UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

Modelos y estadísticas de la evolución

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Matemático

PRESENTA:

Aurelio Bolívar Galván Yttesen

TUTOR

Arnaud Charles Leo Jegousse

Ciudad Universitaria, CD.MX., 2021

1. Datos del alumno
Galván
Yttesen
Aurelio Bolívar
5556799039
Universidad Nacional Autónoma de México
Facultad de Ciencias
Matemáticas
310235592

2. Datos del tutor
Dr
Arnaud Charles Leo
Jegousse

3. Datos del sinodal 1
Dr
Fernando
Baltazar
Larios

4. Datos del sinodal 2
Dra
Verónica de la Santísima Faz
Miró
Pina

5. Datos del sinodal 3
Dra
Laura Clementina
Eslava
Fernández

6. Datos del sinodal 4
Dr
Adrián
González Casanova
Soberón

7. Datos del trabajo escrito
Modelos y estadísticas de la evolución
70 p
2021

Dedicatoria

A todos los estudiantes que les interese y que les pueda servir este trabajo para orientarse y resolver dudas sobre los temas tratados.

Agradecimientos

Agradezco primeramente a mis padres que me han dado todo y gran parte de lo que soy se lo debo a ellos.

Agradezco a mi hermana por ayudarme eventualmente a entender muchos de los temas biológicos que se exponen.

Agradezco a mi asesor que tuvo la paciencia de explicarme simpre que tenía una duda.

Agradezco a mis amigos y todas las personas que me ayudaron, me animaron y me aconsejaron en este trabajo.

Por último agradezco enormemente a mi Alma Máter, la Universidad Nacional Autónoma de México, por todos los años que en sus aulas me ha enseñado.

Índice general

Introducción

La genética de poblaciones es la disciplina que nos ayuda a entender la evolución a través del estudio de la variación en el ADN de una población. Existen cuatro factores que determinan la evolución genética de una especie, estos son, la mutación, la deriva genética, la migración y la selección natural, los cuales serán explicados brevemente en lo sucesivo. Previo a ésto, se expondrá un concepto importante, el alelo, para obtener un mejor entendimiento de los cuatro factores mencionados anteriormente.

Un alelo es cada una de la formas que tiene un gen de expresarse. Supongamos, por ejemplo, que el código genético para formar el color de ojos de un individuo tiene dos posibles expresiones uno de los cuales es el color azul, que llamaremos alelo A y la otra posible expresión del gen es el color café, el cuál llamaremos alelo a. Cuando existen diferentes posibilidades de expresarse un gen como lo es en el ejemplo anterior con los alelos A y a, entonces se dice que el individuo es heterocigoto para este gen. Cuando exite solo un alelo para el gen en cuestión, se dice que el individuo es homocigoto con respecto a dicho gen. Se suelen representar a los alelos con letras mayúsculas o minúsculas dependiendo de si el alelo es dominante, en cuyo caso se asigna la letra mayúscula, o recesivo, al cual se le asigna la letra minúscula.

Dicho lo anterior, los cuatro factores evolutivos son:

- La selección natural es el proceso que resulta de cumplir 3 condiciones:

- Variación fenotípica (rasgos observables de los organismos) entre los individuos de una población.
- Supervivencia o reproducción diferencial asociada a la variación.
- Herencia de la variación.

Cuando ocurren estas 3 condiciones entonces hay una variación genética por selección natural. La selección natural se suele resumir en la frase "la supervivencia del más apto".

- Otro factor evolutivo es la migración pues con esta ocurre un intercambio de genes en la población. Cuando las frecuencias de los alelos de dos poblaciones difieran, entonces con la mezcla de los individuos, estas frecuencias cambiarán.
- La mutación es la escencia del cambio genético, sin ella simplemente no habría evolución. Una mutación es un cambio heredable en el genoma, por lo tanto altera el ADN. Muchas variaciones no tienen éxito y son eliminadas, sin embargo, hay variaciones que prosperan y se integran al genotipo de la especie. La frecuencia con la que se producen mutaciones se le llama *tasa de mutación*.
- En la deriva genética se transmiten los alelos de una generación a la siguiente de forma aleatoria. Supongamos por ejemplo que en la generación n hay dos tipos de alelos A y a, entonces para la siguiente generación $n+1$ estos dos alelos estarán distribuidos aleatoriamente sobre la poblacion según la facilidad con la que predomine un alelo sobre otro. Es por esto que la deriva genética tiene un carácter de azar. Si sucede que en alguna generación un alelo desaparece, éste ya no vuelve a presentarse en las siguientes generaciones. La deriva genética suele llevar a la perdida de la variabilidad genética. Dicho estancamiento de la variablidad se rompe por medio de la mutación.

En 1940 Ronald Fisher crea la teoría sintética moderna la cual define a la evolución como un cambio en la frecuencia de los alelos de una población a lo largo de las generaciones. Para dicha teoría, la evolución consta de dos etapas, la primera es el surgimiento al azar de la variación y la segunda es la selección de las variantes producidas. Sin embargo, uno de los defectos que señalan los científicos que tiene esta teoría es que no considera a las extinciones masivas como una explicación evolutiva. Igualmente en la década de 1940 se logró identificar definitivamente el ADN como el ente responsable de transmitir la información genética, y es hasta 1953 que Watson y Crick dan a conocer la estructura de doble hélice. Todos estos avances sentaron las bases para desarrollar la teoría molecular para que posteriormente en la década de 1970, Motoo Kimura creara la teoría neutralista de la evolución. Ésta señala a la deriva genética y a la mutación como los principales factores evolutivos, incluso por encima de la selección natural. A la teoría neutralista se le considera como el modelo por excelencia para probar cualquier hipótesis selectiva gracias a la simplicidad y las buenas predicciones teóricas que tiene sobre la tasa evolutiva.

A inicios de la década de 1980 varios grupos de investigadores de manera independiente fueron creando lo que se conoce como la teoría de coalescencia que viene a ser un complemento muy importante para la genética de poblaciones, con ella se busca entender las genealogías de una población con teoría probabilística y sobretodo haciendo uso de simulaciones para modelar e inferir los factores que afectaron la evolución de una población, por ejemplo, el tamaño poblacional, la tasa de mutación, la tasa de recombinación, etc. Uno de los principales personajes que aportó a la teoría de coalescencia es John Kingman el cual creó un modelo que se ha usado de manera significativa en este campo. Este modelo recibe el nombre de *coalescente de Kingman*. Existen muchos diferentes modelos aparte del mencionado anteriormente como, por ejemplo, el Λ - *coalescente* pero este trabajo se centrará

más en el coalescente de Kingman.

Capítulo 1

Modelos clásicos de la evolución

1.1. Modelo Wright-Fisher

En la genética de poblaciones hay distintos tipos de modelos probabilísticos que nos ayudan a entender el comportamiento de la evolución de una población por medio de sus variaciones genéticas, uno de ellos es el famoso modelo Wright-Fisher que hace ciertas suposiciones generales sobre la población a estudiar las cuales son:

- Una población con un número de individuos constante N, que en el caso de las especies las cuales tienen su información genética por pares de cromosomas como los humanos (23 pares de cromosomas) se suele encontrar en la literatura que la población es representada por un valor constante 2N.
- Los individuos de la población se mezclan bien, es decir, cada indivuo es propenso a interactuar con cualquier otro de la población.
- No actua la selección sobre la población.

En el modelo Wright-Fisher las generaciones son discretas, no se traslapan y cada una se forma de manera uniforme y con reemplazo respecto a la generación anterior, es decir,

para una población a tiempo $n \in \mathbb{Z}$ cada individuo escoge a su padre de la generación *n-1* de manera uniforme e independiente, por lo que la distribución que sigue la descendencia es Multinomial $M\left(2N, \frac{1}{2N}, ..., \frac{1}{2N}\right)$.

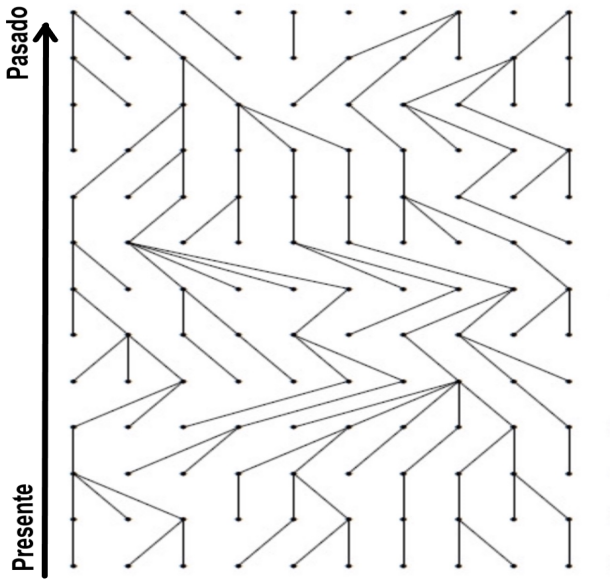

Figura 1.1: En esta figura se observa como la descendencia elige a sus padres aleatoriamente y con reemplazo en una población con $2N$ individuos.

Supongamos ahora que tenemos dos tipos de alelos A y a, llamaremos X_n al número de alelos tipo A que hay en la generación n. Podemos ver que X_n es una cadena de Markov que sigue una distribución binomial, lo anterior sucede puesto que cada nueva generación depende únicamente de lo que sucedió en la generación anterior (no hay saltos entre generaciones). Por lo tanto para saber la cantidad de alelos A que habrá en una generación nos fijamos en la cantidad que hubo en la generación anterior, de esta manera si hay i individuos con el alelo A en la generación n, entonces la probabilidad de tener j alelos en la generación $n+1$ es

$$P(X_{n+1} = j \mid X_n = i) = p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}.$$

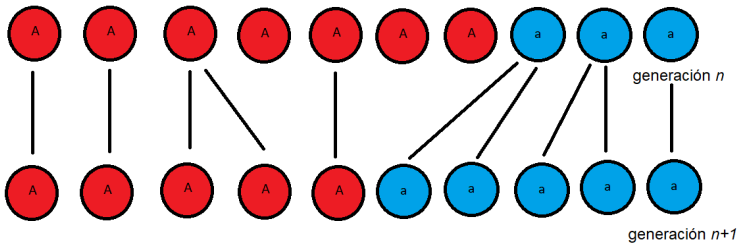

Figura 1.2: Ejemplo de una generación progenitora con descendencia y frecuencia alélica.

En la Figura 1.2 se muestra a modo de ejemplo una generación y la sucesiva con dos tipos de alelos A y a con su respectiva descendencia. En este sencillo caso se muestra el evento de que el proceso pase del estado con 7 alelos A al estado con 5 alelos, es decir, la probabilidad

$$P(X_{n+1} = 5 \mid X_n = 7) = p_{75} = \binom{10}{5}\left(\frac{7}{10}\right)^5\left(1-\frac{7}{10}\right)^{10-5} = .1029.$$

Es bueno recalcar que esta cadena tiene dos estados absorbentes los cualés son cuando el número de alelos A llegue a 0 y el otro es cuando haya $2N$ alelos A. Desde luego es más probable que el tiempo de fijación (tiempo en el que, en la muestra, sólo hay un determinado tipo alelo) llegue más rapido en poblaciones pequeñas o cuando la frecuencia de uno de los alelos sea muy grande.

En la Figura 1.3 se muestra como entre la población es más chica hay menos fluctuaciones con respecto a la frecuencia del alelo A, debido a que se llega al tiempo de fijación con mayor rapidez, mientras que si la población es más grande la frecuencia del alelo A varia más y por mayor tiempo. También se puede observar con facilidad como cuando la población es chica ésta llega más rapido al tiempo de fijación.

Hasta el momento hemos visto la actuación sólo de la deriva génica en el modelo Wright-Fisher que a lo largo de las generaciones propicia la pérdida de la variación, pero el factor que

[1]Oluwafemi Oyamakin, S., Chukwu, A., Oluwaseun, W., Ogunjobi, E. (2019). Allele Based Inference on Evolution and Extinction; A Genetic Drift Approach. Journal of Cancer, Genetics And Biomarkers - 1(4):1-15.

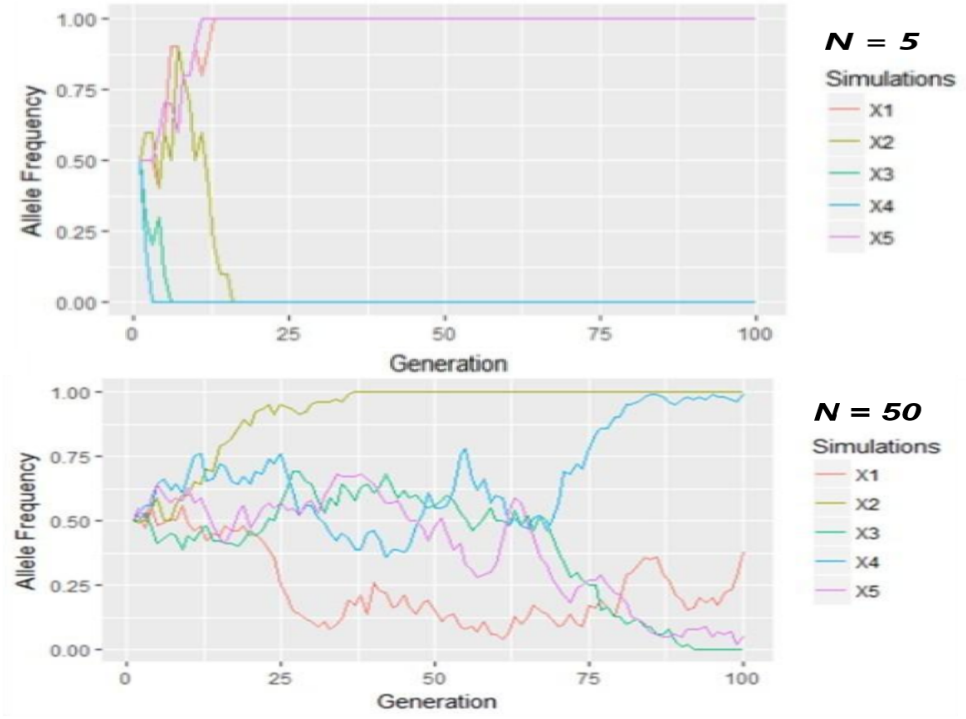

Figura 1.3: Frecuencias alélicas según el tamaño poblacional. [1]

la contrarresta es la mutación. Entenderemos una mutación como el cambio en alguna base nitrogenada en algún segmento de la secuencia del ADN. Llamaremos μ a la *tasa de mutación* en el modelo Wright-Fisher, donde más adelante será explicada la siguiente igualdad:

$$\mu = \frac{\theta}{2N} \quad \text{con} \quad \theta \geq 0.$$

1.2. Genealogías del modelo Wright-Fisher

Ahora bien ¿Cuál es la relación que existe entre el modelo Wright-Fisher y el coalescente de Kingman? La respuesta a esta pregunta viene de reescalar el tiempo de la genealogía. Para esto primero observemos cuál sería la probabilidad de que dos individuos de una muestra provengan del mismo padre en el modelo Wright-Fisher, la cual es fácil ver que es $\frac{1}{2N}$, similarmente la probabilidad para que tres individuos sean hermanos es $\left(\frac{1}{2N}\right)^2$, en general, la probabilidad de que k individuos sean hermanos es $\left(\frac{1}{2N}\right)^{k-1}$.

