Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Estadística

Cadenas de Markov en genética poblacional

Autor:

Esther García Lerma

Tutora:

Lourdes Barba Escribá

Agradecimientos

A mi familia por apoyarme siempre,

A mis compañeros por convertirse en amigos,

A mi pareja por empujarme hacia delante

Y a mi tutora por todo lo aprendido

Gracias.

Índice de contenido

Resumen

El estudio de la genética lleva evolucionando desde principios del siglo XX hasta el punto de haberse convertido en una rama de la biología de gran importancia. En la genética poblacional se pretende estudiar la presencia de los alelos de un locus a lo largo del tiempo, habiendo llegado a desarrollarse los modelos de Fisher-Wright y Moran, que plantean eventos de nacimiento y muerte de maneras distintas, permitiendo así realizar predicciones y estudiar la población a lo largo del tiempo.

Mediante el uso de las cadenas de Markov vamos a estudiar ambos modelos, su comportamiento a largo plazo y lo que representa la manera en la que ocurren los acontecimientos de nacimiento y muerte. Además vamos a realizar comparaciones entre los modelos, entendiendo sus diferencias y la necesidad de ambos para un adecuado estudio de las poblaciones.

Palabras clave: genética poblacional, deriva genética, Fisher-Wright, Moran, cadenas de Markov.

Abstract

Genetic's study has been developing since the start of 20th century until be converted in a very important Biologist's branch. Population genetics pretend to study the allele's presence of a particular locus along the time. It has depeloved two differents models, Fisher-Wright and Moran, which explain differents birth and death events, allowing us to make predictions and to study the population in the time.

Using markov chains we are going to study both models, their behaviour along the time and what meaning have the birth and death events. Also we are going to make comparisions between the models, trying to understand the differencies and the need of them for an appropiate population's study.

Key words: population genetics, genetic drift, Fisher-Wright, Moran, Markov chain.

Capítulo 1: Introducción

La genética viene definida como el campo de la biología que estudia los genes y los mecanismos que regulan la transmisión de caracteres hereditarios. A mediados del siglo XIX, en 1865, Mendel publicó *Experimentos de hibridación en plantas*, donde muestra su teoría sobre la herencia de caracteres, sin embargo no fue hasta principios de siglo XX cuando el redescubrimiento de esta teoría permitió la unión de las ideas de genes y alelos, que Mendel tenía, con la teoría de la evolución que ya el siglo anterior habían desarrollado Charles Darwin y Alfred Russel Wallace.

Fue a partir de este momento en el que el estudio de la genética empezó a tener más auge. Durante la primera mitad del siglo XX se observa un aumento en los descubrimientos en este ámbito que continua hasta nuestros días. Durante los primeros años del siglo se da la síntesis de los trabajos genéticos y citológicos, lo que da lugar a que la genética llegue a considerarse una ciencia propia e independiente. En 1902 T. Boveri, al mismo tiempo que W. Sutton, se percata del paralelismo que encontramos entre los principios mendelianos y la conducta de los cromosomas durante la meiosis. En 1910 se descubre la herencia ligada al cromosoma X, siendo en 1916 cuando se demuestra la teoría cromosómica de la herencia mediante la no disyunción del cromosoma X.

Fue a partir de 1940 cuando se estableció el ADN como la substancia genética, a partir del cual se consigue descubrir el ARN y las proteínas que se sintetizan a partir de él. Siendo en 1953 cuando James Watson y Francis Crick concluyen que la estructura del ADN es una doble hélice, cuyas cadenas son antiparalelas. Este acontecimiento es considerado el más revolucionario y fundamental tanto en genética como en biología. Los avances desde entonces han pasado por el descubrimiento de los distintos tipos de ARN hasta llegar, a principios del siglo XXI, al proyecto Genoma humano, en el que distintos países se propusieron obtener la secuencia completa del genoma humano.

Dentro de toda la genética, al igual que ocurre con el resto de las ciencias, podemos encontrar diversos campos, nosotros vamos a centrarnos en la genética poblacional. Este campo se encargar del estudio de la variación genética en las poblaciones y cómo cambia con el tiempo, así como de las fuerzas que alteran la composición genética de una especie. La idea es investigar patrones de variación genética o estructuras genética dentro y entre grupos de individuos que se cruzan entre si. Al ser la estructura genética la base en los cambios de las poblaciones, esta área ha visto incrementada su importancia a lo largo del tiempo.

La genética de poblaciones se ve afectada por diversos mecanismos como son las mutaciones, la selección natural, las migraciones y la deriva genética. Estos mecanismos tienen una importante componente aleatoria, de manera que la lógica falla a la hora de realizar predicciones, por lo que el estudio de la genética poblacional va obligatoriamente unido al estudio de los procesos estocásticos.

En nuestro caso vamos a utilizar modelos estadísticos como el modelo de Fisher-Wright y el modelo de Moran,que hacen uso de las cadenas de Markov en tiempo discreto para explicar y comprender el comportamiento, a lo largo de las generaciones, de las frecuencias alélicas y de las frecuencias genotípicas en poblaciones pequeñas. Vamos a centrarnos en procesos en los que intervienen la deriva genética y las mutaciones, ya que el primero produce la perdida de alelos en la población, y por lo tanto, la perdida del genotipo heterocigoto y de uno de los genotipos homocigotos, mientras que con las mutaciones seremos capaces de recuperar alelos que se suponían perdidos.

En este documento vamos a explicar lo que son las cadenas de Markov así como sus propiedades mas importantes, que nos permitirán modelizar de diferentes formas nuestra variable de interés, la frecuencia alélica para un locus dado. Además vamos a explicar de manera superficial en qué consiste la genética poblacional y, más en detalle, en qué consiste la deriva genética. Pudiendo así comprender al final los diferentes ejemplos que vamos a proponer, en los cuales, mediante la simulación queremos acercarnos a situaciones reales.

Capítulo 2: Cadenas de Markov

2.1 Introducción a las cadenas de Markov

Un proceso estocástico es una colección de variables aleatorias $\{X_t, \ t \in T \ \}$ definidas sobre un espacio de probabilidad. De manera general el espacio paramétrico T suele ir referido al tiempo, ya que los procesos estocásticos surgieron del estudio de la evolución temporal de fenómenos aleatorios.

Estos procesos pueden clasificarse de cuatro maneras distintas, dependiendo tanto de la variable aleatoria como del parámetro, fijándonos en si estos son continuos o discretos. Los tipos de procesos estocásticos que nos encontramos serán con tiempo y variable discretas, con tiempo discreto y variable continua, con tiempo continuo y variable discreta y con tiempo y variables continuas. Un proceso estocástico puede verse como una función en t y ω.

Nosotros vamos a fijarnos en un proceso estocástico concreto, los procesos de Markov. Los procesos de Markov se han ido utilizando en diversos ámbitos, desde la biología, hasta la economía con la teoría de colas. Estos procesos se caracterizan por cumplir la propiedad de Markov.

Propiedad de Markov: la evolución del proceso depende solamente del pasado inmediato.

$$P\left(X_{tn} \in B | X_{t1}, X_t, \ldots X_{tn-1}\right) = P\left(X_{tn} \in B | X_{tn-1}\right)$$

Lo que estamos diciendo con esta propiedad es que no nos importa cómo hemos llegado hasta el instante de tiempo actual, únicamente nos interesa cuál era la situación en el instante anterior. Esto nos va a permitir predecir el futuro necesitando únicamente la información del presente, no vamos a necesitar la información pasada.

Cuando trabajamos con casos en los que tanto la variable aleatoria como el tiempo toma valores discretos, nos encontramos con que el proceso de Markov recibe el nombre de cadena de Markov.

Una vez que hemos visto lo que son las cadenas de Markov, vamos a pasar a estudiar sus propiedades. Consideramos X_n un proceso de Markov, que por lo tanto cumple la propiedad de Markov, el soporte de X_n, que denotaremos como S_x, se denomina conjunto de estados, que puede ser finito o no, siendo un estado cada uno de los valores que puede tomar la variable en este proceso.

2.2 Probabilidades de transición

Una de las propiedades de las cadenas de Markov, que además nos va a permitir hacer predicciones, es la probabilidad de transición.

La probabilidad de transición (p_{ij}) va a ser denominada como la probabilidad que se tiene de pasar de un estado i a un estado j.

$$p(i,j)=P(X_{n+1}=j|X_n=i, X_{n-1}=i_{n-1},\dots,X_0=i_0)\forall i,j\in S$$

Vamos a tener probabilidades de transición para cada combinación i x j, en casos en los que el conjunto de estados sea finito, podremos representarlas de dos maneras, la primera en una *matriz de transición*, que nos permite ver de manera directa la probabilidad de un cambio concreto.

$$P=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix}$$

Como ya hemos dicho, P es una matriz de probabilidades, por lo que cada $0 \leq p_{ij} \leq 1$ para todo i, j que pertenezca a S. Además, cada una de las filas de la matriz nos expresa la probabilidad de ir desde ese punto hasta alguno de los demás, por lo que para cada fila debe darse que:

$$\sum_j p_{ij}=1$$

Al cumplir las propiedades ya dichas, podemos decir que la matriz de transición es una matriz estocástica. Si además se cumpliera que la suma en columnas es también igual a 1, la matriz se denominará doblemente estocástica.

La segunda es con una *representación gráfica de la cadena*, que nos permite ver más rápidamente las conexiones que hay entre los distintos estados tal, en esta representación vamos a reflejar tanto las uniones que hay entre los estados como las probabilidades de transición que se dan entre los estados, un ejemplo básico lo vemos en la figura:

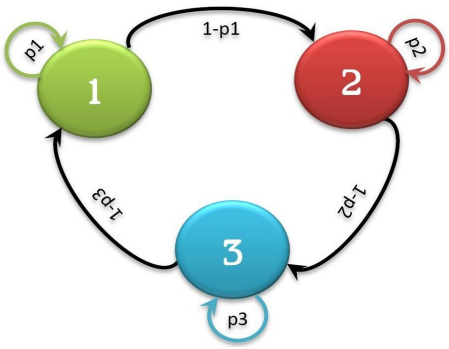

Figura 1: ejemplo de cadena de Markov.

Las probabilidades vistas hasta ahora solamente nos permiten predecir el siguiente paso al que nos encontramos, lo cual aunque útil, nos limita a la hora de trabajar. Por esto, contamos con las probabilidades de transición en varios pasos (p^m_{ij}). Esto quiere decir que vamos a calcular la probabilidad de pasar del estado i al estado j en m pasos, siendo m > 1.

$$p_{ij}^m=P(X_{n+m}=j|X_n=i)$$

Con esta notación estamos mostrando que la manera de obtener las probabilidades de transición en m pasos va a conseguirse mediante potencias m-ésimas de la matriz de transición P. De manera que

$$p_{ij}^{m} = (P^{m})_{ij}$$

Es decir, la probabilidad de transición de i a j en m pasos corresponde con la posición ij de la m-ésima potencia de la matriz de transición. De manera lógica podemos plantear otra forma de obtener estas probabilidades, vamos a verlo con un ejemplo.

Tenemos tres estados distintos A, B y C. En el momento actual nos encontramos en el estado A y queremos conocer la probabilidad de estar en el estado B tras dos unidades de tiempo, para ello voy considerar todas las posibilidades en cada paso. Imaginemos que en el primer paso me mantengo en A, entonces la probabilidad sería la de mantenerse en A en el primer paso y la de pasar a B en el segundo paso. La manera de tener en cuenta todas las posibilidades va a ser la siguiente:

$$p_{AB}^{2} = p_{AA} * p_{AB} + p_{AB} * p_{BB} + p_{AC} * p_{CB} = (P^{2})_{AB}$$

Expresándolo de manera general para cualquier proceso de Markov y para cualquier número de pasos, contamos con las ecuaciones de Chapman-Kolmogorov

$$p_{ij}^{m+n} = \sum_{k} p_{ik}^{m} * p_{kj}^{n}$$

En esta ecuación, k va a representar los estados intermedios posibles que encontramos entre i y j. Si tenemos la matriz de transición en el momento m, obtener la matriz de transición hasta el siguiente momento va conseguirse multiplicando la ya calculada por P, lo cual queda demostrado con estas ecuaciones

$$p_{ij}^{m+1} = \sum_{k} p_{ik}^{m} * p_{kj}$$

Con lo que dejamos demostrado que la obtención de la matriz de transición m pasos hacia adelante se obtiene con la potencia de la misma.

Es a partir de la idea de las probabilidades de transición en varios pasos que somos capaces de plantear cómo van a actuar las cadenas en el futuro, llegando incluso a buscar un comportamiento límite, un comportamiento que se mantenga pasado una cantidad de tiempo importante. Este comportamiento límite podemos estudiarlo gracias a la matriz límite, P^{n} para $n \to \infty$.

2.3 Clasificación de los estados

A la hora de trabajar con los distintos estados que podemos encontrarnos en una cadena de Markov o en cualquier proceso de Markov, tenemos que tener en cuenta el tipo de estado que es, ya que esto nos aportará información importante para entender mejor el comportamiento de la cadena a lo largo del tiempo.

Para poder clasificar los estados, primero tenemos que definir un concepto. El tiempo de retorno al estado y T_y. Este tiempo se define como el tiempo que la cadena tarda en pasar por el estado y sin tener en cuenta si ha empezado o no en dicho estado.

$$T_y = min\{n \geq 1 : X_n = y\}$$

Tenemos dos posibilidades, que no volvamos a pasar por el estado y, lo que significa que $T_y = \infty$, o que en algún momento volvamos a pasar por el estado, lo que hace que este tiempo sea finito. Cómo es lógico, nuestra forma de medir si se pasa o no, de nuevo, por el estado y, va a ser mediante el uso de probabilidades. Por esto definimos $\rho_{yy} = P_y(T_y < \infty)$ es decir, la probabilidad de que, partiendo del estado y, regresemos a él mismo en algún momento.

Para estudiar esto, vamos a introducir un nuevo concepto y una propiedad importante. Decimos que T es un tiempo de parada si la ocurrencia de un evento que nos permite decir que pararemos en el momento n. puede determinarse con la simple observación del proceso en los momentos hasta n. T_y es un tiempo de parada puesto que:

$$\{T_y = n\} = \{X_1 \neq y \dots X_{n-1} \neq y, X_n = y\}$$

Teorema (propiedad fuerte de Markov). *Supongamos que T es un tiempo de parada. Dado $T = n$ y $X_T = y$, cualquier otra información sobre $X_0,\dots, X_T$ es irrelevante para predecir X_{T+k}, $k \geq 0$, ya que el comportamiento será igual al del proceso inicial.*

El planteamiento que emplearíamos para obtener la probabilidad de pasar por el estado y al menos dos veces es intuitivo al utilizar la propiedad fuerte de Markov, hemos llegado por primera vez a y, queremos llegar otra vez, por lo que la probabilidad de hacerlo de nuevo vuelve a ser ρ_{yy}, por lo que la probabilidad de pasar dos veces por ese estado va a ser $\rho_{yy} * \rho_{yy} = \rho_{yy}^2$. Esto podemos generalizarlo para el número de veces que queramos, viendo que la probabilidad de pasar k veces por el estado y, partiendo del mismo, va a ser ρ_{yy}^k.

