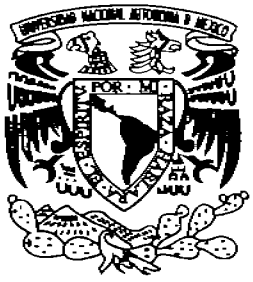

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS
PROGRAMA DE DOCTORADO EN CIENCIAS BIOMÉDICAS

"ANÁLISIS DE LAS HERRAMIENTAS METODOLÓGICAS PARA EVIDENCIAR LIGAMIENTO EN ENFERMEDADES GENÉTICAS COMÚNES"

T E S I S

QUE PARA OBTENER EL GRADO DE:
DOCTORA EN CIENCIAS BIOMÉDICAS

P R E S E N T A:
M. EN MAT. SANDRA ROMERO HIDALGO

DIRECTOR DE TESIS:
DRA. MARÍA TERESA TUSIÉ LUNA

MÉXICO, D.F. 2006

Doctorado en Ciencias Biomédicas

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

pdcb/grad/239**Jur**/2005.

ING. LEOPOLDO SILVA GUTIERREZ
Director General de la
Administración Escolar
P r e s e n t e .

Por medio de la presente me permito informar a usted que en la reunión del Comité Académico del Programa de Doctorado en Ciencias Biomédicas que tuvo lugar el 21 de septiembre de 2005, se acordó designar el siguiente jurado para examen de Doctorado en Ciencias Biomédicas de la M. en Mat. **SANDRA ROMERO HIDALGO** con no. de cuenta 91505518 con la tesis titulada: "Análisis de las herramientas metodológicas para evidenciar ligamiento en enfermedades genéticas comunes", dirigida por la Dra. Ma. Teresa Tusié Luna.

Presidente:	Dr. Ignacio Méndez Ramírez
Secretario:	Dra. Ma. Teresa Tusié Luna
Vocal:	Dr. Pedro Julio Collado Vides
Vocal:	Dr. Ruben Lisker Yourkowitzky
Vocal:	Dr. Pedro Miramontes Vidal
Suplente:	Dra. Ma. Eugenia Gonsebatt Bonaparte
Suplente:	Dra. Eliane Regina Rodríguez Caloni

Sin otro particular por el momento, aprovecho la ocasión para enviarle un cordial saludo.

A t e n t a m e n t e
"Por mi raza hablará el espíritu"
CD. Universitaria, D.F., 15 de diciembre de 2005.

Dr. Edmundo Lamoyi Velázquez
Responsable en el Instituto de
Investigaciones Biomédicas

Dr. J. Javier Espinosa Aguirre
Coordinador del Doctorado
en Ciencias Biomédicas

c.c.p. - Secretaría de Asuntos Escolares

A Franz Emilian

Índice general

Índice de figuras

Índice de cuadros

Prefacio

Este trabajo se desarrolló bajo la dirección de la Dra. María Teresa Tusié Luna (Insituto de Investigaciones Biomédicas) y las cotutorías del Dr. Eduardo Gutiérrez Peña (Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas), de la Dra. Eliane Rodrigues (Instituto de Matemáticas) y de la Dra. Socorro Durán Vargas (Insituto de Investigaciones Biomédicas). Parte de este trabajo se desarrolló en el Insituto de Matemáticas, Unidad Cuernavaca, con la colaboración del Dr. Luis Javier Álvarez Noguera.

Agradezco de manera muy especial a la Dra. María Teresa Tusié por el apoyo recibido para la realización de esta tesis; al Dr. Eduardo Gutiérrez y a la Dra. Eliane Rodrigues por su estrecha y valiosa participación; al Dr. Luis Javier Álvarez por su disposición y tiempo; y a Laura Riba y Samuel Canizales por sus consejos y por proporcionarme los datos utilizados en esta tesis.

Agradezco profundamente a mi familia y amigos por el apoyo moral que siempre me brindaron; particularmente, agradezco a Franz por sus consejos, amor y paciencia.

Finalmente, quiero agradecer al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico que me otorgaron.

Sandra Romero Hidalgo

México D.F., enero 2006

Resumen

El análisis de ligamiento genético tiene como objetivo inferir la posición del gen o genes responsables de una enfermedad relativa a uno o varios marcadores genéticos. Cuando se dispone de familias multigeneracionales y los individuos pueden ser clasificados sin ambigedad como sanos o afectados, la técnica estadística más popular es el método de lod score. Utilizando diferentes algoritmos, este método ha sido implementado en diversos programas de cómputo de dominio público. Dos de los programas más utilizados son GENEHUNTER y SimWalk2. El primero esta limitado a un número reducido de individuos. Una solución es partir a la familia en familias más pequeñas. El segundo no tiene restricción con respecto al número de individuos pero se basa en métodos de simulación, y por lo tanto, proporciona un resultado aproximado. En este trabajo se utilizaron datos simulados, para comparar los resultados de ambos programas; investigar el efecto sobre el valor de lod

score de descartar individuos y dividir a la familia; y evaluar el desempeño de cada uno de ellos ante información faltante. Los resultados muestran que ambos programas producen resultados muy similares; que descartar individuos y dividir a la familia tienen un efecto adverso; y que ambos programas tienen un comportamiento similar ante información faltante. Adicionalmente, se desarrolló e implementó un programa de cómputo que realiza un análisis Bayesiano de ligamiento genético utilizando métodos de Monte Carlo vía cadenas de Markov, el cual considera a todos los parámetros del modelo como desconocidos. Con el fin de evaluar el desempeño del programa se utilizaron tres ejemplos correspondientes a tres distintos escenarios. Los resultados muestran la utilidad del método.

Abstract

Linkage analysis aims to identify genes responsible for certain inherited diseases. When extended multigenerational families are available and the family members are unambiguously clinically characterized, as affected or unaffected, one of the most popular statistical technique is the lod score method. Using different algorithms, this method has been implemented in several public domain computer programs. Two of the most commonly used programs are GENEHUNTER y SimWalk2. The former is limited to a reduced number of individuals. One solution is to split the pedigree into smaller ones. The latter has no limit in the number of individuals but it works based on simulation methods, therefore its results are only approximates. In this study simulated data was used to compare the results of both programs; to evaluate the effect of discarding individuals and splitting the pedigree on the lod score value; and to assess how missing data affect the performance of ea-

ch program. The results show that both programs produce nearly the same results; that either discarding individuals or splitting the pedigree have an adverse effect; and the performance of both programs is similar in the missing data scenario. Additionally, a computer program was developed that does a Bayesian genetic linkage analysis using Markov Chain Monte Carlo methods. The program includes all the parameters of the model as unknown. To measure the performance of the program three examples were used corresponding to three different scenarios. The results show the benefit of the method.

Capítulo 1

Introducción

Muchas de las enfermedades llamadas comunes como son las enfermedades cardiovasculares, hipertensión arterial, y diabetes, entre otras, representan un problema de salud con grandes repercusiones socioeconómicas en México, y en los tres casos estudios previos evidencian la participación de factores genéticos.

Cuando una enfermedad tiene un componente genético involucrado, éste puede ser el resultado de la alteración en un solo gen (enfermedades mendelianas), en un número pequeño de genes (enfermedades oligogénicas) o en un número mayor de genes con un efecto menor de cada uno (enfermedades poligénicas). Estas dos últimas corresponden a las llamadas enfermedades

complejas.

Cuando se trata de determinar si una enfermedad es causada por un gen o un grupo de genes y, aún más, localizar su posición en el genoma, la estadística juega un papel fundamental y con este fin distintos métodos han sido desarrollados.

Particularmente, el análisis de ligamiento genético tiene como objetivo identificar la posición del gen o genes responsables de un rasgo o enfermedad relativa a uno o varios marcadores genéticos. Cuando hay disponibilidad de familias multigeneracionales, cuyos miembros pueden ser caracterizados clínicamente sin ambigüedad, como sanos o afectados, una de las herramientas estadísticas más conocidas es el método de lod (*logarithm of odds*) score. El fundamento teórico de este método descansa sobre el concepto biológico conocido como recombinación.

Utilizando diferentes algoritmos, el método de lod score ha sido implementado en diversos programas de cómputo de dominio público. A la fecha hay tres algoritmos disponibles; estos son: el algoritmo Elston-Stewart, el algoritmo Lander-Green y los métodos de Monte Carlo vía cadenas de Markov. Cada uno de estos algoritmos tiene ventajas y limitaciones. El primero está limitado a un número reducido de marcadores, el segundo a un núme-

ro reducido de individuos, y el tercero, a pesar de que no tiene limitación en cuanto al número de marcadores o individuos, es una estrategia basada en métodos de simulación y por esta razón el resultado que proporciona es aproximado. Las limitaciones propias de cada uno de los programas pueden ocasionar resultados incongruentes cuando se utiliza una misma base de datos. En los últimos años, se han reportado un número considerable de regiones cromosómicas presuntamente ligadas a diferentes rasgos en distintas poblaciones, de las cuales solo unas pocas han sido replicadas y confirmada su participación. En la medida que se tenga mayor conocimiento del efecto que tienen diferentes situaciones sobre el valor de lod score, se tendrá mayor confianza al reportar un hallazgo.

Por otro lado, todos los programas disponibles utilizan un enfoque estadístico clásico o frecuentista donde el parámetro de interés es la fracción de recombinación, el cual representa la distancia genética entre el gen de la enfermedad y un marcador genético. Sin embargo, en un análisis de ligamiento genético hay otros elementos involucrados. Por ejemplo, en el método de lod score implementado en los programas disponibles es necesario especificar de antemano parámetros como la penetrancia y las frecuencias alélicas de los marcadores en la población de estudio. En muchos casos los valores de estos

parámetros son desconocidos y por lo mismo se especifican de manera arbitraria. Utilizando un enfoque estadístico alternativo conocido como enfoque Bayesiano es posible incorporar de manera natural todos estos componentes como parámetros desconocidos con el fin de hacer inferencias sobre ellos también.

El contenido principal de esta tesis está concentrado en cuatro capítulos. Los capítulo 2 y 3 comprenden una introducción de los conceptos, desde el punto de vista biológico y estadístico, respectivamente, sobre los cuales descansa el análisis de ligamiento genético. Particularmente, el capítulo 3 describe brevemente el método de lod score, así como los tres algoritmos principales mencionados previamente, en conjunto con los problemas potenciales que conlleva cada uno de ellos. Por ejemplo, cuando se utilizan métodos de Monte Carlo vía cadenas de Markov en el contexto de análisis de ligamiento genético, un problema potencial es que no siempre es posible garantizar que la cadena de Markov es irreducible y en referencia a esto muchos autores han propuesto diferentes soluciones. Por último, en este capítulo también se incluye una breve descripción del mecanismo que sigue el enfoque estadístico frecuentista en contraste con el que sigue el enfoque estadístico Bayesiano para hacer inferencia sobre un parámetro de interés.

En el capítulo 4 se realiza, utilizando datos simulados, una comparación de dos de los programas más populares utilizados para realizar un análisis de ligamiento a través del método de lod score; estos son GENEHUNTER y SimWalk2. Además de comparar sus resultados, también se estudia el efecto, sobre el valor de lod score, de descatar individuos y de partir a la familia en familias más pequeñas, así como el comportamiento de cada uno de los programas ante la presencia de información faltante.

Por último, en el capítulo 5 se utiliza el enfoque Bayesiano para realizar un análisis de ligamiento genético, el cual es implementado en un programa de cómputo y aplicado a tres ejemplos prácticos. El primero con evidencia fuerte de ligamiento, el segundo con evidencia sugestiva de ligamiento y el tercero con evidencia en contra de ligamiento.

Capítulo 2

Antecedentes genéticos

2.1. Terminología

El ácido desoxirribonucleico (ADN) humano consiste de tres mil millones de pares de bases y está empaquetado en cromosomas. El núcleo de cada una de las células de un individuo contiene 23 pares de cromosomas homólogos, 22 de los cuales se les denomina cromosomas autosomales y un par de cromosomas sexuales. De cada par de cromosomas, uno proviene de la madre del individuo y el otro proviene del padre del individuo.

A lo largo de los cromosomas se encuentran localizados los genes. Los genes son las unidades básicas que contienen la información hereditaria. La

posición que ocupa un gen dentro de un cromosoma se le conoce como *locus* (plural *loci*). Los genes puede presentar variaciones o alelos. En un locus determinado cada individuo tiene dos alelos (uno en cada cromosoma), este par de alelos constituye el genotipo de un individuo. Si el genotipo de un individuo consta de dos alelos iguales entre sí entonces se dice que el individuo es homocigoto, si son alelos distintos entonces se dice que el individuo es heterocigoto.

La expresión o característica observable de un genotipo se define como fenotipo. Dependiendo de la expresión del alelo en el fenotipo, los alelos pueden ser dominantes, recesivos o codominantes.

En un gen dado, la relación entre el genotipo y fenotipo se describe a través de lo que se conoce como penetrancia. La penetrancia es la probabilidad condicional de observar un fenotipo dado un genotipo; en los casos más simples la penetrancia es 0 ó 1. Sin embargo, en una gran variedad de enfermedades se observan valores intermedios de penetrancia, es decir, penetrancia incompleta.

Como ejemplo, considere el gen que determina el grupo sanguíneo (gen ABO) de los individuos, el cual se encuentra localizado en el brazo largo del cromosoma 9. Este gen tiene tres alelos, A, B y O, seis genotipos AA, AB,

	Genotipo					
Fenotipo	AA	AB	AO	BB	BO	OO
A	1	0	1	0	0	0
B	0	0	0	1	1	0
AB	0	1	0	0	0	0
O	0	0	0	0	0	1

Cuadro 2.1: *Penetrancia gen ABO*

AO, BB, BO y OO, y cuatro fenotipos A, B, AB y O. En el Cuadro 2.1 se muestran los valores de penetrancia correspondientes a este gen, estos valores indican que tanto el alelo A como el alelo B son dominantes sobre el alelo O, o lo que es lo mismo, el alelo O es recesivo con respecto al alelo A y al alelo B. Por otro lado, los alelos A y B son codominantes.

A las frecuencias relativas en la población de los diferentes alelos de un locus se les conoce como frecuencias alélicas, y se dice que un locus es polimórfico si su alelo más común en la población tiene una frecuencia relativa menor al 99 %.

Si suponemos que la población es grande, que se reproduce aleatoriamente, y que no existen fuerzas selectivas tales como mezcla entre poblaciones,

mutaciones o selección a favor de una característica genética particular, entonces esta población se dice que está en equilibrio de Hardy-Weinberg. Esto quiere decir que las frecuencias relativas de los genotipos en la población dependen únicamente de las frecuencias alélicas. Como ejemplo, considere un locus con alelos A y a, denotemos por p_A y p_a a la frecuencia relativa en la población de cada uno estos alelos. Si la población se encuentra en equilibrio de Hardy-Weinberg entonces la frecuencia esperada de los genotipos AA, Aa y aa es p_A^2, $2p_Ap_a$ y p_a^2, respectivamente.

Por otro lado, un concepto que se utiliza con mucha frecuencia a lo largo de todo el trabajo es lo que se conoce como marcador genético. Un marcador genético es un segmento (funcional o no-funcional) de ADN bien localizado y polimórfico de tal manera que es posible establecer su segregación entre miembros de una familia. Cabe destacar que un gen corresponde a un segmento funcional de ADN, por lo tanto, un gen puede servir como marcador genético. Por segmento funcional se entiende que contiene información para construir proteínas, la cuales desempeñan un papel estructural en el organismo.

Al conjunto de alelos, correspondientes a varios marcadores genéticos, que un individuo recibe de uno de sus padres se le denomina haplotipo. Por

ejemplo, considere dos loci bialélicos, el locus 1 con alelos A y a, y el locus 2 con alelos B y b, en este caso es posible formar cuatro posibles haplotipos que son AB, Ab, aB y ab. De manera general, para N loci donde n_i es el número de alelos correspondiente al *i-ésimo* locus, $i = 1, 2, \ldots, N$, el número total de posibles haplotipos está dado por el siguiente producto $n_1 \times n_2 \times \ldots \times n_N$.

La segregación de una característica de interés, tal como una enfermedad, entre miembros de la misma familia en ocasiones es interpretada como evidencia de un efecto genético. Este efecto genético puede ser el resultado de un solo gen (enfermedades monogénicas o Mendelianas), o bien puede ser el resultado de un conjunto mayor de genes con un efecto menor cada uno, posiblemente interactuando entre ellos y con el medio ambiente (enfermedades complejas).

Derivado de lo anterior, los árboles genealógicos son particularmente útiles para el estudio de las enfermedades monogénicas. Se dice que una enfermedad tiene un mecanismo de herencia autosómico dominante si el gen responsable está localizado en uno de los cromosomas autosomales y el fenotipo se manifiesta en individuos heterocigotos. En un árbol genealógico este mecanismo de herencia se reconoce cuando hay individuos afectados en diferentes generaciones. Por su parte, se dice que una enfermedad tiene un mecanismo

de herencia autosómico recesivo cuando el gen responsable está localizado nuevamente en uno de los cromosomas autosomales y el fenotipo se manifiesta únicamente en individuos homocigotos. En contraste con el caso anterior, en un árbol genealógico, este mecanismo se reconoce cuando se observan individuos afectados en una sola generación con progenitores son sanos. La excepción, de este último caso, ocurre en poblaciones aisladas donde existe una tasa alta de consanguinidad. Cabe destacar que en muchos casos la distinción entre diferentes mecanismos de herencia es ambigua.

Una estrategia estadística construida con el fin de conocer la ubicación del gen o de los genes responsables de una enfermedad es la que se conoce como análisis de ligamiento genético. Esta estrategia ha sido utilizada ampliamente y con gran exito en el estudio de enfermedades monogénicas.

2.2. Análisis de ligamiento génetico

El análisis de ligamiento genético tiene como objetivo inferir la posición del gen o de los genes responsables de una enfermedad relativa a uno o varios marcadores genéticos.

En las últimas décadas se han desarrollado varios métodos estadísticos

alrededor del mismo objetivo. Uno de los pioneros fue el método propuesto por Morton (1955), cuyo fundamento teórico descansa sobre el concepto biológico de entrecruzamiento.

Durante la meiosis, los cromosomas homólogos se aparean y realizan un proceso de intercambio de material genético denominado entrecruzamiento. En otras palabras, un entrecruzamiento ocurre cuando un padre transmite a su hijo un cromosoma que es una combinación de su cromosoma materno y paterno. Se asume que si dos loci están cerca uno del otro, la probabilidad de que ocurra un entrecruzamiento entre ellos es pequeña, y entre mayor sea la distancia entre ellos, mayor será la probabilidad de observar un entrecruzamiento. De esta forma, la distancia genética entre dos loci en un cromosoma, se define como el número esperado de entrecruzamientos que ocurre entre ellos. La unidad en la que se mide esta distancia se conoce como Morgan.

Por otro lado, cuando los entrecruzamientos no ocurren de manera independiente a lo largo del cromosoma, es decir, que la presencia de un entrecruzamiento afecta la probabilidad de que se presente otro entrecruzamiento en una región cercana se dice que hay interferencia genética.

No es posible identificar de manera directa el número de entrecruzamientos que ocurrieron entre dos loci, por lo que se introduce un concepto adicio-

nal llamado recombinación. Se dice que se observa una recombinación cuando ocurre un número impar de entrecruzamientos entre estos dos loci.

La probabilidad de observar una recombinación entre dos loci se conoce como fracción de recombinación y se denota con la letra griega θ. Para loci lejanos uno del otro se espera una proporción equivalente de individuos recombinantes y no recombinantes en una familia, es decir, $\theta = 1/2$. Sin embargo, cuando estos están cerca se espera que los alelos de los dos loci segreguen juntos de generación en generación. Este fenómeno se conoce como ligamiento genético; en este caso $\theta < 1/2$.

Con el fin de estimar la fracción de recombinación entre dos loci, Morton (1955) estableció, para cierto tipo de familias de dos generaciones, las bases iniciales de lo que hoy conocemos como el método de lod score.

Capítulo 3

Antecedentes estadísticos

3.1. Análisis de ligamiento génetico

3.1.1. Enfoque frecuentista

De manera general, el proceso que sigue la estadística clásica o frecuentista para hacer inferencia sobre un parámetro de interés se puede resumir en los siguientes puntos. Considere como ejemplo que el parámetro de interés es θ, que representa a la fracción de recombinación y por tanto un valor plausible de la distancia genética entre el locus de la enfermedad y un marcador genético.

1. Se desea poner a prueba una hipótesis sobre el posible valor del parámetro, por ejemplo, la hipótesis de no-ligamiento ($\theta = 1/2$).

2. Se construye un modelo de probabilidad para las observaciones dado el valor del parámetro, $p(x|\theta)$. El modelo se construye de tal forma que sea compatible con las características del fenómeno estudiado, en el sentido de que las observaciones deben satisfacer los supuestos del modelo.

3. Se construye la función de verosimilitud, $L(\theta)$, que corresponde a la función de densidad conjunta de los datos vista como función del parámetro de interés, y posiblemente de otros parámetros. Cuando las observaciones son independientes entonces $L(\theta) \propto \prod_{i=1}^{n} p(x_i|\theta)$. Sin embargo, en el caso de ligamiento genético tenemos que los individuos están relacionados entre sí, y por tanto las observaciones no son independientes. Debido a esto, la expresión de la función de verosimilitud es diferente (ver sección 3.2).

4. Un estimador puntual óptimo del parámetro será aquel que maximice el valor de la función de verosimilitud, de ahí su nombre de estimador de máxima verosimilitud.

5. La estadística de prueba es una medida que servirá como criterio para decidir si los datos disponibles apoyan o rechazan la hipótesis. Esto se hace cuantificando la desviación de lo que se observa a través de los datos respecto de lo que se espera según la hipótesis.

6. Se rechazará la hipótesis cuando el valor de la estadística de prueba sea grande, esto es, la desviación entre lo que se observa y lo que se espera es grande.

7. El valor de la estadística de prueba será grande si la probabilidad de que se presente una desviación como la observada o mayor es pequeña, por ejemplo 0.05 ó 0.01. A esta probabilidad se le conoce como nivel de significancia. En otras palabras, el nivel de significancia proporciona una medida de error asociada a la decisión de no rechazar la hipótesis cuando ésta es cierta.

En resumen, el enfoque frecuentista descansa únicamente en la evidencia proporcionada por los datos para hacer inferencia sobre el parámetro de interés.