Llamaremos T a la variable aleatoria que mide el tiempo en el que coalescen dos linajes. T se distribuye geométricamente pues se contará el número de generaciones hasta que los linajes coalezcan, es decir $T \sim Geo\left(\frac{1}{2N}\right)$.

Por otro lado recordemos que

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{nx} = e^{-x} \quad .$$

Así, nos fijamos en la probabilidad de tardar más de m generaciones en coalescer, es decir,

$$P\left(T > m\right) = \left(1-\tfrac{1}{2N}\right)^{m}.$$

Entonces ¿qué pasa si la población tiende a infinito? Esto es que $2N \to \infty$ que implica que de la probabilidad anterior los dos linajes no llegarán a tocarse ya que $P(T > m) \to 1$ para algún m fijo. Es aquí donde buscamos reescalar el tiempo y que por ende éste no se cuente de una generación en una generación de manera discreta, sino de $2N$ generaciones en $2N$ generaciones, esto debido a que en promedio ambos linajes tardarán en juntarse $2N$ generaciones, pues $E(T) = 2N$. De esta manera si multiplicamos por un factor $2N$ el tiempo tenemos

$$P(T > 2Nm) = \left(1-\frac{1}{2N}\right)^{2Nm}$$

así si hacemos $N \to \infty$ tenemos que

$$P(T > 2Nm) = P\left(\frac{T}{2N} > m\right) \to e^{-m}$$

es decir, $\frac{T}{2N}$ converge en distribución a la exponencial de parámetro 1. Por lo tanto, al reescalar pasamos a un tiempo continuo.

Sucede algo similar a lo anterior si vemos lo que pasa cuando queremos que tres linajes se junten al mismo tiempo, pero ahora con el tiempo reescalado. Si llamamos T' a la variable

aleatoria que cuenta el tiempo que tardan en juntarse los linajes de tres individuos simultaneamente, entonces la probabilidad de tardar más de $2Nm$ generaciones para que ocurra la coalescencia múltiple es 1, como podremos ver en la siguiente ecuación

$$P(T' > 2Nm) = \left(1 - \frac{1}{(2N)^2}\right)^{2Nm}$$

de donde el denominador $(2N)^2$ hace referencia a la probabilidad de que 3 individuos compartan el mismo padre, y así, al tender a infinito este denominador crece con un orden distinto al del exponente, entonces

$$P(T' > 2Nm) = P\left(\frac{T'}{2N} > m\right) \to 1 \text{ cuando } N \to \infty$$

por lo que vemos que los 3 linajes no se llegan a juntar en el tiempo, incluso ya reescalado. Con esto podemos ver cómo se cumplen las características que distinguen al coalescente de Kingman a raíz del modelo Wright-Fisher, que son los tiempos de coalescencia a un ritmo exponencial y la coalescencia binaria.

Para saber ahora lo que pasa retomamos con el factor de mutación del modelo Wright-Fisher al coalescente de Kingman, llamaremos T_M el número de generaciones antes de la primer mutación, con una tasa μ , igualmente T_M se distribuye como una variable aleatoria geométrica, así con la escala de tiempo que hemos venido usando observamos que, análogamente

$$P(T_M > 2Nm) = (1-\mu)^{2Nm}$$

por lo tanto para que converja la probabilidad anterior cuando $N \to \infty$ tenemos que hacer que μ converja a una velocidad inversa a la velocidad del exponente, por ende hacemos

$$\mu = \frac{\theta}{2N}$$

para algún θ fijo, con esto vemos que

$$P(T_M > 2Nm) = P\left(\frac{T_M}{2N} > m\right) = (1-\mu)^{2Nm} = \left(1 - \frac{\theta}{2N}\right)^{2Nm} \to e^{-\theta m} \text{ cuando } N \to \infty$$

así hacemos notar que las mutaciones en el coalescente de Kingman aparecen distribuidas en el tiempo según una distribución exponencial con una tasa en el coalescente de Kingman de

$$\boxed{\theta = 2N\mu.}$$

1.3. Cadenas de Markov a tiempo continuo

Se expondrá ahora un tema que servirá especialmente para definir de manera formal el Coalescente de Kingman y también para dar una base sobre la cual está apoyada la teoría detrás del coalescente.

Recordemos primero lo que es una cadena de Markov a tiempo discreto, esto es:

Definición 1.3.1 (Cadena de Markov) *Sea $n = 0, 1, 2, \cdots$. Una cadena de Markov a tiempo discreto es una sucesión de variables aleatorias X_n, las cuales toman valores en un espacio de estados discreto y que satisfacen la propiedad de Markov*

$$P\left(X_{n+1} = x_{n+1} \mid X_0 = x_0, \cdots, X_n = x_n\right) = P\left(X_{n+1} = x_{n+1} \mid X_n = x_n\right)$$

Dicha propiedad señala, en resumen, que para ver como se comporta una variable en el tiempo $n+1$ sólo importa fijarse en el estado anterior n, es decir, todo lo ocurrido en tiempos previos a n carece de importancia.

Ahora, si la idea de cadena de Markov no se limita por un tiempo discreto y se extiende a una donde el tiempo que rige el proceso toma el carácter de variable aleatoria continua,

entonces se llega a una proceso de cadena de Markov a tiempo continuo. Se considerará un espacio de estados discreto, así como se hizo en el proceso con la cadena de Markov anterior. Cabe aclarar que una vez que el proceso se encuentre en un estado i la cadena permanecerá en este estado un tiempo aleatorio T_i, como se muestra en la Figura 1.4; a estos tiempos se

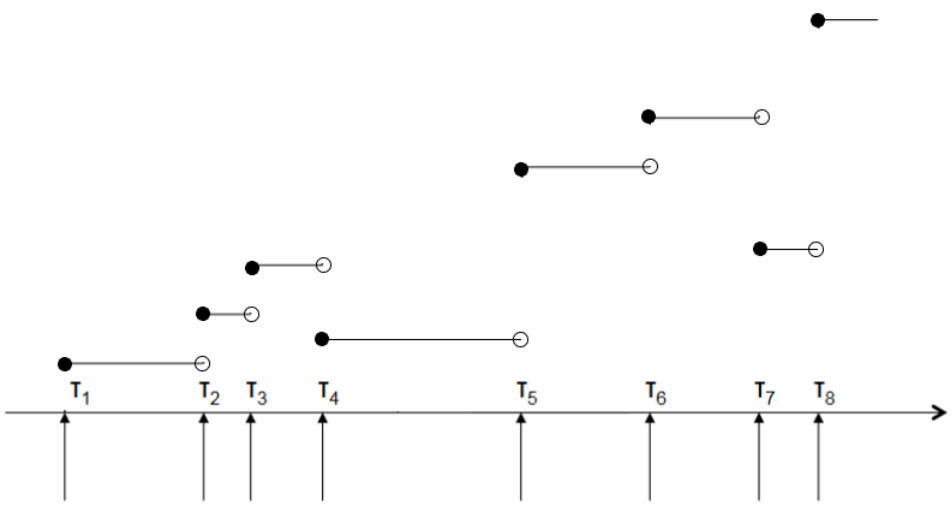

Figura 1.4: Saltos en un proceso a tiempo continuo.

les conoce como tiempos de salto. Sin embargo, con lo que se ha expuesto no se garantiza el cumplimiento de la propiedad de Markov a tiempo continuo, por lo que el modelo requiere de algunas características más; llamemos p_{ij} a la probabilidad de pasar del estado i al estado j cuando el proceso da un salto entonces cumplirá:

- $\sum_j p_{ij} = 1$
- $p_{ij} \geq 0$
- $p_{ii} = 0$

supondremos adicionalmente que un estado tiene dos opciones, puede ser absorbente o no absorbente, implicando que una vez que el valor de algún tiempo T_i tienda a infinito entonces el proceso ya no dará más saltos, aunado a esto se pedirá que los tiempos T_i sean

independientes entre sí. Con esto podemos concluir con un resultado fuerte y no sencillo de probar que señala lo siguiente:

" El proceso de saltos como se construyó anteriormente cumple la propiedad de Markov si y sólo si los tiempos de salto de los estados no absorbentes tienen distribución exponencial con parámetro $\lambda_i > 0$. "[2]

Dicho lo anterior, llamaremos $Q(i,j) = P(X_{T_{k+1}} = j \mid X_{T_k} = i)$ donde $T_0 = 0$ y T_k es una variable aleatoria la cual representa el tiempo en el que la cadena de Markov salta de un estado al siguiente.

Ahora como se vio en la sección 1.2 en la transición del modelo Wright-Fisher al coalescente de Kingman, la variable aleatoria con la que se avanza en el tiempo es una *exponencial*, esto quiere decir que $T \backsim exp(\lambda(i))$, donde $\lambda(i)$ es la suma de ciertos valores que se aclarará más adelante y se escribió de esa forma debido a que se quiere hacer énfasis en que dicho parámetro depende del lugar donde se encuentra situado el proceso. Antes de continuar se introducirá notación.

Notación 1

- *Denotaremos* $[n]$ *al conjunto de los primeros n números naturales* $\{1,2,3,\ldots,n\}$.
- *Llamaremos* $\mathscr{P}_n$ *al conjunto de todas las particiones de* $[n]$.

Para tener una idea clara de cómo es que pasa (el proceso estocástico) de un estado a otro, tenemos que entender que, el primer lugar a donde se va a brincar es a donde el tiempo sea mínimo de entre todas las posibles opciones a donde puede ir el proceso. Antes que nada veremos a continuación un lema muy sencillo de probar para ayudarnos a justificar los resultados siguientes.

[2]Rincón, L. (2013). Cadenas de Markov a tiempo continuo. En Introducción a los procesos estocásticos(p. 147). México: Facultad de Ciencias, UNAM.

Lema 1.3.1 *Sea* $Y = min\{X_1, X_2, \ldots, X_n\}$ *con* $X_i \backsim exp(\lambda_i)$ *variables aleatorias independientes. Entonces*

$$Y \backsim exp\left(\sum_{i=1}^{n} \lambda_i\right)$$

Ahora, para demostrarlo tomemos $Y = min\{X_1, X_2, \ldots, X_n\}$ y sabemos que por ser exponenciales $P(X_i > y) = e^{-\lambda_i y}$, entonces

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = 1 - P(Y > y) = 1 - P(X_1 > y, X_2 > y, \ldots, X_n > y) \\ &= 1 - P(X_1 > y) \cdot P(X_2 > y) \cdots P(X_n > y) \quad \text{por independecia} \\ &= 1 - e^{-\lambda_1 y} \cdot e^{-\lambda_2 y} \ldots e^{-\lambda_n y} \\ &= 1 - e^{\sum_{i=1}^{n} -\lambda_i y} \end{aligned}$$

por lo tanto $Y \backsim exp\left(\sum_{i=1}^{n} \lambda_i\right)$.

Ahora supongamos que estamos en un estado i (una partición con m conjuntos) y queremos pasar al estado j_k (una partición accesible desde i) para algún $k \in \left[\binom{m}{2}\right]$. Hay dos formas en las que podemos entender este salto.

- Si suponemos que estamos en i , entonces a partir de que llegamos empieza a transcurrir un tiempo exponencial con parámetro λ_i y brincamos a j_k con una cierta probabilidad $Q(i, j_k)$.

- La otra forma en la que podemos entender el salto es que si estamos en el estado i y dados los distintos estados posibles a los que se puede brincar, la idea es que todos éstos compiten o lo empiezan a atraer al mismo tiempo, entonces al estado al que brincará el proceso es al que arroje el menor tiempo de entre todos, es decir, de los estados posibles brincará a donde tengamos un tiempo que igualmente es exponencial

pero con una tasa de esta forma $exp\left(\lambda_i \cdot \sum_{k=1}^{m} Q(i,j_k)\right)$, ésto dado que sabemos cuál fue el que ganó.

Para entender mejor se ilustrará con una figura el inciso anterior. Fijémonos en una parte de la Figura 1.6, y vemos que a cada estado accesible desde $i = \{\{1\},\{2,4\},\{3\}\}$ tenemos, supongamos, distintas tasas de transición que al hacer una sola simulación nos arroja diferentes resultados, por ejemplo,

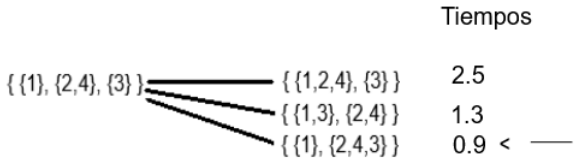

Figura 1.5: Transición entre dos estados con ejemplo de tiempos de coalescencia. Se tomó un caso en el que se hace el brinco de un estado a otros posibles estados del ejemplo de la Figura 1.6 .

para $j_1 = \{\{1,2,4\},\{3\}\}$ tenemos un tiempo *exponencial* $exp(\lambda_i \cdot Q(i,j_1)) \longrightarrow t_1 = 2{,}5$,

para $j_2 = \{\{1,3\},\{2,4\}\}$ tenemos un tiempo *exponencial* $exp(\lambda_i \cdot Q(i,j_2)) \longrightarrow t_2 = 1{,}3$,

para $j_3 = \{\{1\},\{2,4,3\}\}$ tenemos un tiempo *exponencial* $exp(\lambda_i \cdot Q(i,j_3)) \longrightarrow t_3 = 0{,}9$;

por lo que a donde daría el salto sería a j_3 y este tiempo también sería *exponencial*

$$exp\left(\lambda_i \cdot \left[\underbrace{Q(i,j_1) + Q(i,j_2) + Q(i,j_3)}_{=1}\right]\right)$$

y además sabemos que

$$\sum_{k=1}^{m} Q(i,j_k) = 1\ ,\ k \in \left[\binom{m}{2}\right]$$

por tratarse de probabilidades de transición con $\binom{m}{2}$ el número de estados accesibles desde i. De tal forma que al preguntarnos por el mínimo tiempo tenemos lo siguiente,

$$\begin{aligned} exp\left(\lambda_i \cdot \sum_{k=1}^{m} Q(i, j_k)\right) &= exp(\lambda_i \cdot 1) \\ &= exp(\lambda_i) \end{aligned}$$

por lo tanto vemos que recuperamos el parámetro λ_i. La ventaja de ver cómo ocurre el salto entre los estados de la segunda manera es que podemos definir una matriz que caracterice todo el proceso. Definimos M como sigue:

$$M = \begin{cases} \lambda_i \cdot Q(i, j) & \text{si } i \neq j \text{ y } j \text{ un estado accesible desde } i \\ -\lambda_i & \text{si } i = j \\ 0 & \text{cualquier otro caso.} \end{cases}$$

1.4. Coalescente de Kingman

El coalescente de Kingman se puede entender como un proceso estocástico, más en particular como una cadena de Markov a tiempo continuo. Las dos principales cualidades que se expusieron en el apartado 1.2 las cuales caracterizan a este coalescente son:

- Tiempos de coalescencia: los tiempos en que coalescen los linajes ocurren según una distribución exponencial λ_i.
- Coalescencia por pares: no puede ocurrir una coalescencia de tres o más linajes a la vez.