Estas probabilidades son la que nos van a permitir separar los estados en dos clases.

- Si $\rho_{yy} < 1$ entonces nos encontramos ante un **estado transitorio**, es decir, si pensamos en la probabilidad de pasar k veces por el estado y disminuirá con el aumento de k, de manera que cuando $k \to \infty$, $\rho_{yy}^k \to 0$. Estos estados dejarán de ser visitados con el tiempo.
- Si $\rho_{yy} = 1$ entonces nos encontramos ante un **estado recurrente**, en este caso, da igual el número de veces que queramos pasar por el, la potencia de la probabilidad siempre va a ser 1, por esto podemos asegurar que volveremos a llegar al estado y.

Además, los estados recurrentes pueden ser **estados absorbentes** o no, un estado absorbente es aquel del que, una vez llegado a él, no podemos salir.

No solo podemos considerar importante conocer la probabilidad o el tiempo que se tarda en volver a un estado y, también nos resulta útil estudiar la probabilidad o el tiempo de, partiendo de un estado x, llegar al estado y en algún momento. Vamos a definir $P_x(T_y < \infty)$ como la probabilidad de, partiendo del estado x, llegar al estado y alguna vez.

Si suponemos que $P_x(T_y \leq k) \geq \alpha > 0$ para cualquier x perteneciente al espacio de estados, entonces tendremos que

$$P_x(T_y > nk) \leqslant (1-\alpha)^n$$

Partiendo de esto podemos decir que un estado x va comunicar con un estado y, lo que escribiremos como $x \rightarrow y$, si hay una probabilidad positiva de alcanzar el estado y estando en x, esto lo expresamos como

$$\rho_{xy} = P_x(T_y < \infty) > 0$$

Tener en cuenta que no estamos estableciendo que se deba llegar en el paso siguiente, sino que la probabilidad de llegar en algún momento del tiempo no sea 0, pudiendo ser necesarios mas de un paso. Además debemos entender la diferencia existente entre ρ_{xy} y ρ_{yx} ya que x puede conectar con y, pero en el caso, por ejemplo, de que y sea un estado absorbente, no se podría salir de él y por lo tanto no se podría regresar a x.

Si $\rho_{xy} > 0$, pero $\rho_{yx} < 1$, entonces podemos asegurar que el estado x va a ser transitorio, ya que no somos capaces de asegurar que podamos volver a él. Aún así, vamos a probarlo.

Denominamos $K = min\{k : p^k(x,y) > 0\}$ el mínimo número de pasos que podemos dar desde x hasta y. Como $p^K(x,y) > 0$ tiene que haber una secuencia $y_1, \ldots, y_{k-1}$ tal que

$$p(x,y_1)p(y1,y_2)\ldots p(y_{K-1},y) > 0$$

Al ser K mínimo todas las y_i van a ser distintas de y, por lo que tenemos que

$$P_x(T_x = \infty) \geqslant p(x,y_1)p(y_{1,}y_2)\ldots p(y_{K-1},y)(1-\rho_{yx}) > 0$$

por lo que x es un estado transitorio.

Por otro lado, podemos demostrar que si x es recurrente, entonces si $\rho_{xy} > 0$ se tiene que $\rho_{yx} = 1$, ya que como hemos dicho, si ρ_{yx} fuera menor que 1, el estado sería transitorio.

A partir de lo ya explicado podemos llegar a la definición de **conjunto de estados cerrado**. Un conjunto de estados $A \in S$ se considera cerrado si es imposible salir del mismo, es decir, si dado un estado $i \in A$ y otro estado $j \notin A$, entones, tendremos que p(i,j) = 0. Por otro lado, tendremos un **conjunto de estados irreducible** B si $\forall i, j \in B$ i conecta con j.

Si el conjunto de estados C es finito, cerrado e irreducible, entonces todos los estados de C van a ser recurrentes.

Los conjuntos de estados finitos vamos a poder representarlos como la unión de conjuntos disjuntos, de los cuales algunos van a ser transitorios T y otros van a ser recurrentes R

$$S = T \cup R_1 \cup R_2 \cup \ldots \cup R_n$$

Si x es recurrente y además, $x \rightarrow y$, entonces y será también recurrente. Esto tiene sentido ya que si es seguro regresar a x en el futuro, y desde este podemos llegar a y, entonces sabemos que acabaremos llegando a y el mismo número de veces que a x.

Por otro lado, si tenemos un conjunto finito de estados, tiene que haber al menos uno de ellos que sea recurrente, ya que, al tener un número limitado de estados, tendremos dos opciones, o acabar en uno de ellos, o acabar dando vueltas en un conjunto cerrado e irreducible de la cadena. Si todos los estados fueran transitorios, es decir, si con el tiempo, todos los estados fueran a dejar de visitarse, siempre necesitaríamos un estado nuevo al que llegar, que por ser también transitorio, debería dejar de visitarse en algún momento para acabar en otros.

Para demostrar esto necesitamos definir el número de visitas a un estado partiendo del estado x, como N(y), de manera que seamos capaces de calcular la esperanza de esto mismo como EN(y)

$$E_x N(y) = \frac{\rho_{xy}}{1-\rho_{yy}}$$

Vamos a demostrar esta igualdad:

Al ser X una variable que toma valores enteros no negativos, el valor esperado puede calcularse como

$$EX = \sum_{k=1}^{\infty} P(X \geqslant k)$$

Una vez visto esto, vemos que la probabilidad de regresar a y al menos k veces $\{N(y) \geqslant k\}$ es la misma que la probabilidad de que ocurra el k-ésimo retorno al estado y. $\{T_y^k < \infty\}$, usando esto llegamos a que

$$E_x N(y) = \sum_{k=1}^{\infty} P(N(y) \geqslant k) = \rho_{xy} \sum_{k=1}^{\infty} \rho_{yy}^{k-1} = \frac{\rho_{xy}}{1-\rho_{yy}}$$

ya que nos encontramos con una serie geométrica, $\sum_{0}^{\infty} b^n = \frac{1}{1-b}$, siempre que |b| < 1

Si nos encontramos con el caso de que y es un estado recurrente, es decir que $\rho_{yy} = 1$, entonces, observaremos que $E_y N(y) = \infty$. De esta manera llegamos a que un estado y es recurrente si y solo si

$$\sum_{n=1}^{\infty} p^n(y,y) = E_y N(y) = \infty$$

2.4 Distribución estacionaria

Podemos encontrarnos con casos en los que la matriz de transición en un número grande de pasos acabe tendiendo a lo siguiente:

$$\begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & \cdots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix}$$

Es decir, estamos planteando que $P^n(i,j) \overrightarrow{n \to \infty} \pi_j$, por lo que la probabilidad de transición a la larga va a depender únicamente del estado al que queramos llegar y no del que estemos partiendo.

$$P^n(i,j) \approx \pi_j \qquad y \qquad P^n(l,j) \approx \pi_j \qquad i \neq l$$

Vamos a denominar las probabilidades iniciales como q(i), es decir, $q(i) = P(X_0 = i)$

$$q_n(j) = P(X_n = j) = \sum q(i) * p^n(i,j) \overrightarrow{n \to \infty} \sum q(i) * \pi_j = \pi_j$$

Con esto estamos viendo que no es solo es el límite para las probabilidades condicionadas p(i,j), si no que también es el límite para las probabilidades incondicionadas q(i)

$$q_n = q * p^n$$
$$q_{n+1} = q * p^{n+1} = q * p^n * p = q_n * p$$
$$si\ q_n \overrightarrow{n \to \infty} \pi\ entonces\ q_{n+1} \to \pi\ y\ q_n P = \pi P$$

Por lo que llegamos a que

$$\pi = \pi P$$

Que se conoce como **distribución estacionaria**. La mejor manera de estudiar la distribución estacionaria de una cadena de Markov va a ser mediante las ecuaciones de equilibrio.

$$\pi_j = \sum_i \pi_i * p(i,j) \; \forall \, j$$

Además recordar que la suma de las distintas π debe ser igual a 1, ya que esta es la probabilidad de acabar en alguno de los estados.

2.5 Comportamiento límite

El estudio de los estados transitorios no es de importancia, ya que sabemos que la probabilidad de continuar visitandolos va a acabar siendo cero. Por otro lado, es muy interesante estudiar los estados recurrentes, puesto que son los posibles estados a los que va a llegar la cadena de Markov tras el paso del tiempo.

Para estudiar el comportamiento límite de las cadenas, vamos a definir primero lo que es el periodo para un estado. El **periodo** se considera el máximo común divisor de los n tales que es posible volver al estado en n pasos. Cuando una cadena es finita e irreducible, todos los estados van a tener el mismo periodo.

Vamos a considerar una cadena **aperiódica** cuando sea una cadena de periodo 1.

Si $\rho_{xy} > 0$, $x \to y$ será $\rho_{yx} > 0$, entonces x e y van a tener el mismo periodo, siempre que la cadena sea irreducible, aperiódica o todos los estados tengan el mismo periodo.

En las cadenas reducibles podemos encontrar estados con periodos distintos. Una forma de saber fácilmente si una cadena va a ser aperiódica es observando la matriz de probabilidades de transición. Si P no contiene ningún cero, entonces podemos decir que la cadena es aperiódica, ya

que en todos los estados tendremos una posibilidad positiva de quedarnos en él mismo, y entonces en un paso podemos llegar al mismo estado.

Teorema de la convergencia.

Suponemos que una cadena es irreducible aperiódica y con distribución estacionaria, entonces, si $n \to \infty$, $p^n(x,y) \to \pi(y)$.

Teorema de la frecuencia asintótica.

Si la cadena es irreducible y con todos sus estados recurrentes, y llamamos $N_n(y)$ al número de visitas al estado y en el momento n. Entonces podemos expresar la frecuencia relativa de visitas al estado y como $\frac{N_n(y)}{n}$ y su comportamiento límite será

$$\frac{N_n(y)}{n} \xrightarrow{n\to\infty} \frac{1}{E_y T_y}$$

Por lo que debemos tener en cuenta que E_yT_y debe ser finita para que la obtención de la frecuencia relativa a largoplazo sea posible.

Además, si la cadena es irreducible y con distribución estacionaria, tendremos que

$$\pi(y) = \frac{1}{E_y T_y}$$

Para justificar lo anterior, observamos que si

$$N_n(y)=k$$
$$\frac{t_y^1+\ldots+t_y^{k+1}}{k} \leqslant \frac{n}{k} \leqslant \frac{t_y^1+\ldots+t_y^{k+1}}{k+1}$$

Por la ley de los grandes números, todos los términos de la última inecuación van a tender a $E_y(T_y)$ cuando n (y k) tiende a infinito.

2.6 Tiempos de salida

En cadenas de Markov en las que encontremos estados absorbentes puede ser útil querer calcular el tiempo medio que se tarda en llegar a dichos estados, esto es lo que vamos a denominar **tiempo de salida**.

Definimos g(x) como el tiempo que se tarda, partiendo del estado x en llegar a alguno de los estados absorbentes.

En casos en los que tenemos más de un estado absorbente no tiene sentido querer estimar el tiempo que tardaríamos en llegar a uno concreto porque es probable que nunca lleguemos a él ya que acabemos en alguno de los otros estados que son absorbentes.

Vamos a calcular los distintos g(x) haciendo uso del método del primer paso. Lo que pretendemos con este método es pensar, para cada uno de los estados que formen la cadena de Markov, cual podría ser el estado en el siguiente momento temporal. Es decir,

$$g(x) = 1 + \sum_{y} p(x,y) * g(y) \; si \; x \; es \; transitorio$$
$$g(x) = 0 \; si \; x \; es \; absorbente$$

Al obtener esta ecuación para cada uno de los estados, estaremos consiguiendo un sistema de ecuaciones determinado, que podremos resolver para obtener los tiempos medios de salida de la cadena de markov para cada uno de los estados de la misma.

Veamos un ejemplo sencillo que pueda ilustrarnos.

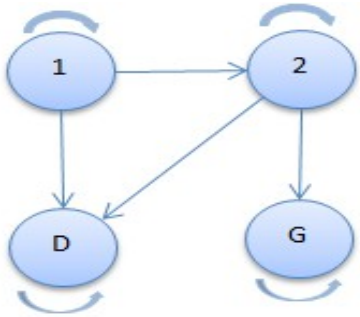

Figura 2: Cadena de Markov

Con esta cadena de Markov estamos representando la trayectoria durante unos estudios de dos años. Los estados representan lo siguiente:

- 1: el alumno está en primer curso.
- 2: el alumno está en segundo curso.
- D: el alumno abandona los estudios.
- G: el alumno acaba los estudios.

Planteamos la siguiente matriz de transición

$$P = \begin{pmatrix} 0,25 & 0,6 & 0 & 0,15 \\ 0 & 0,2 & 0,7 & 0,1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Con el método del primer paso tendríamos las siguientes ecuaciones para cada uno de los estados de la cadena.

$$g(1) = 0,25 * g(1) + 0,6 * g(2) + 1$$
$$g(2) = 1 + 0,2 * g(2)$$
$$0,8 * g(2) = 1 \; ; \; g(2) = \frac{5}{4}$$
$$0,75 * g(1) = 0,6 * \frac{5}{4} + 1 \; ; \; g(1) = \frac{7}{3}$$

Podemos encontrar otras dos formas de obtener estos tiempos. Para la primera de ellas vamos a definir la matriz r como aquella matriz en la que no tenemos en cuenta las filas y columnas que pertenecen a los estados absorbentes. En el ejemplo anterior la matriz r sería

$$r = \begin{pmatrix} 0{,}25 & 0{,}6 \\ 0 & 0{,}2 \end{pmatrix}$$

A partir de esta matriz podemos calcular los tiempos de salida como:

$$g(x) - \sum_{y} r(x, y) * g(y) = 1$$

Lo que podemos reescribir como

$$(I - r)g = \mathbf{1}$$

Donde el **1** va a representar un vector de una dimensión del mismo tamaño que la matriz r en el que todas las componentes van a ser el 1. Por lo que, si despejamos para obtener el valor de las g, tenemos

$$g = (I - r)^{-1}\mathbf{1}$$

La última manera que vamos a ver de obtener los tiempos de salida en una cadena de Markov discreta va a tener en cuenta que N(y) es el número de visitas al estado y en el momento $n \geqslant 0$, por lo que

$$E_x N(y) = \sum_{n=0}^{\infty} r^n(x, y)$$

Capítulo 3: Genética de poblaciones

3.1 Definición

Como ya hemos dicho en la introducción, la genética poblacional o genética de poblaciones se encarga del estudio de las fuerzas que alteran la composición genética de una especie. Entendiéndose la composición genética como la variación que hay en los genes y genotipos en una población, así como la distribución que hay de dicha variación entre los individuos.