Lo que conocemos hoy como método de lod score es la aplicación de un enfoque frecuentista, donde el parámetro sobre el que se desea hacer infe-

rencias es la fracción de recombinación. Este método asume que las observaciones provienen de modelos de probabilidad que dependen de la fracción de recombinación, θ, así como de parámetros tales como la penetrancia y las frecuencias alélicas de los marcadores en la población de estudio. En este tipo de análisis, la construcción de la función de verosimilitud requiere de la especificación del mecanismo de herencia de la enfermedad; por lo tanto, es aplicable principalmente a enfermedades monogénicas o Mendelianas. Por esta razón, a este método también se le conoce como análisis de ligamiento "paramétrico".

Siguiendo los pasos descritos previamente, se desea probar la hipótesis nula de no-ligamiento, i.e. $H_0 : \theta = 1/2$, contra la hipótesis alternativa de ligamiento, i.e. $H_1 : \theta < 1/2$. El lod score, denotado por $Z(\theta)$, corresponde a una versión modificada del cociente de verosimilitudes, donde $L(\theta)$ es la función de verosimilitud correspondiente a la familia y el parámetro de interés es la fracción de recombinación θ,

$$Z(\theta) = log_{10}\frac{L(\theta)}{L(1/2)}.$$

Para m familias independientes, el lod score global tendrá la siguiente forma,

$$Z(\theta) = \sum_{i=1}^{m} Z_i(\theta),$$

donde $Z_i(\theta)$ es el lod score para la *i-ésima* familia.

Es común rechazar H_0 cuando $Z_{max} \geq 3$ y $Z_{max} \geq 2$ para loci autosómicos y sexuales, respectivamente, donde $Z_{max} = max_\theta Z(\theta)$. Estos valores fueron propuestos originalmente por Morton (1955).

Con base en la teoría del cociente de verosimilitudes, bajo la hipótesis nula de no-ligamiento, la distribución aproximada de $4.6 \times Z_{max}$ es una ji-cuadrada con 1 grado de libertad, es decir $4.6 \times Z_{max} \sim \chi^2_1$, y por lo tanto el nivel de significancia asociado a los valores $Z_{max} \geq 3$ y $Z_{max} \geq 2$ es 0.0001 y 0.001, respectivamente, (Ott, 1999).

La descripción anterior corresponde a la versión actual del método de lod score. Esta difiere ligeramente del método propuesto originalmente por Morton (1955). El método sugerido por este autor consiste en recolectar familias secuencialmente mientras $log_{10}B < Z(\theta) < log_{10}A$, donde A y B son constantes que determinan las propiedades de la prueba. Para una prueba con nivel de significancia α y poder $1 - \beta$, Morton sugirió aproximar los valores de A y B a través de $\frac{1-\beta}{\alpha}$ y $\frac{\beta}{1-\alpha}$, respectivamente. Para $\alpha = 0.001$ y

$\beta = 0.99$, se tiene que $A \approx \frac{1}{\alpha}$ y $B \approx \beta$, por lo tanto $A \approx 1000$ y $B \approx 0.01$. De esta forma, la recolección de familias termina una vez que el valor de lod score es mayor o igual a 3 ($log_{10}1000$), o menor o igual a -2 ($log_{10}0.01$). En el primer caso se rechaza la hipótesis nula de no-ligamiento y en el segundo caso se acepta.

Cuando el tamaño de la familia incrementa, en conjunto con el número de marcadores genéticos, obtener la expresión de $L(\theta)$ y calcularla, se vuelve una tarea complicada (ver sección 3.2). La dificultad radica en que la fracción de recombinación especifica probabilidades de transmisión correspondientes a los genotipos de los individuos mientras que la información disponible de cada uno de ellos es su fenotipo.

Cabe destacar también que en el método de lod score el parámetro de interés es únicamente la fracción de recombinación. Otros parámetros como son la penetrancia y las frecuencias alélicas se asumen conocidos y es necesario especificarlos de antemano.

3.1.2. Enfoque Bayesiano

Por su parte, el enfoque Bayesiano se basa en una interpretación subjetiva de la probabilidad, es decir, como una medida de incertidumbre. En

vista de esto, considera al valor del parámetro θ como una variable aleatoria con función de densidad de probabilidad, $p(\theta)$. Esta última se conoce como distribución inicial. A través de la distribución inicial, el investigador puede incorporar conocimiento previo sobre el valor del parámetro, o en condiciones donde no se tiene ese conocimiento previo, se puede utilizar una distribución inicial que refleje la ausencia del mismo, por ejemplo una distribución uniforme. Una vez que la muestra está disponible, el objetivo es actualizar la distribución inicial con la información contenida en la misma, a través de la función de verosimilitud. Recuerde que esta función corresponde a la función de probabilidad conjunta de los datos, $p(x|\theta)$, vista como una función del parámetro de interés, y posiblemente de otros parámetros. De esta manera se obtiene la distribución actualizada del parámetro o distribución final, $p(\theta|x)$. Esto es posible gracias al Teorema de Bayes, que tiene la siguiente forma,

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

La expresión anterior se puede simplificar, debido a que $p(x)$ es una función que no depende del parámetro de interés, obteniendo

$$p(\theta|x) \propto p(x|\theta)p(\theta).$$

La probabilidad de la hipótesis de ligamiento (H_L) se evalúa de la si-

guiente manera,

$$P(H_L) = P(0 \leq \theta < 1/2) = \int_{0 \leq \theta < 1/2} p(\theta|x)\, d\theta.$$

Evidencia de ligamiento significativa bajo este contexto equivaldría a obtener una probabilidad alta para H_L, por ejemplo $P(H_L) = 0.95$.

Pasar del caso unidimensional, de un solo parámetro, al caso multidimensional donde hay dos o más parámetros, no requiere de nueva teoría. Suponga que ahora tenemos un vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ de parámetros sobre los que se desea hacer inferencias. En este caso es necesario especificar una distribución inicial conjunta, $p(\theta)$, que combinada con la función de verosimilitud, $p(x|\theta)$, vía el teorema de Bayes, da lugar a la distribución final conjunta, $p(\theta|x)$.

Si se desea hacer inferencias sobre un parámetro en particular, entonces utilizando la distribución final conjunta se obtiene la distribución final marginal de dicho parámetro, de la siguiente manera,

$$p(\theta_1|x) = \int_{\theta_2} \cdots \int_{\theta_k} f(\theta|x)\, d\theta_2 \ldots d\theta_k.$$

En resumen, el enfoque Bayesiano utiliza la combinación de información previa proporcionada por el investigador y la evidencia muestral. Smith (1959) fue el primero en utilizar un enfoque Bayesiano en el problema de

análisis de ligamiento genético. A partir de entonces varios autores han hecho contribuciones al respecto, incluyendo Thomas y Cortessis (1992), Hauser y Boehnke (1993) y Thomas et al. (1997). Una comparación de ambos enfoques dentro del contexto de análisis de ligamiento genético la proporciona Vieland (1998).

Particularmente, el trabajo de Thomas y Cortessis utiliza un enfoque Bayesiano para hacer un análsis de ligamiento de dos puntos, es decir, infiere la posición del locus de la enfermedad relativa a un solo marcador, que en este caso es bialélico. A diferencia del método de lod score, la propuesta de estos autores incluye no solo la fracción de recombinación como parámetro desconocido sino también parámetros tales como la penetrancia, las frecuencias alélicas y los genotipos de los individuos. Específicamente, utilizan métodos de Monte Carlo vía cadenas de Markov, particularmente el muestreo de Gibbs. Una introducción a estas técnicas se presenta más adelante en este mismo capítulo. Una descripción detallada del esquema propuesto por estos autores se puede encontrar en el capítulo 5, donde se extiende este algoritmo con el fin de permitir un marcador multialélico.

3.2. Función de verosimilitud

3.2.1. Definición

Denotemos como $x = \{x_1, x_2, \ldots, x_n\}$ al conjunto de fenotipos de n individuos en una familia. La función de verosimilitud, $L(\theta)$, corresponde a la función de probabilidad conjunta de los fenotipos de los individuos escrita como función de la fracción de recombinación (θ), la penetrancia (λ) y las frecuencias alélicas (f), es decir $L(\theta) = p_\Theta(x_1, x_2, \ldots, x_n)$ donde $\Theta = (\theta, \lambda, f)$. Ahora considere $g = \{g_1, g_2, \ldots, g_n\}$ como el conjunto de posibles genotipos. Utilizando la regla de la probabilidad total, la función de verosimilitud se puede escribir como

$$p_\Theta(x_1, x_2, \ldots, x_n) = \tag{3.1}$$
$$\sum_{g_1}\sum_{g_2}\cdots\sum_{g_n} p_\Theta(x_1, x_2, \ldots, x_n|g_1, g_2, \ldots, g_n)p_\Theta(g_1, g_2, \ldots, g_n).$$

Es decir, es una suma ponderada sobre todos los posible genotipos consistentes con el fenotipo de los individuos. Cada término de esta expresión se compone de dos factores, la penetrancia, que corresponde a la probabilidad de los fenotipos dada una configuración de genotipos, y la probabilidad de dicha configuración. Con respecto a la penetrancia, se asume que los fenotipos de los individuos son condicionalmente independientes dado los genotipos,

por lo tanto

$$p_\Theta(x_1, x_2, \ldots, x_n | g_1, g_2, \ldots, g_n) = \prod_i^n p_\Theta(x_i | g_i). \tag{3.2}$$

Por su parte, la probabilidad de una configuración de genotipos, se puede descomponer de la siguiente manera

$$p_\Theta(g_1, g_2, \ldots, g_n) = p_\Theta(g_1) \prod_{i=2}^n p_\Theta(g_i | g_1, \ldots, g_{i-1}). \tag{3.3}$$

Los elementos en el producto en la expresión anterior pueden tomar dos formas: para los individuos fundadores (que no tienen padres en la familia), simplemente corresponde a la probabilidad de su genotipo, es decir $p_\Theta(g_i)$, ya que este no depende de los genotipos de otros individuos; y para los individuos no-fundadores (que sí tienen padres en la familia) corresponde a la probabilidad de su genotipo dado el genotipo de sus padres, es decir $p_\Theta(g_i | g_{P_i}, g_{M_i})$, donde los subíndices P_i y M_i hacen referencia al padre y a la madre del individuo i, respectivamente.

Derivado de lo anterior, la función de verosimilitud se puede reescribir como:

$$p_\Theta(x_1, x_2, \ldots, x_n) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_n} \prod_i^n p_\Theta(x_i | g_i) \prod_j^f p_\Theta(g_j) \prod_k^{nf} p_\Theta(g_k | g_{P_k}, g_{M_k}).$$

El primer término integra parámetros propios de la penetrancia que dependen del tipo de variable elegida para representar a la enfermedad, ya sea

como variable dicotómica (por ejemplo, afectado o no-afectado) o variable continua (por ejemplo, nivel de colesterol), y el mecanismo a través del cual se asume se hereda la enfermedad. Para este último, el escenario más sencillo puede ser un gen dominante o recesivo, y el escenario más complejo puede ser varios genes con una contribución menor interactuando entre ellos y con el medio ambiente. Con respecto al segundo término, generalmente se asume que los genotipos de los fundadores están en equilibrio de Hardy-Weinberg, y por lo tanto su frecuencia depende únicamente de la distribución de las frecuencias alélicas en la población. Si este último supuesto no es válido, una alternativa es incluir un parámetro conocido como coeficiente de consanguinidad, (Weir, 1996). Por último, el tercer término integra el parámetro de mayor interés, la fracción de recombinación.

Para dos loci bialélicos, una familia pequeña e información completa de todos los individuos, evaluar manualmente la función de verosimilitud es una tarea tediosa pero posible. Sin embargo, cuando el tamaño y la complejidad de la familia aumenta, en conjunto con el número de loci, y adicionalmente se cuenta con información faltante, evaluar dicha función requiere de algoritmos que realicen esta tarea de manera eficiente. Por esta razón, en las últimas décadas se ha invertido mucho esfuerzo en buscar estrategias con este fin. A

la fecha hay dos algoritmos principales para esto, uno propuesto por Elston y Stewart (1971) y el otro propuesto por Lander y Green (1987).

3.2.2. Algoritmo Elston-Stewart

Elston y Stewart (1971) proporcionaron un algoritmo para evaluar recursivamente la función de verosimilitud para diferentes escenarios, un locus, varios loci no ligados, dos loci ligados, entre otros. Este algoritmo, conocido como *peeling*, permite utilizar familias de múltiples generaciones con estructuras simples (donde el árbol inicia con una pareja fundadora), y no permite consanguinidad en las familias. Suponga que los individuos están ordenados de tal forma que los padres anteceden a los hijos. La siguiente representación de la ecuación 3.2 se puede encontrar en Ott (1974) y ejemplifica de manera explícita la propuesta de Elston y Stewart,

$$p_\Theta(x_1, x_2, \ldots, x_n) =$$

$$\sum_{g_1} p_\Theta(x_1|g_1)p_\Theta(g_1|\cdot) \sum_{g_2} p_\Theta(x_2|g_2)p_\Theta(g_2|\cdot) \ldots \sum_{g_n} p_\Theta(x_n|g_n)p_\Theta(g_n|\cdot)$$

donde $p_\Theta(g_i|\cdot)$ corresponde a $p_\Theta(g_i)$ cuando se trata de un fundador y a $p_\Theta(g_i|g_{P_i}, g_{M_i})$ cuando se trata de un no-fundador, para $i = 1, \ldots, n$. Este algoritmo calcula la verosimilitud, familia nuclear por familia nuclear, em-

pezando por los individuos de la generación más reciente. Para cada familia nuclear primero se obtiene la verosimilitud de los hijos y el resultado se anexa al calculo de la verosimilitud de los padres, donde el último padre procesado será aquel que conecta a la familia nuclear con el resto de la familia.

Ott (1974) extendió este algoritmo para considerar estructuras de familias ligeramente más generales, sin consanguinidad, y lo implementó en un programa llamado LIPED. Sin embargo, el caso aplicable a cualquier estructura de familia que incluya varios casos de consanguinidad fue resuelto por Cannings et al. (1978).

El programa LIPED realiza un análisis de ligamiento genético de dos puntos, es decir, infiere la posición del locus de la enfermedad relativa a un solo marcador. Basados en las mismas ideas, la extensión para el caso de marcadores múltiples fue propuesta por Lathrop et al. (1984) e implementada en un programa llamado LINKAGE.

A la fecha, el programa más flexible disponible es FASTLINK (Cottingham et al. 1993), el cual es simplemente una versión acelerada del programa LINKAGE, que puede manejar estructuras de familias con casos de consaguinidad, y es posible realizar un análisis de dos y múltiples puntos.

Un programa que supera en velocidad a FASTLINK es VITESSE (O'Connell

y Weeks, 1995). Lo que hace que este programa sea más rápido es que tiene una forma más eficiente de representar los genotipos de múltiples loci, y adicionalmente recodifica los genotipos de los individuos en conjuntos de alelos transmitidos y no-transmitidos.

El problema principal del algoritmo Elston-Stewart es el número de posibles genotipos que pueden ser asignados a los individuos. Por ejemplo, para dos loci bialélicos, es posible formar $H = 4$ haplotipos, y por lo tanto cada individuo tiene $H(H+1)/2 = 10$ posibles genotipos, es decir, para n individuos se tienen 10^n posibles configuraciones de genotipos. Por esta razón, los programas existentes basados en este algoritmo permiten un número considerable de individuos en la familia pero sucumben ante el número de marcadores.

3.2.3. Algoritmo Lander-Green

Un algoritmo con la característica contraria a la que posee el algoritmo Elston-Stewart, es decir, que permite un número considerable de marcadores pero el número de individuos es limitado, es el propuesto por Lander y Green (1987). La estrategia utilizada por estos autores es la aplicación del algoritmo esperanza-maximización (EM), (Dempster et al. 1977).

Los supuestos principales del algoritmo Lander-Green es que los marcado-

res están ordenados correctamente y que no hay interferencia genética. Para l loci, $M_1, M_2, \ldots, M_l$, el objetivo es encontrar el vector $\theta = (\theta_1, \theta_2, \ldots, \theta_{l-1})$ que maximiza la probabilidad de observar los datos disponibles, donde θ_j corresponde a la fracción de recombinación entre los loci adyacentes M_j y M_{j+1}, y se utiliza el algoritmo EM para encontrar el vector θ que tiene esa característica. De manera general, el algoritmo EM, para un vector inicial cualquiera de $\theta^\star = (\theta_1^\star, \theta_2^\star, \ldots, \theta_{l-1}^\star)$, itera consecutivamente entre el paso que calcula el número esperado de recombinaciones y no recombinaciones en cada intervalo con base en el vector $\theta^\star$ (esperanza) y, el paso que maximiza y actualiza el vector $\theta^\star$ basado en los valores esperados del paso anterior (maximización), hasta que los valores del vector $\theta^\star$ se estabilizan.

El paso complicado es el de la esperanza, y aquí los autores sugieren utilizar vectores de segregación. Para el locus M_j y nf individuos no-fundadores, sea $V_j = (V_{1j}, V_{2j}, \ldots, V_{mj})$ un vector de segregación, donde cada una de las entradas corresponde a cada una de las meiosis ($m = 2nf$), y contiene una variable indicadora binaria que toma el valor de 0 para indicar que el locus j del individuo proviene de la abuela materna (o paterna) y 1 para indicar que proviene del abuelo materno (o paterno).

Note que la probabilidad de que $V_{ij} \neq V_{ij+1}$ (por ejemplo, $V_{ij} = 1$ y

$V_{ij+1} = 0$) es θ_j, para $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, l$. Con base en esto, los autores proponen una cadena de Markov oculta con una probabilidad de transición que depende justamente del parámetro θ_j, con el fin de obtener la distribución de probabilidad de todas las posibles configuraciones del vector V_j, dado los datos disponibles, y de esta manera obtener el número esperado de recombinaciones en cada intervalo consecutivamente.

Kruglyak et al. (1996) proponen algunas mejoras al modelo de Markov oculto sugerido por Lander y Green (1987) y las implementan en un programa conocido como GENEHUNTER. En la actualidad, este programa es uno de los más populares entre todos los programas de ligamiento genético de distribución gratuita disponibles. Sin embargo, dado que tiene como base el algoritmo Lander-Green, el programa GENEHUNTER está limitado a un número reducido de individuos, aproximadamente 18, mientras que el número de marcadores que puede manejar es considerablemente alto.

Se han propuesto diversas estrategias para optimizar los algoritmos Elston-Stewart y Lander-Green, que se han incorporado en las versiones más recientes de los programas disponibles, sin embargo, las limitaciones de cada uno se mantienen a pesar de todos los esfuerzos. Hasta el día de hoy, estos siguen siendo los dos algoritmos principales para realizar un análisis de ligamiento

genético basado en la fracción de recombinación y donde el resultado que se obtiene es exacto.

3.2.4. Métodos de Monte Carlo vía cadenas de Markov (MCMC)

En los casos en donde el número de marcadores y/o individuos excede el límite de cada uno de los algoritmos discutidos previamente, la alternativa es utilizar lo que se conoce como métodos de Monte Carlo vía cadenas de Markov (MCMC), (Gilks et al. 1996).

Como se vio en la sección de enfoque Bayesiano (3.1.2), evaluar la hipótesis de ligamiento requiere de calcular la integral,

$$P(0 \leq \theta < 1/2) = \int_{0 \leq \theta < 1/2} p(\theta|x)\, d\theta$$

Sin embargo, en ocasiones no es posible realizar este cálculo de manera análitica, y por lo tanto es necesario recurrir a una herramienta numérica, entre las que se encuentra el método de Monte Carlo.

El principio del método de Monte Carlo consiste en escribir la integral como un valor esperado. Recuerde que el valor esperado de una variable aleatoria X es el promedio de X y está dado por:

$$E(X) = \sum_x xp(x) \qquad \text{si } X \text{ es discreta , o}$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx \quad \text{si } X \text{ es continua,}$$

donde $p(x)$ y $f(x)$ son las funciones de probabilidad y de densidad de probabilidad, respectivamente.

Considere que se desea calcular una integral que tiene, de manera general, la siguiente forma,

$$I = \int f(x)\,dx.$$

Note que I se puede escribir como,

$$I = \int \frac{f(x)}{s(x)} s(x)\,dx = E_s\left[\frac{f(x)}{s(x)}\right]$$

donde a $s(x)$ se le conoce como distribución de muestreo por importancia.

Si ahora generamos una muestra $X_1, X_2, \ldots, X_N$ de $s(x)$ entonces podemos aproximar la integral I a través de

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(x_i)}{s(x_i)}$$

La precisión de $\hat{I}$ dependerá del tamaño de la muestra N y de la distribución de muestreo $s(x)$. Esta última generalmente se elige de manera que sea fácil de simular y que tenga una forma similar a $f(x)$, lo cual no siempre es factible. Esta técnica se le conoce como muestreo por importancia.

Sin embargo, cuando la función que se desea integrar es multidimensional una mejor estrategia es utilizar una cadena de Markov apropiada cuya distribución de equilibrio sea la distribución de interés, que en este caso llamaremos $\pi(x)$, y de esta manera aproximar algunas características de dicha distribución, o estimar el valor esperado de una función $f(x)$ con respecto a la distribución $\pi(x)$. A este técnica se le conoce como métodos de Monte Carlo vía cadenas de Markov y se abrevia MCMC por sus siglas en inglés.

En otras palabras, la idea es generar $X^{(1)}, X^{(2)}, \ldots, X^{(t)}, \ldots$ una cadena de Markov homogénea en el tiempo, irreducible y aperiódica con espacio de estados finito Δ y distribución de equilibrio $\pi(x)$. Entonces, conforme $t \to \infty$,

(i) $X^{(t)} \xrightarrow{f} X$, donde $X \sim \pi(x)$;

(ii) $\frac{1}{t}\sum_{i=1}^{t} f(X^{(i)}) \to E_{\pi}[f(X)]$.

En el inciso (i), el término $\xrightarrow{f}$ representa convergencia en distribución y la expresión $X \sim \pi(x)$ indica que la distribución de X es $\pi(x)$, (Feller, 1968;Karlin y Taylor, 1975). En el apéndice A se define lo que es una cadena de Markov homogénea en el tiempo, irreducible y aperiódica.

