８．　無限サイトモデルと中立仮説の検定法

Kingman, Tajima 等によって導入された合祖理論は中立な遺伝子の系図であり、中立説は木村資生によって 1968 年に提唱され、分子進化のメカニズムを説明する説として注目され多くの研究がされてきた。また自然選択の有無を検定するための帰無仮説としても、中立仮説は重要な役割を果たしている。遺伝子系図を利用して、中立仮説あるいは自然選択、個体群動態などに関する検定法が提唱されているが、この章では中立な遺伝子の系図と関連した諸量の分布およびそれを利用した中立仮説の検定法について紹介する。

８．１　サイト頻度および突然変異の年齢

生物のゲノム、例えばヒトゲノムでは３０億個もの塩基から成り、その上に起こる突然変異は、同じサイトに生じることは非常にまれである。ゲノムが無限に長い塩基サイトから成り、同じサイトに突然変異が生じる確率は０と仮定するモデルが良く用いられ、無限サイトモデルと呼ばれている。生物集団の多様性、遺伝子系図が対象としている時間スケールではこの無限サイトモデルが良く利用されている。無限サイトモデルでは、系図上に生じる突然変異は常に異なるサイトに起こる新しい突然変異であり、突然変異数は分離サイトの数と等しくなる。Griffiths and Tavare(1998)は系図上に生じたある突然変異サイトのサンプル中頻度が与えられたとき、その変異の年齢分布、その塩基が祖先型である確率などサンプルの系図と関連した種々の量の分布を求めている。集団から n 個の遺伝子をサンプルしたとき、合祖過程に従い祖先の数は次第に減少して行くが、祖先の数が k である状態の滞在時間を第３章の記法に従い τ_k で表すことにする。Griffiths and Tavare(1998)では第３章５節で紹介した集団サイズの変動も考慮に入れた、以下の三つの条件の下で諸量の分布を考察している。

（１）$\tau_n,\cdots,\tau_2$ は連続確率変数である。

（２）遺伝子系図は、分岐は２分岐であり、k 個の祖先が $k-1$ の状態へ合祖する確率は可換性により、どのペアについても $\binom{k}{2}^{-1}$ である。

（３）突然変異は系図の枝上に単位時間当たり $\theta/2$ の率のポアソン過程に従って生じる。

８．１．１　サンプル中のある分離サイトにおける変異塩基の数の分布

集団から n 個のサンプルを取り出し、それらが k 個の祖先に由来すると仮定する。祖先により k 個の同値類に分割される場合の数は、可換性より区別できない n 個のボールを空の箱が無いように、区別できる k 個の箱（クラス）に分ける場合の数に等しい。これは

$r_1+\cdots+r_k=n,\ r_i\geq 1(i=1,\cdots,k)$ を満たす自然数解 $(r_1,\cdots,r_k)$ の個数に等しく $\binom{n-1}{k-1}$ 通りある。その中である特定のクラス（箱）の中にb個のサンプルが入っている場合の数は b 個のサンプルを除いて、$n-b$ 個のサンプルを $k-1$ 個のクラスに分ければよいので $\binom{n-b-1}{k-2}$ 通り、よって k 個の祖先に由来するとき、ある特定のクラスのサイズが b である確率 $P_{n,k}(b)$ は、

$$P_{n,k}(b)=\frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \quad ,1\leq b\leq n-1。 \tag{8.1}$$

ここで、サンプル頻度 $\frac{b}{n}=x$ を一定として、極限 $n\to\infty$ を取ると次式が成り立つ。

$$\begin{aligned}\lim_{n\to\infty} n\times P_{n,k}(b)&=\lim_{n\to\infty}\frac{n\times(n-b-1)!}{(k-2)!(n-b-k-1)!}\times\frac{(k-1)!(n-k)!}{(n-1)!}\\&=\lim_{n\to\infty}(k-1)\times n\times\frac{(n-b-1)(n-b-2)\cdots(n-b-k+2)}{(n-1)(n-2)\cdots(n-k+1)}\\&=\lim_{n\to\infty}(k-1)\left(1-\frac{b}{n-1}\right)\cdots\left(1-\frac{b}{n-k+2}\right)\times\frac{n}{n-k+1}=(k-1)(1-x)^{k-2}\end{aligned} \tag{8.2}$$

(8.1)及び(8.2)は合祖モデルの可換性と組み合わせ論的議論だけで得られるもので、第3章5節で紹介した集団サイズが変動する場合にも適用できる。

n 個のサンプルを取り出したとき、ある塩基サイトが b 個の突然変異と $n-b$ 個の祖先型に分離している確率を $q_{n,b}$ とすると、次の定理が特定成り立つ。

定理８．１

$$q_{n,b}=\frac{\sum_{k=2}^{n}kP_{n,k}(b)E[\tau_k]}{\sum_{k=2}^{n}kE[\tau_k]}=\frac{(n-b-1)!(b-1)!\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k]}{(n-1)!\sum_{k=2}^{n}kE[\tau_k]} \tag{8.3}$$

（証明）Ethier and Griffiths(1987)に従って、DNA塩基配列を単位長さの線分(0, 1)で表すことにする。新しい突然変異の位置はこの線分上に一様分布で現れる。この線分の部分集合 $M\subset(0,1)$ に対して、$\theta|M|/2$ の発生率で領域 M に突然変異が生じる。すなわち、系図上の長さ L の枝にこの領域 M 上の突然変異が生じる確率はパラメーター $\frac{\theta|M|L}{2}$ のポアソン分布に従うと仮定する。$T=(\tau_2,\cdots,\tau_n)$ を各状態での滞在時間の列、$C_h=C(y,h,b)$ を

