Universidad Nacional
Autónoma de México

Facultad de Ciencias

CADENAS DE MARKOV NO-HOMOGÉNEAS PARA EL MAPEO GENÉTICO DE POBLACIONES MEZCLADAS (*ADMIXTURE MAPPING*)

T E S I S

Que para obtener el título de
MATEMÁTICO

Presenta
JORGE ALAN MORALES MORILLÓN

Directoras de Tesis
DRA. ELIANE REGINA RODRIGUES
DRA. SANDRA ROMERO HIDALGO

2018

CIUDAD UNIVERSITARIA, CD.MX.

Hoja de información

1. **Datos del alumno**
   Morales
   Morillón
   Jorge Alan
   Universidad Nacional Autónoma de México
   Facultad de Ciencias
   Licenciatura en Matemáticas
   309240459

2. **Datos del tutor**
   Dra.
   Sandra
   Romero
   Hidalgo

3. **Datos del co-tutor**
   Dra.
   Eliane Regina
   Rodrigues

4. **Datos del sinodal 1**
   Dr.
   Eduardo Arturo
   Gutiérrez
   Peña

5. **Datos del sinodal 2**
   Dra.
   María Asuncion
   Begoña
   Fernández Fernández

6. **Datos del sinodal 3**
   Dra.
   Ana
   Meda
   Guardiola

7. **Datos del trabajo**
   Cadenas de Markov no-homogéneas
   para el mapeo genético de poblaciones mezcladas
   (*Admixture Mapping*)
   75 pp.
   2018

Agradecimientos

Agradezco a mi madre y dos hermanos, que son lo más importante en mi vida.

A mis amigos y profesores, que me han acompañado y enseñado a ser una mejor persona.

A la Dra. Sandra Romero Hidalgo, por abrirme las puertas al área de la genética, por los conocimientos que me ha compartido y por su paciencia.

A la Dra. Eliane Regina Rodrigues, por compartir sus conocimientos y experiencia conmigo, por el apoyo y comprensión que me ha brindado y por alentarme a seguir estudiando.

Al Dr. Eduardo, por la ayuda y atención que me ha brindado para realizar este proyecto de tesis y por compartir conmigo sus conocimientos.

A las sinodales, Dra. María Asunción Begoña Fernández Fernández y Dra. Ana Meda Guardiola, por revisar este proyecto de tesis y por sus comentarios y sugerencias.

Al Departamento de Matemáticas - Facultad de Ciencias - UNAM, por haberme autorizado una licencia por cláusula 69 (fracción IV) del Contrato Colectivo de Trabajo.

Al Instituto de Matemáticas - UNAM, por haberme brindado una beca de lugar.

Al Instituto Nacional de Medicina Genómica, por haberme brindado acceso a las instalaciones, un espacio de trabajo y por el acceso a datos para mi tesis.

Índice

Introducción

La mezcla genética o mestizaje ocurre cuando dos o más poblaciones, que se encontraban previamente aisladas, se mezclan entre sí dando lugar a una población híbrida. A las poblaciones previamente aisladas se les conoce como poblaciones ancestrales o parentales. Las poblaciones mezcladas poseen características de sus poblaciones ancestrales que pueden ser benéficas o perjudiciales para la salud, y por lo tanto resulta importante estudiarlas.

La medicina genómica es el área de las ciencias que tiene como objetivo estudiar el componente genético de las enfermedades o rasgos de una población. Una de las estrategias consiste en el mapeo genético; en particular, la estrategia específicamente diseñada para poblaciones mezcladas es la conocida como *Admixture Mapping* (AM, por sus siglas en inglés) o mapeo genético de poblaciones mezcladas. Este tipo de mapeo aprovecha la estructura genética de los individuos de una población mezclada para localizar regiones de riesgo; se utiliza principalmente para estudiar enfermedades o rasgos que presentan una mayor prevalencia en alguna de sus poblaciones ancestrales.

El objetivo del presente trabajo de tesis es abordar el AM mediante métodos estocásticos y Bayesianos. Para este fin, en el primer capítulo se introducen los conceptos básicos de genética necesarios para comprender el objetivo de la presente tesis. Se incluye una revisión sobre las leyes de Mendel, seguida de los principios básicos del mapeo genético con

un enfoque para la medicina genómica y, en particular, para el estudio de poblaciones con mezcla genética.

Por su parte, en el segundo capítulo se desarrollan los conceptos matemáticos básicos necesarios divididos en dos secciones: en la primera se desarrollan los conceptos de proceso estocástico y proceso de Markov para describir el caso particular de las cadenas de Markov no-homogéneas y en la segunda se describen conceptos como la función de verosimilitud, distribuciones *a priori* y *a posteriori*.

El tercer capítulo está dedicado a la formulación Bayesiana para el caso de un modelo Markoviano no-homogéneo. Posteriormente, en el cuarto capítulo se propone la aplicación de un modelo estadístico para realiza un análisis de AM utilizando datos reales de dos grupos de individuos mexicanos, un grupo de individuos afectados con neuromielitis óptica y otro grupo de individuos sanos.

Finalmente, presentamos las conclusiones de la tesis en el quinto capítulo.

Capítulo 1

Conceptos básicos de genética

1.1. Leyes de Mendel

El estudio de la genética clásica es fundamental para el entendimiento de la genética moderna. Una serie de experimentos realizados por Gregor Johann Mendel (1822-1884) conforman la base de la genética clásica. En sus experimentos Mendel utilizó plantas de chícharos para estudiar varios rasgos heredables como el color y la forma de sus semillas. Mendel explicó la existencia de ciertas "unidades de información" que ahora se conocen como genes[1]. Los genes pueden tener variantes o alelos;[2] cada organismo recibe dos alelos, uno de su padre y otro de su madre. La pareja de alelos de cada individuo conforman su genotipo[3]. El fenotipo[4] constituye la característica observable que un individuo presenta como resultado de su genotipo. Los resultados que obtuvo en sus experimentos los plasmó en 1865 en los siguientes postulados ahora conocidos como Leyes de Mendel.

- *Ley de uniformidad.* Si se considera un gen $\boldsymbol{G_1}$ con dos alelos $\{G_1, g_1\}$. Dados dos

[1]Unidad física fundamental de la herencia.

[2]Formas alternativas de un gen.

[3]Constitución alélica o genética de un organismo.

[4]Manifestación variable del genotipo de un organismo en un determinado ambiente.

organismos O_1 y O_2 homocigotos[5] para el gen $\boldsymbol{G_1}$, tales que poseen los genotipos G_1G_1 y g_1g_1, respectivamente, entonces cualquier descendiente de esta pareja de organismos tendrá el genotipo G_1g_1 para el gen $\boldsymbol{G_1}$.

- *Ley de segregación.* Si se considera un gen $\boldsymbol{G_1}$ con dos alelos $\{G_1, g_1\}$. Dado un organismo O (heterocigoto[6] para el gen $\boldsymbol{G_1}$) con el genotipo G_1g_1, el alelo G_1 segrega en un gameto[9] y el alelo g_1 segrega en otro gameto distinto.

- *Ley de independencia de segregación.* Aquí se consideran dos genes distintos con dos alelos cada uno, $\{\boldsymbol{G_1}\}$ con alelos $\{G_1, g_1\}$ y $\{\boldsymbol{G_2}\}$ con alelos $\{G_2, g_2\}$. Dados dos organismos O_1 y O_2 homocigotos para el par de genes $\{\boldsymbol{G_1}\}$ y $\{\boldsymbol{G_2}\}$, donde el organismo O_1 posee el genotipo G_1G_1 y G_2G_2, mientras que el organismo O_2 posee el genotipo g_1g_1 y g_2g_2, respectivamente, entonces el genotipo de un organismo O que es descendiente de los organismos O_1 y O_2 queda determinado mediante el genotipo G_1g_1 y G_2g_2. A su vez, el organismo O podrá segregar en gametos (haploides) los siguientes alelos: G_1G_2, G_1g_2, g_1G_2 o g_1g_2.

Los rasgos que estudió Mendel segregan de forma independiente pero posteriormente se identificó que hay genes que no segregan de esta misma manera, es decir que están ligados. Para poder explicar el fenómeno de ligamiento es necesario introducir el concepto de entrecruzamiento. El entrecruzamiento es un proceso que ocurre durante la meiosis[11] y consiste en el intercambio de material genético entre pares de cromosomas homólogos. Una de las principales consecuencias es la introducción de variabilidad en el material genético.

[5]Organismo (célula) con alelos idénticos para un gen o genes de interés.
[6]Organismo (célula) con distintos alelos en uno o varios loci[7].
[7]Plural de locus[8].
[8]Posición fija en un cromosoma.
[9]Célula haploide[10]reproductiva.
[10]Célula que posee un único juego de cromosomas.
[11]Proceso de división celular que tiene como objetivo la producción de células sexuales.

En un cromosoma, tanto el número de entrecruzamientos como el lugar donde ocurren forman parte de un proceso aleatorio. Sin embargo, el número de entrecruzamientos entre dos loci puede servir como medida estocástica de distancia genética en los cromosomas. Si dos genes se encuentran "cerca", hay una probabilidad baja de que ocurra al menos un entrecruzamiento entre ellos; en cambio, si se encuentran "lejos" la probabilidad de que ocurra al menos un entrecruzamiento es alta. Cuando los genes segregan juntos ocurre el ligamiento genético. En conjunto los conceptos previamente descritos conforman la base para la construcción de mapas genéticos [1].

Los mapas genéticos tienen como objetivo identificar la posición de los genes en los cromosomas. Existen distintos tipos de mapas, como los mapas genéticos y los mapas físicos. Los mapas genéticos están conformados por conjuntos ordenados de loci con distancias genéticas estimadas entre loci adyacentes. Este tipo de mapas fueron los primeros en ser desarrollados para ubicar la posición de los genes a lo largo de los cromosomas. El número esperado de entrecruzamientos entre dos loci por meiosis se utiliza como medida de distancia que se representa en centiMorgans(cM), en honor a Thomas Hunt Morgan (1866-1954), genetista estadounidense, quien lo propuso. A diferencia de los mapas genéticos, la distancia de los mapas físicos está representada en unidades de pares de bases (bp)[12] de ADN; las distancias se pueden expresar en kilobases (kb)[13] o megabases (Mb)[14], lo que refleja el hecho de que los cromosomas son cadenas largas de ADN.

Se debe notar que las distancias genética y física no se correlacionan de forma precisa, debido a que la distancia genética no es proporcional a la distancia física ya que el valor esperado de entrecruzamientos entre dos loci varía a lo largo de los cromosomas, lo que significa que la misma distancia genética en distintas regiones de los cromosomas repre-

[12]Unidad de medida que consta de dos nucleobases unidas entre sí por enlaces de hidrógeno.

[13]Equivaletne a mil pares de pares bases.

[14]Equivalente a un millón de pares de bases.

sentan distintas distancias físicas [1]. Sin embargo, sucede que a mayor distancia genética mayor distancia física.

1.2. Mapeo Genético

Fracción de recombinación

Considerando dos loci distintos con dos alelos cada uno, $\boldsymbol{G_1}$ con alelos $\{G_1, g_1\}$ y $\boldsymbol{G_2}$ con alelos $\{G_2, g_2\}$, si una célula diploide[15] contiene el genotipo G_1G_2 y g_1g_2 en una pareja de cromosomas homólogos, entonces en el proceso de la meiosis se pueden obtener los siguientes productos : G_1G_2, G_1g_2, g_1G_2 y g_1g_2. A los genotipos G_1G_2 y g_1g_2 se les conoce como no-recombinantes ya que preservan la misma configuración de los cromosomas homólogos, a los genotipos G_1g_2 y g_1G_2 se les conoce como recombinantes. Para observar una recombinación entre dos marcadores es necesario que ocurra una número impar de entrecruzamientos entre ellos [2].

Definición 1.1. Dados un par de loci $\boldsymbol{G_1}$ y $\boldsymbol{G_2}$ con alelos $\{G_1, g_1\}$ y $\{G_2, g_2\}$, respectivamente, se define la fracción de recombinación $\boldsymbol{\theta}$:

$$\boldsymbol{\theta_{G_1G_2}} = \frac{\#recombinantes}{total \quad de \quad cromosomas} \times 100\,\%.$$

El genetista Alfred Henry Sturtevant (1891-1970) desarrolló una técnica donde por primera vez se utilizó la variación en la magnitud de ligamiento entre genes para dotar de un orden lineal a sucesiones de genes en los cromosomas. Así, Sturtevant construyó el primer mapa genético de un cromosoma en 1913. Éste es un ejemplo de que históricamente la fracción de recombinación se ha utilizado para la construcción de mapas genéticos (mapas de ligamiento). Estos mapas contribuyeron a establecer un orden lineal y lugares específicos de los genes en los cromosomas.

[15]Célula que posee un doble juego de cromosomas.

La fracción de recombinación tiene como rango el intervalo $[0, 0.5]$. Valores cercanos a 0 representan que dos loci están tan cerca uno del otro, que segregan juntos de generación en generación. Mientras que valores cercanos a 0.5 representan que los loci están lejos uno del otro en el mismo cromosoma o que se encuentran en cromosomas distintos, y por lo tanto segregan de forma independiente. Se debe notar que la fracción de recombinación no es propiamente una pseudométrica: cumple con la propiedad $\boldsymbol{\theta_{G_1G_1}} = 0$; cumple con la propiedad de simetría $\boldsymbol{\theta_{G_1G_2}} = \boldsymbol{\theta_{G_2G_1}}$; pero no cumple con la desigualdad del triángulo, $\boldsymbol{\theta_{G_1G_3}} \leq \boldsymbol{\theta_{G_1G_2}} + \boldsymbol{\theta_{G_2G_3}}$. Esto se debe a que la fracción de recombinación no es sensible al suceso de un número par de recombinaciones entre loci considerablemente separados, produciendo así un efecto en el que se subestima el valor real de la fracción de recombinación. Se ha observado que para fracciones de recombinación con un valor menor a 10 % generalmente se preserva la desigualdad del triángulo.

Como consecuencia, se crearon las funciones de mapa que tienen como principal objetivo dotar de una pseudométrica a los cromosomas, la cual sirve para calcular distancias entre loci con base en los valores de recombinación.

Funciones de mapa

El genetista John Burdon Sanderson Haldane (1892-1964), considerado como uno de los padres de la genética de poblaciones, desarrolló en 1919 la primer función de mapa que considera la posibilidad de múltiples entrecruzamientos a lo largo de los cromosomas. Dada una región cromosómica, la probabilidad de recombinación es pequeña por lo que la probabilidad de múltiples entrecruzamientos en la misma región se asume que tiene un comportamiento aleatorio que se ajusta a una distribución Poisson [3]. Así, la función de mapa de Haldane $\boldsymbol{x_h}(\cdot)$ (*Haldane's map function*) [1] relaciona la fracción de

recombinación y la distancia genética de la siguiente forma:

$$\boldsymbol{x_h} = \begin{cases} -\frac{1}{2}\ln(1-2\theta_h), & \text{si} \quad 0 \leq \theta_h < \frac{1}{2} \\ \infty, & \text{en otro caso} \end{cases}$$

con inversa

$$\boldsymbol{\theta_h} = -\frac{1}{2}\big(1 - \exp\{-2|x_h|\}\big).$$

Posteriormente, el matemático Damodar Dharmananda Kosambi (1907-1966) publicó otra función de mapa en 1944 que considera el fenómeno de interferencia, el cual describe que la ocurrencia de un entrecruzamiento hace que sea menos probable la ocurrencia de otro en el mismo gameto. Su modelo describe que la interferencia en los entrecruzamientos depende de la distancia del segmento del cromosoma, disminuye cuando el segmento del cromosoma es suficientemente grande y aumenta cuando el segmento del cromosoma es pequeño. La función de mapa de Kosambi $\boldsymbol{x_k}(\cdot)$ (*Kosambi's map function*) [1] relaciona la fracción de recombinación y la distancia genética de la siguiente forma:

$$\boldsymbol{x_k} = \frac{1}{2}\tanh^{-1}(2\theta_k) = \frac{1}{4}\ln\frac{1+2\theta_k}{1-2\theta_k},$$

con inversa

$$\boldsymbol{\theta_k} = \frac{1}{2}\tanh(2x_k) = \frac{1}{2}\frac{\exp\{4x_k\}-1}{\exp\{4x_k\}+1}.$$

Ambas funciones brindan aditividad en las distancias a lo largo de los cromosomas. Así, las funciones de mapeo generan una pseudométrica a lo largo de los cromosomas, cuya unidad de medida es el centimorgan (cM). Formalmente, se dice que dos marcadores genéticos se encuentran separados por 1 cM si en promedio ocurre un entrecruzamiento por cada 100 meiosis entre dichos marcadores.

1.3. Mapeo genético para la medicina genómica

La medicina genómica es una área de las ciencias que tiene como objetivo estudiar el componente genético de las enfermedades o rasgos de una población, con el fin de mejorar los cuidados de la salud a través de una práctica más personalizada, predictiva y participativa. El mapeo genético es una estrategia utilizada para estudiar el componente genético de las enfermedades. Asimismo el mapeo genético ayuda a detectar la ubicación de genes responsables o relacionados con el desarrollo de enfermedades o rasgos. De forma general, las enfermedades genéticas se pueden clasificar en las siguientes categorías:

- Enfermedades monogénicas o mendelianas.
- Enfermedades poligénicas o multifactoriales.