Para definir más rigurosamente el coalescente de Kingman se darán algunas definiciones y notaciones previas para entender como está construido.

Definición 1.4.1 *Definimos π una partición, es decir, una relación de equivalencia en $\mathscr{P}_n$, de tal manera que las clases de equivalencia de esta partición serán los bloques $B_1, B_2, B_3, \ldots$ los*

cuáles son definidos como sigue: B_1 es el bloque que contiene al 1*, B_2 el bloque que contiene al elemento más chico que no esté en B_1, B_3 el bloque que tiene al elemento más chico de* $[n]$ *que no está contenido ni en B_1 ni B_2, etc.*

Ejemplo 1 *Para entender un poco más esta última definición, hagamos*

- $n = 5$ *implica* $[n] = \{1,2,3,4,5\}$ *una posible partición sería* $\pi = \{B_1, B_2, B_3, B_4, B_5\}$ *donde* $B_1 = \{1\}, \quad B_2 = \{2\}, \quad B_3 = \{3\}, \quad B_4 = \{4\}, \quad B_5 = \{5\}$ *llamamos* singuletes *a los bloques que contienen sólamente un elemento de* $[n]$*, en este caso la partición π es una partición de puros singuletes.*

- $n = 6$ *implica* $[n] = \{1,2,3,4,5,6\}$ *una posible partición sería* $\pi = \{B_1, B_2, B_3, B_4\}$ *donde* $B_1 = \{1,3\}, \quad B_2 = \{2\}, \quad B_3 = \{4\}, \quad B_4 = \{5,6\}$ *podemos identificar dos singuletes los cuales son B_2 y B_3.*

Notación 2 *Se denotará $i \backsim_\pi j$ si i y j están en el mismo bloque de π.*

Definición 1.4.2 (Coalescente de Kingman) *Sea $n \in \mathbb{N} \cup \{\infty\}$. El n-coalescente de Kingman es una cadena de Markov a tiempo continuo, $(\Pi_t^n)_{t \geq 0}$, con valores en $\mathscr{P}_n$, tal que:*

1. *Inicialmente Π_0^n es la partición de n singuletes.*

2. *Las tasas de transición $q(\pi, \pi')$ son positivas si y sólo si π' es obtenida de unir dos bloques de π y en tal caso $q(\pi, \pi') = 1$.*

Ahora se explicará con más profundidad todo lo que implica el coalescente de Kingman y como se trabajó con él en simulaciones de genealogías y mutaciones.

Empecemos por exponer de una forma más clara cómo sería el espacio de estados en este proceso aleatorio. Para entender esto hay que considerar que la forma en que vamos a estudiar los coalescentes es viendo hacia el pasado, lo que significa que a tiempo cero se

empezará con todos los linajes separados; cada bloque del coalescente representa un linaje y cada número un individuo distinto, por lo que inicialmente veremos a cada bloque siendo un singulete. Cuando dos bloques se juntan, o sea cuando tenemos $B_i \cup B_j$ para algún i, j quiere decir que los linajes de los individuos que pertenecen a B_i y B_j se unen. Por tanto como se parte de un estado donde todos son singuletes, más formalmente partimos de una partición

$$\pi_0 = \{\underbrace{\{1\},\{2\},\ldots,\{n\}}_{n}\}$$

y dentro de un cierto tiempo brincamos a otra partición (estado) pero ahora bajo los supuestos del coalescente de Kingman sabemos que sólo se permite coalescer a dos linajes a la vez y cada linaje tiene la misma probabilidad de coalescer con cualquier otro, así se juntará una pareja de los n linajes existentes, con esto obtenemos $\binom{n}{2}$ particiones posibles a las cuales puede brincar nuestro proceso aleatorio. Cabe aclarar que el proceso no puede regresar a su estado original donde todos son singuletes y además el proceso tiene como único estado absorbente

$$\pi = \{1,2,\ldots,n\}.$$

Posterior al primer paso, cuando coalescen dos individuos el numero de linajes presentes se reduce a n-1,

$$\pi_1 = \{\underbrace{\{1\},\{2\},\ldots,\{j,k\},\ldots,\{n\}}_{n-1}\}$$

por lo que ahora existirán $\binom{n-1}{2}$ formas para pasar al siguiente estado y así sucesivamente.

A continuación se presentará un ejemplo con un diagrama para comprender mejor cómo se conforma el espacio de estados y cómo se van formando las particiones respecto a los estados que va tomando el proceso.

Con la Figura 1.6 podemos darnos una idea del inmenso tamaño que tendría la matriz de transición del coalescente de Kingman, solamente en el pequeño caso del ejemplo anterior

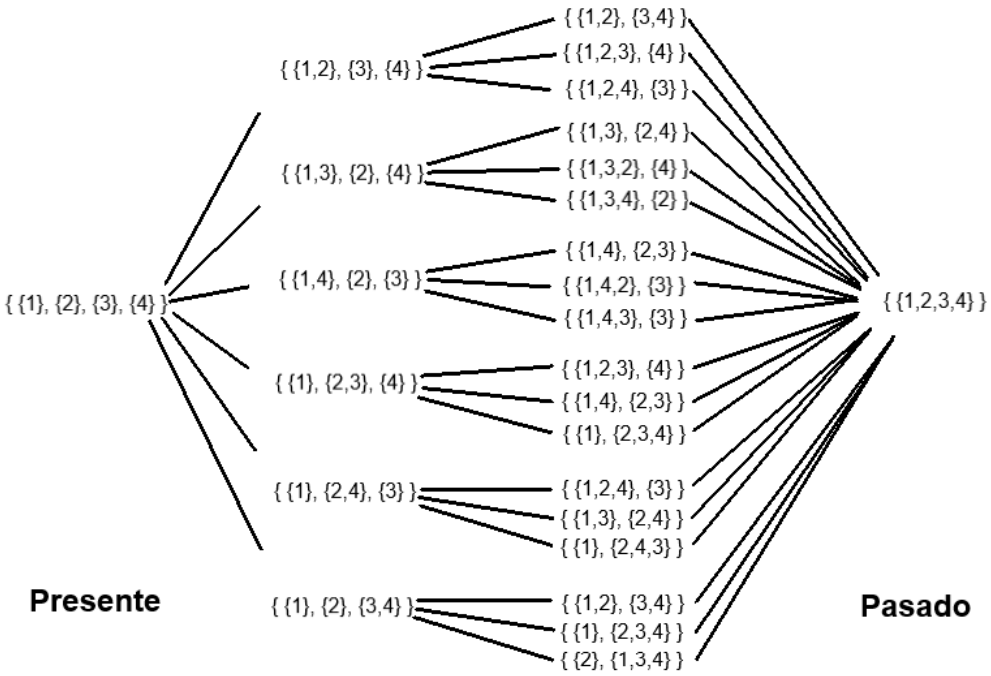

Figura 1.6: En esta imagen podemos ver un ejemplo de las distintas formas en que se comporta el proceso conforme avanza en el tiempo y observar las particiones o estados del proceso, que aunque en este ejemplo muchas de ellas se repitan da una idea clara de las diferentes posibilidades que hay cada que avanza. En este caso n=4 y si nos encontramos al inicio del proceso, o sea donde hay singuletes, tenemos $\binom{n}{2} = \binom{4}{2} = 6$ posibles estados a los que puede ir el proceso, en el siguiente paso vemos que hay $\binom{n-1}{2} = \binom{3}{2} = 3$ posibles estados a los cuales se puede ir y para el último caso hay $\binom{n-2}{2} = \binom{2}{2} = 1$ posibilidades y ahi termina el proceso.

la matriz constaría de 15 renglones y 15 columnas. Por lo que se suele dejar de lado el uso de la matriz de transición.

Continuando con el coalescente de Kingman, una de las cosas que le vamos a pedir, como se vió en la definición 2 es que cada pareja de bloques coalesce a tasa 1, o escrito de otra forma $q(\pi,\pi') = 1$ (recordemos que π' es alguno de los estados accesibles desde π), que por lo visto en la sección 1.3 equivale a lo siguiente

$$\lambda_i \cdot Q(i,j) = 1 \quad .$$

Esto implica que por cada partición a la que podamos acceder desde el estado en el que se encuentre el proceso tendremos una tasa igual a 1, por lo que saber el valor de λ_i se resume a saber a cuantas particiones se tiene acceso, o lo que es lo mismo, dado que estamos en i cuantos parejas de bloques se pueden formar, es así como se conforma la tasa de salto

$$\boxed{\lambda_i = \binom{n}{2} \quad \text{con} \quad n \quad \text{el numero de bloques en el estado} \quad i.}$$

Es así como podemos ver como en cada tiempo tenemos una media $\frac{2}{n(n-1)}$, ya que al tratarse de distribuciones exponenciales la media es el inverso del parámetro.

Una de las ventajas de las simulaciones por computadora es que, por medio de ellas, a través de los coalescentes nos ayudan a estudiar de mejor manera la genética de poblaciones. El coalescente de Kingman es el coalescente más sencillo que podemos encontrar y para simularlo se requieren usar una serie de algoritmos que explicaremos en breve. Primero que nada explicaremos una forma sencilla y efectiva para representar la unión de linajes, debido a que conllevaría un gran problema, de hecho escencial, el no saber visualizar un coalescente, por ejemplo, puede suceder que si lo representamos mal, al momento de querer simularlo computacionalmente el programa sea poco eficiente y conlleve un gran número de horas ejecutarlo. Una de las maneras para representar el Kingman con las que se trabajó es la siguiente:

1. Tomamos un número cualquiera n el cuál representa el tamaño de la muestra y por tanto el número de singuletes con los que empezaremos el procedimiento.

2. Posteriormente consideraremos un vector inicial con n número de "0"(ceros), cada uno de éstos representa a cada uno de los singuletes, de forma tal que la primera entrada del vector será el singulete $\{1\}$, el segundo "0" del vector será el singulete $\{2\}$ y así sucesivamente.

3. Como siguiente paso se formará otro vector con el mismo número de entradas, es decir n, pero ahora con n-2 "0" y dos "1". Lo anterior lo haremos ya que este nuevo vector representa el primer paso donde hay una coalescencia y va a consistir de la siguiente manera; las entradas del vector de los bloques o singuletes que coalescieron tendrán un nuevo valor el cual será "1", pues es el *primer* paso en el cual ocurrió la coalescencia. Todos las demás entradas del vector seguirán siendo "0". Más formalmente, si $i \in B_i$, $j \in B_j$ con i, j singuletes y $\pi_1 = \{B_1, \ldots, B_i \cup B_j, \ldots, B_n\}$, entonces el vector recien formado en las entradas i y j valdrá "1". Refiriéndonos a π_1 la partición a la que salta en el primer paso el coalescente.

4. Para el próximo paso, el número dos, ocurrirán alguna de las siguientes dos opciones, que son:

   a) que coalezcan dos singuletes, o;

   b) que coalezcan el bloque que representa la coalescencia anterior con algún singulete, es decir, que $B_i \cup B_j$ (los cuales son representados por los "1" del vector) se una con algún otro bloque.

   En el caso de que ocurra la primera opción, pasará lo mismo que con lo que se hizo en el paso anterior, que es, a cada entrada del vector de los singuletes coalescidos se

sustituirá el valor de "0" pero ahora por "2" pues vamos en el segundo paso. Si sucede la segunda opción a lo que se precederá es a sustituir el valor "1" y "0" por el "2" del segundo paso, sólo que por cada elemento que coalesció del bloque en el paso anterior en el nuevo vector se cambiará por el número "2". Escribiéndo más formalmente este último procedimiento, llamamos $B_k = B_i \cup B_j$ el bloque donde están i, j, B_l el bloque que contiene al singulete $\{l\}$, y ahora hacemos la unión $B_k \cup B_l$, lo cual nos llevaría a la siguiente partición $\pi_2 = \{B_1, \ldots, B_k \bigcup B_l, \ldots B_n\}$, con esto tenemos que $i \backsim_\pi j \backsim_\pi l$, por lo tanto, las entradas i, j, l del nuevo vector valdrán "2" .

5. Posteriormente ocurrirá lo mismo sólo que sustituyendo en el vector por el número del paso en el que vamos.

6. Sucederá de esta forma hasta que llegue el paso final del proceso, en el cuál el último vector tendrá cada entrada igual a "$n-1$". Esto significa que toda la muestra ya se encuentra dentro de un solo bloque.

En otras palabras, las entradas en el vector de todos los elementos que estén relacionados tendrán el mismo valor.

A continuación se verá a modo de ejemplo lo que se acaba de explicar anteriormente para que quede más claro como se forma el coalescente.

Ejemplo 2 *1. Tomando un caso sencillo con* $[n] = [4]$. *Sabemos que empezamos con una partición llena de singuletes.*

$$\pi_0 = \{\{1\},\{2\},\{3\},\{4\}\} \longrightarrow \upsilon_0 = (0,0,0,0)$$

$$\pi_1 = \{\{1,3\},\{2\},\{4\}\} \longrightarrow \upsilon_1 = (1,0,1,0)$$

$$\pi_2 = \{\{1,3\},\{2,4\}\} \longrightarrow \upsilon_2 = (1,2,1,2)$$

$$\pi_3 = \{\{1,2,3,4\}\} \longrightarrow \upsilon_3 = (3,3,3,3)$$

2. *Otro ejemplo tomando* $[n] = [5]$

$$\pi_0 = \{\{1\},\{2\},\{3\},\{4\},\{5\}\} \longrightarrow \upsilon_0 = (0,0,0,0,0)$$

$$\pi_1 = \{\{1\},\{2,3\},\{4\},\{5\}\} \longrightarrow \upsilon_1 = (0,1,1,0,0)$$

$$\pi_2 = \{\{1\},\{2,3,4\},\{5\}\} \longrightarrow \upsilon_2 = (0,2,2,2,0)$$

$$\pi_3 = \{\{1,2,3,4\},\{5\}\} \longrightarrow \upsilon_3 = (3,3,3,3,0)$$

$$\pi_4 = \{\{1,2,3,4,5\}\} \longrightarrow \upsilon_4 = (4,4,4,4,4)$$

Además de cómo se representará el coalescente de Kingman, es necesario saber cómo se escogen en cada paso los bloques que se van a unir, lo harán de manera aleatoria y uniformemente. Recordemos que, excepto con los singuletes, todos los elementos que están relacionados estarán representados con un mismo numero en el vector de simulación.

Posteriormente se busca en la simulación representar los tiempos de coalescencia, los cuales serán guardados en un vector donde cada entrada significará el tiempo de coalescencia de cada paso. Como ya se vio anteriormente, los tiempos de salto serán regidos por una

ley *exponencial* con parámetro $\binom{m}{2}$ donde m es el número de linajes presentes en cada paso de la simulación.

En la Figura 1.7 se muestra como se vería la realización de un coalescente de Kingman, en este ocasión se representó el segundo caso del Ejemplo 2.