区間$(y,y+h)\subseteq(0,1)$の領域に b 個のサンプルを持つ突然変異が生じる事象、またI_kを合祖過程で k 個の祖先の状態のときに突然変異が生じる事象とする。$\mathrm{T}=(\tau_2,\cdots,\tau_n)$を与えたとき、b個のサンプルを持つ突然変異を$U$で表すと、事象$C_h$の条件付き確率$P(C_h|T)$は

$$P(C_h|T)=\sum_{k=2}^{n}P(C_h\cap I_k|T)=\sum_{k=2}^{n}P_{n,k}(b)P(I_k,U\in(y,y+h)|\mathrm{T})$$
$$=\sum_{k=2}^{n}P_{n,k}(b)\left(k\tau_k\frac{\theta}{2}h+o(h)\right)$$

Tの分布について期待値を取ると　$P(C_h)=\sum_{k=2}^{n}P_{n,k}(b)\times\frac{\theta h}{2}\times kE[\tau_k]+o(h)$。　(8.4)

bについて加えると、突然変異Uが区間$(y,y+h)$に生じる確率になるので

$$P(U\in(y,y+h))=\frac{\theta h}{2}\left(\sum_{k=2}^{n}kE[\tau_k]\right)+o(h)。\qquad(8.5)$$

(8.4)を(8.5)で割り$h\to 0$とすると、突然変異によりある分離サイトが生じたとき、そのサイトでn個のサンプル中に突然変異がb個存在している確率$q_{n,b}$が得られる。よって

$q_{n,b}=\dfrac{\sum_{k=2}^{n}kP_{n,k}(b)E[\tau_k]}{\sum_{k=2}^{n}kE[\tau_k]}$、さらに(8.1)を代入し整理すると(8.3)が得られる。　(証明終わり)

定理８．１は集団サイズが変動する場合も成り立つが、集団サイズが一定のときはτ_kは指数分布$Exp(k(k-1)/2)$に従うので、$E[\tau_k]=\dfrac{2}{k(k-1)}$。これより

(8.3)の分子$=(n-b-1)!(b-1)!\sum_{k=2}^{n}2\binom{n-k}{b-1}=2(n-b-1)!(b-1)!\binom{n-1}{b}$となる。

ただし、$\sum_{k=2}^{n}\binom{n-k}{b-1}=\sum_{k=2}^{n-b-1}\binom{n-k}{b-1}=\binom{n-1}{b}$を利用した。

(8.3)の分母$=(n-1)!\sum_{k=2}^{n}\frac{2}{k-1}=2(n-1)!\sum_{j=1}^{n-1}\frac{1}{j}$、これより

$$q_{n,b}=\frac{(n-b-1)!(b-1)!}{(n-1)!}\times\frac{\binom{n-1}{b}}{\sum_{j=1}^{n-1}j^{-1}}=\frac{1}{b\sum_{j=1}^{n-1}1/j},\quad b=1,2,\cdots,n-1\qquad(8.6)$$

平均μ及び分散σ^2は集団サイズ一定のとき、

$\mu=(n-1)/\sum_{j=1}^{n-1}j^{-1}$, $\sigma^2=n(n-1)/\left(2\sum_{j=1}^{n-1}j^{-1}\right)-\left((n-1)/\sum_{j=1}^{n-1}j^{-1}\right)^2$ となり、nが大きいとき、

$\mu\sim\frac{n}{\log n}$, $\sigma^2\sim\frac{n^2}{2\log n}$が得られる。

8．1．2　分離サイトにおける祖先型確率

n個のサンプル遺伝子の塩基配列中に分離したサイトがあるとき、そのサイトの2種の塩基のどちらが祖先型でどちらが突然変異であるのか、n個の遺伝子の共通祖先MRCAの塩基配列が既知の場合は確定するが、祖先型が不明の場合は祖先型である確率を求めることは興味ある問題である。ある分離したサイトに2種の塩基A,B（実際にはA,T,G,Cの中の2種の塩基）が存在し、Aがa個、Bが$b=n-a$個であったとき、Aが祖先型である確率$P(a,b)$は無限サイトモデルではそのサイトではAまたはBが祖先型であるので、A (B)が祖先型である事象を$E_A(E_B)$とすると求める確率は$P(a,b)=\frac{P(E_A)}{P(E_A)+P(E_B)}$となる。また、

$P(E_A)$＝Bが突然変異である確率＝$q_{n,b}$なので定理8．1より

$$P(a,b)=\frac{q_{n,b}}{q_{n,b}+q_{n,a}}=\frac{\sum_{k=2}^{n}kP_{n,k}(b)E[\tau_k]}{\sum_{k=2}^{n}kP_{n,k}(b)E[\tau_k]+\sum_{k=2}^{n}kP_{n,k}(a)E[\tau_k]}$$
$$=\frac{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k]}{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k]+\sum_{k=2}^{n}k(k-1)\binom{n-k}{a-1}E[\tau_k]} \qquad (8.7)$$

$\frac{a}{n}=x$を固定し、$n\to\infty$の極限を取ると(8.2)より$\lim_{n\to\infty}nP_{n,k}(b)=(k-1)x^{k-2}$,

$\lim_{n\to\infty}nP_{n,k}(a)=(k-1)(1-x)^{k-2}$よって$\lim_{n\to\infty}P(a,b)=P_\infty(x)$とすると、

$$P_\infty(x)=\frac{\sum_{k=2}^{n}k(k-1)x^{k-2}E[\tau_k]}{\sum_{k=2}^{n}k(k-1)(1-x)^{k-2}E[\tau_k]+\sum_{k=2}^{n}k(k-1)x^{k-2}E[\tau_k]}。 \qquad (8.8)$$

集団サイズが一定のとき、$E[\tau_k]=\frac{2}{k(k-1)}$より$P(a,b)=\frac{a}{a+b}=\frac{a}{n}$, $P_\infty(x)=x$となる。

この結果は「頻度xの対立遺伝子(allele)が最も古い対立遺伝子である確率はxである」という Watterson and Guess(1977)の古典的結果と一致する。(8.8)はこの結果の一般化ということができる。