Definición 1.2. Las enfermedades monogénicas son aquellas que son causadas por una mutación[16] en un solo gen con penetrancia alta.

La penetrancia corresponde a la probabilidad de manifestar el fenotipo dado el genotipo. En las enfermedades monogénicas se considera una penetrancia de 100 % (completa), o superior a 80 % (incompleta). Una de sus características principales es que la contribución genética es determinante, y la contribución ambiental es poca o nula, para el desarrollo de estas enfermedades o rasgos. Algunos ejemplos de enfermedades monogénicas son las siguientes: fibrosis quística, enfermedad de Huntington, distrofia muscular de Duchenne, hemofilia, entre otras. La proporción de personas que sufren de estas enfermedades con respecto a la totalidad de la población generalmente es baja y por esta razón también se les conoce como enfermedades raras. Estas enfermedades fueron las primeras en ser estudiadas con los conceptos y mapeos descritos a lo largo de este capítulo.

Uno de los principales métodos que se utilizaron para identificar la posición de los genes o marcadores genéticos involucrados en las enfermedades monogénicas es el método de LOD

[16]Variación en la secuencia de ADN. Presente en menos del 1 por ciento de la población.

Score que consiste en estimar la fracción de recombinación asumiendo que la penetrancia y la frecuencia de la mutación es conocida utilizando familias de múltiples generaciones con varios individuos afectados y sanos [1]. Otros métodos que han tenido éxito en la localización de alelos de riesgo para el desarrollo de enfermedades monogénicas son los estudios de asociación basados en familias. Estos métodos pueden ser empleados utilizando varios diseños como los que están basados en tríos, que consisten en la selección de un caso afectado y compararlo con sus dos progenitores, los que están basados en parejas de hermanos concordantes o discordantes con respecto al fenotipo de estudio o los que están basados en familias [4]. Estos métodos se fueron adaptando para estudiar enfermedades multifactoriales.

Definición 1.3. Las enfermedades multifactoriales son aquellas donde participan polimorfismos[17] en múltiples genes interactuando entre ellos y con el medio ambiente.

En las enfermedades multifactoriales es necesario tomar en cuenta que se involucra la interacción de varios marcadores, en donde cada marcador incrementa ligeramente el riesgo de desarrollar la enfermedad, pero la aportación de cada marcador no es necesariamente la misma. Además, el factor ambiental repercute en el posible desarrollo de la enfermedad. Es por esto que sólo se puede hablar de riesgo en las enfermedades multifactoriales y no de penetrancia. Este riesgo ha de tomar en cuenta la presencia de marcadores de susceptibilidad y factores ambientales necesarios para el desarrollo de la enfermedad. Algunos ejemplos de enfermedades multifactoriales son las siguientes: enfermedad de Alzheimer, hipertensión arterial, obesidad, diabetes mellitus, entre otras. La prevalencia de estas enfermedades en la población es muy alta. La cantidad de enfermedades multifactoriales es mucho mayor a la de las enfermedades monogénicas. El estudio de estas enfermedades ha aumentado en las últimas décadas debido a que se consideran problemas de salud pública y también al gasto que generan en los sistemas de salud. El proyecto del genoma humano, y los avances tecnológicos que se han dado como consecuencia del mismo, han permitido

[17]Variación en la secuencia de ADN en un locus. Presente en al menos 1 por ciento de la población.

el desarrollo de plataformas genómicas que han acelerado la identificación de marcadores genéticos involucrados en las enfermedades multifactoriales. Asimismo, el desarrollo tecnológico ha estado acompañado por el desarrollo y progreso de estrategias y algoritmos computaciones, así como métodos matemáticos y estadísticos, principalmente diseñados para manejar de forma eficiente datos masivos.

Una de las principales estrategias para el estudio de las enfermedades multifactoriales es la conocida como estudio de asociación con individuos no-relacionados. Este tipo de estudio tienen como objetivo identificar asociación entre un polimorfismo y una enfermedad. La asociación entre un polimorfismo y una enfermedad puede ser de tres tipos:

- *Asociación directa*. Cuando se identifica asociación con un polimorfismo que es el factor causal.
- *Asociación indirecta*. Cuando el polimorfismo de estudio no es un factor causal pero está en desequilibrio de ligamiento[18] con el factor causal, es decir que se encuentran muy cerca uno del otro.
- *Estratificación poblacional*. Cuando se identifica asociación con un polimorfismo que tiene diferencias de frecuencias importantes en las poblaciones ancestrales o parentales.

Un caso específico de un análisis de asociación consta de estudiar individuos afectados con un fenotipo de interés (casos) y compararlos con individuos sin el fenotipo de interés (controles). Hay distintos métodos para el análisis de este tipo de datos, por ejemplo: pruebas de asociación o modelos lineales generalizados.

Las pruebas de asociación presentan una limitación para ajustar por factores de confusión como la estratificación poblacional o mezcla genética. La estratificación poblacional se

[18] Asociación no aleatoria de alelos en diferentes loci en una población determinada.

presenta cuando existen diferencias en las frecuencias alélicas entre los grupos de estudio atribuidas a diferencias en la composición de componentes ancestrales. Como consecuencia de esta limitación surge la estrategia de *Admicture Mapping* (*AM*, por sus siglas en inglés). Esta estrategia aprovecha la mezcla genética que se obtiene como producto del mestizaje entre dos o más poblaciones y convierte la composición genética en objeto de estudio.

1.4. Mapeo genético basado en la mezcla de poblaciones

La mezcla genética o mestizaje ocurre cuando dos o más poblaciones, que se encontraban previamente aisladas, se mezclan entre sí dando lugar a una población híbrida. A las poblaciones previamente aisladas se les conoce como poblaciones ancestrales o parentales.

La mezcla genética entre poblaciones que presentan frecuencias alélicas distintas genera gametos que consisten en un mosaico de segmentos heredados de cada población ancestral. Por ejemplo, en el caso de los gametos que provienen de un individuo con ascendencia mixta, se muestra una autocorrelación de los segmentos ancestrales, ya que dados dos loci se tiene que a menor distancia de mapa entre sí, mayor es la probabilidad de que ambos loci pertenezcan a la misma población ancestral. En las regiones donde la frecuencia alélica varía entre poblaciones, esta autocorrelación de ancestría genera una asociación alélica que decae con la distancia. Por otra parte, la estratificación poblacional genera asociaciones alélicas que son independientes de la distancia de mapa. Tanto la mezcla genética como la estratificación presentan oportunidades y retos para la epidemiología genética. Hay oportunidades de explotar la estructura genética de poblaciones mezcladas como estrategia de mapeo ya que puede facilitar la identificación de loci de susceptibilidad de enfermedades. El reto consiste en controlar por la estratificación poblacional generada por la mezcla genética durante el análisis de asociación [2].

El mapeo genético basado en la mezcla de poblaciones es un método para el estudio de enfermedades o rasgos genéticos que muestran una prevalencia diferente en las poblaciones ancestrales de una población mezclada. Por sus características, este tipo de mapeo se ha empleado en poblaciones afroamericanas y latinoamericanas ya que presentan una estructura de mezcla genética reciente; esto se debe a que los segmentos de información genética heredados de las poblaciones ancestrales son más grandes, lo que ayuda a distinguir regiones de susceptibilidad con una mayor precisión. En los últimos años el desarrollo tecnológico y científico en el área de la genética ha impulsado el progreso en los métodos utilizados para la realización de inferencias más robustas sobre la composición genética de un individuo y por tanto de la población en general [5].

El AM es una estrategia de mapeo genético, al igual que los análisis de asociación y los análisis de ligamiento. A continuación se presentan algunas ventajas que pueden presentar los análisis de AM:

- La prevalencia de algunos de los fenotipos de riesgo varían entre poblaciones, lo que permite una mayor precisión en la localización de los posibles factores de riesgo. Por ejemplo, las enfermedades autoinmunes como la esclerosis múltiple tienden a ser más frecuentes en individuos con ancestría europea, mientras que enfermedades como la hipertensión tienden a ser más frecuentes en individuos con ancestría africana [5].

- El desempeño estadístico para la detección de loci de susceptibilidad[19] es mejor para AM en comparación con estudios de ligamiento (de familias) [2].

- En los análisis para detectar alelos causantes de enfermedades con diferencias importantes en las frecuencias alélicas entre poblaciones ancestrales, el AM puede detectar genes de efecto moderado con un desempeño comparable al del mapeo de

[19]Loci que genera variación en el riesgo de la enfermedad entre poblaciones.

haplotipos de genoma completo [6].

- Para los análisis a nivel de genoma completo la densidad de marcadores genéticos requerida es mucho más baja para el AM en comparación a los análisis de asociación [2]. Este tipo de análisis se relacionan con pruebas de comparaciones múltiples [1].

- Los estudios de AM pueden detectar locus siempre y cuando el total de alelos de riesgo involucrados en el estudio estén diferencialmente distribuidos entre las poblaciones ancestrales que contribuyeron a la mezcla de la población de estudio. En contraste, en los análisis de asociación es poco probable encontrar alelos de riesgo [2].

Debemos notar que, al igual que otros tipos de mapeo, el AM es susceptible a factores de confusión como patrones desconocidos de herencia y la interacción entre factores de riesgo genéticos y ambientales. Además se debe reconocer que el AM no asume que el riesgo de las enfermedades multifactoriales está distribuido equitativamente a lo largo de los loci involucrados y que generalmente no se sabe qué tanto del riesgo diferencial por ancestría se debe a factores genéticos o a factores ambientales, por lo que es de suma importancia estudiar estas relaciones.

1.4.1. Estimación de ancestría

Por sus características, el mapeo genético basado en la mezcla de poblaciones AM ha sido considerado como un enfoque eficiente para localizar variantes de riesgo para enfermedades o rasgos que difieren en frecuencia entre las poblaciones ancestrales [6] y correlacionar la composición ancestral de un loci con un fenotipo [5].

Para realizar un análisis de AM es necesario estimar la composición ancestral de los individuos bajo estudio. A la composición ancestral también se le conoce como ancestría. La estimación de ancestría no es una tarea trivial, dado que los grupos continentales han

perdido entre 10-15 % de su diversidad alélica ancestral compartida, lo que sugiere que hay pocos marcadores genéticos que se han fijado de forma diferencial entre poblaciones continentales. Un ejemplo de este tipo de marcadores es el locus FY (Antígeno de Duffy) que tiene una prevalencia muy alta en poblaciones subsaharianas y casi nula en otras poblaciones [2,5]. Por lo tanto no es posible estimar de forma directa la ancestría de locus de un genotipo de interés. Debido a que en los últimos años a aumentado la densidad de marcadores genéticos, se ha impulsado un amplio desarrollo de métodos estadísticos para inferir la ancestría de forma global (a lo largo del genoma) o de forma local (en lugares particulares de los cromosomas).

La ancestría global corresponde a la proporción relativa de bloques de las poblaciones ancestrales a lo largo del genoma del individuo de estudio. Se han desarrollado diversos métodos para hacer estas estimaciones, los cuales han sido implementadas en programas como:

- *ADMIXMAP*, es un programa para modelar mezcla genética que usa datos de marcadores de genotipos y del rasgo de estudio sobre una muestra de una población mestiza, donde los marcadores han sido seleccionados para tener una diferencias extremas en las frecuencias alélicas entre las poblaciones ancestrales involucradas en el mestizaje. Para modelar la ancestría se utilizan modelos multinivel (modelos jerárquicos lineales) [7].

- *STRUCTURE*, es un programa que utiliza datos de múltiples locus de los genotipos de interés para el estudio de la estructura de la población. Utiliza métodos de Monte Carlo vía cadenas de Markov (*Markov chain Monte Carlo*, MCMC por sus siglas en inglés) [8].

- *ADMIXTURE*, es un programa para maximizar la verosimilitud de la estimación de ancestría de individuos a partir de conjuntos de datos multilocus de SNP de genoti-

pos. Implementa algoritmos de esperanza-maximización (*Expectation-Maximization*, EM por sus siglas en inglés) y de MCMC [9].

Por su parte, la ancestría local corresponde a la estimación de ancestría en un lugar particular de un cromosoma del individuo en cuestión. De igual forma, se han desarrollado diversos métodos para hacer estas estimaciones, los cuales han sido implementados en programas como:

- *Hapmix*, es un programa para inferir de forma precisa segmentos de distintas ancestrías en poblaciones mestizas. Emplea técnicas estandar de cadenas de Markov ocultas (*hidden Markov models*, HMM por sus siglas en inglés) con recursión hacia adelante y hacia atrás (*forward-backward recursion*) [10].

- *LAMP-LD*, es un programa para la inferencia de ancestría local en poblaciones mestizas. Implementa un análisis de componentes principales (*principal components analysis*, PCA por sus siglas en inglés) después de haber procesado la información mediante HMM [11].

- *PCAdmix*, es un programa diseñado para inferir ancestría local a lo largo del genoma de un individuo mestizo. Aplica un análisis de PCA para asignar mayor peso a las variantes que son más informativas sobre la ancestría, después se implementa un método de HMM para modelar la probabilidad de ancestría, para cada población ancestral, en segmentos de los cromosomas [12].

Una vez que se tiene la ancestría local de los individuos de estudio para realizar un análisis de *AM*, el siguiente paso es proponer métodos estadísticos para hacer inferencias sobre los factores de riesgo genéticos involucrados en las enfermedades. Es importante notar que hay pocas estrategias reportadas para realizar análisis de *AM*, como la propuesta por Nick paterson et. al [6].

Capítulo 2

Conceptos matemáticos básicos

2.1. Procesos Estocásticos

Esta sección se dedica a la presentación de los conceptos básicos de los procesos estocásticos y, en particular, las cadenas de Markov no-homogéneas. La teoría de los procesos estocásticos tiene como objetivo el estudio de las estructuras que conforman colecciones de variables aleatorias indexadas. Estas estructuras usualmente modelan los resultados de un experimento aleatorio en el transcurso del tiempo.

Definición 2.1. Un **proceso estocástico** es una colección de variables aleatorias $\boldsymbol{X} = \{X_t : t \in \mathcal{T}\}$ con $\mathcal{T}$ su conjunto de índices, asumiendo valores en un conjunto de estados $\mathcal{S}$, llamado espacio de estados [13–15].

De manera general, los procesos estocásticos se pueden clasificar de acuerdo a las características de su conjunto de índices y de su espacio de estados, así como las relaciones de dependencia entre las variables aleatorias que conforman dichos procesos. A continuación se presentan algunas definiciones básicas.

Definición 2.2.

(a) Cuando el conjunto de índices $\mathcal{T}$ es finito o numerable, se dice que el proceso $\boldsymbol{X}$ es de tiempo discreto.

(b) Cuando el conjunto de índices $\mathcal{T}$ es un intervalo de la recta real, se dice que el proceso $\boldsymbol{X}$ es de tiempo continuo.

(c) Cuando el espacio de estados $\mathcal{S}$ es de la forma $\mathcal{S} \subseteq \mathbb{Z}$, se dice que el proceso $\boldsymbol{X}$ es de estado discreto.

(d) Cuando el espacio de estados $\mathcal{S}$ es de la forma $\mathcal{S} \subseteq \mathbb{R}$ y $\mathcal{S}$ es un intervalo, se dice que el proceso $\boldsymbol{X}$ es de estado continuo.

Observación 2.1. Es posible generalizar el conjunto de índices y el espacio de estados [13–15].

Algunos ejemplos clásicos de procesos estocásticos caracterizados por relaciones de dependencia entre sus variables aleatorias son los siguientes:

- Cadenas de Markov,
- Martingalas,
- Procesos de Poisson,
- Movimiento Browniano.

De los distintos tipos de procesos estocásticos nos interesa concentrarnos en los procesos de Markov.

2.1.1. Procesos de Markov

Un proceso de Markov es aquel que tiene la propiedad de que, dado el valor de X_t, el valor de $X_s, t < s$, no depende de los valores de $X_u, u < t$; es decir, la probabilidad de cualquier comportamiento futuro particular del proceso, cuando su estado presente se conoce con exactitud, no es alterada por conocimiento adicional correspondiente a su comportamiento pasado [13].

Para el caso en el que el proceso $\boldsymbol{X}$ es continuo con espacio de estados de tiempo continuo, el proceso de Markov queda determinado de la siguiente manera.

Definición 2.3. Un proceso estocástico $\boldsymbol{X}$ es un proceso de Markov si, para $t_0, t_1, ..., t_n, t \in \mathcal{T}$, tales que, $t_0 < t_1 < \cdots < t_n < t$ y para $a, b \in \mathcal{S}$, tales que, $a < b$ se tiene que

$$P(a < X_t \leq b | X_{t_0} = x_0, X_{t_1} = x_1, ..., X_{t_n} = x_n) = P(a < X_t \leq b | X_{t_n} = x_n).$$

Podemos introducir el concepto de probabilidad de transición para un proceso de Markov como sigue.

Definición 2.4. Sean $\boldsymbol{X}$ un proceso de Markov, $\mathcal{S} \subseteq \mathbb{R}$ su espacio de estados y $A \subseteq \mathcal{S}$ con A un intervalo, entonces la función de probabilidad de transición está dada por:

$$P(x, s; t, A) = P(X_t \in A | X_s = x) \quad s, t \in \mathcal{T}, s < t, x \in \mathcal{S}.$$

Un caso particular de los procesos de Markov son las cadenas de Markov.