Una característica importante de la genética de poblaciones es que nos es indiferente la estructura básica de la genética, es decir, nos basta con el conocimiento de las leyes de la herencia y la acción que tienen sobre ellas las fuerzas evolutivas. Esto ha quedado demostrado ya que la teoría básica de esta ciencia ha sido desarrollada en su mayor parte antes de que llegara a conocerse la estructura del ADN.

En este área el azar tiene un papel significativo, lo cual nos indica que las predicciones que se puedan hacer sobre los modelos utilizados para esta clase de problemas deben formularse en términos de probabilidad, por lo que la herramienta que se va a utilizar son los modelos matemáticos que dejan de lado la intuición, la cual falla en la mayoría de las ocasiones.

Tal y como hemos explicado, la variabilidad que encontramos en la población es una de las bases para este estudio. Tenemos que plantearnos hasta que punto existe tal variedad, y es que, solo con el hecho de observarnos entre los mismos humanos nos podemos percatar de que esta variabilidad es mucho mayor de lo que se puede pensar a primera vista. Dentro de un mismo grupo de personas encontramos colores de ojos distintos, alturas, pesos, etc. Esta variabilidad somos capaces de verla también en animales y resto de seres vivos, tamaños, colores e incluso resistencia ante enfermedades, aún así, al fijarnos solamente en los fenotipos de la población estamos obviando cierta variabilidad, ya que un fenotipo puede darse con uno o más genotipos.

Es por esto que el estudio de las frecuencias genéticas nos va a permitir entender las características de las poblaciones de estudio.

3.2 Frecuencias genéticas y ley de Hardy-Weinberg

La frecuencia genica, también conocida como frecuencia alélica, es la proporción de cada alelo en un locus dado en una población específica. Al estar hablando de proporciones, sabemos que la suma de todas las frecuencias debe darnos 1.

Debemos diferenciar la frecuencia genica de la frecuencia genotípica. La frecuencia genotípica es la proporción de genotipos que tenemos en una población, mientras que la frecuencia genica es la proporción de cada uno de los alelos. Tenemos que tener en cuenta que, dado un locus con dos alelos A y a, los genotipos que podemos tener, para individuos diploides, van a ser AA, Aa y aa. Es decir, que con la frecuencia alélica vamos a trabajar con la cantidad de veces que tenemos que de A o a, mientras que la frecuencia genotípica va a ser la cantidad de AA, Aa y aa que tenemos en

la población. A partir de las frecuencias genotípicas vamos a ser capaces de obtener las frecuencias alélicas, sin embargo, al contrario no podemos obtenerlo. Veamos cómo se consigue.

Suponemos p = proporción del alelo A y q = proporción del alelo a. Claramente p+q = 1. Así pues, si nos fijamos en cada uno de los genotipos observamos que:

- Para el genotipo AA p(A) = 1 y p(a) = 0
- Para el genotipo Aa p(A) = 0,5 = p(a)
- Para el genotipo aa p(A) = 0 y p(a) = 1

Una vez que entendemos esto, de manera lógica llegamos a que las proporciones alélicas van a obtenerse como

$$p = \frac{1*(AA)+0{,}5*(Aa)}{N} \qquad q = \frac{1*(aa)+0{,}5*(Aa)}{N}$$

Que sepamos obtener las proporciones alélicas no nos permite conocer las proporciones genotípicas, esto solamente podemos conseguirlo en condiciones específicas que se dan en las poblaciones que se encuentran en **equilibrio.**

Una población está en equilibrio si los alelos mantienen sus frecuencias a través de las generaciones. Para que una población se considere en equilibrio debe cumplir la **ley de Hardy-Weinberg**, la que también se conoce cómo equilibrio panmíctico. Para que se de esta ley, debemos tener en cuenta las siguientes suposiciones.

1. Los individuos de cualquier genotipo tienen iguales tasas de supervivencia e igual éxito reproductivo.
2. No aparecen nuevos alelos o se convierten uno en otros por mutaciones.
3. No hay migración hacia o desde la población.
4. La población es infinitamente grande, es suficientemente grande como para poder despreciar tanto los errores de muestreo como los efectos aleatorios.
5. Los individuos deben aparearse de manera aleatoria.

En poblaciones en las que se encuentre dicho equilibrio va a ser posible obtener, en base a la probabilidad, las frecuencias de los genotipos esperadas. De nuevo, vamos a considerar p como la probabilidad del alelo A y q la probabilidad del alelo a. La proporción de los gametos puede explicarse con la siguiente tabla.

	Hembras		
Machos	Gametos	p	q
	p	AA p^2	Aa pq
	q	Aa pq	aa q^2

Tabla 1: proporción de gametos.

Es decir, la proporción esperada para el genotipo AA va a ser p^2, para Aa tendremos 2pq y para aa, q^2. Si nos fijamos, la suma de las tres proporciones coincide con el cuadrado de una suma. Es decir:

$$(p+q)^2 = p^2+2pq+q^2 = 1$$

Se sigue manteniendo la propiedad de las proporciones, la suma de las proporciones de las distintas posibilidades va a ser 1.

De esta ley vamos a sacar tres consecuencias claras:

1. Los caracteres dominantes no van a aumentar, por lo que los recesivos tampoco van a disminuir.
2. La variabilidad genética puede mantenerse.
3. Conociendo las frecuencias de un genotipo podemos calcular las frecuencias de los otros dos.

3.3 Alteración del equilibrio

Si las poblaciones se mantuvieran siempre en el equilibrio de Hardy-Weinberg, la genética de poblaciones no sería necesaria porque sería fácil predecir las proporciones alélicas y genéticas que se iban a tener en las generaciones siguientes, sin embargo, las condiciones que se suponen para que una población esté en equilibrio no son fáciles de encontrar. En la realidad tenemos distintos mecanismos que se encargan de los cambios en dichas frecuencias, en esta sección vamos a pasar a estudiar dichos mecanismos.

3.3.1 Mutaciones

La mutación es el único mecanismo que va a producir nuevos alelos. Las mutaciones son aleatorias, no pretenden beneficiar ni perjudicar a los organismos. Por supuesto, la aparición de nuevos alelos va a producir un cambio en las proporciones alélicas, ya que la totalidad va a tener que dividirse entre un mayor número de posibilidades.

Sin embargo, es importante tener en cuenta la proporción en las que se producen las mutaciones. Observar esta proporción puede ser complicado debido al hecho de que las mutaciones suelen ser recesivas en organismos diploides, por lo que la observación de esto debe realizarse mediante cálculo de probabilidades o con programas de análisis a gran escala. La proporción de mutaciones se expresa como el número de nuevos alelos mutantes por cada número de gametos dado.

Las mutaciones son un mecanismo relativamente lento en el cambio de las frecuencias alélicas, únicamente en el caso de las bacterias que las generaciones se miden en minutos pueden verse como un factor importante.

3.3.2 Migraciones

Las migraciones se dan cuando los individuos se desplazan entre las poblaciones, quedando geográficamente separados. Los individuos que migren llevan consigo alelos diferentes, que van a cambiar las frecuencias de la población receptora. Este fenómeno se conoce como flujo génico.

Las frecuencias alélicas de la generación siguiente a la llegada de los inmigrantes van a verse afectada por este hecho. El cambio que se produce en las proporciones alélicas va a ser proporcional a las diferencias en frecuencias entre la población autóctona y la población inmigrante, y la tasa de migración.

Otro efecto que encontramos con las migraciones es que prevenimos la divergencia genética entre poblaciones distintas, esto contrarresta el efecto de la selección natural y la deriva genética que veremos más adelante, manteniendo las poblaciones homogéneas en sus frecuencias alélicas.

3.3.3 Selección natural

La selección natural se da al presentarse la situación en que un genotipo presenta una ventaja selectiva sobre los otros. Esto puede considerarse un proceso de selección que actúa a través de las diferencias de los genotipos en cuanto a variabilidad y a fertilidad, produciendo diferencias en la prolificidad relativa. Si estos rasgos adaptativos que suponen una ventaja tienen una base genética, serán heredados por la descendencia, apareciendo en mayor frecuencia en la generación siguiente.

Otro factor que influye en la selección natural proviene de la mortalidad prerreproductiva, y es que no todos los individuos que forman una generación van a reproducirse, la mayoría de los casos, debido a la muerte de los individuos de manera precoz.

3.3.4 Endogamia

También conocida como endocría se define como el cruzamiento entre individuos emparentados. Esto produce el aumento en las proporciones de los individuos homocigoticos. Cuanto mayor sea el número de casos de endogamia, mayor será la probabilidad de que en una población, para un locus determinado, ambos alelos sean idénticos por ascendencia, es decir, que ambos alelos provengan del mismo gen antecesor.

Un ejemplo claro de la endogamia es la autofecundación, en los casos de genotipos heterocigotos, la progenie podrá tener genotipos tanto hetero como homocigotos en las proporciones 1/4:1/2:1/4, correspondientes a AA:Aa:aa, sin embargo, de los individuos homocigotos solo podrá obtenerse descendencia homocigota. Los individuos homocigotos que provengan de progenitores heterocigotos se sumarán a aquellos que provienen de progenitores homocigotos, aumentando la proporción de los mismos y disminuyendo la de los heterocigotos.

3.3.5 Deriva genética

La deriva genética se define como el cambio en las frecuencias alélicas que resulta del muestreo aleatorio de gametos. Es un suceso que no vamos a observar en poblaciones grandes, ya que estas estarían en el equilibrio de Hardy-Weinberg. Al ser este el tema principal del estudio, vamos a dedicarle un apartado entero para su mejor explicación.

3.4 Deriva genética

Como hemos visto, según la ley de Hardy-Weinberg, en las poblaciones en equilibrio, las frecuencias alélicas se deben mantener constantes de una generación a la siguiente. Ya hemos visto diversos motivos que pueden producir que esta ley no se cumpla, ahora vamos a centrarnos en un mecanismo que altera estas frecuencias y que se basa en el hecho de que las poblaciones no tienen un número infinito de individuos, en muchas ocasiones ni siquiera tienen un número de individuos suficientemente grande como para considerar que se cumple el equilibrio panmíctico.

En poblaciones en las que el número de individuos es limitado, vamos a encontrar un número más limitado de aquellos que sean capaces de reproducirse. Es en estas poblaciones en las que la deriva genética, actuando como un caso particular de los errores de muestreo más se puede apreciar. Sabemos que la magnitud del error de muestreo es inversamente proporcional al tamaño muestral, cuanto menor sea el tamaño de la muestra, mayores serán los efectos de la deriva.

En el caso de la genética encontramos que cuanto menor sea el número de reproductores, mayores serán los cambios en las frecuencias alélicas. Por el contrario, cuanto mayor sea el número de progenitores de una generación, mas cerca estarán las frecuencias observadas (las de los descendientes) a las frecuencias esperadas (las de los progenitores). En el caso de que se escojan unos pocos individuos para dar lugar a la siguiente generación, al ser una muestra pequeña podemos encontrarnos con que no sea una muestra representativa y que, por lo tanto, las frecuencias alélicas de la descendencia varíen con respecto a la generación de los padres.

Podemos suponer que la condición de tener una población pequeña no es algo que encontremos fácilmente en la naturaleza, los humanos somos seres sociales que vivimos en grandes comunidades, al igual que ocurre con muchos otros animales, pero hasta organismos tan simples como las bacterias suelen encontrarse en grande número, por ello la importancia de la deriva genética se ha demostrado principalmente en dos situaciones :

1. **Efecto cuello de botella:** se produce si una población queda reducida en muy pocos individuos por causas ajenas a la propia población, como pueden ser sucesos ambientales,

por ejemplo los tifones, o causas humanas como la caza. Se dice que la población pasa por un cuello de botella por el que pasa solo una pequeña parte de la población.

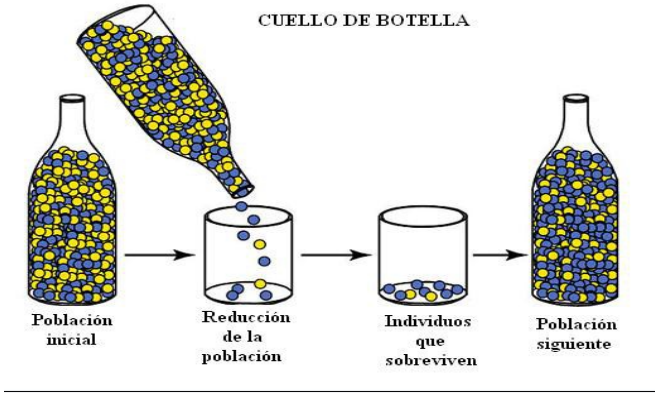

Figura 3: efecto cuello de botella

2. **Efecto fundador:** surge cuando una población pequeña se separa de una población mayor y unos pocos fundadores colonizan una nueva región. Estos fundadores llevan consigo solo una pequeña parte de la variación genética presente en la población original, por lo que los pocos alelo presentes en los descendientes serán los que se encuentren entre los que tenían los colonizadores.

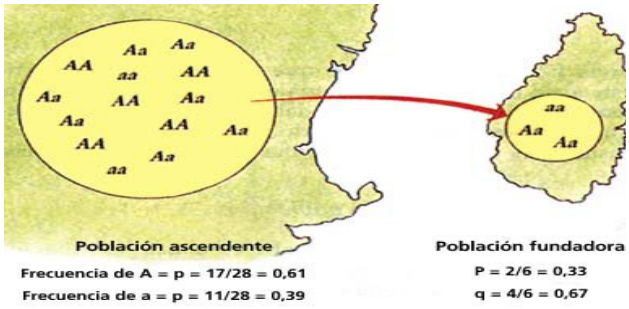

Figura 4: efecto fundador.

Aunque es cierto que no es posible predecir los cambios en frecuencias alélicas en una única población, si vamos a poder predecir el comportamiento promedio de las frecuencias alélicas en un número grande de poblaciones. Para esto vamos a suponer que una población inicialmente grande va a subdividirse en poblaciones más pequeñas de tamaño N que van a cumplir las siguientes condiciones:

1. Existe igual número de machos que de hembras.
2. El tamaño de cada subpoblación es constante.
3. El apareamiento sigue siendo al azar.
4. No se producen procesos de migración.
5. No hay mutaciones.
6. No hay selección natural.

Bajo estas suposiciones, se ha desarrollado un modelo que plantea que un conjunto de subpoblaciones van a actuar de la misma forma y de las cuales sólo una o unas pocas van a tener existencia material. El **modelo de Fisher-Wright** considera las consecuencias del muestreo binomial que ocurre en poblaciones pequeñas a lo largo de las generaciones. De esta forma obtenemos un modelo que predice la distribución de las frecuencias alélicas en un conjunto de poblaciones sometidas a la acción de la deriva genética.

Por otro lado, encontramos el **modelo de Moran** que no supone que la reproducción se haga a la vez en toda la población, sino que la reproducción es individual. Para este modelo consideramos dos eventos posibles, el primero es el evento de nacimiento, que se da con la reproducción de un individuo aleatorio de la población; el segundo es un evento de muerte, por el cual, cualquier individuo de la población muere. Así pues en el modelo de Moran consideramos que se pueden dar eventos de nacimiento y muerte en el mismo instante de tiempo, es decir, cualquier individuo puede ser escogido para reproducirse, en este momento la población tendrá un individuo más que antes, para mantener el tamaño de la población cualquier individuo de la misma puede ser elegido para morir.