８．１．３　分離サイトにおける突然変異の年齢

分離サイトにおける突然変異の年齢を考察しよう。$\xi_{n,b}$でn個のサンプル中にb個（ただし$1 \le b \le n-1$）存在する突然変異の年齢とする。$S_k = \sum_{i=k}^{n} \tau_k$ とする。このサイトの突然変異がサンプルの祖先がk個の状態の時に生じたと仮定する。すなわち滞在時間τ_kの間に発生したと仮定すると、$\tau_k = t$という条件の下で、突然変異が発生した時点は$[0,t]$上の一様分布に従う（カーリン「確率過程講義」p202 参照)。従ってUを$[0,1]$上の一様分布に従う確率変数とすると、τ_kおよびS_{k+1}が与えられたとき、確率変数$\xi_{n,b}$は$\xi_{n,b} = U\tau_k + S_{k+1}$と表せる。$\xi_{n,b}$のラプラス変換を$E[\exp(-\lambda\xi_{n,b})]$とする。定理８．１の証明と同様に突然変異が生じる時点で場合分けして考察すると、次式を得る。

$$E[\exp(-\lambda\xi_{n,b})] = \frac{\sum_{k=2}^{n} kP_{n,k}(b)E[T_k \exp(-\lambda(U\tau_k + S_{k+1})]}{\sum_{k=2}^{n} kP_{n,k}(b)E[\tau_k]} \tag{8.9}$$

ここで、τ_k, S_{k+1}が与えられた条件下でUは$[0,1]$上の一様分布に従うので

$$\begin{aligned} E[T_k \exp(-\lambda(U\tau_k + S_{k+1})] &= E[T_k \exp(-\lambda S_{k+1})E[\exp(-\lambda T_k U)|T_k, S_{k+1}]] \\ &= E\left[T_k \exp(-\lambda S_{k+1})\int_0^1 \exp(-\lambda T_k u)du\right] \\ &= \frac{1}{\lambda}E[\exp(-\lambda S_{k+1})\{1-\exp(-\lambda T_k)\}] \end{aligned} \tag{8.10}$$

ラプラス逆変換により年齢$\xi_{n,b}$の分布密度$g_{n,b}(t)$が次のように得られる。

定理８．２

$A_n(t)$をサンプル数nの合祖過程の過去に遡り時刻tでの祖先の数とする。

ただし、集団サイズが変動する場合（第３章第５節）も含むものとする。

$$g_{n,b}(t)=\frac{\sum_{k=2}^{n}kP_{n,k}(b)P\left(A_n(t)=k\right)}{\sum_{k=2}^{n}kP_{n,k}(b)E[\tau_k]}=\frac{E\left[A_n(t)(A_n(t)-1)\binom{n-A_n(t)}{b-1}\right]}{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k]} \tag{8.11}$$

（証明） $P(A_n(t)=k)=P(S_{k+1}\le t)-P(S_k\le t)$に注意すると、

$$\lambda^{-1}E[\exp(-\lambda S_{k+1})(1-\exp(-\lambda T_k))]=\lambda^{-1}E[\exp(-\lambda S_{k+1})-\exp(-\lambda S_k)]$$
$$=\lambda^{-1}\left\{\int_0^\infty e^{-\lambda t}\frac{d}{dt}P(S_{k+1}\le t)dt-\int_0^\infty e^{-\lambda t}\frac{d}{dt}P(S_k\le t)dt\right\}$$

部分積分により

$$=\int_0^\infty e^{-\lambda t}\left(P(S_{k+1}\le t)-P(S_k\le t)\right)dt=\int_0^\infty e^{-\lambda t}P(A_n(t)=k)dt \tag{8.12}$$

(8.9)に(8.10),(8.12)を代入し、ラプラス逆変換を行うと、(8.11)の最初の等式を得る。

$P_{n,k}(b)=\dfrac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}=\dfrac{(n-b-1)!(b-1)!}{(n-1)!}\times(k-1)\binom{n-k}{b-1}$と書けるので

$$\text{分子}=\sum_{k=2}^{n}kP_{n,k}(b)P(A_n(t)=k)=\frac{(n-b-1)!(b-1)!}{(n-1)!}\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}P(A_n(t)=k)$$
$$=\frac{(n-b-1)!(b-1)!}{(n-1)!}E\left[A_n(t)(A_n(t)-1)\binom{n-A_n(t)}{b-1}\right]$$

同様に分母も$\dfrac{(n-b-1)!(b-1)!}{(n-1)!}\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k]$となるので、(8.11)を得る。

定理８．２より$\xi_{n,b}$のモーメントを求めることができる。j次のモーメントは

(8.11)の分母を$C_{n,b}$で表すと

$$E[(\xi_{n,b})^j]=\int_0^\infty t^j g_{n,b}(t)dt=\frac{1}{C_{n,b}}\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}\left\{\int_0^\infty t^j P(A_n(t)=k)dt\right\}$$

$$=\frac{1}{C_{n,b}}\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}\left\{\int_0^\infty t^j\left(P(S_{k+1}\le t)-P(S_k\le t)\right)dt\right\}$$

部分積分により

$$\int_0^\infty t^j\left(P(S_{k+1}\le t)-P(S_k\le t)\right)dt$$

$$=\left[\frac{t^j}{j+1}\left(P(S_{k+1}\le t)-P(S_k\le t)\right)dt\right]_0^\infty-\int_0^\infty\frac{t^{j+1}}{j+1}\left\{\frac{d}{dt}P(S_{k+1}\le t)-\frac{d}{dt}P(S_k\le t)\right\}dt$$

$$=\frac{1}{j+1}E\left[S_k^{j+1}-S_{k+1}^{j+1}\right]$$

これを上式に代入すると

$$E\left[(\xi_{n,b})^j\right]=\frac{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}\frac{1}{j+1}E\left[S_k^{j+1}-S_{k+1}^{j+1}\right]}{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k]}\qquad(8.13)$$

