FACULTADE DE MATEMÁTICAS

**Traballo Fin de Grao**

# Fundamentos matemáticos de la coalescencia en Biología

Mar Vázquez Rabuñal

2020/2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

**Traballo Fin de Grao**

# Fundamentos matemáticos de la coalescencia en Biología

Mar Vázquez Rabuñal

Julio, 2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

# Trabajo propuesto

| |
|---|
| **Área de Coñecemento:** Estadística e investigación operativa |
| **Título:** Fundamentos matemáticos de la coalescencia en Biología |
| **Breve descrición do contido** |
| La teoría de la coalescencia estudia cómo las variantes genéticas de una población pueden haberse originado a partir de un ancestro común. En el caso más simple, se supone que no hay recombinación, ni selección natural, ni estructura de la población, lo que significa que es igualmente probable que cada variante haya pasado de una generación a la siguiente. El objetivo de este trabajo es introducir al alumno en el estudio de las herramientas matemáticas de la teoría de la coalescencia, que son una colección de modelos estocásticos utilizados para generar predicciones sobre patrones de variación genética, y saber hacer inferencias a partir de muestras de datos genéticos. |
| **Recomendacións** |
| |
| **Outras observacións** |
| |

# Índice general

## Resumen

La teoría de la coalescencia es una disciplina dentro de la genética de poblaciones que estudia los ancestros, y sus relaciones, de una muestra de secuencias de material genético. La coalescencia se basa en la teoría de la probabilidad y realiza sus razonamientos a partir de un modelo de población determinado. Cuanto más complejo sea este modelo, más sofisticado será el procedimiento matemático que permite describir el proceso de coalescencia. En este trabajo comenzaremos con el caso más simple en el que se tiene un modelo con tamaño de población constante, sin estructura social ni geográfica y sin recombinación, a partir del cual asentaremos las bases de esta teoría para después ir completando el modelo permitiendo, por ejemplo, la posibilidad de mutación de los genes. Teniendo esto en cuenta podremos aplicar la teoría de la coalescencia a casos de datos reales y conocer características interesantes de una muestra determinada como el tiempo en el que se tiene su ancestro común más reciente o el tiempo en el que hubo un determinado número de linajes.

## Abstract

Coalescent theory is a discipline within the field of population genetics that studies the ancestors of a sample of genetic material and their relationships. Coalescence is based on probability theory and is studied under a specific population model. The more complex this model, the more sophisticated is the mathematical procedure which makes it possible to describe the coalescence process. In this project we will start with the simplest case, in which there is a model with constant population size, without social or geographical structure and without recombination. From this case, we will establish the main ideas of this theory. Subsequently, we will complete the model enabling, for example, the possibility of gene mutation. Taking this into consideration, it will be possible to apply coalescent theory to real information and to study interesting characteristics of a specific sample such as the time until the most recent common ancestor or the time when there were a particular number of lineages.

# Introducción

La genética de poblaciones es una disciplina dentro de la biología que estudia la evolución de las especies tratando de entender qué es lo que produce y mantiene la variación genética. Fue introducida entre 1920 y 1930 por R. Fisher, J.B.S. Haldane y S. Wright, tres figuras que presentaron las bases que aún siguen guiando este campo de estudio a día de hoy.

En la actualidad, la genética de poblaciones está en pleno auge por la necesidad de herramientas para tratar las grandes cantidades de datos relacionados con el material genético que se han obtenido en los últimos años. Dentro de estos datos destaca la información derivada del *Proyecto Genoma Humano*, que en el año 2003 consiguió determinar la secuencia casi completa del ADN del ser humano.

Dentro de la genética de poblaciones nace alrededor del año 1980, de la mano del matemático inglés John Kingman, la denominada *teoría de la coalescencia*. Esta teoría se basa en la descripción de los ancestros, y de la relación entre ellos, de unos determinados genes. Este estudio se realiza siempre bajo un modelo poblacional específico y permite realizar predicciones sobre los patrones de variación genética de una población.

Una idea clave en la teoría de la coalescencia es que el análisis se realiza desde el presente hacia el pasado, es decir, se parte de una población de genes en el presente y se van estudiando sus ancestros según nos movemos hacia atrás en el tiempo. Una herramienta muy útil para visualizar mejor esta idea es la de pensar la relación entre genes como un árbol genealógico análogo al que podríamos emplear para describir el parentesco de nuestras familias. En la parte baja del árbol estarían los genes que se consideran en el presente y según vamos subiendo por las ramas vamos encontrando a los ancestros. Las ramas de este árbol no serían más que los linajes de los genes considerados.

Cada vez que dos genes encuentran su ancestro común se dice que tiene lugar un evento de coalescencia. Pensando en la idea del árbol, el evento de coalescencia se da cuando dos ramas del árbol se unen para dar lugar a una única rama.

Cuando en lugar de dos genes tenemos más, digamos que tenemos $n$ genes, nos puede resultar interesante conocer el ancestro común de toda la muestra y algunas propiedades

como el tiempo que se tarda en conseguir este ancestro común o cuánto tiempo de la historia de la muestra hubo exactamente un número determinado de linajes. Este proceso fue generalizado por Kingman en sus artículos y se le atribuyó el nombre de *n-coalescencia*.

En este trabajo trataremos de exponer los fundamentos matemáticos básicos que se encuentran debajo de la teoría de la coalescencia. Para ello, introduciremos la coalescencia de manera intuitiva a partir de un modelo poblacional simplificado como es el modelo de Wright-Fisher y después, basándonos en la teoría de la probabilidad, describiremos matemáticamente el proceso que tiene lugar.

Una vez que estas bases estén claras podremos ir un poco más allá y estudiar la coalescencia en un modelo poblacional un poco más sofisticado, en el que fenómenos como las mutaciones tienen lugar.

Con todo esto ya podremos aplicar la teoría de la coalescencia a algún ejemplo sencillo de la vida real. Debemos tener en cuenta que según las hipótesis que consideremos en nuestro modelo poblacional, el análisis matemático que hay detrás será más o menos complicado. Es por ello por lo que, dado el alcance de esta memoria, con las herramientas que vamos a introducir, solo podremos considerar poblaciones ideales que en muchos casos no se ajustarán del todo a los datos reales. Pese a todo, este es el primer paso que hay que seguir para poder después adentrarse en situaciones más complejas de la genética poblacional.

# Capítulo 1

# Conceptos preliminares

Antes de comenzar directamente con el estudio de la coalescencia, en este capítulo inicial se van a presentar una serie de conceptos tanto biológicos como matemáticos que facilitarán la comprensión de lo que se tratará más adelante.

## 1.1. Conceptos biológicos

Para comprender el estudio de la genética poblacional es imprescindible conocer conceptos básicos de la genética. Para empezar, es esencial saber qué es exactamente el material genético. De forma general podría definirse como el conjunto de estructuras que almacena la información hereditaria de un ser vivo. Está formado por un tipo especial de moléculas denominadas *ácidos nucleicos*. Existen dos tipos de ácidos nucleicos: el ácido desoxirribonucleico (ADN) y el ácido ribonucleico (ARN). En los organismos vivos conocidos, el material genético está formado por el ADN, aunque hay algunos virus en los que el ARN es quien cumple la función de llevar la información genética.

Las unidades que constituyen los ácidos nucleicos reciben el nombre de *nucleótidos*. Los nucleótidos presentan tres partes bien diferenciadas que se pueden ver en la Figura 1.1: un grupo fosfato, una pentosa (un azúcar con 5 átomos de carbono) y una base nitrogenada. Existen 5 tipos de bases nitrogenadas: la adenina (A), la guanina (G), la citosina (C), la timina (T) y el uracilo (U). Las tres primeras están presentes tanto en el ADN como en el ARN, mientras que la timina es exclusiva del ADN y el uracilo del ARN. Estableceremos cómo es una cadena de ADN o de ARN simplemente conociendo las bases nitrogenadas de los nucleótidos que la conforman.

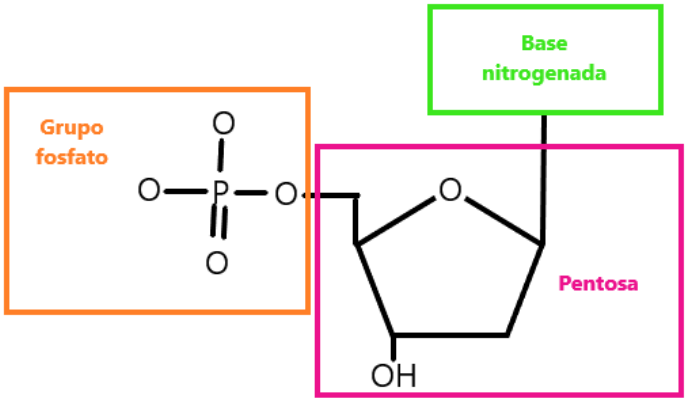



Figura 1.1: Componentes principales de un nucleótido.

La mayor parte del ADN presente en los seres vivos se encuentra compactificado en unas estructuras con forma de X denominadas cromosomas. Según el número de cromosomas que presente un individuo se dirá que es *haploide* o *diploide*. Los individuos haploides son aquellos que presentan un único juego de cromosomas y que para reproducirse se duplican y dividen. Son haploides, por ejemplo, algunas bacterias. Los individuos diploides son aquellos que presentan dos juegos de cromosomas, uno procedente del padre y otro de la madre. El ser humano es diploide y presenta 23 pares de cromosomas, un par de ellos son los denominados cromosomas sexuales X e Y (X hace referencia al sexo femenino e Y al masculino). Si un ser humano presenta estos dos cromosomas como XX será una mujer y si presenta XY será un hombre.

Ahora que ya se conoce cómo es la estructura del material genético a nivel orgánico, es importante centrarse en los conceptos con los que trata la genética de poblaciones. El primero de ellos, y el más importante, es el concepto de *gen*. Una definición estricta establece que un gen es la unidad funcional de la herencia que controla cada carácter de los seres vivos. A nivel estructural un gen no es más que una secuencia o segmento de ADN que codifica una determinada información. Al conjunto de todos los genes de una especie se le denomina *genoma*.

Cada gen que lleva la información de un determinado carácter puede manifestarse de varias formas. A cada una de estas formas se les conoce como *alelos*. Un ejemplo clásico para explicar los alelos es el del color de los guisantes. El gen que lleva la información del color presenta dos alelos: uno que determina el color verde y otro el color amarillo. Según cuál de los alelos presente el gen, el guisante será verde o será amarillo.

En ocasiones las secuencias de ADN experimentan *mutaciones*, que no son más que cambios en la secuencia de nucleótidos que lo conforman. Este tipo de modificaciones

pueden tener efectos muy diversos: desde que sean imperceptibles, hasta que supongan una ventaja o una desventaja para el organismo que la presenta.

Dentro de la genética poblacional es muy común el estudio de la historia de una determinada población, la llamada *genealogía*, que no es más que el análisis de las distintas generaciones de una población. Para facilitar este estudio se suele emplear la estructura de árbol genealógico que conecta los parentescos de los individuos, genes o secuencias de ADN que se estén considerando. Una palabra muy común dentro de este contexto es la de *linaje*, que hace referencia a la línea de antepasados de un determinado individuo.

En la teoría de la coalescencia nos interesa conocer el ancestro común más reciente de una determinada muestra de genes. Dibujando el árbol genealógico sobre los genes de estudio, podremos visualizar con facilidad dónde se encuentra este ancestro común, aunque después nos será necesario establecer un modelo matemático para poder analizar todo el proceso más a fondo.

Teniendo en cuenta todos estos conceptos propios de la estructura biológica del material genético, así como las ideas básicas de la genética poblacional, contamos con los conocimientos necesarios, a nivel biológico, para adentrarnos en la teoría de la coalescencia. Pese a todo, vamos a hacer una pequeña presentación de conceptos matemáticos que nos serán de utilidad para modelizar esta teoría y entenderla. Para ello nos basaremos principalmente en el libro de Sheldon Ross: *A First Course in Probability* ([1]).

## 1.2. Conceptos matemáticos

Consideremos un *experimento aleatorio*, es decir, un experimento del que no se puede saber con certeza cuál va a ser su resultado. Al conjunto de todos los posibles resultados de un experimento se le conoce como *espacio muestral* del experimento y se denota por $\Omega$. A cualquier subconjunto del espacio muestral se le denomina *evento* o *suceso* $(E)$, es decir, un evento es un conjunto de posibles resultados de un experimento.

Podemos definir la *probabilidad de un evento* $E$ del espacio muestral $\Omega$, $P(E)$, de forma que satisface los tres axiomas de Kolmogorov.

**Axioma 1:** $0 \leq P(E) \leq 1$

**Axioma 2:** $P(\Omega) = 1$

**Axioma 3:** Para cualquier secuencia de eventos mutuamente excluyentes $E_1, E_2, \cdots$ (es decir, eventos para los cuales $E_i \cap E_j = \emptyset$ cuando $i \neq j$), se tiene:

$$P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i)$$

En muchas ocasiones, a la hora de llevar a cabo un experimento, no nos interesan los

propios resultados del mismo, sino cantidades definidas a partir de ellos. Estas cantidades, que formalmente son funciones reales sobre el espacio muestral, se denominan *variables aleatorias*. Una variable aleatoria es una función $X$ de la forma: $X : \Omega \longrightarrow \mathbb{R}$.

Como el valor de una variable aleatoria viene determinado por el resultado del experimento, podemos asignar probabilidades a los posibles valores de la variable aleatoria.

Para una variable aleatoria $X$ se define su *función de distribución* $F$ de la siguiente manera:

$$F(x) = P(X \leq x), \qquad \text{con } -\infty < x < \infty \tag{1.1}$$

Por lo tanto, la función de distribución es simplemente aquella función que determina para todos los valores reales de $x$, la probabilidad de que la variable aleatoria sea menor o igual que $x$.

Una variable aleatoria que presenta un número finito o infinito numerable de valores posibles se dice que es *discreta*. Dada una variable discreta $X$, se define la función de masa de probabilidad $p(x)$ de $X$ como:

$$p(x) = P(X = x) \tag{1.2}$$

Por otro lado, diremos que una variable aleatoria $X$ es *continua* cuando puede tomar cualquier valor en algún intervalo (o intervalos) del conjunto de los número reales. En esta situación se denomina función de densidad de probabilidad de $X$ a una función no negativa $f$, definida para todos los números reales $x \in (-\infty, \infty)$, que satisface que para cualquier conjunto $B$ de números reales se tiene:

$$P(X \in B) = \int_B f(x)dx \tag{1.3}$$

Además, esta función debe cumplir:

$$P\left[X \in (-\infty, \infty)\right] = \int_{-\infty}^{\infty} f(x)dx = 1 \tag{1.4}$$

Un concepto muy importante en la teoría de la probabilidad es el de *esperanza* de una variable aleatoria. Si $X$ es una variable aleatoria discreta con función de masa de probabilidad $p(x)$, entonces su valor esperado o esperanza $E(X)$ viene dado por:

$$E[X] = \sum_x xp(x) \tag{1.5}$$

Por otro lado, si $X$ es una variable aleatoria continua se define la esperanza de $X$ como:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \tag{1.6}$$

Una propiedad muy empleada de la esperanza es que, dada una función real $g$, si tenemos una variable discreta $X$ o una variable continua $Y$, se verifican:

$$E[g(X)] = \sum_x g(x)p(x) \qquad\qquad E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy \tag{1.7}$$

Podemos ver la demostración de esta propiedad en el libro *A First Course in Probability* de Sheldon Ross ([1]).

Usando esta última propiedad vemos que, dadas $a$ y $b$ dos constantes determinadas, se tiene que: $E[a + bX] = a + bE[X]$.

La esperanza de una variable aleatoria no nos informa sobre la dispersión de los posibles valores de $X$. Una medida que sí que nos aporta esta información es la llamada *varianza*. La varianza de una variable aleatoria $X$ se define de la siguiente manera:

$$Var(X) = E\left[(X - \mu)^2\right] \tag{1.8}$$

donde $\mu = E[X]$.

Una forma alternativa de escribir $Var(X)$ es:

$$Var(X) = E\left[X^2\right] - (E[X])^2 \tag{1.9}$$

Veamos que esta forma es equivalente a la dada en la definición, utilizando las propiedades de la esperanza:

$$\begin{aligned} Var(X) &= E\left[(X - \mu)^2\right] = E\left[X^2 - 2\mu X + \mu^2\right] \\ &= E\left[X^2\right] - 2\mu E[X] + \mu^2 = E\left[X^2\right] - \mu^2 \end{aligned} \tag{1.10}$$

Otra medida interesante es la llamada *covarianza* de dos variables aleatorias $X$ e $Y$, que nos da información sobre la relación existente entre ellas. Se define como:

$$Cov(X, Y) = E\left[(X - E[X])(Y - E[Y])\right] = E[XY] - E[X]E[Y] \tag{1.11}$$

Para llegar a la segunda igualdad se emplean las propiedades básicas de la esperanza de una variable aleatoria.

Algunas propiedades inmediatas de esta cantidad son que $Cov(X, Y) = Cov(Y, X)$ y que $Var(X) = Cov(X, X)$. Otra propiedad sencilla es la que sigue:

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j) \tag{1.12}$$

Probémosla escribiendo para abreviar $E[X_i] = \mu_i$ y $E[Y_j] = \nu_j$. Además, tengamos en cuenta que $E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mu_i$ y que $E\left[\sum_{j=1}^m Y_j\right] = \sum_{j=1}^m \nu_j$. Así:

$$\begin{aligned} Cov\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) &= E\left[\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right)\left(\sum_{j=1}^m Y_j - \sum_{j=1}^m \nu_j\right)\right] \\ &= \sum_{i=1}^n \sum_{j=1}^m E\left[(X_i - \mu_i)(Y_j - \nu_j)\right] \\ &= \sum_{i=1}^n \sum_{j=1}^m Cov(X_i, Y_j) \end{aligned} \tag{1.13}$$

Ahora que sabemos esto podemos demostrar la siguiente propiedad:

$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i) + 2\sum\sum_{i<j} Cov(X_i, X_j) \tag{1.14}$$

Tenemos:

$$\begin{aligned} Var\left(\sum_{i=1}^n X_i\right) &= Cov\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) = \sum_{i=1}^n \sum_{j=1}^n Cov\left(X_i, X_j\right) \\ &= \sum_{i=1}^n Var(X_i) + 2\sum\sum_{i<j} Cov(X_i, X_j) \end{aligned} \tag{1.15}$$

En el último paso hemos usado que cada par de índices $i, j$, con $i \neq j$, aparece dos veces en la doble suma anterior.

Definamos ahora un concepto también esencial que es el de *independencia de variables aleatorias*. Dadas dos variables aleatorias $X$ e $Y$, decimos que son independientes si para cualquier par de conjuntos de números reales $A$ y $B$ se tiene que:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{1.16}$$

Consideremos la función de densidad conjunta de dos variables continuas $X$ e $Y$, $f(x, y)$, definida por:

$$P\left[(X, Y) \in C\right] = \int\int_{(x,y)\in C} f(x, y)dxdy \tag{1.17}$$

donde $C \subset \mathbb{R}^2$.

Se tiene que $X$ e $Y$ son independientes si, y solo si, $f(x, y) = f_X(x)f_Y(y)$ para todo $x, y \in \mathbb{R}$, donde $f_X$ y $f_Y$ son las funciones de densidad marginales de $X$ y de $Y$, respectivamente.

En el caso de que $X$ e $Y$ sean variables discretas, la condición de independencia es equivalente a $p(x,y) = p_X(x)p_Y(y)$, para todo $x, y \in \mathbb{R}$.