Si una población tienen 2N alelos para un locus dado que pueden ser A o a, podemos describir la población según el número de alelos A que se presenten. Los estados posibles es que encontremos 0, 1, 2, …, 2N alelos A.

Capítulo 4: Ejemplos

Una vez visto dos maneras de modelizar los cambios en las frecuencias alélicas, vamos a pasar a estudiar sus comportamientos mediante el uso de las cadenas de Markov. Para esto vamos a plantear condiciones iniciales ficticias y vamos a realizar simulaciones las mismas, haciendo uso del software R y de la librería markovchain.

Unas recordaciones antes de pasar a las obtenciones de los cálculos.

- Tenemos dos alelos para un locus concreto A y a.
- Llamaremos p = proporción del alelo A y q = proporción del alelo a.
- Al ser proporciones p + q = 1.

4.1 Modelo de Fisher-Wright

Partimos de una población pequeña, de tamaño N = 30 de individuos diploides, por lo que el número total de alelos que vamos a tener es de 2N = 60.

Como hemos fijado el tamaño de la población en todas las simulaciones, la matriz de transición que vamos a tener en los distintos casos van a ser similares. Definimos como proceso de Markov

$$X_t = \textit{número de alelos de tipo A en la generación } t \qquad X_t \in [0,\ 60]$$

Por lo que la matriz de transición de las distintas simulaciones va a ser similar, teniendo 61 estados posibles, donde los estados 0 y 60 van a ser recurrentes y absorbentes, mientras que el resto de los estados van a ser transitorios.

Según el modelo de Fisher-Wright las distintas probabilidades de transición pueden obtenerse como casos de una binomial, ya que estas probabilidades vienen dadas por la expresión:

$$p_{ij} = \binom{2\mathrm{N}}{j}\left(\frac{i}{2\mathrm{N}}\right)^j\left(1-\frac{i}{2\mathrm{N}}\right)^{2\mathrm{N}-j}$$

Para poder realizar estas simulaciones, tenemos que fijar las condiciones en las que se van a producir los pasos de las generaciones.

- Todos los individuos de la población tienen la misma probabilidad de reproducirse con cualquier otro individuo.
- La generación descendiente va a tener el mismo tamaño que la generación actual, debido a la muerte o perdida de capacidad reproductiva de los progenitores.
- No consideramos la aparición de mutaciones a lo largo de las generaciones.

Por lo que la matriz de transición va a ser la misma en las tres simulaciones que vamos a ejecutar. La forma que va a tener la matriz de transición la podemos ver con el siguiente gráfico.

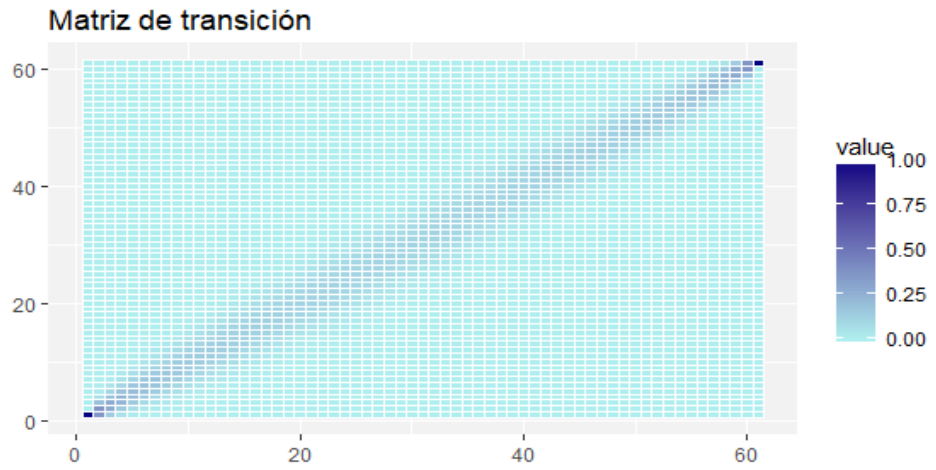

Figura 5: matriz de transición.

Observamos que solamente vamos a tener probabilidades distintas de cero en la zona de la diagonal principal y en los alrededores de esta. Las zonas de azul más claro son las que representan las probabilidades próximas o iguales a cero, mientras que, cuanto más oscura sea la representación, mayor será la probabilidad de transición entre los estados, llegando al azul oscuro que representa una probabilidad de 1 y que solamente se da en los dos estados absorbentes. Este comportamiento es lógico, ya que si en la generación i tenemos, por ejemplo, 15 alelos del tipo A, es improbable que en la siguiente generación los 30 individuos que la formen vayan ser homocigóticos AA.

La cadena de Markov que tenemos en este caso es un ejemplo de una cadena finita, cerrada y con dos estados absorbentes. En este caso no somos capaces de hallar una única distribución estacionaria, sino que al intentar resolver el sistema de ecuaciones encontramos que podemos obtener dos posibles soluciones.

La primera de las distribuciones estacionarias sería

$$\pi_0 = 1 \quad y \quad \pi_i = 0 \quad si \quad i \in [1,60]$$

La segunda distribución estacionaria será

$$\pi_{60} = 1 \quad y \quad \pi_i = 0 \quad si \quad i \in [0,59]$$

Como podemos observar en ambas distribuciones, estamos llegando a la conclusión que esperábamos y es que o desaparece por completo el alelo A o este queda fijado. Realmente no podemos decir que la cadena de Markov tenga distribución estacionaria, sino que hay dos posibles situaciones a largo plazo.

Caso p = 0.1

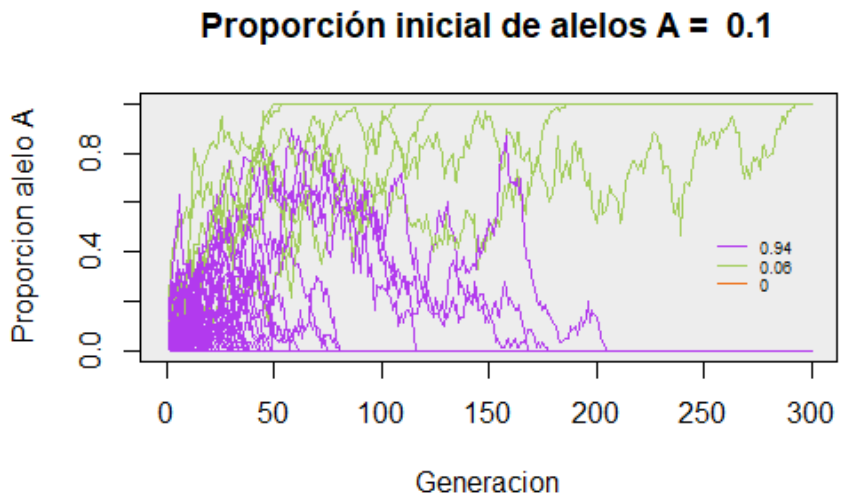

Figura 6: simulación

Como podemos observar en estas simulaciones, al partir de una población en la que la proporción de alelos del tipo A es mucho menor que la de los alelos de tipo a, en la mayor parte de las simulaciones, 94 de 100, el resultado final que obtenemos es la perdida total del alelo A. Observamos que aunque la mayoría de las simulaciones la perdida de este alelo se consigue en menos de 100 generaciones, en aquellas en las que el alelo A acaba fijo necesitan mas tiempo y su proporción fluctúa mucho más a lo largo del tiempo.

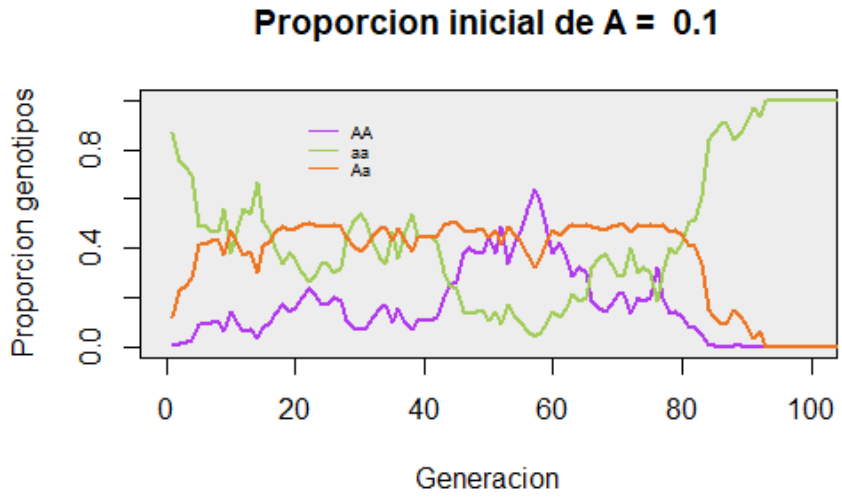

Figura 7: simulación

Como es lógico, en esta simulación sobre la proporción de genotipos encontramos que los genotipos que requieren de alelo A se pierden en el tiempo, en menos de 100 generaciones.

Caso p = 0.5

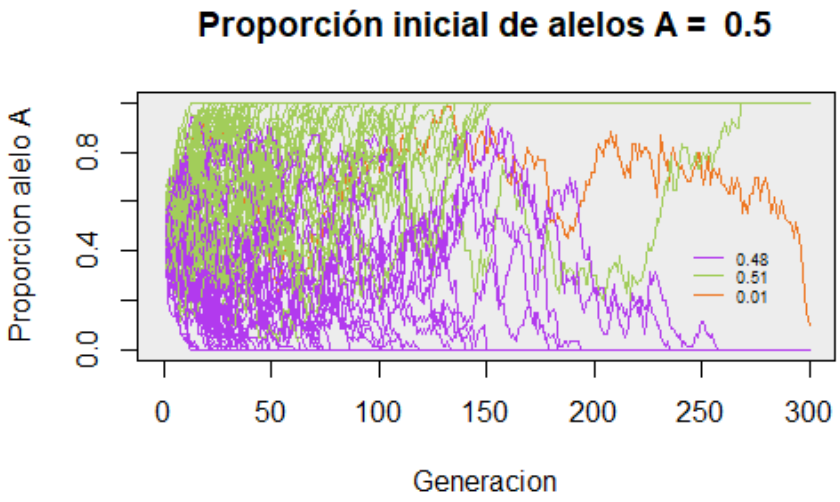

Figura 8: simulación

Para esta situación en la que empezamos con el mismo número de alelos A que de alelos a observamos que la situación cambia considerablemente. Aunque en la mayoría de las simulaciones, pasadas las 300 generaciones alguno de los alelos se ha fijado, tenemos una simulación, que aunque parece que va dirigida a la perdida del alelo A, no llega a perderse. Por otro lado, el número de simulaciones en las que el alelo A se fija y en número de aquellas en las que se pierde es muy similar, esto se debe a haber empezado con un estado que se encuentra de manera equidistante a los dos estados absorbentes.

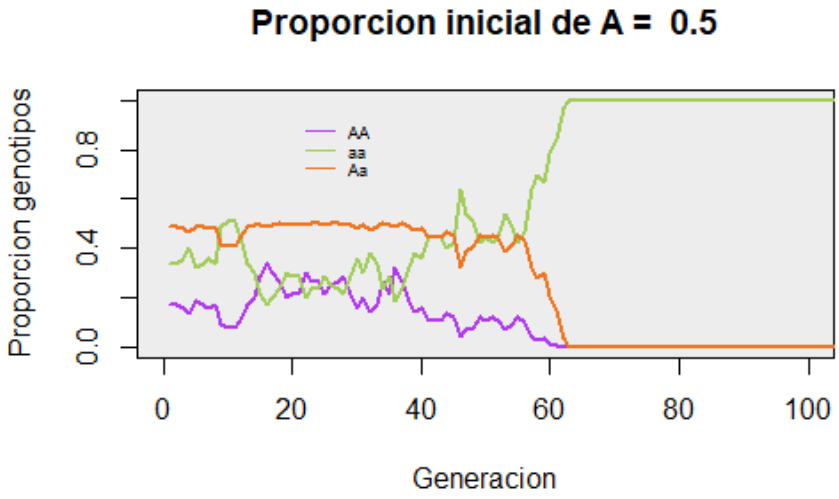

Figura 9: simulación

En la simulación del comportamiento de los genotipos a lo largo del tiempo podemos ver que aunque el genotipo heterocigoto es el que empieza con mayor proporción inicial, en torno al 0.5,

con el paso del tiempo acaba perdiéndose, ya que esta es la consecuencia que se espera de la deriva genética.

Caso p = 0.9

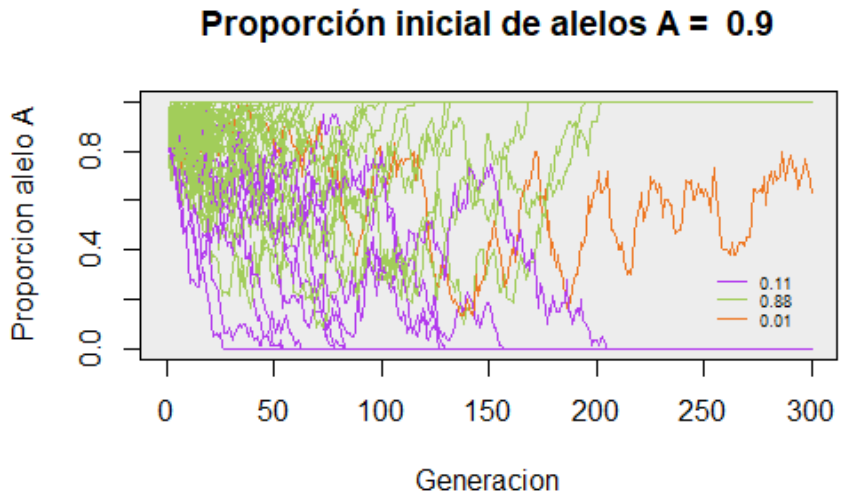

Figura 10: simulación

En esta simulación encontramos una situación contraria a la primera que vimos, al empezar con mayor proporción de alelos A que a la mayor parte de las simulaciones llevan a la fijación de este alelo. Sin embargo, en este caso nos ocurre de nuevo que hay una simulación que no llega ni a la fijación ni a la perdida, y que en el momento de la generación número 300 se encuentra con una proporción de alelo A cercana a 0.6, por lo que no podemos hacernos una idea de lo que podría llegar a pasar en dicha simulación.