$(k-1)\binom{n-k}{b-1}=\frac{(n-1)!}{(n-b-1)!(b-1)!}P_{n,k}(b)$ 及び $\lim_{n\to\infty}nP_{n,k}(b)=(k-1)(1-x)^{k-2}$ より

$\frac{b}{n}=x$ を固定して、$n\to\infty$ の極限を取り $\lim_{n\to\infty}E\left[(\xi_{n,b})^j\right]=E\left[(\xi_x)^j\right]$ とすると、頻度 x の突然変異遺伝子の年齢の j 次モーメントが得られる。

$$E[((\xi_x)^j]=\frac{\sum_{k=2}^{\infty}k(k-1)(1-x)^{k-2}\frac{1}{j+1}E\left[S_k^{j+1}-S_{k+1}^{j+1}\right]}{\sum_{k=2}^{\infty}k(k-1)(1-x)^{k-2}E[\tau_k]}\qquad(8.14)$$

定理８．２より頻度 x の突然変異の年齢の分布密度は

$$g_x(t)=\frac{\sum_{k=2}^{\infty}k(k-1)(1-x)^{k-2}P(A(t)=k)}{\sum_{k=2}^{\infty}k(k-1)(1-x)^{k-2}E[\tau_k]}=\frac{E[A(t)(A(t)-1)(1-x)^{A(t)-2}]}{\sum_{k=2}^{\infty}k(k-1)(1-x)^{k-2}E[\tau_k]}\qquad(8.15)$$

特に集団サイズが一定のとき τ_k は平均 $2/k(k-1)$ の指数分布に従うので、(8.13)より $j=1$ として、期待値を求めると

$$E[\xi_{n,b}]=\frac{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}\frac{1}{2}E[S_k^2-S_{k+1}^2]}{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}\frac{2}{k(k-1)}}=\frac{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[S_k^2-S_{k+1}^2]}{4\binom{n-1}{b}}$$

$S_k=\sum_{i=k}^{n}\tau_i$ より、$E[\exp(-\lambda S_k)]=\prod_{i=1}^{n}E[\exp(-\lambda\tau_i)]=\prod_{i=1}^{n}\frac{\lambda_i}{\lambda+\lambda_i}$, $\lambda_i=\frac{i(i-1)}{2}$。

これより $E[S_k^2-S_{k+1}^2]=\frac{d^2}{d\lambda^2}E[\exp(-\lambda S_k)-\exp(-\lambda S_{k+1})]\big|_{\lambda=0}=\frac{8(n-k+1)}{nk(k-1)^2}$。

以上を代入して次式を得る。

$$E[\xi_{n,b}]=\frac{2}{\binom{n-1}{b}}\sum_{k=2}^{n}\binom{n-k}{b-1}\frac{n-k+1}{n(k-1)} \tag{8.16}$$

さらに、$\frac{b}{n}=x$ （一定）、$n\to\infty$ の極限をとると、$\lim_{n\to\infty}\binom{n-k}{b-1}\Big/\binom{n-1}{b}=x(1-x)^{k-2}$ より

$$E[\xi_x]=\lim_{n\to\infty}E[\xi_{n,b}]=2x\sum_{k=2}^{\infty}\frac{(1-x)^{k-2}}{k-1}=-\frac{2x}{1-x}\log x \tag{8.17}$$

この結果は第４章３節でも得られている。

８．１．４　サンプル構成の条件下での共通祖先までの待ち時間分布

n 個のサンプルを取り出したとき、ある分離サイトに b 個の突然変異と $n-b$ 個の祖先型をもつという条件下でこのサンプルの共通祖先までの時間 $\eta_{n,b}$ を考える。$\eta_{n,b}$ の確率分布密度を $f_{n,b}(t)$ とする。合祖時間を $W_n=\sum_{k=2}^{n}\tau_k$、W_n の分布密度を $f_n(t)$ とする。サンプル構成が $(b,n-b)$ のときの W_n の条件付き分布密度が $f_{n,b}(t)$ である。

定理８．３

$$f_{n,b}(t)=f_n(t)\frac{\sum_{k=2}^{\infty}k(k-1)\binom{n-k}{b-1}E[\tau_k|W_n=t]}{\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k]},\quad t>0 \tag{8.18}$$

（証明）　$E_{n,b}$ を n 個のサンプルの構成が $(b,n-b)$ である事象とする。

$$f_{n,b}(t)=P\left(W_n=t\middle|E_{n,b}\right)=\frac{P(W_n=t,E_{n,b})}{P(E_{n,b})}=\frac{P(W_n=t)P\left(E_{n,b}|W_n=t\right)}{P(E_{n,b})},$$

$P(E_{n,b})=q_{n,b}$, $P(W_n=t)=f_n(t)$ なので、$P(E_{n,b}|W_n=t)$ に定理８．１の証明と同じ議論を用いて(8.18)を得る。

（証明終わり）

$\eta_{n,b}$ の期待値を求めると

$$E[\eta_{n,b}]=\int_0^\infty t\times f_{n,b}(t)dt=\frac{1}{C_{n,b}}\left\{\sum_{k=2}^{\infty}k(k-1)\binom{n-k}{b-1}\int_0^\infty tf_n(t)E[\tau_k|W_n=t]dt\right\}$$

$$=\frac{1}{C_{n,b}}\left\{\sum_{k=2}^{\infty}k(k-1)\binom{n-k}{b-1}\int_0^\infty E[W_nE[\tau_k|W_n]]dt\right\}$$

$$=\frac{1}{C_{n,b}}\left\{\sum_{k=2}^{\infty}k(k-1)\binom{n-k}{b-1}\int_0^\infty E[W_n\tau_k]dt\right\} \qquad (8.19)$$

