SOBRE LA DERIVA GENÈTICA

Escrit per:
Martín Ríos
Dept. d'Estadística
Universitat de Barcelona

Tractarem de descriure aquí, mitjançant un senzill model matemàtic, el fenomen conegut en genètica de poblacions com a *deriva* (o també, a vegades, com a *cosanguinitat*).

Considerem una població d'organismes diploides la mida poblacional dels quals, per simplificar, suposarem constant (N) de generació en generació. També, per simplicitat, tractarem el temps com una variable discreta, quantificat en termes de generacions. Suposarem aparellament a l'atzar (*panmíxia*) i no considerarem ni migració, ni mutació ni selecció.

Anem a estudiar, per a un locus autosòmic determinat amb dos al·lels possibles, H i h, com canvien les freqüències gèniques i genotípiques de generació en generació, especialment la freqüència d'heterozigots Hh.

Sigui $W_j = (X_j, Y_j, Z_j)$ un vector aleatori que representa el número d'individus amb genotips HH, Hh i hh, respectivament, en la generació j. Observem que $X_j + Y_j + Z_j = N$ i que el total d'al·lels per a aquest locus és $2N$.

Com a conseqüència de la hipòtesi de *panmíxia*, i en absència de mutació, selecció o migració, si en la j-èsima generació $X_j = x_j$, $Y_j = y_j$ i $Z_j = z_j$, llavors en formar un zigot de la generació $j+1$, tindrem les probabilitats següents:

$$\begin{aligned} p &= P(HH) = \left(\tfrac{a_j}{2N}\right)^2, \\ q &= P(Hh) = 2\left(\tfrac{a_j}{2N}\right)\left(1-\tfrac{a_j}{2N}\right), \\ r &= P(hh) = \left(1-\tfrac{a_j}{2N}\right)^2 \end{aligned} \qquad (1)$$

Sent $a_j = 2x_j + y_j$, la freqüència absoluta de l'al·lel H en la generació j.

La distribució de $W_j+1 = (X_j+1, Y_j+1, Z_j+1)$ condicionada a $W_j = (X_j, Y_j, Z_j)$ serà una trinomial, i.e.:

$$P[X_{j+1} = x_{j+1}, Y_{j+1} = y_{j+1}, Z_{j+1} = z_{j+1} | X_j = x_j, Y_j = y_j, Z_j = z_j] =$$

$$= \frac{N!}{x_{j+1}!\, y_{j+1}!\, z_{j+1}!}\; p^{x_{j+1}}\, q^{y_{j+1}}\, r^{z_{j+1}}$$

La funció característica de $W_j+1 = (X_j+1, Y_j+1, Z_j+1)$, que ens servirà per trobar l'esperança de Y_j+1, la podrem calcular mitjançant:

$$\begin{aligned}\Phi_{W_{j+1}}(t_1,t_2,t_3) &= E\left(e^{i(t_1X_{j+1}+t_2Y_{j+1}+t_3Z_{j+1})}\right)\\ &= E\Big(E\left(e^{i(t_1X_{j+1}+t_2Y_{j+1}+t_3Z_{j+1})}\big|W_{j+1}\right)\Big)\\ &= \sum_{x_j+y_j+z_j=N} P(X_j=x_j,Y_j=y_j,Z_j=z_j)\;\Phi_{W_{j+1}|W_j}(t_1,t_2,t_3)\\ &= \sum_{x_j+y_j+z_j=N} P(X_j=x_j,Y_j=y_j,Z_j=z_j)\;(pe^{it_1}+qe^{it_2}+re^{it_3})^N\end{aligned}$$

Per tant,

$$\frac{\partial\Phi}{\partial t_2} = i\sum_{x_j+y_j+z_j=N} P(X_j=x_j,Y_j=y_j,Z_j=z_j)N(pe^{it_1}+qe^{it_2}+re^{it_3})^{N-1}\;q\;e^{it_2}$$

i com que $iE(Y_{j+1}) = \frac{\partial\Phi}{\partial t_2}(0,0,0)$ i $p+q+r=1$, tenint en compte (1), resulta

$$\begin{aligned}E(Y_{j+1}) &= \textstyle\sum_{x_j+y_j+z_j=N} P(X_j=x_j,Y_j=y_j,Z_j=z_j)\;a_j\left(1-\frac{a_j}{2N}\right)\\ &= E(A_j)-\tfrac{1}{2N}E(A_j^2)\end{aligned}$$

on A_j és la variable aleatòria número d'al·lels H presents en la generació j.
Alternativament podem escriure també:

$$E(Y_{j+1}) = E(A_j) - \frac{1}{2N}(\mathrm{var}(A_j)+E^2(A_j)). \qquad (2)$$

Observem doncs que $E(Y_j)$ depèn de l'esperança i la variància d'A_j, per la qual cosa procedirem a calcular-les. La variable aleatòria A_j, el número d'al·lels H en la generació j, pot prendre valors 0, 1, ..., $2N$. Suposem que en la generació inicial el número d'al·lels H que hi ha en la població és c, conegut, $A_0 = c$. Anem a calcular els moments de la variable aleatòria A_j a partir de la funció característica, tal i com hem fet amb anterioritat.

$$\begin{aligned}\Phi_{A_{j+1}}(t) &= E\left(e^{itA_{j+1}}\right) = E\left(E\left(e^{itA_{j+1}}\mid A_j\right)\right)\\ &= \sum_{k=0}^{2N} P(A_j=k)\;\Phi_{A_{j+1}|A_j=k}(t)\end{aligned}$$

on $\Phi_{A_{j+1}|A_j=k}(t)$ és la funció característica d'A_j+1 condicionada a $A_j=k$.