2.1.2. Cadenas de Markov

Definición 2.5. Una cadena de Markov $\boldsymbol{X}$ es un proceso de Markov cuyo espacio de estados $\mathcal{S}$ es discreto (finito o infinito numerable).

Observación 2.2. En el caso de cadenas de Markov, la función de transición es llamada probabilidad de transición.

Definición 2.6. La probabilidades de transición de una cadena de Markov discreta $\boldsymbol{X}$ con espacio de estados $\mathcal{S}$ se describe de la siguiente forma:

$$\text{Para} \quad t \in \mathcal{T}, \quad x_i \in \mathcal{S} \quad \forall i \in \{0, 1, ..., t+1\},$$
$$P(X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, ..., X_t = x_t) = P(X_{t+1} = x_{t+1} | X_t = x_t).$$

Decimos que al tiempo t la cadena se encuentra en el i-ésimo estado cuando $X_t = i$.

Definición 2.7. La probabilidad de transición de un paso para cadenas de Markov discretas es la probabilidad de que la cadena se encuentre en el estado j al tiempo $t+1$ dado que se encuentra en el estado i al tiempo t y lo denotamos de la siguiente manera.

$$P_{i,j}^{(t,t+1)} = P(X_{t+1} = j | X_t = i), \quad i, j \in \mathcal{S}, t \geq 0.$$

Definición 2.8. Una cadena de Markov $\boldsymbol{X}$ es homogénea en el tiempo si las probabilidades de transición no dependen del tiempo en el que se encuentra la cadena, es decir $P(X_{t+1} = j | X_t = i) = P(X_{s+1} = j | X_s = i), i, j \in \mathcal{S}, t \neq s, t, s \in \mathcal{T}$. En este caso indicamos $P_{i,j}^{(t,t+1)} = P_{i,j}$.

Definición 2.9. Una cadena de Markov $\boldsymbol{X}$ es no-homogénea en el tiempo si las probabilidades de transición dependen del tiempo en el que se encuentra la cadena, es decir, el tiempo t tiene influencia en el valor de la probabilidad de transición.

Para una cadena de Markov discreta no-homogénea, en un tiempo fijo t, las probabilidades de transición conforman una matriz que condensa dicha información. Esta matriz es conocida como matriz de probabilidad de transición o matriz de transición. La definición formal es la siguiente.

Definición 2.10. La matriz de transición al tiempo $t \in \mathcal{T}$ de una cadena de Markov discreta no-homogénea, indicada por $P^{(t,t+1)}$, está dada por:

$$P^{(t,t+1)} = \left[P_{i,j}^{(t,t+1)} \right]_{i,j \in \mathcal{S}}.$$

Observación 2.3. La distribución de probabilidad de la variable aleatoria X_{t+1} dado que $X_t = i$ es representada en el i-ésimo renglón de la matriz de transición de la cadena de Markov.

Las cadenas de Markov son completamente determinadas por su estado inicial y las probabilidades de transición. Esto permite pensar en una distribución de probabilidad inicial en el espacio de estados. Esta información inicial esta determinada por un vector denominado vector de probabilidades iniciales, que definiremos a continuación.

Definición 2.11. El vector de probabilidades iniciales de una cadena de Markov discreta con espacio de estados $\mathcal{S}$ es $P^{(0)} = \left[P_i^{(0)}\right]_{i\in\mathcal{S}}$, donde $\quad P_i^{(0)} = P(X_0 = i), i \in \mathcal{S}$.

Proposición 2.1. *Sea $\boldsymbol{X}$ una cadena de Markov discreta no-homogénea con espacio de estados $\mathcal{S}$ y conjunto de índices $\mathcal{T}$. Las siguientes afirmaciones valen para su matriz de transición, $i, j \in \mathcal{S}, t \in \mathcal{T}$.*

1. $P_{i,j}^{(t,t+1)} \geq 0$,
2. $\sum_{j\in S} P_{i,j}^{(t,t+1)} = 1$.

Demostración. Sea $\boldsymbol{X}$ una cadena de Markov discreta no homogénea. Tomemos $i, j \in \mathcal{S}$ y $t \in \mathcal{T}$.

1. La afirmación 1 es verdadera debido a que $P_{i,j}^{(t,t+1)}$ es una probabilidad y por lo tanto $P_{i,j}^{(t,t+1)} \geq 0$.

2. Para la afirmación 2, consideremos la siguiente partición del espacio muestral Ω en eventos disjuntos,

$$\Omega = \bigcup_{j\in S}(X_{t+1} = j).$$

Así tenemos que,

$$1 = P(\Omega|X_t = i) = P\Big(\bigcup_{j\in S}(X_{t+1} = j)|X_t = i\Big) = \sum_{j\in S} P(X_{t+1} = j|X_t = i)$$
$$= \sum_{j\in S} P_{i,j}^{(t,t+1)}.$$

□

Proposición 2.2. *Sea* $\boldsymbol{X}$ *una cadena de Markov discreta no homogénea, entonces se cumple la siguiente igualdad.*

$$P(X_0 = x_0, ..., X_t = x_t) = \left[\prod_{r\leq t} P^{(r-1,r)}_{x_{r-1},x_r}\right] P^{(0)}_{x_0}.$$

Demostración. Procederemos por inducción. El caso para cuando el tiempo $t = 0$ es trivial.

- Consideremos $t = 1$. Entonces,

$$P(X_0 = x_0, X_1 = x_1) = P(X_1 = x_1|X_0 = x_0)P(X_0 = x_0) = P^{(0,1)}_{x_0,x_1} P^{(0)}_{x_0}.$$

- Hipótesis inductiva. Supongamos que vale para t, es decir, para $x_i \in \mathcal{S}, i \in \{0, 1, ..., t\}$

$$P(X_0 = x_0, ..., X_t = x_t) = \left[\prod_{r+1\leq t} P^{(r,r+1)}_{x_r,x_{r+1}}\right] P^{(0)}_{x_0}.$$

- Consideremos ahora el caso para $t + 1$, es decir, veamos que el resultado vale para $P(X_0 = x_0, X_1 = x_1, ..., X_t = x_t, X_{t+1} = x_{t+1})$. Note que,

$$\begin{aligned}
P(X_0 = x_0, ..., X_{t+1} = x_{t+1}) &= P(X_{t+1} = x_{t+1}|X_0 = x_0, ...X_t = x_t) \\
&\quad P(X_0 = x_0, ..., X_t = x_t) \\
&= P(X_{t+1} = x_{t+1}|X_t = x_t) \left[\prod_{r+1\leq t} P^{(r,r+1)}_{x_r,x_{r+1}}\right] P^{(0)}_{x_0} \\
&= P^{(t,t+1)}_{x_t,x_{t+1}} \left[\prod_{r+1\leq t} P^{(r,r+1)}_{x_r,x_{r+1}}\right] P^{(0)}_{x_0} \\
&= \left[\prod_{r+1\leq t+1} P^{(r,r+1)}_{x_r,x_{r+1}}\right] P^{(0)}_{x_0}.
\end{aligned}$$

□

2.2. Estadística Bayesiana

Esta sección es dedicada a la presentación de algunos conceptos básicos y generales de la estadística Bayesiana.

Los modelos probabilísticos son utilizados para describir el comportamiento de los resultados obtenidos al realizar algún experimento aleatorio. A partir de este análisis se trata de describir el experimento y con esto estudiar sus propiedades. Los modelos utilizados pueden depender de ciertos parámetros desconocidos que necesitan ser estimados. Una forma de realizar esta tarea es utilizando la inferencia Bayesiana. En este caso, los valores de los parámetros se describen a través de variables aleatorias para las cuales se asigna una distribución.

El planteamiento de un ejemplo clásico para inferencia estadística es en el que se considera una muestra aleatoria $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$ con función de densidad conjunta $p(\boldsymbol{x}|\theta)$, donde las observaciones x_i son independientes e idénticamente distribuidas y θ es un vector de parámetros presentes en el modelo usado para describir el comportamiento de $\boldsymbol{x}$.

En la práctica, es posible que al realizar un análisis estadístico, los investigadores cuenten con información adicional sobre el vector de parámetros θ. Hacer uso de esta información adicional (descrita en términos de una distribución) es un aspecto que distingue a la estadística Bayesiana de la estadística clásica ya que en la estadística clásica no es posible incorporar esta información adicional. Podemos describir un modelo Bayesiano con los siguientes pasos:

- Determinar un modelo muestral, $p(x|\theta)$.
- Determinar una distribución inicial, $p(\theta)$.
- Calcular mediante propiedades de probabilidad condicional (el Teorema de Bayes)

una distribución final, $p(\theta|x)$.

- Resumir la información contenida en $p(\theta|x)$ para hacer inferencias sobre el vector de parámetros θ.

2.2.1. Función de verosimilitud

Al considerarse un modelo para describir un conjunto de observaciones, existe la posibilidad de que los datos observados provengan de un modelo con parámetros θ, es decir, el modelo con el parámetro θ lleva al valor observado $\boldsymbol{x}$.

Definición 2.12. Sea $p(\boldsymbol{x}|\theta)$ la función de probabilidad o de densidad conjunta de la muestra $\boldsymbol{X} = \{X_1, X_2, ..., X_n\}$ dado el modelo con parámetros θ. Dada $\boldsymbol{X} = \boldsymbol{x}$ la **función de verosimilitud** está definida por $\mathcal{L}(\boldsymbol{x}|\theta) \propto p(\boldsymbol{x}|\theta)$. Es decir, $\mathcal{L}(\boldsymbol{x}|\theta)$ es proporcional a la probabilidad o densidad de que los datos observados vienen del modelo descrito por $p(\boldsymbol{x}|\theta)$ con vector de parámetros θ.

2.2.2. Distribución *a priori*

La distribución inicial de un modelo Bayesiano es conocida como **distribución *a priori*** y describe la información que se tiene sobre el comportamiento del parámetro θ antes de que el valor de la muestra aleatoria sea observado. Determinar esta distribución nos permite incorporar, si lo deseamos, la información adicional previa sobre el parámetro de interés θ. En ocasiones no se cuenta con información adicional inicial o simplemente no se desea incorporar al análisis. En ambos casos, cuando se incorpora o no información inicial para el análisis es necesario que la distribución *a priori* describa el parámetro θ. Usualmente se denota de la siguiente forma:

- Distribución *a priori* informativa, es aquella en la que incorporamos información adicional previa sobre el parámetro θ.

- Distribución *a priori* no-informativa, es aquella en la que no incorporamos información adicional previa sobre el parámetro θ.

A continuación presentamos un ejemplo.

Ejemplo 2.1. Supongamos que contamos con un espacio de parámetros discreto para θ. Sea $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ el espacio parametral. La siguiente distribución sobre Θ no incorpora información adicional sobre el valor de θ.

$$p(\theta_i) = \frac{1}{n}, \quad i \in \{1, 2, ..., n\},$$

Porque la información de cada una de los elementos del espacio parametral Θ es la misma, por lo tanto es una distribución *a priori* no-informativa.

2.2.3. Distribución *a posteriori*

Para definir la distribución final, comúnmente conocida como **distribución *a posteriori***, hacemos uso de las propiedades de la probabilidad condicional. A continuación se presenta el teorema que permite obtener la distribución *a posteriori* de un vector de parámetros θ.

Teorema 2.1. *La distribución a posteriori del vector de parámetros θ presentes en el modelo que describe el fenómeno que produjo una muestra $\boldsymbol{x}$ está determinada por*

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})}, \tag{2.1}$$

donde

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)p(\theta)d\theta.$$

Demostración. Por definición de probabilidad condicional se tiene que,

$$p(\boldsymbol{x}, \theta) = p(\theta|\boldsymbol{x})p(\boldsymbol{x}).$$

También podemos escribir $p(\boldsymbol{x}, \theta) = p(\theta, \boldsymbol{x})$; a su vez $p(\theta, \boldsymbol{x}) = p(\boldsymbol{x}|\theta)p(\theta)$. Entonces,

$$p(\theta|\boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x}|\theta)p(\theta).$$

Por lo tanto tenemos que,

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})}.$$

□

Proposición 2.3. *La función de densidad a posteriori es proporcional al producto de la función de verosimilitud multiplicando a la densidad a priori, es decir,* $p(\theta|\boldsymbol{x}) \propto \mathcal{L}(\boldsymbol{x}|\theta)p(\theta)$.

Demostración. Por (2.1) tenemos que

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})},$$

donde $p(\boldsymbol{x})$ no depende de θ y por lo tanto es una constante respecto a θ. De esta forma podemos escribir

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\theta)p(\theta) \\ &\propto \mathcal{L}(\boldsymbol{x}|\theta)p(\theta). \end{aligned} \tag{2.2}$$

□

Para facilitar el cálculo de la distribución *a posteriori*, se buscan combinaciones de distribuciones *a priori* y de verosimilitudes que sean convenientes. Esta idea nos lleva a las llamadas familias conjugadas.

2.2.4. Familias conjugadas

Para simplificar el análisis en un modelo Bayesiano es común buscar distribuciones *a priori* y *a posteriori* que pertenezcan a la misma familia de distribuciones. Esto resulta en una gran ventaja para los procedimientos de inferencia, ya que el análisis queda restringido a

un subconjunto de todas las posibles distribuciones a tratar. Así podemos pasar de una distribución *a priori* a una *a posteriori* mediante un cambio en los hiperpárametros de las distribuciones. Esto permite que el análisis de las propiedades del vector de parámetros θ pueda ser realizado de manera más sencilla en comparación con métodos más complejos como los de Monte Carlo vía cadenas de Markov.

Definición 2.13. Una familia de distribuciones P es conjugada para un modelo muestral F si para toda distribución *a priori* en P y para modelos muestrales $f \in F$, la distribución *a posteriori* también pertenece a P.

Un ejemplo muy importante es el de las distribuciones conjugadas de la familia exponencial ya que en ella se encuentran varias de las distribuciones que se utilizan frecuentemente en la práctica.

Decimos que una densidad de probabilidad $f(x|\theta)$ pertenece a la familia exponencial de un parámetro si se puede expresar de la siguiente forma [16].

$$f(x|\theta) = a(x)\exp\{\phi(\theta)t(x) + b(\theta)\}.$$

En esta forma se encuentran, entre otras:

- Distribución normal con σ^2 conocida, $N(\theta, \sigma^2)$.
  $N(\theta, \sigma^2) = \exp\{\frac{x^2}{2\sigma^2}\}\exp\{\frac{\theta}{\sigma^2}x - \frac{\theta^2}{2\sigma^2}\}$
  donde,
  $a(x) = \exp\{\frac{x^2}{2\sigma^2}\}, \quad \phi(\theta) = \frac{\theta}{\sigma^2}, \quad t(x) = x, \quad b(\theta) = -\frac{\theta^2}{2\sigma^2}$
- Distribución exponencial, $\exp\{\theta\}$.
  $\exp(\theta) = \exp\{-\theta x + \log(\theta)\}$
  donde,
  $a(x) = 1, \quad \phi(\theta) = -\theta, \quad t(x) = x, \quad b(\theta) = \log(\theta)$

Una familia conjugada que será importante en este trabajo es la que involucra las distribuciones multinomial y la de Dirichlet. Trataremos esto a continuación.

Definición 2.14. Decimos que un vector aleatorio discreto $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ se distribuye de forma **multinomial** con parámetros N y $(p_1, p_2, ..., p_k)$, indicado por, $\boldsymbol{X} \overset{\mathfrak{D}}{=} \text{Mult}(p_1, p_2, ..., p_k, N)$ con $N \in \mathbb{Z}_+$ y $\sum_{i=1}^k x_i = N$, si para $(X_1, X_2, ..., X_k) = (x_1, x_2, ..., x_k)$ tal que $x_i \in \mathbb{Z}_+$ y $\sum_{i=1}^k p_i = 1$, su función de probabilidad es

$$p(x_1, x_2, ..., x_k | p_1, p_2, ..., p_k, N) = \frac{\Gamma((\sum_{i=1}^k x_i) + 1)}{\prod_{i=1}^k \Gamma(x_i + 1)} \prod_{i=1}^k p_i^{x_i}, \tag{2.3}$$

donde $\Gamma(\cdot)$ es la función gamma, dada por

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx,$$

que, en el caso cuando a es un entero positivo, es dada por

$$\Gamma(a) = (a-1)!$$

Definición 2.15. Decimos que una vector aleatorio $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ tiene distribución **Dirichlet** con parámetros $\alpha_1, \alpha_2, ..., \alpha_k$, indicado por, $\boldsymbol{X} \overset{\mathfrak{D}}{=} \text{Dir}(\alpha_1, \alpha_2, ..., \alpha_k)$, con $\alpha_i > 0, i \in \{0, 1, 2, ..., k\}$, $k \geq 2$, si el soporte de esta distribución es

$$\Delta_k = \Big\{ (x_1, x_2, ..., x_k) \in \mathbb{R}_+^k, \sum_{i=1}^k x_i = 1 \Big\},$$

y para $(X_1, X_2, ..., X_k) = (x_1, x_2, ..., x_k) \in \Delta_k$ la función de densidad de probabilidad es

$$p(x_1, x_2, ..., x_k | \alpha_1, \alpha_2, ..., \alpha_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}. \tag{2.4}$$

Veamos que la familia de distribuciones de Dirichlet es una familia conjugada de la distribución multinomial.