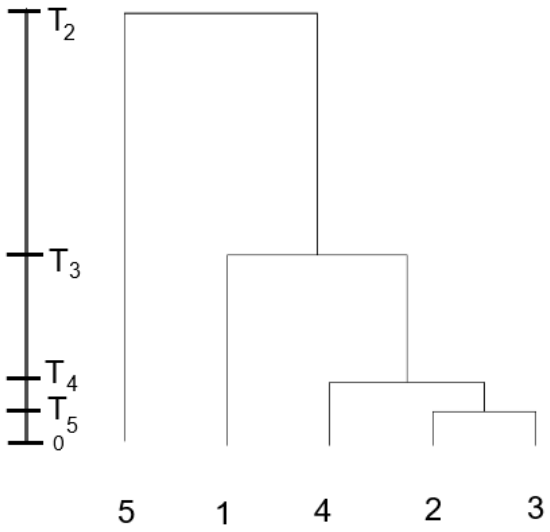

Figura 1.7: Grafica hecha a partir de un coalescente tomado del segundo inciso del Ejemplo 2 con referencia de tiempos de coalescencia.

Después se buscó simular las mutaciones en cada parte del coalescente según los tiempos y se contabilizó el número de mutaciones que influían sobre los linajes. A continuación se hablará sobre las mutaciones en el coalescente de Kingman.

Antes de la siguiente subsección donde se seguirá hablando de las mutaciones en el coalescente de Kingman, daremos un pequeño contexto para entender un poco como además del Kingman, existen diversas formas de coalescentes, un ejemplo de ellos son los Λ – *coalescentes*, este tipo de modelos ya permite la unión de dos o más individuos simultáneamente, mientras que exiten otros modelos más generales que permiten que para dos o más grupos los individuos de cada uno de esos grupos coalezcan simultaneamente.

Hacemos notar que existen más tipos de modelos de coalescentes, como se puede obser-

var en la Figura 1.8[3], unos por ejemplo son particularidades de otros que son más generales. Como el Kingman que es una particularidad del $\Lambda - coalescente$.

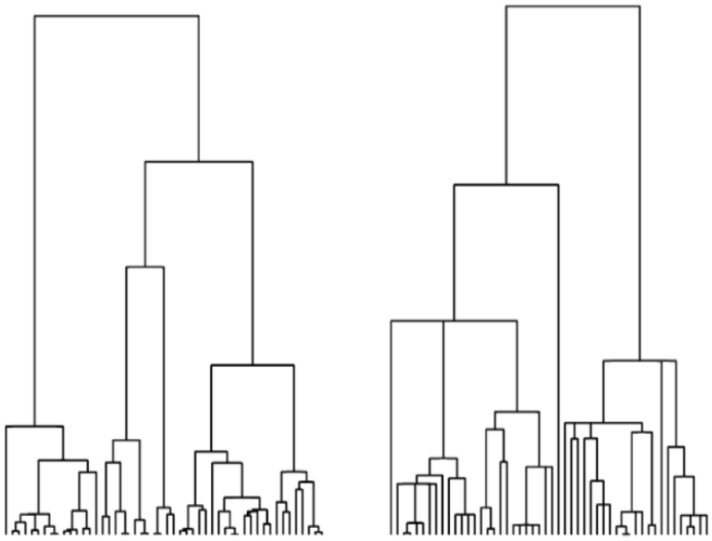

Figura 1.8: Se muestran dos simulaciones, la de la izquierda pertenece a un coalescente de Kingman y la de la derecha a un Λ - coalescente, ambas con 50 individuos. [3]

1.5. Agregar mutaciones al Coalescente de Kingman

En la teoría de coalescencias entenderemos a una mutación como un cambio en la secuencia genética del ADN, lo que significa un cambio en alguna base nitrogenada de un segmento o segmentos de ADN en los individuos de la muestra de la población a estudiar.

Como se vió anteriormente, las mutaciones en el coalescente de Kingman nacen también del modelo Wright-Fisher. Recordemos de la subsección 1.2 que la tasa de mutación es de la forma $\theta = 2N\mu$, desde luego entre más grande sea θ más mutaciones habrá. La diferencia es

[3]Kersting, G. & Wakolbinger, A. (2020). Probabilistic aspects of Λ - coalescents in equilibrium and in evolution. Cornell University. arXiv:2002.05250.

que en el coalescente de Kingman estamos contando al tiempo por una unidad $2N$, por lo que entre cada coalescencia hay $2N$ generaciones en las cuáles en cada una de ellas pueden surgir mutaciones, es de esta manera que se podría decir que en cualquier momento y en cualquier linaje pueden surgir mutaciones en el coalescente de Kingman de acuerdo a la tasa.

A la hora de simular el coalescente se quiso recrear, también, el número total de mutaciones que afectarían a los linajes del Kingman, lo cual se buscó hacer por medio de una ley Poisson, pero ésta ley incluye otro factor en sus parámetros además del parámetro θ, el cual es la longitud del coalescente, pues el número de mutaciones también estará directamente relacionado con el tiempo en lo que se tarda en encontrar el ancestro común más cercano de la muestra. Por lo que el total de mutaciones en el proceso, denotado por S, tiene una distribución

$$S \backsim Poi(\theta L)$$

donde L es la longitud de todas las ramas del árbol, es decir, la longitud del coalescente de Kingman.

Adicionalmente existe un conjunto de estadísticos igual de importantes de analizar y que están estrechamente relacionados con el número total de mutaciones S, estos son los estadísticos S_i. Observando la Figura 1.9 podemos entender dichos estadísticos de la siguiente forma: si contamos el número de mutaciones que afectaron a un solo individuo estaríamos contando las mutaciones que afectaron a los singuletes únicamente, en este caso S_1 sería el total de todas esas mutaciones; para S_2 nos preguntamos ahora por los linajes que mutaron y que dichas mutaciones influyen o afectan a dos individuos de la muestra original entonces estas mutaciones son transmitidas necesariamente de un padre a sus dos descendientes; para S_3 consideraremos pensar en una mutación que afectó a 3 individuos, entonces puede

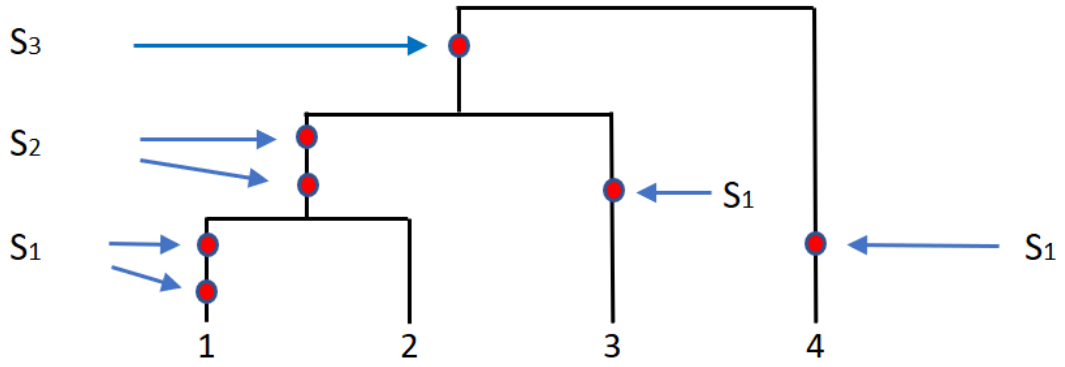

Figura 1.9: Visualización de los S_i.

que ésta haya sido heredada, por ejemplo, de un gen a dos descendencias una de las cuales tiene a su vez otras dos descendencias, y así sucesivamente. Si prestamos atención, si sumamos cada una de estas formas de contar las mutaciones nos dará exactamente el número total S. Más formalmente:

Definición 1.5.1 *En una muestra de n individuos con sus respectivas secuencias de ADN, decimos que una mutación es de nivel i, $1 \leq i \leq n-1$, si esa mutación se ve reflejada en i secuencias de ADN de la muestra. Por lo tanto llamaremos S_i a la cantidad total de mutaciones de nivel i que ocurrieron.*

A lo dicho anteriormente le sigue un concepto muy importante que se introduce en esta teoría y es a lo que vamos a llamar el *Espectro de Frecuencias de Sitio* que es el vector conformado en cada entrada por los valores S_i, o lo que es lo mismo el vector

$$(\quad S_1 \quad , \quad S_2 \quad , \quad \ldots \quad , \quad S_{n-1} \quad).$$

Partiendo como base de la Figura 1.7 ejemplificaremos como se verían reflejadas las mutaciones gráficamente y a su vez como se vería al momento de observar la muestra, es decir,

al momento de observar los códigos genéticos de la generación actual.

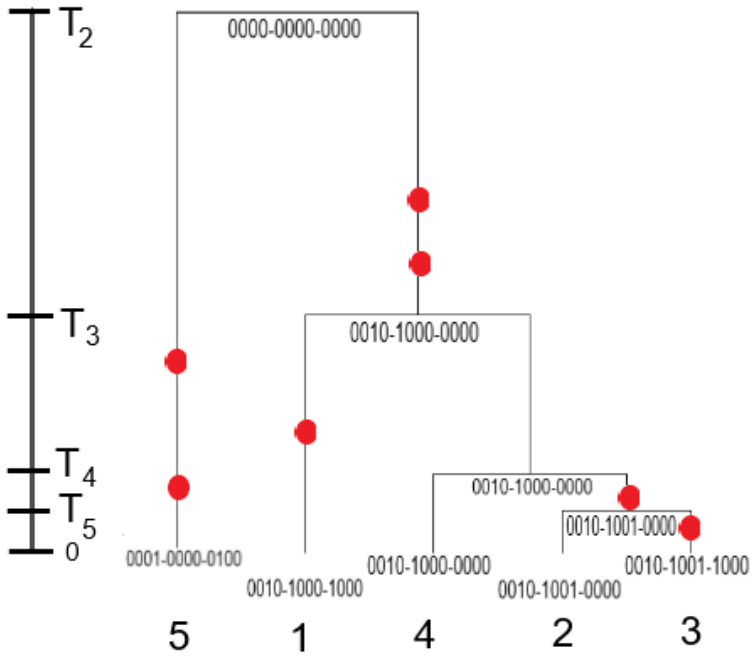

Figura 1.10: Mutaciones y como se heredan.

Ejemplo 3 *Notemos a continuación un árbol de coalescencias en el cual representaremos unas mutaciones gráficamente con un punto rojo como en la Figura 1.10, en el caso de este ejemplo tendríamos un código genético ancestral dado, representaremos a este código ancestral como tres grupos de cuatro ceros donde cada* 0 *será alguna letra de las bases nitrogenadas del ADN, o sea, A - Adenina, T - Timina, G - Guanina, C - Citosina, y una mutación en este código será denotada por un* 1*, lo que implica que si algún lugar o sitio del código de la descendencia tiene un* 1 *significa que en ese sitio una mutación hizo cambiar una letra del código por alguna otra que lo diferenciaría del código ancestral. De manera que tenemos lo siguiente:*

0000 – 0000 – 0000 *es el código ancestral y éste a su vez tiene dos descendientes, de los cuáles el individuo 5 mutó dos veces, por lo cual se podría ver una mutación del siguiente estilo* 0001 – 0000 – 0100*; el otro descendiente (que da origen a los individuos 1, 2, 3, 4) también mutó dos veces* 0010 – 1000 – 0000*; este individuo en la siguiente generación tiene a su vez*

dos descendientes; el individuo 1 que tiene una nueva mutación, por lo que se podría ver de la siguiente forma $0010-1000-1000$ *(se puede observar como se hereda la misma mutación que tuvo el padre), el otro descendiente que da origen a los individuos 2, 3, 4 no presenta mutaciones, es así como su código queda tal cual cómo el del padre* $0010-1000-0000$*; a su vez este individuo tiene dos descendientes, el individuo 4 no presenta mutación de tal manera que su código se vería como el del padre y abuelo* $0010-1000-0000$*, el otro descendiente que da origen a los individuos 2 y 3 presenta una mutación de la siguiente forma* $0010-1001-0000$ *y de sus dos descendientes, el individuo 2 no mutó por lo que queda* $0010-1001-0000$ *y el individuo 3 mutó de la siguiente forma* $0010-1001-1000$.

De esta manera podemos observar como sería nuestro arbol genealógico:

Es así como podemos ver que el número de mutaciones termina siendo como sigue:

$$S_1 = 4, \quad S_2 = 1, \quad S_3 = 0, \quad S_4 = 2$$

siendo en total $S = 7$.

A continuación veremos cuál es la esperanza del número total de sitios segregados S, para esto recordemos que los tiempos de coalescencia T_k se distribuyen exponencialmente

$$T_k \sim exp\left(\frac{k(k-1)}{2}\right),$$

por lo tanto tienen una media $\frac{2}{k(k-1)}$ donde k es el número de linajes o bloques que hay al momento; supongamos que tomamos θ una tasa de mutación; ahora bien, definamos el largo total del coalescente L como la suma de todas sus ramas, es decir, si tomamos por ejemplo una muestra de 4 individuos L será

$$4 \cdot T_4 + 3 \cdot T_3 + 2 \cdot T_2$$

pues es el tiempo que tarda en suceder una coalescencia por el número de linajes existentes,

entonces, la longitud del coalescente está dada por

$$L = \sum_{k=2}^{n} kT_k.$$

Así, para encontrar el valor esperado del número total de mutaciones debemos considerar que implica tomar en cuenta todas las ramas del coalescente pues en cada una de ellas puede ocurrir una mutación y también, desde luego, considerar la tasa en la que ocurren las mutaciones que sería θ, por lo tanto,

$$\begin{aligned}
E(S) &= E(\theta \cdot L) \quad \text{por ser Poisson} \\
&= \theta E(L) \\
&= \theta E(\sum_{k=2}^{n} kT_k) \\
&= \theta \sum_{k=2}^{n} k \cdot E(T_k) \\
&= \theta \sum_{k=2}^{n} k \cdot \frac{2}{k(k-1)} \\
&= \theta \sum_{k=2}^{n} \frac{2}{k-1} \\
&= 2\theta \sum_{i=1}^{n-1} \frac{1}{i} \quad \text{haciendo un cambio de variable} \quad i = k-1.
\end{aligned}$$

Con esto tenemos un resultado importante

$$\boxed{E(S) = 2\theta \sum_{i=1}^{n-1} \frac{1}{i}.}$$

Ahora bien, en la simulación se buscó lanzar mutaciones a cada paso que realiza el coalescente por medio de un *Proceso Poisson* con tasa $\theta \cdot T_i \cdot i$, donde θ es la tasa de mutación, T_i el tiempo que transcurrió para que coalescieran los siguientes dos linajes y i son el número de ramas que hay en el tiempo i.

1.6. Código del Coalescente de Kingman

A continuación se mostrará el código que se creó para simular un coalescente de Kingman en el lenguaje de programación R, incluye las matrices de coalescencia, tiempos de coalescencia y mutaciones. Además se explicará, previo al código, como es que se fue constituyendo la simulación.

1. La simulación se construyó con base en la representación que se expuso en el ejemplo 2, de esta manera se construyó en el código una matriz que representa un Coalescente de Kingman, ésta se construyó a partir de un vector que se le llamó "x" en el código. De ese vector "x" con la ayuda de un vector auxiliar "$vecaux$" se escogían *uniformemente* dos individuos para que coalescieran.

2. Posteriormente se programaron los tiempos de coalescencia, que para efectos de practicidad en la simulación se decidió invertir el orden de los subíndices de los tiempos, es decir, en el vector que simula los tiempos en el código, la primera entrada representa el tiempo en el que coalescen los primeros dos individuos, la segunda entrada será el tiempo en el que ocurrió la segunda coalescencia y así sucesivamente. Los tiempos se programaron con base en *exponenciales* de parámetro $\binom{n-i}{2}$ donde i son los linajes o bloques que se encuentran en ese momento.