La distribució de la variable aleatòria A_j+1 condicionada a $A_j=k$, com a conseqüència de la hipòtesi de *panmíxia* i en absència de mutació, selecció o migració, serà una Binomial, $B(2N, \frac{k}{2N})$, per tant

$$\Phi_{A_{j+1}}(t) = \sum_{k=0}^{2N} P(A_j=k)(\frac{k}{2N}e^{it}+1-\frac{k}{2N})^{2N}$$

$$\Phi'_{A_{j+1}}(t) = i\sum_{k=0}^{2N} P(A_j=k)2N(\frac{k}{2N}e^{it}+1-\frac{k}{2N})^{2N-1}\frac{k}{2N}\;e^{it}$$

i com que $iE(A_{j+1}) = \Phi'_{A_{j+1}}(0)$, resulta:

$$E(A_{j+1}) = \sum_{k=0}^{2N} P(A_j=k)\;k = E(A_j)$$

la qual cosa implica, per recurrència, que

$$E(A_{j+1}) = E(A_j) = \ldots = E(A_0) = c.$$

És a dir, que el *valor mitjà* de les freqüències dels al·lels H roman constant durant totes les generacions.
Per obtenir la variància d'A_{j+1}, calcularem primer $E(A_{j+1}^2)$ a través de la derivada segona de la funció característica, ja que $-E(A_{j+1}^2) = \Phi''_{A_{j+1}}(0)$.

$$\begin{aligned}\Phi''_{A_{j+1}}(t) &= -\sum_{k=0}^{2N} P(A_j = k)\left(2N(2N-1)(\frac{k}{2N}e^{it} + 1\frac{k}{2N})^{2N-2}\frac{k^2}{4N^2}\ e^{2it}\right.\\ &+ \left. 2N(\tfrac{k}{2N}e^{it} + 1 - \tfrac{k}{2N})^{2N-1}\tfrac{k}{2N}\ e^{it}\right)\end{aligned}$$

per tant

$$\begin{aligned}E(A_{j+1}^2) &= \sum_{k=0}^{2N} P(A_j = k)\left(\frac{2N(2N-1)}{4N^2}\ k^2 + k\right)\\ &= \left(1 - \tfrac{1}{2N}\right)E(A_j^2) - E(A_j)\\ &= \left(1 - \tfrac{1}{2N}\right)\left(\mathrm{var}(A_j) + E(A_j)^2\right) - E(A_j)\\ &= \left(1 - \tfrac{1}{2N}\right)\left(\mathrm{var}(A_j) + c^2\right) - c.\end{aligned}$$

i

$$\mathrm{var}(A_{j+1}) = \left(1 - \frac{1}{2N}\right)(\mathrm{var}(A_j) + c^2) + c - c^2 = \left(1 - \frac{1}{2N}\right)\mathrm{var}(A_j) + c\left(1 - \frac{c}{2N}\right)$$

Reiterant el procés, en ser A_0 una constant, $\mathrm{var}(A_0) = 0$, tindrem

$$\begin{aligned}\mathrm{var}(A_1) &= c\left(1 - \tfrac{c}{2N}\right)\\ \mathrm{var}(A_2) &= c\left(1 - \tfrac{c}{2N}\right)\left(1 + \left(1 - \tfrac{1}{2N}\right)\right)\\ \mathrm{var}(A_3) &= c\left(1 - \tfrac{c}{2N}\right)\left(1 + \left(1 - \tfrac{1}{2N}\right) + \left(1 - \tfrac{1}{2N}\right)^2\right)\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \mathrm{var}(A_n) &= c\left(1 - \tfrac{c}{2N}\right)\left(1 + (1 - \tfrac{1}{2N}) + (1 - \tfrac{1}{2N})^2 + \ldots + (1 - \tfrac{1}{2N})^{n-1}\right)\end{aligned}$$

i tenint en compte la fórmula de la suma dels *n* primers termes d'una progressió geomètrica, finalment tindrem

$$\mathrm{var}(A_n) = 2Nc\left(1 - \frac{c}{2N}\right)\left(1 - \left(1 - \frac{1}{2N}\right)^n\right)$$

