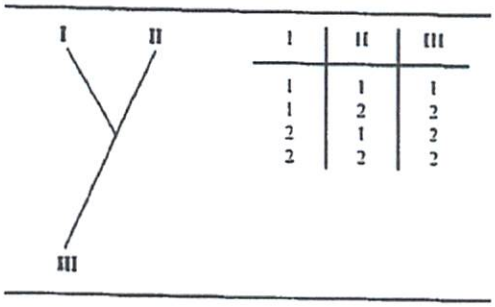


表 5-1

I は Incoming path、II は Continuing path、III は結果のタイプを表す。



I	II	III
1	1	1
1	2	2
2	1	2
2	2	2

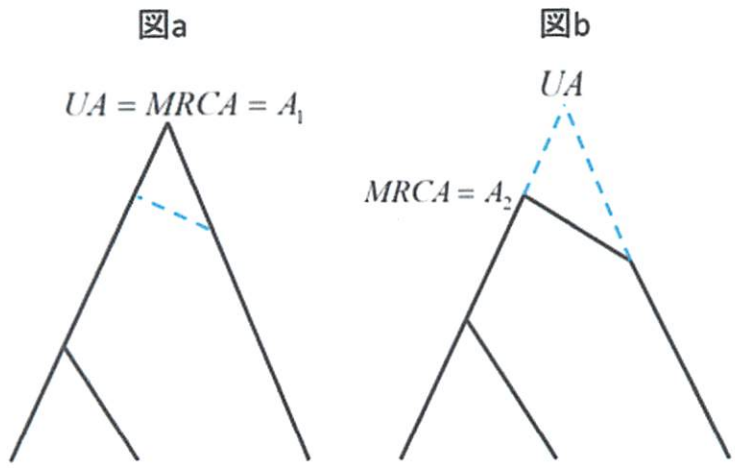
図 5-4 のサンプル数 3 の場合を例として説明する。$A_3(0)=3$ であり、祖先の数 $A_3(t)$ は $3\to2\to3\to2\to1$ と推移し、ジャンプ時刻は $0<T_3^1<T_3^1+T_2^1<T_3^1+T_2^1+T_3^2<T_3^1+T_2^1+T_3^2+T_2^2$ となる。どの個体が合祖あるいは分枝をするかは一様に等確率である。

手順(1)~(3)に従ってＡSG を作成し、その上に突然変異を発生させた後、手順(4)の系図の確定がどのようになされるのか説明しよう。最終祖先(UA)の状態 $Y_1^{(0)}$ を定常分布 $\phi(x)$ に従って選ぶ。時間 T_2^2 はパラメーター $\sigma+1$ の指数分布に従い、各 branch にパラメー

ター$\frac{\theta}{2}$のポアソン過程を発生させて、状態$Y_1^{(1)}, Y_2^{(1)}$を決定する。次に時間T_3^2は再びパラメーター$\sigma+1$の指数分布に従い、各 branch 上に突然変異を発生させて状態$Y_1^{(2)}, Y_2^{(2)}, Y_3^{(2)}$を決定する。ただし、$Y_1^{(2)}, Y_2^{(2)}$は状態$Y_1^{(1)}$を指定した上で独立に決まり、$Y_3^{(2)}$は状態$Y_2^{(1)}$を指定した上で決まる。分枝を生じている箇所で連続路と入来路のどちらの branch が実際の系図であるかということになるが、これを説明したのが表 5-1 である。入来路からのタイプ$Y_2^{(2)}$と連続路からのタイプ$Y_3^{(2)}$のどちらが選択され系図に繋がるのか、その結果が Table 1 のⅢに示されている。入来路が選択されるのは、入来路から分枝点に入ってくる祖先タイプがアレルタイプA_2の場合なので、表 5-1 において、I=1のときは連続路が選択され「結果Ⅲのタイプ=Ⅱのタイプ」となる。他方 I =2 のときは、入来路が選択され、「結果Ⅲのタイプ= I のタイプ」となる。

突然変異が無い場合、最終祖先 UA がアレルタイプA_1であるか、A_2であるかに従って、入来路あるいは連続路のどちらが選択されるのかを示したのが図 5-5 である。濃い実線で示されたのが、確定した系図となる。図 a ではサンプルの 3 個体は全てアレルタイプA_1となり、図 b では、全てアレルタイプA_2となる。

図 5-5



共通祖先(MRCA)に到達する時刻をT_{MRCA}と書く。図 a では$T_{UA} = T_{MRCA}$、図 b では$T_{UA} > T_{MRCA}$である。

５．１．５　合祖時間（Coalescence Time）および Identity by Descent

(1)T_{MRCA}について

明らかに n 個のサンプルについて、　$E_n[T_{MRCA}] \leq E_n[T_{UA}]$が成り立つ。

祖先サイズ過程 $A_n(t)$ の状態の推移を $\vec{A}=(a_0,a_1,\cdots)$ とする。ここで $a_0,a_1,\cdots$ は Ancestral Selection Graph の推移する祖先の数を表す。特に分枝が起きないときのサイズ変化を $\vec{a_0}=(n,n-1,n-2,\cdots,2,1)$ とすると、

$$E_n[T_{MRCA}]=E_n[E_n[T_{MRCA}|A]]=\sum_a E_n[T_{MRCA}|A=a]P_n(A=a)$$
$$\geq E_n[T_{MRCA}|A=\vec{a_0}]P(A=\vec{a_0})$$