Dos de los algoritmos para generar cadenas de Markov son el algoritmo de Metropolis-Hastings (Metropolis et al. 1953; Hastings, 1970) y el muestreo

de Gibbs (Geman y Geman, 1984). A continuación se presenta una breve introducción del mecanismo general de cada uno de estos algoritmos.

3.2.4.1. Algoritmo de Metropolis-Hastings

Este algoritmo construye una cadena de Markov definiendo probabilidades de transición de $X^{(t)} = x$ al siguiente estado de la cadena, $X^{(t+1)}$, de la siguiente manera:

Sea $q(x, x^\star)$ una distribución de transición (arbitraria), se define

$$\alpha(x, x^\star) = min\left\{1, \frac{\pi(x^\star)q(x^\star, x)}{\pi(x)q(x, x^\star)}\right\}$$

Entonces, dado un valor inicial $X^{(0)} = x^0$ cualquiera, el valor $X^{(t+1)}$ en la *t-ésima* iteración se obtiene a través de los siguientes pasos,

1. generar una observación $x^\star$ utilizando $q(x, x^\star)$;

2. generar una observación u utilizando una distribución uniforme en el intervalo $(0, 1)$;

3. si $u \leq \alpha(x, x^\star)$, se acepta $x^\star$, es decir $X^{(t+1)} = x^\star$; en caso contrario, se rechaza $x^\star$ y por lo tanto $X^{(t+1)} = x$.

3.2.4.2. Muestreo de Gibbs

Al igual que el algoritmo de Metropolis-Hastings, el muestreo de Gibbs permite simular una cadena de Markov $X^{(1)}, X^{(2)}, \ldots$ con distribución de equilibrio multidimensional $\pi(x) = \pi(x_1, x_2, \ldots, x_k)$.

Sea $x = (x_1, x_2, \ldots, x_k)$ un vector de k variables, y sea x_{-i} el vector que contiene a todos los elementos de x excluyendo al elemento x_i, es decir $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$. Ahora suponga que las distribuciones condicionales completas, $\pi(x_i|x_{-i})$, de cada uno de los componentes x_i dado los valores del resto de los componentes x_{-i}, están disponibles para muestreo. En otras palabras, es posible generar observaciones de estas distribuciones condicionales completas.

El algoritmo procede como sigue, sean $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_k^{(0)})$ valores iniciales arbitrarios,

genera $x_1^{(1)}$ utilizando $\pi(x_1|x_2^{(0)}, \ldots, x_k^{(0)})$;

genera $x_2^{(1)}$ utilizando $\pi(x_2|x_1^{(1)}, x_3^{(0)}, \ldots, x_k^{(0)})$;

genera $x_3^{(1)}$ utilizando $\pi(x_3|x_1^{(1)}, x_2^{(1)}, x_4^{(0)}, \ldots, x_k^{(0)})$;

$\vdots$

genera $x_k^{(1)}$ utilizando $\pi(x_k|x_{-k}^{(1)})$.

En este punto una iteración queda completa, después de t iteraciones se tiene $(x_1^{(t)}, x_2^{(t)}, \ldots, x_k^{(t)})$. Entonces, conforme $t \to \infty$

$$(X_1^{(t)}, X_2^{(t)}, \ldots, X_k^{(t)}) \xrightarrow{f} (X_1, X_2, \ldots, X_k) \sim \pi(x_1, x_2, \ldots, x_k).$$

En otras palabras, para t suficientemente grande la distribición de $(X_1^{(t)}, X_2^{(t)}, \ldots, X_k^{(t)})$ es $\pi(x_1, x_2, \ldots, x_k)$.

3.2.4.3. Análisis de ligamiento genético vía MCMC

Como vimos en las secciones anteriores, los métodos de Monte Carlo vía cadenas de Markov son algoritmos para generar muestras de una distribución de interés, con el fin de conocer características de la misma tal como el valor esperado de una función con respecto a dicha distribución. En el caso particular del análisis de ligamiento genético, un punto fundamental que no se puede resolver ni con el algoritmo Elston-Stewart ni con el algoritmo Lander-Green es la evaluación de la función de verosimilitud para un número arbitrario de marcadores e individuos.

Note que la función de verosimilitud se puede escribir de la siguiente manera

$$L(\theta) = p_\Theta(x) = \sum_g p_\Theta(x|g) p_\Theta(g) = E(p_\Theta(x|g))$$

donde $x = \{x_1, x_2, \ldots, x_n\}$ y $g = \{g_1, g_2, \ldots, g_n\}$.

Poniendo el análisis de ligamiento genético en el contexto de los métodos de Monte Carlo vía cadenas de Markov, la idea es generar $G^{(1)}, G^{(2)}, G^{(3)}, \ldots$ una cadena de Markov homogénea en el tiempo, irreducible y aperiódica con espacio de estados finito Δ y distribución de equilibrio $p_\Theta(g)$, donde $G^{(k)} = (G_1^k, G_2^k, \ldots, G_n^k)$ es una configuración de genotipos para todos los miembros de la familia. En la *t-ésima* iteración, sea $G^{(t)} = g^{(t)}$ entonces conforme $t \to \infty$,

(i) $g^{(t)} \xrightarrow{f} g$, donde $g \sim p_\Theta(g)$

(ii) $\frac{1}{t}\sum_{i=1}^{t} p_\Theta(x|g^{(i)}) \to E(p_\Theta(x|g))$

En otras palabras, si podemos generar una muestra suficientemente grande de configuraciones de genotipos provenientes de la distribución $p_\Theta(g)$ (ecuación 3.3), entonces será posible estimar la función de verosimilitud, obteniendo $p_\Theta(x|g^{(k)})$ (ecuación 3.2) para cada una de estas configuraciones y calculando su promedio, es decir

$$\hat{L}(\theta) = \frac{1}{t}\sum_{i=1}^{t} p_\Theta(x|g^{(i)})$$

Una manera de generar genotipos para todos los miembros de una familia es a través del algoritmo conocido como *gene-dropping* propuesto por

MacCluer et al. (1986). Este algoritmo asigna genotipos a todos los individuos fundadores de acuerdo a las frecuencias alélicas y segrega alelos de los padres a los hijos de acuerdo a las leyes de Mendel. El problema con este procedimiento es que la gran mayoría de las configuraciones generadas serán incompatibles con el fenotipo de los individuos, es decir $p_\Theta(x|g^{(k)}) = 0$.

Por otro lado, Lange y Matthysse (1989) utilizan el algoritmo de Metropolis para realizar una caminata aleatoria sobre el espacio de configuraciones de genotipos, a los que llaman estados de descendencia genética (*genetic descent states*). Estos autores proponen cuatro reglas de transición para saltar de una configuración a otra, modificando el genotipo de varios individuos simultáneamente en cada transición.

Para un locus bialélico, Lange y Matthysse (1989) demuestran que todas las configuración compatibles (o legales) se comunican entre sí, es decir, dado una configuración inicial legal, cualquier otra configuración legal puede ser alcanzada a través de una secuencia finita de transiciones. Si esto ocurre entonces la cadena es irreducible. Sin embargo, para el caso de un locus multialélico es necesario un tratamiento especial para garantizar que la cadena tiene esta propiedad.

Con el fin de ilustrar lo anterior, considere una familia nuclear como

la que se presenta en la Figura 3.1. Esta familia consta de 4 individuos, dos fundadores y dos no-fundadores. Suponga que el genotipo de los padres no está disponible y el genotipo de los hijos es conocido. La Figura 3.1.a corresponde a un locus con dos alelos A y B, y la Figura 3.1.b corresponde a un locus con tres alelos A, B y C.

Figura 3.1: *Familia nuclear con padres desconocidos, a. locus bialélico y b. locus multialélico*

Considere primero la Figura 3.1.a, donde el genotipo de los individuos 3 y 4 es AA y AB, respectivamente, es decir $g_3 = AA$ y $g_4 = AB$. Note que el genotipo de los hijos permite determinar los posibles genotipos de los padres, es decir $g_1 \in \{AA, AB\}$ y $g_2 \in \{AA, AB\}$. El espacio de posibles configuraciones de genotipos compatibles consta de tres elementos, $(g_1 = AA, g_2 = AB, g_3 = AA, g_4 = AB)$, $(g_1 = AB, g_2 = AB, g_3 = AA, g_4 = AB)$

y $(g_1 = AB, g_2 = AA, g_3 = AA, g_4 = AB)$. En este caso, es posible pasar de cualquiera de estas configuraciones a cualquier otra modificando el genotipo de uno de los padres en cada iteración.

Considere ahora la Figura 3.1.b. De la misma manera que en el caso anterior, el genotipo de los hijos, $g_3 = AA$ y $g_4 = BC$, permite determinar los posibles genotipos de los padres, $g_1 \in \{AB, AC\}$ y $g_2 \in \{AB, AC\}$. En este caso, el espacio de posibles configuraciones de genotipos compatibles consta de únicamente dos elementos, $(g_1 = AB, g_2 = AC, g_3 = AA, g_4 = BC)$ y $(g_1 = AC, g_2 = AB, g_3 = AA, g_4 = BC)$. Observe que, en este caso, no es posible pasar de una configuración a la otra modificando el genotipo de solamente uno de los padres, ya que inevitablemente llegamos una configuración incompatible.

Por otro lado, Sheehan y Thomas (1993) proporcionan un algoritmo para actualizar los genotipos, individuo por individuo, utilizando el muestreo de Gibbs. Dada una configuración inicial de genotipos compatible, en la *k-ésima* iteración se selecciona al individuo i al azar, y se calcula $p(G_i = g|g_{-i}, x) = p(G_i = g|g_{\delta_i}^{(k)}, x_i)$ para todos los posibles genotipos g, donde g_{-i} es el vector de genotipos de todos los individuos excepto el genotipo del individuo i, y g_{δ_i} corresponde al vector de genotipos de los vecinos in-

mediatos al individuo i (padres, esposos e hijos). Utilizando esta distribución de probabilidad se obtiene $G_i^{(k+1)} = g_i^{(k+1)}$ y se actualiza el vector $g^{(k+1)} = (g_1^{(k)}, g_2^{(k)}, \ldots, g_i^{(k+1)}, \ldots, g_n^{(k)})$. La *k-ésima* iteración termina cuando se ha actualizado el genotipo de todos los individuos.

Estos autores demuestran que para un locus con un número arbitrario de alelos es posible construir una cadena de Markov irreducible si se tiene una penetrancia positiva para todos los fenotipos x y genotipos g. Por lo tanto, Sheehan y Thomas (1993) sugieren asignar una probabilidad pequeña positiva, γ, a las penetrancias que bajo del modelo genético son iguales a cero, es decir

$$p^\star(x|g) = \begin{cases} p(x|g) & \text{si} \quad p(x|g) > 0 \\ \gamma & \text{si} \quad p(x|g) = 0 \end{cases}$$

y rechazar aquellas configuraciones que son inconsistentes con el fenotipo de los individuos.

En teoría esta propuesta efectivamente resuelve el problema de la falta de irreducibilidad. Sin embargo, en la práctica no es una estrategia eficiente, ya que valores muy pequeños de γ provocan una mala comunicación entre configuraciones, en el sentido de que aquellas configuraciones que no se comunicaban entre sí, ahora la probabilidad de pasar de una a otra es positiva pero

muy pequeña. Por otro lado, para valores de γ no tan pequeños se presenta una tasa de rechazo alta; por ejemplo, en la aplicación que proporcionan los autores, para un valor de $\gamma = 0.05$ la tasa de rechazo es del 99.11 %. Sin embargo, ellos argumentan que una tasa de rechazo alta disminuye la correlación entre configuraciones y por lo tanto arroja mejores aproximaciones.

Lin et al. (1993) basado en la propuesta de Sheehan y Thomas (1993), sugiere disminuir el espacio de posibles configuraciones modificando únicamente las penetrancias de los genotipos heterocigotos, a los que les llama puentes. Los puentes conectan a subconjuntos de individuos, con sus correspondientes genotipos, que no comunican entre sí. Estos subconjuntos se les llama islas. Adicionalmente, proponen correr cadenas de Markov acopladas donde en cada una utiliza el algoritmo de Metropolis acelerado.

Brevemente, el algoritmo de Metropolis para cadenas de Markov acopladas es una estrategia que se utiliza para alcanzar irreducibilidad. Considere el caso más sencillo, dos muestreos de Gibbs actuando en paralelo, uno de ellos esta restringido a los parámetros del modelo, mientras que el otro tiene parámetros relajados, especificados de tal forma que garantizan irreducibilida. Periódicamente, se intercambian el estado en el que se encuentran las dos cadenas, aceptando o rechazando el cambio de acuerdo a cierta probabilidad.

Una vez que la cadena es irreducible pueden existir configuraciones que son muy poco probables, una solución a este problema es aumentar la velocidad de convergencia usando el algoritmo de Metropolis acelerado.

Desde el punto de vista teórico, la solución de Lin et al. (1993) es más eficiente que la de Sheehan y Thomas (1993), sin embargo, en la práctica es necesario definir el número de cadenas y así como el factor de aceleración (temperatura) de cada una de ellas para cada problema en particular. Además, cuando el número de alelos aumenta, aumenta también el número de puentes necesarios para asegurar que todas las islas están conectadas, y por lo tanto este algoritmo se vuelve impráctico. Derivado de lo anterior, Lin et al. (1994) propone identificar primero todas las islas y construir puentes específicos basados en individuos clave, esto con el fin de disminuir aún más el espacio de posibles configuraciones.

El algoritmo proporcionado por Lin et al. (1994) dice lo siguiente:

1. Forme una secuencia de familias nucleares empezando por la generación más reciente.

2. Para cada familia nuclear,

   a) si ambos padres no tienen genotipo, encuentre todas las islas y

verifique si existe consistencia con otros miembros ya procesados;

b) si únicamente un padre no tiene genotipo, utilice la información de los hijos y conyuge para inferir el genotipo de este individuo, y verifique si existe consistencia con otros miembros ya procesados;

c) si ambos padres tienen genotipo continue con la siguiente familia nuclear.

3. Almacene la información de la secuencia en la que fueron procesada las familias nucleares, el número de islas, así como los individuos forzados a tomar un genotipo.

En este reporte, los autores proporcionan cinco diferentes ejemplos para que el lector entienda el razonamiento a través del cual estos autores identifican las islas. Sin embargo, cabe destacar que el algoritmo que proporcionan no es un procedimiento que identifica las islas puntualmente de manera automática, sino es simplemente un procedimiento para maniobrar con la familia de tal forma que la busqueda de las islas sea más organizada.

Posteriormente, Lin (1995) propone un esquema de muestreo para construir una cadena de Markov irreducible saltando entre islas, evitando del todo pasar por configuraciones inconsistentes. Para esto es necesario tener

identificadas las islas de antemano. Finalmente, Lin (1996), lo extiende al caso de marcadores múltiples.

Por su parte, Jensen y Sheehan (1998) argumentan que el algoritmo propuesto por Lin et al. (1994) no es capaz de identificar las islas para cualquier locus polimórfico en cualquier estructura de familia. Para respaldar su argumento proporcionan diferentes contraejemplos y concluyen diciendo que el principal problema con el algoritmo que identifica las islas es considerar que las islas solamente son creadas por los genotipos de los hijos donde los genotipos de los padres son desconocidos.

Por otro lado, Thompson (1994) sugiere una estrategia completamente distinta utilizando los vectores de segregación introducidos por Lander y Green (1987). Recuerde que V_{ij} es una variable binaria que toma el valor de 0 ó 1 para indicar que la meiosis i en el locus j recibió el alelo materno o paterno, respectivamente, de su padre (madre). Esta autora propone utilizar el algoritmo de Metropolis para obtener muestras de $V = \{V_{ij}\}$, seleccionando i y j al azar, modificando V_{ij} de 0 a 1 o de 1 a 0, según sea el caso, para obtener $V^\star = v^\star$ y, aceptando o rechazando el cambio de acuerdo a la siguiente probabilidad,

$$\alpha = min\left\{1, \frac{p_\Theta(x|v^\star)p_\Theta(v^\star)}{p_\Theta(x|v)p_\Theta(v)}\right\}. \tag{3.4}$$

Note que la propuesta modifica únicamente el estatus de recombinación/no-recombinación en los dos intervalos adyacentes al locus j y la probabilidad condicional de los fenotipos en el locus j. Por lo tanto, la probabilidad anterior se puede escribir como:

$$\alpha = min\left\{1, \frac{p_\Theta(x_j|v_j^\star)}{p_\Theta(x_j|v_j)}\left(\frac{\theta_{j-1}}{1-\theta_{j-1}}\right)^{T_{j-1}}\left(\frac{\theta_j}{1-\theta_j}\right)^{T_j}\right\},$$

donde T_{j-1} es una variable indicadora que toma el valor de 1 ó -1 para añadir o remover, respectivamente, una recombinación entre los loci $j-1$ y j. De la misma manera se define la variable T_j. Los componentes principales en el término $p_\Theta(x_j|v_j)$ son la penetrancia y el conjunto de todas las configuraciones alélicas compatibles con v_j.

Este algoritmo es más eficiente por el simple hecho de que el espacio de posible vectores de segregación, V, es menor que el espacio de posibles configuraciones de genotipos, G. Sin embargo, Sobel y Lange (1996) proporcionan un ejemplo muy particular en donde el algoritmo propuesto por Thompson (1994) no es irreducible e introducen otro algoritmo.

El trabajo de Sobel y Lange (1996) es una extensión al trabajo Lange y Matthysse (1989) para el caso de marcadores multialélicos múltiples, utilizando ahora gráficas de descendencia génetica (GDG, por las siglas en inglés de *genetic descent graphs*) en lugar de estados de descendencia genética (GDS,

por las siglas en inglés de *genetic descent states*). En la Figura 3.2, se presenta un ejemplo que ilustra la diferencia entre una y la otra. Ambas pueden verse como gráficas donde cada individuo esta representado por dos nodos, uno materno y otro paterno, y las aristas indican la segregación o el flujo de los nodos fundadores a través de la familia. De esta manera, la etiqueta (alelos) de cada uno de los nodos fundadores en combinación con los vectores de segregación representados en la gráfica a través de las aristas, especifican una configuración completa de genotipos.

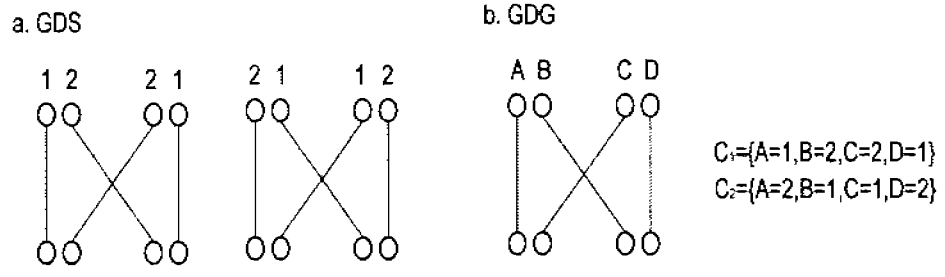

Figura 3.2: *Ejemplo que ilustra cómo dos estados de descendencia genética (GDS) se pueden representar a través de una gráfica de descendencia genética (GDG).*

Como se muestra en la Figura 3.2, hay varias configuraciones de alelos fundadores compatibles con una GDG, cada una de estas configuraciones es

una GDS, por esta razón el espacio de posibles GDG es menor que el espacio de posibles GDS, ya que en este último cada uno de los nodos está fijo con el valor de cada uno de los alelos del individuo.

En esencia, el algoritmo de Sobel y Lange (1996) es muy parecido al algoritmo de Thompson (1994). La diferencia radica en que este último mueve una sola arista en cada paso de la cadena, mientras que el primero consta de cuatro reglas de transición que mueven varias aristas simultáneamente. Para garantizar irreducibilidad, Sobel y Lange (1996) proponen emplear transiciones múltiples en cada paso; particularmente ellos utilizan una distribución geométrica con media 2 para determinar el número de transiciones. En cada paso de la cadena seleccionan al azar el número de transiciones, el tipo de transición y el locus, e igualmente aceptan o rechazan el paso de acuerdo con el cociente de Metropolis (ecuación 3.4). Estos autores implementan su algoritmo en un programa de distribución gratuita llamado SimWalk2.

Otras propuestas motivan el uso del muestreo de Gibbs por bloques. Por ejemplo, E.A. Thompson en colaboración con otros autores implementan, en un programa llamado MORGAN[1], dos algoritmos, el L-sampler y el M-sampler, los cuales hacen una actualización por bloques. El L-sampler, origi-

[1]http://www.stat.washington.edu/thompson/Genepi/

nalmente desarrollado por Kong (1991), actualiza de manera conjunta todas las meiosis, locus por locus. El M-sampler, propuesto por Thompson y Heath (1999), actualiza de manera conjunta todos los loci, meiosis por meiosis. El programa MORGAN, en cada iteración, selecciona uno de los dos algoritmos con una cierta probabilidad (por ejemplo, 20 % L-sampler y 80 % M-sampler).

Jensen y Kong (1999) utilizan también el muestreo de Gibbs por bloques. Sea S el conjunto de variables que serán actualizadas en cada iteración del algoritmo. El conjunto S contiene variables múltiples asociadas con cada individuo de la familia, por ejemplo cada una de las entradas del vector de segregación es una variable. Estos autores proporcionan en su reporte un ejemplo donde para una familia con 73 individuos el conjunto S contiene 915 variables.

Al conjunto S lo agrupan en p bloques, $B_1, B_2, \ldots, B_p$, y utilizan el algoritmo de Gibbs para actualizar consecutivamente el bloque i, para $i = 1, 2, \ldots, p$, dado el conjunto S_{-B_i}, donde S_{-B_i} contiene los elementos de S excluyendo los elementos de B_i. Estos bloques no son necesariamente disjuntos pero su unión debe contener a todos los elementos de S, es decir $B_1 \cup B_2 \cup \ldots \cup B_p = S$, y además deben cumplir con tres criterios, (1) cada bloque debe contener el mayor número de variables posible; (2) todas las

variables deben ser muestreadas con la misma frecuencia y (3) deben de ser de tal forma que el algoritmo sea irreducible.