○集団サイズが一定の場合

補題８．４

$h_{n,b}(t)=\frac{1}{C_{n,b}}\sum_{k=2}^{\infty}k(k-1)\binom{n-k}{b-1}\exp\left(-\frac{k(k-1)}{2}t\right)$, $t>0$ とすると、

$$f_{n,b}(t)=f_n(t)*h_{n,b}(t), \quad (*\text{は畳み込み}) \qquad (8.20)$$

（証明）ラプラス変換を取って $L\{f_{n,b}(t)\}=L\{f_n(t)\}L\{h_{n,b}(t)\}$ が成り立つことを示せばよい。

$$L\{f_{n,b}(t)\}=\frac{1}{C_{n,b}}\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}\int_0^\infty e^{-\lambda t}f_n(t)E[\tau_k|W_n=t]dt$$

$$=\frac{1}{C_{n,b}}\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k\exp(-\lambda W_n)]$$

$$=\frac{1}{C_{n,b}}\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E\left[\tau_k\exp(-\lambda\tau_k)\left(\prod_{\substack{i=2\\(i\neq k)}}^{n}\exp(-\lambda\tau_i)\right)\right]$$

$$=\frac{1}{C_{n,b}}\sum_{k=2}^{n}k(k-1)\binom{n-k}{b-1}E[\tau_k\exp(-\lambda\tau_k)]\prod_{\substack{i=2\\i\neq k}}^{n}E[\exp(-\lambda\tau_i]$$

$$= \frac{1}{C_{n,b}} \sum_{k=2}^{n} k(k-1) \binom{n-k}{b-1} \frac{\lambda_k}{(\lambda+\lambda_k)^2} \left(\prod_{\substack{i=2 \\ i \neq k}}^{n} \frac{\lambda_i}{\lambda+\lambda_i} \right)$$

$$= \left(\prod_{i=2}^{n} \frac{\lambda_i}{\lambda+\lambda_i} \right) \left(\frac{1}{C_{n,b}} \sum_{k=2}^{n} k(k-1) \binom{n-k}{b-1} \frac{1}{(\lambda+\lambda_k)} \right) = L\{f_n(t)\} L\{h_{n,b}(t)\}$$

（証明終わり）

$\eta_{n,b}$のラプラス変換が$E\left[\exp(-\lambda\eta_{n,b})\right] = \left(\prod_{i=2}^{n} \frac{\lambda_i}{\lambda+\lambda_i} \right) \left(\frac{1}{C_{n,b}} \sum_{k=2}^{n} k(k-1) \binom{n-k}{b-1} \frac{1}{(\lambda+\lambda_k)} \right)$

なので、

$$E[\eta_{n,b}] = -\frac{d}{d\lambda} E\left[\exp(-\lambda\eta_{n,b})\right]_{\lambda=0} = 2\left(1-\frac{1}{n}\right) + 2\binom{n-1}{b-1}^{-1} \sum_{k=2}^{n} \binom{n-k}{b-1} \frac{1}{k(k-1)}$$

$\frac{b}{n} = x$を固定して、$n \to \infty$の極限をとると、$\lim_{n\to\infty} \frac{\binom{n-k}{b-1}}{\binom{n-1}{b}} = x(1-x)^{k-2}$より、

$$E[\eta_{\infty,x}] = \lim_{n\to\infty} E[\eta_{n,b}] = 2 + \frac{2x}{1-x} \sum_{j=1}^{\infty} \frac{(1-x)^j}{j(j+1)} = 2 + \frac{2x}{1-x}\left(1 + \frac{x}{1-x}\log x\right) \qquad (8.21)$$

８．２ 分離サイト数と塩基多様度

集団からサンプルしたｎ本の遺伝子DNA配列について多型サイト（分離サイト）の数(the number of segregating sites)をSとする。またｎ本の配列の${}_nC_2$通りのすべてのペアで配列を比較し異なるサイトの数の合計を塩基多様度(nucleotide diversity)と呼びΠで表す。図８．１のような１０本の配列が与えられた場合、多型サイトは*印の６サイトあるので、$S=6$。任意交配を行っている集団サイズ一定の単一集団から中立な遺伝子をサンプルした場合、多型サイト数Sの分布は定理３．６あるいは(3.15)で示されている。また、その母関数は定理３．５において$G(0,z)$で与えられるので

$G(0,z) = E\left[z^S\right] = \prod_{j=2}^{n} \frac{(j-1)}{(j-1)+\vartheta(1-z)}$。平均$E[S]$、分散$V[S]$は

$$E[S] = \frac{d}{dz} G(0,z)\Big|_{z=1} = \vartheta \sum_{j=2}^{n} \frac{1}{j-1},$$

また$\frac{d^2}{dz^2}G(0,z)=E\left[S(S-1)z^{S-2}\right]$より

$V[S]=\frac{d^2}{dz^2}G(0,z)|_{z=1}+E[S]-(E[S])^2=\vartheta\sum_{j=2}^{n}\frac{1}{j-1}+\vartheta^2\sum_{j=2}^{n}\frac{1}{(j-1)^2}$を得る。

塩基多様度Πとは、n本の配列から取り出した2本(i,j)ペアの塩基相違度d_{ij}の総和を組み合わせ総数$\binom{n}{2}=\frac{n(n-1)}{2}$で割った量を$\Pi=\frac{\sum_{i<j}d_{ij}}{n(n-1)/2}$とする。これはn本の配列について$\ell$番目のサイトで塩基$A,T,G,C$の数を順に$n_1(\ell),n_2(\ell),n_3(\ell),n_4(\ell)$とすると、そのサイトの異なる塩基ペアの数は$\sum_{p<q}n_p(\ell)n_q(\ell)$、これをペア組み合わせ数$\binom{n}{2}$で割った量を$h_\ell=\frac{\sum_{p<q}n_p(\ell)n_q(\ell)}{n(n-1)/2}$とする。$h_\ell=\frac{2\sum_{p<q}(n_p(\ell)/n)(n_q(\ell)/n)n^2}{n(n-1)}=\frac{n}{n-1}\sum_{p<q}2\left(\frac{n_p(\ell)}{n}\right)\left(\frac{n_q(\ell)}{n}\right)$