ここで、分枝が起こらずに最終祖先に到達する確率 $P(A=a_0)$ は常に Coalesce が起こることなので $P(A=a_0)=\prod_{k=2}^{n}\frac{{}_kC_2}{{}_kC_2+k\sigma/2}$。また、状態 k での滞在時間は平均 $\frac{1}{{}_kC_2+k\sigma/2}$ の指数分布に従う。以上の式を代入すると、次の不等式を得る。

$$E_n[T_{MRCA}]\geq E_n[T_{MRCA}|A=a_0]P(A=a_0)=(\sum_{k=2}^{n}\frac{1}{{}_kC_2+(k\sigma/2)})\{\prod_{k=2}^{n}\frac{{}_kC_2}{{}_kC_2+(k\sigma/2)}) \quad (5.9)$$

(2) Identical by Descent について

Identity by descent (IBD) とは、2個のサンプル遺伝子について、それらの共通祖先(MRCA)まで系図上に突然変異が起こらず状態が同じという意味である。この事象が起こる確率を考えてみよう。2個のサンプル遺伝子について、考えられる ASG を全て考え、それを G で表す。個々のグラフは小文字 g で表示する。その幾つかのグラフが図 5-6 に示されている。さらに各グラフの祖先サイズの変化を $\vec{a}=(a_0,a_1,\cdots)$ とする。グラフ g をそのサイズ変化によって分類し、サイズ変化 $\vec{a}$ に属すグラフの集合を $E_{\vec{a}}$ で表す。このとき、

$$\mathrm{P(IBD)}=E[P(IBD|G)]=\sum_{\vec{a}}\sum_{g\in E_{\vec{a}}}P(IBD|G=g)P(G=g) \quad (5.10)$$

特に $\vec{a}=(2,1)$ のとき　グラフは $g(0)$ のみであり、

$$P(G=g(0))=P(A=(2,1))=\frac{{}_2C_2}{{}_2C_2+\sigma}=\frac{1}{1+\sigma} \quad (5.11)$$

このとき、このグラフ上に共通祖先まで突然変異を起こさない確率は

$$P(IBD\,|G=g(0))=\int_0^\infty e^{-\theta t}(1+\sigma)e^{-(1+\sigma)t}dt=\frac{1+\sigma}{1+\sigma+\theta} \quad (5.12)$$

1回のみ分枝が生じる場合 $\vec{a}=(2,3,2,1)$ となるが、この $\vec{a}$ に属すグラフが図 5-6 の g(1)から g(6)である。g(1)と g(2), g(3)と g(4), g(5)と g(6)は互いに鏡像関係にある。6個のグラフは

等確率で生じるので

$$P(G=g(i))=\frac{1}{6}P\left(\vec{A}=(2,3,2,1)\right)=\frac{1}{6}\frac{\sigma}{1+\sigma}\frac{3}{3+3\sigma/2}\frac{1}{1+\sigma}=\frac{\sigma}{3(1+\sigma)^2(2+\sigma)},\ (i=1,2,..,6)\quad (5.13)$$

各グラフについて、連続路と入来路のどちらを選択するかは突然変異によって決まるアレルタイプによって決まり、サンプル遺伝子の確定した系図上で突然変異が生じていない場合に IBD となる。

図 5-6

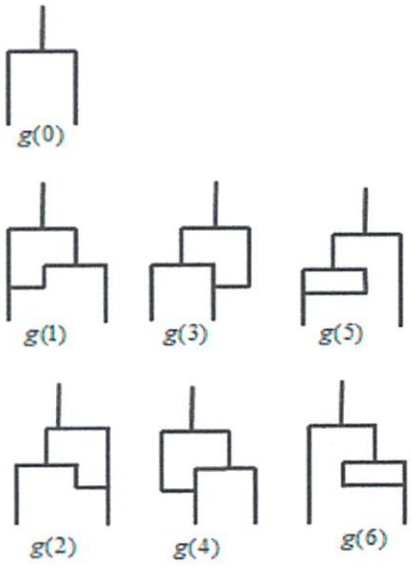


図 5-7

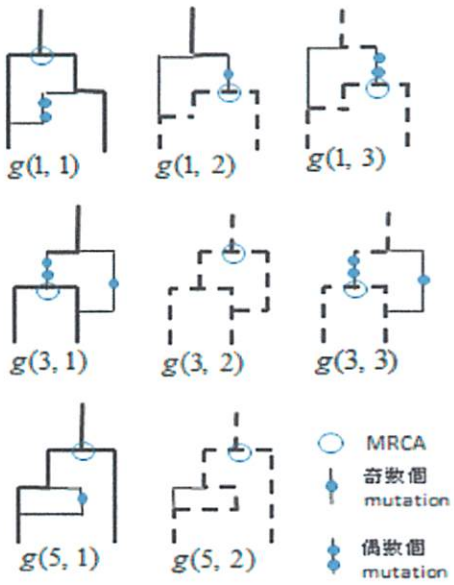


図 5-8（g(1,2)の拡大図）

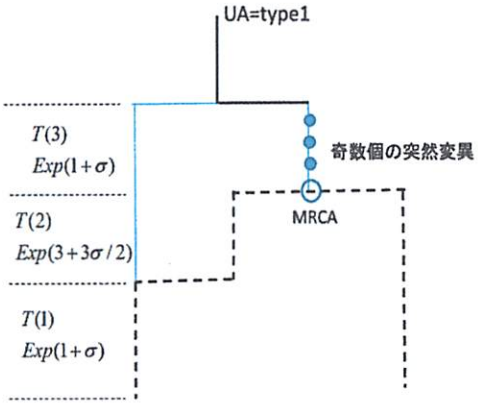


図 5-7, 5-8 について、実線で表されるは突然変異を含まないタイプ A_1 の枝を表し、点線は突然変異を含まないタイプ A_2 の枝を表す。また、枝上の青い点は突然変異を表す。図 5-7 の○は MRCA を表す。