Proposición 2.4. *La familia de distribuciones de Dirichlet forma una familia conjugada para una función de verosimilitud de la forma multinomial; es decir, la distribución a posteriori también es de la familia Dirichlet.*

Demostración. Sean $\boldsymbol{p} = \{p_1, p_2, ..., p_k\} \sim \text{Dir}(\alpha_1, \alpha_2, ..., \alpha_k)$ y $\boldsymbol{y} = \{y_1, y_2, ..., y_k\} \sim \text{Mult}(p_1, p_2, ..., p_k; N)$, entonces, para $\theta = (p_1, p_2, ..., p_k)$ y $\boldsymbol{D} = \boldsymbol{y}$, tenemos

$$\begin{aligned} p(\theta|\boldsymbol{D}) &\propto \mathcal{L}(\boldsymbol{D}|\theta)p(\theta) \\ &\propto \Big(p(y_1, y_2, ..., y_k|p_1, p_2, ..., p_k)\Big)\Big(p(p_1, p_2, ..., p_k|\alpha_1, \alpha_2, ..., \alpha_k)\Big) \\ &\propto \prod_{i=1}^{k} p_i^{y_i} \prod_{i=1}^{k} p_i^{\alpha_i - 1} \\ &= \prod_{i}^{k} p_i^{\alpha_i + y_i - 1}. \end{aligned}$$

Así, si definimos como hiperparámetros a $\alpha_i' = \alpha_i + y_i$ tenemos que

$$p(\theta|\boldsymbol{D}) \propto Dir(\alpha_1', \alpha_2', ..., \alpha_k')$$

□

Observación 2.4. Podemos decir que un análisis Bayesiano nos brinda la información de nuestro interés condensada en una función de distribución de θ (distribución *a posteriori*), esto nos permite incorporar métodos probabilísticos en nuestros análisis.

Debemos notar que cada problema de inferencia tiene sus particularidades y dificultades propias; por ejemplo, es posible que aun cuando se haya encontrado una familia conjugada para las distribuciones *a priori* y *a posteriori*, el cálculo de la distribución *a posteriori* sea analíticamente complejo y para este (caso como para el caso en el que no sea posible encontrar una familia conjugada) puede ser necesario incorporar métodos numéricos o técnicas computacionales más sofisticados para poder hacer inferencias.

Capítulo 3

Inferencia Bayesiana para modelos Markovianos

En este capítulo se describirá la formulación Bayesiana para el caso de un modelo Markoviano no-homogéneo. Primero se presentarán algunos resultados relacionados con la función de verosimilitud del modelo y después se derivará la distribución *a posteriori* de los parámetros que están presentes en el modelo.

Supongamos que se tienen M realizaciones de una cadena de Markov $X = \{X_t : t \in \mathcal{T}\}$ no-homogénea con espacio de estados $\mathcal{S}$. Definimos las siguientes cantidades [17, 18],

- $n_i(t)$, el número de cadenas tales que al tiempo t se encuentran en el estado i, $i \in \mathcal{S}$, $t \in \mathcal{T}$.
- $n_{ij}(t)$, el número de cadenas tales que al tiempo t se encuentran en el estado i y que en el tiempo $t+1$ se encuentran en el estado j, con $i, j \in \mathcal{S}$, $t \in \mathcal{T}$.
- $n_{x_0,x_1,...,x_T,...}$, el número de cadenas que tienen la siguiente realización: $X = \{X_0 = x_0, X_1 = x_1, ..., X_T = x_T, ...\}, x_i \in \mathcal{S}, T \geq 0$.

Observación 3.1. Note que para $i, j \in \mathcal{S}, t \in \mathcal{T}$ tenemos [17],

$$n_{ij}(t) = \sum_{\substack{x_r \in \mathcal{S} \\ r \in \mathcal{T} \setminus \{t, t+1\}}} n_{x_0, x_1, \ldots, x_{t-1}, i, j, x_{t+2}, \ldots, x_T, \ldots}$$

Una forma de obtener las cantidades $n_i(t)$ y $n_{ij}(t), i, j \in \mathcal{S}, t \in \mathcal{T}$ es ordenando las M cadenas de Markov en forma de matriz. Así, sean $\boldsymbol{X}^{(1)} = \{X_0^{(1)}, X_1^{(1)}, ...\}, \boldsymbol{X}^{(2)} = \{X_0^{(2)}, X_1^{(2)}, ...\}, ..., \boldsymbol{X}^{(M)} = \{X_0^{(M)}, X_1^{(M)}, ...\}$ las M cadenas, $\boldsymbol{D} = \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, ..., \boldsymbol{X}^{(M)}\}$. Las cadenas en $\boldsymbol{D}$ se pueden arreglar de la siguiente forma, para realizar el conteo correspondiente.

$$\begin{array}{cccccccc}
X_0^{(1)} & X_1^{(1)} & X_2^{(1)} & \cdots & X_t^{(1)} & X_{t+1}^{(1)} & \cdots \\
X_0^{(2)} & X_1^{(2)} & X_2^{(2)} & \cdots & X_t^{(2)} & X_{t+1}^{(2)} & \cdots \\
X_0^{(3)} & X_1^{(3)} & X_2^{(3)} & \cdots & X_t^{(3)} & X_{t+1}^{(3)} & \cdots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots \\
X_0^{(M)} & X_1^{(M)} & X_2^{(M)} & \cdots & X_t^{(M)} & X_{t+1}^{(M)} & \cdots
\end{array}$$

3.1. Función de verosimilitud para modelos Markovianos

Antes de obtener la expresión de la función de verosimilitud para el caso no-homogéneo consideremos un ejemplo para el caso homogéneo.

Ejemplo 3.1. Suponga que $\mathcal{S} = \{0, 1, 2, 3\}$ es el espacio de estados de la cadena $\boldsymbol{X}$ y sea $\boldsymbol{D}$ una realización de esta cadena. Por ejemplo,

$$\boldsymbol{D} = \{0, 1, 1, 1, 2, 2, 3, 1, 0, 0, 2, 0, 0, 3, 1, 1, 2, 2, 3, 3\}.$$

En este caso $n_i(0)$ se transforma en $n_i = \sum_{i \in \mathcal{S}} I_i(X_0)$, donde $I_i(X_0)$ es igual a 1 si $X_0 = i$ y es igual a 0 en otro caso. También tenemos que $n_{i,j}(t) = n_{i,j}$ que es el número de transiciones del estado i al j, $i, j \in \mathcal{S}, t \in \mathcal{T}$, de la cadena X hasta el tiempo $t = 20$.

Por lo tanto, como se está asumiendo un modelo Markoviano homogéneo la función de verosimilitud está dada por

$$\begin{aligned} L(D|\theta) &\propto P_0^{(0)} P_{0,1} P_{1,1} P_{1,1} P_{1,2} P_{2,2} P_{2,3} P_{3,1} P_{1,0} P_{0,0} P_{0,2} P_{2,0} P_{0,0} P_{0,3} P_{3,1} P_{1,1} P_{1,2} P_{2,2} P_{2,3} P_{3,3} \\ &= \left(P_0^{(0)}\right)\left(P_1^{(0)}\right)^0\left(P_2^{(0)}\right)^0\left(P_3^{(0)}\right)^0 P_{0,0}^2 P_{0,1} P_{0,2} P_{0,3} P_{1,0} P_{1,1}^3 P_{1,2}^2 \left(P_{1,3}\right)^0 P_{2,0}\left(P_{2,1}\right)^0 \\ &\quad P_{2,2}^2 P_{2,3}^2 \left(P_{3,0}\right)^0 P_{3,1}^2 \left(P_{3,2}\right)^0 P_{3,3}. \end{aligned}$$

Se puede notar que en el presente caso se pueden obtener el vector con los valores $n_i(0)$ y la matriz con los valores $n_{i,j}$ de la siguiente forma:

$$n_i(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad n_{i,j} = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 3 & 2 & 0 \\ 1 & 0 & 2 & 2 \\ 0 & 2 & 0 & 1 \end{pmatrix}.$$

En un modelo Markoviano los parámetros del modelo son el vector de distribución inicial y la matriz de transición. Debido a que, en el caso no-homogéneo, el interés es obtener las distribuciones *a posteriori* tanto del vector de probabilidades iniciales $P^{(0)}$ como de las matrices de transición $P^{(t,t+1)}$ para $t \in \mathcal{T}$ se presenta el siguiente resultado [19, 20].

Proposición 3.1. *La función de verosimilitud para un modelo Markoviano no-homogéneo con espacio de estados $\mathcal{S}$ y conjunto de índices $\mathcal{T}$ es*

$$\begin{aligned} \mathcal{L}(\boldsymbol{D}|\theta) &= \mathcal{L}\left(\boldsymbol{D}\middle|P^{(0)}, P^{(t,t+1)}, t \in \mathcal{T}\right) \\ &= \mathcal{L}^{(0)}\left(\boldsymbol{D}\middle|P^{(0)}\right)\left(\prod_{t\in\mathcal{T}} \mathcal{L}^{(t)}\left(\boldsymbol{D}\middle|P^{(t,t+1)}\right)\right), \end{aligned} \tag{3.1}$$

donde

$$\mathcal{L}^{(0)}\left(\boldsymbol{D}\middle|P^{(0)}\right) \propto \prod_{i\in\mathcal{S}} \left(P_i^{(0)}\right)^{n_i(0)} = \prod_{i\in\mathcal{S}} (P(X_0 = i))^{n_i(0)} \tag{3.2}$$

y para $t \in \mathcal{T}$

$$\mathcal{L}^{(t)}\left(\boldsymbol{D}\middle|P^{(t,t+1)}\right) \propto \prod_{i\in\mathcal{S}}\prod_{j\in\mathcal{S}}\left(P_{i,j}^{(t,t+1)}\right)^{n_{i,j}(t)} = \prod_{i\in\mathcal{S}}\prod_{j\in\mathcal{S}}(P(X_{t+1}=j|X_t=i))^{n_{i,j}(t)}, \tag{3.3}$$

donde $\boldsymbol{D} = \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, ..., \boldsymbol{X}^{(M)}\}$.

Demostración. Por simplicidad tomamos el caso en el que $\mathcal{T} = \{0, 1, 2, ..., T\}$.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{D}|\theta) &= \prod_{i\in\mathcal{S}}\left(P_i^{(0)}\right)^{n_i(0)} \prod_{\substack{x_t\in\mathcal{S}\\ t\in\mathcal{T}}}\left(P_{x_0,x_1}^{(0,1)}, P_{x_1,x_2}^{(1,2)}, ..., P_{x_{T-1},x_T}^{(T-1,T)}\right)^{n_{x_0,x_1,...,x_T}} \\
&= \prod_{i\in\mathcal{S}}\left(P_i^{(0)}\right)^{n_i(0)} \prod_{\substack{x_t\in\mathcal{S}\\ t\in\mathcal{T}}}\left(P_{x_0,x_1}^{(0,1)}\right)^{n_{x_0,x_1,...,x_T}} \prod_{\substack{x_t\in\mathcal{S}\\ t\in\mathcal{T}}}\left(P_{x_1,x_2}^{(1,2)}\right)^{n_{x_0,x_1,...,x_T}} \\
&\quad ... \prod_{\substack{x_t\in\mathcal{S}\\ t\in\mathcal{T}}}\left(P_{x_{T-1},x_T}^{(T-1,T)}\right)^{n_{x_0,x_1,...,x_T}} \\
&= \prod_{i\in\mathcal{S}}\left(P_i^{(0)}\right)^{n_i(0)} \prod_{x_0,x_1}\left(P_{x_0,x_1}^{(0,1)}\right)^{n_{x_0,x_1}(1)} \prod_{x_1,x_2}\left(P_{x_1,x_2}^{(1,2)}\right)^{n_{x_1,x_2}(2)} \\
&\quad ... \prod_{x_{T-1},x_T}\left(P_{x_{T-1},x_T}^{(T-1,T)}\right)^{n_{x_{T-1},x_T}(T)} \\
&= \prod_{i\in\mathcal{S}}\left(P_i^{(0)}\right)^{n_i(0)} \prod_{t\in\mathcal{T}\setminus\{T\}}\prod_{i\in\mathcal{S}}\prod_{j\in\mathcal{S}}\left(P_{i,j}^{(t,t+1)}\right)^{n_{i,j}(t)} \\
&= \mathcal{L}^{(0)}\left(\boldsymbol{D}|P^{(0)}\right)\left(\prod_{t\in\mathcal{T}\setminus\{T\}}\mathcal{L}^{(t)}\left(\boldsymbol{D}|P^{(t,t+1)}\right)\right).
\end{aligned}$$

□

Proposición 3.2. *Si el vector de probabilidades iniciales* $P^{(0)}$ *de una cadena de Markov con espacio de estados* $\mathcal{S} = \{0, 1, 2, ..., K\}$ *tiene como distribución a priori una distribución de Dirichlet, entonces su distribución a posteriori también es de Dirichlet.*

Demostración. Asumimos que el vector de probabilidades iniciales tiene una distribución *a priori* de Dirichlet, es decir, $P^{(0)} = (P_0^{(0)}, P_1^{(0)}, ..., P_K^{(0)}) \overset{\mathfrak{D}}{=} Dir(\alpha_0(0), \alpha_1(0), ..., \alpha_K(0))$.

Usando (2.2) tenemos la siguiente relación de proporcionalidad.

$$\begin{aligned} p(P^0|\boldsymbol{D}) &\propto \mathcal{L}(\boldsymbol{D}|P^{(0)})p(P^{(0)}) \\ &\propto \left[\prod_{i=0}^{K}(P_i^{(0)})^{n_i(0)}\right]\left[\frac{\Gamma(\sum_{i=0}^{K}\alpha_i(0))}{\prod_{i=0}^{K}\Gamma(\alpha_i(0))}\prod_{i=0}^{K}(P_i^{(0)})^{\alpha_i(0)-1}\right] \\ &\propto \prod_{i=0}^{K}(P_i^{(0)})^{\alpha_i(0)-1}(P_i^{(0)})^{n_i(0)} \\ &\propto \prod_{i=0}^{K}(P_i^{(0)})^{n_i(0)+\alpha_i(0)-1}. \end{aligned}$$

Por lo tanto $p(P^{(0)}|\boldsymbol{D}) \propto \text{Dir}(n_0(0)+\alpha_0(0), n_1(0)+\alpha_1(0), ..., n_K(0)+\alpha_K(0))$. □

Proposición 3.3. *Asumiendo que los renglones de la matriz de transición $P^{(t,t+1)}, t \in \mathcal{T}$ de una cadena de Markov no-homogénea con espacio de estados $\mathcal{S} = \{0, 1, 2, ..., K\}$ son independientes y para $t, s \in \mathcal{T}$, $t \neq s$, $P^{(t,t+1)}$ y $p^{(s,s+1)}$ también son independientes. Si la matriz de transición $P^{(t,t+1)}, t \in \mathcal{T}$ tiene como distribución a priori una distribución de Dirichlet, entonces su distribución a posteriori también es de Dirichlet.*

Demostración. Asumimos que los renglones de la matriz de transición $P^{(t,t+1)}$ son independientes con distribución *a priori* de Dirichlet, es decir, para $t \in \mathcal{T}, i \in \mathcal{S}$ fijos tenemos que $(P_{i,0}^{(t,t+1)}, P_{i,1}^{(t,t+1)}, ..., P_{i,K}^{(t,t+1)}) \overset{\mathfrak{D}}{=} \text{Dir}(\alpha_{i,0}(t), \alpha_{i,1}(t), ..., \alpha_{i,K}(t))$.

Usando (2.2) tenemos la siguiente relación de proporcionalidad.

$$\begin{aligned} p(P^{(t,t+1)}|\boldsymbol{D}) &\propto \mathcal{L}(\boldsymbol{D}|P^{(t,t+1)})p(P^{(t,t+1)}) \\ &\propto \left[\prod_{i=0}^{K}\prod_{j=0}^{K}(P_{i,j}^{(t,t+1)})^{n_{i,j}(t)}\right]\left[\prod_{i=0}^{K}\left[\frac{\Gamma(\sum_{j=0}^{K}\alpha_{i,j}(t))}{\prod_{j=0}^{K}\Gamma(\alpha_{i,j}(t))}\prod_{j=0}^{K}(P_{i,j}^{(t,t+1)})^{\alpha_{i,j}(t)-1}\right]\right] \\ &\propto \prod_{i=0}^{K}\left[\prod_{j=0}^{K}(P_{i,j}^{(t,t+1)})^{\alpha_{i,j}(t)-1}(P_{i,j}^{(t,t+1)})^{n_{i,j}(t)}\right] \\ &\propto \prod_{i=0}^{K}\left[\prod_{j=0}^{K}(P_{i,j}^{(t,t+1)})^{n_{i,j}(t)+\alpha_{i,j}(t)-1}\right]. \end{aligned}$$

Por lo tanto $p(P^{(t,t+1)}|\boldsymbol{D}) \propto \prod_{i=0}^{K} \mathrm{Dir}(n_{i,0}(t)+\alpha_{i,0}(t), n_{i,1}(t)+\alpha_{i,1}(t), ..., n_{i,K}(t)+\alpha_{i,K}(t))$.