En esta situación, es decir, si $X$ e $Y$ son independientes se tiene que $Cov(X,Y) = 0$. Veámoslo probando que $E[XY] = E[X]E[Y]$ y considerando que $X$ e $Y$ son continuas (si fuesen discretas el razonamiento sería muy parecido):

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy = E[X]E[Y] \end{aligned} \tag{1.18}$$

Por lo tanto, si $X_1, X_2, \cdots, X_n$ son variables aleatorias mutuamente independientes, entonces, aplicando que la covarianza entre dos de ellas es siempre 0, se tiene, observando la ecuación (1.15), que:

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) \tag{1.19}$$

Un concepto que también usaremos a lo largo del trabajo es el de *probabilidad condicionada*. Sean dos eventos $A$ y $B$ del espacio muestral $\Omega$ con $P(A) > 0$, entonces la probabilidad condicionada de $B$ dado $A$, es definida como:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1.20}$$

En relación a esto tenemos el *teorema de las probabilidades totales*, que usaremos más adelante, y que nos dice: dados $A_1, A_2, \ldots, A_k$ formando una partición de $\Omega$ (es decir, $A_i \cap A_j = \emptyset$ si $i \neq j$ y $\sum_{i=1}^{k} A_i = \Omega$) y teniendo $0 < P(A_i) < 1$ para todo $i$, entonces dado un evento $B$ en $\Omega$:

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i) \tag{1.21}$$

Algo que también nos va a ser de utilidad es calcular la función de distribución de la suma de dos variables aleatorias. Consideremos así dos variables aleatorias $X$ e $Y$ y calculemos la función de distribución de $Z = X+Y$, es decir, la convolución de las funciones de distribución de $X$ e $Y$:

$$\begin{aligned} F_Z(z) = P(X+Y \leq z) &= \int_{-\infty}^{\infty} P(X+Y \leq z|Y=y)f_Y(y)dy \\ &= \int_{-\infty}^{\infty} P(X \leq z-y|Y=y)f_Y(y)dy = \int_{-\infty}^{\infty} F_{X|Y}(z-y)f_Y(y)dy \end{aligned} \tag{1.22}$$

donde $F_{X|Y}$ es la función de distribución condicionada de $X$ dado $Y = y$.

Si $X$ e $Y$ son independientes, entonces $F_{X|Y} = F_X$. De esta forma:

$$F_Z(z) = \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy \tag{1.23}$$

Derivando esta expresión respecto a $z$ llegamos a la función de densidad de la suma:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy \tag{1.24}$$

Ahora que ya conocemos todos estos conceptos y razonamientos esenciales de la teoría de la probabilidad, vamos a presentar algunas de las distribuciones de probabilidad de variables aleatorias más comunes y que nos van a ser de gran utilidad más adelante.

### 1.2.1. Distribuciones de probabilidad útiles en este trabajo

#### Distribución de Bernoulli

Sea X una variable aleatoria que puede tomar los valores 0 y 1, haciendo referencia al fracaso o al éxito, respectivamente. La probabilidad de tener un éxito es $p$ y la de obtener un fracaso es $1-p$.

Decimos que una variable aleatoria $X$ sigue una distribución de Bernoulli, $X \sim Ber(p)$, si su función de masa de probabilidad viene dada por:

$$P(X = x) = p^x (1-p)^{1-x}, \quad \text{con } x = 0, 1 \tag{1.25}$$

Podemos calcular fácilmente su esperanza y su varianza:

$$E[X] = \sum_{x \in \{0,1\}} x \cdot P(X = x) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = p \tag{1.26}$$

$$Var(X) = E[X^2] - E[X]^2 = \sum_{x \in \{0,1\}} x^2 \cdot P(X = x) - p^2 = p - p^2 = p(1-p) \tag{1.27}$$

#### Distribución binomial

Consideremos ahora $n$ intentos independientes de Bernoulli de parámetro $p$, es decir, cada uno de ellos tiene probabilidad de éxito $p$ y probabilidad de fracaso $1-p$. Si $X$ representa el número de éxitos que tienen lugar en los $n$ intentos, entonces $X$ se dice que sigue una distribución de probabilidad binomial con parámetros $(n, p)$, $X \sim Bi(n, p)$.

La función de masa de probabilidad de una variable aleatoria binomial de parámetros $(n, p)$ viene dada por:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{con } x = 0, 1, 2, \ldots, n \tag{1.28}$$

Obtengamos la esperanza y la varianza de una variable aleatoria de este tipo. Para ello calculemos antes de nada $E[X^k]$ con $k \in \{1, 2, 3, \ldots\}$.

$$\begin{aligned} E[X^k] &= \sum_{x=0}^{n} x^k \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x^k \binom{n}{x} p^x (1-p)^{n-x} \\ &= np \sum_{x=1}^{n} x^{k-1} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} \\ &= np \sum_{i=0}^{n-1} (i+1)^{k-1} \binom{n-1}{i} p^i (1-p)^{n-i-1} \\ &= npE[(Y+1)^{k-1}] \end{aligned} \tag{1.29}$$

donde hemos usado que $i = x - 1$ y donde $Y$ es una variable aleatoria binomial de parámetros $(n-1, p)$.

Sustituyendo para $k = 1$ llegamos a que $E[X] = np$ y, por tanto, para $k = 2$ tenemos:

$$E[X^2] = npE[Y+1] = np\left[(n-1)p + 1\right] \tag{1.30}$$

Así, la varianza viene dada por:

$$Var(X) = E[X^2] - E[X]^2 = np\left[(n-1)p + 1\right] - (np)^2 = np(1-p) \tag{1.31}$$

**Distribución geométrica**

Consideremos una serie de intentos independientes, cada uno de ellos con dos posibles resultados: éxito (con probabilidad $p$) o fracaso (con probabilidad $q = 1 - p$). Sea $X$ la variable que describe el número de intentos necesarios para conseguir el primer éxito, decimos que $X$ sigue una distribución geométrica de parámetro $p$, $X \sim Geo(p)$. Esta variable tiene una función de masa de probabilidad dada por:

$$P(X = x) = p(1-p)^{x-1}, \quad \text{con } x = 1, 2, \ldots \tag{1.32}$$

Para las variables que siguen esta distribución podemos calcular la esperanza como se muestra a continuación:

$$\begin{aligned} E[X] &= \sum_{x=1}^{\infty} xpq^{x-1} = \sum_{x=1}^{\infty} (x-1+1)pq^{x-1} = \sum_{x=1}^{\infty} (x-1)pq^{x-1} + \sum_{x=1}^{\infty} pq^{x-1} \\ &= q \sum_{i=0}^{\infty} ipq^{i-1} + 1 = qE[X] + 1 \end{aligned} \tag{1.33}$$

Despejando esto llegamos a que $E[X] = \frac{1}{p}$. Calculemos ahora $E[X^2]$ para poder obtener después $Var(X)$.

$$\begin{aligned} E[X^2] &= \sum_{x=1}^{\infty} x^2 pq^{x-1} = \sum_{x=1}^{\infty} (x-1+1)^2 pq^{x-1} \\ &= \sum_{x=1}^{\infty} (x-1)^2 pq^{x-1} + \sum_{x=1}^{\infty} 2(x-1)pq^{x-1} + \sum_{x=1}^{\infty} pq^{x-1} \\ &= q\sum_{i=0}^{\infty} i^2 pq^{i-1} + 2q\sum_{i=0}^{\infty} ipq^{i-1} + 1 \\ &= qE[X^2] + 2qE[X] + 1 = qE[X^2] + \frac{2q}{p} + 1 \end{aligned} \tag{1.34}$$

Despejando de esta última expresión llegamos a que $E[X^2] = \frac{q+1}{p^2}$. Obtengamos ahora $Var(X)$:

$$Var(X) = E[X^2] - E[X]^2 = \frac{q+1}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2} \tag{1.35}$$

Para probar una serie de propiedades interesantes, nos va a ser útil conocer $P(X \leq x_0)$:

$$P(X \leq x_0) = \sum_{x=1}^{x_0} pq^{x-1} = p\sum_{i=0}^{x_0-1} q^i = p\left(\frac{1-q^{x_0}}{1-q}\right) = 1-(1-p)^{x_0} \tag{1.36}$$

Ahora que ya sabemos esto, vamos a considerar la propiedad que hace referencia a la *falta de memoria* de esta distribución:

$$P(X > x_2 + x_1 | X > x_1) = P(X > x_2) \tag{1.37}$$

Para entender bien esta propiedad de manera general imaginémonos que $X$ hace referencia al número de veces que sale cara antes de que salga cruz al tirar una moneda. La probabilidad de que salga cruz cuando han salido $x_1 + x_2$ caras, sabiendo que ya han salido $x_1$ caras, es la misma que la probabilidad inicial de que salga cruz después de $x_2$ caras. Es decir, se ha "olvidado" que ya han salido $x_1$ caras.

Veamos su demostración:

$$\begin{aligned} P(X > x_2 + x_1 | X > x_1) &= \frac{P(X > x_2 + x_1, X > x_1)}{P(X > x_1)} = \frac{P(X > x_2 + x_1)}{P(X > x_1)} \\ &= \frac{(1-p)^{x_2+x_1}}{(1-p)^{x_1}} = (1-p)^{x_2} = P(X > x_2) \end{aligned} \tag{1.38}$$

**Distribución de Poisson**

Sea $X$ una variable aleatoria que mide el número de eventos que tienen lugar en un determinado intervalo. Se dice que $X$ sigue una distribución de Poisson de parámetro $\lambda$,

$X \sim Po(\lambda)$, si $\lambda$ es el número promedio de veces que se espera que ocurra el fenómeno en un intervalo dado. Se tiene que la función de masa de probabilidad es:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad \text{con } x = 0, 1, 2, \ldots \tag{1.39}$$

donde $x$ es el número veces que ocurre el fenómeno.

Podemos obtener la esperanza y la varianza de una variable aleatoria que siga esta distribución:

$$E[X] = \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!} = \lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = \lambda e^{-\lambda}e^{\lambda} = \lambda \tag{1.40}$$

Para determinar la varianza demos primero el valor de $E[X^2]$:

$$\begin{aligned} E[X^2] &= \sum_{x=0}^{\infty} x^2\frac{e^{-\lambda}\lambda^x}{x!} = \lambda\sum_{x=1}^{\infty} x\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda\sum_{i=0}^{\infty}(i+1)\frac{e^{-\lambda}\lambda^i}{i!} \\ &= \lambda\left[\sum_{i=0}^{\infty} i\frac{e^{-\lambda}\lambda^i}{i!} + \sum_{i=0}^{\infty}\frac{e^{-\lambda}\lambda^i}{i!}\right] = \lambda(\lambda+1) \end{aligned} \tag{1.41}$$

De esta forma, restando, tenemos: $Var(X) = \lambda$

**Distribución exponencial**

Sea $X$ una variable continua que sigue una distribución exponencial de parámetro $\lambda$, $X \sim Exp(\lambda)$. Su función de densidad viene dada por:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & si \quad x \geq 0 \\ 0 & si \quad x < 0 \end{cases} \tag{1.42}$$

La función de distribución se obtiene de la siguiente manera:

$$F(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda x}dx = 1 - e^{-\lambda x}, \quad \text{con } x \geq 0 \tag{1.43}$$

Podemos obtener la esperanza y la varianza de una variable aleatoria que sigue una distribución exponencial. Para ello calculemos antes de nada $E[X^k]$, con $k > 0$.

$$\begin{aligned} E[X^k] &= \int_0^{\infty} x^k\lambda e^{-\lambda x}dx = -x^k e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} kx^{k-1}e^{-\lambda x}dx \\ &= 0 + \frac{k}{\lambda}E[X^{k-1}] = \frac{k}{\lambda}E[X^{k-1}] \end{aligned} \tag{1.44}$$

Sustituyendo para $k = 1$ y después para $k = 2$ llegamos a:

$$E[X] = \frac{1}{\lambda} \qquad E[X^2] = \frac{2}{\lambda^2} \tag{1.45}$$

Calculemos ahora la varianza de esta variable aleatoria:

$$Var(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \tag{1.46}$$

Otra propiedad que nos será de utilidad es que si $X$ sigue una distribución de parámetro $\lambda$ y $c$ es una constante mayor que 0, entonces:

$$Y = cX \sim Exp\left(\frac{\lambda}{c}\right) \tag{1.47}$$

Probémoslo calculando su función de distribución que, como veremos a continuación se corresponde con la de una variable que sigue una distribución exponencial de parámetro $\frac{\lambda}{c}$:

$$F(y) = P(Y \leq y) = P\left(X \leq \frac{y}{c}\right) = \int_0^{\frac{y}{c}} \lambda e^{-\lambda y} dy = 1 - e^{-\frac{\lambda}{c}y} \tag{1.48}$$

Además, dados $x_1, x_2 \geq 0$, en esta distribución también se tiene la propiedad de la *falta de memoria* (que ya hemos visto en la distribución geométrica):

$$P(X > x_2 + x_1 | X > x_1) = P(X > x_2) \tag{1.49}$$

Veamos la prueba de esta propiedad:

$$\begin{aligned} P(X > x_2 + x_1 | X > x_1) &= \frac{P(X > x_2 + x_1, X > x_1)}{P(X > x_1)} = \frac{P(X > x_2 + x_1)}{P(X > x_1)} \\ &= \frac{e^{-\lambda(x_1+x_2)}}{e^{-\lambda x_1}} = e^{-\lambda x_2} = P(X > x_2) \end{aligned} \tag{1.50}$$

Sea $Y$ otra variable que sigue una distribución exponencial de parámetro $\lambda'$ y asumamos que $X$ e $Y$ son independientes. Se tienen las dos siguientes propiedades:

$$P(X < Y) = \frac{\lambda}{\lambda + \lambda'} \tag{1.51}$$

$$\min(X, Y) \sim Exp(\lambda + \lambda') \tag{1.52}$$

Veamos que se tiene la primera de las propiedades considerando la probabilidad condicionada:

$$\begin{aligned} P(X < Y) &= \int_0^\infty P(X < Y | X = x) \lambda e^{-\lambda x} dx = \int_0^\infty P(Y > x) \lambda e^{-\lambda x} dx \\ &= \int_0^\infty e^{-\lambda' x} \lambda e^{-\lambda x} dx = \int_0^\infty \lambda e^{-(\lambda+\lambda')x} dx = \frac{\lambda}{\lambda + \lambda'} \end{aligned} \tag{1.53}$$

La segunda propiedad la podemos demostrar estudiando la función de distribución de $\min(X, Y)$:

$$\begin{aligned} P(\min(X, Y) \leq x) &= 1 - P\left(\min(X, Y) > x\right) = 1 - P(X > x) P(Y > x) \\ &= 1 - e^{-\lambda x} e^{-\lambda' x} = 1 - e^{-(\lambda+\lambda')x} \end{aligned} \tag{1.54}$$

**Distribución multinomial**

Consideremos una serie de $n$ intentos independientes, en cada uno de los cuales solamente se puede observar uno de los $k$ siguientes eventos excluyentes $E_1, E_2, \ldots, E_k$, y en los que las probabilidad de que tenga lugar el evento $E_j$ en cualquiera de los intentos es $p_j$ (se tiene que $p_1 + p_2 + \cdots + p_k = 1$). Sean $X_1, X_2, \ldots, X_k$ las variables aleatorias que denotan el número de veces que ocurren los eventos $E_1, E_2, \ldots, E_k$, respectivamente, en los $n$ intentos, con $\sum_{i=1}^{k} X_i = n$. Por tanto, la distribución conjunta de $X_1, X_2, \ldots, X_k$ viene dada por: [2]

$$\begin{aligned} P\left[X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k\right] &= \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \\ &= \binom{n}{x_1, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \end{aligned} \tag{1.55}$$

con $x_i \geq 0$ y $\sum_{i=1}^{k} x_i = n$.

Esta distribución recibe el nombre de multinomial con parámetros $(n; p_1, p_2, \ldots, p_k)$.

Mencionemos ahora algunas propiedades de la distribución que nos serán de utilidad. Para empezar, por la propia definición de la distribución, cada $X_i$ sigue una distribución binomial de parámetros $n$ y $p_i$. Por lo tanto:

$$E[X_i] = np_i \qquad\qquad Var(X_i) = np_i(1 - p_i) \tag{1.56}$$

Para calcular $E[X_i X_j]$ y $Cov(X_i, X_j)$ emplearemos la *función generatriz de momentos*. Dada una variable aleatoria $X$ se define su función generatriz de momentos como:

$$M_X(t) = E\left[e^{tX}\right], \qquad\qquad \text{con } t \in \mathbb{R} \tag{1.57}$$

Sea:

$$S = \left\{ x_1, x_2, \ldots, x_k \middle| x_i \in \{0, 1, 2, \ldots\}, \sum_{i=1}^{k} x_i = n \right\} \tag{1.58}$$

La función generatriz de momentos de una variable multinomial viene dada por:

$$\begin{aligned} M_{X_1, X_2, \ldots, X_k} &= E\left[e^{t_1 X_1 + t_2 X_2 + \cdots + t_k X_k}\right] = \sum_{S} \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} p_i^{x_i} e^{t_i x_i} \\ &= \sum_{S} \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} \left(p_i e^{t_i}\right)^{x_i} = \left(\sum_{i=1}^{k} p_i e^{t_i}\right)^n \end{aligned} \tag{1.59}$$

En el último paso hemos empleado el teorema multinomial que establece que dados $y_1, \ldots, y_k \in \mathbb{R}$, $k$ entero positivo y $n$ entero no negativo, entonces:

$$(y_1 + y_2 + \cdots + y_k)^n = \sum_{S} \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} y_i^{x_i} \tag{1.60}$$

Hagamos uso del siguiente resultado, cuya demostración para el caso de una variable aleatoria se puede ver en [3] y para el caso de un vector aleatorio en [4].

Si $(X_1, X_2, \dots X_k)$ es un vector aleatorio que presenta una función generatriz de momentos $M_{X_1,X_2,\dots,X_k}(t_1, t_2, \dots, t_k)$, entonces:

$$E\left[X_1^{r_1} \cdot X_2^{r_2} \cdot \dots \cdot X_k^{r_k}\right] = \left.\frac{\partial^{r_1+r_2+\dots+r_k} M_{X_1,X_2,\dots,X_k}(t_1, t_2, \dots, t_k)}{\partial t_1^{r_1} \partial t_2^{r_2} \dots \partial t_k^{r_k}}\right|_{t_1=0,t_2=0,\dots,t_k=0} \tag{1.61}$$

Con esto ya podemos calcular el valor de $E[X_i X_j]$, con $i \neq j$, y después también el de $Cov(X_i, X_j)$, pues $Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$. Así:

$$E\left[X_i \cdot X_j\right] = \left.\frac{\partial^2 M_{X_1,X_2,\dots,X_k}(t_i, t_j, \dots, t_k)}{\partial t_i \partial t_j}\right|_{t_1=0,t_2=0,\dots,t_k=0} = n(n-1)p_i p_j \tag{1.62}$$

$$Cov(X_i, X_j) = n(n-1)p_i p_j - n^2 p_i p_j = -n p_i p_j \tag{1.63}$$

Con esto terminamos este primer capítulo de introducción a los conceptos básicos tanto biológicos como matemáticos, que usaremos de ahora en adelante.

# Capítulo 2

# Coalescencia en el modelo de Wright-Fisher

Ahora que ya conocemos los conceptos necesarios para poder entender las bases de la teoría de la coalescencia, vamos a comenzar a estudiarla. La teoría de la coalescencia se enmarca dentro de la genética de poblaciones y nace a partir del estudio de uno de los modelos poblacionales más comunes: el de Wright-Fisher.

Para comenzar el estudio de la coalescencia vamos a presentar en detalle este modelo de poblaciones y después veremos cómo se deduce la coalescencia a partir de él. Tanto en este capítulo como en los que siguen nos guiaremos por dos libros: *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory* ([5]) y *Coalescent Theory: An Introduction* ([6]).

## 2.1. Modelo de Wright-Fisher

Wright y Fisher introducen, alrededor de 1930, un modelo de poblaciones sencillo que describe la relación genealógica entre los genes. Este modelo proporciona una descripción de la evolución de una población idealizada y de la transmisión de genes de una generación a la siguiente.

Se puede aplicar tanto a poblaciones de individuos haploides como diploides y, para facilitar la comparación de los modelos en ambos casos, vamos a considerar una población de $2N$ genes, de forma que en el caso haploide tendremos $2N$ individuos y en el diploide $N$.