図 5-6 のグラフ g(1)について、確定した系図が連続路あるいは入来路のいずれの枝を使うかによって二つのサンプルが IBD となる 3 つのパターンが生じる。これを表したのが図 5-7 の g(1,1),g(1,2),g(1,3)である。同様に、各グラフ $g(i)$ について、突然変異も含めて IBD となるグラフを $g(i,1),\cdots,g(i,n(i))$ とする。突然変異も含めたグラフの集合を $\tilde{G}$ とすると、

$$P(IBD|G=g(i))=\sum_{j=1}^{n(i)}P(IBD,\ \tilde{G}=g(i,j)|G=g(i)) \qquad (5.14)$$

まず、g(1, 2)について考えてみよう。図 5-8 において各滞在時間 $T(1), T(2), T(3)$ はそれぞれ指数分布 $Exp(1+\sigma), Exp(3+3\sigma/2), Exp(1+\sigma)$ に従う。グラフが g(1)という条件なので、この待ち時間の後にそれぞれ、分枝、合祖、合祖が順に起こっている。Fig 8 より、グラフ g(1,2)の場合は UA がアレルタイプ A_1 なので、UA から MRCA を結ぶ枝上に奇数個の突然変異が生じ、その下の点線上には突然変異は生じていないならば、二つのサンプルはタイプ A_2 で IBD となる。（もし、点で表された突然変異が偶数個ならば、MRCA は A_1 となり、入来路は選択されず、系図は g(1,1)となる。しかし、確定した系図上に突然変異が生じているので、二つのサンプルは IBD ではない。）以上より、時間 t の間に 1 本の枝上に生じる突然変異の数を $M(t)$ と表すと、

$$\begin{aligned}
&P(IBD,\ \tilde{G}=g(1,2)|G=g(1))\\
&=E\left[\exp\left(-\frac{\theta}{2}\times 2(T(1)+T(2))\right)E\left[P(M(T(3))\text{が奇数}|T(3)\right]\right]\times P(UA=1)\\
&=E[\exp(-\theta T(1))]E[\exp(-\theta T(2))]E\left[P(M(T(3))\text{が奇数}|T(3)\right]\times P(UA=1) \qquad (5.15)
\end{aligned}$$

ここで、$E[\exp(-\theta T(1))]=\int_0^\infty e^{-\theta t}(1+\sigma)e^{-(1+\sigma)}dt=\frac{1+\sigma}{1+\sigma+\theta}$

$$E[\exp(-\theta T(2))]=\int_0^\infty e^{-\theta t}\left(3+\frac{3\sigma}{2}\right)e^{-(3+3\sigma/2)t}dt=\frac{3+3\sigma/2}{3+3\sigma/2+\theta}$$

$$E\left[P(M(T(3))\text{が奇数}|T(3)\right]=(1+\sigma)\int_0^\infty P(M(t)\text{が奇数})e^{-(1+\sigma)t}dt$$

ここで、$P(M(t)\text{が奇数})=\sum_{k:\text{奇数}}P(M(t)=k)=\sum_{k:\text{奇数}}e^{-(\theta/2)t}\frac{(\theta t/2)^k}{k!}$

$$=e^{-(\theta/2)t}\frac{e^{(\theta/2)t}-e^{-(\theta/2)t}}{2}=\frac{1-e^{-\theta t}}{2}$$

よって、$E\left[P(M(T(3))\text{が奇数}|T(3)\right]=(1+\sigma)\int_0^\infty\frac{1-e^{-\theta t}}{2}e^{-(1+\sigma)t}dt=\frac{\theta}{2(1+\sigma+\theta)}$

以上の結果を(5.15)に代入すると

$$P(IBD,\tilde{G}=g(1,2)|G=g(1))=\frac{1+\sigma}{1+\sigma+\theta}\times\frac{3+3\sigma/2}{3+3\sigma/2+\theta}\times\frac{\theta}{2(1+\sigma+\theta)}\times P(UA=1) \qquad (5.16)$$

最後に $P(UA=1)$ については、アレル A_1 の頻度を x とすると、$P(UA=1)$ はサンプルした 1 個の遺伝子がアレル A_1 である確率なので、定常分布は $\phi(x)=Kx^{\theta-1}(1-x)^{\theta-1}\exp(-\alpha x)$ より、

$P(UA=1)=\int_0^1 x\phi(x)dx=K\int_0^1 x^\theta(1-x)^{\theta-1}e^{-\alpha x}dx$ となる。ここで、Kummer の合流型超幾何関数

$M(a,b,z)=\dfrac{\Gamma(b)}{\Gamma(b-a)\Gamma(a)}\int_0^1 e^{zt}t^{a-1}(1-t)^{b-a-1}dt$ を用いると、

$$K=1/\int_0^1 e^{-\alpha x}x^{\theta-1}(1-x)^{\theta-1}dx=1/\left\{\frac{(\Gamma(\theta))^2}{\Gamma(2\theta)}M(\theta,2\theta,-\sigma)\right\},$$

$$\int_0^1 x^{\theta}(1-x)^{\theta-1}e^{-\alpha x}dx=\frac{(\Gamma(\theta+1))^2}{\Gamma(2\theta+1)}M(\theta+1,\,2\theta+1,-\sigma)$$

以上より、$P(UA=1)=\dfrac{\Gamma(2\theta)(\Gamma(\theta+1))^2}{\Gamma(2\theta+1)(\Gamma(\theta))^2}\dfrac{M(1+\theta,1+2\theta,-\sigma)}{M(\theta,2\theta,-\sigma)}=\dfrac{1}{2}\dfrac{M(1+\theta,1+2\theta,-\sigma)}{M(\theta,2\theta,-\sigma)}$。 (5.17)