Con respecto al criterio (1), entre mayor sea el número de variables en cada bloque más eficiente será el muestreo y la convergencia será también más rápida. Los autores proporcionan una estrategia para hacer esto. Con respecto al criterio (2), se desea evitar que algunas variables sean actualizadas muchas veces mientras que otras muy pocas veces (esto se da gracias a que los bloques no son disjuntos) y también los autores proporcionan un algoritmo para esto. Por último, para asegurar irreducibilidad, criterio (3), es necesario que los bloques contengan las variables correspondientes a los individuos que ocasionan el problema de reducibilidad. En la aplicación que ellos presentan, los bloques fueron diseñados manualmente con el fin de garantizar irreducibilidad. Sin embargo, los autores lamentan que no haya a la fecha un algoritmo que identifique, para una estructura de familia arbitraria, los conjuntos de individuos, con sus correspondientes genotipos, que no comunican entre sí, ya que debido a esto no es posible probar que su algoritmo es irreducible en un caso general.

Por último, Sisson (2004) proporciona un método para identificar estos conjuntos mínimos de individuos, con sus correspondientes genotipos, que

no comunican entre sí, o clases no-comunicadoras como él las llama. Este es una generalización al trabajo de Lin et al. (1994) y utiliza como ejemplo dos familias proporcionadas por Jensen y Sheehan (1998). Una de las limitantes de este algoritmo es que puede arrojar clases no-comunicadoras vacías o duplicadas.

Capítulo 4

Comparación GENEHUNTER y SimWalk2

4.1. Planteamiento

La mayoría de las propuestas para modificar, extender o hacer más eficiente un algoritmo ya existente van acompañadas de una implementación en un programa de cómputo, que usualmente se pone a disposición del público interesado de manera gratuita. En la página de Internet que mantiene el Laboratorio de Genética Estadística de la Universidad de Rockefeller[1] es posible

[1] http://linkage.rockefeller.edu/

encontrar una lista actualizada de los programas disponibles. Por ejemplo, el algoritmo Elston-Stewart fue implementado en el programa llamado LINKAGE, el algoritmo Lander-Green fue implementado en el programa llamado GENEHUNTER. Estos dos programas proporcionan un resultado exacto, sin embargo, como se comentó previamente tienen limitaciones con respecto al número de marcadores e individuos, respectivamente. Una alternativa que puede manejar un número considerablemente grande de marcadores e individuos simultáneamente son los métodos de Monte Carlo vía cadenas de Markov. Estos métodos proporcionan un resultado aproximado aunque la aproximación puede ser tan precisa como se quiera. Dos de los programas que realizan un análisis de ligamiento genético utilizando estos métodos son SimWalk2 y MORGAN.

En la actualidad, el programa más utilizado es GENEHUNTER, debido principalmente a tres factores: (1) tiene una interface amigable, (2) es muy rápido, es decir, en pocos minutos se tiene el resultado disponible, y (3) el número de marcadores permitido es considerablemente alto. Cabe destacar que actualmente un escaneo completo del genoma humano puede contener fácilmente más de 400 marcadores. Sin embargo, la limitación más importante de este programa es el número máximo de individuos permitido, que

depende de la estructura de la familia pero que no supera 12 individuos no-fundadores. Aún cuando el tamaño de la familia excede el límite permitido por GENEHUNTER, en muchos casos se utiliza igualmente este programa. Cuando esto ocurre el programa descarta los individuos excedentes empezando por los sanos. Con el fin de incluir el mayor número de individuos posible, los autores del programa recomiendan partir manualmente a la familia en familias más pequeñas y tratarlas como si éstas fueran independientes.

A pesar de las bondades de los métodos de Monte Carlo vía cadenas de Markov, la implementación de estos métodos en el análisis de ligamiento genético no ha tenido el impacto que los autores que los propusieron hubieran querido en la comunidad de investigadores usuarios de estos programas. Esto puede deberse a que estos métodos requieren que el usuario tenga un conocimiento teórico más profundo con el fin de especificar los parámetros necesarios para hacer el análisis. De los programas que utilizan métodos de Monte Carlo vía cadenas de Markov, el más destacado es SimWalk2. Este programa puede manejar familias de aproximadamente 258 individuos, sin embargo, su mayor limitación es el tiempo de procesamiento y una interface poco amigable, ya que requiere de archivos de entrada que tienen un formato complicado. La manera más sencilla de generar estos archivos es utilizando el

programa MEGA2 (Mukhopadhyay et al., 1999). Debido a que cada programa que se pone a disposición del público en general requiere de archivos de entrada con formatos específicos; MEGA2 fue creado justamente con el fin de facilitar a los usuarios la creación de estos archivos. MEGA2 requiere de tres archivos de entrada con un formato particular, mismos que convierte en formatos alternativos requeridos por otros programas, entre ellos SimWalk2.

Por otro lado, la materia prima de un análisis de ligamiento genético son familias multigeneracionales, conocidas como pedigríes. Un problema inherente es la información faltante debido a que algunos miembros de las generaciones iniciales no están disponibles. En algunos casos es posible determinar, a través de un sondeo entre los familiares, el fenotipo de los individuos ausentes, así como en ocasiones es posibles inferir el genotipo de estos individuos a través de los genotipos de otros miembros de la familia. Sin embargo, en la práctica no se contará con información completa y esto inevitablemente afectará el resultado del análisis. A la fecha, no hay en la literatura un estudio que investigue este problema, es decir, cómo la información faltante afecta el valor de lod score.

Se ha observado que, utilizando la misma base de datos, GENEHUNTER y SimWalk2, pueden arrojar resultados diferentes, encaminando a conclusio-

nes contrarias. En este caso es necesario evaluar cuidadosamente cuál resultado es el más confiable. Un ejemplo de un caso particular se presentó durante el desarrollo de un proyecto de investigación en donde el fenotipo de interés era hipercolesterolemia familiar. En este estudio, cuando 17 marcadores del cromosoma 2 fueron analizados (ver Figura 4.1(a)), los resultados de GENEHUNTER y SimWalk2 arrojaban resultados contradictorios. Se obtuvo un lod score de 1.2 a través de GENEHUNTER y de -4.9 a través de SimWalk2. Fueron estos resultados en conjunto con la limitación de GENEHUNTER concerniente al número de individuos lo que motivó el trabajo del presente capítulo.

El efecto, sobre el valor del lod score, de descartar individuos tal y como lo hace GENEHUNTER, así como el efecto de partir a la familia en familias independientes más pequeñas ha sido estudiado previamente por Goedken et al. (2000). Estos autores compararon los resultados obtenidos por GENEHUNTER y VITESSE (O'Connell and Weeks, 1995) para seis familias reales. El tamaño de las familias varía en un rango de 17 a 39 individuos, y constan de entre 3 y 13 individuos afectados por familia. Las conclusiones generales alcanzadas por estos autores son que hay una perdida de información de ligamiento cuando GENEHUNTER descarta individuos. Con respecto a dividir

a la familia, ellos concluyen que partir a la familia tiene un efecto negativo mayor que descartar individuos.

4.2. Objetivo

Este capítulo tiene como objetivo, utilizar datos simulados para:

1. Comparar los resultados de GENEHUNTER y SimWalk2 para el caso de una familia de tamaño moderado.

2. Evaluar el efecto de descartar individuos.

3. Evaluar el efecto de partir a la familia en familias independientes más pequeñas.

4. Evaluar el efecto de información faltante.

Se simularon dos modelos de enfermedad basados en la estructura de una familia real. Ambos modelos corresponden a un gen dominante con penetrancia alta (90 %) y baja (50 %). Para cada uno de estos modelos se simularon dos escenarios, presencia y ausencia de ligamiento entre el gen responsable de la enfermedad, clasificada cualitativamente, y un marcador.

4.3. Metodología

A partir de este punto GENEHUNTER y SimWalk2 serán denominados GH y SW2, respectivamente.

Basado en la estructura de una familia real (Figura 4.1), se simularon genotipos para los miembros de la familia. La estructura y caracterización de esta familia corresponde a una versión preliminar de la familia utilizada en un estudio de ligamiento real (Canizales-Quinteros et al., 2003). El supuesto que se utilizó en dicho estudio fue que el fenotipo en esta familia está determinado por un sólo gen con un mecanismo de herencia autosómico dominante con penetrancia incompleta del 90 %. Fueron considerados diecisiete marcadores polimórficos (microsatélites) a lo largo del cromosoma 2. La familia consta de 72 individuos distribuidos en cinco generaciones, 11 de los cuales están afectados con hipercolesterolemia familiar (Figura 4.1(a)), 31 individuos con genotipo desconocido, 22 individuos con estatus de la enfermedad desconocido, y de estos últimos, 19 tiene también genotipo desconocido. La familia que mantiene estas mismas características se le denomina "original" en el resto del texto, incluyendo cuadros y figuras.

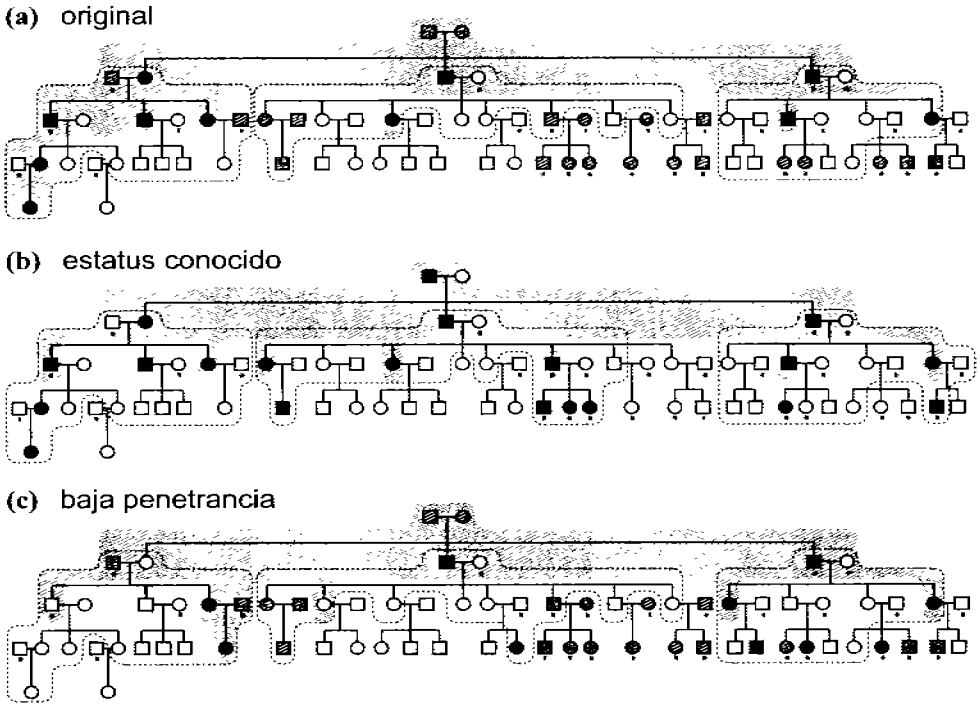

Figura 4.1: *Familias estudiadas, donde los símbolos en negro corresponden a individuos afectados, los símbolos en blanco corresponden a individuos sanos y los símbolos en gris corresponden a individuos con estatus de la enfermedad desconocido. Los individuos no disponibles para su genotipificación están indicados con un asterisco. El área sombreada incluye a los individuos que GENEHUNTER mantiene en el análisis. Cada una de las tres familias (a-c) se dividió en tres familias más pequeñas, las líneas discontinuas indican para cada una de estas tres familias los individuos que GENEHUNTER incluye en el análisis.*

Adicionalmente, otros dos casos fueron considerados, cada uno de los cuales fue producido modificando una característica en particular de la descripción anterior. En el primero caso se redujo la penetrancia a un 50 %. Utilizando esta penetrancia y la misma estructura familiar, los estatus de los individuos fueron reasignados. La familia resultante (Figura 4.1(c)) consta de 8 individuos afectados y será denominada "baja penetrancia". En el segundo caso, la penetrancia del 90 % se mantiene pero se les asigna aleatoriamente estatus de la enfermedad a los 22 individuos que no lo tenían en el caso de la familia "original". Como se observa en la Figura 4.1(b), la familia resultante consta de 20 individuos afectados. A este último caso se le denominará "estatus conocido".

Para cada uno de los tres casos descritos, "original", "baja penetrancia" y "estatus conocido", se simularon dos modelos. El primer modelo asume que el gen de la enfermedad no se encuentra ligado al cromosoma 2 (no-ligamiento). El segundo modelo asume que el gen de la enfermedad está ligado al marcador 14 (posición 168.33 cM) en el cromosoma 2 a una fracción de recombinación de cero, es decir $\theta = 0$ (ligamiento). Ambos modelos fueron desarrollados considerando la penetrancia correspondiente y el estatus especificado en la Figura 4.1.

Usando las frecuencias alélicas estimadas para los diecisiete marcadores y bajo el supuesto de equilibrio de Hardy-Weinberg, se generaron genotipos para aquellos individuos que no tienen padres en la familia (fundadores). Para generar genotipos de los individuos que sí tienen padres en la familia, el número de recombinaciones, así como su posición entre los marcadores fue simulada de acuerdo a un proceso Poisson como el descrito por Terwilliger et al., 1993.

Usando este procedimiento, para cada uno de los dos modelos (no-ligamiento y ligamiento) se generaron 100 réplicas de las tres diferentes familias presentadas en la Figura 4.1. Estas réplicas reflejan el escenario hipotético donde los genotipos de cada individuo de la familia son conocidos. Con base en estos dos grupos de réplicas, se crearon 5 grupos adicionales, donde cada uno representa un escenarios en particular. En total se generaron 6 diferentes grupos de 100 réplicas, para cada uno de los dos modelos estudiados (Cuadro 4.1). En el grupo 1, el genotipo de los 72 miembros de la familia es conocido. En el grupo 2, el genotipo de 31 de los 72 miembros de la familia es desconocido, tal y como ocurre en la familia real. En el grupo 3, la familia consta únicamente de los individuos que GH incluye en el análisis, todos con genotipo conocido. En el grupo 4, 8 individuos del grupo 3 tienen genotipo desconocido. En el

grupo 5, la estructura original de la familia se dividió en 3 subfamilias de 19, 31 y 20 individuos cada una, todos con genotipo conocido. Por último, en el grupo 6 se consideró la misma partición que en el grupo 5 pero en este último escenario el genotipo de 6 individuos de la primera subfamilia, 12 de la segunda y 11 de la tercera es desconocido.

Se realizó un análisis de ligamiento genético de múltiples puntos utilizando los programas GENEHUNTER v.2 beta y/o SimWalk2 v.2.82. En el Cuadro 4.1 presenta las principales características de los 6 diferentes escenarios descritos y el programa utilizado en cada escenario. Con el fin de resumir la descripción de cada uno de los 6 escenarios, la última columna del Cuadro 4.1 se proporciona la siguiente notación: T para referirse a la familia que consta de todos los individuos, D para referirse a la familia en la que se descartaron individuos y P para referirse a la familia que se dividió en familias más pequeñas, mientras que C e I se utilizan para denotar los escenarios que tienen información completa o incompleta, respectivamente. Por información completa se entiende que todos los individuos de la familia tienen genotipo conocido, mientras que por información incompleta se entiende que algunos individuos tienen genotipo desconocido.

Grupo	No. familias	No. individuos (O, EC, BP)	Individuos faltantes (O, EC, BP)	Programa	Notación
1	1	(72, 72, 72)	(0, 0, 0)	SW2	TC
2	1	(72, 72, 72)	(31, 31, 31)	SW2	TI
3	1	(18, 16, 16)	(0, 0, 0)	GH/SW2	DC
4	1	(18, 16, 16)	(8, 8, 8)	GH/SW2	DI
5	3	(45, 50, 46)	(0, 0, 0)	GH	PC
6	3	(45, 50, 46)	(14, 20, 15)	GH	PI

Cuadro 4.1: *Descripción de los seis grupos generados para cada uno de los dos modelos (ligamiento y no-ligamiento), para cada una de las tres familias (O: original, EC: estatus conocido y BP: baja penetrancia). Las columnas dos, tres y cuatro indican, para cada uno de los grupos, el número de familias consideradas, el número de individuos incluidos en el análisis, así como el número de individuos cuyo genotipo no está disponible, respectivamente. La columna cinco indica el programa utilizado en cada caso, es decir, GENEHUNTER (GH) y/o SimWalk2 (SW2), y la última columna indica la notación utilizada tanto en el texto como en las gráficas relacionadas.*

4.4. Resultados

4.4.1. Comparación de GENEHUNTER y SimWalk2 (familia de tamaño moderado)

Con el fin de comparar el resultado de GH y SW2 para el caso de una familia de tamaño moderado, cada una de las 100 réplicas del grupo 4 (DI) fue analizada con ambos programas. Como se muestra en la Figura 4.2, los valores de lod score obtenidos por GH y SW2, para los 17 marcadores, son muy semejantes para el modelo de no-ligamiento. Por su parte, cuando hay ligamiento se observa un comportamiento similar tomando en consideración únicamente los valores de lod score obtenidos para la posición del gen de la enfermedad, es decir, en el marcador 14. Las Figuras 4.2(a), 4.2(b), y 4.2(c) corresponden a las tres familias consideradas, "original", "baja penetrancia" y "estatus conocido", respectivamente.

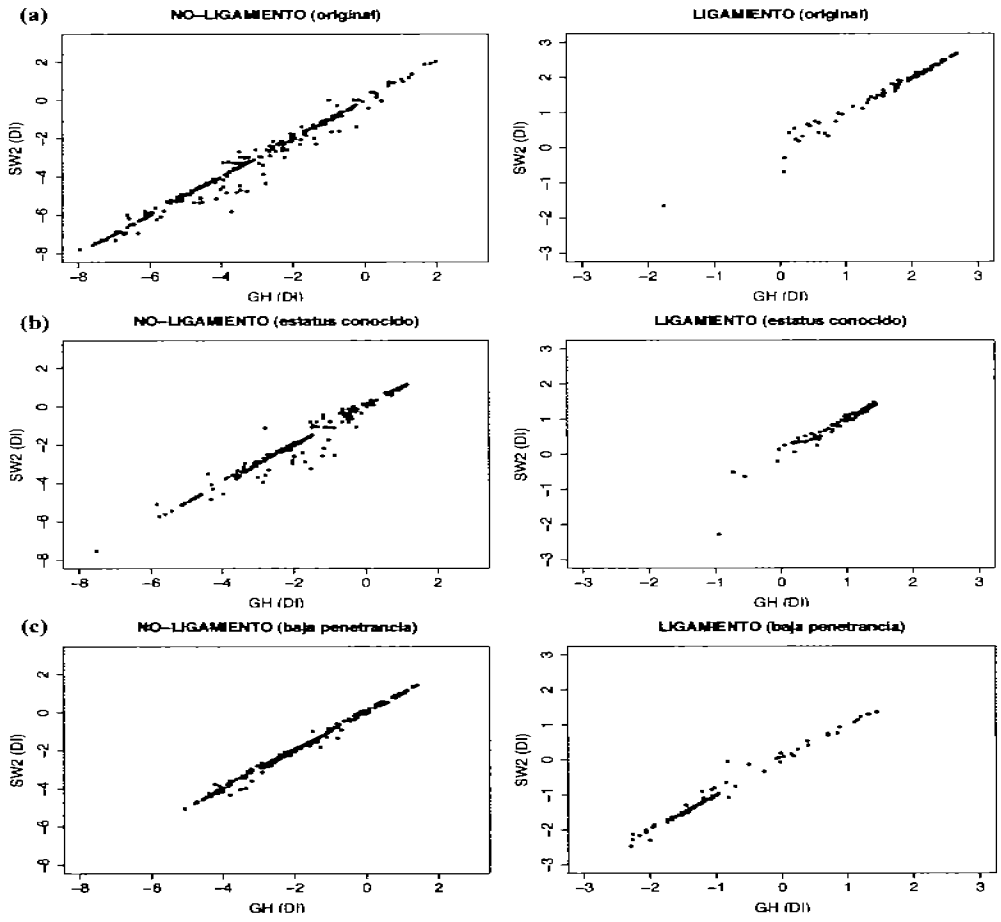

Figura 4.2: *En presencia de información faltante, comparación de los valores de lod score obtenidos a través de GH y SW2 para las tres familias de tamaño moderado. Los resultados correspondientes a los 17 marcadores fueron utilizados en el modelo de no-ligamiento, mientras que únicamente los resultados correspondientes al marcador 14 fueron utilizados en el modelo de ligamiento.*

Al comparar los resultados puntuales obtenido a través de GH y SW2, se observan pocos casos en donde el resultado de ambos programas difiere. Estos casos, en la Figura 4.2, son aquellos puntos que no caen exactamente en la diagonal. En aproximadamente el 95 % de los casos se observó una diferencia no mayor a ± 0.5.

La fracción de recombinación estimada a partir de los dos programas se muestra en la Figura 4.3. Únicamente se presentan los escenarios donde hay información incompleta. Estos valores fueron calculados a partir de la diferencia, en "centiMorgans", entre la posición que produjo el valor de lod score máximo y la posición del marcador 14. Posteriormente esta diferencia se transformó a su correspondiente fracción de recombinación utilizando la función de mapeo correspondiente. Como se observa en la Figura 4.3, la distribución en las tres familias es muy similar para ambos programas.

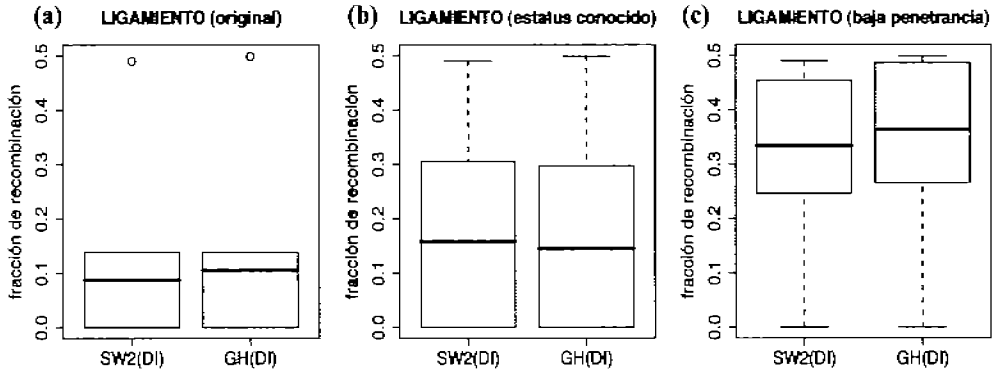

Figura 4.3: *En el modelo de ligamiento y en presencia de información faltante, comparación de la distribución muestral de la fracción de recombinación estimada a través de GH y SW2 para cada una de las familias de tamaño moderado.*

El Cuadro 4.2 muestra la potencia, también conocido como poder, de cada uno de los programas para detectar ligamiento cosiderando las tres familias estudiadas. Los valores del Cuadro 4.2 corresponden a la proporción de valores de lod score que fueron mayores o iguales a 3 en el modelo de ligamiento. Las celdas sin valores corresponden a escenarios que no fueron considerados. En el caso de GH no es posible utilizar la familia completa, y como SW2 permite realizar el análisis con la familia completa, los grupos 5 y 6, es decir aquellos en donde se dividió la familia en subfamilias, no se

corrieron con este programa.