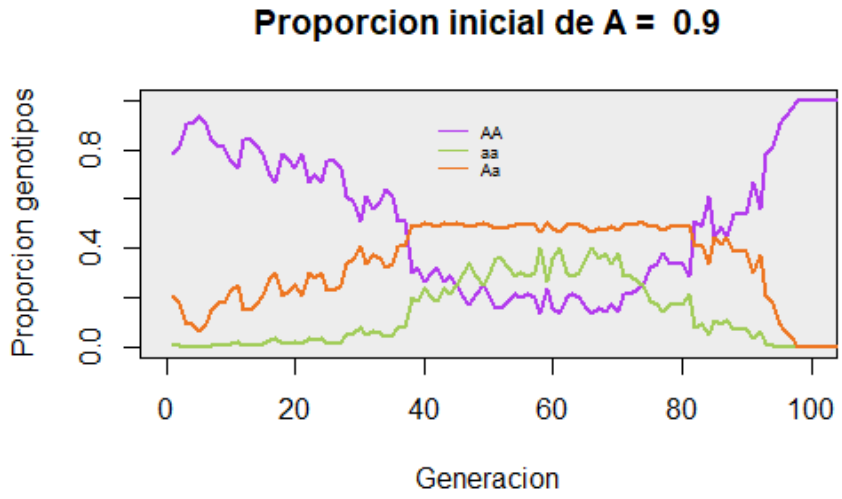

Figura 11: simulación

Además de observar como se puede comportar la cadena de Markov dependiendo de la frecuencia alélica de la generación inicial, también podemos considerar importante estudiar el tiempo esperado de salida de la cadena de Markov, en nuestro caso, vamos a considerar que salimos de la cadena cuando hemos llegado a uno de los estados absorbentes, es decir, cuando hemos llegado a que todos los alelos de la población son A (estado 2N) o cuando llegamos a que todos los estados de la población son a (estado 0).

x	g(x)	x	g(x)	x	g(x)	x	g(x)	x	g(x)
1	9,499	**13**	61,048	**25**	79,659	**37**	78,051	**49**	55,556
2	16,56	**14**	63,497	**26**	80,259	**38**	77,038	**50**	52,488
3	22,689	**15**	65,762	**27**	80,724	**39**	75,883	**51**	49,183
4	28,152	**16**	67,852	**28**	81,055	**40**	74,582	**52**	45,622
5	33,097	**17**	69,773	**29**	81,254	**41**	73,133	**53**	41,779
6	37,618	**18**	71,532	**30**	81,32	**42**	71,532	**54**	37,618
7	41,779	**19**	73,133	**31**	51,254	**43**	69,773	**55**	33,097
8	45,622	**20**	74,582	**32**	81,055	**44**	67,852	**56**	28,152
9	49,183	**21**	75,883	**33**	80,724	**45**	65,762	**57**	22,689
10	52,488	**22**	77,038	**34**	80,259	**46**	63,497	**58**	16,56
11	55,556	**23**	78,051	**35**	79,659	**47**	61,048	**59**	9,499
12	58,404	**24**	78,924	**36**	78,0924	**48**	58,404		

Tabla 2: tiempos de salida

La representación gráfica de la tabla va a ser la siguiente:

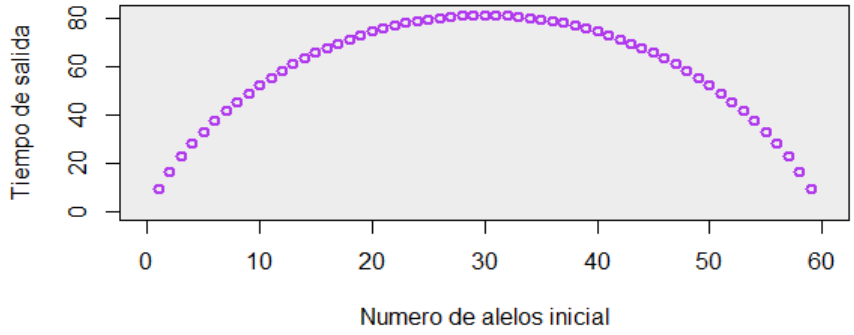

Figura 12: tiempos de salida

Tal y como podríamos esperar, los tiempos de salida tienen una representación simétrica, es decir, el tiempo estimado que tardaremos en llegar a un estado absorbente va a ser el mismo si empezamos con 10 alelos de tipo A que si empezamos con 50, ya que el punto medio de la gráfica es el valor de 30, que coincide con la situación en la que la mitad de la población contara con alelos A y la otra mitad con alelo a.

Vamos a repetir el proceso, pero con una población haploide del mismo tamaño que para el caso diploide, N = 30. Aunque el tamaño poblacional sea el mismo, el conjunto de estados se ha visto muy reducido, $S_x = \{0,...,30\}$. Las probabilidades de transición se obtienen de nuevo con los cálculos de la binomial, ya que el modelo es el mismo.

De nuevo nos encontramos con una cadena de markov es finita, cerrada y con dos estados absorbentes, ya que como hemos dicho la diferencia entre una población haploide o diploide para este modelo no la encontramos en la matriz de transición, sino en las inferencias sobre genotipos que podemos obtener. Por esto mismo, obtenemos resultados similares al caso diploide cuando queremos hallar la distribución estacionaria, con la diferencia de que el número de estados en ambos casos, va a ser distinto.

Ahora las posibles distribuciones que encontramos serán:

$$\pi_0 = 1 \quad y \quad \pi_i = 0 \quad si \quad i \in [1,60]$$
$$o$$
$$\pi_{60} = 1 \quad y \quad \pi_i = 0 \quad si \quad i \in [0,59]$$

Al igual que veíamos para el caso diploide, no podemos considerar que nuestra matriz de transición tenga distribución estacionaria.

Caso p = 0.1

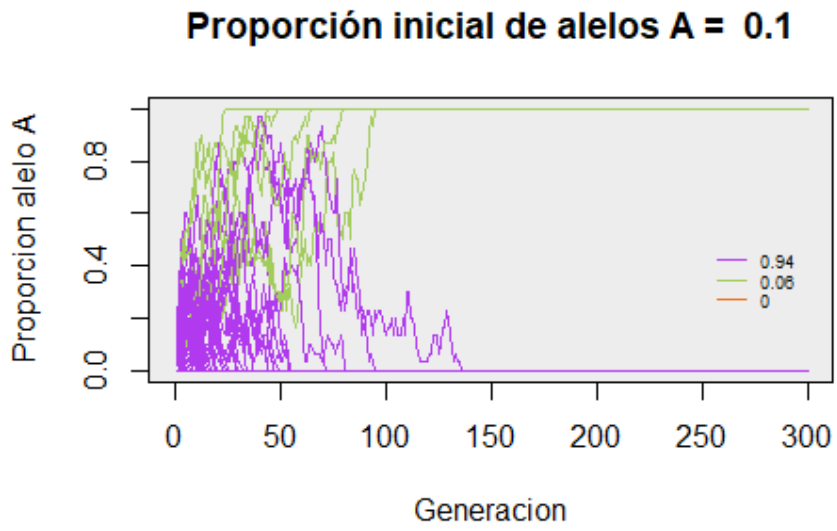

Figura 13: simulación

Como podemos observar, al haber cambiado de una población diploide a una población haploide y al haber reducido el número de estados posibles, encontramos que se necesita menor número de generaciones para llegar a uno de los estados absorbentes. Al igual que en la población diploide, al empezar con una proporción 0,1 de alelos de tipo A, la mayor parte de las simulaciones acaban en la extinción de este alelo.

Caso p = 0.5

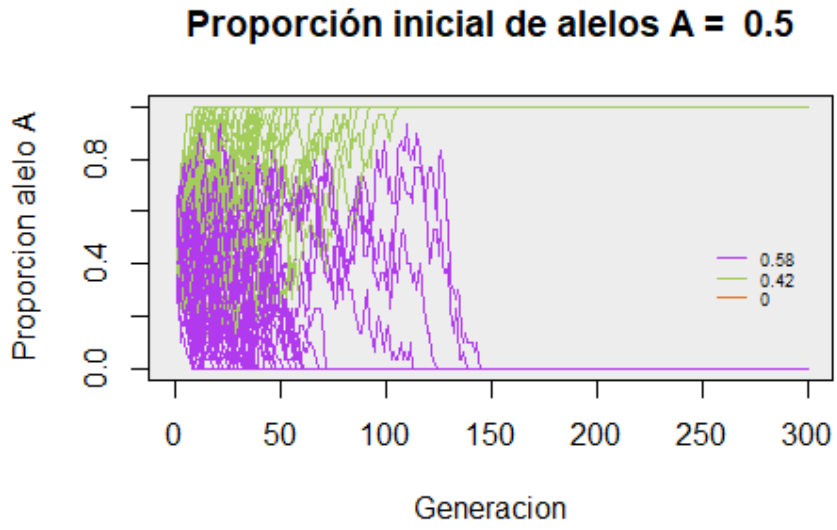

Figura 14: simulación

Igual que nos ocurría en el caso anterior, no son necesarias tantas generaciones para que todas las simulaciones lleguen a alguno de los estados absorbentes. También, al igual que hemos observado en en la población diploide, tendremos las mismas probabilidades de que se fije el alelo A que de que se extinga.

Caso p = 0.9

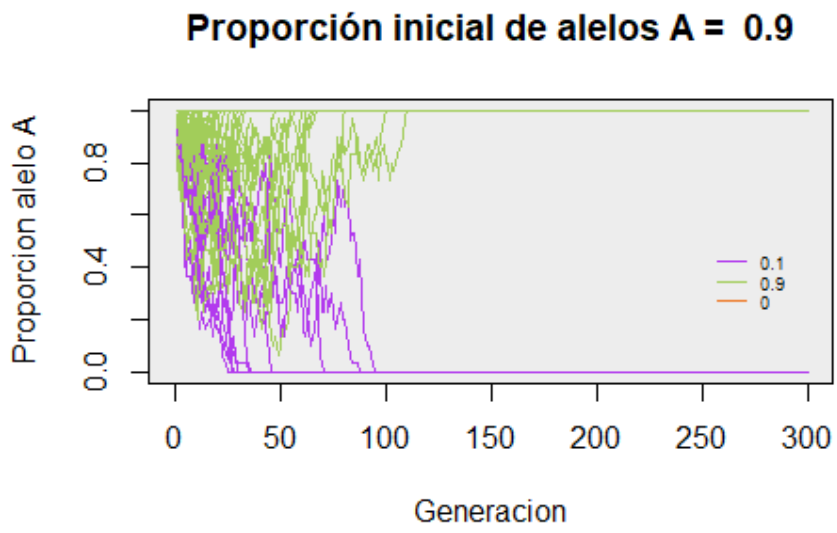

Figura 15: simulación

Nos encontramos de nuevo con la situación de que con menor número de generaciones lleagamos a los estados absorbentes.

Ahora vamos a pasar, al igual que en el caso de los diploides, a observar los tiempos esperados de salida para los distintos estados de la cadena. Por las simulaciones que ya hemos visto podemos esperar que estos tiempos sean mucho menores que en el caso anterior.

x	g(x)	x	g(x)	x	g(x)	x	g(x)	x	g(x)
1	8,076	**7**	31,13	**13**	39,434	**19**	37,837	**25**	25,684
2	13,692	**8**	33,286	**14**	39,828	**20**	36,62	**26**	22,292
3	18,344	**9**	35,109	**15**	39,96	**21**	35,109	**27**	18,34
4	22,292	**10**	36,62	**16**	39,828	**22**	33,286	**28**	13,692
5	25,684	**11**	37,837	**17**	39,434	**23**	31,13	**29**	8,076
6	28,61	**12**	38,772	**18**	38,772	**24**	28,61		

Tabla 3: tiempos de salida

La representación gráfica de estos tiempos de salida va a ser similar a la ya obtenida para la población diploide.

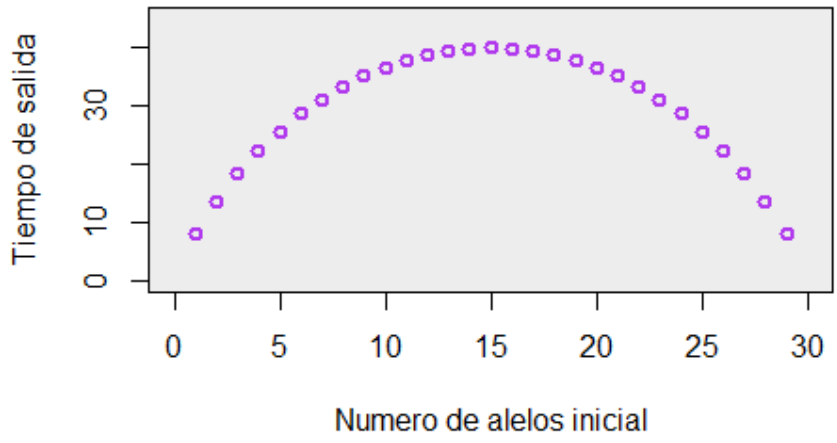

Figura 16: tiempos e salida

A primera vista parece que la gráfica parece ser igual que en el caso diploide, sin embargo al fijarnos en el tiempo máximo que se necesitaría para este caso encontramos una clara diferencia, ya que es hasta el doble del máximo necesario para una población haploide.

En este caso que la población sea haploide o diploide solamente se diferencian en el tamaño de la población. Los resultados obtenidos para una población diploide de 30 individuos serán los mismos que los obtenidos para una población haploide de 60. Sin embargo, el conocer cómo son los individuos de la población si nos puede resultar interesante para obtener las representaciones gráficas de los genotipos en las simulaciones, ya que estas no tienen sentido en organismos haploides.

4.2 Modelo de Moran

Al igual que hemos hecho con el modelo de Fisher-Wright, para el modelo de Moran también hemos simulado datos de una población haploide con tamaño N = 30. Al igual que en el caso anterior, vamos a fijarnos en el estudio de un locus concreto con dos posibles alelos A y a. En este caso la cadena de Markov que vamos a representar es la siguiente:

$$X_t = \textit{número de copias del alelo A en el momento } t$$

En este caso el conjunto de estados de la cadena de Markov va a ser los número enteros desde 0 hasta el tamaño de la población, es decir, hasta 30, por lo que $S_x = \{0,...,30\}$. Tal y cómo se plantea este modelo, en cada instante tenemos solamente tres opciones, perder un alelo de tipo A, ganar un alelo de tipo A o mantenernos tal y como estábamos en el instante anterior.

Las probabilidades de transición vamos a obtenerlas de la siguiente manera:

$$P(i,j) = \begin{cases} (1-\frac{i}{N})^2 + (\frac{i}{N})^2 & si \quad j = i \\ (1-\frac{i}{N}) * (\frac{i}{N}) & si \quad j = i+1 \quad ó \quad si \quad j = i-1 \end{cases}$$

Tal y como vemos, solamente van a tener probabilidades distintas de cero los estados adyacentes al estado actual, además del estado actual, por lo que podemos considerar esta cadena de Markov como un camino aleatorio con barreras absorbentes, ya que al llegar a los estados 0 o N nos mantendremos durante el resto de las generaciones en dicho estado. Por lo que la forma que podemos esperar de la matriz de transición será como el que representamos a continuación.