□

Capítulo 4

Aplicación a datos reales

Identificar los factores genéticos involucrados en las enfermedades complejas o multifactoriales es de gran importancia porque este conocimiento podrá contribuir a nuevas estrategias de prevención, métodos de diagnóstico y/o mejorar tratamientos. En el caso particular para el estudio de enfermedades de la población mexicana se debe considerar que esta población posee una estructura de mezcla genética reciente de distintos grupos étnicos continentales, como son las poblaciones nativa americana (indígena), europea y africana. Esta característica de la población puede representar una limitación en los estudios de asociación o una ventaja en los estudios de mapeo genético basado en la mezcla de poblaciones (*Admixture Mapping*, AM por sus siglas en inglés). El AM está diseñado para identificar factores genéticos de riesgo asociados a enfermedades o rasgos multifactoriales que muestran diferencias importantes de prevalencia entre las poblaciones ancestrales.

En este capítulo se propone aplicar un modelo estadístico que implementa cadenas de Markov no-homogeneas y un enfoque Bayesiano para realizar un mapeo genético basado en la mezcla de poblaciones, es decir, identificar regiones de enriquecimiento de alguna ancestría en particular en un grupo de individuos. La aplicación del modelo se basa en la simulación de cadenas de Markov no-homogéneas, debido a que la distribución de las

ancestrías locales asignadas varía a lo largo de los cromosomas, es decir, el comportamiento de la asignación de ancestría no es uniforme a lo largo de los cromosomas. Se asume que las distribuciones *a priori* de los vectores de probabilidad inicial y de los renglones de las matrices de transición son de Dirichlet. Además, las distribuciones *a posteriori* también tienen una distribución de Dirichlet dado que esta distribución forma parte de una familia conjugada para el modelo multinomial, que es la base del modelo de Markov no-homogéneo que se considera. Otra ventaja es que el dominio de la distribución de Dirichlet es un simplex, que es donde están definidos los vectores de probabilidad inicial y los renglones de las matrices de transición. Esta distribución permite incorporar información inicial de los hiperparámetros del modelo mediante su vector de parámetros. También permite la simulación de transiciones que posiblemente no fueron observadas, cuando se trata de una muestra pequeña, de una forma consistente con la muestra observada. Se podrían utilizar algunas otras distribuciones como distribuciones *a priori*, sin embargo podría ser que la distribución *a posteriori* correspondiente no sea una distribución de forma conocida y por lo tanto se deberían utilizar otros métodos más complejos. A continuación se describe el modelo y algoritmo desarrollados.

Para realizar un análisis de AM utilizando datos reales de dos grupos de individuos, un grupo de individuos sanos a los que se denominan "controles" y otro grupo de individuos afectados por una enfermedad que llamada Neuromielitis Óptica a los que se denominan "casos". A continuación se presentara una breve descripción de la enfermedad bajo estudio.

La Neuromielitis Óptica (NMO) es una enfermedad desmielinizante, inflamatoria, crónica del sistema nervioso central que afecta selectivamente al nervio óptico y la médula espinal, causando ceguera y parálisis en individuos de entre 35 y 47 años, y se presenta con mayor frecuencia en mujeres.

Durante mucho tiempo se consideró como una variante de la Esclerosis Múltiple. Se considera que muchos pacientes con NMO han sido mal diagnosticados con Esclerosis Múltiple [21], por lo tanto la prevalencia en el mundo no está bien establecida. Sin embargo, a pesar de la falta de estimaciones epidemiológicas precisas, se ha sugerido que la NMO es más frecuente en poblaciones de origen no-europeo [22]. En México, se reportó una prevalencia de 1.3 por cada 100,000 pacientes de origen mestizo mexicano, y se sugiere que la enfermedad parece tener una presentación más grave en nuestra población [23].

Como muchas otras enfermedades autoinmunes, la NMO pertenece a un grupo de desórdenes multifactoriales que resultan de interacciones complejas entre factores genéticos y ambientales. Para NMO existen muy pocos estudios con respecto a los factores genéticos; se estima que al menos el 3 % de los pacientes tiene un familiar afectado, lo que respalda la participación de factores genéticos [24].

Aún cuando los estudios epidemiológicos para NMO son escasos, de manera consistente sugieren diferencias basadas en el origen étnico. En contraste con Esclerosis Múltiple, la NMO parece ser más frecuente en población de origen no-europeo, lo que podría sugerir la participación de factores genéticos propios de la población indígena y poblaciones descendientes.

Objetivo

Implementar un análisis de *Admixture Mapping* utilizando un modelo basado en cadenas de Markov no-homogéneas y un enfoque Bayesiano.

4.1. Modelo matemático

Se consideran los siguientes parámetros, sean $M_c \geq 0$ un número natural que representa el tamaño de muestra observada de individuos de la población afectada (casos) y $M_s \geq 0$

un número natural que representa el tamaño de la muestra observada de individuos de la población sana (controles).

Los datos están conformados por la asignación de ancestría en cada uno de los segmentos a partir de una partición en $T \geq 0$ segmentos de cada uno de los cromosomas homólogos, considerando tres poblaciones ancestrales: africana, europea y nativa. La información de ancestría local asignada a la pareja de cromosomas homólogos de cada individuo de ambas muestras puede ser representada mediante dos sucesiones de indicadores de ancestría en cada uno de los segmentos de la partición, donde la información de ancestría está codificada de la siguiente forma:

- A: corresponde a la población africana.
- E: corresponde a la población europea.
- N: corresponde a la población nativa americana.

Así, para cada individuo X, la información de ancestría se describe mediante dos cadenas $X^{(0)} = \{X^{(0)}(t), t \geq 0\}$ y $X^{(1)} = \{X^{(1)}(t), t \geq 0\}$, donde cada cadena es la representación de cada uno de los cromosomas homólogos. Los valores de ancestría asignados a cada uno de los segmentos t de la partición de la pareja de cromosomas homólogos son representados por $X^{(0)}(t) = x^{(0)}(t)$ y $X^{(1)}(t) = x^{(1)}(t)$.

Formalmente podemos describir esta información como se presenta a continuación.
Sean:

- $\mathcal{M}_c = \{0, 1, ..., M_c\}$
- $\mathcal{M}_s = \{0, 1, ..., M_s\}$
- $\mathcal{T} = \{0, 1, ..., T\}$
- $\mathcal{S}' = \{\ A,\ E,\ N\ \}$

Para $m_i \in \mathcal{M}_i, i \in \{c, s\}$ se define:

$$X_{m_i} = \{X_{m_i}^{(0)}, X_{m_i}^{(1)}\},$$

donde

$$X_{m_i}^{(j)} = \{X_{m_i}^{(j)}(t), t \in \mathcal{T}\}, j \in \{0, 1\},$$

es la cadena que describe la sucesión de ancestría en el cromosoma j (de la pareja de cromosomas homólogos) del individuo m_i y con espacio de estados $\mathcal{S}'$.

Así las observaciones de las muestras pueden denotarse de la siguiente forma:

$$X_{m_i}^{(j)} = x_{m_i}^{(j)} \quad \text{donde} \quad x_{m_i}^{(j)} = \{x_{m_i}^{(j)}(t) = s' : s' \in \mathcal{S}', t \in \mathcal{T}\}, j \in \{0, 1\}.$$

Debido a que nos interesa identificar regiones de enriquecimiento de una ancestría en particular, es necesario re-codificar la información de ancestría local en la pareja de cromosomas homólogos para cada individuo de las muestras de forma adecuada.

Definimos el nivel de ancestría como la representación de ancestría de interés asignada a un segmento t de la partición en ambos cromosomas homólogos para cada individuo. Esta representación se identifica con los números naturales 0 si la ancestría de interés no está asignada en ningún cromosoma de la pareja homóloga, 1 si la ancestría de interés está asignada en exactamente un cromosoma de la pareja de cromosomas homologos y 2 si la ancestría de interés está asignada en ambos cromosomas homólogos.

Es decir, para la pareja de cromosomas homólogos de un individuo se tienen las siguientes asignaciones de ancestría local en cualquier segmento t,

$$\begin{array}{ccc} (N,N) & (N,E) & (N,A) \\ (E,N) & (E,E) & (E,A) \\ (A,N) & (A,E) & (A,A) \end{array}$$

y si el interés es en la ancestría nativa americana (N) las asignaciones de nivel de ancestría corresponden a la siguiente representación,

$$\begin{matrix} 2 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{matrix}$$

Podemos observar que para cada individuo los valores de nivel de ancestría describen una sucesión sobre la partición del cromosoma. Además, ya que el fenómeno de recombinación ocurre de forma aleatoria a lo largo de los cromosomas, el nivel de ancestría puede ser representado mediante una cadena de Markov. Para dicha representación, el conjunto $\mathcal{S} = \{0, 1, 2\}$ es el espacio de estados de la cadena de Markov no-homogénea que registra la información de ancestría presente en ambos cromosomas homólogos y para $m_k \in \mathcal{M}_k, k \in \{c, s\}, t \in T$, definimos la cadena como: $X'_{m_k} = \{X'_{m_k}(t) : t \in T\}$. De esta forma, para $s' \in S'$ el nivel de ancestría de interés es el siguiente:

$$\begin{aligned} X'_{m_k}(t) = 0 \quad & \text{si} \quad X^{(0)}_{m_k}(t) \neq s' \quad \text{y} \quad X^{(1)}_{m_k}(t) \neq s' \\ X'_{m_k}(t) = 1 \quad & \text{si} \begin{cases} X^{(0)}_{m_k}(t) \neq s' \quad \text{y} \quad X^{(1)}_{m_k}(t) = s' \\ X^{(0)}_{m_k}(t) = s' \quad \text{y} \quad X^{(1)}_{m_k}(t) \neq s' \end{cases} \\ X'_{m_k}(t) = 2 \quad & \text{si} \quad X^{(0)}_{m_k}(t) = s' \quad \text{y} \quad X^{(1)}_{m_k}(t) = s' \end{aligned}$$

Una vez que se cuenta con la información de nivel de ancestría de los casos y los controles, se obtienen las frecuencias absolutas de los valores de nivel nivel de ancestría como se describe a continuación.

Definición 4.1. Consideremos las siguientes frecuencias absolutas sobre las muestras observadas (casos y controles).

- $n_i^{(k)}(t) := |\{m_k \in \mathcal{M}_k : X'_{m_k}(t) = i\}| \quad t \in \mathcal{T}$,
- $n_{i,j}^{(k)}(t) := |\{m_k \in \mathcal{M}_k : X'_{m_k}(t) = i, X'_{m_k}(t+1) = j\}| \quad t \in \mathcal{T} \setminus \{T\}$,

para $i, j \in S, k \in \{c, s\}$, donde para A un conjunto, $|A|$ se entiende por su cardinalidad.

Las frecuencias absolutas en el primer segmento de la partición del cromosoma son:

$$n^k(0) = [n_0^{(k)}(0), n_1^{(k)}(0), n_2^{(k)}(0)] \quad k \in \{c, s\}.$$

Las frecuencias absolutas de transición describen una matriz,

$$n^{(k)}(t) = \begin{pmatrix} n_{0,0}^{(k)}(t) & n_{0,1}^{(k)}(t) & n_{0,2}^{(k)}(t) \\ n_{1,0}^{(k)}(t) & n_{1,1}^{(k)}(t) & n_{1,2}^{(k)}(t) \\ n_{2,0}^{(k)}(t) & n_{1,2}^{(k)}(t) & n_{2,2}^{(k)}(t) \end{pmatrix} \quad k \in \{c, s\}.$$

Cada individuo se puede describir mediante la realización de una cadena de Markov no-homogénea. Para poder estudiar el comportamiento de dichas muestras es necesario obtener las distribuciones de los vectores de probabilidiad inicial y de las matrices de transición que regulan a estas cadenas.

Suponga que los vectores de probabilidad inicial para cada uno de las cadenas de Markov tienen una distribución *a priori* de Dirichlet con parámetros $(\alpha_0(0), \alpha_1(0), \alpha_2(0))$. Considerando la Proposición 3.2 se tiene que los vectores de probabilidad inicial tienen una distribución *a posterirori* de Dirichlet, es decir,

$$P_k^{(0)} \sim \text{Dir}\Big(n_0^{(k)}(0) + \alpha_0(0), n_1^{(k)}(0) + \alpha_1(0), n_2^{(k)}(0) + \alpha_2(0)\Big), \quad k \in \{c, s\}.$$

De manera análoga, suponga que los renglones de las matrices de transición son independientes y tienen una distribución *a priori* de Dirichlet con parámetros $(\alpha_{i,0}(t), \alpha_{i,1}(t), \alpha_{i,2}(t))$, tal que $i \in \{0, 1, 2\}, t \in \mathcal{T} \setminus \{T\}$. Nuevamente, por la Proposición 3.3 se tiene que las matrices de transición tienen una distribución de Dirichlet, es decir,

$$P_k^{(t,t+1)} \sim \prod_{i=0}^{2} \text{Dir}\Big(n_{i,0}^{(k)}(t) + \alpha_{i,0}(t), n_{i,1}^{(k)}(t) + \alpha_{i,1}(t), n_{i,2}^{(k)}(t) + \alpha_{i,2}(t)\Big), \quad k \in \{c, s\}.$$

Dado que se tienen totalmente determinadas las cadenas de Markov que describen a los elementos de las muestras observadas, se pueden hacer inferencias sobre la distribución del nivel de ancestría en cada segmento de la partición mediante un método de simulación de las probabilidades iniciales y las matrices de transición.

4.2. Algoritmo

A continuación se describe el algoritmo para aproximar la distribución de la proporción del nivel de ancestría en cada segmento cromosómico.

- Obtenga una muestra aleatoria de la distribución *a posteriori* del vector de probabilidad inicial $P_k^{(0)}$.

$$\left(P_{0_k}^{(0)}, P_{1_k}^{(0)}, P_{2_k}^{(0)}\right) = \left(p_{0_k}^{(0)}, p_{1_k}^{(0)}, p_{2_k}^{(0)}\right), \quad k \in \{c, s\}.$$

- Obtenga una muestra aleatoria de las matrices de transición $P_k^{(t,t+1)}$ para cada segmento $t \in \mathcal{T} \setminus \{T\}, k \in \{c, s\}$.

$$\begin{pmatrix} P_{0,0_k}^{(t,t+1)} & P_{0,1_k}^{(t,t+1)} & P_{0,2_k}^{(t,t+1)} \\ P_{1,0_k}^{(t,t+1)} & P_{1,1_k}^{(t,t+1)} & P_{1,2_k}^{(t,t+1)} \\ P_{2,0_k}^{(t,t+1)} & P_{2,1_k}^{(t,t+1)} & P_{2,2_k}^{(t,t+1)} \end{pmatrix} = \begin{pmatrix} p_{0,0_k}^{(t,t+1)} & p_{0,1_k}^{(t,t+1)} & p_{0,2_k}^{(t,t+1)} \\ p_{1,0_k}^{(t,t+1)} & p_{1,1_k}^{(t,t+1)} & p_{1,2_k}^{(t,t+1)} \\ p_{2,0_k}^{(t,t+1)} & p_{2,1_k}^{(t,t+1)} & p_{2,2_k}^{(t,t+1)} \end{pmatrix}$$

- Obtenga las probabilidades de nivel de ancestría en cada segmento, esto se hace de forma recursiva como se muestra a continuación. Para $i, j \in \mathcal{S}, t \in \mathcal{T}, k \in \{c, s\}$ se tiene que,

$$P_{i_k}(t) = \begin{cases} p_{i_k}^{(0)} & \text{si} \quad t = 0 \\ \sum_{j \in \mathcal{S}} p_{j,i_k}^{(t-1,t)} p_{j_k}(t-1) & \text{si} \quad t \neq 0 \end{cases}$$

- Calcule el valor esperado de nivel de ancestría en cada segmento.

$$\mathbf{E}_k(t) = \left(0P_{0_k}(t) + 1P_{1_k}(t) + 2P_{2_k}\right) \quad t \in \mathcal{T}, k \in \{c, s\}.$$

A continuación se presentan las iteraciones de este algoritmo para obtener muestras de los valores esperados de ancestría para las dos poblaciones (casos y controles).

Para $N \in \mathbb{N}, N > 0$ y $k \in \{c, s\}$ en cada iteración obtenemos los datos descritos en el algoritmo y para distinguirlos se indexan con $n \in \{0, 1, ..., N\}$ de la siguiente forma:

- Muestra aleatoria del vector de probabilidad inicial, $P_k^{(0),n}$.
- Muestra aleatoria de las matrices de transición, $P_k^{(t,t+1),n}$.
- Probabilidades de nivel de ancetría en cada segmento, $P_{i_k}^n(t)$.
- Valores esperados del nivel de ancestría en cada segmento, $\mathbf{E}_k^n(t)$.

Notemos que con los valores esperados de ancestría simuladas describimos una matriz donde cada renglón corresponde a la información obtenida de las iteraciones.

$$\mathbf{E}_k = \begin{pmatrix} E_k^1(0) & E_k^1(1) & \cdots & E_k^1(t) & \cdots E_k^1(T) \\ E_k^2(0) & E_k^2(1) & \cdots & E_k^2(t) & \cdots E_k^2(T) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ E_k^N(0) & E_k^N(1) & \cdots & E_k^N(t) & \cdots E_k^N(T) \end{pmatrix}.$$

4.3. Aplicación del modelo

Se cuenta con información genotipificada de la pareja de cromosomas homólogos correspondientes al cromosoma 1 para un grupo de 83 individuos afectados con NMO (casos) y un grupo de 97 individuos sanos (controles); es decir, se tiene información genotipificada de 166 y 194 cromosomas homólogos respectivamente. La información corresponde a una estimación de ancestría local para cada uno de estos cromosomas; esta estimación se obtuvo mediante el programa *PCAdmix* [12].