と表せるのでh_ℓはそのサイトの母集団におけるヘテロ接合度の推定値と考えられる。

$\frac{n}{n-1}$は不偏推定のための係数である。塩基多様度は、Lを塩基サイト数とすると

$\Pi=\sum_{\ell=1}^{L}h_\ell$と書けるので、塩基多様度$\Pi$はヘテロ接合度の総和を表している。

図８．１の１０本の配列の場合、上の配列から順に1,2,...,10と番号付けると

$d_{12}=2, d_{13}=1,\cdots,d_{9,10}=3$となる。他方、分離しているサイトは左から３，１４，２４，３６，４４，５８番目の計６サイトのみなのでこれらのサイトのヘテロ接合度より

$h_3=\frac{2\times8}{45}=\frac{16}{45}, h_{14}=\frac{1\times9}{45}=\frac{9}{45}, h_{24}=\frac{3\times7}{45}=\frac{21}{45}, h_{36}=\frac{4\times6}{45}=\frac{24}{45}, h_{44}=\frac{1\times9}{45}=\frac{9}{45},$

$h_{58}=\frac{1\times9}{45}=\frac{9}{45}$なので$\Pi=\frac{16+9+21+24+9+9}{45}=\frac{88}{45}=1.955\cdots$となる。

中立な遺伝子の場合、d_{ij}は2本の配列について分離したサイトの数に等しいので

$E[d_{ij}]=\vartheta$、よって$E[\Pi]=\frac{\sum_{i<j}d_{ij}}{n(n-1)/2}=E[d_{ij}]=\vartheta$。分散はTajima(1983)により

$V[\Pi]=\frac{n+1}{3(n-1)}\vartheta+\frac{2(n^2+n+3)}{9n(n-1)}\vartheta^2$ となる。

図８．１

```
AGGCTCGAACTGCACCAGTCAGCTTGAACTAGCAGTCACATGCAGTTCGCACGTCAGGTCAGCT
AGCCTCGAACTGCACCAGTCAGCTTGAACTAGCAGCCACATGCAGTTCGCACGTCAGGTCAGCT
AGGCTCGAACTGCACCAGTCAGCATGAACTAGCAGTCACATGCAGTTCGCACGTCAGGTCAGCT
AGGCTCGAACTGCACCAGTCAGCTTGAACTAGCAGCCACATGCAGTTCGCACGTCAGGTCAGCT
AGGCTCGAACTGCACCAGTCAGCTTGAACTAGCAGTCACATGCAGTTCGCACGTCAGGTCAGCT
AGCCTCGAACTGCACCAGTCAGCTTGAACTAGCAGCCACATGCAGTTCGCACGTCAGGTCAGCT
AGGCTCGAACTGCACCAGTCAGCATGAACTAGCAGTCACATGCTGTTCGCACGTCAGGTCAGCT
AGGCTCGAACTGCACCAGTCAGCTTGAACTAGCAGTCACATGCAGTTCGCACGTCAGCTCAGCT
AGGCTCGAACTGCTCCAGTCAGCTTGAACTAGCAGCCACATGCAGTTCGCACGTCAGGTCAGCT
AGGCTCGAACTGCACCAGTCAGCATGAACTAGCAGTCACATGCAGTTCGCACGTCAGGTCAGCT
   *           *           *           *           *           *
```

８．３　Tajima's D テスト

８．３．１　Tajima's D の定義

Tajima(1989)は現在 Tajima's　D と呼ばれている統計量を導入し、中立仮説の検定統計量として、広く使われている。まず、Tajima(1989)に従って、次の記号を導入する。

$a_n=\sum_{i=1}^{n-1}\frac{1}{i}$, $a_2=\sum_{i=1}^{n-1}\frac{1}{i^2}$、$b_1=\frac{n+1}{3(n-1)}$, $b_2=\frac{2(n^2+n+3)}{9n(n-1)}$, $c_1=b_1-\frac{1}{a_1}$, $c_2=b_2-\frac{n+2}{a_1 n}+\frac{a_2}{a_1^2}$

$e_1=\frac{c_1}{a_1}$, $e_2=\frac{c_2}{a_1^2+a_2}$。

前節の結果より、$E[S]=a_1\vartheta$, $V[S]=a_1\vartheta+a_2\vartheta^2$, $E[\Pi]=\vartheta$, $V[\Pi]=b_1\vartheta+b_2\vartheta^2$ と表示できる。これより、$E[\Pi]=E\left[\frac{S}{a_1}\right]=\vartheta$、すなわち $E\left[\Pi-\frac{S}{a_1}\right]=0$ となり、中立な遺伝子の場合 $d=\Pi-\frac{S}{a_1}$ の値は 0 の近傍の値を取ることが期待される。その分散を計算すると

$$V[d]=V\left[\Pi-\frac{S}{a_1}\right]=V[\Pi]-\frac{2}{a_1}Cov(\Pi,S)+\frac{1}{a_1^2}V[S]。$$

$V[\Pi]=b_1\vartheta+b_2\vartheta^2$, $V[S]=a_1\vartheta+a_2\vartheta^2$ であり、$Cov(\Pi,S)$ は Tajima(1989)によると $Cov(\Pi,S)=\vartheta+\left(\frac{1}{2}+\frac{1}{n}\right)\vartheta^2$ となる。これより $V(d)=c_1\vartheta+c_2\vartheta^2$ が得られる。