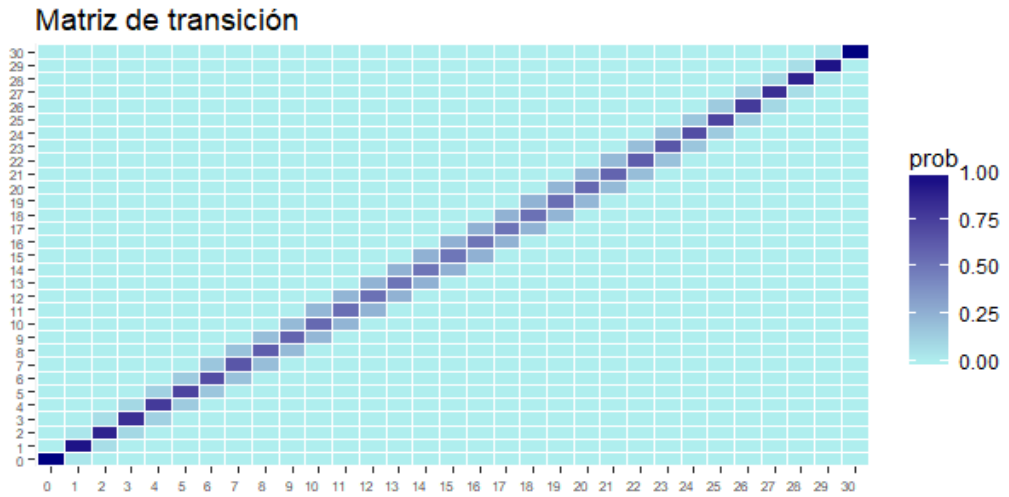

Figura 17: matriz de transición

En cualquier instante vamos a tener mayor probabilidad de quedarnos con el mismo número de alelos A que en el instante anterior, que llegar a perder o ganar un individuo del alelo A.

De nuevo, nos encontramos con una cadena de Markov finita cerrada y con estados absorbentes, y al igual que en el modelo de Fiser-Wright, no vamos a ser capaces de encontrar la distribución estacionaria debido a que podemos obtener dos conjuntos de soluciones del sistema de ecuaciones. Estas soluciones van a ser las mismas que las que se obtienen con Fisher-Wright para una población haploide, y que recordamos son las siguientes:

$$\pi_0 = 1 \quad y \quad \pi_i = 0 \quad si \quad i \in [1,60]$$
$$o$$
$$\pi_{60} = 1 \quad y \quad \pi_i = 0 \quad si \quad i \in [0,59]$$

Con el paso del tiempo, o el alelo se extinguirá o el alelo se fijará en la población.

Al igual que con el modelo de Fisher-Wright, vamos a pasar a realizar simulaciones a partir de la matriz de transición que está completa en el anexo. De nuevo, vamos a escoger distintas proporciones iniciales para el alelo A para así observar el comportamiento de la cadena de Markov.

Caso p = 0.1

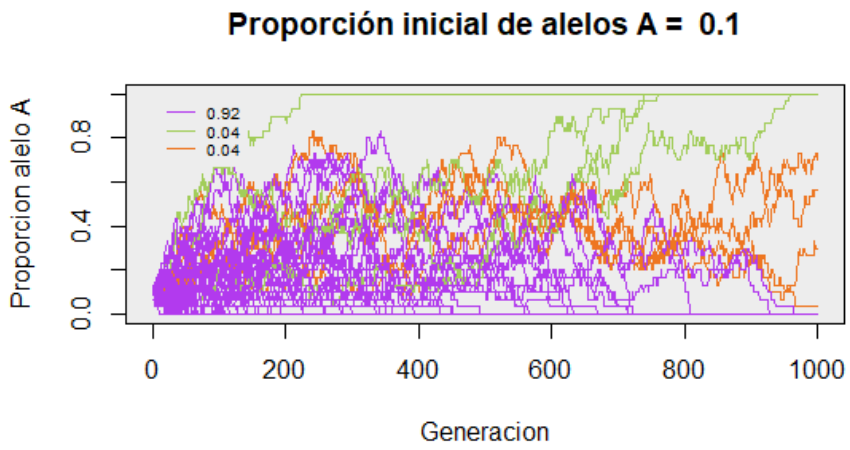

Figura 18: simulación

Como podemos observar en el gráfico, para conseguir que se llegue a los estados absorbentes de la cadena de Markov, es decir, que se llegue a que todos los alelos sean A o que ninguno de ellos lo sean, se necesita más tiempo que en con el modelo anterior. Con 1000 generaciones observamos incluso cuatro simulaciones que no llegan a un estado absorbente.

Caso p = 0.5

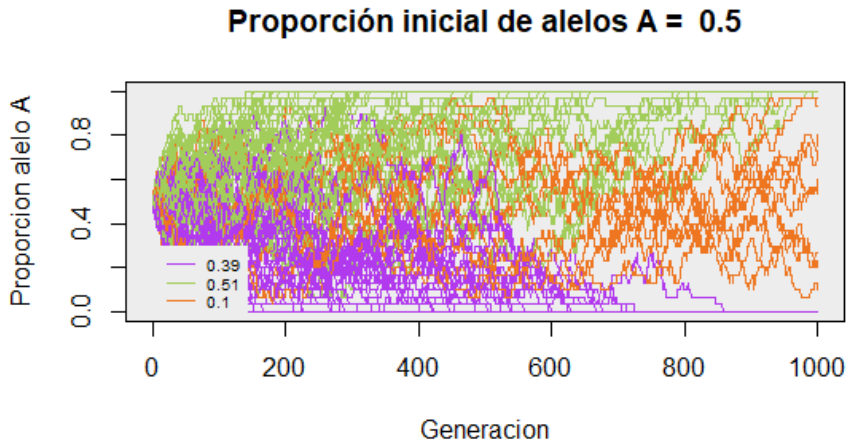

Figura 19: simulación

En este caso son muchas mas las simulaciones que no llegan a un estado absorbente. Esto tiene sentido ya que es más complicado llegar a cualquiera de los estados absorbente partiendo de un estado que es equidistante a 0 y a N, en nuestro caso, que tenemos una población de 30, habríamos empezado con 15 individuos de alelo A y otros 15 de alelo a.

Caso p = 0.9

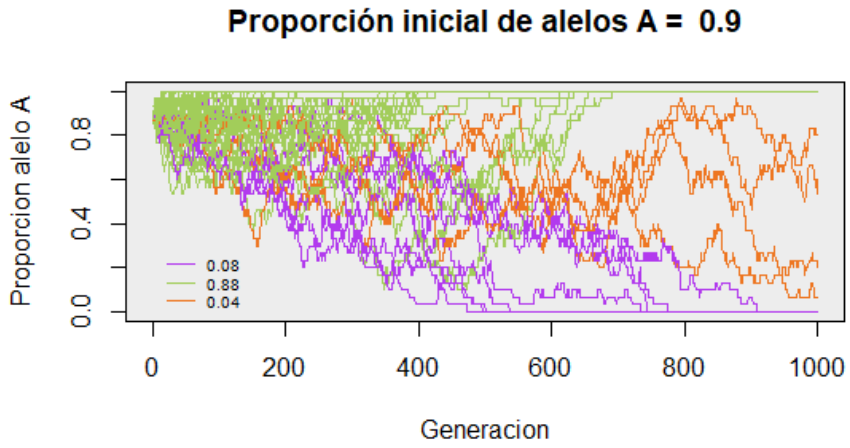

Figura 20: simulación

Nos encontramos con la situación contraria, al partir de una proporción inicial de alelo A tan cercana a 1, en este caso contamos con que hay más simulaciones que acaban con una población homogénea de individuos A que el caso contrario.

De nuevo, vamos a pasar a obtener los tiempos esperados de salida. Al igual que en el caso anterior, el tiempo de salida lo entendemos como el tiempo esperado en llegar a que todas las

células sean A o a, ya que los estados absorbentes van a ser el 0 y el N. Los tiempos medios que se obtienen para cada uno de los estados de esta cadena de Markov los vemos en la tabla.

x	g(x)	x	g(x)	x	g(x)	x	g(x)	x	g(x)
1	118,85	**7**	474,327	**13**	601,064	**19**	576,717	**25**	391,02
2	206,665	**8**	507,266	**14**	607,082	**20**	558,154	**26**	339,041
3	278,408	**9**	535,091	**15**	609,082	**21**	535,091	**27**	278,408
4	339,041	**10**	558,154	**16**	607,082	**22**	507,266	**28**	206,665
5	391,02	**11**	576,717	**17**	601,064	**23**	474,327	**29**	118,85
6	435,799	**12**	590,974	**18**	590,974	**24**	435,799		<

Tabla 4: tiempos de salida

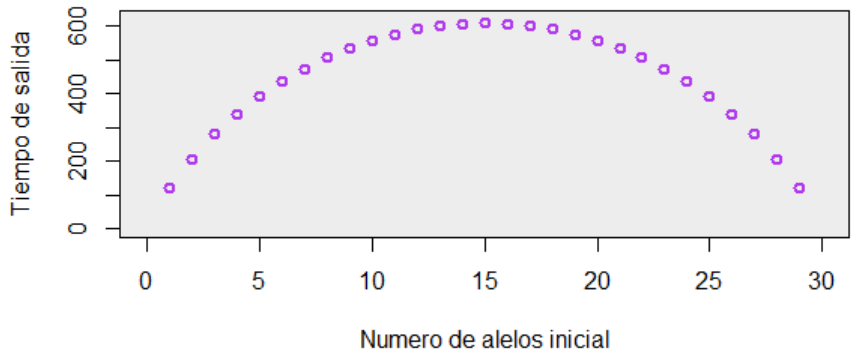

Figura 21: tiempos de salida

Podemos observar que la forma de la gráfica es igual que en con el modelo de Fisher-Wright, sin embargo, los tiempos de salida son mucho mayores en este modelo que en el anterior. Esto se debe a que con el modelo de Moran la probabilidad de quedarse en el mismo estado es mayor que la de pasar a alguno de los estados adyacentes.

4.3 Modelo de Moran con barreras reflectantes

Otra manera que encontramos de plantear el modelo de Moran es haciendo que los estados 0 y N no sean absorbentes, sino que actúen como una barrera reflectante. Esto quiere decir que cada vez que lleguemos al estado 0 o N, en el siguiente momento pasaremos con probabilidad igual a 1 al estado 1 o al estado N-1 respectivamente. De esta manera consideramos imposible la perdida de cualquiera de los alelos, ya que cuando se fije uno de ellos, en el instante siguiente, debido a las mutaciones, volverá a aparecer un alelo del tipo contrario al fijado.

Por lo que las probabilidades de transición para este caso podemos escribirlas de la siguiente forma:

$$P(i,j) = \begin{cases} 1 \;\; si \;\; i = 0 \;\; j = 1 \\ 1 \;\; si \;\; i = N \;\; j = N-1 \\ (1-\frac{i}{N})^2 + (\frac{i}{N})^2 \;\; si \;\; j = i \\ (1-\frac{i}{N}) * (\frac{i}{N}) \;\; si \;\; j = i+1 \;\; ó \;\; si \;\; j = i-1 \end{cases}$$

La matriz de transición que encontramos es muy similar a la del caso anterior

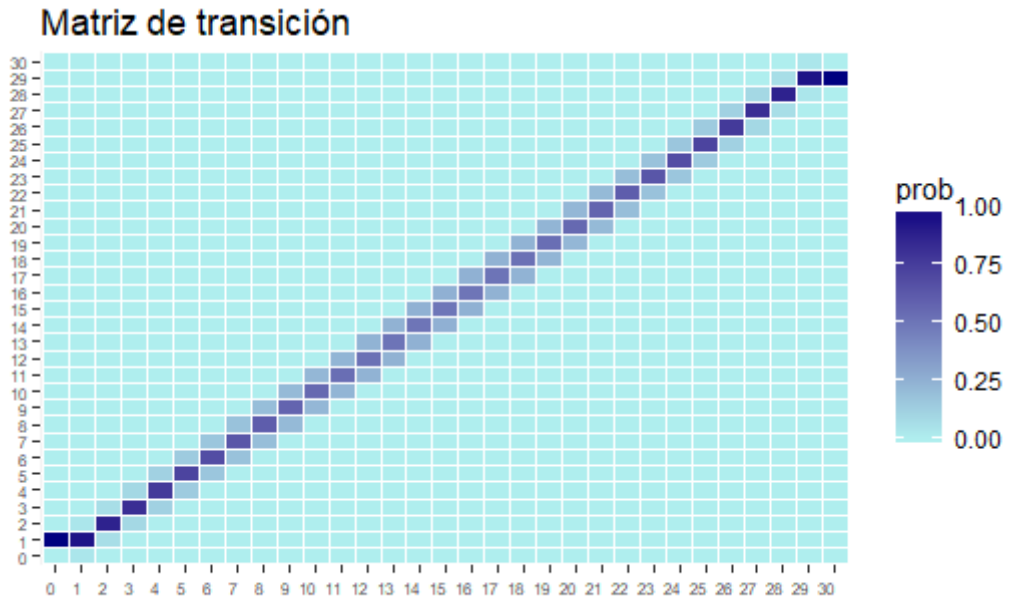

Figura 22

Observamos que del estado 0 pasamos, como hemos dicho, directamente al estado 0, aunque parezca que no se puede llegar del estado 1 al estado 0 esto no es cierto, sino que la probabilidad de esto va a ser muy pequeña.

La cadena de Markov con la que trabajamos ahora tiene todos sus estados recurrentes no absorbentes. Con este modelo si vamos a ser capaces de encontrar una distribución estacionaria única, la cual vemos representada en la siguiente tabla:

x	π(x)	x	π(x)	x	π(x)	x	π(x)	x	π(x)
0	0,00417	**7**	0,02332	**14**	0,01676	**21**	0,01986	**28**	0,06704
1	0,12947	**8**	0,02133	**15**	0,01668	**22**	0,02133	**29**	0,12947
2	0,06704	**9**	0,01986	**16**	0,01676	**23**	0,02332	**30**	0,00417
3	0,04635	**10**	0,01877	**17**	0,01698	**24**	0,02607		
4	0,0361	**11**	0,01796	**18**	0,01738	**25**	0,03003		
5	0,03003	**12**	0,01738	**19**	0,01796	**26**	0,0361		
6	0,02607	**13**	0,01698	**20**	0,01877	**27**	0,04635		

Tabla 5: distribución estacionaria

La representación de las probabilidades de la distribución estacionaria quedaría de la siguiente manera:

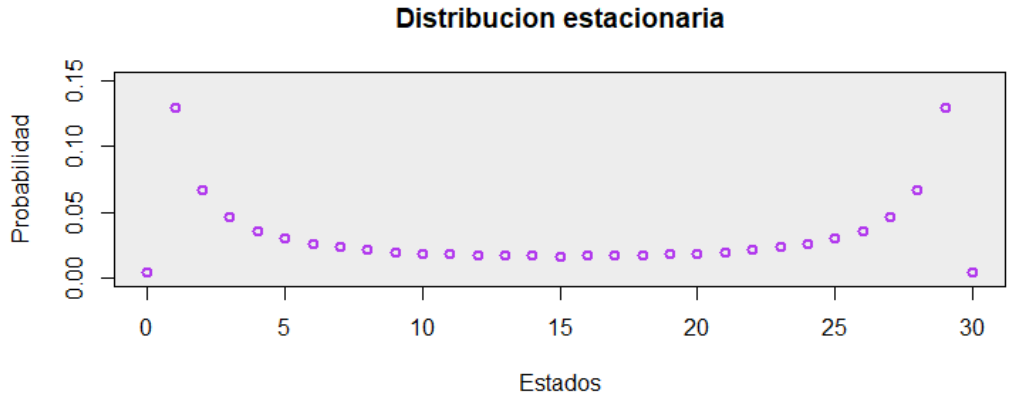

Figura 23: distribución estacionaria

Tal y como observamos, independientemente del estado en el que nos encontremos, tras el paso del tiempo, la probabilidad de estar en los estados 0 y 30 son muy cercanas a 0, sin embargo, los estados en los que sería más probable estar van a ser los estados 1 y 29.