次に g(1,1)を考えてみよう。g(1,1)は連続路を選択しているので、２つのサンプル遺伝子のアレルタイプはA_1であり、○印=UA=MRCA=A_1となる。このとき、確定した系図上には突然変異はなく、入来路には 0 個も含めて偶数個の突然変異が生じている。なぜなら、入来路に奇数個の突然変異が生じた場合、入来路から分枝点に入ってくる祖先のタイプはA_2となり、入来路が選択されて系図は g(1,2)の形となり、しかも IBD ではない。以上より

$$P\left(IBD,\,\tilde{G}=g(1,1)\middle|G=g(1)\right)=E\left[e^{-(\theta/2)\times 2(T(1)+T(2)+T(3))}E\left[P\left(M(T(2))が偶数\middle|T(2)\right)\right]\right]\times P(UA=1)$$

$$P\left(M(t)が偶数\right)=\frac{1+e^{-\theta t}}{2}\ なので$$

$$P\left(IBD,\tilde{G}=g(1,1)\middle|G=g(1)\right)=E\left[e^{-\theta T(1)}\right]E\left[e^{-\theta T(3)}\right]E\left[e^{-\theta T(2)}\times\frac{1+e^{-\theta T(2)}}{2}\right]\times P(UA=1)$$

$$=\left(\frac{1+\sigma}{1+\sigma+\theta}\right)^2\times\left\{\int_0^1 e^{-\theta t}\left(\frac{1+e^{-\theta t}}{2}\right)\times\left(3+\frac{3\sigma}{2}\right)e^{-(3+3\sigma/2)t}dt\right\}\times P(UA=1)$$

$$=\left(\frac{1+\sigma}{1+\sigma+\theta}\right)^2\frac{6+3\sigma+3\theta}{(3+3\sigma/2+\theta)(3+3\sigma/2+2\theta)}\times P(UA=1) \qquad (5.18)$$

同様にして

$$P\left(IBD,\tilde{G}=g(1,3)\middle|G=g(1)\right)=\frac{1+\sigma}{1+\sigma+\theta}\times\frac{3+3\sigma/2}{3+3\sigma/2+\theta}\times\frac{2+2\sigma+\theta}{2+2\sigma+2\theta}\times P(UA=2) \qquad (5.19)$$

ただし、$P(UA=2)=1-P(UA=1)$。

グラフ g(3),g(5)についても同様に求めることができるが、結果は煩雑となるので、詳しくは Krone and Neuhauser(1997)の Appendix B を参照していただきたい。さらに分枝が２回以上生じる系図も考えられるが、分枝の率σが小さいときには、分枝が２回以上起こる確率は$O(\sigma^2)$となる。σが小さいとき Kummer の合流型超幾何関数より

$P(UA=1)=\dfrac{1}{2}-\dfrac{1}{4(1+2\theta)}\sigma+O(\sigma^2)$ なので、以上の結果をまとめると次の定理が成り立つ。

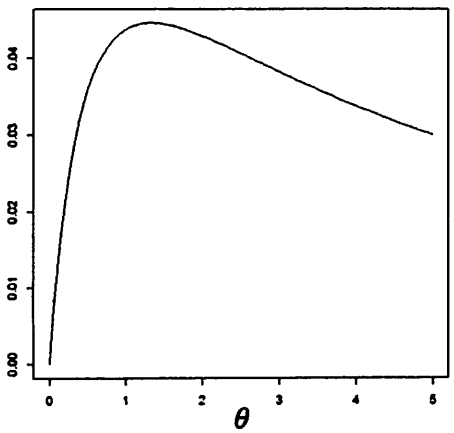
定理５．５

σが小さいとき、$P(IBD)=\frac{1}{1+\theta}-\frac{\theta(5+2\theta)}{4(1+\theta)^2(3+\theta)(3+2\theta)}\sigma+O(\sigma^2)$ (5.20)

σが小さいとき、$P(IBD)$は中立の時の値$\frac{1}{1+\theta}$よりも小さくなるが、

$P(IBD)=\frac{1}{1+\theta}\left\{1-\frac{\theta(5+2\theta)}{4(1+\theta)(3+\theta)(3+2\theta)}\sigma\right\}+O(\sigma^2)$と変形して、

σの係数$f(\theta)=\frac{\theta(5+2\theta)}{4(1+\theta)(3+\theta)(3+2\theta)}$のグラフを描いてみると、



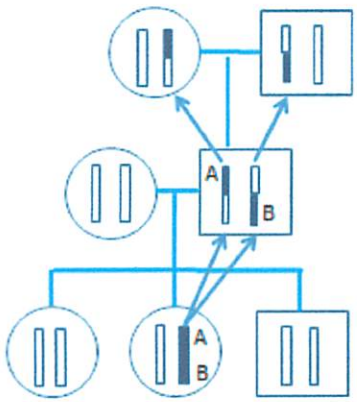
左図のようになり、最大でも0.04を少し超える程度であることが分かる。従って、σが小さいとき中立のときと比べてIBDが減少する割合は、わずかである。

５．２　組み換えを含む遺伝子系図（Ancestral Recombination Graph）

５．２．１．　２遺伝子座モデルと祖先組み換えグラフ(ARG)

この節では同じ染色体上に連鎖した２つの遺伝子座の系図を考える。そこでN個の染色体から成る集団を考えよう。ある個体の１本の染色体を取出し、その上の２つの遺伝子座A,Bに注目する。

図 5-9



この染色体はその父親の二つの染色体の組み換え体であるかもしれない。このとき、染色体の祖先は二つに分枝(branching)する。各々の染色体は祖先を遡ると、さらに分枝あるいは合祖(coalescence)を起こしたりする。