En el modelo haploide, cada gen de la generación $t+1$ se obtiene mediante la copia del gen de un individuo aleatorio de la generación $t$. Este proceso se repite independientemente hasta que se tienen los $2N$ genes de la nueva generación. Cada gen en la generación $t+1$

tiene un padre en la generación $t$, pero un gen en la generación $t$ puede no tener ningún descendiente en la generación $t+1$ y, por tanto, su linaje muere. En la Figura 2.1, a la izquierda, se muestra el procedimiento.

En el caso de la reproducción diploide de especies con sexos separados, se asumen dos subpoblaciones (hembras y machos) de tamaños $N_f$ y $N_m$, respectivamente, con $N = N_f + N_m$, representando de nuevo $2N$ genes. Cada individuo de la generación $t+1$ escoge un macho (padre) y una hembra (madre) de la generación $t$. De cada padre y de cada madre, uno de los dos genes que presentan es escogido con probabilidad $\frac{1}{2}$. Este esquema reproductivo se muestra en la Figura 2.1, a la derecha. Como en el modelo haploide, cada gen tiene un gen predecesor (en un macho o en una hembra), pero ahora cada individuo tiene dos progenitores.

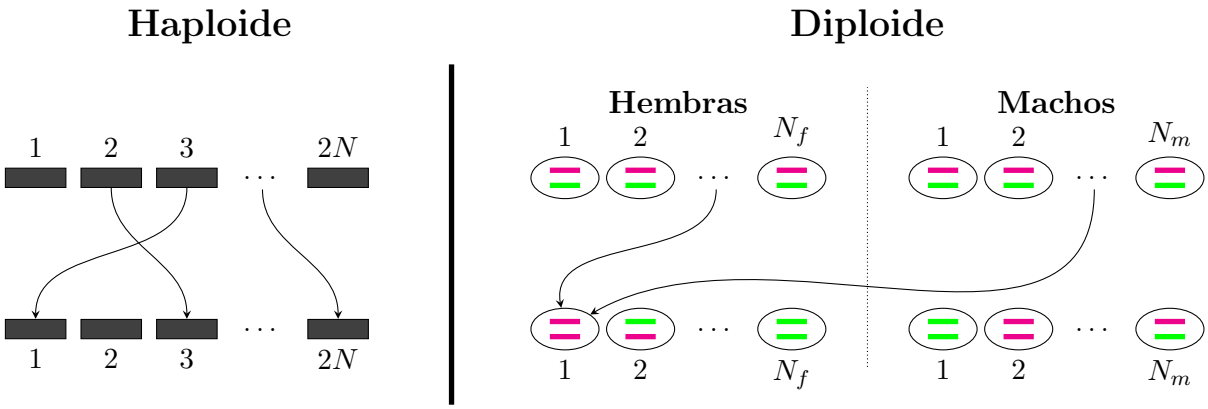



Figura 2.1: Propiedades básicas del modelo en el caso haploide y diploide.

El estudio de los ancestros de dos genes es ligeramente distinto si estamos en el modelo haploide o en el diploide pero, en este trabajo vamos a considerar el modelo haploide por su mayor sencillez a la hora de presentar algunos conceptos.

Establezcamos ahora las hipótesis en las que se basa este modelo de reproducción de Wright-Fisher:

1. Generaciones discretas y no superpuestas. En el caso de los seres humanos esto es equivalente a considerar que todos tenemos la misma esperanza de vida desde el nacimiento hasta el momento de la reproducción, y que la reproducción y la muerte ocurren al mismo tiempo y de forma síncrona para todos los individuos. De todas formas, modelos que consideran "generaciones que se solapan" obtienen genealogías probabilísticamente similares.

2. Individuos haploides o dos subpoblaciones (machos y hembras), como se ve en la Figura 2.1.

3. El tamaño de la población es constante. Esta es una hipótesis esencial en el modelo, pues si la población aumentase, oscilase o disminuyese habría importantes modificaciones en las predicciones del modelo.

4. Todos los individuos tienen la misma eficacia biológica o aptitud (la misma capacidad de tener descendencia). Esta hipótesis es importante a la hora de introducir los conceptos básicos, pero deja de lado fenómenos importantes como la selección natural.

5. La población no tiene estructura geográfica o social. La elección de padres de forma aleatoria como en este modelo no es un mecanismo real en ninguna población, por lo que, en el análisis de datos reales, la estructura de la población puede afectar significativamente a las genealogías.

6. Los genes en una población no se recombinan. Esta es una asunción importante que en muchos casos tiene que ser relajada para el estudio de datos reales. El problema es que relajar esta hipótesis hace que el desarrollo matemático sea mucho más complejo y que ya no se pueda estudiar mediante un árbol genealógico, sino que haya que emplear estrategias más sofisticadas como algunos gráficos específicos o colecciones de árboles.

Ahora que ya hemos presentado las hipótesis por las que se rige este modelo, vamos a estudiar el proceso general de descendencia de un gen en una población dada.

### 2.1.1. Número de descendientes de un gen en una generación

El número de descendientes de un determinado gen, $i$, en la generación $t$ es una variable aleatoria. Cada vez que aparece un gen en la generación $t+1$, tiene una probabilidad $\frac{1}{2N}$ de tener como padre al gen $i$ de la generación $t$. Este muestreo se realiza repetidamente $2N$ veces con reemplazamiento para conseguir los $2N$ genes presentes en la generación $t+1$. Por lo tanto, la variable aleatoria que indica el número de descendientes del gen $i$ en la generación $t$, y a la que denotaremos por $V_i$, sigue una distribución binomial con parámetros $2N$ y $\frac{1}{2N}$. Así, la probabilidad de que en la generación $t+1$ haya $v$ descendientes del gen $i$ es:

$$P(V_i = v) = \binom{2N}{v} \left(\frac{1}{2N}\right)^v \left(1 - \frac{1}{2N}\right)^{2N-v} \tag{2.1}$$

Ya conocemos los momentos de una distribución binomial por lo que se tiene:

$$E(V_i) = 2N \cdot \frac{1}{2N} = 1 \tag{2.2}$$

$$Var(V_i) = 2N \cdot \frac{1}{2N} \cdot \left(1 - \frac{1}{2N}\right) = 1 - \frac{1}{2N} \tag{2.3}$$

Que la media de $V_i$ sea 1 es consecuencia de que el tamaño de la población es constante: si la media del número de descendientes de un gen fuese mayor o menor que 1, entonces indicaría que la población está aumentando o disminuyendo de tamaño, respectivamente.

Además, la distribución conjunta del número de descendientes de los $2N$ genes en una generación dada, es una multinomial de parámetros $2N$ y $p_1 = p_2 = \cdots = p_{2N} = \frac{1}{2N}$. Los posibles sucesos característicos de esta distribución son que un determinado gen $i$ (con $1 \leq i \leq 2N$) sea el padre de un miembro de la siguiente generación. Así, podemos calcular la probabilidad de que el gen 1 tenga $v_1$ descendientes en la generación $t+1$, el 2 tenga $v_2$ y así sucesivamente hasta que el gen $2N$ tenga $v_{2N}$ descendientes en esa generación (recordemos que se tiene que verificar que $v_1 + v_2 + \cdots + v_{2N} = 2N$). Esta probabilidad viene dada por:

$$\begin{aligned} P(V_1 = v_1, \cdots, V_{2N} = v_{2N}) &= \frac{(2N)!}{v_1! \cdots v_{2N}!} \left(\frac{1}{2N}\right)^{v_1} \cdots \left(\frac{1}{2N}\right)^{v_{2N}} \\ &= \frac{(2N)!}{v_1! \cdots v_{2N}!} \left(\frac{1}{2N}\right)^{2N} \end{aligned} \tag{2.4}$$

También podemos obtener la covarianza y el coeficiente de correlación del número de descendientes de dos genes $i$ y $j$:

$$Cov(V_i, V_j) = -2Np_ip_j = -\frac{1}{2N} \tag{2.5}$$

$$Cor(V_i, V_j) = \frac{Cov(V_i, V_j)}{\sqrt{Var(V_i)Var(V_j)}} = -\frac{1}{2N-1} \tag{2.6}$$

Vemos que para valores grandes de $2N$, $V_i$ y $V_j$ presentan una correlación baja. Además, es de esperar que la covarianza entre $V_i$ y $V_j$ sea negativa pues, si el gen $i$ deja muchos descendientes en la nueva generación, el gen $j$ es más probable que deje pocos. Esto es una consecuencia de que el tamaño de la población sea constante e igual a $2N$.

Cuando el número $2N$ es muy elevado, podemos ver que la distribución de $V_i$ tiende a una Poisson con media y varianza 1, $Po(1)$:

$$\begin{aligned} \lim_{2N\to\infty} P(V_i = v) &= \lim_{2N\to\infty} \binom{2N}{v} \left(\frac{1}{2N}\right)^v \left(1 - \frac{1}{2N}\right)^{2N-v} \\ &= \frac{1}{v!} \lim_{2N\to\infty} \frac{2N!}{(2N-v)!} \frac{1}{(2N-1)^v} \left(1 - \frac{1}{2N}\right)^{2N} = \frac{1}{v!} e^{-1} \end{aligned} \tag{2.7}$$

De esta forma, podemos expresar de manera informal que cuando $2N$ es suficientemente grande, la distribución que sigue el número de descendientes de un gen $i$ en la generación siguiente es aproximadamente una Poisson de parámetro 1:

$$P(V_i = v) \approx \frac{1}{v!}e^{-1} \tag{2.8}$$

Teniendo esto en cuenta, vemos que la probabilidad de que un gen no deje descendientes en una determinada generación es de aproximadamente $e^{-1} \approx 0{,}37$ y, por tanto, hay una probabilidad de 0,63 de que sí que los deje. Dada, por ejemplo, una población de tamaño $2N = 10000$, si estudiamos el número de genes ancestrales hace $t = 15$ generaciones veremos que tan solo tenemos $0{,}63^{15}10000 \approx 10$ genes. Es decir, el resto de genes, unos 9990, han perdido su linaje en estas 15 generaciones.

Estos razonamientos solamente son válidos si estamos considerando un número muy grande de genes, en otro caso, si buscamos el ancestro común de un grupo pequeño de todos estos genes recurriríamos a la teoría de la coalescencia.

Para ilustrar esto consideremos una población de 6 genes de los cuales buscamos el ancestro común más reciente (MRCA) del primero, del tercero y del sexto. Como se observa en la Figura 2.2, dos generaciones hacia atrás el primer gen y el tercero encuentran su ancestro común y cuatro generaciones hacia atrás lo encuentran los tres genes que estábamos considerando.

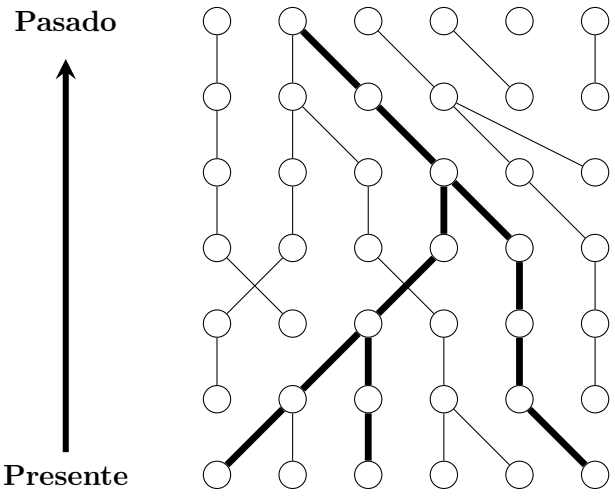



Figura 2.2: Genealogía de tres genes en una población que presenta en total seis.

Modelizando matemáticamente las distintas variables que intervienen en estos procesos, podremos llegar a conocer cómo es la estructura general de las genealogías de este tipo y algunas cantidades que resultan de interés, como el número de linajes que hay en un determinado instante de tiempo.

## 2.2. Fundamentos de la coalescencia

Como ya hemos mencionado, es interesante estudiar cuánto tiempo pasa hasta que un conjunto de genes encuentran su ancestro común más reciente. Si medimos el tiempo en forma discreta, es decir, en generaciones, y consideramos la variable aleatoria $T$ que da información sobre el tiempo de espera hasta que aparece el MRCA de dos genes tenemos que $T$ sigue una distribución geométrica. El éxito tendrá lugar si se encuentra el ancestro común de esos dos genes, y el fracaso si no es así (con probabilidades $p$ y $1-p$ respectivamente). Así, $T \sim Geo(p)$:

$$P(T = t) = p(1-p)^{t-1} \qquad \text{con } t = 1, 2, \ldots \tag{2.9}$$

Que $T = t$ implica que ha habido $t-1$ fracasos ($t-1$ generaciones en las que no se ha encontrado un ancestro común) antes del primer éxito, el encuentro de su MRCA.

### 2.2.1. Coalescencia en tiempo discreto

Una vez que ya conocemos qué distribución sigue la variable que hace referencia al número de generaciones que pasan hasta que se encuentra el ancestro común más reciente de dos genes, podemos estudiar cómo es este proceso de coalescencia en muestras de dos genes y después generalizarlo para muestras con un número mayor de genes.

#### Coalescencia para una muestra de dos genes

Lo primero que haremos será estudiar la distribución del tiempo de espera hasta que se obtenga el MRCA de dos genes en un modelo haploide con $2N$ genes. La probabilidad de que estos dos genes encuentren un ancestro común en la primera generación hacia atrás en el tiempo es de $\frac{1}{2N}$ pues el primer gen puede elegir a su padre con libertad, pero el segundo debe escoger el mismo que el primero, es decir, solo puede escoger 1 de las $2N$ posibilidades existentes. La probabilidad de que dos genes tengan distintos ancestros es, por tanto, $1-\frac{1}{2N}$. Como las selecciones de los padres en distintas generaciones son procesos independientes, la probabilidad de que dos genes encuentren un ancestro común $t$ generaciones atrás en el tiempo es:

$$\left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} \tag{2.10}$$

Esto representa que en las primeras $t-1$ generaciones no se encuentra el ancestro común pero en la t-ésima sí. Por lo tanto, el tiempo de coalescencia hasta que dos genes encuentren su MRCA, que denotaremos $T_2$, sigue una distribución geométrica con parámetro $\frac{1}{2N}$:

$$P(T_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} \qquad \text{con } t = 1, 2, \ldots \tag{2.11}$$

El valor esperado de $T_2$ viene dado por: $E(T_2) = \frac{1}{\frac{1}{2N}} = 2N$. Esto indica que si nos vamos tantas generaciones hacia atrás en el tiempo, como genes hay en la población, se espera encontrar el ancestro común más reciente de dos genes.

**Coalescencia en una muestra de n genes**

Generalicemos ahora lo obtenido en el apartado anterior para una muestra de $n$ genes, donde $n$ se considera mucho menor que $2N$, el tamaño de la población. Empecemos obteniendo la distribución del tiempo de espera hasta que $k(\leq n)$ genes presenten menos de $k$ linajes ancestrales, es decir, hasta que alguno de los $k$ genes comparta un ancestro común con otro gen de ese conjunto en la anterior generación. La probabilidad de que $k$ genes tengan $k$ ancestros distintos en la anterior generación se obtiene de forma similar al caso de dos genes: el primer gen puede escoger libremente entre los $2N$ genes, el segundo tiene que escoger un padre distinto y entonces solo puede escoger entre $2N-1$, el tercero entre $2N-2$ y así se sigue hasta que el último puede escoger entre $2N-(k-1)$. Por tanto, esta probabilidad viene dada por:

$$\begin{aligned}\frac{(2N-1)}{2N}\frac{(2N-2)}{2N}\cdots\frac{(2N-k+1)}{2N} &= \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{k-1}{2N}\right)\\ &= 1-\sum_{i=1}^{k-1}\frac{i}{2N}+O\left(\frac{1}{N^2}\right)\\ &= 1-\binom{k}{2}\frac{1}{2N}+O\left(\frac{1}{N^2}\right) \qquad (2.12)\end{aligned}$$

Para pasar de la primera línea a la segunda hemos realizado los productos y agrupado los sumandos con potencias mayores o iguales a $\frac{1}{N^2}$ en el término $O\left(\frac{1}{N^2}\right)$, verificándose entonces que $\lim_{N\to\infty}\frac{O\left(\frac{1}{N^2}\right)}{\frac{1}{N^2}} = \text{cte}$. Para pasar de la segunda línea a la tercera hemos usado que $\sum_{i=1}^{k-1} i = \binom{k}{2} = \frac{k(k-1)}{2}$. Probemos esta igualdad por inducción en $k$, siendo $k \geq 2$.

**Paso 1:** Estudiemos la validez de la fórmula para el caso $k = 2$:

$$\frac{2(2-1)}{2} = 1 \qquad (2.13)$$

**Paso 2:** Consideremos que la propiedad se cumple para un determinado $k$ y veamos que se verifica también para $k+1$.

Así, la hipótesis de inducción será:

$$\sum_{i=1}^{k-1} i = \frac{k(k-1)}{2} \qquad (2.14)$$

Veamos qué ocurre para $k+1$:

$$\sum_{i=1}^{k} i = \sum_{i=1}^{k-1} i + k = \frac{k(k-1)}{2} + k = \frac{k(k-1)+2k}{2} = \frac{(k+1)((k+1)-1)}{2} \tag{2.15}$$

Con esto concluimos la prueba y establecemos que la fórmula es válida para todo número entero $k$ mayor o igual que 2.

Como se asume que $n$ es mucho menor que $N$, los términos con potencias de $1/N^2$ o mayores, es decir, $O\left(\frac{1}{N^2}\right)$, son despreciables y pueden ser ignorados. Esta aproximación es equivalente a ignorar la posibilidad de que más de un par de genes encuentren un ancestro común en la misma generación. Por lo tanto, cuando $n$ es mucho menor que $N$, la probabilidad de que no se produzca ninguna coalescencia, es decir, que dados $k$ genes todos tengan un ancestro distinto en la anterior generación, es:

$$1 - \binom{k}{2}\frac{1}{2N} \tag{2.16}$$

y, por tanto, la probabilidad de que tenga lugar una coalescencia en una generación dada es:

$$\binom{k}{2}\frac{1}{2N} \tag{2.17}$$

En consecuencia, denotando por $T_k$ el tiempo de coalescencia hasta que dos genes, de los $k$ considerados, encuentren su MRCA, se tiene que la probabilidad de que dos de los $k$ genes considerados encuentren un ancestro común $t$ generaciones en el pasado (con $t = 1, 2, \dots$) es, aproximadamente:

$$P\left(T_k = t\right) \approx \left[1 - \binom{k}{2}\frac{1}{2N}\right]^{t-1}\binom{k}{2}\frac{1}{2N} \tag{2.18}$$

Por lo tanto, vemos que la variable aleatoria $T_k$ sigue, aproximadamente, una distribución geométrica con parámetro $\frac{\binom{k}{2}}{2N}$.

En la Figura 2.3 hemos considerado 5 genes y vemos el árbol de coalescencia asociado a su genealogía. Con este árbol buscamos comprender bien a qué nos referimos cuando hablamos de $T_k$, es decir, del tiempo en el que hay $k$ ancestros de los 5 genes seleccionados. En la imagen vemos diferenciadas las distintas etapas $T_k$ con $k = 2, 3, 4, 5$.

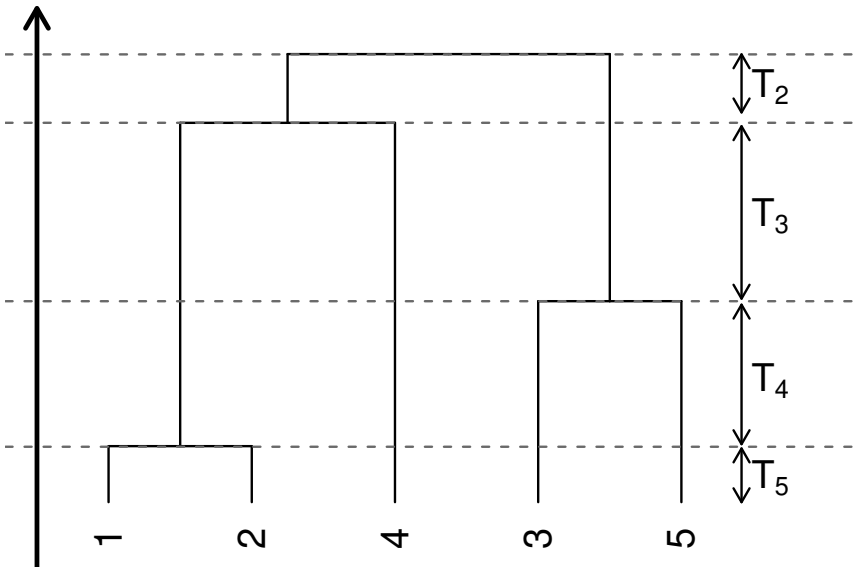



Figura 2.3: Árbol de coalescencia con los distintos $T_k$ señalados, siendo $k = 2, 3, 4, 5$.