Podem observar, a partir de l'expressió anterior, que fixada la mida poblacional N, la variància de la variable aleatòria, número d'al·lels H, augmenta de generació en generació. Quan el número de generacions tendeix a infinit, la variància serà:

$$\lim_{n\longrightarrow\infty} \mathrm{var}(A_n) = 2Nb\left(1 - \frac{b}{2N}\right) = 2N\mathrm{var}(A_1) \qquad (3)$$

Tornant ara a l'expressió (2), finalment tindrem

$$E(Y_{j+1}) = c\left(1 - \frac{c}{2N}\right)\left(1 - \frac{1}{2N}\right)^{j+1}$$

Si suposem que la població incialment ja estava en equilibri Hardy-Weinberg, les freqüències relatives dels diferents genotips a l'inici són:

$$f_r(HH) = \left(\frac{c}{2N}\right)^2, \quad f_r(Hh) = 2\left(\frac{c}{2N}\right)\left(1 - \frac{c}{2N}\right), \quad f_r(hh) = \left(1 - \frac{c}{2N}\right)^2$$

Per tant

$$Y_0 = 2N\left(\frac{c}{2N}\right)\left(1 - \frac{c}{2N}\right)$$

i resulta finalment a:

$$E(Y_n) = Y_0\left(1 - \frac{1}{2N}\right)^n$$

La qual cosa significa que l'esperança de la freqüència del genotip *Hb* en la generació *n*, depén de la quantitat inicial d'heterozigots i va disminuint a mesura que passen les generacions. Si calculem el límit quan n $\to \infty$ tindrem:

$$\lim_{n\longrightarrow\infty} E(Y_n) = \lim_{n\longrightarrow\infty} Y_0\left(1 - \frac{1}{2N}\right)^n = 0$$

Cosa que demostra que, en les condicions estudiades, els heterozigots s'extingueixen.

A continuació exposem unes gràfiques on es posen de manifest els resultats teòrics obtinguts anteriorment.

A la **Figura 1** realitzem simulacions amb 1.000 individus i amb 250 al·lels *H* al començament. Observem que a les generacions, 216, 832, 1748 l'al·lel H ha desaparegut i que a la generació 1721, amb les mateixes condicions inicials, tots els al·lels de la població són *H*.

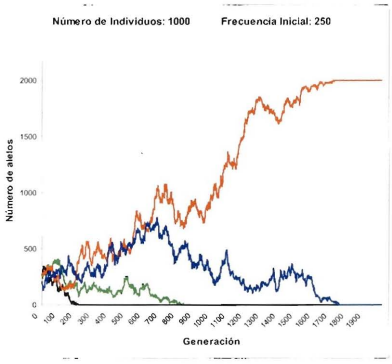

Figura 1:

A la **Figura 2**, podem veure que a les generacions 299, 678, 1798 tots els al·lels són *H* i que amb les mateixes condicions experimentals, els al·lels *H* han desaparegut a la generació 1367. En general, si el número d'al·lels inicial *b*, està més proper a zero que a 2*N*, és més fàcil que tots els al·lels acabin sent *H*. A més, generalment, quan més a prop és el número d'al·lels *H* de l'estat absorbent, 0 ó 2*N*, o més petita sigui la mida de la població, més ràpidament s'arriba a aquesta situació.

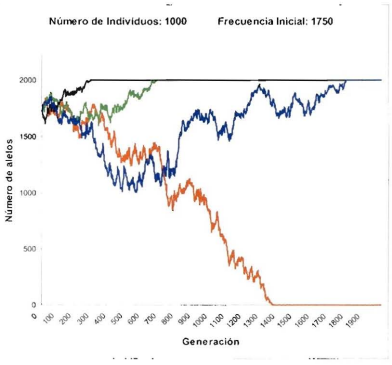

Figura 2:

Podem veure a la **Figura 3** que amb 250 individus i amb el 50% d'al·lels *H* al començament, tots acaben sent *H* a la generació 578 i que amb 1.000 individus i amb el 50% d'al·lels *H* al començament, acaben sent tots *H* a la generació 1620.

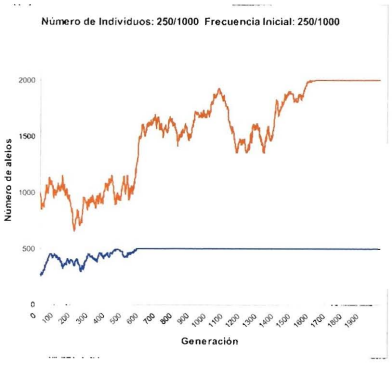

Figura 3:

Martín Ríos és llicenciat en Matemàtiques per la Universitat de Saragossa i llicenciat i doctor en Medicina per la Universitat de Barcelona. Forma part del Departament d'Estadística de la Facultat de Biologia de la Universitat de Barcelona on treballa en el camp de la geometria i l'estadística, l'anàlisi de dades, la regressió, la predicció i l'estadística mèdica. Pertany al Grup d'Anàlisi Estadística Multivariant i Computacional.