ここでは、二つの遺伝子座 A,B の内部では組み換えは起こさず、遺伝子座間の組み換えのみ考えることにする。

２つの遺伝子座間で１世代あたり、組み換えする確率を$r=\frac{\rho}{2N}$とする。毎世代ランダムに２本の染色体が接合体をつくり、組み換えを上記の確率で起こすものとする。

k 個のサンプルが、１世代で組み換え(recombination)も合祖 coalescent も起きない確率は$(1-r)^k(1-\frac{1}{N})(1-\frac{2}{N})....(1-\frac{k-1}{N})=1-\frac{k\rho}{2N}-\frac{k(k-1)}{2N}+O(\frac{1}{N^2})$となる。組み換えを起こすと親となる染色体（サンプルした染色体上にある遺伝子の少なくとも１個は含んでいる染色体）の数は増加するが、　組み換えのため親が$k+1$になる確率は$\frac{k\rho}{2N}+O(\frac{1}{N^2})$,

合祖のため、ｋ個の個体がｋ－１個の親に由来する確率は$\frac{k(k-1)}{2N}+O(\frac{1}{N^2})$となる。

離散時間において上記の祖先染色体の数に着目したマルコフ連鎖を$A_n^N(Nt)$と表し、$N\to\infty$の極限において得られる連続時間のマルコフ連鎖を$A_n^\rho(t)$で表すことにする。

n 個のサンプルの祖先の数は、推移確率を$P(n,j)=P(A_n^\rho(t)=j\,|\,A_n^\rho(0)=n)$とすると、次のコルモゴロフ後ろ向き方程式で表現される出生死滅過程となる。

$$\frac{d}{dt}P_t(n,j)=\frac{n\rho}{2}P_t(n+1,j)+\frac{n(n-1)}{2}P_t(n-1,j)-\frac{n\rho+n(n-1)}{2}P_t(n,j) \qquad (5.21)$$

方程式(5.21)は自然選択 ASG の(5.5)式と同じ形となる。n 個のサンプルに対して、上記の出生死滅過程においてサンプルの最大祖先数をM_n, 共通な一つの祖先に達するまでの待ち時間をτ_nとする。

補題 5．6

$$E[\tau_n]=\frac{2}{\rho}\int_0^1(\frac{1-x^{n-1}}{1-x})(e^{\rho(1-x)}-1)dx \qquad (5.22)$$

$$P(M_n\le k)=\frac{\sum_{j=n-1}^{k-1}j!\rho^{-j}}{\sum_{j=0}^{k-1}j!\rho^{-j}} \quad ,k\ge n \qquad (5.23)$$

（証明）状態 1 を吸収壁とする出生死亡過程を考え、状態 1 へ吸収されるまでの待ち時間をτ_nとする。$\rho_i=\frac{\lambda_2....\lambda_{i-1}}{\mu_2....\mu_i}$, $i\ge 2$とすると$\lambda_i=\frac{i\rho}{2}$, $\mu_i=\frac{i(i-1)}{2}$より$\rho_i=2\rho^{i-2}/i!$、

第 1 章の定理 1．10より $E[\tau_n]=\sum_{r=1}^{n-1}(\prod_{k=2}^{r}\frac{\mu_k}{\lambda_k})\{\sum_{j=r+1}^{\infty}\rho_j\}$、ただし$r=1$のとき$\prod_{k=2}^{r}\frac{\mu_k}{\lambda_k}=1$

$$E[\tau_n]=\sum_{r=1}^{n-1}(\prod_{k=1}^{r}\frac{k-1}{\rho})\left(\sum_{j=r+1}^{\infty}\frac{2\rho^{j-1}}{j!}\right)=\sum_{r=1}^{n-1}\frac{(r-1)!}{\rho^{r-1}}(\sum_{j=r+1}^{\infty}\frac{2\rho^{j-2}}{j!})=2\sum_{r=1}^{n-1}\sum_{j=r+1}^{\infty}\frac{(r-1)!}{j!}\rho^{j-r-1}$$

$$=2\sum_{m=2}^{n}\sum_{k\ge 0}\rho^k\frac{(m-2)!}{(k+m)!}=\frac{2}{\rho}\int_0^1(\frac{1-x^{n-1}}{1-x})(e^{\rho(1-x)}-1)dx \qquad (j-r-1=k, r+1=m)$$

最大の祖先数の分布を$p_n(k)=P(M_n\le k)$, ただし$p_1(k)=1$, $k\ge 1$ かつ

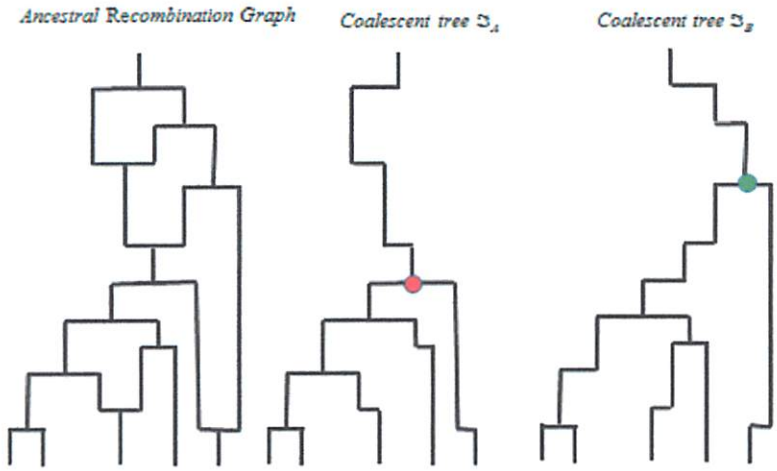
$p_n(k)=0$ *if* $n>k$ とすると Ancestral process の最初のジャンプに注目して次式を得る。

$$p_n(k)=\frac{n-1}{\rho+n-1}p_{n-1}(k)+\frac{\rho}{\rho+n-1}p_{n+1}(k) \qquad (5.24)$$