La exactitud de las aproximaciones realizadas para poblaciones de gran tamaño, lleva a una formulación de la coalescencia en la que se emplea un modelo basado en la continuidad del tiempo y que es independiente de $2N$. Es aquí donde introduciremos que el tiempo (continuo) hasta encontrar el MRCA de dos genes sigue una distribución exponencial.

### 2.2.2. Coalescencia en tiempo continuo

Una de las formas más naturales de introducir el tiempo en escala continua en la teoría de la coalescencia es considerando que una unidad de tiempo se corresponde con el tiempo medio que tardan dos genes en encontrar un ancestro común (ya hemos visto que son $2N$ generaciones). Esta transformación del tiempo hace que la coalescencia se vuelva independiente del tamaño de la población.

Para derivar el proceso continuo de la coalescencia, se considera $t_c = \frac{t}{2N}$ donde $t$ es el tiempo medido en generaciones. Obviamente, se puede pasar el tiempo continuo a generaciones simplemente despejando $t$; si el valor que se obtiene $t = 2Nt_c$ no es un entero entonces $t$ se trunca al entero menor más cercano.

Ahora veremos que la distribución geométrica, empleada en el apartado anterior, puede

ser aproximada por una distribución exponencial. Sea $T \sim Geo(p)$, se tiene:

$$P(T > t) = 1 - P(T \leq t) = 1 - \left[1 - (1-p)^t\right] = (1-p)^t \tag{2.19}$$

Considerando $a = p2N$, podemos reescribir $(1-p)^t$ como:

$$\left(1 - \frac{p2N}{2N}\right)^{\frac{2Nt}{2N}} = \left(1 - \frac{a}{2N}\right)^{t_c 2N} \tag{2.20}$$

Así, en el límite de $2N$ muy grande se tiene:

$$\lim_{2N\to\infty} P(T > t) = \lim_{2N\to\infty} P\left(\frac{T}{2N} > t_c\right) = \lim_{2N\to\infty} \left(1 - \frac{a}{2N}\right)^{t_c 2N} = e^{-at_c} \tag{2.21}$$

De esta forma, la variable que definiremos por $T^c = \frac{T}{2N}$, sigue una distribución exponencial con parámetro $a$ en el límite de $2N$ tendiendo a $\infty$.

Denotaremos el tiempo hasta que $k$ genes tengan $k-1$ ancestros en el caso continuo como $T_k^c$, que es simplemente una variable dada por $\frac{T_k}{2N}$, donde $T_k$ es el tiempo medido en generaciones, para que $k$ genes tengan $k-1$ ancestros. Recordemos que $T_k \sim Geo\left(\frac{\binom{k}{2}}{2N}\right)$ y, por lo que acabamos de ver, se tiene que: $T_k^c \sim Exp\left(\binom{k}{2}\right)$. Por lo tanto:

$$P(T_k^c \leq t_c) = 1 - e^{-\binom{k}{2} t_c} \tag{2.22}$$

Por estar distribuido exponencialmente, la media y la varianza de estos tiempos de coalescencia $T_k^c$ vienen dadas por:

$$E(T_k^c) = \frac{2}{k(k-1)} \tag{2.23}$$

$$Var(T_k^c) = \left(\frac{2}{k(k-1)}\right)^2 \tag{2.24}$$

De la ecuación (2.23), vemos que según disminuye $k$, es decir, el número de genes entre los que esperamos que dos encuentren su ancestro común, aumenta el valor esperado de $T_k^c$. Por tanto, el tiempo de coalescencia cuando solamente quedan dos genes para encontrar su ancestro común es el que se espera que sea más largo.

Es posible describir un algoritmo que crea genealogías para $n$ genes, considerando la expresión continua del tiempo. A continuación vemos los distintos pasos que definen el algoritmo.

**Algoritmo 1**

1. Empezar con $k = n$ genes.
2. Simular el tiempo de espera $T_k^c$ hasta el siguiente evento de coalescencia, sabiendo que $T_k^c \sim Exp\left(\binom{k}{2}\right)$.

3. Escoger aleatoriamente una pareja $(i, j)$ con $1 \leq i < j \leq k$ entre los $\binom{k}{2}$ posibles pares.

4. Convertir $i$ y $j$ en un único gen y disminuir el tamaño de la muestra en una unidad, $k \to k - 1$.

5. Si $k > 1$ ir al segundo paso, en otro caso parar.

Con los pasos de este algoritmo podemos crear un código en R, como el que se muestra en el Apéndice A.1, para generar un árbol.

Así, a continuación presentamos un árbol de una muestra de 5 genes, generado con este algoritmo, donde en la izquierda vemos el tiempo continuo, y en la derecha la correspondencia en generaciones.

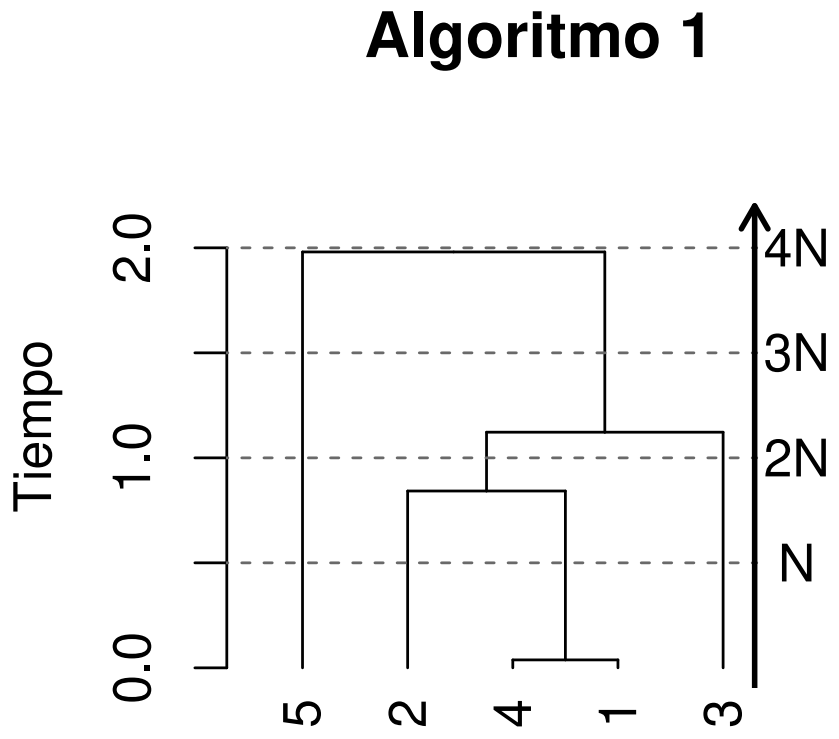



Figura 2.4: Árbol generado a partir del algoritmo 1. A la izquierda tenemos el tiempo continuo y a la derecha las generaciones a las que se corresponde.

### 2.2.3. Modelo alternativo: Modelo de Moran

Ya hemos comentado que la teoría de la coalescencia surge basándose principalmente en el modelo de Wright-Fisher, pero existe otro modelo muy estudiado que también permite la derivación de esta teoría: el modelo de Moran, que fue creado en 1958.

La principal razón de su importancia es que, al contrario del modelo de Wright-Fisher, este sí que considera generaciones que se solapan. Además, desde el punto de vista matemático, muchos resultados obtenidos con exactitud bajo el modelo de Moran, eran simplemente aproximaciones en el de Wright-Fisher.

Consideremos de nuevo, por simplicidad, una población de $2N$ individuos o genes. La nueva generación se formará a partir de la anterior mediante la selección aleatoria de un gen, para dar lugar a otro gen, y de un gen para morir. El gen que muere no puede ser el gen que va a dar lugar a otro nuevo gen y todos los demás sobreviven a la siguiente generación. Cabe destacar que la forma en la que está construido este modelo impide que más de dos genes encuentren a su ancestro común en la anterior generación.

En la Figura 2.5 se percibe mejor la manera en la que se desarrolla el modelo. Los puntos verdes representan genes que mueren y los rosas, genes que dan lugar a otros nuevos.

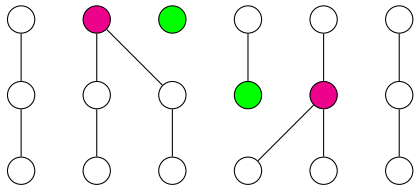


Figura 2.5: Ejemplo del modelo de Moran.

La probabilidad de que dos genes encuentren su ancestro común en la anterior generación es $\frac{1}{N(2N-1)}$. Esto se debe a que tenemos $\binom{2N}{2} = N(2N-1)$ posibles parejas y solamente una de ellas encuentra su ancestro común. Por lo tanto, el tiempo de espera hasta encontrar el MRCA de dos genes, sigue una distribución geométrica de parámetro $\frac{1}{N(2N-1)}$ y, entonces, realizando un razonamiento análogo al que se llevó a cabo para el modelo de Wright-Fisher, se puede estudiar este modelo midiendo el tiempo en unidades de $N(2N-1)$ generaciones.

Vemos, por tanto, que el análisis de la coalescencia podría ser llevado a cabo también con este modelo, de una manera bastante similar a la empleada en el modelo de Wright-Fisher.

## 2.3. Medidas del tamaño de una genealogía

Volviendo al modelo de Wright-Fisher, que es el que hemos analizado con más precisión, como ya sabemos cómo tiene lugar la coalescencia entre genes de una población, vamos a analizar un par de resultados interesantes relacionados con el tamaño y la forma de una genealogía. El tiempo $T_{MRCA}$ hasta que se encuentra el ancestro común más reciente es de claro interés a nivel biológico a la hora de estudiar una población, así como también lo es otra medida, $T_{total}$, que es la longitud total de la genealogía. $T_{total}$ también tiene una gran importancia biológica porque indica el tiempo en el que las mutaciones pueden haber ocurrido en la historia de una muestra.

Por lo tanto estudiemos desde un punto de vista matemático estas dos cantidades, considerando una muestra de $n$ genes: el tiempo hasta el ancestro común más reciente de la muestra completa (que al final es simplemente la altura del árbol), $T_{\text{MRCA}}$, y la longitud total de todas las ramas de la genealogía, $T_{total}$. Como $T_k^c$ es el tiempo en la historia de la muestra en el que hubo exactamente $k$ linajes se tiene:

$$T_{\text{MRCA}} = \sum_{k=2}^{n} T_k^c \tag{2.25}$$

$$T_{total} = \sum_{k=2}^{n} k T_k^c \tag{2.26}$$

En el árbol representado en la Figura 2.4 podemos ver que $T_{\text{MRCA}} \approx 1{,}98$ y que $T_{total} = 2T_2^c + 3T_3^c + 4T_4^c + 5T_5^c \approx 2 \cdot 0{,}86 + 3 \cdot 0{,}28 + 4 \cdot 0{,}80 + 5 \cdot 0{,}04 = 5{,}96$.

Como $T_{\text{MRCA}}$ y $T_{total}$ son funciones de variables aleatorias exponenciales, se pueden obtener sus valores esperados simplemente usando las propiedades básicas de la esperanza y recordando que $T_k^c \sim Exp\left(\binom{k}{2}\right)$:

$$\begin{aligned} E\left[T_{\text{MRCA}}\right] &= \sum_{k=2}^{n} E(T_k^c) = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) \\ &= 2\left(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \cdots + \frac{1}{n-1} - \frac{1}{n}\right) \\ &= 2\left(1 - \frac{1}{n}\right) \xrightarrow[n\to\infty]{} 2 \end{aligned} \tag{2.27}$$

$$E\left[T_{total}\right] = \sum_{k=2}^{n} k E(T_k^c) = \sum_{k=2}^{n} k \frac{2}{k(k-1)} = 2\sum_{k=1}^{n-1} \frac{1}{k} \tag{2.28}$$

Cuando el tamaño de la muestra $n$ tiende a $\infty$, vemos que $E\left[T_{\text{MRCA}}\right]$ tiende a 2, mientras que $E\left[T_{total}\right]$ aumenta sin límite conforme $n$ crece.

De forma parecida a como hicimos con la esperanza, aplicando que los $T_k^c$ son variables aleatorias independientes, calculemos la varianza de $T_{\text{MRCA}}$, teniendo en cuenta que $Var(T_{\text{MRCA}}) = \sum_{k=2}^{n} Var(T_k^c)$.

$$Var(T_{\text{MRCA}}) = \sum_{k=2}^{n} Var(T_k^c) = \sum_{k=2}^{n} \left( \frac{2}{k(k-1)} \right)^2 = 8 \sum_{k=2}^{n} \frac{1}{k^2} - 4 \left( 1 - \frac{1}{n} \right)^2 \tag{2.29}$$

Probemos esta última igualdad por inducción en $n$, siendo $n \geq 2$.

**Paso 1:** Estudiemos la validez de la fórmula para el caso $n = 2$:

$$\left( \frac{2}{2(2-1)} \right)^2 = 1 = 8\frac{1}{2^2} - 4 \left( 1 - \frac{1}{2} \right)^2 \tag{2.30}$$

**Paso 2:** Consideremos que la propiedad se cumple para un determinado $n$ y veamos que se verifica también para $n + 1$.

Así, la hipótesis de inducción será:

$$\sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} = 8 \sum_{k=2}^{n} \frac{1}{k^2} - 4 \left( 1 - \frac{1}{n} \right)^2 \tag{2.31}$$

Veamos qué ocurre para $n + 1$:

$$\begin{aligned}
\sum_{k=2}^{n+1} \frac{4}{k^2(k-1)^2} &= \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} + \frac{4}{(n+1)^2 n^2} = 8 \sum_{k=2}^{n} \frac{1}{k^2} - 4 \left( 1 - \frac{1}{n} \right)^2 + \frac{4}{(n+1)^2 n^2} \\
&= 8 \sum_{k=2}^{n} \frac{1}{k^2} - 4 \left( 1 - \frac{1}{n} \right)^2 + 4 \left( \frac{2}{(n+1)^2} + \frac{-2n^2 + 1}{n^2 (n+1)^2} \right) \\
&= 8 \sum_{k=2}^{n+1} \frac{1}{k^2} + 4 \left[ - \left( \frac{n-1}{n} \right)^2 + \frac{-2n^2 + 1}{n^2 (n+1)^2} \right] \\
&= 8 \sum_{k=2}^{n+1} \frac{1}{k^2} + 4 \left( \frac{-(n-1)^2 (n+1)^2 - 2n^2 + 1}{n^2 (n+1)^2} \right) \\
&= 8 \sum_{k=2}^{n+1} \frac{1}{k^2} + 4 \left( \frac{-n^2}{(n+1)^2} \right) = 8 \sum_{k=2}^{n+1} \frac{1}{k^2} - 4 \left( \frac{n}{n+1} \right)^2 \\
&= 8 \sum_{k=2}^{n+1} \frac{1}{k^2} - 4 \left( 1 - \frac{1}{n+1} \right)^2
\end{aligned} \tag{2.32}$$

Con esto concluimos la prueba y establecemos que la fórmula es válida para todo número entero $n$ mayor o igual que 2.

Pasemos ahora a calcular la varianza de $T_{total} = \sum_{k=2}^{n} kT_k^c$. Para ello definamos una nueva variable $T_k^* = kT_k^c$, que sabemos que también sigue una distribución exponencial pero en este caso con parámetro $\frac{\binom{k}{2}}{k} = \frac{k-1}{2}$. Para establecer esto hemos hecho uso de la propiedad básica de la distribución exponencial dada por la ecuación (1.47): $\left[X \sim Exp(\lambda) \Rightarrow cX \sim Exp\left(\frac{\lambda}{c}\right)\right]$. Por seguir una distribución exponencial se tiene que $Var(T_k^*) = \frac{4}{(k-1)^2}$. Así:

$$Var(T_{total}) = \sum_{k=2}^{n} Var(T_k^*) = 4\sum_{k=2}^{n} \frac{1}{(k-1)^2} = 4\sum_{k'=1}^{n-1} \frac{1}{k'^2} \tag{2.33}$$

En el último paso hemos hecho simplemente un cambio de variable $k' = k - 1$.

En conclusión, tenemos:

$$Var(T_{\text{MRCA}}) = 8\sum_{k=2}^{n} \frac{1}{k^2} - 4\left(1 - \frac{1}{n}\right)^2 \tag{2.34}$$

$$Var(T_{total}) = 4\sum_{k=1}^{n-1} \frac{1}{k^2} \tag{2.35}$$

Veamos a qué tienden estas dos cantidades cuando $n \to \infty$. Para ello, lo primero es tener en cuenta que:

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \tag{2.36}$$

La obtención del valor de esta serie ha sido un problema famoso en la historia de las matemáticas y recibe el nombre de Problema de Basilea. Su prueba se puede ver en [7].

Sabiendo esto y fijándonos en dónde empiezan las sumas es fácil ver que cuando $n \to \infty$, $Var(T_{\text{MRCA}})$ tiende a $\frac{4\pi^2}{3} - 12 \approx 1{,}16$, mientras que $Var(T_{total})$ lo hace a $\frac{2\pi^2}{3} \approx 6{,}58$.

$$\begin{aligned} \lim_{n\to\infty} Var(T_{\text{MRCA}}) &= \lim_{n\to\infty} \left[8\sum_{k=2}^{n} \frac{1}{k^2} - 4\left(1 - \frac{1}{n}\right)^2\right] \\ &= 8\left(\sum_{k=1}^{\infty} \frac{1}{k^2} - 1\right) - 4 = \frac{4\pi^2}{3} - 12 \approx 1{,}16 \end{aligned} \tag{2.37}$$

$$\lim_{n\to\infty} Var(T_{total}) = \lim_{n\to\infty} 4\sum_{k=1}^{n-1} \frac{1}{k^2} = 4\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{2\pi^2}{3} \approx 6{,}58 \tag{2.38}$$

Cabe destacar que la altura del árbol, $T_{\text{MRCA}}$, tiene una varianza bastante grande comparada con la media y esta fracción no se va reduciendo según aumenta el tamaño

de la muestra. En cambio, la longitud del árbol, $T_{total}$, tiene una varianza que se vuelve despreciable, comparándola con la media, según $n$ aumenta. Todo esto se puede observar en la Figura 2.6.

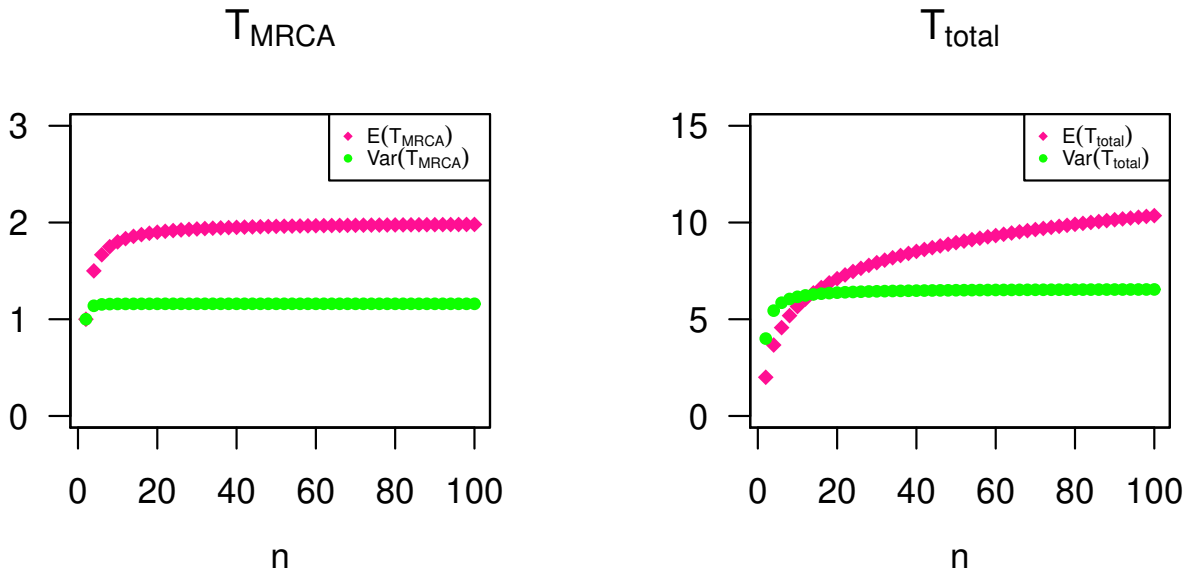



Figura 2.6: Representación de la esperanza y de la varianza para $T_{\text{MRCA}}$ y para $T_{total}$.