	T		P		D	
	C	I	C	I	C	I
SW2						
Original	0.83	0.75	—	—	0.00	0.00
Estatus conocido	0.82	0.43	—	—	0.00	0.00
Baja penetrancia	0.02	0.01	—	—	0.00	0.00
GH						
Original	—	—	0.61	0.48	0.00	0.00
Estatus conocido	—	—	0.65	0.14	0.00	0.00
Baja penetrancia	—	—	0.00	0.00	0.00	0.00

Cuadro 4.2: *En el modelo de ligamiento y cuando se dispone de información completa (C) e incompleta (I), la proporción de valores de lod score mayores o iguales a 3 para el caso de la familia que incluye a todos los individuos (T), la familia donde se descartaron individuos (D) y la familia dividida en familias más pequeñas (P).*

4.4.2. Efecto de descartar individuos y partir a la familia

Con el fin de evaluar el efecto de descartar individuos y partir a la familia, se compararon los resultados de los grupos 4 (DI) y 6 (PI) con los resultados del grupo 2 (TI) para cada una de las familias, "original", "baja penetrancia" y "estatus conocido".

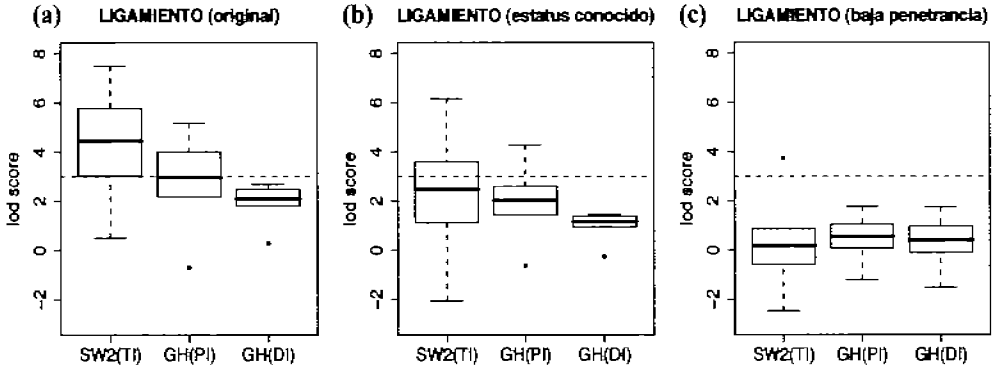

Figura 4.4: *En el modelo de ligamiento y en presencia de información faltante, comparación de los valores de lod score máximos obtenidos a través de SW2 para la familia completa, y obtenidos a través de GH para la familia donde se descartaron individuos y la familia dividida en familias más pequeñas.*

El número de individuos incluidos en cada una de las tres familias se

presenta en el Cuadro 4.1. La Figura 4.4 muestra, para cada una de las tres familias, el comportamiento de los valores máximos de lod score correspondiente al modelo de ligamiento para los siguientes escenarios: SW2 familia completa (TI), GH familia dividida (PI) y GH descartando individuos (DI). Las diferencias entre las cajas en la Figuras 4.4(a) y 4.4(b) representan el efecto negativo de dividir a la familia o descartar individuos comparado con el escenario que incluye a todos los individuos, cuando la enfermedad tiene una penetrancia considerablemene alta (90 %). De estos resultados es claro que dividir a la familia es una mejor alternativa que descartar individuos del análisis.

En el caso de la familia de baja penetrancia (50 %), como se muestra en la Figura 4.4(c), los valores de lod score obtenidos a través de SW2 utilizando a la familia entera son más bajos que aquellos producidos por GH para los escenarios donde se divide a la familia o se descartan individuos. Esto se debe a que en el análisis de la familia entera, se incluyen individuos sanos no-penetrantes como consecuencia de la penetrancia baja, y son estos individuos los que generalmente son descartados por GH cuando la familia se excede en tamaño. Con el fin de minimizar el efecto de los individuos sanos no-penetrantes sobre el valor de lod score, una estrategia es asignar

estatus desconocido a todos los individuos sanos e incluirlos en el anásis. Esta estrategia se le conoce en la literatura como análisis de sólo afectados (*affecteds-only analysis*).

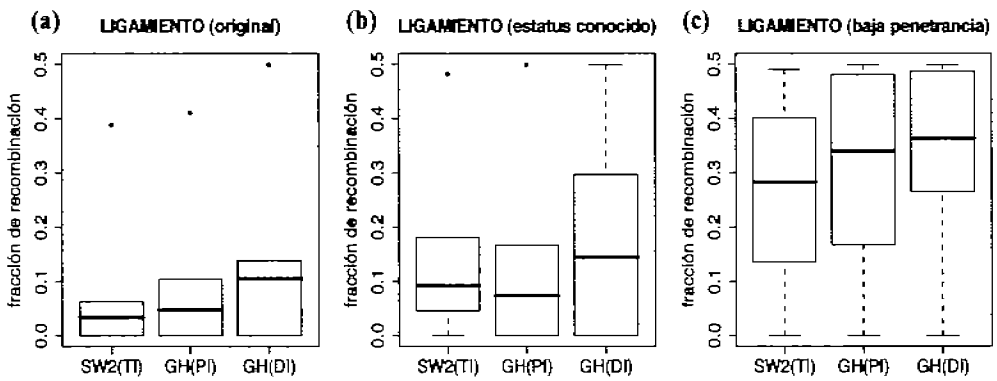

Figura 4.5: *En el modelo de ligamiento y en presencia de información faltante, comparación de la distribución muestral de la fracción de recombinación estimada a través de SW2 para la familia completa, y estimada a través de GH para la familia donde se descartaron individuos y la familia dividida en familias más pequeñas.*

La Figura 4.5 muestra la distribución de la fracción de recombinanción estimada para los mismos escenarios de la Figura 4.4. Es posible observar que utilizar la familia entera produce mejores estimaciones para el caso de la

familia "original" y "baja penetrancia", mientras que en el caso de la familia "estatus conocido" las estimaciones producidas por SW2 utilizando a todos los individuos son similares a los producidos por GH dividiendo la familia. El poder de cada programa para detectar ligamiento en este contexto se encuentra en el Cuadro 4.2.

4.4.3. Efecto de información faltante

Con el fin estudiar el efecto de información faltante sobre el valor de lod score, se realizaron comparaciones para cada una de las familias estudiadas, "original", "estatus conocido" y "baja penetrancia". El número de individuos en cada caso en particular se encuentra especificado en el Cuadro 4.1.

Para SW2, la comparación se realizó entre los resultados obtenidos en el grupo 1 (TC) y el grupo 2 (TI). La Figura 4.6(1a) y 4.6(1c) muestra que la información faltante no tiene un impacto fuerte en los casos de la familia "original" y "baja penetrancia". Sin embargo, como se puede observar en la Figura 4.6(1b), cuando todos los individuos faltantes tiene estatus conocido, la información faltante tiene un efecto considerable en contraste con las otras dos familias. Un comportamiento similar se puede observar con respecto a las estimaciones de la fracción de recombinación (Figura 4.7(1a-c)).

El efecto derivado de la presencia de información faltante se puede apreciar también en el Cuadro 4.2, donde el poder para detectar ligamiento pasa 0.83 a 0.75 en el caso de la familia "original" y de 0.82 a 0.43 en el caso de la familia "estatus conocido". Recordemos que la diferencia entre las familias "original" y "estatus conocido" es que en la primera, 19 de los 31 individuos que tienen información faltante también tiene estatus de la enfermedad desconocido, mientras que en la segunda todos los individuos tiene estatus de la enfermedad conocido. Por lo tanto, la diferencia en el efecto que tiene la presencia de información faltante en estas dos familias se debe al estatus de los individuos no disponibles. Desde el punto de vista teórico, este fenómeno tiene una explicación en términos de los componentes que intervienen, en uno y otro caso, en la función de verosimilitud. Si se compara el poder para detectar ligamiento en el escenario hipotético de información completa tenemos 0.83 y 0.82 para las familias "original" y "estatus conocido", respectivamente. Este último resultado pudiera sugerir que los individuos con estatus desconocido pero genotipo conocido proporcionan la misma información de ligamiento que los individuos afectados o sanos con genotipo conocido. Esta última observación apoya la estrategia de asignar estatus desconocido a todos los individuos sanos con el fin de reducir el efecto adverso de los

individuos sanos no-penetrantes en el caso de una enfermedad con baja penetrancia (análisis de sólo afectados). En otras palabras, los individuos con estatus desconocido no repercuten sobre el análisis sino al contrario, pueden ser de utilidad en situación particulares (ver Figuras 4.6(1a-c)).

Para GH, la comparación se realizó entre los resultados obtenidos en el grupo 3 (DC) y el grupo 4 (DI). Las gráficas correspondientes obtendidas por SW2 no se incluyeron ya que son muy similares a las obtenidas por GH. Como se observa en las Figuras 4.6(2a-c), la información faltante produce un consistente subestimación de los valores de lod score, siendo ésta más pronunciada en el caso de la familia "estatus conocido". Con respecto a las estimaciones de la fracción de recombinanción, la Figura 4.7(2a) y 4.7(2b) muestra que en las familias "original" y "estatus conocido" tiene el efecto de producir estimaciones menos precisas, mientras que en la familia "baja penetrancia" los resultados son similares (Figura 4.7(2c)).

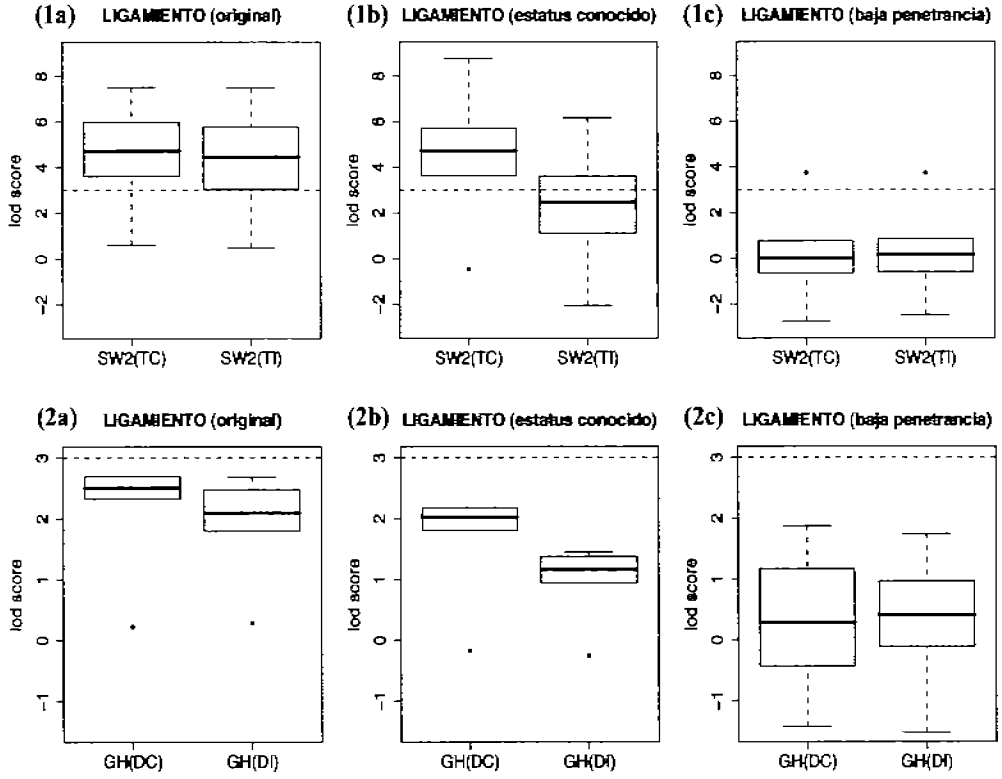

Figura 4.6: *En el modelo de ligamiento, comparación de los valores de lod score máximos, cuando se dispone de información completa e incompleta, obtenidos a través de SW2 para la familia completa y a través de GH para la familia donde se descartaron individuos.*

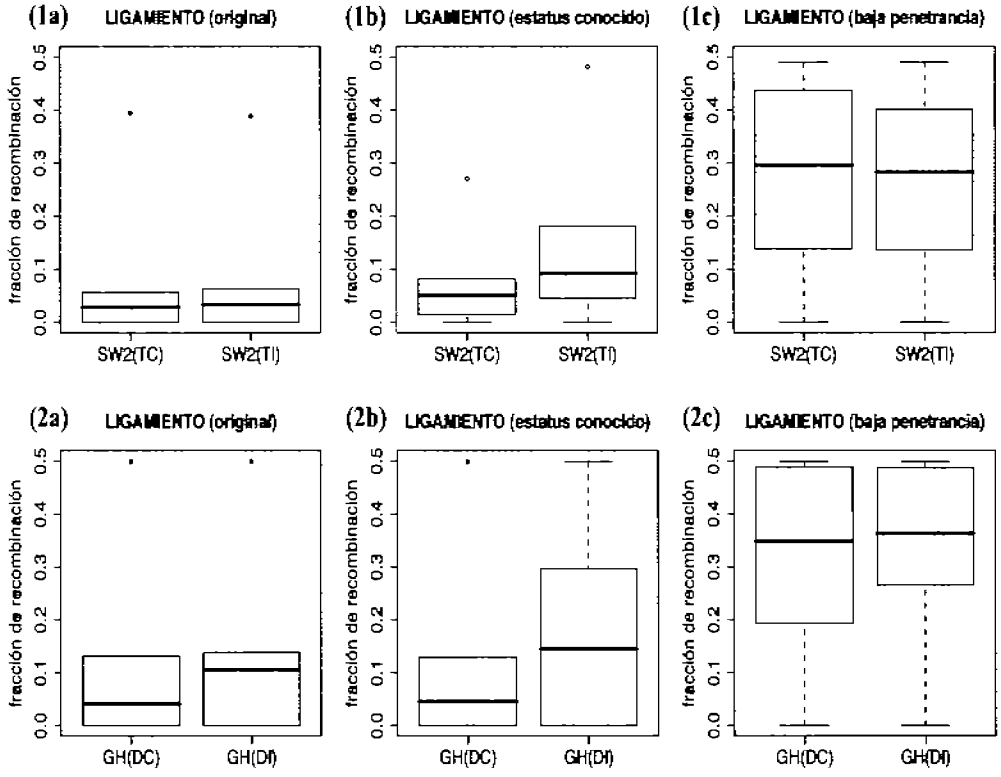

Figura 4.7: *En el modelo de ligamiento, comparación de la distribución muestral de la fracción de recombinación estimada, cuando se dispone de información completa e incompleta, obtenidos a través de SW2 para la familia completa y a través de GH para la familia donde se descartaron individuos.*

Capítulo 5

Análisis Bayesiano de ligamiento génetico

5.1. Planteamiento

En la actualidad, todos los programas disponibles que realizan un análisis de ligamiento genético requieren de la especificación previa de factores tales como la penetrancia, las frecuencias alélicas poblacionales de cada uno de los marcadores, y la frecuencia de la enfermedad en la población. Si los valores de estos parámetros no están disponibles deberán ser estimados. En la práctica, las frecuencias alélicas se estiman a partir de los genotipos de los

individuos fundadores disponibles, mientras que la penetrancia en muchos casos se determina de manera arbitraria, lo anterior no por negligencia sino por falta de información con respecto a los valores reales de este parámetro. En relación con la frecuencia de la enfermedad en la población, generalmente se obtiene con base en experiencia clínica.

Cuando el mecanismo de herencia de la enfermedad es desconocido, algunos autores utilizan lo que se conoce como "mod" score, que consiste en obtener el valor de lod score para una gran variedad de modelos que incluyen, por ejemplo, diferentes valores de penetrancia o diferentes mecanismos de transmisión (dominante, recesivo), y seleccionar aquel que arroja el valor de lod score más alto. Aún cuando algunos autores han mostrado que en algunas situaciones utilizar el modelo incorrecto no ocasiona una sobrevaluación del valor de lod score (Clerget-Dapoux et al. 1986; Greenberg, 1989; Elston, 1989), otros lo han cuestionado estudiando situaciones en donde sí se presenta una sobrevaluación (Weeks et al., 1990). Hodge y Elston (1994) estudian bajo qué condiciones es válido utilizar esta estrategia.

Un enfoque más realista consiste en suponer que, tal como la fracción de recombinación, estos factores también son parámetros desconocidos, y de esta forma también hacer inferencias sobre ellos. Para esto, desde el enfoque

Bayesiano, es necesario obtener la distribución final conjunta de todos los parámetros dadas las observaciones. Recuerde que los métodos de Monte Carlo vía cadenas de Markov, en particular el muestreo de Gibbs, permite generar muestras de una distribución multidimensional. Entonces, utilizando este algoritmo es posible generar muestras de la distribución final conjunta de interés, y de esta manera hacer inferencias sobre cualquiera de los parámetros involucrados. Los métodos de Monte Carlo vía cadenas de Markov en el contexto del análisis de ligamiento genético, además de ofrecer la posibilidad de estudiar familias extensas, también ofrecen una solución relativamente simple a un problema complejo en el campo de la estadística Bayesiana.

Thomas y Cortessis (1992) proponen un esquema para realizar un análisis Bayesiano de ligamiento genético de dos puntos, es decir, inferir la posición del locus de la enfermedad relativa a un sólo marcador, donde éste es bialélico. Para esto, utilizan MCMC, particularmente el muestreo de Gibbs. Los parámetros que estos autores consideran como desconocidos son la fracción de recombinación, la penetrancia, las frecuencias alélicas tanto del marcador como del locus de la enfermedad, y los genotipos de los individuos.

En un enfoque Bayesiano es posible incorporar información previa a través de la distribución inicial, lo que permite reflejar conocimiento que se tenga

acerca del parámetro. Por ejemplo, para aquellos parámetros que, bajo un enfoque frecuentista están fijos y que en la práctica se estiman de manera rudimentaria, estas estimaciones pueden ser incorporadas en la distribución inicial del parámetro correspondiente. Con respecto a la fracción de recombinación, Smith (1959) asume una distribución inicial mixta para θ, es decir $p(\theta) = 1/22$ para $\theta < 1/2$, y $p(\theta) = 21/22$ para $\theta = 1/2$. En otras palabras, bajo esta distribución inicial, la probabilidad de ligamiento entre un marcador cualquiera y el locus de la enfermedad es de 0.045, y por lo tanto la probabilidad de que no estén ligados es de 0.954.

Thomas y Cortessis (1992) consideran únicamente el escenario de un marcador bialélico, ya que sólo en este caso está demostrado que se cumple la condición de irreducibilidad necesaria cuando se utiliza MCMC (Sheehan y Thomas, 1993). Cuando se trata de un marcador multialélico, como se vio en la sección 3.2.4, no es posible garantizar irreducibilidad y en torno a este problema varios autores han propuesto diferentes soluciones. Entre ellos se encuentran Sobel y Lange (1996) que proponen una caminata aleatoria, que realiza transiciones múltiples en cada paso, sobre el espacio de gráficas genéticas de descendencia.

5.2. Objetivo

Implementar y extender la propuesta de Thomas y Cortessis (1992) para el caso de un marcador multialélico. Es decir, implementar en un programa de cómputo un análisis Bayesiano de ligamiento genético de dos puntos, para un marcador multialélico, utilizando métodos de Monte Carlo vía cadenas de Markov, particulamente el algoritmos de Metropolis-Hastings y muestreo de Gibbs, considerando como parámetros desconocidos, además de la fracción de recombinación, la penetrancia, los genotipos de cada uno de los individuos, así como las frecuencias alélicas del marcador o marcadores y la frecuencia poblacional de la enfermedad.

5.3. Metodología

5.3.1. Notación

Considere una familia con n individuos, y sea $i = 1, 2, \ldots, n$ el índice correspondiente a cada uno de estos individuos. Si el individuo i es no-fundador, entonces sus padres serán denotados por M_i y P_i, correspondiendo a su madre y padre, respectivamente. El número total de no-fundadores y fundadores en la familia será denotado por n_{nf} y n_f, respectivamente. Sea

$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ y $\mathbf{m} = (m_1, m_2, \ldots, m_n)$ el vector de fenotipos del locus de la enfermedad y el marcador, respectivamente.

Para un locus con m alelos, el número de posibles genotipos que se pueden formar es $m(m+1)/2$. Sea $\mathbf{g_x} = (g_{x_1}, g_{x_2}, \ldots, g_{x_n})$ y $\mathbf{g_m} = (g_{m_1}, g_{m_2}, \ldots, g_{m_n})$ el vector de genotipos correspondientes al locus de la enfermedad y el marcador, respectivamente.

En el caso que nos concierne, se asume que el gen de la enfermedad consta de dos alelos d y D, y por lo tanto tres posibles genotipos dd, dD y DD, donde D representa el alelo que desencadena la enfermedad. Por otra parte, el fenotipo de la enfermedad está categorizado como sano y afectado, es decir $x_i = 1$ si el individuo i es sano y $x_i = 2$ si el individuo i es afectado. Recordemos que la relación entre el fenotipo y el genotipo de un gen se describe a través de la penetrancia, por lo tanto denotemos por $\mathrm{f} = (f_0, f_1, f_2)$ al vector de penetrancias correspondiente al locus de la enfermedad, donde f_0, f_1 y f_2 corresponde a la probabilidad de que un individuo con genotipo dd, dD y DD, respectivamente, esté afectado. Por último, se denota por r a la frecuencia del alelo D en la población, e implícitamente la frecuencia del alelo d es $(1 - r)$.

Con respecto al marcador, se asume que su fenotipo es medido sin error

y que todos los alelos que lo componen son codominantes. Suponga ahora que el marcador tiene m alelos, los cuales serán denotados por $a_1, a_2, \ldots, a_m$, donde el vector $\mathbf{p_a} = (p_{a_1}, p_{a_2}, \ldots, p_{a_m})$ contiene las frecuencias de cada uno de los alelos en la población y satisface $\sum_{i=1}^{m} p_{a_i} = 1$.