En este caso, al hacer simulaciones, no tiene sentido diferenciar aquellas que acaban con la fijación del alelo A de aquellas que acaban con la extinción del mismo, ya que al llegar a cualquiera de los estados que representan estos sucesos se pasa a un estado con la presencia de los dos alelos. La probabilidad de encontrarnos en el estado 0 o en el estado N tras un número t de generaciones va a ser muy pequeña, las que corresponderían a las proporciones iniciales de alelo A con las que hemos trabajado el resto de modelos serían las siguientes.

	0	30
0.1	1.057874e-02	2.127841e-07
0.5	3.854498e-04	3.854498e-04
0.9	2.127841e-07	2.127841e-07

Tabla 6: probabilidades de transición.

Debido a esto, para las simulaciones vamos a diferenciar en este caso aquellas que tienen proporciones muy altas, superiores a 0,75; muy bajas, inferiores a 0,25 y medias, entre 0,26 y 0,74. En las simulaciones que vamos a ver a continuación el color morado va a representar aquellas que tengan proporciones altas, el color verde va a representar las proporciones bajas y el naranja las medias.

Caso p = 0.1

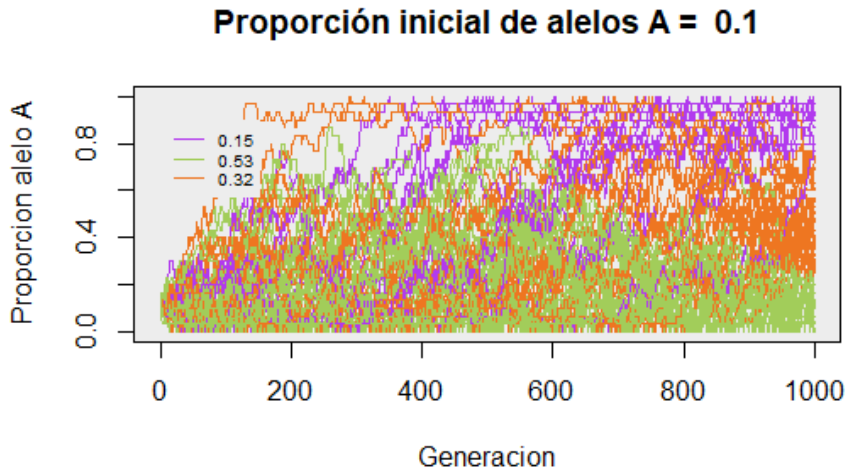

Figura 24: simulación.

Observamos que la mayor parte de las simulaciones acaban con proporciones del alelo A menores que 0.25, lo cual tiene sentido ya que empezamos con una proporción inicial de 0.1 que es bastante baja. Por otro lado encontramos mayor número de simulaciones que tras mil generaciones acaban con una proporción media de los que acaban con una proporción alta. Por lo que, a pesar de que existe la posibilidad de, llegado a la extinción del alelo, este vuelva a resurgir, es mucho más probable que en una población que inicialmente tenga un 10% de presencia del alelo A, sea el alelo a el que predomine.

Caso p = 0.5

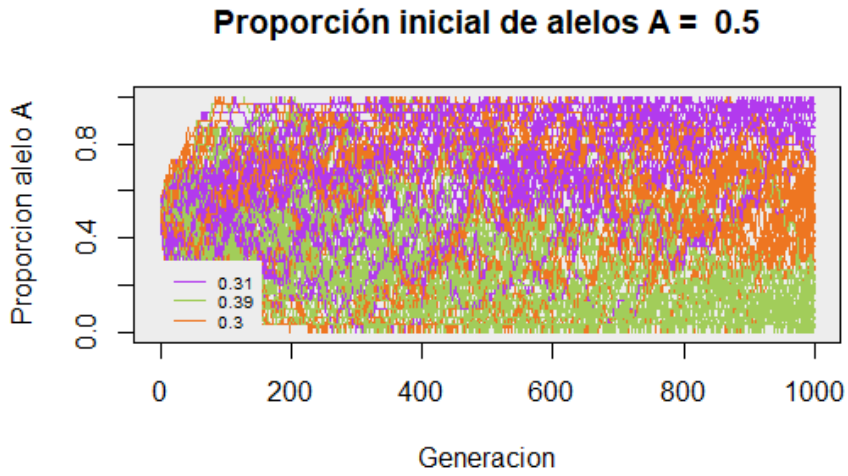

Figura 25: simulación.

Al tener una proporción inicial de 0.5 observamos que hay un cambio significativo con respecto a las simulaciones anteriores. En este caso las proporciones finales son más parecidas, siendo mayor para valores medios ya que alberga más casos posibles.

Caso p = 0.9

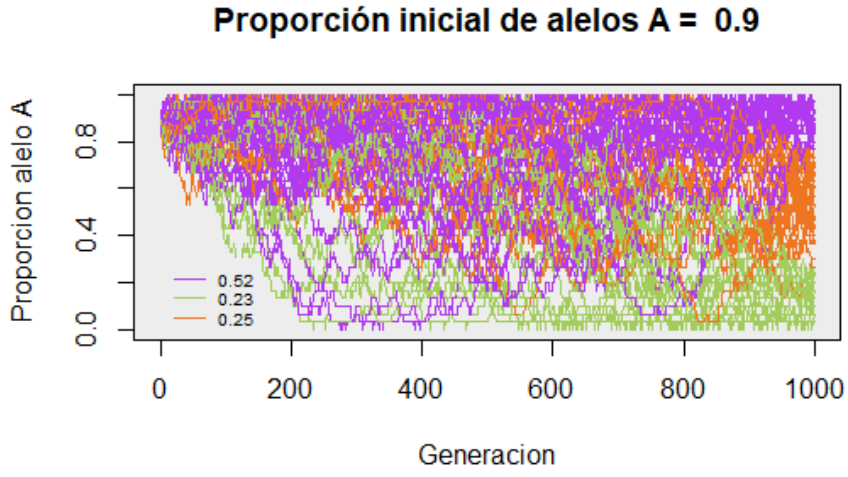

Figura 26: simulación.

En este último caso encontramos una situación similar simétrica respecto a la primera. Tenemos muchos más casos que acaban con proporciones de alelo A altas, aunque algunas de ellas llegan incluso a momentos en los que este alelo desaparece.

4.4 Comparación de los modelos

Como hemos visto, podemos comparar los dos modelos estudiados ya que en ambos estamos planteando cadenas de Markov equivalentes. La diferencia entre los modelos es la forma en la que se considera la reproducción de los individuos de la población. Así pues, vamos a empezar comparando los tiempos medios de salida para ambos casos. Para esto, consideramos poblaciones haploides de tamaño N = 30.

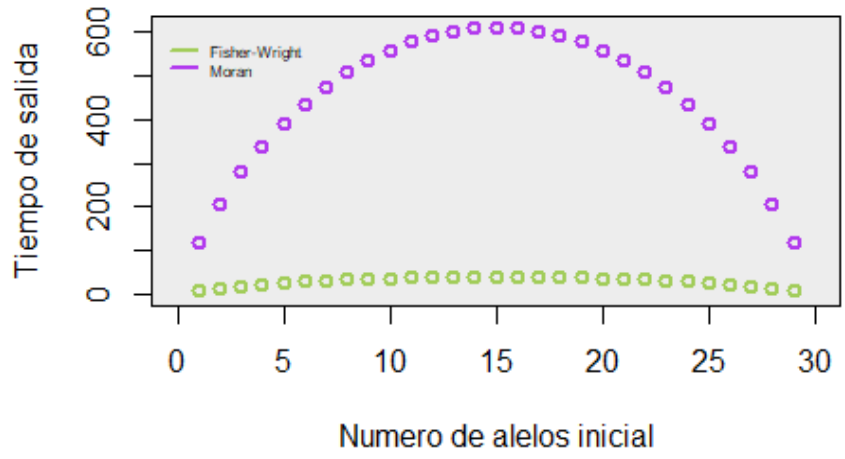

Figura 27:tiempos de salida

Tal y como podemos observar en el gráfico el tiempo que se necesita para cada uno de los modelos varía de manera muy significativa. Mientras que para el modelo de Fisher-Wright no llegamos ni siquiera a las 50 generaciones, empezando con 15 individuos con alelo de tipo A, en el modelo de Moran, para el mismo número de alelos necesitamos más de 600 generaciones. Esta diferencia tan grande es debido al número de reproducciones que se produce en cada instante, mientras que en el modelo de Fisher-Wright consideramos que en cada instante del tiempo se están produciendo 30 individuos nuevos, en el modelo de Moran, por cada instante de tiempo estamos suponiendo un único individuo nuevo, por lo que, en realidad deberíamos contar que una generación según el modelo de Fisher-Wright corresponde a 30 generaciones según el modelo de Moran.

Conclusiones

Las frecuencias alélicas que según el equilibrio de Hardy-Weinberg puede calcularse sin posibilidad de error en cualquier población, se ven significativamente alteradas por los siguientes mecanismos: mutaciones, migraciones, selección natural, endogamia y la deriva genética.

Tanto las mutaciones como la deriva genética tienen una importante carga estocástica que los modelos de Fisher-Wright y Moran han tenido en cuenta para llegar a obtener predicciones sobre las generaciones futuras.

Hemos sido capaces de modelizar poblaciones diploides y haploides con el modelo de Fisher-Wright, sin embargo las suposiciones iniciales que se realizan para poder realizar los cálculos limitan mucho las poblaciones a las que se puede aplicar el modelo. Por ejemplo, no sería lógico plantear este modelo para poblaciones humanas, ya que no estamos considerando el sexo de los distintos individuos, que hace que no sea posible que cualquier individuo se reproduzca con cualquier otro. Tampoco se da el caso, en los humanos, de que todos los individuos de la población se reproduzcan en el mismo instante temporal.

Por otro lado el modelo de Moran parece adaptarse mejor a la realidad, ya que tiene en consideración que no toda la población se reproduce a la vez, sin embargo, este modelo solamente nos permite ajustar las frecuencias alélicas en poblaciones haploides, por lo que, de nuevo estamos limitando mucho su aplicación.

Aunque ambos modelos han conseguido dar resultados que nos permiten calcular predicciones y que nos permiten tener una idea del posible comportamiento de los alelos a lo largo del tiempo, considero que dejan de lado un gran número de poblaciones y posibilidades, ya que los organismos que mejor se ajustarían a los requisitos que ambos modelos exigen son células procariotas o células eucariotas que conformen un individuo, pero en ninguno de los casos estaríamos hablando de poblaciones de individuos multicelulares como pueden ser animales o plantas.

Esto plantea la posibilidad de seguir estudiando modelos que requieran menos restricciones, permitiendo así ajustarse mejor a poblaciones mas complejas, ya que no solamente los organismos unicelulares están sujetos a los cambios alélicos producidos por los métodos ya comentados. Un estudio en humanos nos permitiría comprender como han ido desapareciendo algunas enfermedades ligadas a genes recesivos y llegar a estudiar la probabilidad de que enfermedades actuales acaben desapareciendo.

Bibliografía

- Pierce, B. (2007) *Genética, un enfoque conceptual* Ed. Médica Panamericana.
- http://bioinformatica.uab.es/genetica/curso/Historia.html
- Durrentt, R. (2012) *Essentials of stochastic processes* Springer.
- http://uvigen.fcien.edu.uy/utem/Popgen/popintro.html
- Sanchez-Monge, E., Jouve, N. (1984) *Genética* Omega
- https://glosarios.servidor-alicante.com/genetica
- http://yazmin97.blogspot.com.es/2011/10/deriva-genica.html
- Otto, S. P, Day T,(2007) *A bioloist's Guide to Mathematical Modeling in Ecology and Evolution*, Princeston University Press.
- Etheridge, A.(2012) *Some Mathematical Model from Population Genetics*, Springer-Verlag
- R Core Team. *R: A lenguage and Enviroment for Statistical Computing.* R Foundation for Statistical Computing

• Lista de figuras

Lista de figuras

• Lista de tablas

Anexo I: Glosario

- **Alelo:** cada una de las variantes de un locus. Cada alelo aporta diferentes variaciones al carácter que afecta. En organismos diploides (2n) los alelos de un mismo locus se ubican físicamente en los pares de cromosomas homólogos.
- **Autosómico:** Se refiere a cualquiera de los cromosomas que no son los determinantes del sexo (es decir, X e Y) o a los genes de esos cromosomas.
- **Diploide:** que tiene doble juego de cromosomas (2n). Características de las células somáticas.
- **Eucariota:** organismo uni o multicelular cuyas células poseen un núcleo limitado por una membrana nuclear, se dividen por mitosis y pueden entrar en meiosis.
- **Fenotipo:** rasgos físicos que reflejan las características normales o anormales de los genes (genotipo).
- **Genotipo:** par de alelos (normal o mutado) que posee un individuo para un gen específico.
- **Haploide:** célula con una sola dotación cromosómica, u organismo compuesto de tales células.
- **Heterocigoto:** célula o individuo diploide con alelos diferentes en uno o más loci de cromosomas homólogos.
- **Homocigosis:** unión de gametos con idénticas características que lógicamente producen sujetos de raza pura (homocigotos)
- **Homocigoto:** estado en el que se portan un par de alelos idénticos en un locus.
- **Locus:** Ubicación del gen en un cromosoma. Para el locus puede haber varios alelos posibles. Su plural es loci.
- **Panmíctico**: todos los individuos tienen la misma probabilidad de aparearse y el apareamiento es al azar.
- **Procariota:** perteneciente al super-reino Procariotas, que incluye a los mircoorganismos que se multiplican por división binaria y carecen de núcleo delimitado por envoltura nuclear.
- **Prolicifidad:** cualidad del que se reproduce o es capaz de reproducirse en abundancia.
- **Recesivo:** Carácter genético hereditario latente, que no se manifiesta externamente en la descendencia si no es transmitido por los dos reproductores a la vez.