El hecho de que haya estas diferencias en la relación entre la esperanza y la varianza de estas dos cantidades se puede explicar por el hecho de que $T_{total}$ acaba siendo dominado por el gran número de ramas pequeñas que hay cerca de los nodos terminales del árbol, mientras que $T_{\text{MRCA}}$ se ve más afectado por las dos ramas largas cercanas al nodo raíz. [8]

Con el algoritmo descrito anteriormente, vamos a generar 6 árboles genealógicos de muestras de 25 genes cada una. En la Figura 2.7 percibimos que la altura de los árboles está, efectivamente, muy ligada a $T_2$ (que hemos resaltado con flechas rosas), pues en la mayoría de los árboles más de la mitad de la altura viene dada por esta cantidad.

Además, fijándonos en los valores numéricos de la altura de los árboles, vemos que hay bastante variación entre un árbol y otro. Esta variación viene cuantificada en la varianza de $T_{\text{MRCA}}$.

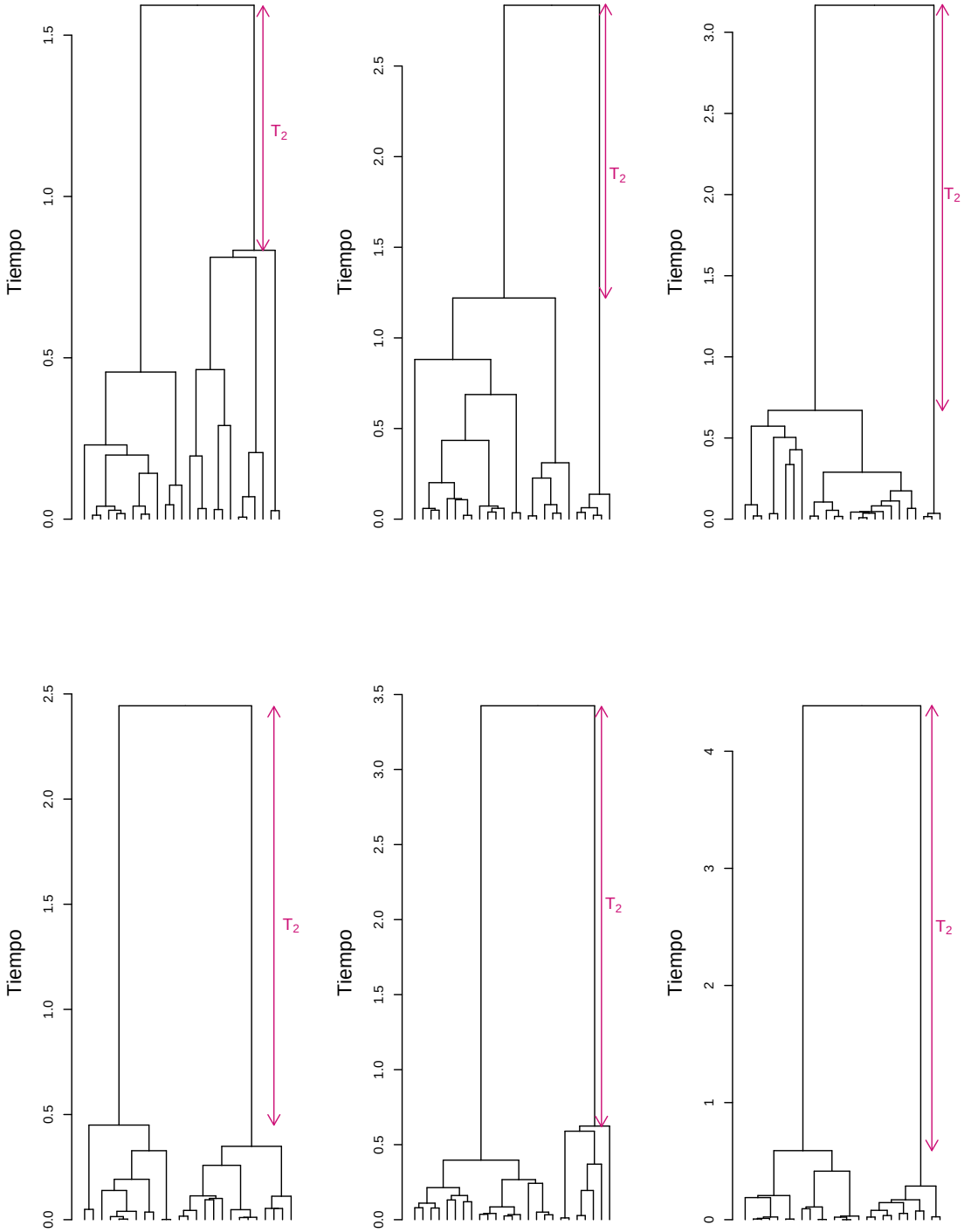



Figura 2.7: Árboles de coalescencia de seis muestras de 25 genes cada una.

El hecho de que los distintos $T_k^c$ sean independientes entre sí, hace que derivar las distribuciones de probabilidad de $T_{\text{MRCA}}$ y de $T_{total}$ no sea muy complicado. La distribución de $T_{\text{MRCA}}$ es simplemente la suma de $n-1$ variables aleatorias exponenciales independientes, $T_k^c$, con parámetros $\binom{k}{2}$, con $2 \leq k \leq n$. Estudiemos así cuál es la distribución de probabilidad de $T_{\text{MRCA}}$.

Nos va a ser de utilidad obtener la función de densidad de la suma de $n$ variables exponenciales con parámetro $\lambda_i$ (con $i = 1, 2, \ldots, n$) independientes. Lo primero que haremos será calcular la función de densidad de la suma de dos de las variables empleando la convolución y después, por inducción, se llega a la función de densidad de la suma buscada. Así:

$$\begin{aligned} f_{X_1+X_2}(x) &= \int_0^x f_{X_1}(s) f_{X_2}(x-s) ds = \int_0^x \lambda_1 e^{-\lambda_1 s} \lambda_2 e^{-\lambda_2 (x-s)} ds \\ &= \frac{\lambda_1}{\lambda_1 - \lambda_2} \lambda_2 e^{-\lambda_2 x} + \frac{\lambda_2}{\lambda_2 - \lambda_1} \lambda_1 e^{-\lambda_1 x} \end{aligned} \tag{2.39}$$

Por lo tanto, para las $n$ variables llegamos a:

$$f_{\sum_{i=1}^n X_i}(x) = \sum_{i=1}^n \lambda_i e^{-\lambda_i x} \prod_{j=1, j\neq i}^n \frac{\lambda_j}{\lambda_j - \lambda_i} \tag{2.40}$$

De esta forma, podemos calcular la función de densidad de $T_{\text{MRCA}} = \sum_{k=2}^n T_k^c$, simplemente aplicando esta ecuación (2.40).

$$f_{T_{\text{MRCA}}}(t_c) = \sum_{i=2}^n \binom{i}{2} e^{-\binom{i}{2} t_c} \prod_{j=2, j\neq i}^n \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \tag{2.41}$$

Realicemos un procedimiento completamente análogo para obtener la función de densidad de $T_{total} = \sum_{k=2}^n T_k^*$, recordando que ya vimos que $T_k^* \sim Exp\left(\frac{k-1}{2}\right)$ .

Así, teniendo todo esto en cuenta junto a la ecuación (2.40) llegamos a que:

$$f_{T_{total}}(t_c) = \sum_{i=2}^n \frac{i-1}{2} e^{-\frac{i-1}{2} t_c} \prod_{j=2, j\neq i}^n \frac{j-1}{j-i} \tag{2.42}$$

Esta ecuación es equivalente a la que sigue:

$$f_{T_{total}}(t_c) = \sum_{i=2}^n (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} e^{-\frac{i-1}{2} t_c} \tag{2.43}$$

Veámoslo, probando que $\prod_{j=2,j\neq i}^{n} \frac{j-1}{j-i} = (-1)^i \binom{n-1}{i-1} = (-1)^i \frac{(n-1)!}{(i-1)!(n-i)!}$:

$$\begin{aligned}\prod_{j=2,j\neq i}^{n} \frac{j-1}{j-i} &= \left(\frac{2-1}{2-i}\right)\left(\frac{3-1}{3-i}\right)\cdots\left(\frac{i-2}{-1}\right)\left(\frac{i}{1}\right)\left(\frac{i+1}{2}\right)\cdots\left(\frac{n-1}{n-i}\right) \\ &= (-1)^i \left(\frac{i}{1}\right)\left(\frac{i+1}{2}\right)\left(\frac{i+2}{3}\right)\cdots\left(\frac{n-1}{n-i}\right) \\ &= (-1)^i \frac{(n-1)!}{(i-1)!(n-i)!} \end{aligned} \tag{2.44}$$

Otra fórmula equivalente es la que se muestra a continuación:

$$f_{T_{total}}(t_c) = \frac{n-1}{2} e^{-\frac{t_c}{2}} (1 - e^{-\frac{t_c}{2}})^{n-2} \tag{2.45}$$

Veamos que se tiene la equivalencia:

$$\begin{aligned} f_{T_{total}}(t_c) &= \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} e^{-\frac{i-1}{2}t_c} = \frac{e^{\frac{t_c}{2}}}{2} \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} (i-1) e^{-\frac{i}{2}t_c} \\ &= \frac{(n-1)e^{\frac{t_c}{2}}}{2} \sum_{i=2}^{n} (-1)^i \binom{n-2}{i-2} e^{-\frac{i}{2}t_c} = \frac{(n-1)e^{\frac{t_c}{2}}}{2} \sum_{j=0}^{n-2} (-1)^j \binom{n-2}{j} e^{-\frac{j+2}{2}t_c} \\ &= \frac{(n-1)e^{\frac{t_c}{2}}e^{-t}}{2} \sum_{j=0}^{n-2} \binom{n-2}{j} \left(-e^{-\frac{1}{2}t_c}\right)^j = \frac{n-1}{2} e^{-\frac{t_c}{2}} (1 - e^{-\frac{t_c}{2}})^{n-2} \end{aligned} \tag{2.46}$$

En el último paso hemos usado el teorema del binomio: $(p+q)^m = \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j}$, considerando que $p = -e^{-\frac{1}{2}t_c}$, $q = 1$ y $m = n-2$.

# Capítulo 3

# Coalescencia en el modelo de Wright-Fisher con mutaciones

## 3.1. Modelo de Wright-Fisher con mutaciones

En el análisis de datos genéticos reales es esencial tener en cuenta las mutaciones que pueden experimentar los genes. Consideremos entonces, en el modelo de Wright-Fisher explicado en el capítulo anterior, la posibilidad de mutación, es decir, cada gen que se reproduce es susceptible de pasar un proceso de mutación con probabilidad $u$. Por lo tanto, tenemos que un gen puede ser copiado a su descendencia sin cambios con probabilidad $1 - u$ y, con probabilidad $u$, puede ser modificado por una mutación. Cabe destacar que según el tipo de datos que estemos considerando $u$ hace referencia a la tasa de mutación por generación, por locus (posición fija en un cromosoma) o por sitio (ubicación en una secuencia de ADN).

En la Figura 3.1 vemos tres generaciones del modelo de Wright-Fisher con mutación. En la segunda fila, es decir, en la segunda generación, vemos dos genes que son copias mutadas de genes de la anterior generación. En la generación del presente también se observa un gen que es una copia modificada del de la generación previa.

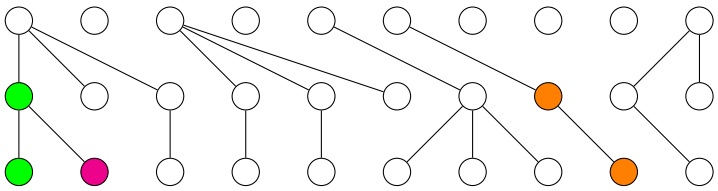


Figura 3.1: Modelo de Wright-Fisher con mutaciones.

Siguiendo un determinado linaje vemos que hay una probabilidad $u$ de que la clase del gen padre en la generación $t$ sea distinta de la clase del gen hijo en la generación $t+1$. Por lo tanto, el número de generaciones que pasan (comenzando desde el presente) hasta que tiene lugar la primera mutación, que denotaremos por $T_M$, sigue una distribución geométrica de parámetro $u$. Así, la probabilidad de que un linaje experimente su primera mutación $t$ generaciones hacia el pasado es:

$$P(T_M = t) = u(1-u)^{t-1} \tag{3.1}$$

Considerando, como ya se hizo para el modelo de Wright-Fisher sin mutación, que el tiempo es medido en unidades de $2N$ generaciones, y denotando a este tiempo hasta que se tiene la primera mutación como $T_M^c$, tenemos que, para valores grandes de $2N$, podemos aproximar la distribución de $T_M^c$ por una exponencial. Para ello recordemos que por seguir $T_M$ una distribución geométrica se tiene que $P(T_M > t) = (1-u)^t$. Así:

$$\begin{aligned} \lim_{2N\to\infty} P\left(T_M > t\right) &= \lim_{2N\to\infty} P\left(\frac{T_M}{2N} > t_c\right) = \lim_{2N\to\infty} \left(1 - \frac{2Nu}{2N}\right)^{t_c 2N} \\ &= 1 - e^{-\theta t_c/2} = P\left(T_M^c > t_c\right) \end{aligned} \tag{3.2}$$

En esta ecuación hemos definido dos nuevas variables: $t_c = \frac{t}{2N}$, que describe el tiempo en unidades de $2N$ generaciones, y $\theta = 4Nu$, que es el llamado parámetro de mutación o tasa poblacional de mutación. Este parámetro se puede interpretar como el número esperado de mutaciones que tienen lugar en dos linajes distintos antes de que encuentren su ancestro común. Recordemos que el tiempo esperado para que tenga lugar una coalescencia entre dos genes es $2N$ generaciones y entonces, en ese tiempo, es de esperar que tengan lugar $2Nu$ mutaciones en cada linaje, es decir, $4Nu$ mutaciones en total. Vemos que $T_M^c$ sigue una distribución exponencial de parámetro $\frac{\theta}{2}$.

A lo largo de este capítulo consideraremos que estamos trabajando con poblaciones de tamaño grande, de forma que podremos considerar la distribución exponencial del tiempo de espera hasta la primera mutación.

Por seguir $T_M^c$ esta distribución exponencial de parámetro $\frac{\theta}{2}$, el número $M$ de mutaciones que tienen lugar en un intervalo de tiempo de duración $t_c$ sigue una distribución de Poisson de parámetro $\frac{\theta t_c}{2}$. La formalización de esta afirmación, relacionada con los procesos de Poisson, se puede encontrar en la página 297 del libro *Probability and Measure* de Patrick Billingsley [9].

Así, la probabilidad de que tengan lugar $m$ mutaciones en un tiempo $t_c$ viene dada por:

$$P(M = m|T_M^c = t_c) = \frac{\left(\frac{\theta t_c}{2}\right)^m}{m!} e^{-\frac{\theta t_c}{2}}, \qquad \text{con } m = 0, 1, 2, \ldots \tag{3.3}$$

Además:

$$E[M|T_M^c = t_c] = Var(M|T_M^c = t_c) = \frac{\theta t_c}{2} \tag{3.4}$$

Debemos tener en cuenta que estas mutaciones que estamos estudiando no aportan ningún tipo de ventaja o desventaja a los linajes, son mutaciones que reciben el nombre de neutras porque no alteran los patrones de reproductividad que hay en una población y, por tanto, son independientes del proceso genealógico.

Para estudiar este tipo de mutaciones se han creado varios modelos matemáticos, pero nosotros nos vamos a centrar en uno de ellos: en el modelo de sitios infinitos.

### 3.1.1. Modelo de sitios infinitos

Este modelo apareció en 1969 de la mano del biólogo y matemático japonés Kimura. En él se considera que los genes son simplemente secuencias de ADN. Además, se asume que las mutaciones tienen lugar siempre en una nueva posición (considerando posición como cada uno de los nucleótidos presentes en una secuencia de ADN), es decir, las mutaciones tienen lugar siempre en un nucleótido distinto.

Se puede usar este modelo para describir la evolución de las cadenas de ADN muy largas que presentan una baja tasa de mutación en cada posición. Desde el punto de vista biológico esta baja tasa se justifica teniendo en cuenta que, en general, el número de sitios que varían en una muestra de secuencias reales suele ser bastante menor que el número de sitios que son idénticos en todas las secuencias.

En el modelo de sitios infinitos siempre habrá uno o dos posibles alelos en una posición de un conjunto de secuencias, pero nunca más, porque cada posición muta como mucho una vez.

El modelo también establece que todas las mutaciones que ocurren en algún momento de la historia de la muestra pueden ser recuperadas pues, como solamente puede haber un cambio en un nucleótido específico, si ese cambio tiene lugar se podrá percibir en todo momento.

Para entender mejor este modelo se añade la Figura 3.2 que nos muestra el árbol que indica el proceso evolutivo de una secuencia de ADN. Vemos que a lo largo del tiempo tienen lugar cuatro mutaciones (marcadas con puntos negros). Como cada una de ellas ocurre en un nucleótido diferente, vamos marcando con puntos rosas las posiciones diferenciadoras o segregadoras de los nucleótidos donde tienen lugar. Así, tenemos tantas posiciones diferenciadoras como mutaciones hay en la historia de la secuencia.

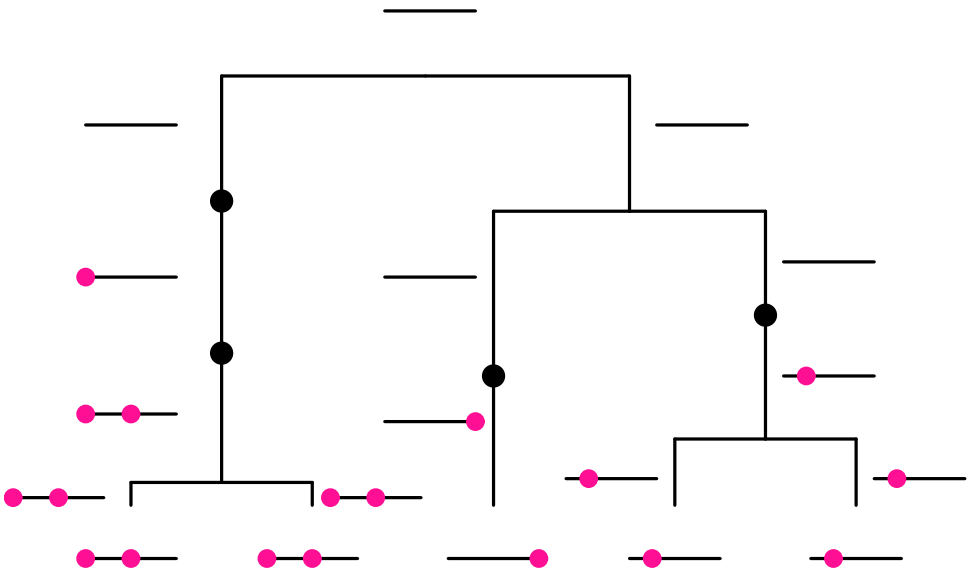


Figura 3.2: Ejemplo del modelo de sitios infinitos.