Utilizando estos datos, se identifican los parámetros del modelo Markoviano como sigue: $M_c = 83$ que corresponde al tamaño de la muestra del grupo de la población afectada, $M_s = 97$ que corresponde al tamaño de la muestra de la población sana, $T = 783$ que corresponde al tamaño de la partición de la pareja de cromosomas homólogos.

La ancestría nativa de interés para este trabajo corresponde a la población nativa americana, por tanto el archivo inicial es codificado respecto a esta población ancestral de la siguiente forma:

$$
\begin{aligned}
X'_{m_k}(t) = 0 \quad &\text{si} \quad X^{(0)}_{m_k}(t) \neq N \quad \text{y} \quad X^{(1)}_{m_k}(t) \neq N \\
X'_{m_k}(t) = 1 \quad &\text{si} \begin{cases} X^{(0)}_{m_k}(t) \neq N \quad \text{y} \quad X^{(1)}_{m_k}(t) = N \\ X^{(0)}_{m_k}(t) = N \quad \text{y} \quad X^{(1)}_{m_k}(t) \neq N \end{cases} \\
X'_{m_k}(t) = 2 \quad &\text{si} \quad X^{(0)}_{m_k}(t) = N \quad \text{y} \quad X^{(1)}_{m_k}(t) = N
\end{aligned}
$$

Después de realizar la asignación correspondiente del nivel de ancestría se procede a calcular las frecuencias absolutas de los valores de nivel de ancestría.

Para llevar a cabo la aplicación del modelo, se utilizaron dos distribuciones *a priori* con un fin comparativo, una que se denominarán informativa y otra no-informativa. A continuación se presentan las distribuciones para los modelos correspondientes a estas distribuciones.

4.3.1. Modelo no-informativo

Para la distribución *a priori* de Dirichlet no-informativa los valores de los hiperparámetros poseen el mismo valor, es decir, $\alpha_i(0) = c$ y $\alpha_{ij}(t) = d$ para toda $i, j \in S$, $k \in \{c, s\}$ y toda $t \in T$. En el presente caso se eligen $c = d = \frac{1}{3}$, por lo tanto las distribuciones *a posteriori* para el vector de probabilidades iniciales y para las matrices de transición de ambas poblaciones, $k \in \{0, 1\}$, son las siguientes:

$$
\begin{aligned}
P^{(0)}_k &\propto Dir\Big(n^k_0(0) + \frac{1}{3}, n^k_1(0) + \frac{1}{3}, n^k_2(0) + \frac{1}{3}\Big), \\
P^{(t,t+1)}_k &\propto \prod_{i=0}^{2} Dir\Big(n^k_{i,0}(t) + \frac{1}{3}, n^k_{i,1}(t) + \frac{1}{3}, n^k_{i,2}(t) + \frac{1}{3}\Big).
\end{aligned}
$$

4.3.2. Modelo informativo

Para la distribución *a propri* informativa los valores de los hiperparámetros están determinados de forma arbitraria mediante una función de distancia que asigna valores enteros positivos proporcionales a las frecuencias absoluta obtenidas. Para este trabajo se toman valores enteros entre 2 y 5 en la asignación de los hiperparámetros. Para el vector de frecuencias absolutas iniciales para cada renglón (vector) de las matrices de frecuencias absolutas de transición la función asigna el valor de 2 al valor mínimo del vector de frecuencias absolutas y un valor entero entro entre 2 y 5 dependiendo de la distancia del valor mínimo y el valor máximo del vector de frecuencias absolutas. Por ejemplo, al vector $[23, 43, 31]$ se le asignan los valores de $[2, 5, 3]$. Por lo tanto, las distribuciones *a posteriori* para el vector de probabilidades iniciales y para las matrices de transición de ambas poblaciones, $k \in \{c, s\}$, son las siguientes: para $\alpha_i(t), \alpha_{i,j}(t) \in \{2, 3, 4, 5\}, i, j = 0, 1, 2$.

$$P_k^{(0)} \sim \text{Dir}\Big(n_0^k(0) + \alpha_0(0), n_1^k(0) + \alpha_1(0), n_2^k(0) + \alpha_2(0)\Big),$$
$$P_k^{(t,t+1)} \sim \prod_{i=0}^{2} \text{Dir}\Big(n_{i,0}^k(t) + \alpha_{i0}(t), n_{i,1}^k(t) + \alpha_{i1}(t), n_{i,2}^k(t) + \alpha_{i2}(t)\Big).$$

4.4. Resultados

Se realizaron $N = 10,000$ iteraciones del algoritmo para ambas poblaciones de estudio (casos y controles) y para cada una de las distribuciones *a priori* consideradas (informativa y no-informativa). A continuación se presentan algunos resultados sobre el valor esperado del nivel de ancestría para ambos modelos, el informativo y el no-informativo.

En las Figuras 4.1 y 4.2 se muestran los percentiles 0.025, 0.5 y 0.975 de lo valores esperados de nivel de ancestría $\mathbf{E_0}$ y $\mathbf{E_1}$. Se obtienen los intervalos de credibilidad sobre el valor esperado del nivel de ancestría para cada uno de los segmentos del cromosoma 1.

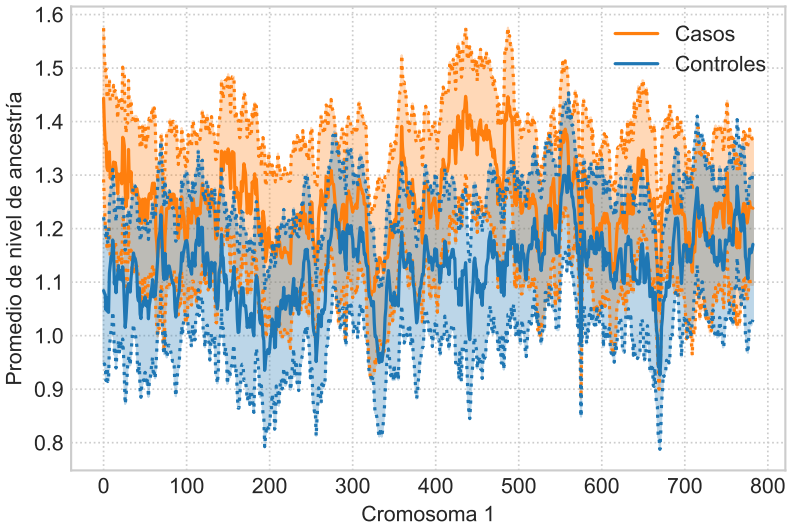


Figura 4.1: Promedio de nivel de ancestría para las poblaciones casos y controles en el modelo informativo.

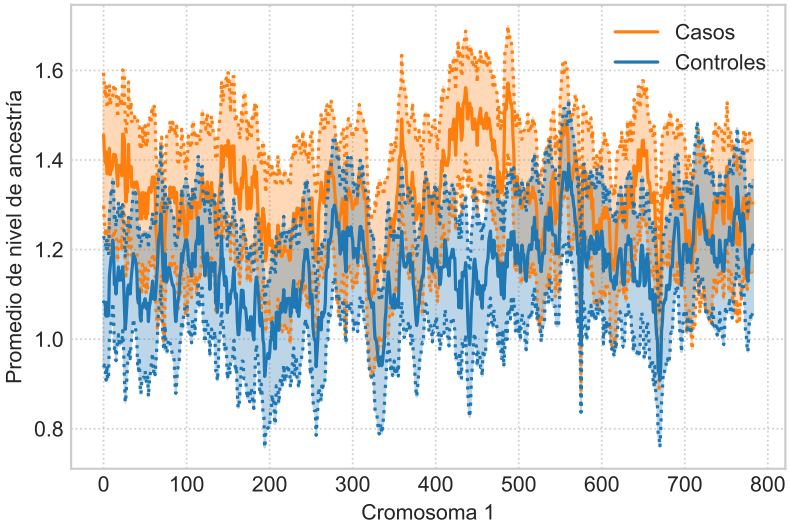


Figura 4.2: Promedio de nivel de ancestría para las poblaciones casos y controles en el modelo no-informativo.

De forma general, se puede observar que en general los casos presentan una ancestría nativo americana mayor en comparación con los controles a lo largo del cromosoma 1; esto indica una diferencia de etnicidad entre ambos grupos.

En algunas secciones del cromosoma 1 los intervalos de credibilidad se encuentran ligeramente separadas (no se intersectan). En las Figuras 4.3 y 4.4 se muestran los segmentos a lo largo del cromosoma en donde los intervalos de credibilidad de los dos grupos no se intersectan. En el caso particular del modelo no-informativo (figura 4.2) se detectaron 65 segmentos donde no se intersectan los intervalos de credibilidad y la región donde existe una mayor separación ente los intervalos de credibilidad se describe del segmento 433 al segmento 445 alcanzando su máxima separación en el segmento 441. De forma análoga, para el caso particular del modelo informativo (figura 4.2) se detectaron 34 segmentos donde no se intersectan los intervalos de credibilidad y la región donde existe una mayor separación ente los intervalos de credibilidad se describe del segmento 433 al segmento 443, alcanzando su máxima separación en el segmento 441.

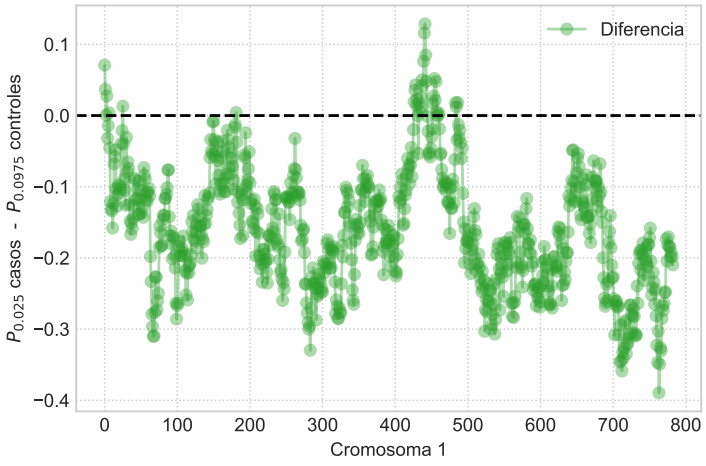


Figura 4.3: Diferencia de nivel de ancestría esperado para las poblaciones casos y controles en el modelo informativo.

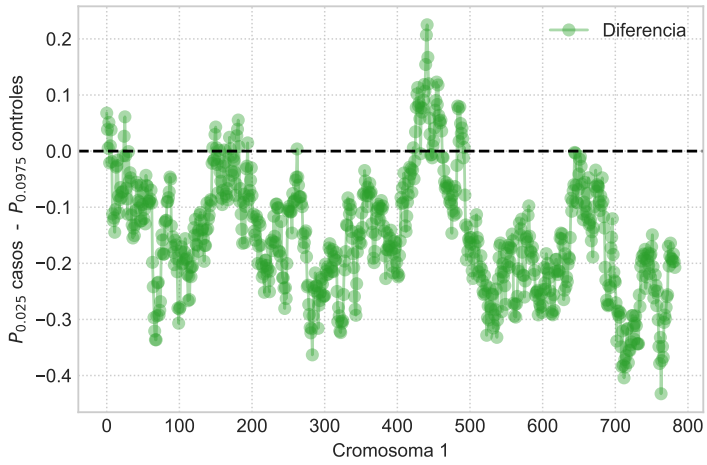


Figura 4.4: Diferencia de nivel de ancestría esperado para las poblaciones casos y controles en el modelo no-informativo.

En las Figuras 4.5 y 4.6 se presentan los histogramas del valor esperado del nivel de ancestría específicamente en el segmento 441 para el grupo de casos y el grupo de controles. Se puede observar el comportamiento de las frecuencias de los valores de nivel de ancestría con respecto a la media observada de nivel de ancestría en los datos reales.

En el modelo informativo se observa que la frecuencia de nivel de ancestría para los controles se distribuye alrededor de la media observada en los datos reales; en contraste, el nivel de ancestría para los casos se distribuye por debajo de la media observada en los datos reales (Figura 4.5). Por su parte, en el modelo no-informativo se observa que el nivel de ancestría tanto para los casos como para los controles la frecuencia de nivel de ancestría se distribuyen alrededor de sus respectivas medias observadas en los datos reales (Figura 4.6).

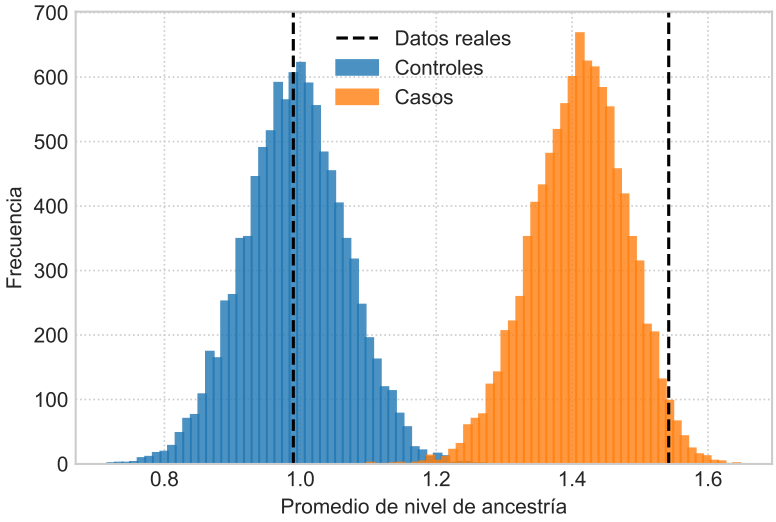


Figura 4.5: Distribución del valor esperado de nivel de ancestría en el segmento 441 de las poblaciones casos y controles para el modelo informativo.

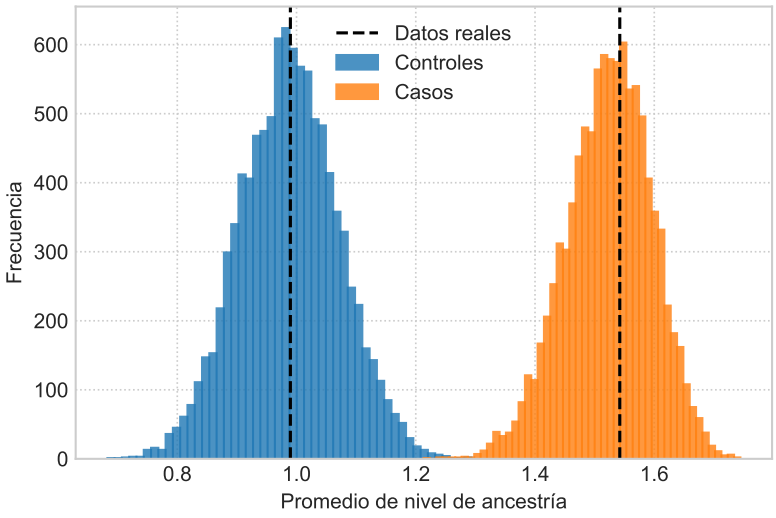
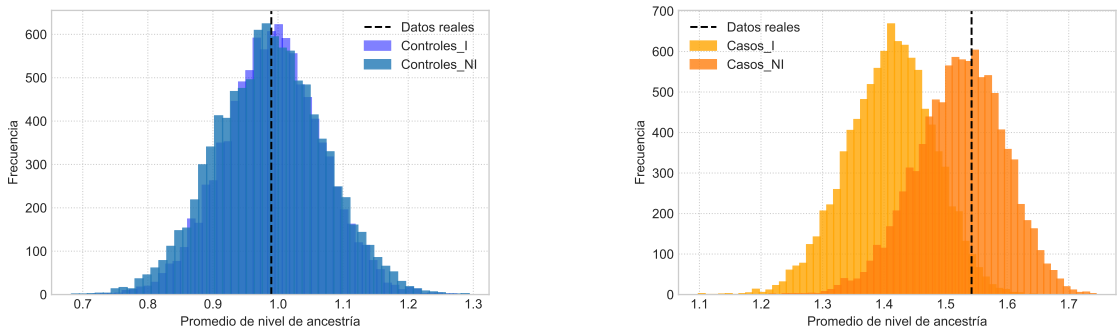


Figura 4.6: Distribución del valor esperado de nivel de ancestría en el segmento 441 de las poblaciones casos y controles para el modelo no-informativo.

Se realiza una comparación entre los modelos informativo y no-informativo de las distribuciones de frecuencias del nivel de ancestría, para los casos y para los controles (Figura 4.7).



(a) Controles **(b)** Casos

Figura 4.7: Frecuencia del valor de nivel esperado de ancestría en el segmento 441 de las poblaciones controles y casos para los modelos informativo (I) y no-informativo (NI).

Para seguir el análisis sobre el segmento 441, en el Cuadro 4.1 se presenta la distribución de probabilidad y el valor esperado de nivel de ancestría nativo americano para los casos.

Casos	$P(0)$	$P(1)$	$P(2)$	$\mathbb{E}$
Datos observados	0.096	0.265	0.639	1.534
Datos simulados I	0.153	0.282	0.564	1.41
Datos simulados NI	0.105	0.268	0.627	1.522

Cuadro 4.1: Probabilidad de nivel de ancestría para la población de casos, ventana 441.