$E[S]=\vartheta$, $E[S^2]-E[S]=(a_1^2+a_2)\vartheta^2$ なので、ϑ 及び ϑ^2 の不偏推定量として S および

$\frac{S(S-1)}{a_1^2+a_2}$を用いて、Tajima(1989)は分散$V[d]$の推定量として$\hat{V}[d]=e_1S+e_2S(S-1)$を用いて、$d=\Pi-\frac{S}{a_1}$を標準化した次の統計量Dを提案した。

$$D=\frac{d}{\sqrt{\hat{V}[d]}}=\frac{\Pi-S/a_1}{\sqrt{e_1S+e_2S(S-1)}} \tag{8.22}$$

Tajima(1989)はDが取りえる値は$D_{\min}=\frac{(2/n)-1/a_1}{\sqrt{e_2}}$, $D_{\max}=\frac{1/2+1/2(n-1)-1/a_1}{\sqrt{e_2}}$の範囲であり、塩基の遺伝的変異が中立な場合$D$の分布が区間$[D_{\min}, D_{\max}]$のベータ分布に近いことをコンピューター・シミュレーションで示した。

Griffiths and Tavare(1998)の結果との関連で次の様なTajima's Dの表現式も紹介しておこう。ω_iをn個のサンプル中にi個存在する突然変異の数とする。明らかに

$S=\sum_{i=1}^{n-1}\omega_i$, $\Pi=\frac{2}{n(n-1)}\sum_{i=1}^{n-1}i(n-i)\omega_i$。これより、

Tajima's Dの分子$=\sum_{i=1}^{n-1}\left(\frac{2i(n-i)}{n(n-1)}-\frac{1}{a_n}\right)\omega_i$と表現できる。一つの塩基サイトが$b$個の突然変異と$n-b$個の祖先型に分離している確率$q_{n,b}$を定理８．１で求めたが、期待値$E[\omega_b]$は$q_{n,b}$を用いて表すことができる。系図上で生じる突然変異数の期待値は$\frac{\theta}{2}E[L]$ (Lは全枝の長さ）なので、$E[\omega_b]=\frac{\theta}{2}E[L]\times q_{n,b}=\frac{\theta}{2}\left(\sum_{k=2}^{n}kE[\tau_k]\right)\times q_{n,b}$。定理８．１を用いて

$$E[\omega_b]=\frac{\theta}{b}\binom{n-1}{b}^{-1}\sum_{k=2}^{n-b-1}\binom{k}{2}\binom{n-k}{b-1}E[\tau_k] \tag{8.23}$$

集団サイズが一定のときは、(8.6)と同様にして$E[\omega_b]=\frac{\theta}{b}$。 (8.24)

８．３．２　Tajima's D テストの解釈

中立仮説の下では、D の値は 0 近傍の値を取ることが期待される。より正確に述べると以下の条件を満たす必要がある。

（ｉ）単一の任意交配集団で個体数が一定

（ii）他の集団との間で個体の移住がない。

（iii）自然選択がない。

よって、D の値が有意に正又は負の大きな値を取る時は、上記の条件のいずれかが満たされていない可能性が生まれる。D の値が有意に 0 から外れる原因として考えられるのか要因を紹介しよう。

（1） Tajima's D が正となる要因

$D>0$ となるのは $\Pi > \frac{S}{a_1}$、すなわちヘテロ接合度が高いサイトが多い場合である。個体群動態及び自然選択の両面から、このような塩基多様性が生じるような状況を考える。

① 個体群動態の要因

・集団サイズの急激な減少がある場合低い頻度の対立遺伝子は消失し、対立遺伝子数すなわちSは減少するが、頻度の高い対立遺伝子はそれほど減らないのでΠは維持される。

・遺伝子頻度構成の大きく異なる複数の集団が多数個体づつ混合した場合、ヘテロ接合度の高い塩基頻度組成サイトが即座に出現し、$D>0$となる。

② 自然選択の要因

ヘテロ接合度を高める自然選択としては平衡選択、すなわち超優性選択、頻度が低い対立遺伝子が有利になる負の頻度依存性選択、ニッチによって多様な選択が働くニッチ多様性選択などがある。

（2） Tajima's D が負となる要因

$D<0$ となるのは $\Pi < \frac{S}{a_1}$、すなわちヘテロ接合度が低いサイトが多い場合である。

① 個体群動態の要因

・集団サイズの急激な増大。集団が大きくなるにつれ新たに突然変異により多型サイトが生じるが最初は1個(singleton)の状態から始まるので、どれも頻度が低くΠの値は小さくSの値は比較的大きい。

・ボトルネック等によって急激に集団サイズが減少し、遺伝的多様性がほとんど 0 の状態から出発して間もないころ、突然変異で生じた対立遺伝子はほとんど低頻度なので、Sと比較してΠの値は小さい。

② 自然選択の要因

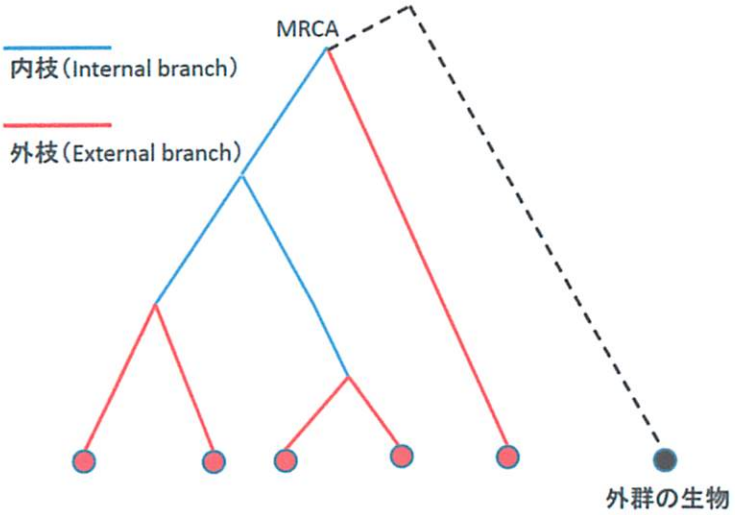
・ある遺伝子に非常に有利な突然変異が生じ急速に集団内で頻度を増したとき、その突然変異の周辺領域も組み換えが起こるより速く頻度を増す。そのため突然変異近傍の広い領域に亘って遺伝的多様性が低い領域が生じる。これを自然選択的一掃(Selective sweep)という。その領域では①の集団サイズの急激な増大と類似な状況になる。また、ある遺伝子領域のどこに突然変異が生じても弱有害な効果を及ぼす場合、すなわち浄化淘汰(purifying selection)がその遺伝子領域全体に働いている場合、生じ