数学的帰納法により解は(5.23)で与えられることが分かる。

組み換えを考慮に入れた 2 つの遺伝子座の遺伝子系図は合祖(coalescence)と分枝(branching)を含む出生死滅過程となり、図示すると下図の最左図のようになり、祖先組み換えグラフ(Ancestral Recombination Graph)と呼ばれる。略して ARG と表記される。

図 5-10

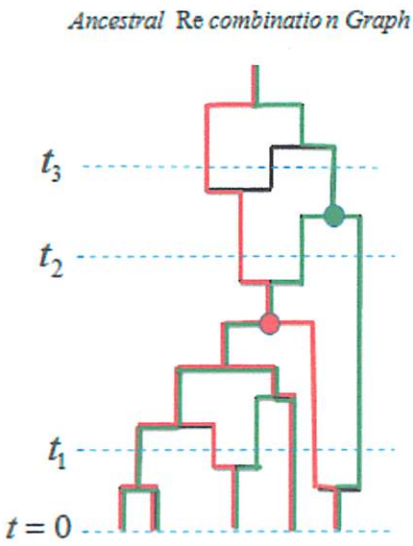


左図の ARG は 5 本のサンプル染色体の系図を表示したものであるが、現在から過去に遡るとき、分枝が生じたときは遺伝子座 A の系図は左の枝を取り、遺伝子座 B の系図は右の枝を辿ることによって得られる。上図の中央および右の図はそれぞれ、そのようにして得られる遺伝子座 A および B の系図を表している。赤丸と緑丸は遺伝子座 A 及び B の MRCA を表している。遺伝子座 A 及び B の系図を $\mathfrak{I}_A$ および $\mathfrak{I}_B$ で表す。

5．2．2．　ARG の構造

ARG の系図を G、グラフ G の枝の集合を $\varepsilon(G)$ で表す。また、遺伝子座 A,B の系図の枝の集合を $\varepsilon(\mathfrak{I}_A)$, $\varepsilon(\mathfrak{I}_B)$ で表示する。遺伝子座 A, B の系図をそれぞれ赤および緑で着色して ARG の図に重ねると下図のようになる。

図 5-11



赤だけの枝、緑だけの枝、赤と緑が重なった枝、何も着色されていない黒い枝の 4 種類に分類される。
すなわち、次の記号で表される。

$$\tilde{\mathrm{A}} = \varepsilon(\mathfrak{I}_A) \cap \varepsilon(\mathfrak{I}_B)^C \ ; \tilde{\mathrm{B}} = \varepsilon(\mathfrak{I}_A)^C \cap \varepsilon(\mathfrak{I}_B) \ ;$$
$$\tilde{C} = \varepsilon(\mathfrak{I}_A) \cap \varepsilon(\mathfrak{I}_B) \ ; \tilde{D} = \varepsilon(G) \cap \varepsilon(\mathfrak{I}_A)^C \cap \varepsilon(\mathfrak{I}_B)^C$$

ある時刻tにおけるARGの切片を考え、その枝との交点の集合を$\varepsilon(G_t)$とする。
その時刻における4種の枝$\tilde{A},\tilde{B},\tilde{C},\tilde{D}$との交点の数を以下のように表す。

$$n_{\tilde{A}}(t)=\left|\varepsilon(G_t)\cap\tilde{\mathrm{A}}\right|,\ n_{\tilde{B}}(t)=\left|\varepsilon(G_t)\cap\tilde{\mathrm{B}}\right|,\ n_{\tilde{C}}(t)=\left|\varepsilon(G_t)\cap\tilde{C}\right|,\ n_{\tilde{D}}(t)=\left|\varepsilon(G_t)\cap\tilde{D}\right|$$

$m(t)=(n_{\tilde{A}}(t),n_{\tilde{B}}(t),n_{\tilde{C}}(t),n_{\tilde{D}}(t)),\ t\geq 0$とする。
例えば上図において、$m(0)=(0,0,5,0),\ m(t_1)=(2,2,2,0),\ m(t_2)=(1,2,0,0),$
$m(t_3)=(1,1,0,1)$となる。

明らかに、$n_{\tilde{A}}(t)+n_{\tilde{B}}(t)+n_{\tilde{C}}(t)+n_{\tilde{D}}(t)=\left|\varepsilon(G_t)\right|=A_n^{\rho}(t)$であり、

$n_{\tilde{A}}(t)+n_{\tilde{C}}(t)=\left|\varepsilon(\Im_A(t))\right|\equiv A_n(t),\ \ n_{\tilde{B}}(t)+n_{\tilde{C}}(t)=\left|\varepsilon(\Im_B(t))\right|\equiv B_n(t)$、$A_n(t),B_n(t)$は遺伝子座A,Bの系図の祖先の数を表している。

マルコフ過程$m(t)$の生成作用素（すなわち推移率）は以下のようになる。
$m(t)=(a,b,c,d)$とすると、

$$\begin{cases} (a+1,b+1,c-1,d) & & c\rho/2\ \ (\mathrm{Rec}) \\ (a-1,b-1,c+1,d) & & ab\ \ (Coal) \\ (a-1,b,c,d) & & ac+a(a-1)/2\ \ (Coal) \\ (a,b-1,c,d) & at\ rate & bc+b(b-1)/2\ \ (Coal) \\ (a,b,c-1,d) & & c(c-1)/2\ \ (Coal) \\ (a,b,c,d+1) & & (a+b+d)\rho/2\ \ (\mathrm{Rec}) \\ (a,d,c,d-1) & & d(a+b+c)+d(d-1)/2\ \ (Coal) \end{cases} \tag{5.25}$$