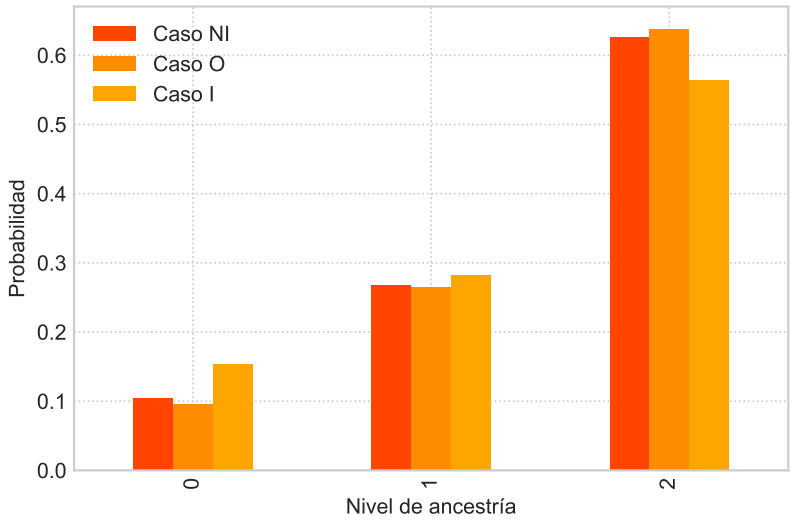


Figura 4.8: Distribución de probabilidad del valor de nivel de ancestría en la segmento 441.

Se puede observar que en las distribuciones de probabilidad simuladas con las distribuciones *a priori* informativa y no-informativa y la observada de los datos reales correspondientes al segmento 441, el valor más probable es 2, seguido del valor 1 y por último el valor menos probable es 0. Además, se puede notar que la distribución de probabilidad simulada mediante la distribución *a priori* no-informativa es semejante a la distribución observada en los datos reales; esto refleja el hecho de que no se incorporó ningún tipo de información previa adicional al modelo. Por otro lado, la distribución de probabilidad simulada mediante la distribución *a priori* informativa se comporta diferente en comparación con las otras dos distribuciones, ya que aumenta ligeramente la probabilidad de los valores de ancestría 0 y 1 y disminuye ligeramente la probabilidad del nivel de ancestría 2, de modo que las probabilidades reflejan un efecto de homogeneización. Esto se puede observar en la Figura 4.8.

De forma análoga, en el Cuadro 4.2 se presenta la distribución de probabilidad y el valor esperado de nivel de ancestría nativo americano para los controles.

Controles	$P(0)$	$P(1)$	$P(2)$	$\mathbb{E}$
Datos observados	0.361	0.289	0.351	.991
Datos simulados I	0.355	0.298	0.347	.992
Datos simulados NI	0.361	0.290	0.349	.988

Cuadro 4.2: Probabilidad de nivel de ancestría para la población control, ventana 441.

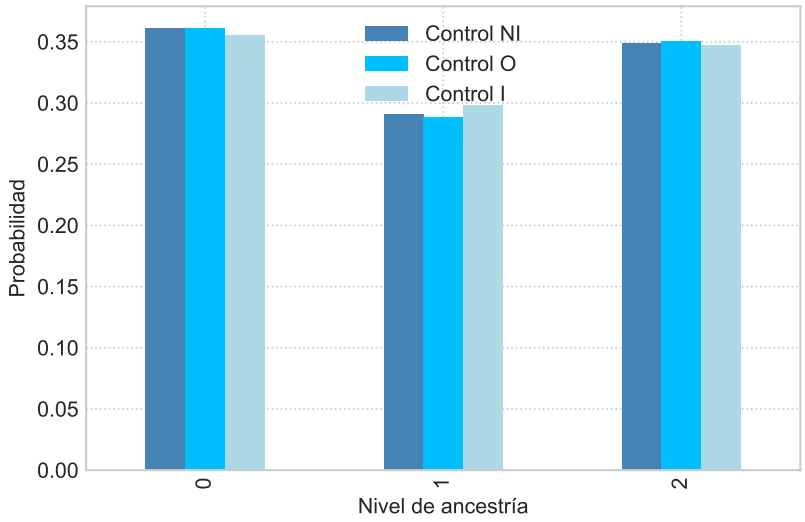


Figura 4.9: Distribución de probabilidad del valor de nivel de ancestría en la ventana 441 para datos pbservados (O) y modelos informativo (I) y no-informativo (NI).

En el Cuadro 4.2, que se refiere al grupo de los controles, se puede observar que la distribución de probabilidad del nivel de ancestría observada de los datos reales tiene un

comportamiento homogéneo, los valores de probabilidad de los niveles 2 y 0 son considerablemente parecidos y el valor de probabilidad del nivel 1 se encuentra ligeramente por debajo de las probabilidades restantes. Esto sugiere que no hay algún tipo de información ancestral adicional en la región cromosómica que determina el segmento 441. La distribución de probabilidad simulada mediante la distribución *a priori* no-informativa es sumamente semejante a la distribución de probabilidad observada. En el caso de la distribución de probabilidad simulada mediante la distribución *a priori* informativa también produce un efecto de homogeneización que es menos perceptible en comparación al grupo de los casos. Esto se pude observar en la Figura 4.9

Por último, se incorpora la Figura 4.10 que corresponde a los resultados obtenidos de un análisis frecuentista realizado mediante un modelo de regresión logística; el modelo parte de la misma asignación de nivel de ancestría, que es empleada como la variable independiente del modelo.

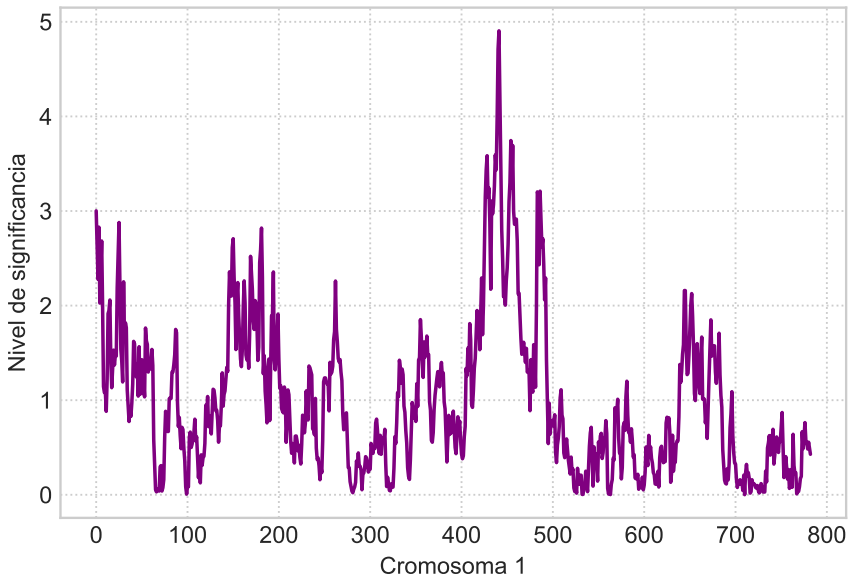


Figura 4.10: Resultados del análisis de un modelo frecuentista.

Se puede corroborar que hay un valor significativo en el segmento 441 del cromosoma 1, el cual es consistente con los resultados obtenidos de la aplicación del modelo en el presente trabajo.

4.5. Discusión

El estudio de enfermedades genéticas, en particular de las enfermedades multifactoriales resulta una tarea compleja debido a sus características. Como ya se ha mencionado en la Sección 1.3, uno de los principales métodos de estudio para este tipo de enfermedades es el análisis de asociación. Una de las principales desventajas de los estudios de asociación se presenta cuando hay diferencias de composición ancestral en los grupos de estudio, lo que se conoce como estratificación poblacional. Si no se toma en cuenta esta característica, puede convertirse en un factor de confusión y puede dar lugar a asociaciones falsas.

El método de *Admixture Mapping* (AM) está diseñado para aprovechar la estructura genética que tienen las poblaciones mezcladas para la localización de regiones cromosómicas asociadas al riesgo del desarrollo de enfermedades o rasgos multifactoriales. Las enfermedades o rasgos de interés ideales para un análisis de AM son aquellas cuya prevalencia es diferencial entre las poblaciones que se consideran ancestrales, debido a que estos métodos localizan regiones donde los individuos que tienen la enfermedad o rasgo presentan una mayor proporción de ancestría de la población ancestral donde dicha enfermedad o rasgo es mas frecuente.

Actualmente hay muy pocas propuestas metodológicas para los análisis de AM. Una de las propuestas más importantes es la del Dr. Nick Patterson en su artículo *Methods for High-Density Admixture Mapping of Disease Genes* [6], en la cual reporta una metodología que consta de cuatro puntos principales: 1) Proporciona un nuevo método para la estimación de ancestría local. 2) Evalúa el desempeño del método, particularmente en un

grupo de individuos afroamericanos. 3) Prueba el comportamiento del método mediante un proceso de simulaciones computacionales exhaustivas. 4) Explora el poder estadístico del AM para detectar loci de enfermedades en diversos escenarios de diferenciación en la frecuencia alélica y efectos genéticos, con datos reales y datos simulados. Observa que para alelos causantes de enfermedad con una amplia diferencia en las frecuencias alélicas entre poblaciones ancestrales, el AM puede detectar genes de un efecto moderado para la enfermedad [6].

Los métodos estadísticos empleados para la estimación de ancestría local en la propuesta del Dr. Nick Patterson hacen uso de modelos de Markov ocultos (HMM, por sus siglas en inglés) y algoritmos a través de métodos de Monte Carlo vía Cadenas de Markov (MCMC, por sus siglas en inglés). Para el análisis de la información utiliza un enfoque frecuentista basado en la comparación de medias de los parámetros de su modelo para obtener un nivel de significancia. Una de las principales desventajas del método que propone es que está diseñado para el estudio de poblaciones que son el resultado de la mezcla genética entre dos poblaciones, por lo que el modelo no se puede implementar en poblaciones como la mexicana debido a que su composición ancestral comprende más de dos poblaciones ancestrales.

En el presente trabajo se implementa un modelo de Markov no-homogéneo y un enfoque Bayesiano tomando como base datos reales de dos grupos de individuos, un grupo de 83 casos y otro de 97 controles. Se comparan dos distribuciones *a priori*, una la llamamos informativa y la otra no-informativa, ambas de Dirichlet. El modelo se implementa para realizar inferencias sobre la ancestría de una población con mezcla genética para poder hacer un análisis de AM. Se reporta la implementación de un método para comparar el nivel de ancestría de una población afectada con el de una población sana (respecto a la enfermedad). Posteriormente se presenta la implementación del modelo a datos reales

de dos poblaciones mexicanas, una afectada por neuromielitis óptica (NMO) y la otra sana.

El método se prueba a través de simulaciones computacionales de dos versiones del modelo, una simula distribuciones *a posteriori* de los vectores de probabilidad inicial y de las matrices de transición mediante la distribución *a priori* informativa y la otra mediante la distribución *a priori* no-informativa, utilizando las frecuencias de los valores de ancestría observados. La comparación entre las versiones informativa y no-informativa de la implementación del modelo permiten observar la sensibilidad de éste a los parámetros de las distribuciones *a priori* de Dirichlet.

Al aplicar el modelo a datos reales de las poblaciones antes mencionadas, una afectada por NMO y otra sana, se pueden observar diferencias en el comportamiento del nivel de ancestría nativa americana. Los intervalos de credibilidad obtenidos, mediante las versiones informativa y no-informativa del modelo, sugieren que, a lo largo del cromosoma 1, el grupo de los casos presenta un mayor nivel de ancestría nativa americana en comparación con el grupo de controles. En particular, se puede notar que en ambas versiones del modelo, la región simulada que presenta una mayor diferencia de ancestría corresponde al segmento 441, lo que sugiere que en esta región del cromosoma 1 existe un enriquecimiento de ancestría nativa americana de los casos sobre los controles. En los datos reales se observa un enriquecimiento de ancestría nativa americana de los casos sobre los controles correspondiente a 27.1 %. En los datos simulados mediante la versión no-informativa del modelo el enriquecimiento corresponde a 26.6 %. Por último, en los datos simulados mediante la versión informativa del modelo el enriquecimiento corresponde a 20.9 %.

Al comparar los resultados obtenidos mediante la implementación del modelo en sus versiones informativa y no-informativa, se observa que el modelo sí es sensible a los hiperparámetros de la distribución *a priori* de Dirichlet. Esto lo podemos observar en el

comportamiento de las distribuciones de probabilidad de nivel de ancestría (Figuras 4.8 y 4.9). La sensibilidad del modelo con respecto a los parámetros sugiere que en presencia de información real *a priori* podría mejorar el desempeño del modelo.

Es importante notar que en este caso los resultados obtenidos mediante la propuesta de la implementación del modelo de la presente tesis son consistentes con los obtenidos mediante un modelo frecuentista, demostrando que mediante dos acercamientos distintos del mismo problema, uno con un enfoque bayesiano y otro con un enfoque frecuentista, se puede obtener un mismo resultado.

Capítulo 5

Conclusión

En este trabajo se estudian conceptos básicos de genética y matemáticos necesarios para abordar un análisis de *Admixture Mapping* mediante un enfoque Bayesiano.

Se propone utilizar un modelo estadístico que implementa la simulación de distribuciones *a posteriori* de vectores de probabilidad inicial y de matrices de transición para cadenas de Markov no-homogéneas. Aunque los principios del modelo son conceptos simples, la elaboración de éste comprende una tarea compleja que exige implementar de forma adecuada los conocimientos adquiridos de las áreas matemática, genética y computación para obtener un buen resultado.

Los resultados obtenidos mediante la aplicación del modelo son consistentes con respecto a los resultados obtenidos mediante un modelo frecuentista, verificando que hay un enriquecimiento de ancestría nativa americana en la región determinada por la ventana 441 del cromosoma 1. Lo anterior muestra que mediante dos acercamientos distintos, uno con un enfoque Bayesiano y otro con un enfoque frecuentista, se puede obtener un mismo resultado.

Se pretende continuar con el trabajo para poder generalizar la aplicación del modelo y poder incorporar otro tipo de información inicial, así como lograr mejorar el desempeño del modelo en los aspectos que sean necesarios. Después de haber esto, se puede aspirar a proponer un modelo para la asignación de la ancestría local a poblaciones con mezcla genética.

En un panorama global, realizar este proyecto de tesis me permitió un primer acercamiento al trabajo de investigación, siendo ésta mi primer experiencia trabajando en matemáticas aplicadas. Como consecuencia, adquirí conocimiento nuevo sobre las áreas involucradas en el desarrollo de la tesis, lo que me brindó un mejor panorama sobre la necesidad y la importancia del trabajo multidisciplinario para abordar problemas reales.