た突然変異は頻度を増やすことができず、低頻度なアレルの割合が増える。S と比較してヘテロ接合度Π は小さい。

Dの値が大きく０から外れたとき、個体群動態による要因と自然選択の要因の二つの原因が考えられるが、個体群動態による場合はゲノム全体にその影響が現れるのに対して、自然選択は自然選択が働いている遺伝子の近傍の領域に影響が限定される。従って多くの遺伝子でテストすることによって二つの要因を区別することができる。

８．４　その他の検定法

（１）Fu and Li's D



Fu and Li(1993)は系図の内枝と外枝に生じた突然変異の数の分布に注目して、検定統計量を提案した。内枝とは系図のその枝の先には１個のサンプル配列しか無い枝、外枝とはその先に２個以上のサンプルが付いている枝をいう。外枝に生じた突然変異はサンプル中に１個(Singleton)のみ出現する。しかし、singleton 変異が外枝に生じた変異か否かの判定は祖先配列、すなわち MRCA の配列が必要となる。

外群のサンプル配列があるとき共通祖先 MRCA の配列を推定できる。従って、Singleton の変異がある時、それが祖先型か突然変異か区別できる。それが MRCA の配列の塩基と異なる場合は外枝に生じた突然変異と推定できる。そこで外枝に生じた突然変異の数をη_e、η_iを内枝に生じた突然変異の数とする。また$\eta=\eta_e+\eta_i=S$を総突然変異数とするとき、Fu and Li(1993)は次の二つの統計量を提案した。

$$D_{F.L.}=\frac{\eta-a_n\eta_e}{\sqrt{u_D\eta+v_D\eta^2}},\ \text{ただし}v_D=1+\frac{a_n^2}{b_n+a_n^2}\left(c_n-\frac{n+1}{n-1}\right),\quad u_D=a_n-1-v_D$$

$$F_{F.L.}=\frac{\Pi-\eta_e}{\sqrt{u_F\eta+v_F\eta^2}},\ \text{ただし}v_F=\left[c_n+\frac{2(n^2+n+3)}{9n(n-1)}-\frac{2}{n-1}\right]\Big/\left(a_n^2+b_n\right)$$

$$u_F=\left[1+\frac{n+1}{3(n-1)}-4\frac{n+1}{(n-1)^2}\left(a_{n+1}-\frac{2n}{n+1}\right)\right]\Big/a_n-v_F$$

外群の配列が未知のときは Singleton の変異があるときそれが突然変異なのか祖先型なのか不明なので、$\eta_e=\frac{n-1}{n}\eta_s$ (η_sは singleton 変異の数)でη_eを推定する。これより

$$D^*_{F.L.}=\frac{\left(\frac{n}{n-1}\right)\eta-a_n\eta_s}{\sqrt{u_{D^*}\eta+v_{D^*}\eta^2}},\ \text{ただし}v_{D^*}=\left[\frac{b_n}{a_n^2}-\frac{2}{n}\left(1+\frac{1}{a_n}-a_n+\frac{a_n}{n}\right)-\frac{1}{n^2}\right]\Big/\left(a_n^2+b_n\right)$$

$$u_{D^*}=\left[\left(\frac{n-1}{n}-\frac{1}{a_n}\right)\Big/a_n\right]-v_{D^*}$$

$$F^*_{F.L.}=\frac{\Pi-\frac{n-1}{n}\eta_s}{\sqrt{u_{F^*}\eta+v_{F^*}\eta^2}},$$

$$\text{ただし}v_{F^*}=\left[\frac{2n^3+110n^2-255n+153}{9n^2(n-1)}+\frac{2(n-1)a_n}{n^2}-\frac{8b_n}{n}\right]\Big/\left(a_n^2+b_n\right)$$

$$u_{F^*}=\left\{4n^2+19n+3-12(n+1)a_{n+1}/3n(n-1)a_n\right\}-v_{F^*}$$

Fu and Li の検定量については Tajima's D と同様な性質があるが、Simonsen et al(1995)はこれらの統計量の感度、検出力の比較を行っている。

（2） Fay andWu's H

Fay and Wu(2000)は次のような統計量 H を提案した。

$\hat{\theta}_\pi=\frac{2}{n(n-1)}\sum_{k=1}^{n}\omega_k k(n-k)$,　$\hat{\theta}_H=\frac{2}{n(n-1)}\sum_{k=1}^{n-1}\omega_k k^2$ とするとき、$H=\hat{\theta}_\pi-\hat{\theta}_H$。

集団サイズ一定の中立な遺伝子の場合、明らかに$E[\hat{\theta}_\pi]=E[\hat{\theta}_H]=0$なので、$E[H]=0$となる。あるサイトでの突然変異が有利で集団中に急速に広がるとき、それに連鎖したゲノム領域も引きずられて集団中に広がる(selective sweep)。するとその領域全体のホモ接合度が他の領域に比べて大きくなる。その結果 H は大きく負の値を取る。Fay and Wu の H は Selective sweep の検出に良く用いられる。