## 3.2. Mutaciones y coalescencia

Ahora que ya conocemos cómo se pueden introducir las mutaciones en el modelo de Wright-Fisher y que tenemos un modelo para describir el proceso de mutación, como es el modelo de sitios infinitos, vamos a relacionar el proceso de mutación con la coalescencia.

Tenemos que el tiempo $T_M^c$ de espera hasta que se tiene una mutación en un linaje dado, sigue una distribución exponencial de parámetro $\frac{\theta}{2}$. Si tenemos $k$ linajes, en cada uno de ellos podremos considerar el tiempo de espera hasta la primera mutación, que en todos los casos sigue una distribución exponencial de parámetro $\frac{\theta}{2}$ y son independientes los unos de los otros. En esta situación podemos calcular el tiempo de espera hasta que se tiene una mutación en alguno de los $k$ linajes ($T_{M_k}^c$), que no será más que el mínimo de los tiempos de espera de cada linaje por separado.

Ya hemos visto que dadas dos variables aleatorias $X$ e $Y$ que siguen una distribución exponencial de parámetros $\lambda$ y $\lambda'$, respectivamente, entonces $\min(X, Y) \sim Exp(\lambda + \lambda')$. Así, en este caso $T_{M_k}^c \sim Exp\left(\frac{k\theta}{2}\right)$.

Recordemos además que el tiempo de espera hasta que dos de los $k$ linajes considerados encuentran a su ancestro común, $T_k^c$, sigue una distribución exponencial de parámetro $\binom{k}{2} = \frac{k(k-1)}{2}$.

El tiempo de espera para que ocurra alguno de estos dos eventos (o de coalescencia o

de mutación), es una variable aleatoria que es el mínimo de dos variables exponenciales independientes de parámetros $\frac{k\theta}{2}$ y $\frac{k(k-1)}{2}$. Así este tiempo de espera sigue una distribución exponencial de parámetro $\frac{k\theta}{2} + \frac{k(k-1)}{2} = \frac{k(k-1+\theta)}{2}$

Ahora que ya sabemos cómo es la distribución del tiempo de espera hasta un evento de coalescencia, hasta un evento de mutación y hasta un evento de cualquiera de los dos tipos, podemos calcular la probabilidad de que el primer evento que tenga lugar en una muestra con $k$ linajes sea de coalescencia o de que sea de mutación.

En general, podemos calcular la probabilidad de que un primer evento sea de un determinado tipo considerando que tenemos dos variables exponenciales de parámetros $\lambda_1$ y $\lambda_2$. Busquemos entonces la probabilidad de que el primer evento sea un proceso de parámetro $\lambda_1$, es decir, la probabilidad de que el tiempo $T_1$ a un evento exponencial ($\lambda_1$) sea menor que el tiempo $T_2$ a un evento exponencial ($\lambda_2$).

Ya hemos visto en la ecuación (1.51) que dadas dos variables exponenciales $X$ e $Y$ de parámetros $\lambda$ y $\lambda'$ respectivamente, se tiene $P(X < Y) = \frac{\lambda}{\lambda+\lambda'}$. Por lo tanto en esta situación tenemos:

$$P(T_1 < T_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \tag{3.5}$$

Vemos que la probabilidad de que el primer evento sea de un tipo determinado, viene dada por la tasa relativa de ese evento.

Aplicando esto a nuestro caso, en el que estamos considerando procesos de coalescencia o de mutación, tenemos que la probabilidad de que un evento de coalescencia sea el primero en ocurrir viene dado por:

$$P(\text{coalescencia} \mid \text{coalescencia o mutación}) = \frac{\frac{k(k-1)}{2}}{\frac{k\theta}{2} + \frac{k(k-1)}{2}} = \frac{k-1}{\theta + k - 1} \tag{3.6}$$

mientras que la probabilidad de que una mutación sea el primer evento que ocurra es:

$$P(\text{mutación} \mid \text{coalescencia o mutación}) = \frac{\theta}{\theta + k - 1} \tag{3.7}$$

Calculemos ahora la distribución del número de eventos que tienen lugar hasta el primer evento de coalescencia (incluyéndolo) entre $k$ linajes.

Escribámoslo en forma general, considerando que estamos ante dos procesos descritos por variables exponenciales con parámetros $\lambda_1$ y $\lambda_2$ y sabiendo que la probabilidad de que el siguiente evento sea de "tipo 1" es $\frac{\lambda_1}{\lambda_1+\lambda_2}$ y la de que sea de "tipo 2" es $\frac{\lambda_2}{\lambda_1+\lambda_2}$. Como los tiempos de espera vienen descritos por variables exponenciales y esta distribución ya hemos visto que carece de memoria, entonces la probabilidad de que suceda un evento de

un determinado tipo, una vez ha ocurrido otro, viene dada por las mismas fórmulas que las obtenidas al estudiar cuál es el primer evento.

Por lo tanto, los eventos forman una serie de intentos de Bernoulli con probabilidad de éxito $\frac{\lambda_1}{\lambda_1+\lambda_2}$ pues estamos interesados en encontrar el primer evento de tipo 1. Por lo tanto, el número de eventos $B$ que han tenido lugar cuando el primer evento de tipo 1 ocurre, sigue una distribución geométrica:

$$P(B=b)=\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{b-1}\frac{\lambda_1}{\lambda_1+\lambda_2} \tag{3.8}$$

Por lo tanto, la distribución del número de eventos hasta la primera coalescencia (incluyéndola) entre $k$ linajes sigue una distribución geométrica de parámetro $\frac{k-1}{\theta+k-1}$. Es decir, tenemos:

$$P(B=b)=\left(\frac{\theta}{\theta+k-1}\right)^{b-1}\frac{k-1}{\theta+k-1} \tag{3.9}$$

## 3.3. Algoritmos de generación de genealogías

Como ya conocemos las características de los eventos de mutación y de coalescencia, vamos a generar un par de algoritmos que nos permitan simular genealogías en las que tengan lugar estos dos procesos.

Comencemos con el que denotaremos por algoritmo 2.

**Algoritmo 2**

1. Empezar con $k=n$ genes, siendo $n$ el tamaño de la muestra.
2. Considerar una variable exponencial con parámetro $\frac{k(k-1+\theta)}{2}$ que determinará cuándo ocurre un evento.
3. Con probabilidad $\frac{k-1}{k-1+\theta}$ el evento es un evento de coalescencia y con probabilidad $\frac{\theta}{k-1+\theta}$ es un evento de mutación.
4. Si ocurre un evento de coalescencia, escoger aleatoriamente un par de genes para encontrar el ancestro común. Actualizar $k$: $k \to k-1$.
5. Si ocurre un evento de mutación, escoger un linaje para mutar. Dejar el número $k$ invariante.
6. Continuar hasta que $k$ es 1.

Este algoritmo es una extensión del algoritmo 1 que ya hemos visto. Para determinar las características de los genes que hay en el presente, se parte del ancestro común y se van viendo las mutaciones que experimenta hasta llegar a cada uno de los genes que hay en la actualidad. Como estamos considerando el modelo de sitios infinitos, cada vez que

hay una mutación, el gen, que recordemos que se interpreta como una secuencia de ADN, experimenta una modificación en uno de sus nucleótidos.

A continuación, en la Figura 3.3, mostramos un ejemplo de una genealogía de una muestra de 4 genes generada con este algoritmo con ayuda de R (ver código en el Apéndice A.2). En ella vemos un evento de mutación y tres eventos de coalescencia. En la derecha vemos la probabilidad que había de que ocurriese un evento del tipo indicado,

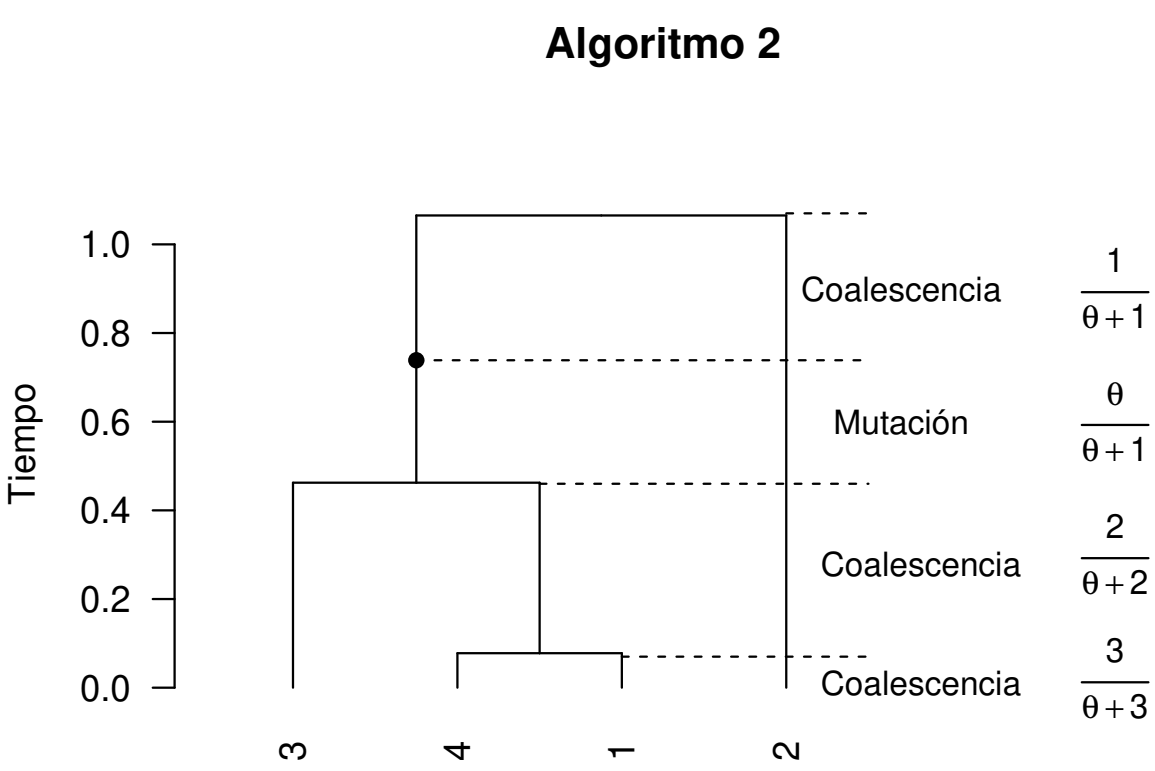



Figura 3.3: Árbol genealógico diseñado con el algoritmo 2. En la derecha se muestra el tipo de evento y la probabilidad de que ocurra en cada caso.

Pasemos ahora a otro algoritmo, el algoritmo 3, que genera genealogías similares a las que acabamos de ver pero el procedimiento seguido es ligeramente distinto.

**Algoritmo 3**

1. Simular la genealogía de $n$ genes de acuerdo con el proceso de coalescencia con parámetro $\binom{k}{2}$, siendo $k$ el número de linajes considerados en cada etapa. Esta simulación se puede llevar a cabo con el Algoritmo 1.

2. Para cada rama considerar el número de mutaciones, $M$, que viene dado por una distribución de Poisson de parámetro $\frac{t_c\theta}{2}$, con $t_c$ la longitud de la rama.

3. Para cada rama los tiempos a los que ocurren los $M$ eventos de mutación son escogidos aleatoriamente.

Este algoritmo se basa en el hecho de que el número de mutaciones tiene distribución de Poisson. Considera además que las mutaciones pueden ser introducidas en la genealogía una vez que esta ya está generada. Esto se debe al hecho de que las mutaciones que estamos considerando son neutras y no afectan al patrón reproductivo.

A continuación, en la Figura 3.4, presentamos un ejemplo de este algoritmo generado con el código creado en R que se puede ver en el Apéndice A.3. Además, resaltamos la idea de que inicialmente se genera la genealogía sin considerar los procesos de mutación y después simplemente se van añadiendo las mutaciones a cada una de las ramas considerando su distribución de Poisson asociada.

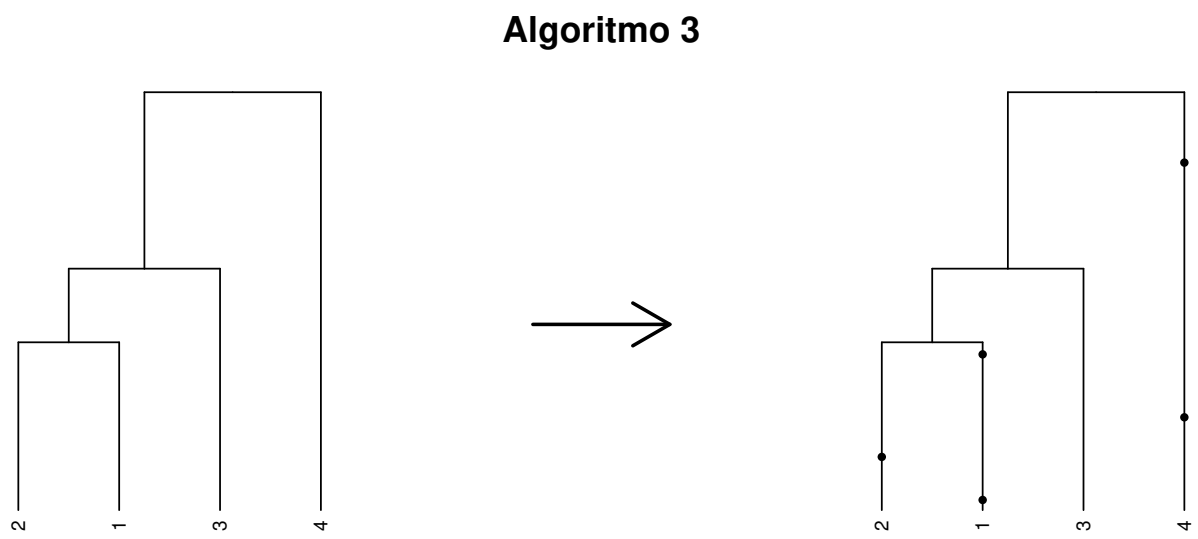



Figura 3.4: Árbol genealógico diseñado con el algoritmo 3. En la izquierda presentamos la genealogía inicial, sin considerar el proceso de mutación y en la derecha tenemos ya introducidas las mutaciones.

## 3.4. Medidas de polimorfismos en una secuencia de ADN

Ahora que ya sabemos cómo se pueden modelizar las mutaciones en una genealogía, vamos a emplear el modelo de sitios infinitos para estudiar un concepto esencial en genética: los *polimorfismos en una secuencia de ADN*. Antes de nada debemos definir qué es un polimorfismo. Los polimorfismos en el ADN son las diferentes secuencias de ADN entre individuos, grupos o poblaciones. Incluyen distintos niveles de variación, desde un único cambio en la base nitrogenada de un nucleótido, cambios en varias bases o cambios en secuencias. Aquí vamos a considerar los polimorfismos de un único nucleótido (SNP), es decir, aquellos en los que se produce una única variación en la base del nucleótido. [10]

Una de las formas más simples de estudiar los polimorfismos del ADN es obteniendo el número de sitios segregadores de una muestra, que denotaremos por $S$. Los sitios segregadores no son más que las posiciones de los nucleótidos en los que se produce la mutación. Para comprender de formas más visual lo que son los sitios segregadores, presentamos en la Figura 3.5 cinco secuencias en las que se observan en color tres sitios segregadores, es decir, $S = 3$.

Secuencia 1: A C G C T A G T C A
Secuencia 2: A G G C T A G T C A
Secuencia 3: A G G C T A G T C T
Secuencia 4: A G G C A A G T C T
Secuencia 5: A C G C A A G T C T

Figura 3.5: Secuencias de ADN con los sitios segregadores señalados.

Recordemos que en el modelo de sitios infinitos, cada mutación ocurre en una única posición, por lo que toda mutación que ocurra en la evolución de la muestra, será un sitio segregador.

Así, el número $S$ de sitios segregadores en una muestra de tamaño $n$ es igual al número de mutaciones en la evolución de la muestra en el modelo de sitios infinitos. Tenemos que tener en cuenta que estamos considerando una genealogía de longitud total $T_{total}$. Entonces, el número de mutaciones en una genealogía de esta longitud sigue una distribución de Poisson de parámetro $\frac{\theta T_{total}}{2}$.

Como conocemos la función de densidad de $T_{total}$:

$$f_{T_{total}}(t_c) = \sum_{k=2}^{n} (-1)^k \binom{n-1}{k-1} \frac{k-1}{2} e^{-\frac{k-1}{2} t_c} \tag{3.10}$$

y la probabilidad de que tengan lugar $m$ mutaciones en un tiempo $T_{total}$:

$$P(M = m|T_{total}) = \frac{\left(\frac{\theta T_{total}}{2}\right)^m}{m!} e^{-\frac{\theta T_{total}}{2}}, \qquad \text{con } m = 0, 1, 2, \ldots \tag{3.11}$$

podemos calcular la distribución de $S$.

Para ello, tendremos en cuenta que dadas dos variables aleatorias $X_1$ y $X_2$, su función de densidad conjunta es $f_{X_1,X_2}(x_1, x_2) = f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2)$. La función de densidad marginal de $X_1$ viene dada por:

$$\begin{aligned} f_{X_1}(x_1) &= \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)dx_2 \\ &= \int_{-\infty}^{\infty} f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2)dx_2 \end{aligned} \tag{3.12}$$

Así, en este caso considerando que $X_1$ es el número de sitios segregadores $S$ y que $X_2$ es la longitud total de la genealogía $T_{total}$, tenemos:

$$\begin{aligned} P(S = s) &= \int_0^{\infty} P(S = s|T^c = t_c) f_{T_{total}}(t_c)dt_c \\ &= \int_0^{\infty} \frac{\left(\frac{\theta t_c}{2}\right)^s}{s!} e^{-\frac{\theta t_c}{2}} \sum_{k=2}^{n} (-1)^k \binom{n-1}{k-1} \frac{k-1}{2} e^{-\frac{k-1}{2}t_c} dt_c \\ &= \left(\frac{\theta}{2}\right)^s \sum_{k=2}^{n} (-1)^k \binom{n-1}{k-1} \frac{k-1}{2} \int_0^{\infty} \frac{t^s e^{-\frac{\theta+k-1}{2}t_c}}{s!} dt_c \\ &= \left(\frac{\theta}{2}\right)^s \sum_{k=2}^{n} (-1)^k \binom{n-1}{k-1} \frac{k-1}{2} \left(\frac{2}{\theta+k-1}\right)^{s+1} \\ &= \sum_{k=2}^{n} (-1)^k \binom{n-1}{k-1} \frac{k-1}{\theta+k-1} \left(\frac{\theta}{\theta+k-1}\right)^s \end{aligned} \tag{3.13}$$

En el paso de la tercera línea a la cuarta hemos empleado el cálculo de la función gamma $\Gamma(n) = \int_0^{\infty} x^{n-1}e^{-x} = (n-1)!$ donde $n$ es un número entero. En nuestro caso hemos hecho el cambio de variable $x = \frac{\theta+k-1}{2}t_c$ y hemos resuelto la integral usando la expresión de la función gamma para $n = s + 1$.

A continuación presentamos un diagrama de barras en 3D de la probabilidad $P(S = s)$ según el número $n$ de secuencias en la muestra, para una población con parámetro de mutación $\theta = 3$ (este diagrama fue generado con R y podemos ver el código en el Apéndice A.4).

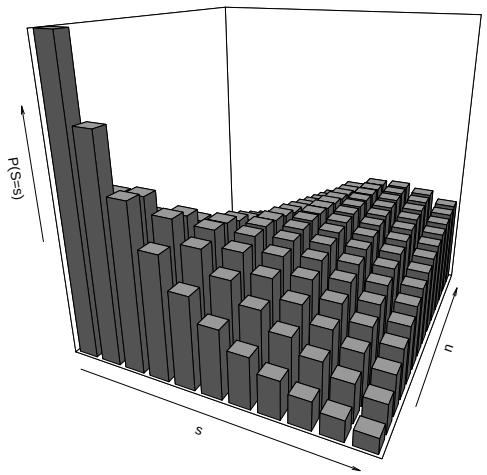



Figura 3.6: Diagrama de barras de la función de probabilidad para el número de sitios segregadores en una muestra de $n$ secuencias con parámetro de mutación $\theta = 3$.