(Rec：組み換え(Recombination)，　Coal：合祖(Coalescence))

$n(t)=(n_{\tilde{A}}(t),n_{\tilde{B}}(t),n_{\tilde{C}}(t))$とすると、この3成分の変化に$d$は寄与しないので$n(t)$はマルコフ連鎖になっており、$n(t)=(a,b,c)$とすると、その生成作用素は次式で与えられる。

$$(a_1,b_1,c_1)=\begin{cases} (a+1,b+1,c-1) & r_1=c\rho/2 \\ (a-1,b-1,c+1) & r_2=ab \\ (a-1,b,c) \quad at\ rate & r_3=ac+a(a-1)/2 \\ (a,b-1,c) & r_4=bc+b(b-1)/2 \\ (a,b,c-1) & r_5=c(c-1)/2 \end{cases} \tag{5.26}$$

総推移率は$d_n=r_1+r_2+r_3+r_4+r_5=\dfrac{c\rho}{2}+\dfrac{n(n-1)}{2},\quad n=a+b+c$となり、(5.21)式の推移率と一致する。

５．２．３．　遺伝子系図の全枝長の相関

L^A, L^B を遺伝子 A,B の系図 $\Im_A, \Im_B$ の全枝長とする。さらに、この共分散を $F(a.b.c;\rho)=Cov(L^A,L^B)$, ただし $n(0)=(a,b,c)$ と置く。連続時間マルコフ連鎖 $n(t)$ のジャンプ過程を $N(.)$ とする。(5.26)に現れる５つの状態を取るベクトル値確率変数 Z を状態(a,b,c)からのジャンプ過程の推移確率より、次式のように定義する。

$$p_i = P(Z=z_i)=\frac{r_i}{d_n}, \quad z_i=(a_1,b_1,c_1) \quad ,i=1,2,3,4,5 \tag{5.27}$$

補題５．７ (Pluzhnikov(1997))

$Z=(a_1,b_1,c_1)$ の条件の下で　　$L^A = X_A + T_A$ (5.28)

ただし　（ｉ）$X_A \sim L_{a_1+c_1}$,　　（L_m は m-coalescent の全枝長）

（ii）　$T_A \sim n_1 T$,　（Tは $Exp(d_n)$ に従う確率変数,　$n_1=a+c$）

（iii）　X_A, T_A は独立である。

同様に　$L^B = X_B + T_B \sim L_{b_1+c_1} + n_2 T$,　$n_2=b+c$。ここで $X \sim Y$ は X と Y が同じ分布に従うという意味である。

(証明)

X_A はサンプル数 a_1+c_1 のA遺伝子の系図の全長であり、初期状態 (a,b,c) での滞在時間Tはパラメーター d_n の指数分布に従うので状態 (a,b,c) から (a_1,b_1,c_1) の間の全枝の長さの和は n_1T。従って（ｉ）（ii）が成り立ち、明らかに X_A, T_A は独立である。

$n_1=a+c\le 1$ のとき、遺伝子座 A は MRCA に到達しているので、$L_A=0$。よって $F(a.b.c;\rho)=Cov(L^A,L^B)=0$、同様に $n_2=b+c\le 1$ のとき $F(a.b.c;\rho)=Cov(L^A,L^B)=0$ である。また、$a<0$, $b<0$, $c<0$ のいずれかが成り立つとき、$F(a.b.c;\rho)=Cov(L^A,L^B)=0$ と定義する。これより、次の定理を得る。

定理５．８

共分散 $F(a,b,c;\rho)=Cov(L^A,L^B)$, $\rho\in[0,\infty)$ は以下の方程式を満たす。

$$\begin{aligned} d_n F(a,b,c;\rho) &= r_1 F(a+1,b+1,c-1;\rho)+r_2 F(a-1,b-1,c+1;\rho) \\ &\quad + r_3 F(a-1,b,c,;\rho)+r_4 F(a,b-1,c;\rho)+r_5 F(a,b-1,c;\rho)+R_n \end{aligned} \tag{5.29}$$

ただし、$n=a+b+c$, $d_n=\dfrac{n(n-1)+c\rho}{2}$, $R_n=\dfrac{2c(c-1)}{(n_1-1)(n_2-1)}$, r_i は(5.26)で定義される定数。

方程式(5.29)は境界条件：「$n_1\le 1$, $n_2\le 1$, $a<0$, $b<0$, $c<0$ のいずれかが成り立つとき、$F(a,b,c;\rho)=0$」の下で一意に解を持つ。

（証明）

一般に、確率変数 X, Y, Z について

$$Cov(X,Y) = E[(X-\bar{X})(Y-\bar{Y})] = E[E[(X-\bar{X})(Y-\bar{Y})|Z]]$$
$$= E[E[\{(X-E[X|Z])+(E[X|Z]-\bar{X})\}\{(Y-E[Y|Z])+(E[Y|Z]-\bar{Y})\}|Z]]$$
$$= E[E[\{X-E[X|Z]\}\{Y-E[Y|Z]\}|Z]] + E[E[\{E[X|Z]-\bar{X}\}\{E[Y|Z]-\bar{Y}\}|Z]]$$
$$= E[Cov(X,Y|Z)] + Cov(E[X|Z], E[Y|Z])$$

ただし $E[E[\{X-E[X|Z]\}\{E[Y|Z]-\bar{Y}\}|Z]] = E[\{E[Y|Z]-\bar{Y}\}E[X-E[X|Z]|Z]] = 0$