3. Las mutaciones que se simularon en cada etapa de coalescencia se representaron con un vector llamado "poi" con una tasa $\theta \cdot T_i \cdot (n-i)$, se usó $n-i$ para ajustarlo al código. Así, las entradas del vector serán el número de mutaciones en cada etapa de coalescencia.

4. Después para efectos del código se usó una vector "$contador$" para contar del vector "x" cuántos singuletes habían, los cuales se representaron en "$contador$" con un 1,

además este vector contaba de cada grupo donde hubieran coalescido individuos el número de linajes o bloques que hubiera dentro de dicho grupo, es decir, si tuviéramos un arreglo de la siguiente forma el vector contador se vería:

$v_0 = (0,0,0,0,0) \rightarrow c_0 = (1,1,1,1,1)$pues son puros singuletes

$v_1 = (0,1,1,0,0) \rightarrow c_1 = (1,1,1,2)$hay 1 grupo con 2 números uno y 3 singuletes

$v_2 = (0,1,1,2,2) \rightarrow c_2 = (1,2,2)$2 grupos con dos números cada uno y 1 singulete

$v_3 = (0,3,3,3,3) \rightarrow c_3 = (1,4)$hay un grupo con 4 números tres y 1 singulete

$v_4 = (4,4,4,4,4) \rightarrow c_4 = (5)$

5. Con base en ese vector "$contador$" se tomó una muestra con repetición, de tamaño igual a la entrada del vector "poi" correspondiente según la etapa de coalescencia. Por ejemplo, si resulta que para v_2 hubo 5 mutaciones, entonces, se toma de c_2 una muestra uniforme con repetición de 5 elementos. Así, si por ejemplo, se seleccionara dos veces el 1, dos veces el primer 2 y una vez el segundo 2, significaría que $S_1 = 2$ pues hubo dos mutaciones afectando un singulete, $S_2 = 3$, asumiendo desde luego que no hubieron más mutaciones en las otras etapas.

6. De manera que como se dijo en el punto anterior se formó el vector de S_i que representa al espectro de frecuencias de sitio.

Aunado a esta explicación se añadieron comentarios en la mayoría de los pasos del código para dejar más clara su construcción.

A continuación se muestra el código que se realizó para simular el coalescente de Kingman:

```
#### Coalescente de Kingman simple hecho en una matriz,  con tiempos
#### exponenciales para cada coalescencia, numero de mutaciones,
#### vector S (espectro de frecuencia de sitios)

### Se inserta un número n entero positivo que corresponde con el
###tamaño de la muestra y p un numero real positivo que corresponde
###con la tasa de mutacion teta

CKingmanS <- function(n,p){

  ###Variables declaradas para la funcion
  ###El vector x representa los individuos. 0 son singuletes, 1
  ### representa a los individuos que coalescieron en la primera etapa,
  ### 2 los que lo hicieron en la segunda, etc.
  cero <- rep( 0 , n-2 )
  x <- c( 1 , 1 , cero )
  vecaux <- rep( 0 , 2 )
  MatrizKingman <- matrix( 0 , ncol = n , nrow = n-1 )
  MatrizContadora <- matrix( NA , ncol = n , nrow= n )

  ###Vector x y matriz  del Coalescente de Kingman
  x <- sample( x , n )
  MatrizKingman[1,] <- x
```

```
###Vector de tiempos y realización del primer tiempo de coalescencia
tiempo <- rep( 0 , n )
tiempo[1] <- rexp( 1 , choose( n , 2 ) )

###Vector Poisson que simula el numero de mutaciones por etapa de
###coalescencia y realización del primer número de mutaciones
###correspondiente al tiempo donde sólo  hay singuletes
poi <- rep( 0 , n )
poi[1] <- rpois( 1 , p*tiempo[1]*n )


###Vector y matriz contador del numero de linajes por nivel
###Aqui se hará el contador del vector y el lanzamiento de mutaciones
### primero sobre el nivel donde no hay coalescencias
###El vector S representa al espectro de frecuencias de sitios
contador <- rep( 1 , n )
s <- rep( 0 , n-1 )
saux <- sample( contador , poi[1] , replace = TRUE )
s[1] <- length( which( saux == 1 ) )
MatrizContadora[1,] <- contador

###Aqui es donde se hará el lanzamiento de mutaciones para el nivel
###donde ocurre la primer coalescencia
contadoraux <- rep( 0 , n )
```

```
contadoraux[1:2] <- c(length(which(x==0) ),length(which(x==1) ) )

contador <- c( rep( 1 , contadoraux[1] ) , contadoraux[2:n] )

contador <- contador[ -which( contador == 0 ) ]

MatrizContadora[ 2 , 1:length( contador ) ] <- contador


tiempo[2] <- rexp( 1 , choose( n-1 , 2 ) )

poi[2] <- rpois( 1 , p*tiempo[2]*(n-1) )

saux <- sample( contador , poi[2] , replace = TRUE )

s[1] <- s[1] + length( which( saux == 1 ) )

s[2] <- length( which( saux == 2 ) )



##Aquí inicia la simulación para la segunda etapa de coalescencia y

##el resto de las etapas

for(i in 2:(n-1)){


 ##El vector auxiliar tomará una muestra de numeros del vector x

 ##que serán los siguientes individuos que coalescerán

 vecaux <- sample( which( x <= i ) , 2 )


 ##El if actua para que el programa no presente warnings

 if( i != n-1 ){

    tiempo[i+1] <- rexp( 1 , choose( n-i , 2 ) )

    poi[i+1] <- rpois( 1, p*tiempo[i+1]*(n-i) )
```

```
 }

##Con este while nos aseguramos que las entradas del vector
##auxiliar sean distintas o unicamente sean iguales cuando sean
##iguales a cero. Así nos aseguramos de no volver elegir a los
##mismos dos individuos que ya coalescieron
while( !xor( x[ vecaux[1] ] != x[ vecaux[2] ] ,
          ( x[ vecaux[1] ] == 0 & x[ vecaux[2] ] == 0) )   ){

  vecaux <- sample(which(x <= i),2)
 }

 ##Aqui se pedirá que a las entradas de x que tengan el mismo valor
 ##que las entradas 1 y 2 del vector auxiliar sean cambiadas por el
 ##nuevo valor de i(la etapa en la que va el coalescente)
 x[ which(  (x[vecaux[1]] == x) & x != 0 ) ] <- i
 x[ which(  (x[vecaux[2]] == x) & x != 0 ) ] <- i
 x[vecaux] <- i

 ##Con este for la entrada 1 del contadoraux cuenta los 0 que tenga
 ## x, la entrada 2 cuenta cuantos 1 hay en x y asi sucesivamente
 for(j in 0:n ){
  contadoraux[j+1] <- length( which( x == j ) )
 }
```

```
##Ajustes finales para que el vector contador ordene a la izquierda
##los 1's que (en este vector) representan cada singulete, y al
##final del vector la cantidad de bloques que coalescieron en
##cada grupo.
contador <- c(  rep( 1 , contadoraux[1] )  ,
                contadoraux[ 2:length( contadoraux ) ]   )

contador <- contador[-which( contador == 0 ) ]
MatrizContadora[ i+1 , 1:length( contador ) ] <- contador

##Suelta uniformemente mutaciones sobre el vector contador según
##el número de mutaciones que se hayan obtenido en el tiempo en el
##que se encuentra el proceso
saux <- sample( contador , poi[i+1] , replace = TRUE )

##A cada entrada del vector S le agrega las mutaciones respectivas
for(k in 1:n-1){
  s[k] <- s[k]+length( which( saux == k))
}

##Representación de la simulación en una Matriz
MatrizKingman[i,] <- x
```

```
  }

  listaresultados <- list( "Tiempos" = tiempo , "Matriz" = MatrizKingman
                           ,  "Contador" = MatrizContadora , "Mutaciones"
                           = poi , "VectorS" = s )
  return(listaresultados)


}
```

A continuación se presentará un ejemplo de una simulación tomando 10 individuos y tasa de mutación igual a 3. Se añaden los resultados que arroja R a través de la consola.

```
> a <- CKingmanS(10 , 3)
> a$Matriz
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    0    1    0    0    0    1    0    0     0
 [2,]    0    0    2    0    0    2    2    0    0     0
 [3,]    0    3    3    0    0    3    3    0    0     0
 [4,]    0    4    4    0    0    4    4    4    0     0
 [5,]    0    5    5    0    0    5    5    5    0     5
 [6,]    0    5    5    6    6    5    5    5    0     5
 [7,]    0    7    7    7    7    7    7    7    0     7
 [8,]    0    8    8    8    8    8    8    8    8     8
 [9,]    9    9    9    9    9    9    9    9    9     9
```

```
> a$Contador
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    1    1    1    1    1    1    1    1    1     1
 [2,]    1    1    1    1    1    1    1    1    2    NA
 [3,]    1    1    1    1    1    1    1    3   NA    NA
 [4,]    1    1    1    1    1    1    4   NA   NA    NA
 [5,]    1    1    1    1    1    5   NA   NA   NA    NA
 [6,]    1    1    1    1    6   NA   NA   NA   NA    NA
 [7,]    1    1    6    2   NA   NA   NA   NA   NA    NA
 [8,]    1    1    8   NA   NA   NA   NA   NA   NA    NA
 [9,]    1    9   NA   NA   NA   NA   NA   NA   NA    NA
[10,]   10   NA   NA   NA   NA   NA   NA   NA   NA    NA
> a$Tiempos
 [1] 0.0596719114 0.0009603377 0.0616048436 0.0029267820 0.0692108506 0.0963445846
 [7] 0.0444017133 0.2446070416 1.2650440381          NaN
> a$Mutaciones
 [1] 1 0 0 0 0 3 0 2 0 0
> a$VectorS
[1] 5 0 0 0 0 0 0 1 0
```

Capítulo 2

Estadísticas de la diversidad

2.1. ¿Cómo observar los datos?

Hay que considerar que desde el punto de vista de la teoría de coalescencias se consideran muestras que no contemplan individuos en sí sino alguna porción del material genético (o ADN) de dichos individuos, a través de distintas técnicas se observa el material genético de un gen, lo que nos ayudará a observar las diferentes variaciones que pudieran existir entre los individuos de la muestra. Ahora, de una manera más práctica se pueden observar los datos de la siguiente manera, vea la Figura 2.1[1]:

[1]Anderson, S., Bankier, A., Barrell, B. et al. (1981). Sequence and organization of the human mitochondrial genome. Nature 290, 457–465.

```
   TCCGCTCTGTCCCCGCCCTGTTCTTA
 1 .......CA.T....T.......... 3
 2 .........A.T....T.......... 2
 3 ..........T....T.......... 1
 4 ..........T....T........C. 1
 5 .T.A..T...T.....T..A....C. 2
 6 .T.A......T........A....C. 2
 7 CT.A......T..T.....A....C. 1
 8 .T.A......T.....T..A....C. 2
 9 CT........T.....T..A....C. 2
10 .T........T.....T..A....CG 1
11 .T........T.....T..A....C. 5
12 .T..............T..A....C. 9
13 .T......A.......T..A....C. 1
14 .T........T....TT..A....C. 1
15 .T........T....TT..AC...C. 2
16 ..........TT..........T.C. 1
17 ....T.....T.........C...C. 1
18 ....T.....T.........C..C. 2
19 ..T.......T......T...C..C. 1
20 .....C....T...A......C..C. 3
21 ..........T..........C..C. 3
22 C.........T..........C.... 3
23 ..........TT......C..C.... 1
24 ..........T.......C..CT... 7
25 ..........TT......C..CTC.. 3
26 ...........T......C..CTC.. 1
27 .......C.C.................. 1
28 .......C.C..T.............. 1
```

Figura 2.1: Diferenciaciones de secuencias genéticas entre individuos. [1]

Como se muestra en la Figura 2.1, tenemos en la parte superior una secuencia de bases nitrogenadas que representan un gen, también llamada *haplotipo*, este haplotipo pertenece al individuo que servirá como referencia y debajo de esta secuencia se encuentran las representaciones de ese mismo gen que tienen los diferentes individuos tomados de una muestra, los espacios con puntos representan las bases nitrogenadas donde no ocurrieron cambios, sin embargo, en los sitios donde ocurrió una mutación se representó por medio de la base nitrogenada a la cual mutó. Además del lado derecho de cada uno de los 28 diferentes haplotipos se tiene escrito un número, el cual significa el número de veces que se observó esa misma secuencia en la muestra que se tomó, por ejemplo, si observamos el haplotipo número 12 vemos que al final de la secuencia está escrito un 9: Lo que significa que hubo 9 individuos que presentaron esas mismas cuatro mutaciones en los mismo sitios que se presentan en el haplotipo 12 y que mutaron hacia las mismas bases nitrogenadas. La secuencia genética que se presenta consta de 26 sitios y se tomó una muestra de 63 individuos en total.

2.2. Modelo de Sitios Infinitos

Motoo Kimura en su afán por desarrollar la teoría neutral de la evolución construyó dos modelos, los cuales nos ayudan a entender la evolución molecular, la permanencia de los alelos y la variación genética. Éstos son el *Modelo de Alelos Infinitos* que fue creado en 1964 y el *Modelo de Sitios Infinitos* el cual fue construido en 1969.

El *Modelo de Alelos Infinitos* supone que cada mutación genera un alelo distinto a cualquier otro anterior. Cabe destacar que este modelo llegó en una época en la que para obtener el material genético de algún individuo se hacía por medio de métodos no tan directos, como la electroforesis, que tiene sus inicios a principios del siglo XIX y que a grandes rasgos hace uso de campos eléctricos en soluciones acuosas para separar moléculas y determinar

sus composiciones.

Por otro lado el *Modelo de Sitios Infinitos* o *ISM* por sus siglas en inglés (*Infinite Sites Model*) es un modelo que asume que cada que ocurre una mutación ésta misma solamente ocurre en un sitio nuevo, o lo que es lo mismo, una nueva mutación no ocurrirá en un mismo sitio donde ya ha habido alguna mutación anteriormente. A modo de ejemplo, tomemos una secuencia de bases nitrogenadas, ACCGTAACC la cuál será la secuencia ancestral.

ACCGTAACC

ACCTTAACC

ACCGTAACC

Si suponemos que cada renlgón que tiene una secuencia de bases nitrogenadas es una generación, y el primer renglón representa al código ancestral, entonces podemos ver que el descendiente del segundo renglón si cumple con el supuesto del modelo pues presenta una mutación en el cuarto sitio, pero el ultimo descendiente no cumple con el supuesto principal del Modelo de Sitios Infinitos pues presenta una mutación sobre el cuarto sitio que además lo hace regresar al código ancestral. Esta suposición resulta bastante significativa por problemas como el anterior en la cuál no podríamos ver la mutación de la segunda secuencia en este caso.