En la Figura 3.7 vemos cuatro gráficos para 4 valores del parámetro de mutación $\theta$. En cada uno de ellos se representa la probabilidad de que el número de sitios segregadores tome un determinado valor para 4 valores distintos del tamaño de la muestra $n$.

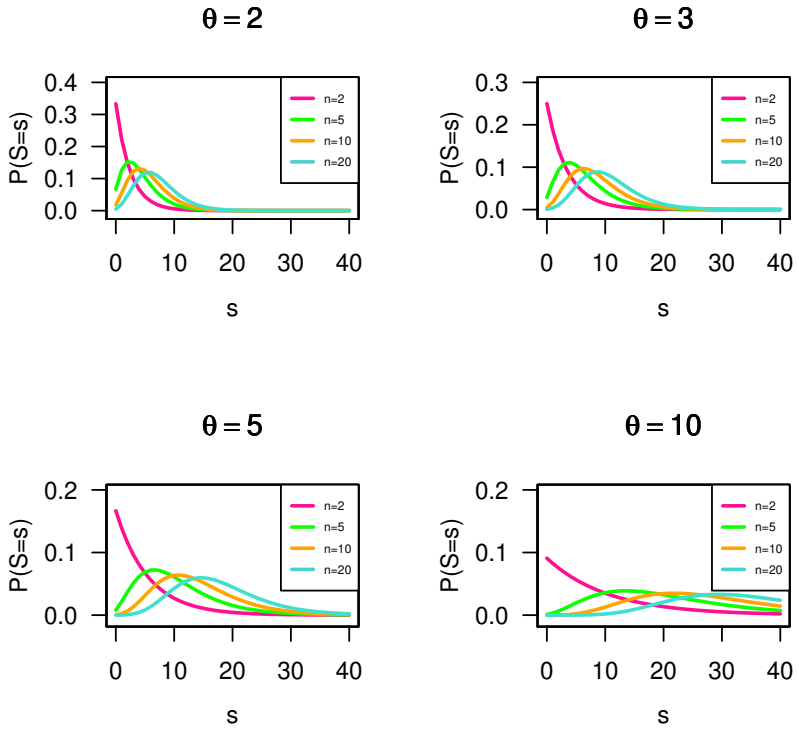



Figura 3.7: Gráficas de la probabilidad de que el número de sitios segregadores tome un determinado valor para distintos valores de $\theta$ y de $n$.

Tanto en estas gráficas como en el histograma vemos que para valores pequeños de $n$, la distribución de $S$ tiene una forma parecida a una L, mientras que según $n$ crece va adquiriendo una forma particular, según el valor de $\theta$, con un máximo en un valor de $s$ mayor que 0.

Para una muestra con $n = 2$ tenemos que la ecuación (3.13) se reduce a una distribución geométrica de parámetro $p = \frac{1}{\theta+1}$:

$$P(S = s) = \frac{1}{\theta + 1}\left(\frac{\theta}{\theta + 1}\right)^s \tag{3.14}$$

Otra forma de obtener las probabilidades de $S$, sería teniendo en cuenta que el número de eventos $B$ hasta la primera coalescencia (incluyéndola) entre $k$ linajes tiene distribución geométrica. Recordemos que:

$$P(B = b) = \left(\frac{\theta}{\theta + k - 1}\right)^{b-1} \frac{k-1}{\theta + k - 1} \tag{3.15}$$

De esta forma tenemos que la distribución del número de sitios segregadores generados por mutaciones que tienen lugar durante el tiempo en el que hay $k$ linajes en la muestra $(S_k)$, viene dada por:

$$P(S_k = s_k) = \left(\frac{k-1}{\theta + k - 1}\right)\left(\frac{\theta}{\theta + k - 1}\right)^{s_k} \tag{3.16}$$

Como $S = \sum_{k=2}^{n} S_k$, se podría obtener a partir de esto $P(S = s)$ haciendo la convolución de los $S_k$. Esta forma fue la que empleó Watterson en 1975 para obtener esta cantidad [11].

Ahora que ya sabemos cómo obtener la distribución del número de sitios segregadores $S$, vamos a estudiar algunas de sus características esenciales. En particular, obtendremos su valor esperado y su varianza, condicionando $S$ en la longitud total del árbol $T_{total}$.

Como ya hemos mencionado, el número de mutaciones en la historia de una muestra coincide con el número de sitios segregadores $S$ en el modelo de sitios infinitos. El número de mutaciones en un intervalo de tiempo $t_c$ seguirá una distribución de Poisson de parámetro $\frac{t_c\theta}{2}$. Por lo tanto, podemos calcular la esperanza y la varianza de $S$ condicionando su valor a la longitud total del árbol $T_{total}$. Así:

$$
\begin{aligned}
E[S] &= \sum_{s=0}^{\infty} sP(S=s) = \sum_{s=0}^{\infty} s \int_0^{\infty} P(S=s|T_{total}=t_c) f_{T_{total}}(t_c) dt_c \\
&= \int_0^{\infty} \sum_{s=0}^{\infty} sP(S=s|T_{total}=t_c) f_{T_{total}}(t_c) dt_c \\
&= \int_0^{\infty} E[S|T_{total}=t_c] f_{T_{total}}(t_c) dt_c \\
&= \frac{\theta}{2} \int_0^{\infty} t_c f_{T_{total}}(t_c) dt_c = \frac{\theta}{2} E[T_{total}] \\
&= \frac{\theta}{2} \left( 2 \sum_{k=1}^{n-1} \frac{1}{k} \right) = \theta \sum_{k=1}^{n-1} \frac{1}{k}
\end{aligned} \tag{3.17}
$$

Para calcular la varianza, obtengamos primero el valor de $E[S^2]$:

$$
\begin{aligned}
E[S^2] &= \sum_{s=0}^{\infty} s^2 P(S=s) = \sum_{s=0}^{\infty} s^2 \int_0^{\infty} P(S=s|T_{total}=t_c) f_{T_{total}}(t_c) dt_c \\
&= \int_0^{\infty} \sum_{s=0}^{\infty} s^2 P(S=s|T_{total}=t_c) f_{T_{total}}(t_c) dt_c \\
&= \int_0^{\infty} E[S^2|T_{total}=t_c] f_{T_{total}}(t_c) dt_c \\
&= \int_0^{\infty} \left( \frac{\theta t_c}{2} + \frac{\theta^2 t_c^2}{4} \right) f_{T_{total}}(t_c) dt_c = \frac{\theta}{2} E[T_{total}] + \frac{\theta^2}{4} E[T_{total}^2] \\
&= \frac{\theta}{2} E[T_{total}] + \frac{\theta^2}{4} \left( Var(T_{total}) + E[T_{total}]^2 \right)
\end{aligned} \tag{3.18}
$$

Así, sustituyendo:

$$
\begin{aligned}
Var(S) &= E[S^2] - E[S]^2 = \frac{\theta}{2} E[T_{total}] + \frac{\theta^2}{4} \left( Var(T_{total}) + E[T_{total}]^2 \right) - \frac{\theta^2}{4} E[T_{total}]^2 \\
&= \frac{\theta}{2} E[T_{total}] + \frac{\theta^2}{4} Var(T_{total}) \\
&= \frac{\theta}{2} \left( 2 \sum_{k=1}^{n-1} \frac{1}{k} \right) + \frac{\theta^2}{4} \left( 4 \sum_{k=1}^{n-1} \frac{1}{k^2} \right) \\
&= \theta \sum_{k=1}^{n-1} \frac{1}{k} + \theta^2 \sum_{k=1}^{n-1} \frac{1}{k^2}
\end{aligned} \tag{3.19}
$$

En el gráfico de la Figura 3.8 vemos representadas estas dos cantidades frente al tamaño de la muestra $n$ considerando que estamos ante una población con $\theta = 3$.

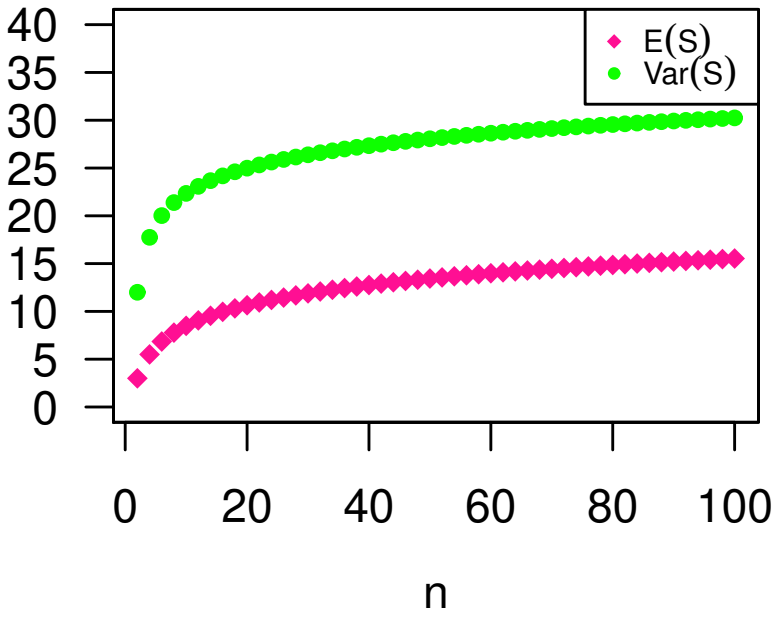



Figura 3.8: Gráfica en la que se representa la esperanza y la varianza de $S$ frente a $n$ para una población con $\theta = 3$.

# Capítulo 4

# Aplicaciones de la teoría de la coalescencia

La teoría de la coalescencia se puede aplicar a un gran número de problemas actuales y a muestras de datos genéticos reales. Un ejemplo que se encuentra a la orden del día es el de encontrar el ancestro común más reciente de varias variantes de coronavirus. Un estudio de este tipo fue llevado a cabo en Francia [12], usando algunos conceptos de la teoría de la coalescencia que incluyen razonamientos más complicados que los que hemos tratado en este trabajo, como poblaciones con un tamaño que no se mantiene constante en el tiempo. Pese a que este estudio y otros estudios actuales no podamos presentarlos en esta memoria por el uso de herramientas mucho más avanzadas y sofisticadas, lo que sí vamos a hacer es finalizar este trabajo con un ejemplo de la teoría de la coalescencia aplicada a la relación entre humanos y neandertales que despertó un gran interés en el año 1998 y que hoy en día sigue estudiándose.

En el año 1997 Krings *et al.* [13] analizaron la primera muestra de material genético procedente de un neandertal, más específicamente analizaron una secuencia de ADN mitocondrial. Los neandertales son una especie extinta del género *Homo* que se sabe que ha coexistido con los humanos hasta hace unos 30000 años. Krings comparó esta secuencia con 986 secuencias de ADN mitocondrial de humanos actuales, estudiando las diferencias que había entre ellas. Haciendo esto estableció que el ancestro común de las 986 muestras de ADN del ser humano actual y de la muestra del neandertal se encuentra en un tiempo $T_r$ que es 4 veces mayor que el tiempo $T_e$ que tardan todas las muestras consideradas del humano moderno en encontrar su ancestro común.

Tras esta publicación, en 1998, Nordborg [14] decidió emplear los fundamentos de la teoría de la coalescencia para estudiar la relación que tienen los seres humanos actuales y los neandertales. La muestra neandertal se considera que fue datada en un tiempo $t_s$ que oscila entre los 30000 y los 100000 años. Como el ADN mitocondrial se transmite de madres a hijos, asumiendo que la población de mujeres tenía un tamaño de 3400 y que cada generación tiene una vida de unos 20 años, entonces $t_s$ toma valores entre 0,44 y 1,47 en la escala de tiempo de coalescencia. Debemos mencionar que en todo este apartado vamos a considerar que el tiempo es continuo y trabajaremos con el formalismo continuo de la coalescencia.

Lo que realmente quería analizar Nordborg con su estudio era la hipótesis nula de que, cuando los seres humanos y los neandertales coexistían, había emparejamiento aleatorio entre ellos.

Dos de los factores que Krings recalcó y que hicieron a Nordborg creer que esta hipótesis era falsa eran, por un lado que $T_r$ fuese 4 veces mayor que $T_e$, y por otro lado, que la estructura del árbol fuese tal que el neandertal no se juntase con ningún linaje humano hasta que solamente quedaba uno. En un principio Nordborg se planteó la idea de que esta forma del árbol ya permitiese rechazar la hipótesis nula, pero observó que no era una condición suficiente pues si el neandertal y el último linaje humano encuentran su ancestro común en un tiempo muy pequeño no podríamos concluir que no hubiese emparejamiento aleatorio entre ellos. Fue por esto por lo que se planteó considerar también el hecho de que $T_r$ fuese mayor o igual que $4T_e$.

Así, con todo esto, Nordborg consideró como p-valor para analizar esta hipótesis nula, la probabilidad de que se diese un modelo tan extremo o más que el establecido por Krings, es decir, la probabilidad de que se diese un árbol con la forma establecida anteriormente y de que $T_r \geq 4T_e$, lo que escribiremos como:

$$P(\text{árbol y } T_r \geq 4T_e) \tag{4.1}$$

En la Figura 4.1 vemos representado el árbol que hace referencia a la situación que estamos considerando, en donde tenemos 986 muestras de humanos actuales y una de un neandertal, que aparece en un tiempo $t_s$. Tenemos también representado el tiempo de coalescencia de todos los humanos ($T_e$) y el de los humanos y el neandertal ($T_r$).

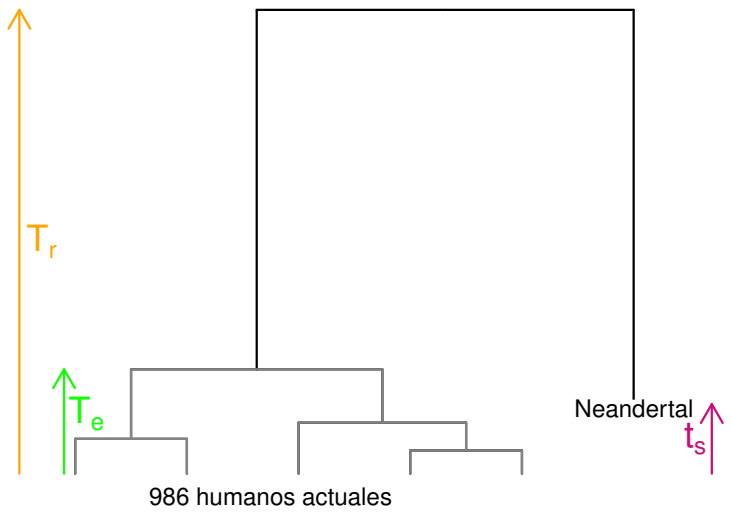



Figura 4.1: Árbol que representa la situación entre las muestras de los humanos y la muestra neandertal.

Usaremos $A_n(t)$ para denotar el número de linajes ancestrales que existen en el tiempo $t$ en el pasado, de una muestra de tamaño $n$ tomada en el presente (en nuestro caso $n = 986$). Nordborg razonó que, conocido el valor de $A_n(t)$, la probabilidad de que el árbol tenga esa forma y de que $T_r \geq 4T_e$ son independientes. Por lo tanto podemos escribir:

$$P(\text{árbol y } T_r \geq 4T_e) = \sum_{k=1}^{986} P(\text{árbol}|A_n(t_s) = k)P(T_r \geq 4T_e|A_n(t_s) = k)P(A_n(t_s) = k) \qquad (4.2)$$

Lo primero que haremos será calcular la probabilidad de que en un tiempo $t$, partiendo de una muestra de tamaño $n$, se tengan $k$ linajes, es decir, trataremos de calcular $g_{n,k}(t) = P(A_n(t) = k)$. Si nos fijamos, esta probabilidad es la misma que la de que antes del tiempo $t$, hayan tenido lugar exactamente $n - k$ eventos de coalescencia. Para calcular $g_{n,k}$ haremos un estudio ligeramente distinto si nos encontramos en el caso $k = 1$ de los demás.

Comencemos con $g_{n,1}(t)$. En este caso estamos buscando la probabilidad de que hayan tenido lugar $n - 1$ eventos de coalescencia antes del tiempo $t$, es decir, la probabilidad de que todos los individuos de la muestra de tamaño $n$ hayan encontrado a su ancestro común más reciente. Sabiendo esto y recordando que ya hemos obtenido la distribución de $T_{\text{MRCA}}$

(ver ecuación (2.41)), se tiene:

$$g_{n,1}(t) = \int_0^t f_{T_{\text{MRCA}}}(x)dx = \int_0^t \sum_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2}x} \prod_{j=2,j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} dx$$
$$= \sum_{i=2}^{n} \left(1 - e^{-\binom{i}{2}t}\right) \prod_{j=2,j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \tag{4.3}$$

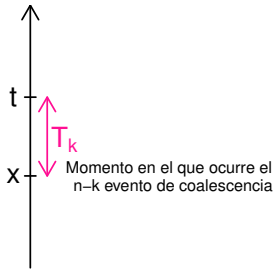
En este procedimiento simplemente hemos sustituido el valor de $f_{T_{\text{MRCA}}}$, que ya habíamos calculado en apartados anteriores, y hemos empleado la integral de una función exponencial.

Pasemos ahora al caso $2 \leq k \leq (n-1)$. En esta situación, para calcular $g_{n,k}(t)$ es necesario tener en cuenta que el $(n-k)$-ésimo evento de coalescencia ocurre antes de $t$, pero, sin embargo, el $(n-k+1)$-ésimo ocurre después de $t$. Vamos a considerar una nueva variable: $T_{n,k} = \sum_{i=k+1}^{n} T_i^c$, que denota el tiempo que pasa hasta que tiene lugar el $(n-k)$-ésimo evento de coalescencia. Una interpretación equivalente de esta variable es que describe el tiempo que pasa desde que tenemos $n$ linajes hasta que tenemos $k$. La distribución de esta nueva variable se obtiene de la misma manera que obtuvimos $f_{T_{\text{MRCA}}}$ en su momento, considerando la convolución de variables aleatorias exponenciales independientes. Así, llegamos, para $2 \leq k < (n-1)$, a que:

$$f_{T_{n,k}}(t) = \sum_{i=k+1}^{n} \binom{i}{2} e^{-\binom{i}{2}t} \prod_{j=k+1,j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \tag{4.4}$$

En el caso $k = n-1$ tenemos que $T_{n,n-1}$ no es más que el tiempo en el que hay $n$ linajes, es decir, no es más que $T_n^c$, y por tanto $f_{T_{n,n-1}}(t) = \binom{n}{2} e^{-\binom{n}{2}t}$.

Para calcular $g_{n,k}(t)$ simplemente tendremos que considerar que el $(n-k)$-ésimo evento de coalescencia ocurre en un instante $x$ antes de $t$ y que el tiempo en el que hay $k$ linajes ($T_k$) es mayor que $t-x$ pues el $(n-k+1)-$ésimo evento de coalescencia tiene que ocurrir después de $t$. Esta idea se plasma en la flecha temporal de la derecha.




En esta situación, para $2 \leq k \leq (n-1)$ tenemos que:

$$g_{n,k}(t) = \int_0^t f_{T_{n,k}}(x) \left[\int_{t-x}^{\infty} f_{T_k^c}(y)dy\right] dx \tag{4.5}$$

Así, en el caso $2 \leq k < (n-1)$ simplemente sustituyendo las funciones de densidad de las variables, integrando y haciendo una serie de cálculos llegamos a la siguiente expresión:

$$g_{n,k}(t) = \sum_{i=k}^{n} \frac{\binom{i}{2}}{\binom{k}{2}} e^{-\binom{i}{2}t} \prod_{j=k, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \tag{4.6}$$

Para $k = n-1$, integrando llegamos a:

$$g_{n,n-1}(t) = \frac{\binom{n}{2}}{\binom{n-1}{2} - \binom{n}{2}} \left[ e^{\left[\binom{n-1}{2} - \binom{n}{2}\right]t} - 1 \right] e^{-\binom{n-1}{2}t} \tag{4.7}$$

Para el caso $k = n$, $g_{n,k}(t)$ indica la probabilidad de que cuando estemos en el tiempo $t$ aún tengamos los $n$ linajes iniciales, es decir, que no se haya producido ningún evento de coalescencia en todo ese tiempo. Esto es sinónimo a que el tiempo $T_n^c$ sea mayor que $t$. Así:

$$g_{n,n}(t) = \int_t^{\infty} f_{T_n^c}(x)dx = e^{-\binom{n}{2}t} \tag{4.8}$$

De esta forma, ya conocemos la probabilidad de que en el tiempo $t$ haya $k$ linajes, para cualquier valor de $k$, partiendo de $n$ linajes iniciales, es decir, ya conocemos $g_{n,k}(t)$.