同様に $E[E[\{Y-E[Y|Z]\}\{E[X|Z]-\bar{X}\}|Z]] = 0$ を使った。

よって $Cov(X,Y) = E(Cov(X,Y|Z)) + Cov(E(X|Z), E(Y|Z))$,

ここで $X = L^A$, $Y = L^B$ および Z を(5.27)で定義した確率変数とすると

$$E[\mathrm{cov}(X,Y|Z)] = \sum_{i=1}^{5} p_i Cov(X,Y|Z=z_i),$$
$$Cov(X,Y|Z=z_1) = Cov(X_A+T_A, X_B+T_B) = Cov(X_A,X_B) + Cov(T_A,T_B)$$
$$= F(a+1,b+1,c-1;\rho) + n_1 n_2 Var(T)$$
$$= F(a+1,b+1,c-1;\rho) + \frac{n_1 n_2}{d_n^{\,2}} \qquad (5.30)$$

以下同様にして

$$E[Cov(X,Y|Z)] = r_1 F(a+1,b+1,c-1;\rho) + r_2 F(a-1,b-1,c+1;\rho)$$
$$+ r_3 F(a-1,b,c,;\rho) + r_4 F(a,b-1,c;\rho) + r_5 F(a,b-1,c;\rho) + n_n n_2 d_n^{-2} \qquad (5.31)$$

$f(Z) = E[X|Z] - \bar{X}$, $g(Z) = E[Y|Z] - \bar{Y}$ とすると $f(Z), g(Z)$ の確率分布は以下のようになる。

Z	$f(Z)$	$g(Z)$	$P(Z=z_i)$
$(a+1,b+1,c-1)$	n_1/d_n	n_2/d_n	p_1
$(a-1,b-1,c+1)$	n_1/d_n	n_2/d_n	p_2
$(a-1,\ b,\ c)$	$n_1/d_n - 2/(n_1-1)$	n_2/d_n	p_3
$(a,\ b-1,c)$	n_1/d_n	$n_2/d_n - 2/(n_2-1)$	p_4
$(a,b,c-1)$	$n_1/d_n - 2/(n_1-1)$	$n_2/d_n - 2/(n_2-1)$	p_5

これより、

$$Cov(E[X|Z],E[Y|Z]) = E[f(Z)g(Z)] = \sum_{i=1}^{5} p_i f(z_i)g(z_i)$$

$$= -\frac{n_1 n_2}{d_n^{\,2}} + \frac{2c(c-1)}{d_n(n_1-1)(n_2-1)} \qquad (5.32)$$

(5.31)と(5.32)を合わせて(5.29)を得る。

$F(a,b,c;\rho)$ の次数を $a+b+2c$ で定義すると(10.2.4)の右辺の最高次数は左辺の次数に等しい。従って、低次のものから順次解くことができる。よって解は一意である。

特に $a+b+2c=4$ となる組み合わせは $(a,b,c)=(0,0,2),(2,2,0),(2,0,1),(0,2,1),(1,1,1)$ の５通りあるが、境界条件より明らかに $F(2,0,1;\rho)=F(0,2,1;\rho)=0$ である。残りの３つについて(5.29)より方程式を立てると

$$\begin{cases} (1+\rho)F(0,0,2;\rho) = \rho F(1,1,1;\rho)+4 \\ 6F(2,2,0;\rho) = 4F(1,1,1;\rho) \\ (3+\rho/2)F(1,1,1;\rho) = (\rho/2)F(2,2,0;\rho)+F(0,0,2;\rho) \end{cases}$$

これより、

$$F(0,0,2;\rho) = \frac{4(\rho+18)}{\rho^2+13\rho+18}, \quad F(1,1,1;\rho) = \frac{24}{\rho^2+13\rho+18}, \quad F(2,2,0;\rho) = \frac{16}{\rho^2+13\rho+18} \qquad (5.33)$$

最後に共通祖先に到達するまでに生じる組み換えの数について考える。

R_n を共通な一つの祖先に到達するまでに起こる組み換えの数とすると **Ancestral process** は出生率 $\lambda_i = \frac{i\rho}{2}$、死亡率 $\mu_i = \frac{i(i-1)}{2}$ の出生死亡過程なので、そのジャンプチェーンの推移確率は $p(k,k+1) = \frac{\rho}{\rho+k-1}$, $P(k,k-1) = \frac{k-1}{\rho+k-1}$。これより次の定理が成り立つ。

定理５．９

$$E[R_n] = \rho\int_0^1 \frac{1-(1-x)^{n-1}}{x}\exp(\rho x)dx \qquad (5.34)$$

(証明)

$E(n) = E[R_n]$ とすると　　$E(n) = \frac{\rho}{n-1+\rho}\{1+E(n+1)\} + \frac{n-1}{n-1+\rho}E(n-1)$

これは吸収壁がある場合の吸収までの待ち時間の式(A.6)と類似の方程式である。

従って(5.22)と同様にして(5.34)が導かれる。また、数学的帰納法により(5.34)が上記の方程式を満たすことも示される。

さらにサンプル数 n に対する組み換え数 R_n の母関数を $E_n[\varsigma^R]$, $(0 \le \varsigma \le 1)$ とすると

$E_n[\varsigma^R] = Q_n(\varsigma)/Q_1(\varsigma)$,　$Q_n(\varsigma) = x^{\rho(1-\varsigma)-1}(1-x)^{n-1}e^{-\rho\varsigma(1-x)}dx$　と表される。

(証明は Ethier,S.N. and Griffiths,R.C.(1990))
組み換えを考慮にいれた２つの遺伝子の系図が求まると、その上にポアソン過程に従って突然変異を生じさせることができる。１遺伝子座の場合と同様にサンプリング公式等について議論できるが、詳しくは Ethier and Griffiths(1990)を参照されたい。