El ISM, como cualquier modelo es un intento de representar la realidad. Desde luego, si se pueden presentar muchos casos como el expuesto en el ejemplo donde haya mutaciones en el mismo sitio. Esta suposición es importante porque aún cuando se tome de una muestra grande unos códigos genéticos relativamente extensos, por ejemplo, de 300 caracteres puede suceder que una mutación toque al mismo sitio en más de una ocasión pues con una tasa θ lo suficientemente grande y una cadena de bases nitrogenadas lo suficientemente corta, este caso se parecería al famoso ejemplo de probabilidad del cumpleaños, donde, de un grupo

dos o más personas cumplen años el mismo día. Por lo que la probabilidad de que exista más de una mutación en el mismo sitio puede llegar a ser alta.

Ahora bien, para este modelo un concepto muy importante es uno que se trató anteriormente que resulta ser el número total de mutaciones S. Este concepto va a dar paso a uno de los estadísticos más importantes para conocer el valor de la tasa de mutación θ que es el estimador de Waterson. Antes de empezar a hablar de este estimador continuemos extrayendo más información del número total de sitios segregados S. Ya se ha determinado la esperanza de S en la sección 1.5, que sabemos fue $E(S) = 2\theta \sum_{i=1}^{n-1} \frac{1}{i}$, a continuación, daremos a conocer la varianza de S. Primero observemos que la distribución de $S \mid L \backsim Poi(\theta L)$ y sabemos que $E(X) = E(E(X \mid Y))$.

Ahora recordemos que $V(S) = E\left(S^2\right) - E^2(S)$,

$$
\begin{aligned}
\text{entonces} \quad E\left(S^2\right) &= E\left(E\left(S^2 \mid L\right)\right) = E\left(\theta^2 L^2 + \theta L\right), \\
\text{así} \quad V(S) &= E\left(S^2\right) - E^2(S) = E\left(\theta^2 L^2 + \theta L\right) - E^2(\theta L) = V(\theta L) + E(\theta L), \\
\text{pero la varianza} \quad V(\theta L) &= \theta^2 V(L) = \theta^2 V\left(\sum_{k=2}^{n} k T_k\right) \\
&= \theta^2 \sum_{k=2}^{n} V(k T_k) \quad \text{por ser variables aleatorias independientes} \\
&= \theta^2 \sum_{k=2}^{n} k^2 V(T_k) = \theta^2 \sum_{k=2}^{n} k^2 \left(\frac{1}{\left(\frac{k(k-1)}{2}\right)^2}\right) \\
&= \theta^2 \sum_{k=2}^{n} k^2 \left(\frac{4}{k^2(k-1)^2}\right) \\
&= 4\theta^2 \sum_{k=2}^{n} \frac{1}{(k-1)^2} \\
&= 4\theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \quad \text{con} \quad i = k-1,
\end{aligned}
$$

por lo que

$$\boxed{V(S) = 2\theta \sum_{i=1}^{n-1} \frac{1}{i} + 4\theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}.}$$

Todo lo anterior nos servirá para conformar y entender al estimador de Watterson.

2.3. Estimador de Watterson

El estimador de Watterson fue desarrollado en la década de 1970 por Margaret Wu y Geoffrey Anton Watterson, su propósito principal es estimar el valor de θ tomando en cuenta el tamaño de la muestra y el número de sitios segregados. Este estimador es muy usado debido a su simplicidad. Se define de la seguiente manera:

$$\boxed{\theta_W = \frac{S}{2\sum_{i=1}^{n-1}\frac{1}{i}}.}$$

Bajo las condiciones del Modelo de Sitios Infinitos, el estimador de Watterson, con una muestra suficientemente grande y estandarizándolo su distribución se aproxima a la normal.

Como podemos ver, el estimador de Watterson es de hecho un estimador insesgado para θ, ya que

$$E(\theta_W) = \theta.$$

Su varianza

$$\boxed{V(\theta_W) = \frac{\theta}{2\sum_{i=1}^{n-1}\frac{1}{i}} + \frac{\theta^2\sum_{i=1}^{n-1}\frac{1}{i^2}}{\left(\sum_{i=1}^{n-1}\frac{1}{i}\right)^2}.}$$

Ejemplo 4 *Tomaremos una muestra de tamaño 5, para así poder usar después este mismo ejemplo. También, la primer secuencia representará al elemento de referencia y el "0"significará que no hubo cambio en ese sitio.*

1. TTACG CCGAA TGGAC

2. 00C0T 00000 A0000

3. A0000 00000 A0000

4. 00000 GG000 00T00

5. 00C0T G000C 00000

Se obtendrá ahora el estimador de Watterson para este ejemplo. Tenemos así 8 mutaciones diferentes en los 15 sitios que hay. Con base en estos datos podemos determinar que el número de sitios segregantes

$$S = 8$$

y al tener 5 elementos en la muestra tenemos que el denominador es $\sum_{i=1}^{5-1} \frac{1}{i}$*, por lo que*

$$\begin{aligned} \theta_W &= \frac{8}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} \\ &= \frac{8}{2{,}0833} = 3{,}84. \end{aligned}$$

Observemos que no se contaron todas las mutaciones que se observaron sino el número total de sitios donde ocurrieron dichas mutaciones, ya que en total se contabilizarían 12 mutaciones y el número de sitios segregantes es de 8.

El estimador de Watterson, además, servirá para conformar un estadístico de prueba, el cual tiene distintas interpretaciones tanto biológicamente como matematicamente. A este estadístico de prueba se le conoce como el *D de Tajima*. Se ha explicado como funciona el estimador de Watterson, pero, cabe aclarar que este estimador tiene ciertas deficiencias en cuanto a su precisión al estimar θ, esto debido a que depende mucho del tamaño de la muestra.

2.4. Espectro de Frecuencias de Sitio y Estimador de Fu y Li

En la subsección 1.5 se mencionó el concepto del *espectro de frecuencias de sitio* que consiste en el vector

$$(S_1, S_2, S_3, \cdots, S_{n-1})$$

recordando que el valor S_i corresponde al número de mutaciones que afectan a i individuos de la muestra. A continuación se detallará más acerca del *espectro de frecuencias de sitios* y se hablará del estimador de Fu y Li.

Recordemos que bajo la hipótesis del *Modelo de Sitios Infinitos* tenemos la siguiente identidad, $\sum S_i = S$, esto puesto que cada mutación ocurre en un sitio distinto de tal manera que cuando se suman las mutaciones de cada nivel i no se está contando varias veces una misma mutación. Con dicha identidad obtenemos una relación del espectro de frecuencias de sitios con el estimador de Watterson y es que

$$\boxed{\theta_W = \frac{\sum_{i=1}^{n-1} S_i}{\sum_{i=1}^{n-1} \frac{1}{i}}.}$$

Debido a la importancia de los S_i Fu estuvo trabajando para extraerles más información, entre otras cosas, de acuerdo con Durrett[2] se dieron cuenta de que muchos estimadores insesgados para θ se podían escribir de la siguiente forma

$$\widehat{\theta} = \sum_{i=1}^{n-1} \alpha_i \cdot S_i,$$

donde α_i es un factor que depende de i y en muchos casos de n también. Igualmente, otro punto muy importante que se desarrolló es la esperanza, además de la varianza y la covarianza de los S_i. Empezando por la esperanza se tiene que

$$\boxed{E(S_i) = \frac{2\theta}{i}.}$$

[2] Durrett, R. (2008). Probability Models for DNA Sequence Evolution(p. 59-60). Springer-Verlag New York.

Para probar esto nos fijaremos en algún nivel k. Un nivel k lo definiremos como el tiempo t en el coalescente en el cual hay k linajes o bloques en el árbol. Llamaremos J_l^k el número de individuos de la muestra que son descendientes de la rama o linaje número l en el nivel k. Para entender el siguiente lema y parte de la prueba explicaremos brevemente el concepto de la variable J_l^k.

Ejemplo 5 *Supongamos que tenemos una muestra $n = 10$ y estamos en el nivel $k = 5$ en el árbol y tenemos una partición de la siguiente manera*

$$\{\{1,2,5\},\{4,7\},\{3,6\},\{8,9\},\{10\}\}$$

podemos ver como hay 5 linajes y la rama 1 del árbol correspondería con $\{1,2,5\}$, así $J_1^5 = 3$. Por lo tanto el vector de enteros positivos $(j_1, j_2, j_3, j_4, j_5)$ que suma hasta n y que corresponde con esta partición es

$$(J_1^5 = j_1, J_2^5 = j_2, J_3^5 = j_3, J_4^5 = j_4, J_5^5 = j_5) = (J_1^5 = 3, J_2^5 = 2, J_3^5 = 2, J_4^5 = 2, J_5^5 = 1).$$

Se hará uso de un resultado sencillo que viene demostrado y explicado más a profundidad en el Durrett[3] y es el siguiente,

Lema 2.4.1 *La distribución del vector de enteros positivos que suma hasta n, es decir, el vector $(J_1^5, J_2^5, J_3^5, J_4^5, J_5^5)$ tiene una distribución uniforme, a saber*

$$P\left(J_1^5 = j_1, J_2^5 = j_2, J_3^5 = j_3, J_4^5 = j_4, J_5^5 = j_5\right) = \frac{1}{\binom{n-1}{k-1}}.$$

Explicando la igualdad del lema anterior retomemos el ejemplo 5. Supongamos que los 10 individuos, sin importar su etiqueta, son representados con un 0 y vamos a separar cada

[3] Durrett, R. (2008). Probability Models for DNA Sequence Evolution(p. 54). Springer-Verlag New York.

linaje por medio de una línea |, por lo que la partición $\{\{1,2,5\},\{4,7\},\{3,6\},\{8,9\},\{10\}\}$ se vería así,

$$000 \mid 00 \mid 00 \mid 00 \mid 0$$

fijémonos que las líneas las podemos colocar únicamente en el espacio que hay entre cada dos 0, en este caso tenemos 9 espacios por ser en total 10 elementos, lo que significa que, en general, para n individuos tenemos $n-1$ espacios y tenemos $k-1$ líneas | pues hay k linajes por ser nivel k. Por lo que buscamos un solo arreglo de un total de $\binom{n-1}{k-1}$ que son los casos totales. Con esto dicho, si ahora nos fijamos en la probabilidad de que algún J_l^k tenga un valor en particular i, esto es , para $i \in [1, n-k+1]$

$$P\left(J_l^k = i\right) = \frac{\binom{n-1-i}{k-2}}{\binom{n-1}{k-1}} \quad , \tag{2.1}$$

lo cual, para entenderlo volvamos al ejemplo 5 y veamos la probabilidad de que, por ejemplo, $J_2^5 = 2$, de esta forma digamos que apartamos esos dos individuos de los ocho restantes, por lo que, tenemos ocho 0 que no han sido separados

$$00 \mid 00000000$$

con lo cual, para los ocho individuos restantes tenemos $n-1-i = 7$ espacios donde colocar las líneas y como ya se usó una línea para separar el linaje con dos elementos, entonces sólo se pueden ocupar $k-2 = 3$ líneas. Definimos L_i como la longitud de todas las ramas que tienen i descendientes. Con esto nos preguntaremos por la esperanza de L_i para así concluir con que $E(S_i) = E(\theta L_i)$. Por lo que, como es de esperarse, se tiene que probar que $E(L_i) = \frac{2}{i}$.

Pero previo a esto revisemos dos resultados importantes que servirán para calcular $E(L_i)$.

$$\text{Primero demostremos que} \quad \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}\frac{1}{k-1} = \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}}\frac{1}{i}. \tag{2.2}$$

La demostración es la siguiente,

$$\begin{aligned}
\frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}\frac{1}{k-1} &= \frac{\frac{(n-i-1)!}{(n-i-1-(k-2))!(k-2)!}}{\frac{(n-1)!}{(n-1-(k-1))!(k-1)!}}\frac{1}{k-1} \\
&= \frac{(n-i-1)!(n-k)!(k-1)!}{(n-1)!(n-i-k+1)!(k-2)!\cdot(k-1)} \\
&= \frac{(n-i-1)!(n-k)!}{(n-1)!(n-i-k+1)!} \quad \text{cancelando los} \quad (k-1)! \\
&= \frac{\frac{(n-k)!}{(n-k-(i-1))!(i-1)!}}{\frac{(n-1)!}{(n-1-i)!i!}}\cdot\frac{1}{i} \quad \text{multiplicando por un} \quad 1 = \frac{i!}{i(i-1)!} \\
&= \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}}\frac{1}{i}.
\end{aligned}$$

Segundo, observemos la siguiente ecuación

$$\sum_{k=2}^{n}\frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} = 1 \tag{2.3}$$

y veamos que cuando escojamos i objetos de $n-1$ y suponiendo que el mínimo en escogerse es $k-1$, entonces faltan ahora $i-1$ objetos por escoger de las siguientes $n-k = n-1-(k-1)$, lo anterior por cada $k \in [2,n]$.

Ahora bien tomemos la esperanza de L_i, para esto tendremos en cuenta la $P\left(J_l^k = i\right)$, consideraremos además la esperanza del tiempo que dura el proceso en el nivel k que es $\frac{2}{k(k-1)}$ y por último también se tomará en cuenta el número de ramas que hay en el árbol en

el nivel k, que son, evidentemente, k ramas. Resumiendo,

$$E(L_i) = \sum_{k=2}^{n} \underbrace{k}_{\text{No. de ramas}} \cdot \underbrace{\frac{2}{k(k-1)}}_{E(T_k)} \cdot \underbrace{\frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}}_{P\left(J_l^k = i\right)}$$

Por lo tanto,

$$\begin{aligned}
E(L_i) &= \sum_{k=2}^{n} k \cdot \frac{2}{k(k-1)} \cdot \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \\
&= \sum_{k=2}^{n} 2 \cdot \frac{1}{k-1} \cdot \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \\
\text{por la ecuación (2.2)} \quad &= \sum_{k=2}^{n} 2 \cdot \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} \frac{1}{i} = 2 \cdot \frac{1}{i} \cdot \frac{\sum_{k=2}^{n} \binom{n-k}{i-1}}{\binom{n-1}{i}} \\
\text{por la ecuación (2.3)} \quad &= \frac{2}{i} \cdot 1 = \frac{2}{i}.
\end{aligned}$$

Por lo tanto si consideramos $E(S_i) = \theta E(L_i)$ tenemos que

$$E(S_i) = \frac{2\theta}{i} \quad .$$

2.5. Estimador de Tajima

El estimador de Tajima fue propuesto por el investigador japonés Fumio Tajima en 1983. Se basa en las diferencias entre bases nitrogenadas que hay en las secuencias de ADN entre los individuos de la muestra, esto lo hace por pares de secuencias, es decir, compara las diferencias entre los sitios donde hubo alguna mutación. Al comparar el número de sitios

segregados por cada dos individuos implica que se tiene que involucrar en el estimador el conteo del total de parejas que se pueden formar dada una muestra de tamaño n, o lo que es lo mismo, las combinaciones de $\binom{n}{2}$, de hecho, se puede decir que toma un promedio de todas las diferencias en los sitios de mutaciones por cada par que puede haber. Por lo que el estimador de Tajima está dado por la siguiente ecuación,

$$\theta_T = \frac{\sum\sum_{i<j} k_{ij}}{\binom{n}{2}},$$

donde k_{ij} son el número de diferencias de bases nitrogenadas que hay entre la $i-ésima$ y $j-ésima$ secuencia de ADN de la muestra y n el tamaño de la muestra de individuos.