Pasemos ahora a calcular la probabilidad relacionada con la forma del árbol. Esta probabilidad hace referencia al hecho de que los neandertales no comparten ningún ancestro común con el ser humano antes de que haya un único linaje de los humanos. Así, calcularemos esta probabilidad condicionada al hecho de que $A_{986}(t_s) = k$. Por lo tanto, buscamos la probabilidad de que un linaje particular (el linaje del neandertal), en una muestra de tamaño $k+1$ (considerando los $k$ linajes humanos presentes en el tiempo $t_s$ y el linaje neandertal) no encuentre el ancestro común con ningún otro linaje hasta el final, cuando solamente queden dos linajes. Sabemos que en general si tenemos $j$ linajes, hay $\binom{j}{2} = \frac{j(j-1)}{2}$ posibles parejas que pueden encontrar su ancestro común. Dentro de ellas $j-1$ involucrarán al linaje del neandertal. Así, hay $\binom{j}{2} - (j-1)$ posibles parejas de linajes que pueden encontrar su ancestro común de entre las $\binom{j}{2}$ parejas existentes. Entonces, la probabilidad que buscamos será simplemente el producto desde $j = 3$ (cuando hay 2 linajes humanos y uno neandertal) hasta $j = k+1$, (cuando hay $k$ linajes humanos y uno neandertal) de los eventos de coalescencia que no involucran al linaje del neandertal entre los posibles totales que se podrían dar en general: $\frac{\binom{j}{2} - (j-1)}{\binom{j}{2}}$.

$$\begin{aligned} P(\text{árbol}|A_{986}(t_s) = k) &= \prod_{j=3}^{k+1} \left( \frac{\binom{j}{2} - (j-1)}{\binom{j}{2}} \right) = \prod_{j=3}^{k+1} \left( 1 - \frac{(j-1)}{\frac{j(j-1)}{2}} \right) \\ &= \prod_{j=3}^{k+1} \left( 1 - \frac{2}{j} \right) = \prod_{j=3}^{k+1} \left( \frac{j-2}{j} \right) \\ &= \frac{1 \cdot 2 \cdots (k-3)(k-2)(k-1)}{3 \cdot 4 \cdots (k-2)(k-1)k(k+1)} = \frac{2}{k(k+1)} \end{aligned} \tag{4.9}$$

Por último nos queda calcular la probabilidad de que $T_r$ sea mayor o igual que 4 veces $T_e$, sabiendo que en el tiempo $t_s$ se tienen $k$ linajes humanos. Al igual que hicimos antes para calcular $g_{n,k}$ vamos a distinguir el caso $k = 1$ de los demás.

Comencemos para $k \geq 2$. Así:

$$\begin{aligned} P(T_r \geq 4T_e | A_n(t_s) = k) &= P(T_r - 4T_e \geq 0 | A_n(t_s) = k) \\ &= P\left((T_r - t_s) - 4T_e \geq -t_s | A_n(t_s) = k\right) \\ &= P(T_{k+1,1} - 4T_{n,1} \geq -t_s) \\ &= P\left(\frac{T_{k+1,1}}{4} - T_{n,1} \geq \frac{-t_s}{4}\right) \end{aligned} \tag{4.10}$$

En este razonamiento hemos restado $t_s$ a ambos lados de la desigualdad y hemos usado que $T_r - t_s = T_{k+1,1}$, pues es el tiempo que pasa entre que tenemos $k+1$ linajes (el neandertal y los $k$ humanos) hasta que solamente tenemos 1. Además también hemos empleado que $T_e = T_{n,1}$, ya que esta variable determina el tiempo en el que se pasa de los $n$ linajes humanos iniciales a un único linaje humano. Finalmente hemos dividido todo entre 4. En la Figura 4.2 tenemos el árbol con $k = 2$ y vemos marcado $T_r - t_s$.

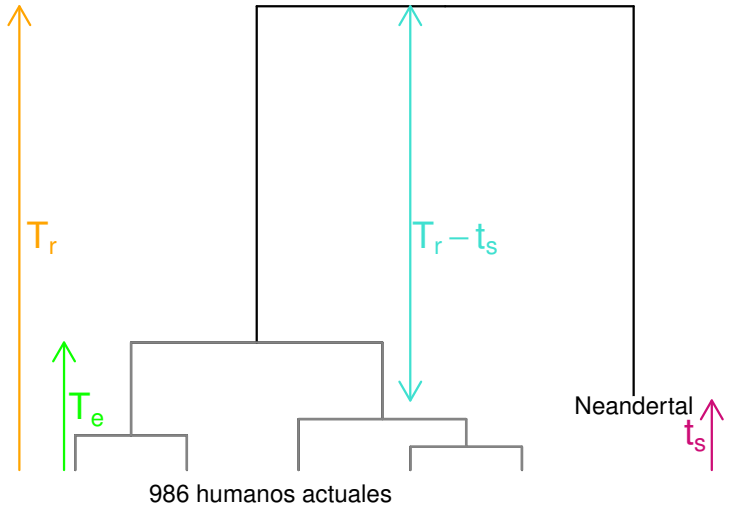



Figura 4.2: Árbol que representa la relación entre humanos y el neandertal junto con $T_r - t_s$ señalado.

Para obtener la distribución de $\frac{T_{k+1,1}}{4}$ calculamos su función de distribución y derivamos con el fin de obtener la función de densidad. Así llegamos a:

$$f_{\frac{T_{k+1,1}}{4}}(t) = \sum_{i=2}^{k+1} 4\binom{i}{2} e^{-4\binom{i}{2}t} \prod_{j=2, j\neq i}^{k+1} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \tag{4.11}$$

La distribución de $T_{n,1}$ ya la hemos visto antes en su fórmula general:

$$f_{T_{n,1}}(t) = \sum_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2}t} \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \tag{4.12}$$

Las dos variables $T_e$ y $T_{k+1,1}$ son independientes, pues estamos considerando por un lado a los humanos y por otro a los humanos más el neandertal. Así, ahora que ya sabemos todo esto, podemos calcular el valor de la probabilidad buscada. Para ello consideramos la siguiente expresión (en la que ya hemos tenido en cuenta que las dos variables son independientes) y empleamos el cambio de variable $y = z - x$:

$$\begin{aligned} P(T_r \geq 4T_e | A_n(t_s) = k) &= \int_{\frac{t_s}{4}}^{\infty} \left[ \int_{x-\frac{t_s}{4}}^{\infty} f_{T_{n,1}}(x) f_{\frac{T_{k+1,1}}{4}}(z) dz \right] dx \\ &+ \int_{0}^{\frac{t_s}{4}} \left[ \int_{0}^{\infty} f_{T_{n,1}}(x) f_{\frac{T_{k+1,1}}{4}}(z) dz \right] dx \\ &= \sum_{i=2}^{n} \sum_{r=2}^{k+1} \left[ 1 - \frac{4\binom{r}{2}}{\binom{i}{2} + 4\binom{r}{2}} e^{-\binom{i}{2}\frac{t_s}{4}} \right] p_i p_r \end{aligned} \tag{4.13}$$

donde:

$$p_i = \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \qquad p_r = \prod_{s=2, s\neq r}^{k+1} \frac{\binom{s}{2}}{\binom{s}{2} - \binom{r}{2}} \tag{4.14}$$

Para el caso $k = 1$ se tiene que $T_{k+1,1} = T_{2,1} = T_2^c$. Como la distribución de $T_2^c$ sabemos que es exponencial de parámetro 1, podemos calcular fácilmente la probabilidad buscada. Así:

$$\begin{aligned} P(T_r \geq 4T_e | k = 1) &= \int_{\frac{t_s}{4}}^{\infty} \left[ \int_{x-\frac{t_s}{4}}^{\infty} f_{T_{n,1}}(x) f_{\frac{T_2}{4}}(z) dz \right] dx \\ &= \sum_{i=2}^{n} \left[ 1 - \frac{4}{\binom{i}{2} + 4} e^{-\binom{i}{2}\frac{t_s}{4}} \right] \prod_{j=2, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \end{aligned} \tag{4.15}$$

Ahora ya tenemos todas las componentes necesarias para calcular el p-valor deseado. De esta forma, sustituyendo con ayuda de R (ver código en el Apéndice A.5) para $n = 986$ y recordando que consideramos que la población de mujeres era de 3400 y que cada generación duraba 20 años, podemos calcular las siguientes cantidades:

| | $t_s$**(en años)** | |
|---|---|---|
| | **30000** | **100000** |
| $\boldsymbol{E[A_{986}(t_s)]}$ | 4,87 | 1,75 |
| $\boldsymbol{P}$**(árbol)** | 0,085 | 0,56 |
| $\boldsymbol{P}$**(árbol y** $\boldsymbol{T_r \geq 4T_e}$**)** | 0,0020 | 0,023 |

Vemos que $E[A_{986}(t_s)]$ toma valores bastante pequeños para los dos valores de $t_s$ considerados. Esto indica que el número esperado de linajes humanos que siguen presentes en un tiempo $t_s$ es bastante pequeño (comparado con los 986 linajes que había inicialmente).

Además observamos que el p-valor que hemos definido toma valores suficientemente pequeños (menores que 0,05) para los dos $t_s$ considerados. Por lo tanto, teniendo esto en cuenta, podemos rechazar la posibilidad de que haya habido emparejamiento aleatorio entre los seres humanos y los neandertales al nivel del $0,2\,\%$ (en el caso de $t_s = 30000$) y al nivel del $2,3\,\%$ (en el caso de $t_s = 100000$).

Desde el año 1998 hasta la actualidad el estudio de los antepasados del ser humano moderno ha avanzado significativamente. Hoy en día se tiene mucha más información sobre este tema y han sido encontradas más muestras paleontológicas de homínidos antiguos. Además, se ha conseguido secuenciar al completo el genoma humano y el ADN de varios individuos neandertales. Con todo esto, la mayoría de los expertos en el tema establecen que sí que hay contribución de los neandertales en los seres humanos actuales ([15]), mientras que sigue habiendo alguno ([16]) que basándose en la comparación de ADN mitocondrial, establece que en el caso de que hubiese contribución esta sería pequeña. En general, la tendencia hoy en día, con todos los datos que se tienen, es la de creer que sí que hubo un momento en el que los neandertales y los humanos convivieron y se emparejaron entre sí.

# Apéndice A

# Códigos de R

## A.1. Algoritmo 1

```
##Algoritmo 1

#Número de genes
n=5

# Creación del cluster
dd <- dist(scale(seq(1:n)), method = "euclidean")
hc <- hclust(dd, method = "ward.D2")
hc$labels=c(1:n)
hc$order=c(1:n)

#Inicialización de variables
x=c(-n:-1)
i=1
T_k=0

#Algoritmo
while(length(x)>1){
  lambda=choose(length(x),2)
  T_k=rexp(1,rate=lambda)+T_k
  hc$height[i]=T_k
  c=sample(x,2,replace=F)
```

```
  x=setdiff(x,c)
  x=append(x,i,after=0)
  hc$merge[i,1]=c[1]
  hc$merge[i,2]=c[2]
  i=i+1
}

#Representación del árbol
hcd <- as.dendrogram(hc)
plot(hcd, type = "rectangle", main="Algoritmo 1")
```

## A.2. Algoritmo 2

```
##Algoritmo 2

#Número de genes
n=6

# Creación del cluster
dd <- dist(scale(seq(1:n)), method = "euclidean")
hc <- hclust(dd, method = "ward.D2")
hc$labels=c(1:n)
hc$order=c(1:n)


#Inicialización de variables
x=c(-n:-1)
i=1
T_k=0
theta=2
k=n
tm=c()

#Algoritmo
while(length(x)>1){
```

```
  lambda=k*(k-1+theta)/2
  T_k=rexp(1,rate=lambda)+T_k
  r=runif(1,0,1)
  l1=(k-1)/(k-1+theta)
  if(r<=l1){
      hc$height[i]=T_k
      c=sample(x,2,replace=F)
      x=setdiff(x,c)
      x=append(x,i,after=0)
      hc$merge[i,1]=c[1]
      hc$merge[i,2]=c[2]
      i=i+1
      k=k-1
  }
  else{
      tm=c(tm,T_k)
  }
}

# Gráfico
hcd <- as.dendrogram(hc)
plot(hcd, type = "rectangle", ylab = "Tiempo",main="Algoritmo 2")

#Dibujar los puntos de las mutaciones
library(ggdendro)
segmentos=dendro_data(hcd)$segments
d=c()
if (length(tm)>0){
  for (i in 1:length(segmentos$x)){
    if (segmentos$y[i]==segmentos$yend[i]){
      d=append(d,i)
    }
  }
  segmentos=segmentos[-d,]
  m=nrow(segmentos)
  n1=c()
```

```
  for (j in 1:length(tm)){
    for (i in 1:m){
      if (tm[j]>=segmentos$yend[i] & tm[j]<=segmentos$y[i]){
        n1=c(n1,i)
      }
    }
    inew=sample(n1,1)
    n1=c()
    points(segmentos$x[inew],tm[j],pch=16)
  }
}
```

## A.3. Algoritmo 3

```
##Algoritmo 3

#Número de genes
n=5

# Creación del cluster
dd <- dist(scale(seq(1:n)), method = "euclidean")
hc <- hclust(dd, method = "ward.D2")
hc$labels=c(1:n)
hc$order=c(1:n)

#Inicialización de los parámetros
x=c(-n:-1)
i=1
T_k=0

#Algoritmo
while(length(x)>1){
  lambda=choose(length(x),2)
  T_k=rexp(1,rate=lambda)+T_k
  hc$height[i]=T_k
```

```
  c=sample(x,2,replace=F)
  x=setdiff(x,c)
  x=append(x,i,after=0)
  hc$merge[i,1]=c[1]
  hc$merge[i,2]=c[2]
  i=i+1

}

# Gráfico
hcd <- as.dendrogram(hc)
plot(hcd, type = "rectangle", ylab = "Tiempo",main="Árbol")

#Colocación de las mutaciones
library(ggdendro)
segmentos=dendro_data(hcd)$segments
d=c()
for (i in 1:length(segmentos$x)){
  if (segmentos$y[i]==segmentos$yend[i]){
    d=append(d,i)
  }
}
theta=2
segmentos=segmentos[-d,]
m=nrow(segmentos)
for (i in 1:m){
  t=abs(segmentos$yend[i]-segmentos$y[i])
  Mt=rpois(1,t*theta/2)
  x1=runif(Mt,segmentos$yend[i],segmentos$y[i])
  x2=rep(segmentos$x[i],Mt)
  points(x2,x1,pch=16)
}
```

## A.4. Diagrama de barras 3D

```
library(plot3D)
k=seq(0,10,by=1)
n=seq(2,15,by=1)

nk=length(k)
nn=length(n)
theta=3
p=c()
A=matrix(0L,nrow=nk,ncol=nn)

for (i in 1:nk){
  for (j in 1:nn){
  A[i,j]=funcionp_S(n[j],k[i],theta)
  }
}
hist3D(k,n,A, theta=25,phi=20,xlab="s",ylab="n",zlab="P(S=s)")

funcionp_S=function(n,k,theta){
  suma=0
  for(i in 2:n){
    suma=suma+(-1)^i*choose(n-1,i-1)*(i-1)/(theta+i-1)*(theta/(theta+i-1))^k
  }
  return(suma)
}
```

## A.5. Neandertales

```
#p-valor
n=986
pv=0
for (k in 1:n){
  pv=pv+2/(k*(k+1))*prob_An(0.44,k,n)*prob_t(0.44,k,n)
}
pv
```

```
#Esperanza de A_n
E=0
for (k in 1:n){
  E=E+k*prob_An(0.44,k,n)
}
E

#Probabilidad árbol
arb=0
for (k in 1:n){
  arb=arb+2/(k*(k+1))*prob_An(0.44,k,n)
}
arb

#Función probabilidad An
prob_An=function(t,k,n){
  if (k==1){
    suma=0
    for (i in 2:n){
      p=1
      for (j in 2:n){
        if (j!=i){
          p=p*choose(j,2)/(choose(j,2)-choose(i,2))
        }
      }
      suma=suma+exp(-choose(i,2)*t)*p
    }
    return(1-suma)
  } else if (k==n){
    suma=exp(-choose(n,2)*t)
    return(suma)
  } else{
    suma2=0
    for (i in k:n){
      p2=1
      for (j in k:n){
```

```
        if (j!=i){
          p2=p2*choose(j,2)/(choose(j,2)-choose(i,2))
        }
      }
      suma2=suma2+choose(i,2)*exp(-choose(i,2)*t)*p2
    }
    return(suma2/choose(k,2))
  }
}

#Función probabilidad T

prob_t=function(t,k,n){
  if (k==1){
    suma=0
    for (i in 2:n){
      p=1
      for (j in 2:n){
        if (j!=i){
          p=p*choose(j,2)/(choose(j,2)-choose(i,2))
        }
      }
      suma=suma+p*(1-4/(4+choose(i,2))*exp(-choose(i,2)*t/4))
    }
    return(suma)
  } else{
    suma6=0
    for (i in 2:(n)){
      p2=1
      for (j in 2:(n)){
        if (j!=i){
          p2=p2*choose(j,2)/(choose(j,2)-choose(i,2))
        }
      }
      suma5=0
      for (r in 2:(k+1)){
```

```
        p5=1
        for (s in 2:(k+1)){
          if (s!=r){
            p5=p5*choose(s,2)/(choose(s,2)-choose(r,2))
          }
        }
        suma5=suma5+p2*p5*((1-4*choose(r,2)*
        exp(-choose(i,2)*t/4)/((choose(i,2)+choose(r,2)*4))))
      }
      suma6=suma6+suma5
    }

    return(suma6)
  }
}
```

# Bibliografía


[1] Sheldon Ross. *A First Course in Probability*. Pearson, 2014.

[2] Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Discrete multivariate distributions*, volume 165. Wiley New York, 1997.

[3] Anirban DasGupta. *Fundamentals of probability: a first course*. Springer Science & Business Media, 2010.

[4] Marco Taboga. *Joint moment generating function*. `https://www.statlect.com/fundamentals-of-probability/joint-moment-generating-function`, 2017.

[5] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, USA, 2004.

[6] John Wakeley. *Coalescent Theory: An Introduction*. Roberts and Company Publishers, 2009.

[7] José Manuel Sánchez Muñoz. El problema de Basilea. *Lecturas matemáticas*, 35(2):199–228, 2014.

[8] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

[9] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 1995.

[10] Yamin Liu. *Genetic Diversity and Disease Susceptibility*. BoD–Books on Demand, 2018.

[11] GA Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276, 1975.

[12] G Danesh, B Elie, and S Alizon. Early phylodynamics analysis of the covid-19 epidemics in france using 194 genomes. Technical report, 2020.

[13] Matthias Krings, Anne Stone, Ralf W Schmitz, Heike Krainitzki, Mark Stoneking, and Svante Pääbo. Neandertal DNA sequences and the origin of modern humans. *cell*, 90(1):19–30, 1997.

[14] Magnus Nordborg. On the probability of neanderthal ancestry. *American journal of human genetics*, 63(4):1237, 1998.

[15] Anders Bergström, Chris Stringer, Mateja Hajdinjak, Eleanor ML Scerri, and Pontus Skoglund. Origins of modern human ancestry. *Nature*, 590(7845):229–237, 2021.

[16] David Serre, André Langaney, Mario Chech, Maria Teschler-Nicola, Maja Paunovic, Philippe Mennecier, Michael Hofreiter, Göran Possnert, and Svante Pääbo. No evidence of neandertal mtdna contribution to early modern humans. In *Early modern humans at the Moravian gate*, pages 491–503. Springer, 2006.