Por último, sea θ la fracción de recombinación, y por lo tanto el parámetro de mayor interés.

Un supuesto importante es que los dos loci, el locus de la enfermedad y el marcador, están en equilibrio de ligamiento, es decir, que la frecuencia de los genotipos correspondientes a un locus es independiente de la frecuencia de los genotipos del otro locus.

Resumiendo, los factores considerados como parámetros desconocidos en este análisis son la fracción de recombinación, θ; la penetrancia, $\mathbf{f} = (f_0, f_1, f_2)$; la frecuencia alélica del locus de la enfermedad, r; las frecuencias alélicas del marcador, $\mathbf{p_a} = (p_{a_1}, p_{a_2}, \ldots, p_{a_m})$; y los genotipos de los individuos, tanto para el locus de a enfermedad como para el marcador, $\mathbf{g_x} = (g_{x_1}, g_{x_2}, \ldots, g_{x_n})$ y $\mathbf{g_m} = (g_{m_1}, g_{m_2}, \ldots, g_{m_n})$. Finalmente denotemos por η al conjunto de todos estos parámetros, es decir $\eta = (\theta, \mathbf{f}, r, \mathbf{p_a}, \mathbf{g_x}, \mathbf{g_m})$.

5.3.2. Distribuciones iniciales

Dado que el trabajo objeto del presente capítulo está basado en el esquema proporcionado por Thomas y Cortessis (1992), las distribuciones iniciales utilizadas en este trabajo son las sugeridas por estos autores con las modificaciones pertinentes.

Para la fracción de recombinación, estos autores se basan parcialmente en la sugerencia de Smith (1959) en el sentido de que utilizan una distribución inicial mixta para θ. En este caso, la probabilidad inicial de que los dos loci no estén ligados, es decir $\theta = 1/2$, es de $21/22 = 0.954$, y se utiliza una distribución Beta (con hiperparámetros τ_1 y τ_2) para valores de $\theta < 1/2$. Lo anterior se puede resumir de la siguiente manera,

$$p(\theta) = w\mathbb{I}_{\{0.5\}}(\theta) + (1-w)2Beta(2\theta|\tau_1, \tau_2),$$

donde $w = 21/22$ y $\mathbf{I}_{\{0.5\}}(\theta)$ es una función indicadora que vale 1 cuando $\theta = 0.5$ y vale 0 en cualquier otro caso. Por otro lado, la notación $Beta(2\theta|\tau_1, \tau_2)$ se refiere a la distribución Beta cuyas características se describen brevemente en el apéndice B.

Por otro lado, se asume que cada uno de los elementos que componen el vector de penetancias tiene también una distribución Beta con hiper-

parámetros propios, es decir $p(f_0) = Beta(\alpha_1, \alpha_2)$, $p(f_1) = Beta(\beta_1, \beta_2)$, y $p(f_2) = Beta(\delta_1, \delta_2)$. De la misma manera se utiliza una distribución Beta como distribución inicial para la frecuencia poblacional del alelo D denotada por r, es decir $p(r) = Beta(\gamma_1, \gamma_2)$.

Por último, para las frecuencias alélicas correspondientes al marcador se utilizará como distribución inicial una distribución Dirichlet con hiper-parámetros $\rho = (\rho_1, \rho_2, \ldots, \rho_m)$. Es decir,

$$p(p_{a_1}, p_{a_2}, \ldots, p_{a_m}) = Dirichlet(\rho_1, \rho_2, \ldots, \rho_m).$$

Una descripción breve de las características de una distribución Dirichlet se puede encontrar en el apéndice B.

Cabe destacar que, dependiendo de los valores que tomen los hiper-parámetros de la distribuciones especificadas, éstas pueden tomar una gran variedad de formas, y es a través de estos valores que se va a reflejar el conocimiento que se tiene acerca de cada uno de los parámetros de interés.

5.3.3. Distribuciones condicionales completas

Sea T el número de meiosis en la familia, es decir $2n_{nf}$, y sea t el número de recombinaciones ocurridas en estas $2n_{nf}$ meiosis. Entonces la distribución

condicional completa de θ tiene la siguiente forma,

$$p(\theta|\mathbf{f}, r, \mathbf{p_a}, \mathbf{g_x}, \mathbf{g_m}, \mathbf{x}, \mathbf{m}) = w^{\star}\mathbb{I}_{\{0.5\}}(\theta) + (1 - w^{\star})p^{\star}(\theta|\tau_1, \tau_2, t, T)$$

donde

$$w^{\star} = \frac{w(0.5)^T}{w(0.5)^T + (+w)\int_0^{0.5} \theta^t(1-\theta)^{T-t}2Beta(2\theta|\tau_1, \tau_2)\, d\theta}$$

y

$$p^{\star}(\theta|\tau_1, \tau_2, t, T) = \frac{\theta^t(1-\theta)^{T-t}2Beta(2\theta|\tau_1, \tau_2)}{\int_0^{0.5} \theta^t(1-\theta)^{T-t}2Beta(2\theta|\tau_1, \tau_2)\, d\theta}$$

Note que $w^{\star}$ representa una probabilidad actualizada de que no haya ligamiento entre los dos loci. Brevemente el algoritmo que obtiene muestras de θ de la distribución condicional completa anterior procede como sigue,

1. Dado una configuración de genotipos compatible, obtener T, t y $w^{\star}$.

2. Generar una observación u utilizando una distribución uniforme en el intervalo $(0, 1)$,

   a) si $u \leq w^{\star}$ entonces $\theta = 0.5$,

   b) si $u > w^{\star}$ entonces generar θ de $p^{\star}(\theta|\tau_1, \tau_2, t, T)$ utilizando el algoritmo de Metropolis-Hasting.

Note que para obtener los valores de de t y T solo se necesita la configuración de genotipos correspondientes al marcador y a la enfermedad, por lo tanto $p(\theta|\mathbf{g_x}, \mathbf{g_m}, \mathbf{x}, \mathbf{m})$.

Con respecto al vector de penetancias, $\mathbf{f}$, considere primero el caso de f_0. Sea n_0 el número de individuos afectados con genotipo *dd* y sea N_0 el número de individuos con estatus de la enfermedad conocido y genotipo *dd*. Entonces la distribución condicional completa correspondiente a la penetrancia f_0 tiene una distribución Beta con hiperparámetros $(\alpha_1 + n_0, \alpha_2 + N_0 - n_0)$. Note que esta distribución sólo depende del genotipo de los individuos para el locus de la enfermedad, lo que se puede expresar de la siguiente manera, $p(\mathbf{f}|\mathbf{g_x}, \mathbf{x}, \mathbf{m}) = Beta(\alpha_1 + n_0, \alpha_2 + N_0 - n_0)$. De la misma manera se obtienen las distribuciones condicionales completas para f_1 y f_2.

Por otro lado, sea n_D el número de alelos D en los genotipos de los individuos fundadores, por lo tanto la distribución condicional completa correspondiente al parámetro r tiene una distribución Beta con hiperparámetros $(\gamma_1 + n_D, \gamma_2 + F - n_D)$. De la misma manera que en el caso de la penetrancia, esta distribución únicamente depende del genotipo de los individuos para el locus de la enfermedad, es decir $p(r|\mathbf{g_x}, \mathbf{x}, \mathbf{m}) = Beta(\gamma_1 + n_D, \gamma_2 + F - n_D)$.

El procedimiento para obtener la distribución condicional completa corres-

pondiente a las frecuencias alélicas del marcador es similar al utilizado en el caso del parámetro r. Sea n_{a_1}, n_{a_2},...,n_{a_m} el número de alelos a_1, a_2,...,a_m en los genotipos de los individuos no-fundadores en el marcador, respectivamente. Entonces la distribución condicional completa en este caso se puede expresar como $p(\mathbf{p_a}|\mathbf{g_m}, \mathbf{x},) = Dirichlet(\rho_1 + n_{a_1}, \rho_2 + n_{a_2}, \ldots, \rho_m + n_{a_m})$.

Para generar configuraciones del genotipo de los individuos correspondiente al locus de la enfermedad se utiliza el algoritmo sugerido por Lange y Matthysse (1989), mientras que para generar configuraciones del genotipo correspondiente al marcador se utiliza el algoritmo propuesto de Sobel y Lange (1996). Este último es una generalización del primero aplicable al caso de un marcador multialélico, ambos se encuentran brevemente explicados en la sección 3.2.4, y en ambos casos se utiliza el algoritmo de Metropolis-Hastings.

Cabe mencionar que en ocasiones es posible obtener más de una configuración alélica compatible con una gráfica de descendencia genética (ver Figura 3.2), por lo que en cada iteración una de ellas es seleccionada al azar asumiendo que todas son igualmente probables.

5.3.4. Aplicación de muestreo de Gibbs

Incorporando todos los elementos en el contexto del algoritmo general, que es muestreo de Gibbs, tenemos que el objetivo es obtener muestras de la distribución final conjunta de todos los parámetros dado los datos disponibles, $p(\theta, \mathbf{f}, r, \mathbf{p_a}, \mathbf{g_x}, \mathbf{g_m}|\mathbf{x}, \mathbf{m})$, de la siguiente manera.

Sea $\eta^{(0)} = (\theta^{(0)}, \mathbf{f}^{(0)}, r^{(0)}, \mathbf{p_a}^{(0)}, \mathbf{g_x}^{(0)}, \mathbf{g_m}^{(0)})$ valores iniciales arbitrarios,

1. generar $\mathbf{g_m}^{(1)}$ de $p(\mathbf{g_m}|\mathbf{p_a}^{(0)}, \mathbf{g_x}^{(0)}, \theta^{(0)}, \mathbf{x}, \mathbf{m})$;

2. generar $\mathbf{p_a}^{(1)}$ de $p(\mathbf{p_a}|\mathbf{g_m}^{(1)}, \mathbf{x}, \mathbf{m})$;

3. generar $\theta^{(1)}$ de $p(\theta|\mathbf{g_m}^{(1)}, \mathbf{g_x}^{(0)}, \mathbf{x}, \mathbf{m})$;

4. generar $r^{(1)}$ de $p(r|\mathbf{g_x}^{(0)}, \mathbf{x}, \mathbf{m})$;

5. generar $\mathbf{f}^{(1)}$ de $p(\mathbf{f}|\mathbf{g_x}^{(0)}, \mathbf{x}, \mathbf{m})$;

6. generar $\mathbf{g_x}^{(1)}$ de $p(\mathbf{g_x}|\theta^{(1)}, \mathbf{g_m}^{(1)}, \mathbf{f}^{(1)}, r^{(1)}, \mathbf{x}, \mathbf{m})$.

En este punto una iteración queda completa, después de t iteraciones se tiene $(\theta^{(t)}, \mathbf{f}^{(t)}, r^{(t)}, \mathbf{p_a}^{(t)}, \mathbf{g_x}^{(t)}, \mathbf{g_m}^{(t)})$. Entonces, conforme $t \to \infty$

$$(\theta^{(t)}, \mathbf{f}^{(t)}, r^{(t)}, \mathbf{p_a}^{(t)}, \mathbf{g_x}^{(t)}, \mathbf{g_m}^{(t)}) \xrightarrow{f} (\theta, \mathbf{f}, r, \mathbf{p_a}, \mathbf{g_x}, \mathbf{g_m}) \sim p(\theta, \mathbf{f}, r, \mathbf{p_a}, \mathbf{g_x}, \mathbf{g_m}|\mathbf{x}, \mathbf{m}).$$

En otras palabras, para t suficientemente grande la distribición de $(\theta^{(t)}, \mathbf{f}^{(t)}, r^{(t)}, \mathbf{p_a}^{(t)}, \mathbf{g_x}^{(t)}, \mathbf{g_m}^{(t)})$ es $p(\theta, \mathbf{f}, r, \mathbf{p_a}, \mathbf{g_x}, \mathbf{g_m}|\mathbf{x}, \mathbf{m})$.

En la sección 3.2.4 correspondiente a Métodos de Monte Carlo vía cadenas de Markov, se describieron dos algoritmos, el algoritmo de Metropolis-Hastings y el muestreo de Gibbs. Este último nos proporciona el algoritmo general que nos permitirá obtener muestras de la distribución final conjunta de todos los parámetros dadas las observaciones. Sin embargo, el algoritmo de Metropolis-Hastings se utiliza particularmente en los pasos 1, 3 y 6, para generar muestras de los genotipos de los individuos correspondientes al marcador y a la enfermedad, así como para generar muestras de θ. Cabe descatar que en los pasos 3 y 6 se utilizan 100 iteraciones del algoritmo de Metropolis-Hastings por cada iteración del muestreo de Gibbs, mientras que para generar genotipos correspondientes al marcador (paso 1) se utiliza únicamente una iteración del algoritmo de Metropolis-Hastings en cada iteración del muestreo de Gibbs.

Los valores iniciales de los parámetros, θ, $\mathbf{f}$, r y $\mathbf{p_a}$, se pueden obtener aleatoriamente a partir de las distribuciones iniciales correspondientes. Por otro lado, como configuración inicial de los genotipos correspondientes al locus de la enfermedad, es posible utilizar la configuración canónica sugerida por Lange y Matthysse (1989). La configuración canónica referida consiste en que todos los individuos son heterocigotos, donde el alelo materno del indivi-

duo proviene de la abuela materna y el alelo paterno del abuelo paterno. Por su parte, para obtener una configuración incial de genotipos correspondientes al marcador se utiliza el algoritmo conocido como eliminación de genotipos (*genotype-elimination*), Sobel et al. (1995) e implementado en el programa SimWalk2.

La implementación de esta metodología se realizó utilizando el lenguaje de programación conocido como FORTRAN.

5.4. Resultados

Con el fin de evaluar la implementación objeto de presente capítulo, se utilizaron tres ejemplos. El primero con evidencia contundente de ligamiento obteniendo un valor de lod score de 7.02 (ejemplo A), el segundo con evidencia sugestiva de ligamiento con un valor de lod score de 2.50 (ejemplo B), y por último, un tercero ejemplo con evidencia de ausencia de ligamiento con un valor de lod score negativo de -6.57 (ejemplo C). En los tres casos el valor de lod score se obtuvo utilizando el programa SimWalk2.

Los tres ejemplos surgen de dos aplicaciones con datos de marcadores y fenotipos reales. El ejemplo A corresponde a una familia estudiada por

Leppert et. al (1986) donde a través de ligamiento genético se asoció por primera vez la hipercolesterolemia con el receptor LDL (por sus siglas en inglés, *low-density lipoprotein*). Esta familia fue utilizada también por Thomas y Cortessis (1992) para introducir el muestreo de Gibbs en el contexto de análisis de ligamiento genético y se utiliza también en este trabajo para fines comparativos con los resultados de estos autores. Dicha familia consta de 58 individuos distribuidos en cinco generaciones, de los cuales 22 presentan el fenotipo considerado como afectado.

El ejemplo B corresponden a una familia estudiada por Canizales-Quinteros et al. (2003) donde a través de ligamiento genético se sugiere la participación de una región del cromosoma 6 con niveles elevados de HDL (por sus siglas en inglés, *high-density lipoprotein*). Esta familia consta de 47 individuos distribuidos en cinco generaciones, de los cuales 10 presentan el fenotipo considerado como afectado. El valor de lod score máximo se encuentra localizado entre los marcadores D6S1960 y D6S1662. Para fines de este ejemplo se utilizó el primero, es decir el marcador D6S1960, que consta de 4 alelos. Por otro lado, para el ejemplo C se utilizó esta misma familia y el marcador D6S1038 (4 alelos), para el cual se obtuvo un valor de lod score negativo de -6.57.

Para cada uno de los tres ejemplos descritos se realizó un análisis Bayesiano de ligamiento genético utilizando la metodología descrita en la sección anterior. El conocimiento inicial que se tiene acerca de cada uno de los parámetros se ve reflejado a través de los valores que toman los hiperparámetros de cada una de las distribuciones inciales. En el Cuadro 5.1 se especifican dichos valores para cada uno de los tres ejemplos. Note que en todos los casos el supuesto inicial es que la enfermedad sigue un mecanismo de herencia dominante, ya que $f_1 = f_2$, con un valor esperado de penetrancia del 90 %. En los tres ejemplos se asume también que la frecuencia esperada del alelo que causa la enfermedad en la población es de alrededor de 1 %. Los valores de los hiperparámetros correspondientes a las frecuencias alélicas del marcador surgen de las estimaciones iniciales utilizadas para correr el método de lod score.

Parámetro	Ejemplo A	Ejemplo B	Ejemplo C
θ	$\tau = (1,1)$	$\tau = (1,1)$	$\tau = (1,1)$
f_0	$\alpha = (99,1)$	$\alpha = (99,1)$	$\alpha = (99,1)$
f_1	$\beta = (10,90)$	$\beta = (10,90)$	$\beta = (10,90)$
f_2	$\delta = (10,90)$	$\delta = (10,90)$	$\delta = (10,90)$
r	$\gamma = (99,1)$	$\gamma = (99,1)$	$\gamma = (99,1)$
$\mathbf{p_a} = (p_{a_1}, p_{a_2}, \ldots, p_{a_m})$	$\rho = (80,20)$	$\rho = (20,60,10,10)$	$\rho = (9,41,41,9)$
$\mathbf{g_x} = (g_{x_1}, g_{x_2}, \ldots, g_{x_n})$	configuración canónica		
$\mathbf{g_m} = (g_{m_1}, g_{m_2}, \ldots, g_{m_n})$	algoritmo eliminación de genotipos		

Cuadro 5.1: *Valores de los hiperparámetros correspondientes a las distribuciones iniciales*

El muestreo de Gibbs descrito en la sección anterior genera muestras de cada uno de los parámetros de interés. De acuerdo con la teoría presentada en la sección 3.2.4, después de un cierto número de iteraciones t suficientemente grande las muestras generadas pueden considerarse como muestras de la distribución de interés. Por lo tanto, el número total de muestras se puede dividir en dos grupos: aquellas que pertenecen al periodo de calentamiento y que por lo tanto serán desechadas; y aquellas que en teoría servirán para ha-

cer inferencia sobre los parámetros. La duración del periodo de calentamiento dependerá de la velocidad de convergencia de la cadena de Markov a la distribución de interés. La velocidad de convergencia a su vez dependerá de los valores iniciales. Una estrategia para verificar que la cadena ha convergido es utilizar lo que se conoce como promedio ergódico. Este promedio no es otra cosa que el promedio aritmético de las muestras conforme va avanzando la cadena, es decir, en la *t-ésima* iteración el promedio ergódico corresponde al promedio aritmético de las t muestras generadas hasta ese momento. Se asume que la cadena de Markov ha convergido una vez que el promedio ergódico se ha estabilizado. La base de este procedimiento se discutió en la sección 3.2.4. En la Figura 5.1 se presentan los promedios ergódicos correspondientes a la fracción de recombinación (θ) para cada uno de los tres ejemplos considerados, donde la línea discontinua vertical indica la longitud del periodo de calentamiento en cada caso.

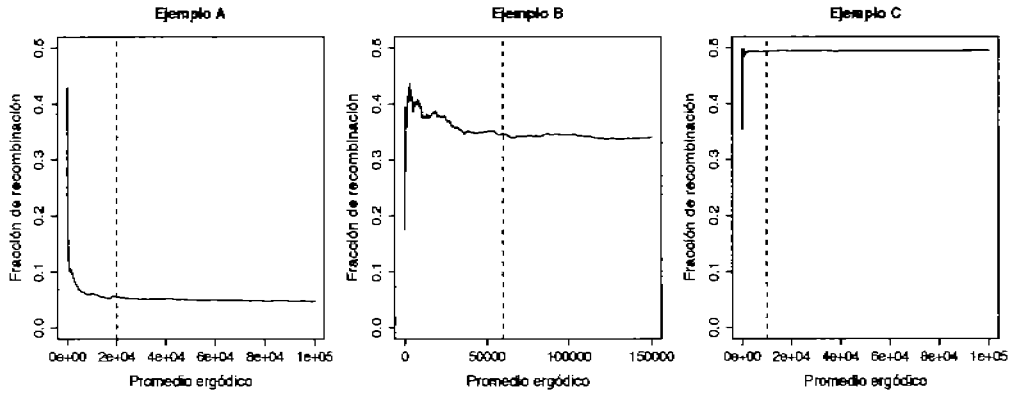

Figura 5.1: *Promedio ergódico correspondiente a la fracción de recombinación para cada uno de los tres ejemplos considerados. La línea discontinua vertical indica la longitud del periodo de calentamiento.*

Adicionalmente, es deseable que las muestras que serán utilizadas para hacer inferencias cumplan con la característica de ser independientes entre sí. Debido a esto es necesario evaluar si existe correlación entre las muestras. Una estrategia es utilizar una medida de correlación conocida como autocorrelación. La correlación nos da una medida de la dependencia entre dos series de valores. La autocorrelación representa la dependencia entre una serie de valores con ella misma. En la Figura 5.2 se presentan las gráficas de autocorrelación para cada uno de los tres ejemplos, donde la línea discontinua vertical indica el número de iteraciones necesarias para reducir la

autocorrelación a niveles aceptables.