Ejemplo 6 *Tomaremos en cuenta el ejemplo 4 usado para el estimador de Watterson. Con* $n = 5$ *y el número total de casos que vamos a manejar* $\binom{5}{2} = 10$. *A continuación contaremos en cuantos sitios difieren las secuencias de ADN.*

1. TTACG CCGAA TGGAC

2. 00C0T 00000 A0000 entre estas 2 secuencias hay 3 diferencias (o polimorfismos).

1. TTACG CCGAA TGGAC

3. A0000 00000 A0000 entre estas 2 secuencias hay 2 polimorfismos.

1. TTACG CCGAA TGGAC

4. 00000 GG000 00T00 entre estas 2 secuencias hay 3 polimorfismos.

1. TTACG CCGAA TGGAC

5. 00C0T G000C 00000 entre estas 2 secuencias hay 4 polimorfismos.

2. 00C0T 00000 A0000

3. A0000 00000 A0000 entre estas 2 secuencias hay 3 polimorfismos distintos.

2. 00C0T 00000 A0000

4. 00000 GG000 00T00 entre estas 2 secuencias hay 6 polimorfismos distintos.

2. 00C0T 00000 A0000

5. 00C0T G000C 00000 entre estas 2 secuencias hay 3 polimorfismos.

3. A0000 00000 A0000

4. 00000 GG000 00T00 entre estas 2 secuencias hay 5 polimorfismos.

3. A0000 00000 A0000

5. 00C0T G000C 00000 entre estas 2 secuencias hay 6 polimorfismos.

4. 00000 GG000 00T00

5. 00C0T G000C 00000 entre estas 2 secuencias hay 5 polimorfismos.

Con esto podemos ya calcular el estimador de Tajima para la muestra en cuestion, que sería,

$$\begin{aligned}\theta_T &= \frac{3+2+3+4+3+6+3+5+6+5}{\binom{5}{2}} \\ &= \frac{40}{10} = 4.\end{aligned}$$

Podemos ver, con el ejemplo, como el estimador de Tajima dependiendo de que tan grande sea el tamaño de la muestra conllevaría a un proceso más tardado para obtenerse. Afortu-

nadamente existe, entre otros, un método que en casos de n grande nos permitiría obtener más facilmente θ_T y es el siguiente,

$$\theta_T = \sum_{i=1}^{S} h_i \quad \text{, con} \quad h_i = \frac{n\left(1 - \sum_j x_{ji}^2\right)}{n-1},$$

y donde x_{ji}^2 es la frecuencia del $j-ésimo$ alelo de la muestra en el $i-ésimo$ sitio de mutación y S el número de sitios segregados. Es decir, para obtener x_{ji} tomaremos la frecuencia de la base nitrogenada que "no mutó", que es el alelo de la secuencia de referencia y a esto le sumamos la frecuencia de la base nitrogenada que si mutó, todo lo anterior sobre el mismo sitio de mutación. Se explicará a continuación obteniendo el mismo resultado del ejemplo anterior pero con este otro método.

Ejemplo 7 *Tomemos la misma muestra con $n = 5$ y obtengamos θ_T.*

1. TTACG CCGAA TGGAC

2. 00C0T 00000 A0000

3. A0000 00000 A0000

4. 00000 GG000 00T00

5. 00C0T G000C 00000

La primer mutación se encuentra en el sitio 1. Obtengamos h_1

$$h_1 = \frac{5\left(1 - \left(\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right)\right)}{4} = \frac{2}{5}.$$

Notemos que $x_{11} = \frac{1}{5}$, pues en el primer sitio de mutación tenemos solo una base nitrogenada A de un total de cinco que representa a la mutación y $x_{21} = \frac{4}{5}$, ya que representa a la base nitrogenada de referencia T, la cual, se encuentra en el primer sitio, en cuatro de las cinco secuencias. Ahora la segunda mutación se encuentra en el sitio 3. Obtengamos h_2,

$$h_2 = \frac{5\left(1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right)\right)}{4} = \frac{3}{5},$$

entonces, similarmente, $x_{12} = \frac{2}{5}$ *pues en el tercer sitio encontramos que dos bases nitrogenadas mutaron a C y* $x_{22} = \frac{3}{5}$ *porque en el tercer sitio la base nitrogenada que no mutó, o sea, A se encuentra en tres de los cinco sitios. Análogamente sucede lo mismo con* h_3, h_4, h_5, h_6, h_7, h_8, *por lo que*

$$\theta_T = \frac{2}{5} + \frac{3}{5} + \frac{3}{5} + \frac{3}{5} + \frac{2}{5} + \frac{2}{5} + \frac{3}{5} + \frac{2}{5} = \frac{20}{5} = 4.$$

Con este resultado, vemos como para muestras de tamaño grande no es necesario revisar pareja por pareja de secuencias las diferencias en las mutaciones.

2.6. Tamaño del Clado Mínimo

Por último, uno de los temas que se han estudiado en los años recientes debido a su efectividad para medir la calidad del modelo es el *tamaño del clado mínimo* que hace referencia a la familia o grupo más chico con el cual un individuo se relaciona, explicándolo más formalmente, en una muestra el tamaño del clado mínimo de cierto individuo siendo este un singulete es el número de elementos de la muestra que están en el bloque en el cual ocurrió la coalescencia de dicho singulete.

Ejemplo 8 *Supongamos para una muestra de tamaño 10 que el proceso se encuentra en un tiempo* T_k. *Considerando un coalescente de Kingman observemos los linajes en el tiempo* k,

$$\pi_{T_k} = \{\{1\}, \{2,3,4\}, \{5\}, \{6,7,8,9,10\}\},$$

ahora, fijémonos en los linajes para el tiempo $k+1$

$$\pi_{T_{k+1}} = \{\{1\}, \{2,3,4\}, \{5,6,7,8,9,10\}\}.$$

Fijémonos en el individuo 5, *podemos ver que la familia con la cual ocurrió la coalescenscencia constaba de cinco individuos por lo que el tamaño de clado mínimo para el individuo*

número 5 *es seis. Además notemos que por tratarse de una coalescencia tipo Kingman podemos inferir que la familia a la cual se une el individuo* 5 *proviene de varias subfamilias pequeñas.*

Además, hace falta recalcar que el tamaño del clado mínimo es una manera con la cual podemos medir que tan distante es un individuo con respecto del resto de la muestra a estudiar, o lo que es lo mismo, nos ayuda a ver que elementos están mas relacionados en cuanto a parentesco genético.

Se tiene que hacer énfasis en uno de los problemas sustanciales que tiene este concepto y es que el tamaño del grupo más chico no es observable puesto que para saber que elementos de la muestra están más emparentados con el gen a estudiar se tiene que saber las mutaciones en común que tienen tanto el gen como los demás genes que estarían dentro de la familia más próxima; por lo cual, si no tenemos una mutación en común para el clado mínimo no tenemos la certeza de cuales son los individuos más emparentados. A continuación se presenta la Figura 2.2 en la que se distinguirá de mejor manera el hecho de no observar una mutación en el clado mínimo. Fijémonos en el individuo 3, notemos que en el árbol genealógico del lado izquierdo tenemos dos mutaciones una de ellas es un S_3 y la otra es un S_1, y en el árbol del lado derecho tenemos una mutación de tipo S_7 y otra de tipo S_1. De esta manera podemos observar que para ambos árboles para el individuo 3 tenemos un tamaño del clado mínimo igual a 3 puesto que se junta con el bloque que tiene a los elementos 1 y 2; pero a la hora de observar únicamente los elementos de la muestra, con el árbol de la izquierda debido a la mutación del tipo S_3 podemos ver que el tamaño observable del clado mínimo es 3, sin embargo, para el árbol de la derecha tenemos una mutación de tipo S_7 lo cual le da a los primeros 7 individuos el mismo código genético que el del individuo 3. Con esto podemos decir que el tamaño del clado mínimo no es observable y dependemos de que

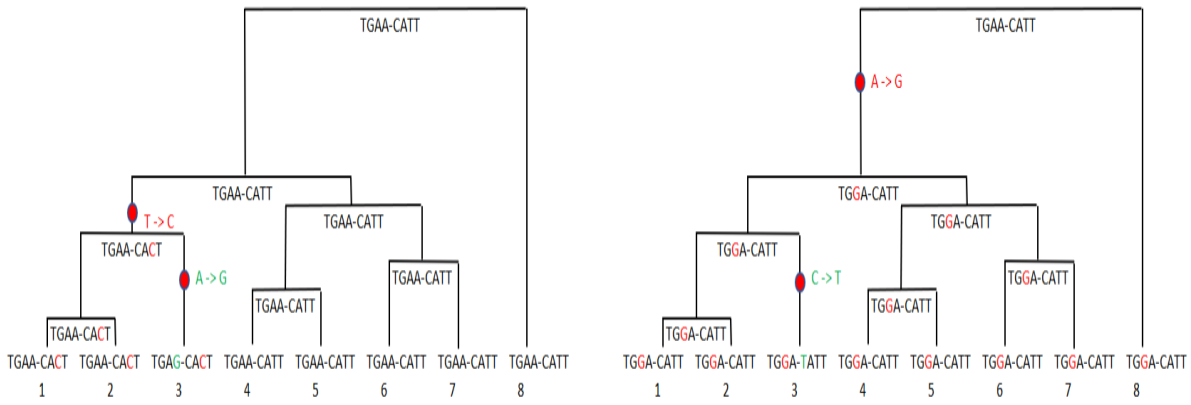

Figura 2.2: Con una muestra de tamaño 8 tenemos dos arboles genealógicos idénticos. Se propone una pequeña secuencia genética con 8 sitios, cada uno con una base nitrogenada. Se distinguen los cambios en la descendencia generados por mutaciones con diferentes colores sobre las bases nitrogenadas.

ocurra alguna mutación para un familia en la cual se encuentre el gen de interés.

A pesar de la dificultad que existe para observar el tamaño del clado mínimo se puede saber que su distribución es, de hecho, una Yule de parámetro 2. Si llamamos X_n al tamaño de la familia más chica de un individuo de la muestra escogido aleatoriamente pero fijo, se probará que

$$\boxed{P(X_n = x) = \frac{4}{(x-1)x(x+1)}.}$$

Para demostrar la igualdad anterior se darán una serie de preámbulos, primero buscaremos encontrar la probabilidad de que el número de linajes sea k cuando dicho gen coalezca, es decir, buscamos $P(K = k)$ con $k = 2, 3, \cdots, n-1, n$.

Para empezar supongamos que nos fijamos en un individuo arbitrario pero fijo de la muestra al cual llamaremos el gen 1. Buscamos una partición $\pi_k \in \mathscr{P}_n$ la cual tenga k clases de equivalencia o bloques $B_1, B_2, \ldots, B_k$.

Definición 2.6.1

- *Se define el tiempo de coalescencia del gen* 1 *con el resto de la muestra como*

$$\tau_1 = sup\{t \geq 0 \mid \{1\} \in \Pi_t^n\}.$$

- *Llamaremos también* $L_k = T_n + \cdots + T_k$.
- *Un clado es un bloque* $B_* \in \Pi_i^n$ *con* $i = 1, \ldots, n-1$ *de tal forma que el tiempo desde el ancestro común más reciente de los individuos en* B_* *es* L_{i+1}. *En este caso i corresponde al número de ancestros en la muestra al momento de la coalescencia.*
- *El clado mínimo del gen* 1 *será el bloque* B_1 *en la partición* $\Pi_{\tau_1}^n$ *de tal manera que* $1 \in B_1$.
- *Definimos el tamaño del clado mínimo como* $X_n = \mid B_1 \mid$.

Ahora bien el nivel de coalescencia K, donde el gen 1 se junta con el resto de la muestra, es una variable aleatoria tal que $K = k$ si y solo si $\tau_1 = L_k$ con $k = 2, \ldots, n$. Introduciremos a continuación un resultado clásico e importante con el fin de obtener la distribución de K.

Lema 2.6.1 $P(\pi_k = \beta) = \frac{(n-k)!k!(k-1)!}{n!(n-1)!} \cdot n_1! \cdots n_k!$.

donde $\beta \in \mathscr{P}_n$ con β_i los bloques o clases de equivalencia de la partición β, y cada bloque con cardinalidad $\mid \beta_i \mid = n_i$ de modo que $n_1 + \cdots + n_k = n$. Observemos para nuestro gen 1 que para que ocurra la coalescencia a un nivel más bajo que k, o sea, $k-1$ clases por ejemplo, se puede ver como la probabilidad de que 1 se encuentre como singulete en una partición π_k, es decir,

$$P(K \leq k) = P(\{1\} \in \pi_k).$$

Pero podemos dividir la partición π_k de la siguiente manera $\pi_k = \beta \cup \{1\}$ donde β es una partición con $k-1$ clases, o sea, $\mid \beta \mid = k-1$, por lo que habría que considerar todas las particiones

del conjunto $[n]-\{1\}$, por lo que tenemos

$$P(K \leq k) = P(\{1\} \in \pi_k) = \sum_{\beta} P(\pi_k = \beta \cup 1).$$

Ahora bien, gracias al Lema 2.6.1 como tenemos una partición β con $k-1$ bloques y un singulete, tenemos que

$$\sum_{\beta} P(\pi_k = \beta \cup 1) = \sum_{\beta} \frac{(n-k)!k!(k-1)!}{n!(n-1)!} \cdot n_1! \cdots n_{k-1}!$$

ahora, desglocemos un poco el lado derecho de la igualdad anterior

$$\sum_{\beta} \frac{(n-k)!k!(k-1)!}{n!(n-1)!} \cdot n_1! \cdots n_{k-1}! = \sum_{\beta} \frac{(n-k)!(k)(k-1)!(k-1)(k-2)!}{(n)(n-1)!(n-1)(n-2)!} \cdot n_1! \cdots n_{k-1}!$$

como la suma del total de probabilidades da 1 y tenemos que $n_1 + \cdots + n_{k-1} = n-1$, entonces tenemos que

$$\sum_{\beta} \frac{(n-k)!(k-1)!(k-2)!}{(n-1)!(n-2)!} \cdot n_1! \cdots n_{k-1}! = 1,$$

con lo cual vemos que

$$\sum_{\beta} \frac{(n-k)!(k)(k-1)!(k-1)(k-2)!}{(n)(n-1)!(n-1)(n-2)!} \cdot n_1! \cdots n_{k-1}! = \frac{k(k-1)}{n(n-1)}$$

ya con esto tenemos lo que se quería probar respecto a la distribución de K, que

$$P(K \leq k) = \frac{k(k-1)}{n(n-1)}.$$

Así, para obtener la probabilidad de $K = k$ hacemos una resta

$$P(K = k) = P(K \leq k) - P(K \leq k-1) = \frac{k(k-1)}{n(n-1)} - \frac{(k-1)(k-2)}{n(n-1)} = \frac{(k-1)2}{n(n-1)}.$$

A continuación debemos probar otro lema para posteriormente demostrar la probabilidad del clado mínimo,

Lema 2.6.2 $\sum_{k=x-2}^{n-3} k(k-1)\cdots(k-x+3)(n-k-1)(n-k-2) = 2\frac{n(n-1)\cdots(n-x)}{(x-1)x(x+1)}$, *donde* x *es un número entero fijo,* $4 \leq n$ *y se cumple el siguiente requisito* $3 \leq x < n$.