Código fuente

Simulación

```
import pandas as pd
import numpy as np
import scipy.stats as sps

#El código es para la simulación de datos con los modelos informativo ó no informativo
#Archivo inicial
f = pd.read_csv('infile.txt', header = None, sep = ' ')

#Asignación de 'cantidad de información nativa'
b = []
f_t = f.transpose()
for i in range(0,f_t.shape[1], 2):
    a = zip(f_t[i], f_t[i+1])
    l = []
    for j in range(f_t.shape[0]):
        if (a[j][0] == 0) & (a[j][-1] == 0):
            l.append(2)
        elif (a[j][0] != 0) & (a[j][-1] != 0):
            l.append(0)
        else:
            l.append(1)
    b.append(l)
b = pd.DataFrame(b)

#Frecuencias estados iniciales
l_i = [[list(b[0]).count(0), list(b[0]).count(1), list(b[0]).count(2)]]

#Frecuencias de transición
```

```
29 L0 = []
30 L1 = []
31 L2 = []
32 for i in range(0, b.shape[1]-1):
33     a = zip(b[i], b[i+1])
34     L0.append([a.count((0,0)), a.count((0,1)), a.count((0,2))])
35     L1.append([a.count((1,0)), a.count((1,1)), a.count((1,2))])
36     L2.append([a.count((2,0)), a.count((2,1)), a.count((2,2))])
37 
38 #Parámetros a priori no informativa
39 al_i = [[1/3. for i in range(len(l_i[j]))] for j in range(len(l_i))]
40 al_0 = [[1/3. for i in range(len(L0[j]))] for j in range(len(L0))]
41 al_1 = [[1/3. for i in range(len(L1[j]))] for j in range(len(L1))]
42 al_2 = [[1/3. for i in range(len(L2[j]))] for j in range(len(L2))]
43 
44 #Parámetros a priori informativa
45 #Función de distancia para asignar valores de parametros
46 def alfa_distance(v):
47     val = [2,5]
48     d_ar = np.linspace(min(v), max(v), max(val))
49     l =[]
50     for i in v:
51         if i == min(v):
52             l.append(min(val))
53         else:
54             for j in range(len(d_ar)):
55                 if d_ar[j] < i <= d_ar[j+1]:
56                     break
57             l.append(min(val)+j)
58     return l
59 #Parametros
60 al_i = (map(alfa_distance, l_i))
61 al_0 = (map(alfa_distance, L0))
62 al_1 = (map(alfa_distance, L1))
63 al_2 = (map(alfa_distance, L2))
64 
65 #Dependiendo de la simulación que se quiera hacer se elige el tipo de parámetros a priori
66 #Debe elegirse únicamente un tipo de parámetros: informativo ó no-informativo
67 
68 #Parámetros a posteriori (no informativa ó informativa)
69 def add_(x,y): return [(sum(z)) for z in zip(x,y)]
70 Am_i = (map(add_, l_i, al_i))
```

```
Am_0 = (map(add_, L0, al_0))
Am_1 = (map(add_, L1, al_1))
Am_2 = (map(add_, L2, al_2))

#Función para muestreo aleatorio de distribución a porsteriori
def dir_prob(L): return[sps.dirichlet.rvs(L[i])[0].tolist() for i in range(len(L))]

#Muestreo aleatorio de distribución a porsteriori
n = 10000
qi = [sps.dirichlet.rvs(Am_i[i], n).tolist() for i in range(len(Am_i))][0]
q0 = [sps.dirichlet.rvs(Am_0[i], n).tolist() for i in range(len(Am_0))]
q0.insert(0,0)
q1 = [sps.dirichlet.rvs(Am_1[i], n).tolist() for i in range(len(Am_1))]
q1.insert(0,0)
q2 = [sps.dirichlet.rvs(Am_2[i], n).tolist() for i in range(len(Am_2))]
q2.insert(0,0)

P = []
for j in range(10000):
    l = []
    for i in range(783):
        if i == 0:
            l.append(qi[j])
        else:
            l.append([sum([q0[i][j][0]*(l[i-1][0]), q1[i][j][0]*(l[i-1][1]),
                           q2[i][j][0]*(l[i-1][2])]),
                      sum([q0[i][j][1]*(l[i-1][0]), q1[i][j][1]*(l[i-1][1]),
                           q2[i][j][1]*(l[i-1][2])]),
                      sum([q0[i][j][2]*(l[i-1][0]), q1[i][j][2]*(l[i-1][1]),
                           q2[i][j][2]*(l[i-1][2])])])
    P.append(l)

#Probabilidad promedio de niveles de ancestría en cada ventana
prob = [list(pd.DataFrame([P[i][j] for i in range(10000)]).mean()) for j in range(b.shape[1])]
pd.DataFrame(prob).to_csv('PROB.txt', header = None, index = False, sep = ' ')

#Esperanza por tiempo (promedios)
M = []
for i in range(10000):
    l = []
    for j in range(b.shape[1]):
        l.append(sum([0*P[i][j][0], 1*P[i][j][1], 2*P[i][j][2]]))
```

```
    M.append(l)
M83 = pd.DataFrame(M)
M83.to_csv('ESP.txt', header = None, index = False, sep = ' ')
```

Gráficas

```
import pandas as pd
import numpy as np
import matplotlib as mtl
import matplotlib.pyplot as plt
mtl.style.use('seaborn-whitegrid')


#Esperanzas por tiempo de los modelos informativo y no-informativo
#Casos
M83_I = pd.read_csv('ESP_CAS_I.txt', header = None, sep = ' ')
M83_NI = pd.read_csv('ESP_CAS_NI.txt', header = None, sep = ' ')
#Controles
M97_I = pd.read_csv('ESP_CON_I.txt', header = None, sep = ' ')
M97_NI = pd.read_csv('ESP_CON_NI.txt', header = None, sep = ' ')

#Las siguientes gráficas se obtienen de la misma forma para los modelos
#informativo y no-informativo
#Se presenta el caso del modelo no-informativo

#Calculo de intervalos de probabilidad
X = range(783)
Y1 = M97_NI.quantile(q = .025)
Y2 = M97_NI.quantile(q = .5)
Y3 = M97_NI.quantile(q = 0.975)
Y4 = M83_NI.quantile(q = 0.025)
Y5 = M83_NI.quantile(q = 0.5)
Y6 = M83_NI.quantile(q = 0.975)

#Gráfica de intervalos de probabilidad
#Promedio de ancestría
#Cromosoma 1 (783 ventanas)
#Casos
p4=plt.plot(X, Y4, 'C1:', label='_nolegend_')
p5=plt.plot(X, Y5, 'C1', label = 'Casos')
```

```
35 p6=plt.plot(X, Y6, 'C1:', label='_nolegend_')
36 plt.fill_between(X, Y4, Y6, color = 'C1', alpha = 0.3)
37 #Controles
38 p1=plt.plot(X, Y1, 'C0:', label='_nolegend_')
39 p2=plt.plot(X, Y2, 'C0', label = 'Controles')
40 p3=plt.plot(X, Y3, 'C0:', label='_nolegend_')
41 plt.fill_between(X, Y1, Y3, color = 'C0', alpha = 0.3)
42 #plt.title('Casos vs Controles\n$A-priori$ no informativa')
43 plt.xlabel('Cromosoma 1')
44 plt.ylabel(u'Promedio de nivel de ancestría')
45 plt.legend()
46 plt.grid(linestyle= ':')
47 plt.savefig('PC_NI.pdf', format='pdf')
48
49 #Gráfica de percentiles 0.025 con 0.975
50 q1 = Y4-Y3
51 plt.plot(q1, color = 'C2', marker = 'o', alpha = .4, label='Diferencia')
52 plt.axhline(y = 0, color = 'black', linestyle='--')
53 plt.grid(linestyle=':')
54 plt.xlabel('Cromosoma 1')
55 plt.ylabel(u'Diferencia de percentil 0.025 casos\ncon percentil 0.975 contoles')
56 plt.legend()
57 plt.savefig('N_INTERSEC_NI.pdf', format='pdf')
58
59 #Gráfica de distribución del valor de nivel de ancestría en la ventana 441
60 y1 = M83_NI[441]
61 y2 = M97_NI[441]
62 p2 = plt.hist(y2,50, label = 'Controles', COLOR = 'C0',alpha = .8)
63 p1 = plt.hist(y1,50, label = 'Casos', COLOR = 'C1',alpha = .8)
64 plt.axvline(x = 1.5421686746987953, color ='black',linestyle='--')
65 plt.axvline(x = 0.9896907216494846, color ='black', linestyle='--')
66 plt.xlabel(u'Promedio de nivel de ancestría')
67 plt.ylabel('Frecuencia')
68 plt.legend()
69 plt.grid(linestyle = ':')
70 plt.savefig('H441_NI.pdf', format='pdf')
71 ###
72
73 #comparacion casos informativa casos no-informativa ventana 441
74 y1 = M83_I[441]
75 y2 = M83_NI[441]
76 p1 = plt.hist(y1,50, label = 'Casos_I', COLOR = 'orange',alpha = .8)
```

```
77 p2 = plt.hist(y2,50, label = 'Casos_NI', COLOR = 'C1',alpha = .8)
78 plt.axvline(x = 1.5421686746987953, color ='black',linestyle='--')
79 plt.xlabel(u'Promedio de nivel de ancestría')
80 plt.ylabel('Frecuencia')
81 plt.legend()
82 plt.grid(linestyle = ':')
83 plt.savefig('H441_CAS.pdf', format='pdf')
84
85 #comparacion controles informativa casos no-informativa ventana 441
86 y1 = M97_I[441]
87 y2 = M97_NI[441]
88 p1 = plt.hist(y1,50, label = 'Controles_I', COLOR = 'blue',alpha = .5)
89 p2 = plt.hist(y2,50, label = 'Controles_NI', COLOR = 'C0',alpha = .8)
90 plt.axvline(x = 0.9896907216494846, color ='black', linestyle='--')
91 plt.xlabel(u'Promedio de nivel de ancestría')
92 plt.ylabel('Frecuencia')
93 plt.legend()
94 plt.grid(linestyle = ':')
95 plt.savefig('H441_CON.pdf', format='pdf')
```

Glosario

Ácido desoxirribonucléico (ADN): Una macromolécula que usualmente esta compuesta de polímeros nucleótidos y azúcar desoxirribosa que comprenden cadenas antiparalelas. El principal portador de la información genética.

Índice de fijación: Medida de diferenciación entre poblaciones debido a su estructura genética.

Alelo: Una de las posibles formas alternativas de un gen, normalmente un alelo se distingue de otros por sus efectos fenotípicos.

Célula diploide: Célula que posee un doble juego de cromosomas.

Célula haploide: Célula que posee un único juego de cromosomas.

Crómatidas no hermanas: Dos hebras distintas donde cada una proviene de la pareja de cromosomas homólogos duplicados en cuestión.

Cromátida: Cada una de las dos hebras hermanas de un cromosoma en división, que darán lugar a un cromosoma completo en cada célula fija.

Cromátidas hermanas: Las dos hebras de un cromosoma duplicado.

Deriva genética Mecanismo de la evolución en el que las frecuencias alélicas de una población cambian a lo largo de varias generaciones debido al azar (error de muestreo).

Desequilibrio de ligamiento: Asociación no aleatoria de alelos en diferentes loci en una población determi-

nada.

Desequilibrio de ligamiento: Propiedad de algunos genes de no segregar de forma independiente, es decir su fracción de recombinación es menor a 50 %.

Desmielizante: Proceso patológico que daña la capa de mielina de las fibras nerviosas.

Desoxirribosa: Aldopentosa derivada de la ribosa, que participa en la estructura de los ácidos desoxirribonucleicos.

Dextrógira: Que gira en el mismo sentido de las agujas del reloj.

Entrecruzamiento: El proceso de intercambio de secciones iguales de ADN entre cromosomas homólogos.

Enzima de restricción: Es una enzima capaz de reconocer secuencias particulares de nucleótidos dentro de una molécula de ADN y cortar la cadena en lugares específicos de la cadena de ADN llamados sitios o dianas de restricción.

Estratificación poblacional: Presencia de una diferenciación sistemática en las frecuencias alélicas entre subpoblaciones de una población.

Etnicidad: Carácter distintivo de una etnia.

Fenotipo: Manifestación variable del genotipo de un organismo en un determinado ambiente.

Flujo genético: Es el intercambio de genes entre dos poblaciones, debido a la dispersión de gametos como resultado de la migración de individuos entre poblaciones.

Fracción de recombinación: La probabilidad de que ocurra el fenómeno de recombinación entre dos loci.

Gameto: Célula haploide reproductiva especializada.

Gen: Unidad física fundamental de la herencia.

Genoma: El conjunto de información hereditaria codificada en el ADN de un organismo.

Genotipificación: Proceso de caracterización de genotipos en organismos.

Genotipo: Constitución alélica o genética de un organismo; normalmente se hace referencia a la composición alélica de un número limitado de genes.

Haplotipo: Conjunto de alelos con loci ligados estrechamente de un individuo que se heredan como una unidad.

Heterozigóto: Organismo con distintos alelos en uno o varios loci.

Homocigóto: Organismo con alelos idénticos para un gen o genes de interés. Estos individuos producen gametos idénticos con respecto al gen o genes en cuestión.

Kilobase (kb): Equivalente a mil pares de pares bases.

Loci: Plural de locus.

Locus: Lugar en un cromosoma donde se encuentra un gen en particular.

Marcador genético: Segmento de ADN con locus identificable en un cromosoma tal que su herencia se puede rastrear.

Marcador genético: Segmento de ADN con una ubicación física conocida en un cromosoma. Los marcadores genéticos pueden ayudar a vincular una enfermedad hereditaria con el gen responsable.

Megabase (Mb): Equivalente a un millón de pares de bases.

Meiosis: Proceso de división celular que tiene como objetivo la producción de células sexuales.

Microarreglos de ADN: Arreglo ordenado de secuencias de ADN o de oligonucleótidos en un substrato (o en cristal),

Mutación: Alteración en la secuencia de ADN de un organismo, su presencia es menor al 1 por ciento en la población.

Nucleósido: Una base de purina o pirimidina unida a una molécula de

azúcar ribosa o desoxiribosa.

Nucleótido: Un nucleosido unido a uno o más grupos fosfato.

Nucleobases (bases nitrogenadas): Compuestos orgánicos cíclicos que incluyen dos o más átomos de nitrógeno.

Oligonucleótido: Secuencia de 10 a 20 nucelotidos

Par de bases (bp): Unidad de medida que consta de dos nucleobases unidas entre sí por enlaces de hidrógeno.

Penetrancia genética: La frecuencia, expresada como porcentaje, con la cual los individuos de un genotipo dado manifiestan al menos cierto grado de un fenotipo mutante especifico o asociado a un rasgo.

Polimorfismo: Variación en la secuencia de ADN en un locus. Presente en al menos 1 por ciento de la población.

Polimorfismo de un dolo nucleótido (SNP): Variación en un único par de bases en la secuencia de ADN. Presente en al menos 1 por ciento de la población.

Prevalencia: La proporción de personas que sufren de estas enfermedades con respecto a la totalidad de la población en estudio.

Recombinación: Proceso de redistribución de los genes en la descendencia, que presenta en consecuencia caracteres distintos a los de sus progenitores.

Segregación: La separación de las parejas de cromosomas homólogos paternos y maternos en gametos durante la meiosis.

Segregación: La separación de los cromosomas homologos en gametos durante la meiosis.

Bibliografía

[1] J. Ott, *Analysis of human genetic linkage*. JHU Press, 1999.

[2] D. J. Balding, M. Bishop, and C. Cannings, *Handbook of statistical genetics*. John Wiley & Sons, 2008.

[3] P. Casares, "A corrected haldane's map function to calculate genetic distances from recombination data," *Genetica*, vol. 129, no. 3, pp. 333–338, 2007.

[4] E. Flores-Alfaro, A. I. Burguete-García, and E. Salazar-Martínez, "Diseños de investigación en epidemiología genética," 2012.

[5] D. Shriner, "Overview of admixture mapping," *Current protocols in human genetics*, pp. 1–23, 2013.

[6] N. Patterson, N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, S. L. Hauser, M. W. Smith, S. J. O'Brien, D. Altshuler, *et al.*, "Methods for high-density admixture mapping of disease genes," *The American Journal of Human Genetics*, vol. 74, no. 5, pp. 979–1000, 2004.

[7] G. Montana and C. Hoggart, "Statistical software for gene mapping by admixture linkage disequilibrium," *Briefings in bioinformatics*, vol. 8, no. 6, pp. 393–395, 2007.

[8] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[9] D. H. Alexander, J. Novembre, and K. Lange, "Fast model-based estimation of ancestry in unrelated individuals," *Genome research*, vol. 19, no. 9, pp. 1655–1664, 2009.

[10] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers, "Sensitive detection of chromosomal segments of distinct ancestry in admixed populations," *PLoS genetics*, vol. 5, no. 6, p. e1000519, 2009.

[11] Y. Baran, B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux, C. Eng, W. Rodriguez-Cintron, R. Chapela, J. G. Ford, P. C. Avila, *et al.*, "Fast and accurate inference of local ancestry in latino populations," *Bioinformatics*, vol. 28, no. 10, pp. 1359–1367, 2012.

[12] A. Brisbin, K. Bryc, J. Byrnes, F. Zakharia, L. Omberg, J. Degenhardt, A. Reynolds, H. Ostrer, J. G. Mezey, and C. D. Bustamante, "Pcadmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations," *Human biology*, vol. 84, no. 4, pp. 343–364, 2012.

[13] S. Karlin, *A first course in stochastic processes*. Academic press, 2014.

[14] G. Grimmett and D. Stirzaker, *Probability and random processes*. Oxford university press, 2001.

[15] S. M. Ross, *Stochastic processes. 1996*. Wiley, New York, 1996.

[16] D. Gamerman and H. F. Lopes, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.

[17] Anderson, Theodore W y Goodman, Leo A, "Statistical inference about markov chains," *The Annals of Mathematical Statistics*, pp. 89–110, 1957.

[18] T. R. Fleming and D. P. Harrington, "Estimation for discrete time nonhomogeneous markov chains," *Stochastic processes and their applications*, vol. 7, no. 2, pp. 131–139, 1978.

[19] E. R. Rodrigues, M. H. Tarumoto, and G. Tzintzun, "A non-homogeneous markov chain model to study ozone exceedances in mexico city," in *Current Air Quality Issues*, InTech, 2015.

[20] L. M. B. Zamora, "Las cadenas de markov con aplicaciones a problemas de contaminación atmosférica," 2016.

[21] S. Jarius, K. Ruprecht, B. Wildemann, T. Kuempfel, M. Ringelstein, C. Geis, I. Kleiter, C. Kleinschnitz, A. Berthele, J. Brettschneider, *et al.*, "Contrasting disease patterns in seropositive and seronegative neuromyelitis optica: a multicentre study of 175 patients," *Journal of neuroinflammation*, vol. 9, no. 1, p. 14, 2012.

[22] D. M. Wingerchuk, V. A. Lennon, C. F. Lucchinetti, S. J. Pittock, and B. G. Weinshenker, "The spectrum of neuromyelitis optica," *The Lancet Neurology*, vol. 6, no. 9, pp. 805–815, 2007.

[23] J. F. Rivera, J. F. Kurtzke, V. A. Booth, and T. Corona, "Characteristics of devic's disease (neuromyelitis optica) in mexico," *Journal of neurology*, vol. 255, no. 5, pp. 710–715, 2008.

[24] M. Matiello, H. Kim, W. Kim, D. Brum, A. Barreira, D. Kingsbury, G. Plant, T. Adoni, and B. G. Weinshenker, "Familial neuromyelitis optica," *Neurology*, vol. 75, no. 4, pp. 310–315, 2010.