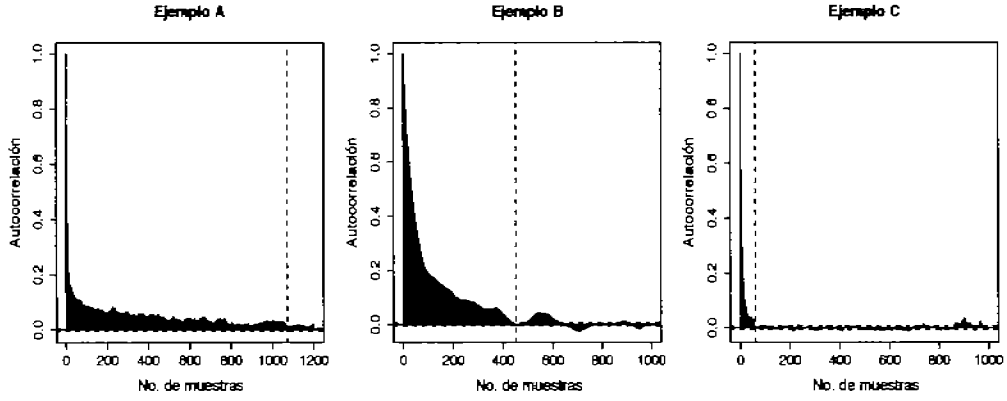

Figura 5.2: *Autocorrelación correspondiente a las muestras generadas para la fracción de recombinación para cada uno de los tres ejemplos considerados. La línea discontinua vertical indica el número de iteraciones necesarias para reducir la autocorrelación a niveles aceptables.*

Como se observa en la Figura 5.2 el número de muestras consecutivas que están correlacionadas entre sí es del orden de 1070 para el ejemplo A, de 450 para el ejemplo B y de 60 para el ejemplo C. Entonces, el procedimiento para seleccionar las muestras que serán utilizadas para hacer inferencias sobre el parámetro de interés consiste en lo siguiente: del número total de muestras generadas, se desechan las correspondientes al periodo de calentamiento y

posteriormente se toma una muestra cada 1070 iteraciones (ejemplo A) con la finalidad de reducir la autocorrelación. En el Cuadro 5.2 se resume para cada uno de los tres ejemplos, el número total de muestras generadas, así como el número de muestras disponibles utilizadas para hacer inferencias. En la Figura 5.3 se presentan las gráficas de autocorrelación para cada uno de los tres ejemplos, donde se puede apreciar que la autocorrelación entre las muestras ha desaparecido.

Muestras	**Ejemplo A**	**Ejemplo B**	**Ejemplo C**
Generadas	1,200,000	1,000,000	1,000,000
Calentamiento	20,000	60,000	10,000
Autocorrelación	1,070	450	60
Inferencia	1,102	2,088	16,500

Cuadro 5.2: *Número de muestras*

Thomas y Cortessis (1992) proponen un esquema diferente. Estos autores sugieren correr varias cadenas con diferentes valores iniciales para minimizar la influencia de éstos. Específicamente, para el ejemplo A, ellos utilizan 100 cadenas con valores iniciales distintos, cada una con un total de 120 iteraciones, 20 de calentamiento y 100 para hacer inferencias. Al final utilizan

indistintamente las 10,000 muestras generadas a través de las 100 cadenas y con ellas obtienen sus resultados. Sin embargo, de acuerdo con el Cuadro 5.2, en nuestro caso el periodo de calentamiento no coincide ni cercanamente con el que ellos proponen. Las principales diferencias entre el algoritmo propuesto por Thomas y Cortessis (1992) con el implementado en nuestro programa son: la especificación de la distribución inicial de la penetrancia, así como la generación de los genotipos de la enfermedad y el marcador. Estos autores actualizan ambos loci de manera conjunta, mientras que en nuestro caso se actualizan de manera individual, lo anterior dado que cuando se trata de un marcador multialélico no es posible actualizar ambos loci de manera conjunta, tal y como ellos lo proponen, ya que nos encontramos con el problema de reducibilidad de la cadena.

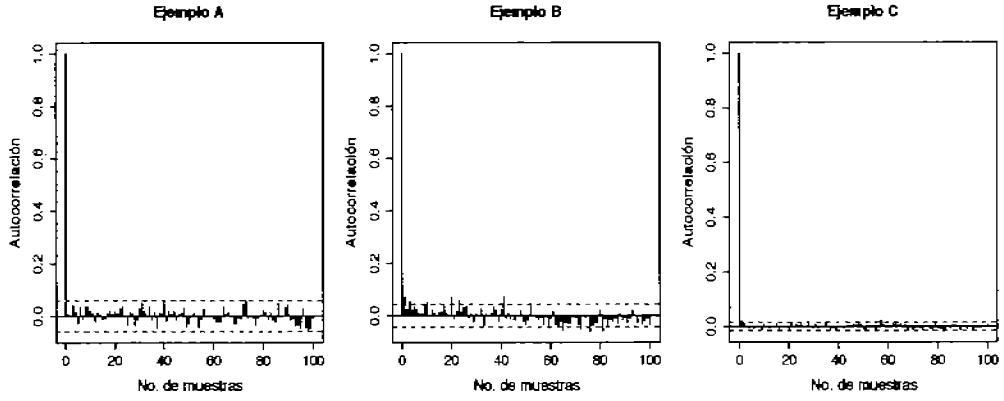

Figura 5.3: *Autocorrelación correspondiente a las muestras generadas para la fracción de recombinación para cada uno de los tres ejemplos considerados, tomando únicamente las muestras generadas de acuerdo al esquema que se presenta en el Cuadro 5.2.*

Finalmente, utilizando las muestras restantes en cada ejemplo es posible hacer inferencias sobre la fracción de recombinación. De particular interés es la probabilidad final de ligamiento. Esta probabilidad se puede estimar a través del número de θ's que son estrictamente menores que 0.5 entre el número total de muestras en cada caso. El Cuadro 5.3 contiene la estimación de dicha probabilidad para cada uno de los tres ejemplos considerados.

	Ejemplo A	**Ejemplo B**	**Ejemplo C**
$P(H_L) = P(0 \leq \theta < 0.5)$	100 %	49.14 %	3.56 %

Cuadro 5.3: *Probabilidad final de la hipótesis de ligamiento*

Recuerde que la probabilidad inicial de ligamiento para los tres ejemplos fue de $1/22 = 4.5\,\%$. Los resultados del Cuadro 5.3 indican que en el ejemplo A, para efectos prácticos, tenemos absoluta certeza de la presencia de ligamiento entre la enfermedad y el marcador estudiado. Mientras que, en el ejemplo C, tenemos bastante certeza de que no hay ligamiento entre el marcador en cuestión y la enfermedad. Por su parte, en el ejemplo B, se puede decir que existe evidencia de la presencia de ligamiento, ya que observamos un incremento considerable de la probabilidad de ligamiento. Sin embargo, existe también un grado considerable de incertidumbre. Un resultado de este tipo motiva la cautela cuando se trata de reportar un hallazgo y proporciona confianza para continuar estudiando la región correspondiente.

Si ahora el interés es hacer inferencias sobre el valor de la fracción de recombinación, desde un punto de vista Bayesiano, la información correspondiente se encuentra en la distribución final del parámetro. Sin embargo, recuerde que en el caso particular de la fracción de recombinación, se trata

de una distribución de probabilidad mixta. Por esta razón, en este caso es necesario utilizar la distribución final de θ condicional en $\theta < 0.5$. En la Figura 5.4 se presentan los histogramas correspondientes a los ejemplos A y B, ya que son estos los dos ejemplos en donde hay evidencia de ligamiento.

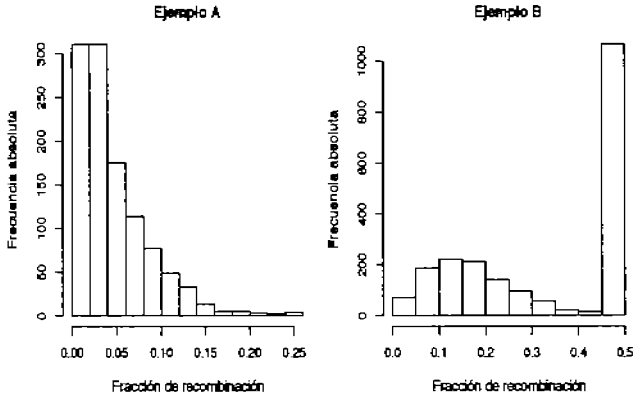

Figura 5.4: *Distribución final de la fracción de recombinación.*

En el ejemplo A, dado que el 100 % de las muestras presentaron un valor de θ menor a 0.5, la distribución final θ condicional en $\theta < 0.5$ es la que se observa en la Figura 5.4. Sin embargo, en el ejemplo B únicamente el 50 % de las muestras correspondieron a valores de $\theta < 0.5$, por lo tanto con base en la distribución de este 50 % es que sería posible hacer inferencias sobre el valor del parámetro. En este caso, la distribución final de θ condicional

en $\theta < 0.5$ es proporcional a la parte de la distribución que corresponde a valores de θ menores a 0.5 en la Figura 5.4. Considerando lo anterior, si se desea una estimación puntual sobre el valor de la fracción de recombinación entonces el valor más plausible es aquel en donde la distribución final de θ condicional en $\theta < 0.5$ alcanza su valor máximo.

Por otro lado, un análisis similar al realizado para la fracción de recombinación se debe llevar a cabo para cada uno de los parámetros involucrados. De hecho el periodo de calentamiento se debe establecer con base en la convergencia de la cadena a nivel global. Sin embargo, como no es posible evaluar la convergencia de la cadena para todos los parámetros de manera conjunta es que se realiza de manera individual parámetro por parámetro. El parámetro más inestable y correlacionado determinará el periodo de calentamiento, así como el número de iteraciones entre muestra y muestra necesarias para reducir la autocorrelación. En los tres ejemplos presentados en esta sección el parámetro más inestable y correlacionado fue la fracción de recombinación. En las Figuras 5.5, 5.6 y 5.7 se presentan, únicamente para el ejemplo B, los promedios ergódicos, así como los histogramas correspondientes a la penetrancia, las frecuencias alélicas correspondientes al marcador y al locus de la enfermedad, respectivamente.

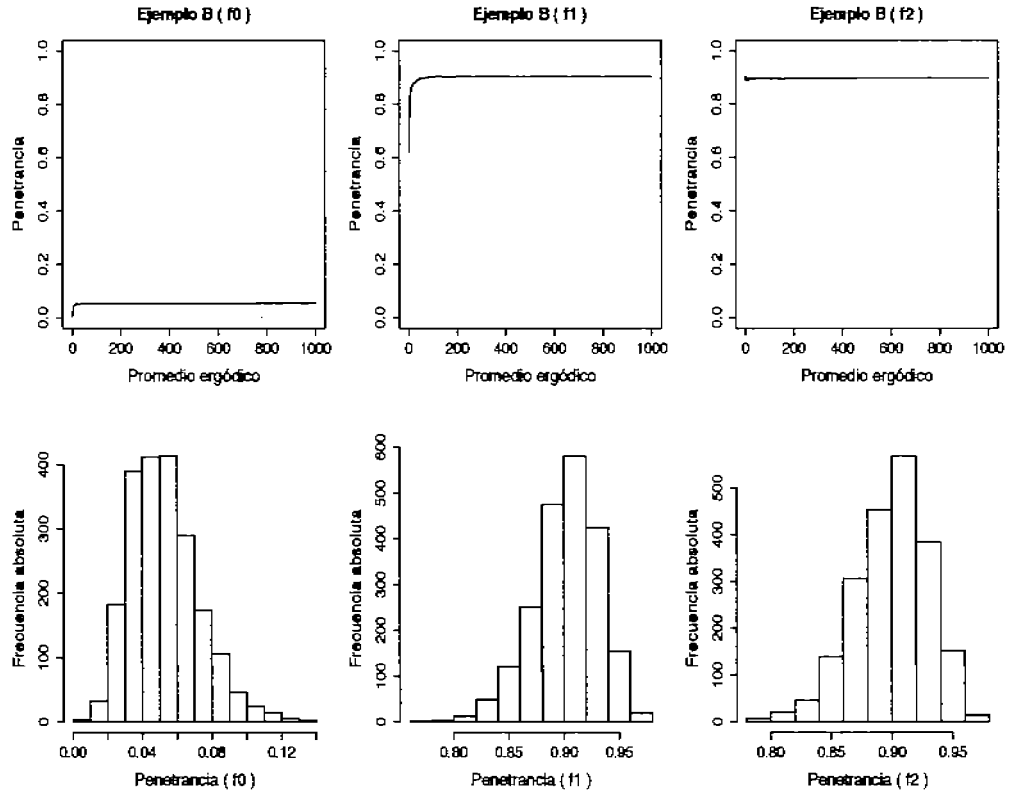

Figura 5.5: *Promedios ergódicos y distribución final de la penetrancia correspondiente al ejemplo B.*

Recuerde que el número de alelos del marcador estudiado en el ejemplo B es de cuatro. En la Figura 5.7 se presentan únicamente los promedio ergódicos y las distribuciones finales correspondientes a tres alelos. Esto se debe a toda la información relativa al cuarto alelo se puede obtener a través de la

información de los primeros tres primeros alelos.

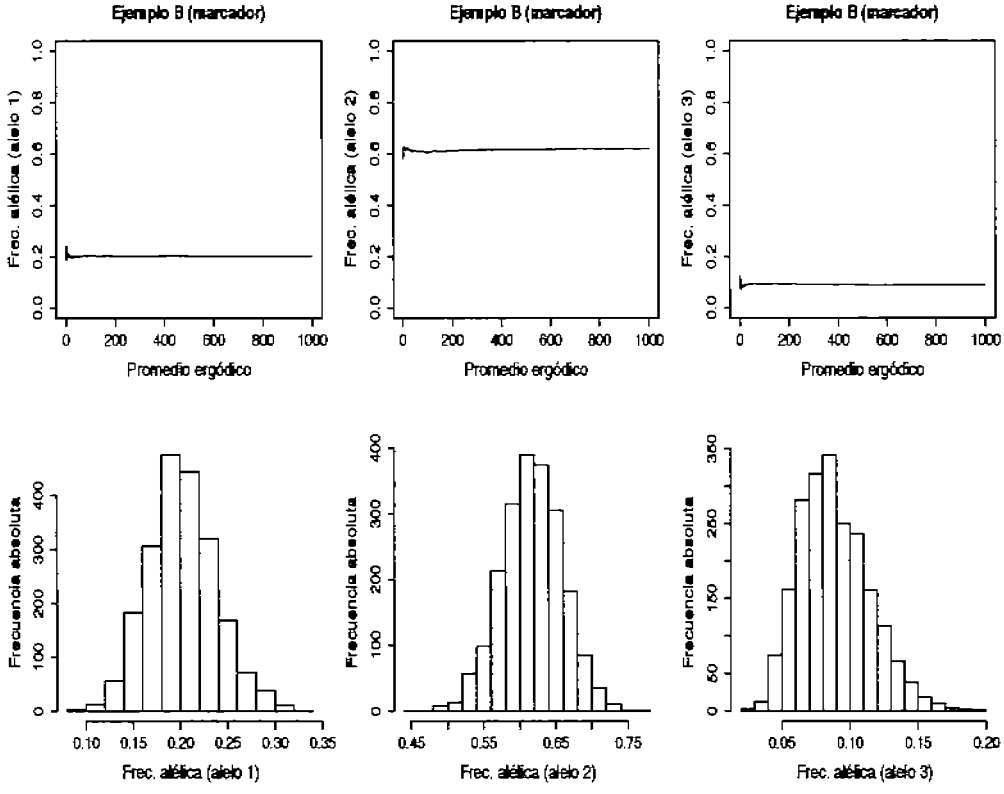

Figura 5.6: *Promedios ergódicos y distribución final de las frecuencias alélicas del marcador correspondiente al ejemplo B.*

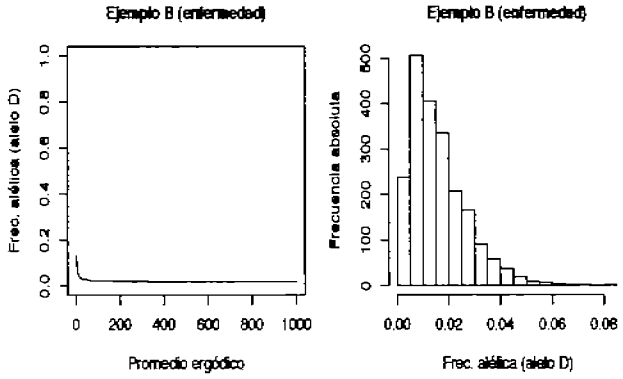

Figura 5.7: *Promedios ergódicos y distribución final de las frecuencias alélicas de la enfermeda correspondiente al ejemplo B.*

De los tres ejemplos presentados en esta sección, el que despierta mayor interés es el ejemplo B. Recuerde que el valor de lod score correspondiente al ejemplo B fue de 2.5, es decir, muy cercano al valor crítico de 3 sugerido por la teoría sobre la que descansa el método de lod score. Desde el punto de vista frecuentista, un valor de lod score mayor o igual a 3 representa que hay suficiente evidencia en favor de ligamiento, es decir que $\theta < 0.5$. Sin embargo, utilizando el enfoque Bayesiano, una probabilidad de ligamiento del 50 % no proporciona la misma seguridad que la que proporciona el valor de lod score de 2.5. Esto se debe a que en un enfoque frecuentista la evidencia en favor de una u otra hipótesis depende en gran medida del tamaño de la muestra.

Por ejemplo, si tenemos una familia de tamaño moderado, aún cuando haya mucha evidencia de ligamiento será difícil alcanzar el valor de lod score de 3, mientras que para una familia de tamaño considerable, un valor de lod score de 3 puede no representar suficiente evidencia, como es el caso particular del ejemplo B.

Por otro lado, el valor de lod score de 2.5 mencionado se obtuvo utilizando valores de penetrancia y frecuencias alélicas fijos. En un caso como este, es muy factible alcanzar el valor de lod score de 3 modificando el valor de alguno de estos parámetros. Desde el punto de vista estadístico, lo inquietante es que dicha modificación se haga de manera arbitraria y tendenciosa. Uno de los propósitos de incluir a estos parámetros en el análisis es que la información proporcionada por sus correspondientes distribuciones finales sirva para tomar decisiones menos arbitrarias y mejor fundamentadas sobre la manipulación de estos parámetros.

Los resultados de este capítulo ponen en evidencia la necesidad de herramientas metodológicas complementarias que proporcionen mayor seguridad cuando se tiene un resultado de lod score determinado.

Capítulo 6

Conclusiones

6.1. Comparación GENEHUNTER y SimWalk2

El método de lod score ha sido implementado en diversos programas de cómputo disponibles gratuitamente para el público en general. El lector interesado puede visitar la página http://linkage.rockefeller.edu. Entre todos los programas, sin duda el más popular es GENEHUNTER aún cuando el número de individuos que admite es limitado. Esto se debe principalmente a que es de fácil operación y a que puede manejar un número considerablemente grande de marcadores, que es el caso de cualquier escaneo completo del genoma humano, en donde se tiene información alrededor de 400 marcadores

genéticos. La práctica usual es que si la familia excede el número de individuos disponible, el programa descarta el número de individuos excedente empezando por los sanos. Una alternativa propuesta por los autores del programa GENEHUNTER es dividir a la familia en familias más pequeñas y tratarlas como si estas fueran independientes ignorando que no lo son. Por otro lado, otro programa no tan popular como lo es GENEHUNTER, es SimWalk2. Este último utiliza métodos de Monte Carlo vía cadenas de Markov y su mayor ventaja es que permite un número considerablemente grande de individuos y marcadores simultáneamente. Entre los factores que ocasionan que este programa no sea utilizado con mayor frecuencia pueden mencionarse: que no es tan amigable, ni tan rápido y además requiere de archivos de entrada que tienen un formato complicado. Cabe destacar que este último inconveniente se puede superar utilizando el programa MEGA2.

Goedken et al. (2000) estudió previamente el efecto de descartar individuos y dividir manualmente una familia. Estos autores compararon los resultados arrojados por los programas GENEHUNTER y VITESSE para seis familias con datos reales. Derivado de su estudio ellos concluyen que hay una mayor perdida de información cuando se divide la familia que cuando se descartan individuos. Sin embargo, los resultados presentados en este traba-

jo, sección 4.4, muestran que para familias extensas esto no necesariamente ocurre. Con respecto a esto, en este trabajo se pueden alcanzar tres conclusiones principales: (1) para familias de tamaño moderado, GENEHUNTER y SimWalk2 producen resultados muy similares; (2) en una familia extensa, cuando se utiliza GENEHUNTER, tanto descartar individuos como dividir una familia en familias más pequeñas tienen un efecto adverso, siendo el primero más grave; y (3) el comportamiento de cada uno de los programas es cualitativamente similar ante la presencia de información faltante. Esta conclusiones están basadas en lo observado en la distribución del valor de lod score, así como en la distribución del valor estimado de la fracción de recombinación.

Cuando se compararon los resultados de ambos programas para una familia de tamaño moderado, en pocas instancias el resultado de ambos programas difirió uno del otro. En aproximadamente el 95 % de los casos las diferencias no excedieron ± 0.5. Por lo tanto, cuando se dispone de familias de tamaño moderado, se recomienda uitlizar GENEHUNTER.

Con respecto al efecto de descartar individuos o dividir a la familia en familias más pequeñas con el propósito de incluir en el análisis el mayor número de individuos posible, no es posible obtener conclusiones cuantitati-

vas generales sobre la magnitud de subestimación en el valor del lod score, ya que ésta dependerá de la estructura y el tamaño de la familia estudiada. Aún cuando los resultados observados en este trabajo sugieren que dividir a la familia es una mejor estrategia que descartar individuos, es necesario ser cuidadosos ya que dividir a la familia de cualquier forma implica pérdida de información relativa al análisis que considera todos los individuos.

Cuando el gen de la enfermedad en cuestión tiene una penetrancia alta, todos los individuos, incluyendo afectados y sanos, porporcionan información útil de ligamiento y no deberían ser descartados bajo ninguna circunstancia. Sin embargo, cuando el gen de la enfermedad tiene una penetrancia baja, aparentemente utilizar la familia completa es desventajoso debido a la presencia de individuos sanos no-penetrantes, que son los que generalmente son descartados por GENEHUNTER cuando la familia se excede en tamaño. Con el fin de evitar esta situación, una estrategia puede ser asignar estatus desconocido a todos los individuos sanos e incluirlos en el análsis. Por lo tanto, en situaciones donde el número de individuos excede el limite permitido por GENEHUNTER, SimWalk2 proporcionará resultados más confiables. Cabe destacar que hay otros programas adems de SimWalk2 que pueden manejar familias extensas, una comparacin de algunos de estos se puede encontrar en

Wijsman (2003).

Con respecto al efecto de la información faltante, los resultados presentados en la sección 4.4 sugieren que, en las familias "original" y "baja penetrancia", la información faltante tiene un efecto menor sobre el valor máximo de lod score cuando existe ligamiento. Sin embargo, para el caso de la familia "estatus conocido" el efecto debido a información faltante es considerablemente mayor. Como se discutió en la sección correspondiente esta diferencia se debe al estatus de los individuos no disponibles. Estos resultados pudieran sugerir que los individuos con estatus desconocido pero genotipo conocido proporcionan la misma información de ligamiento que los individuos afectados o sanos con genotipo conocido. Esta última observación apoya la estrategia de asignar estatus desconocido a todos los individuos sanos con el fin de evitar el efecto adverso de los individuos sanos no-penetrantes en el caso de una enfermedad con baja penetrancia.