Para demostrar este teorema veámoslo de forma más general. Empecemos por considerar dos enteros k, y tal que $1 \leq y < k$. Ahora revisemos la siguiente diferencia,

$$\begin{aligned}(k+1)k\cdots(k-y+1) - k(k-1)\cdots(k-y+1)(k-y) &= \big[k+1-(k-y)\big]\big[k(k-1)\cdots(k-y+1)\big] \\ &= (y+1)k(k-1)\cdots(k-y+1),\end{aligned}$$

donde cada término de dicha resta tiene $y+1$ factores y exactamente y de estos coinciden, así podemos interpretar la siguiente suma como una suma telescópica, para $m > y$ tenemos

$$\begin{aligned}\sum_{k=y+1}^{m}(y+1)k(k-1)\cdots(k-y+1) &= \sum_{k=y+1}^{m}(k+1)k\cdots(k-y+1) - \sum_{k=y+1}^{m}k(k-1)\cdots(k-y+1)(k-y) \\ &= (m+1)m\cdots(m-y+1) - (y+1)!.\end{aligned}$$

Luego, en el caso particular cuando $k = y$ tenemos $k(k-1)\cdots(k-y+1) = y!$, por lo que para todo par de enteros $1 \leq y < m$ tenemos

$$\begin{aligned}\sum_{k=y}^{m}k(k-1)\cdots(k-y+1) &= y! + \frac{(m+1)m\cdots(m-y+1) - (y+1)!}{y+1} \\ &= \frac{(m+1)m\cdots(m-y+1)}{y+1}.\end{aligned}$$

Para acercarnos a la expresión del lema, definimos las funciones $F_0(m,y), F_1(m,y)$ y $F_2(m,y)$ con $1 \leq y \leq m$,

$$\begin{aligned}F_0(m,y) &= \sum_{k=y}^{m}k(k-1)\cdots(k-y+1) \\ F_1(m,y) &= \sum_{k=y}^{m}k(k-1)\cdots(k-y+1)(m+1-k) \\ F_2(m,y) &= \sum_{k=y}^{m}k(k-1)\cdots(k-y+1)(m+1-k)(m+2-k).\end{aligned}$$

Según nuestra construcción hemos demostrado que $F_0(m,y) = \frac{(m+1)m\cdots(m+1-y)}{y+1}$. Ahora bien, tomando $F_1(m,y) = \sum_{k=y}^{m} k(k-1)\cdots(k-y+1)(m+1-k)$ fijémonos en que podemos partir la suma en dos distribuyendo el factor $(m+1-k)$ viéndolo como $m+2$-veces $\sum_{k=y}^{m} k(k-1)\cdots(k-y+1)$ menos el último término (notemos que $m+1-k = m+2-(k+1)$), entonces

$$F_1(m,y) = (m+2)\sum_{k=y}^{m} k(k-1)\cdots(k-y+1) - \sum_{k=y}^{m} k(k-1)\cdots(k-y+1)(k+1).$$

Ahora, haciendo un cambio de variable $j = k+1$ tenemos

$$\begin{aligned} F_1(m,y) &= (m+2)\sum_{k=y}^{m} k(k-1)\cdots(k-y+1) - \sum_{j=y+1}^{m} j(j-1)\cdots(j-(y+1)+1) \\ &= (m+2)F_0(m,y) - F_0(m+1,y+1). \end{aligned}$$

Sustituyendo los valores que tenemos para F_0,

$$\begin{aligned} F_1(m,y) &= \frac{(m+2)(m+1)\cdots(m+1-y)}{y+1} - \frac{(m+2)(m+1)\cdots(m+1-(y+1)+1)}{y+2} \\ &= \frac{(m+2)(m+1)\cdots(m+1-y)}{(y+2)(y+1)}\left[(y+2)-(y+1)\right] \\ &= \frac{(m+2)(m+1)\cdots(m+1-y)}{(y+2)(y+1)}. \end{aligned}$$

Análogamente para $F_2(m,y)$ tenemos que $m+2-k = m+3-(k+1)$ y

$$\begin{aligned} F_2(m,y) &= (m+3)\sum_{k=y}^{m} k(k-1)\cdots(k-y+1)(m+1-k) - \sum_{k=y}^{m} k(k-1)\cdots(k-y+1)(m+1-k)(k+1) \\ &= (m+3)\sum_{k=y}^{m} k(k-1)\cdots(k-y+1)(m+1-k) - \sum_{j=y+1}^{m} j(j-1)\cdots(j-(y+1)+1)(m+1-j+1) \\ &= (m+3)F_1(m,y) - F_1(m+1,y+1). \end{aligned}$$

Sustituyendo los valores para F_1 tenemos,

$$\begin{aligned} F_2(m,y) &= \frac{(m+3)(m+2)\cdots(m+1-y)}{(y+2)(y+1)} - \frac{(m+3)\cdots(m+1-(y+1)+1)}{(y+3)(y+2)} \\ &= \frac{(m+3)(m+2)\cdots(m+1-y)}{(y+3)(y+2)(y+1)}\left[(y+3)-(y+1)\right] \\ &= 2\frac{(m+3)(m+2)\cdots(m+1-y)}{(y+3)(y+2)(y+1)}. \end{aligned}$$

Para concluir con la prueba, observemos que si $3 \leq x < n$ entonces $1 \leq x-2$ y de esta forma sustituyendo $m = n-3$ y $y = x-2$ tenemos

$$F_2(n-3,x-2) = \sum_{k=x-2}^{n-3} k(k-1)\cdots(k-x+3)(n-k-1)(n-k-2)$$

y $F_2(n-3,x-2) = 2\frac{n(n-1)\cdots(n-x)}{(x-1)x(x+1)}$. Con esto concluye la prueba del Lema.

Ya con estas demostraciones tenemos las herramientas para iniciar en la prueba del tamaño del clado mínimo. Recordemos primeramente la ecuación (2.1) de la subsección 2.4 que hace referencia a la probabilidad de que haya i individuos de la muestra descendientes de un linaje l en un nivel k, es decir,

$$P\left(J_l^k = i\right) = \frac{\binom{n-1-i}{k-2}}{\binom{n-1}{k-1}}$$

donde $1 \leq i \leq n-k+1$. Recordemos que nos estamos fijando en el gen 1, lo que quiere decir que "lo vamos a apartar del resto de los individuos de la muestra", entonces si condicionamos a que el nivel (número de linajes al momento de la coalescencia) sea $K = k$ tenemos que el tamaño del clado mínimo tiene una distribución de la siguiente forma

$$X_n \backsim 1 + J_l^{k-1}$$

de donde 1 es evidentemente el tamaño del bloque del gen 1 y J_l^{k-1} toma valores no sobre los n individuos de la muestra sino sobre los $n-1$ restantes y con $k-1$ linajes pues el k-ésimo

linaje es el linaje del gen 1. Entonces tenemos lo siguiente, cuando $k = 2$

$$P(X_n = n \mid K = 2) = 1$$

y cuando $k = 3, \ldots, n$ y $1 \leq i \leq n - k + 1$

$$P(X_n = 1 + i \mid K = k) = \frac{\binom{n-2-i}{k-3}}{\binom{n-2}{k-2}}.$$

Sin embargo, si no condicionamos tenemos los siguientes casos; el primero de ellos

$$P(X_n = n) = \frac{1}{\binom{n}{2}} = \frac{1}{\frac{n(n-1)}{2}} = \frac{2}{n(n-1)},$$

notemos que $P(X_n = n) = P(K = 2)$ ya que $X_n = n$ si y sólo si $K = 2$, ahora bien, este es el caso de la última coalescencia en el proceso, por lo que hay una posibilidad de entre $\binom{n}{2}$-combinaciones para que el gen 1 sea el último en coalescer con el resto de la muestra. Para el otro caso, de manera más general tenemos la siguiente probabilidad,

$$P(X_n = 1 + i) = \sum_{k=3}^{n-i+1} P(K = k) P(X_n = 1 + i \mid K = k),$$

observemos que la suma corre sobre k que representa al número de linajes que hay, entonces la probabilidad anterior equivale a

$$P(X_n = 1 + i) = \sum_{k=3}^{n-i+1} \frac{2(k-1)}{n(n-1)} \cdot \frac{\binom{n-2-i}{k-3}}{\binom{n-2}{k-2}}. \tag{2.4}$$

Estudiemos rápidamente un caso importante de la probabilidad anterior, y es que cuan-

do tomamos $i = 1$ obtenemos lo que sigue

$$\begin{aligned} P(X_n = 2) &= \sum_{k=3}^{n} \frac{2(k-1)}{n(n-1)} \frac{\binom{n-3}{k-3}}{\binom{n-2}{k-2}} \\ &= \sum_{k=3}^{n} \frac{2(k-1)}{n(n-1)} \frac{\frac{(n-3)!}{(n-k)!(k-3)!}}{\frac{(n-2)!}{(n-k)!(k-2)!}} \\ &= \sum_{k=3}^{n} \frac{2(k-1)(k-2)}{n(n-1)(n-2)} = \frac{2}{3}, \end{aligned}$$

lo que significa que lo más probable es que la primer coalescencia de nuestro gen 1 sea con solamente un individuo.

Tomando la ecuación (2.4) hacemos un cambio de variable $x = i+1$ donde nos queda lo que sigue

$$P(X_n = x) = \sum_{k=3}^{n+2-x} \frac{2(k-1) \cdot \binom{n-1-x}{k-3}}{n(n-1) \cdot \binom{n-2}{k-2}},$$

y haciendo un cambio de variable adicional de k por $n-k$ obtenemos lo siguiente

$$P(X_n = x) = \sum_{k=x-2}^{n-3} \frac{2(n-k-1) \cdot \binom{n-1-x}{n-k-3}}{n(n-1) \cdot \binom{n-2}{n-k-2}} = \frac{2(n-k-1) \cdot (n-1-x)! \cdot k! \cdot (n-k-2)!}{n(n-1) \cdot (n-2)! \cdot (k-x+2)! \cdot (n-k-3)!}.$$

Así para cualquier entero $x \in [3, n)$ tenemos

$$P(X_n = x) = \sum_{k=x-2}^{n-3} \frac{2k(k-1)\cdots(k-x+3)(n-k-1)(n-k-2)}{n(n-1)(n-2)\cdots(n-x)},$$

entonces

$$P(X_n = x) = \frac{2\sum_{k=x-2}^{n-3} k(k-1)\cdots(k-x+3)(n-k-1)(n-k-2)}{n(n-1)(n-2)\cdots(n-x)},$$

observemos que podemos aplicar el Lema 2.6.2, así sustituyendo tenemos

$$P(X_n = x) = \frac{2 \cdot \frac{2n(n-1)\cdots(n-x)}{(x-1)x(x+1)}}{n(n-1)(n-2)\cdots(n-x)}$$

por lo tanto finalmente tenemos

$$P(X_n = x) = \frac{4}{(x-1)x(x+1)}.$$

Que es una distribución Yule con parámetro 2 cuando $n \longrightarrow \infty$.

Conclusiones

En la presente tesis se dio a conocer, a partir del modelo Wright-Fisher, el coalescente de Kingman como un modelo que se puede ajustar a una población que se quiera estudiar, de entre otros diversos coalescentes el de Kingman es un modelo clásico dentro de la teoría y que se usa con mucha frecuencia. Sin embargo hay poblaciones que dada la naturaleza de la misma o debido a algún suceso extraordinario no podemos ajustarle este modelo, por ejemplo, suponiendo que estamos estudiando una población y a ésta le haya ocurrido una catastrofe como una extinción masiva y, por consecuencia, se tuviera un cuello de botella reflejado en la genealogía de la población, en este caso el comportamiento se asemejaría más a un coalescente múltiple. Dicho esto, a partir del supuesto del modelo de sitios infinitos se definieron diversas estadísticas que nos sirven para verificar si el modelo del coalescente de Kingman hace un buen ajuste a nuestros datos. Por ejemplo, se presentó tanto el estimador de Watterson

$$\theta_W = \frac{S}{2\sum_{i=1}^{n-1}\frac{1}{i}} \quad ,$$

y también se mostró al estimador de Tajima

$$\theta_T = \frac{\sum\sum_{i<j} k_{ij}}{\binom{n}{2}} \quad ,$$

que juntos son la escencia del famoso *D de Tajima* que está conformado como sigue,

$$D = \frac{\theta_T - \theta_W}{\sqrt{V(\theta_T - \theta_W)}}.$$

Con este estadístico de prueba tenemos, que, para no rechazar nuestro modelo esperamos que $D \approx 0$, o lo que es lo mismo, que $\theta_T \approx \theta_W$; ahora bien, si $D < 0$ se puede suponer, por ejemplo, que hubo un crecimiento en la población lo cual rompería con el supuesto de que la pobalción sea constante; y si $D > 0$ lo que se podría suponer, por ejemplo, es que hubo un reciente cuello de botella.

Entre los estadísticos que se formularon aquí, uno de los que más se ha usado a lo largo de los años en esta teoría es el Espectro de Frecuencias de Sitios (SFS), el cual a través de su media

$$E(S_i) = \frac{2\theta}{i}$$

podemos también saber si nuestro modelo es el correcto viendo si nuestros datos se aproximan a las respectivas esperanzas de los S_i.

Por último se mostró un tema de reciente creación, el Tamaño del Clado Mínimo, en el que pudimos obtener la distribución que sigue el tamaño de la familia más chica con respecto a un gen arbitrario pero fijo de la muestra, la distribución es la siguiente,

$$P(X_n = x) = \frac{4}{(x-1)x(x+1)} \quad .$$

Con estos dos últimos temas, el Tamaño del Clado Mínimo y el SFS, al juntar sus beneficios podemos tener una estadística muy buena debido a que puede reducir el error a la hora de seleccionar el modelo.

Bibliografía

1. Oluwafemi Oyamakin, S., Chukwu, A., Oluwaseun, W., Ogunjobi, E. (2019). Allele Based Inference on Evolution and Extinction; A Genetic Drift Approach. Journal of Cancer Genetics And Biomarkers - 1(4):1-15.

2. Rincón, L. (2013). Introducción a los procesos estocásticos. México: Facultad de Ciencias, UNAM.

3. Kersting, G. & Wakolbinger, A. (2020). Probabilistic aspects of Λ - coalescents in equilibrium and in evolution. Cornell University. arXiv:2002.05250.

4. Anderson, S., Bankier, A., Barrell, B. et al. (1981). Sequence and organization of the human mitochondrial genome. Nature 290, 457–465.

5. Hamilton, M. (2009). Population Genetics. United Kingdom: Wiley-Blackwell.

6. Durrett, R. (2008). Probability Models for DNA Sequence Evolution. New York: Springer-Verlag.

7. Berestycki, N. (2009). Recent progress in coalescent theory. arXiv:math.PR/0909.3985.

8. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

9. Blum, Michael François, Olivier. (2005). Minimal clade size and external branch length under the neutral coalescent. Advances in Applied Probability. 37. 10.1239/aap/1127483740.

10. Tavaré, S. (2004). Ancestral inference in population genetics. In *Lectures on Probability Theory and Statistics* (Lecture Notes Math. 1837), Springer, Berlin, pp. 1-188.