Anexo 2: Código de R

```
library(markovchain)

library(ggplot2)

set.seed(100)

#Simulaciones de trayectorias

simulacion<- function(cm, p0, nsim, ngen){

        e0 <- as.character(p0*(dim(cm)-1))

        D <- array(NA, dim = c(ngen, nsim))

        for(iin 1:nsim){

                D[,i]<- as.numeric(rmarkovchain(ngen, cm, t0 = e0))

        }

        D <- D/(dim(cm)-1)

        plot(NA, xlim = c(0, ngen),ylim = c(0,1),main = paste("Proporción inicialde alelosA = ",p0),ylab =
"       Proporcion aleloA", xlab = "Generacion")

        u <- par("usr")

        rect(u[1],u[3],u[2],u[4], col = "grey93")

        for(iin 1:nsim){

                if(D[ngen,i]==0){

                        points(1:ngen, D[,i],type = "l", col = "darkorchid2")

                }else if(D[ngen,i]== 1){

                        points(1:ngen, D[,i],type = "l", col = "darkolivegreen3")

                }else{

                        points(1:ngen, D[,i],type = "l", col = "chocolate2")

                }

        }

        P1 <- round(sum(D[ngen,]==1)/nsim,3)

        P0 <- round(sum(D[ngen,]==0)/nsim,3)
```

```
        Pa <- round(1-P1-P0,3)
        return(c(P0,P1, Pa)
)
#Simulación con estado no absorbente
sim.noabs<- function(cm, p0, nsim, nge){
        e0 <- as.character(p0*(dim(cm)-1))
        D <- array(NA,dim = c(ngen,nsim))
        for(iin 1:nsim){
                D[,i]<- as.numeric(rmarkovchain(ngen,cm, t0 = e0))
        }
        D <- D/(dim(cm)-1)
        plot(NA, xlim = c(0,ngen),ylim = c(0,1),main = paste("Proporción inicialde alelosA = ",p0),ylab = "
        Proporcion aleloA", xlab = "Generacion")
        u <- par("usr")
        rect(u[1],u[3],u[2],u[4], col = "grey93")
        for(iin 1:nsim){
                if(D[ngen,i]>= 0.75){
                        points(1:ngen, D[,i],type = "l", col = "darkorchid2")
                }else if(D[ngen,i]<= 0.25){
                        points(1:ngen, D[,i],type = "l", col = "darkolivegreen3")
                }else{
                        points(1:ngen, D[,i],type = "l", col = "chocolate2")
                }
        }
        P1 <- round(sum(D[ngen,]>=0.75)/nsim,3)
        P0 <- round(sum(D[ngen,]<=0.25)/nsim,3)
        P <- 1-P1-P0
        return(c(P1,P0, P))
```

```
}

#Tiempos de salida

}tiempo.salida<- function(X){

        d <- dim(X)[1]

        R <- X[-c(1,d),-c(1,d)]

        aux <- diag(nrow = dim(R)[1]) - R

        g <- solve(aux)

        tmed <- round(g%*%rep(1,dim(R)[1]),3)

        return(tmed)

}
```

Modelo de Fisher-Wright

```
#Modelo Fisher-Wright

mTransFW <- function(n){

        X <- array(NA,dim = rep(n,2))

        for(iin 1:n){

                p <- (i-1)/(n-1)

                for(jin 1:dim(X)[2]){

                        X[i,j]<- dbinom((j-1),n-1,p)

                }

        }

        return(X)

}

sim.genotipos<- function(cm,p0,gen){

        e0 <- as.character(p0*(dim(cm) - 1))

        A <- as.numeric(rmarkovchain(gen,cmFWD, t0 = e0))

        A <- A/(2*N)

        pp <- A*A

        qq <- (1 - A)*(1 - A)
```

Anexo 2: Código de R

```
        pq <- 2*A*(1 - A)

        plot(NA, xlim = c(0, nsim), ylim = c(0,1), xlab = "Generacion", ylab = "Proporcion genotipos", main =
        paste("Proporcion inicialde A = ", p0))

        u <- par("usr")

        rect(u[1],u[3],u[2],u[4], col = "grey93")

        points(1:gen,pp, type = "l", col = "darkorchid2", lwd = 2)

        points(1:gen,qq, type = "l", col = "darkolivegreen3", lwd = 2)

        points(1:gen,pq, type = "l", col = "chocolate2", lwd = 2)

}

#Calculo de la matriz de transición

N <- 30

########## Diploide#######

Pdiploide<- mTransFW(2*N+1)

#Trabajando con cadenas de markov

cmFWD <- new("markovchain", states= as.character(0: (2*N)),

transitionMatrix= Pdiploide,name = "FisherWright")

#Representación de la matriz de transición en heatmap

datos <- as(cmFWD, "data.frame")

a <-ggplot(datos,aes(t0,t1))+geom_tile(aes(fill= prob),color="white")+

scale_fill_gradient(low = "paleturquoise", high = "navy")+

labs(x = "", y= "")+

ggtitle("Matrizde transición")+

theme(axis.text.x= element_text(hjust=1,vjust=0.5,size= 6),

axis.text.y= element_text(size= 6))+

theme(panel.background = element_rect(fill= 'grey95'))

plot(a)

#Simulaciones de trayectorias

nsim <- 100
```

```
ngen <- 300

#Distribución estacionaria

de.FWD<- steadyStates(cmFWD)

#Proporción inicial de alelo A 0.1

P10D <- simulacion(cmFWD,0.1,nsim,ngen)

legend(250,0.5, legend = c(P10D),col= c("darkorchid2","darkolivegreen3", "chocolate2"),lty= 1, cex = 0.55,box.lty= 0,bg ="grey93")

#Proporción inicial de 0.5

P50D <- simulacion(cmFWD, 0.5, nsim, ngen)

legend(240,0.45, legend = c(P50D),col= c("darkorchid2","darkolivegreen3", "chocolate2"),lty= 1, cex = 0.55, box.lty= 0, bg ="grey93")

#Proporción inicial de 0.9

P90D <- simulacion(cmFWD,0.9,nsim,ngen)

legend(250, 0.35, legend = c(P90D),col= c("darkorchid2", "darkolivegreen3", "chocolate2"),lty= 1, cex = 0.55, box.lty= 0, bg = "grey93")

#Simulación de genotipos

#Proporción inicial de 0.1

sim.genotipos(cmFWD, 0.1, ngen)

legend(20, 0.95, legend = c("AA", "aa", "Aa"),col= c("darkorchid2", "darkolivegreen3", "chocolate2"),lty= 1, cex = 0.55, box.lty= 0, bg = "grey93")

#Proporción inicial de 0.5

sim.genotipos(cmFWD, 0.5,ngen)

legend(20, 0.95, legend = c("AA", "aa", "Aa"),col= c("darkorchid2", "darkolivegreen3", "chocolate2"),lty= 1, cex = 0.55 ,box.lty= 0, bg = "grey93")

#Proporción inicial de 0.9

sim.genotipos(cmFWD,0.9,ngen)

legend(40, 0.95, legend = c("AA", "aa", "Aa"),col= c("darkorchid2", "darkolivegreen3", "chocolate2"),lty= 1, cex = 0.55, box.lty= 0, bg ="grey93")

#Obtención de los tiempos de salida

tmedD <- tiempo.salida(cmFWD)

plot(NA, ylim = c(0,82), xlim = c(0,60), xlab = "Numero de alelos inicial", ylab = "Tiempo de salida")
```

Anexo 2: Código de R

```
u <- par("usr")

rect(u[1],u[3],u[2],u[4],col = "grey93")

points(1:59, tmedD,col = "darkorchid2", lwd = 2)

########## Haploide#######

Phaploide<- mTransFW(N + 1)

cmFWH <- new("markovchain", states= as.character(0:N),

transitionMatrix= Phaploide,name = "FisherWright")

#Distribución estacionaria

de.FWH<- steadyStates(cmFWH)

#Simulaciones de trayectorias

#Proporción inicial de alelo A 0.1

P10H <- simulacion(cmFWH, 0.1, nsim, ngen)

legend(250, 0.5, legend = c(P10H),col = c("darkorchid2", "darkolivegreen3", "chocolate2"), lty= 1, cex = 0.55,
box.lty= 0, bg = "grey93")

#Proporción inicial de 0.5

P50H <- simulacion(cmFWH, 0.5, nsim,ngen)

legend(250, 0.5, legend = c(P50H),col = c("darkorchid2", "darkolivegreen3", "chocolate2"), lty= 1, cex = 0.55,
box.lty= 0, bg = "grey93")

#Proporción inicial de 0.9

P90H <- simulacion(cmFWH,0.9,nsim,ngen)

legend(250,0.5, legend = c(P90H),col = c("darkorchid2","darkolivegreen3", "chocolate2"), lty= 1, cex =
0.55,box.lty= 0,bg = "grey93")

#Obtención de los tiempos de salida

tmedH <- tiempo.salida(cmFWH)

plot(NA, xlim = c(0,N),ylim = c(0,45), xlab = "Numero de alelosinicial", ylab = "Tiempo de salida")

u <- par("usr")

rect(u[1],u[3],u[2],u[4], col = "grey93")

points(1:29, tmedH, col = "darkorchid2", lwd = 2)
```

Modelo de Moran

```
#Modelo de moran

mTransMoran <- function(n){

	X <- array(0,dim = rep(n,2))

	for(iin 1:dim(X)[1]){

		a <- i- 1

		for(jin 1:dim(X)[2]){

			if(j== i){

				X[i,j]<- (1 - a/N)^2 + (a/N)^2

			}else if(j== i-1 || j== i+1){

				X[i,j]<- (1 - a/N)*(a/N)

			}

		}

	}

	return(X)

}

#Sin estados absorbentes

mTransMoranR <- function(n){

	X <- array(0,dim = rep(n,2))

	X[1,2] <- 1

	X[n+1,n]<- 1

	for(iin 2:n){

		a <- i-1

		for(jin 1:dim(X)[2]){

			if(j== i){

				X[i,j]<- (1 - a/n)^2 + (a/n)^2

			}else if(j== i-1 || j== i+1){

				X[i,j]<- (1 - a/n)*(a/n)
```

```
                    }
                }
            }
        return(X)
}
N <- 30
P <- mTransMoran(N + 1)
cmM <- new("markovchain", states= as.character(0:N), transitionMatrix= P, name = "Moran")
#Matriz de transición en heatmap
datos <- as(cmM, "data.frame")
a <-ggplot(datos,aes(t0,t1))+geom_tile(aes(fill= prob),color= "white")+
scale_fill_gradient(low = "paleturquoise", high = "navy")+
labs(x = "", y= "")+
ggtitle("Matrizde transición")+
theme(axis.text.x= element_text(hjust=1,vjust=0.5,size= 6),
axis.text.y= element_text(size= 6))+
theme(panel.background = element_rect(fill= 'grey95'))
plot(a)
#Distribución estacionaria
de.M- steadyStates(cmM)
#Simulación de trayectorias con cadenas markov
nsim <- 100
ngen <- 1000
#Empezamos con un 10% de alelos de tipo A
P10 <- simulacion(cmM, 0.1, nsim, ngen)
legend(0, 1, legend = round(P10,3), col = c("darkorchid2", "darkolivegreen3", "chocolate2"), lty= 1, cex = 0.6,
bg = "grey93", box.lty= 0)
#Para un 50% de alelos de tipo A
```

```
P50 <- simulacion(cmM, 0.5, nsim, ngen)

legend(0, 0.3, legend = P50, col = c("darkorchid2", "darkolivegreen3", "chocolate2"), lty= 1, cex = 0.6, bg =
"grey93", box.lty= 0)

#Para un 90% de alelos de tipo A

P90 <- simulacion(cmM, 0.9, nsim, ngen)

legend(0, 0.3, legend = P90, col = c("darkorchid2", "darkolivegreen3", "chocolate2"), lty= 1,cex = 0.6, bg =
"grey93", box.lty= 0)

#Tiempo medio de salida

tmed <- tiempo.salida(cmM)

plot(NA, xlim = c(0,N),ylim = c(0,620), xlab = "Numero de alelosinicial", ylab = "Tiempo de salida")

u <- par("usr")

rect(u[1],u[3],u[2],u[4],col = "grey93")

points(1:29,tmed, col = "darkorchid2", lwd = 2)

####### Barrera reflectante######

PR <- mTransMoranR(N+1)

cmMR <- new("markovchain", states= as.character(0:N), transitionMatrix= PR, name = "Moran R")

datos <- as(cmMR, "data.frame")

a <-ggplot(datos,aes(t0,t1))+geom_tile(aes(fill= prob),color= "white")+

scale_fill_gradient(low = "paleturquoise", high = "navy")+

labs(x = "", y= "")+

ggtitle("Matrizde transición")+

theme(axis.text.x= element_text(hjust=1,vjust=0.5,size= 6),

        axis.text.y= element_text(size= 6))+

theme(panel.background = element_rect(fill= 'grey95'))

plot(a)

#Distribución estacionaria

de.MR<- steadyStates(cmMR)

plot(NA, xlim = c(0,N),ylim = c(0,0.15), xlab = "Estados",ylab =

"Probabilidad", main = "Distribucion estacionaria")
```

Anexo 2: Código de R

```
u <- par("usr")

rect(u[1],u[3],u[2],u[4],col = "grey93")

points(0:N,de.MR,lwd = 2, col = "darkorchid2")

PR.nsim <- cmMR^nsim

Pprob0 <- c(PR.nsim[4,1],PR.nsim[16,1],PR.nsim[28,1])

probN <- c(PR.nsim[4,31],PR.nsim[16,31],PR.nsim[28,1])

proba <- cbind(prob0,probN)

rownames(proba) <- c("0.1","0.5","0.9")

#Simulaciones

P10R <- sim.noabs(cmMR, 0.1, nsim, ngen)

legend(-0.5, 0.9, legend = P10R, col = c("darkorchid2", "darkolivegreen3", "chocolate2"), lty= 1, cex = 0.6, bg
= "grey93", box.lty= 0)

P50R <- sim.noabs(cmMR,0.5,nsim,ngen)

legend(0, 0.3, legend = P50R, col = c("darkorchid2", "darkolivegreen3", "chocolate2"), lty= 1, cex = 0.6, bg =
"grey93", box.lty= 0)

P90R <- sim.noabs(cmMR,0.9,nsim,ngen)

legend(0, 0.3, legend = P90R, col = c("darkorchid2", "darkolivegreen3", "chocolate2"), lty= 1, cex = 0.6, bg =
"grey93", box.lty= 0)
```

Comparación de modelos

```
#Comparación de los modelos en poblaciones haploides

plot(NA, xlim = c(0,30), ylim = c(0,610), xlab = "Numero de alelosinicial", ylab = "Tiempo de salida")

u <- par("usr")

rect(u[1],u[3],u[2],u[4],col = "grey93")

points(1:29, tmed, col = "darkorchid2", lwd = 2)

points(1:29, tmedH, col = "darkolivegreen3", lwd = 2)

legend(-1, 600, legend = c("Fisher-Wright","Moran"), lwd = 2, col =c("darkolivegreen3", "darkorchid2"), lty= 1,
cex = 0.5, bg = "grey93", box.lty= 0)
```