En lo que se refiere a las comparación de GENEHUNTER y SimWalk2 en el contexto de una familia multigeneracional y un rasgo cualitativo, consideramos que este trabajo será de utilidad para los usuarios de estos programas en la selección del programa que mejor se adapte a las necesidades y datos disponibles. Con este trabajo se desea también motivar el uso de los métodos

de Monte Carlo vía cadenas de Markov, ya que aún cuando estan basados en métodos de simulación y el resultado proporcionado es aproximado, dicha aproximación puede ser tan precisa como se desee.

6.2. Análisis Bayesiano de ligamiento genético

La herramienta estadística más eficaz para identificar genes con una contribución mayor en el desarrollo de una enfermedad ha sido sin duda el método de lod score. La primera versión de este método fue propuesto por Morton (1955), fue planeada para familias nucleares y un esquema de reclutamiento de familias secuencial. Este autor fue el primero en proponer un valor de lod score de 3 como evidencia a favor de ligamiento y un valor de -2 como evidencia en contra de ligamiento. A pesar de que el método de lod score ha sufrido modificaciones con respecto a su versión original, estos valores críticos para rechazar o aceptar ligamiento siguen prevaleciendo. Actualmente obtener un valor de lod score mayor o igual a 3, independientemente del número de familias, la estructura y tamaño de las mismas, es una garantía para publicar el hallazgo en una revista con arbitraje internacional.

Exiten un número considerable de reportes científicos en donde se argumenta la evidencia de ligamiento entre una enfermedad y alguna región cromósomica. Sin embargo, sólo en pocos se han replicado y confirmado estos resultados. Lo anterior se debe en gran parte a la complejidad propia del problema, ya que en la mayoría de las enfermedades son muchos los factores involucrados; por ejemplo, múltiples genes interactuando entre ellos, heterogeneidad, factores ambientales, entre otros. Sin embargo, aunado a la complejidad del problema esta el uso incuestionable del método estadístico disponible.

Adicionalmente, en el método de lod score el parámetro de mayor interés es la fracción de recombinación, la cual contiene la información relativa a la distancia que hay entre el locus de la enfermedad y el marcador genético. Sin embargo, hay otros parámetros involucrados como son la penetrancia y las frecuencias alélicas tanto del marcador como del locus de la enfermedad, de los cuales se tiene poco conocimiento, y por lo mismo en la práctica se modifican y fijan arbitrariamente. Utilizando un enfoque Bayesiano, en conjunto con los métodos de Monte Carlo vía cadenas de Markov, es posible incorporar estos parámetros al análisis de manera natural.

En el presente trabajo se implementó un programa de cómputo capaz de

realizar un análisis Bayesiano de ligamiento genético de dos puntos para un marcador multialélico. Se utilizaron tres ejemplos: el primero con evidencia contundente de ligamiento con un valor de lod score de 7.02, el segundo con evidencia sugestiva de ligamiento con un valor de lod score de 2.50, y el tercero con evidencia de ausencia de ligamiento con un valor de lod score negativo de -6.57. En los tres casos se utilizó una probabilidad inicial de ligamiento de $1/22 = 4.5\,\%$, obteniendo una probabilidad final de ligamiento de $100\,\%$, $49.1\,\%$ y $3.6\,\%$, respectivamente, para cada uno de los tres ejemplos. En otras palabras, en el primer caso, para efectos prácticos, se tiene absoluta certeza de la presencia de ligamiento entre la enfermedad y el marcador estudiado; en el segundo caso, se tiene evidencia de la presencia de ligamiento, ya que observamos un incremento considerable de la probabilidad de ligamiento, sin embargo, existe también un porcentaje considerable de incertidumbre; y en último ejemplo, se tiene una bastante certeza para considerar que no hay ligamiento y en consecuencia descartar el marcador.

De los tres ejemplo, el segundo es el que despierta mayor interés, ya que la conclusión a la que se llega vía un enfoque frecuentista no es compatible con la conclusión a la que se llega vía un enfoque Bayesiano. Es decir, por el método de lod score, un valor de 2.5 es muy cercano al valor crítico de 3

que representa suficiente evidencia en favor de ligamiento, mientras que una probabilidad de ligamiento del 50 % no proporciona la misma confianza que en el caso anterior, lo cual motiva la cautela cuando se trata de reportar un hallazgo.

En la actualidad el punto de mayor debate en el enfoque Bayesiano es la especificación de la distribución inicial del parámetro, ya que diferentes distribuciones iniciales darán lugar a diferentes conclusiones. El argumento es que el enfoque Bayesiano es vulnerable a los intereses particulares, es decir, a través de la distribución inicial es posible manipular de alguna manera los resultados de los estudios. Sin embargo, cualquier procedimiento estadístico, Bayesiano o no, está sujeto a este tipo de crítica dado que depende de una serie de supuestos para su derivación y aplicación. El procedimiento será adecuado entonces en la medida en que dichos supuestos puedan verificarse en la práctica. Por otro lado, desde el punto de vista Bayesiano es posible utilizar distribuciones iniciales imparciales (no informativas), las cuales impiden que las conclusiones se vean sesgadas debido a información inicial dudosa.

La implementación de los métodos de Monte Carlo vía cadenas de Markov, así como la utilización del enfoque Bayesiano en el análisis de ligamiento genético no ha tenido el impacto, que los autores que los propusieron hubie-

ran querido, en la comunidad de investigadores usuarios de estos programas. Esto puede deberse a que en ambos casos se requiere que el usuario tenga un conocimiento teórico más profundo con el fin de especificar los parámetros necesarios para hacer el análisis. Aún cuando esto es cierto, también es común que problemas complejos requieran también de soluciones complejas.

En esta parte del trabajo se implementó y realizó un análisis Bayesiano de ligamiento genético de dos puntos para un marcador multialélico. La extensión natural de este caso es un análisis Bayesiano de ligamiento genético de múltiples puntos, el cual será objeto de futuros trabajos.

Apéndice A

Cadenas de Markov con espacio de estados finito

Un proceso estocástico es una colección de variables aleatorias $\{X_t\}$, donde el subíndice t toma valores de un conjunto T dado. Con frecuencia T son unidades discretas de tiempo, es decir $T = \{0, 1, 2, \ldots\}$.

Los posibles valores que X_t puede tomar son llamados estados y al conjunto S de estos valores se le conoce como espacio de estados.

Por otra parte, un proceso de Markov es un proceso estocástico $X = \{X_t : t \in T\}$ si para todo $t_0 < t_1 < \ldots < t_n < t$ se tiene que

$$P(X_t = x | X_{t_0} = x_0, X_{t_1} = x_1, \ldots, X_{t_n} = x_n) = P(X_t = x | X_{t_n} = x_n).$$

Entonces, una cadena de Markov es un proceso de Markov $X = \{X_t : t \in T\}$ tal que S es finito o infinito numerable.

Para $x \in S$, $X_n = x$ quiere decir que al tiempo n la cadena toma asume el valor x o bien que esta en el estado x.

La probabilidad de que X_{n+1} estará en el estado y dado que X_n está en el estado x es llamada probabilidad de transición en un paso y se denota de la siguiente manera

$$P_{xy}^{(n,n+1)} = P(X_{n+1} = y | X_n = x) \text{ para todo } x, y \in S.$$

Cuando la probabilidad de transición en un paso $P_{xy}^{(n,n+1)}$, $x, y \in S$ no depende del tiempo n, entonces se dice que X es una cadena de Markov homogenea en el tiempo.

De manera general, a la matriz $P = (P_{xy})_{x,y \in S}$ formada por las probabilidades de transición se le conoce como matriz de trancisión. Esta matriz de transición satisface dos condiciones,

(i) $P_{xy} \geq 0$ para todo $x, y \in S$;

(i) $\sum_{y \in S} P_{xy} = 1$ para todo $x \in S$.

Sea $P_{xy}^{(m)}$ la probabilidad de transición en m pasos de una cadena de

Markov $X = \{X_n : n = 0, 1, \ldots\}$ homogenea en el tiempo, es decir

$$P_{xy}^{(m)} = P(X_{n+m} = y | X_n = x) \text{ para todo } x, y \in S$$

y sea P^m a la matriz de transición de m pasos.

Por otro lado, sea $X = \{X_n : n = 0, 1, \ldots\}$ una cadena de Markov homogenea en el tiempo con espacio de estado S, matriz de transición $P = (P_{xy})_{xy \in S}$. Se dice que x y y, para $x, y \in S$ se comunican entre sí si existen $n, m \geq 0$ tales que $P_{xy}^n > 0$ y $P_{yx}^m > 0$.

Entonces, sea $X = \{X_n : n = 0, 1, \ldots\}$ una cadena de Markov homogenea en el tiempo con espacio de estado S, matriz de transición $P = (P_{xy})_{xy \in S}$. Se dice que X es irreducible si todos sus estados se comunican entre sí.

Por último, sea $X = \{X_n : n = 0, 1, \ldots\}$ una cadena de Markov homogenea en el tiempo con espacio de estado S, matriz de transición $P = (P_{xy})_{xy \in S}$. Se define entonces el periodo $d(x)$ de un estado $x \in S$ como el mayor entero tal que $P_{xx}^n = 0$, para todo $n \neq d(x), 2d(x), \ldots$. Se dice que x es aperiódico si $d(x) = 1$. Si X tiene todos sus estados aperiódicos, entonces X es aperiódica.

Apéndice B

Algunas distribuciones continuas de probabilidad

B.1. Distribución Beta

Se dice que una variable aleatoria X tiene una distribución Beta si su función de densidad de probabilidad está dada por,

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha,\beta > 0.$$

Los parámetros de la distribución son α y β. Distintos valores de los parámetros darán lugar a distintas formas de la distribución beta. Por ejemplo, si α y β son ambos menores que uno, la distribución beta tendrá una

forma de "U", si $\alpha < 1$ y $\beta \geq 1$, la distribución beta tendrá una forma de "J" transpuesta, y si $\alpha \geq 1$ y $\beta < 1$, la distribución beta tendrá una forma de "J". Cuando α y β son ambos mayores que uno, la distribución tendrá forma de "U" invertida y alcanzará su máximo en $x = (\alpha - 1)/(\alpha + \beta - 2)$. Finalmente, la distribución es simétrica cuando $\alpha = \beta$

La media y varianza de la distribución beta son:

$$E(X) = \frac{\alpha}{\alpha + \beta}$$

$$V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

B.2. Distribución Dirichlet

La distribución Dirichlet es una generalización de la distribución beta para k variables aleatorias, $\mathbf{X} = (X_1, X_2, \ldots, X_k)$.

La función de densidad de probabilidad del vector aleatorio $\mathbf{X}$ se define como,

$$f(\mathbf{x}; \mathbf{u}) = \frac{\Gamma(\sum_{i=1}^{n} u_i)}{\prod_{i=1}^{u} \Gamma(u_i)} \prod_{i=1}^{n} x_i^{u_i - 1}$$

donde $x_1, x_2, \ldots, x_k \geq 0$; $\sum_{i=1}^{n} x_i = 1$ y $u_1, u_2, \ldots, u_k > 0$.

El vector de parámetros de la distribución es $\mathbf{u} = (u_1, u_2, \ldots, u_k)$. Sea

$u_0 = \sum_{i=1}^{n} u_i$, entonces la media y varianza de la distribución Dirichlet son:

$$E(x_i) = \frac{u_i}{u_0}$$

$$V(x_i) = \frac{u_i(u_0 - ui)}{u_0^2(u_0 + 1)}.$$

Bibliografía

Canizales-Quinteros, S., Aguilar-Salinas, C., Reyes-Rodríguez, E., Riba, L., Rodríguez-Torres, M., Ramírez-Jiménez, S., Huertas-Vázquez, A., Fragoso-Ontiveros, V., Zentella-Dehesa, A., Ventura-Gallegos, J., Vega-Hernández, G., López-Estrada, A., Aurón-Gómez, M., Gómez-Pérez, F., Rull, J., Cox, N., Bell, G. & Tusie-Luna, M. (2003), 'Locus on chromosome 6p linked to elevated HDL cholesterol serum levels and to protection against premature atherosclerosis in a kindred with familial hypercholesterolemia', *Circulation Research* **92**, 569–576.

Cannings, C., Thompson, E. & Skolnick, M. (1978), 'Probability functions on complex pedigrees', *Advances in Applied Probability* **10**, 26–61.

Clerget-Darpoux, F., Bonaiti-Pellie, C. & Hochez, J. (1986), 'Effects of misspecifying genetic parameters in lod score analysis', *Biometrics* **42**, 393–

399.

Cottingham, R., Idury, R. & Schaffer, A. (1993), 'Faster sequential genetic linkage computations', *American Journal of Human Genetics* **53**, 252–263.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society* **B39**, 1–38.

Elston, R. (1989), 'Man bites dog? The validity of maximizing lod scores to determine mode of inheritance', *American Journal of Medical Genetics* **34**, 487–488.

Elston, R. & Stewart, J. (1971), 'A general model for the genetic analysis of pedigree data', *Human Heredity* **21**, 523–542.

Feller, W. (1968), *An introduction to probability theory and its applications*, John Wiley and Sons, EUA.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Gilks, W., Richardson, S. & Spiegelhalter, D. (1996), *Markov Chain Monte Carlo In Practice*, Chapman and Hall, EUA.

Goedken, R., Ludington, E., Crowe, R., Fyer, A., Hodge, S., Knowles, J., Vieland, V. & Weissman, M. (2000), 'Drawbacks of GENEHUNTER for larger pedigrees: application to panic disorder', *American Journal of Medical Genetics* **96**, 781–783.

Greenberg, D. (1989), 'Inferring mode of inheritance by comparison of lod scores', *American Journal of Medical Genetics* **34**, 480–486.

Hastings, W. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Hauser, E. & Boehnke, M. (1993), 'The posterior probability of linkage', *American Journal of Human Genetics* **Suppl 53**, 1012.

Hodge, S. & Elston, R. (1994), 'Lods, Wrods, and Mods: The interpretation of lod scores calculated under different models', *Genetic Epidemiology* **11**, 329–342.

Jensen, C. & Kong, A. (1999), 'Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops', *American Journal of Human Genetics* **65**, 885–901.

Jensen, C. & Sheehan, N. (1998), 'Problems with determination of noncommunicating classes for Monte Carlo Markov chain applications in pedigree analysis', *Biometrics* **54**, 416–425.

Karlin, S. & Taylor, H. (1975), *A first course in stochastic processes*, Academic Press, EUA.

Kong, A. (1991), Analysis of pedigree data using methods combining peeling and Gibbs sampling, *in* E. Keramidas & S. Kaufman, eds, 'Computer Science and Statistics', Proceedings of the 23rd Symposium on the Interface, Interface Foundation of North America, Fairfax Station, VA, USA, pp. 379–385.

Kruglyak, L., Daly, M., Reeve-Daly, M. & Lander, E. (1996), 'Parametric and nonparametric linkage analysis: A unified multipoint approach', *American Journal of Human Genetics* **56**, 519–527.

Lander, E. & Green, P. (1987), 'Construction of multilocus genetic maps in humans', *Proceedings of the National Academy of Science, USA* **84**, 2363–2367.

Lange, K. & Matthysse, S. (1989), 'Simulation of pedigree genotypes by random walks', *American Journal of Human Genetics* **45**, 959–970.

Lathrop, G., Lalouel, J., Julier, C. & Ott, J. (1984), 'Strategies for multilocus linkage analysis in humans', *Proceedings of the National Academy of Science* **81**, 3443–3446.

Leppert, M., Hasstedt, S., Holm, T., O'Connell, P., Wu, L., Ash, O., Williams, R. & White, R. (1986), 'A DNA probe for the LDL receptor gene is tightly linked to hypercholesterolemia in a pedigree with early coronary disease', *American Journal of Human Genetics* **39**, 300–306.

Lin, S. (1995), 'A scheme for constructing an irreducible Markov chain for pedigree data', *Biometrics* **51**, 318–322.

Lin, S. (1996), 'Multipoint linkage analysis via Metropolis jumping kernels', *Biometrics* **52**, 1417–1427.

Lin, S., Thompson, E. & Wijsman, E. (1993), 'Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data', *IMA Journal of Mathematics Applied in Medicine and Biology* **10**, 1–17.

Lin, S., Thompson, E. & Wijsman, E. (1994), 'Finding noncommunicating sets for Markov chain Monte Carlo estimations on pedigrees', *American Journal of Human Genetics* **54**, 695–704.

MacCluer, J., Vandeburg, J., Read, B. & Ryder, O. (1986), 'Pedigree analysis by computer simulation', *Zoo Biology* **5**, 147–160.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953), 'Equations of state calculations by fase computing machines', *Journal of Chemical Physics* **21**, 1087–1092.

Morton, N. (1955), 'Sequential tests for the detection of linkage', *American Journal of Human Genetics* **7**, 277–318.

Mukhopadhyay, N., Almasy, L., Schroeder, M., Mulvihill, W. & Weeks, D. (1999), 'MEGA2, a data-handling program for facilitating genetic linkage and association analyses', *American Journal of Human Genetics* **65**, A436.

O'Connell, J. & Weeks, D. (1995), 'The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance', *Nature Genetics* **11**, 402–408.

Ott, J. (1974), 'Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies', *American Journal of Human Genetics* **26**, 588–597.

Ott, J. (1999), *Analysis of Human Genetic Linkage*, 3 edn, Johns Hopkins University Press, Baltimore.

Sheehan, N. & Thomas, A. (1993), 'On the irreducibility of a Markov chain defined on a space of genotype configurations by sampling scheme', *Biometrics* **49**, 163–175.

Sisson, S. (2004), 'An algorithm to characterize non-communicating classes on complex genealogies', *publicación electrónica* .

Smith, C. (1959), 'Some comments on the statistical methods used in linkage investigations', *American Journal of Human Genetics* **11**, 289–304.

Sobel, E. & Lange, K. (1996), 'Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics', *American Journal of Human Genetics* **58**, 1323–1327.

Sobel, E., Lange, K., O'Connell, J. & Weeks, D. (1995), Haplotyping algorithms, *in* T. Speed & M. Waterman, eds, 'Genetic mapping and DNA sequencing', Vol. 81 of *IMA volumes in mathematics and its applications*, Springer-Verlag, New York, NY, USA, pp. 89–110.

Terwilliger, J., Speer, M. & Ott, J. (1993), 'Chromosome-based method for rapid computer simulation in human genetic linkage analysis', *Genetic Epidemiology* **10**, 217–224.

Thomas, D. & Cortessis, V. (1992), 'A Gibbs sampling approach to linkage analysis', *Human Heredity* **42**, 63–76.

Thomas, D., Richardson, S., Gauderman, J. & Pitkäniemi, J. (1997), 'A Bayesian approach to multipoint mapping in nuclear families', *Genetic Epidemiology* **14**, 903–908.

Thompson, E. (1994), 'Monte Carlo likelihood in genetics mapping', *Statistical Science* **9**, 355–366.

Thompson, E. & Heath, S. (1999), Estimation of conditional multilocus gene identity among relatives, *in* F. Seillier-Moiseiwitsch, ed., 'Statistics in Molecular Biology and Genetics', Vol. 33 of *Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology. IMS Lecture Note-Monograph Series*, Institute of Mathematical Statistics, Hayward, CA, USA, pp. 95–113.

Vieland, V. (1998), 'Bayesian linkage analysis, or: how I learned to stop worying and love the posterior probability of linkage', *American Journal of Human Genetics* **63**, 947–954.

Weeks, D., Lehner, T., Squires-Wheeler, E. & Kaufmann, C. (1990), 'Measuring the inflation of the lod scores due to its maximization over model parameter values in human linkage analysis', *Genetic Epidemiology* **7**, 237–243.

Weir, B. (1996), *Genetic data analysis II*, Sinauer Associates, Inc., Sunderland, MA.

Wijsman, E. (2003), 'Summary of Group 8: Development and extension of linkage methods', *Genetic Epidemiology* **25**, S64–S71.

Artículo publicado

Romero-Hidalgo, S., Rodrigues, E.R., Gutierrez-Peña, E., Riba, L. & Tusié-Luna, M.T. (2005), 'GENEHUNTER versus SimWalk2 in the context of an extended kindred and a qualitative trait locus', *Genetica* **123**, 235-244.

Genetica (2005) 123: 235–244

GENEHUNTER versus SimWalk2 in the context of an extended kindred and a qualitative trait locus

Sandra Romero-Hidalgo[1,*], Eliane R. Rodrigues[2], Eduardo Gutiérrez-Peña[3], Laura Riba[1] & María Teresa Tusié-Luna[1]

[1]*Unidad de Biología Molecular y Medicina Genómica, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico;* [2]*Instituto de Matemáticas, Universidad Nacional Autónoma de México, Mexico City, Mexico;* [3]*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de Mexico, Mexico City, Mexico;* **Address for correspondence; Unidad de Biología Molecular y Medicina Genómica, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Vasco de Quiroga # 15, Colonia Sección 16, Delegación Tlalpan, México D.F. 14000, México (Phone: +5255-56-55-00-11; E-mail: srhidalgo@biomedicas.unam.mx)*

Received 25 July 2004 Accepted 14 August 2004

Key words: linkage analysis, linkage software, lod score, multigenerational extended kindred, pedigree, recombination fraction

Abstract

GENEHUNTER and SimWalk2 are among the most commonly used software for parametric multipoint linkage analysis. In the context of extended kindred analysis, GENEHUNTER has a limitation in terms of the number of individuals it can handle. One solution is to manually split the kindred into smaller pedigrees. SimWalk2 can handle a much larger number of individuals. However, its major drawback is the time it takes to process the data when compared to GENEHUNTER. Aside from the limitations of each program, when studying extended kindreds researchers are typically confronted with missing data. In this work we used simulated genotype data based on the structure of a real extended pedigree in order to compare the results obtained through GENEHUNTER and SimWalk2, evaluate the effect of discarding individuals and splitting the kindred on the logarithm of odds (lod) score, and to assess how missing data affect the performance of each program. Our results show that (1) for pedigrees of a moderate size, GENEHUNTER and SimWalk2 produce nearly the same results; (2) when using GENEHUNTER, either splitting the kindred into smaller sub-pedigrees or discarding individuals has an adverse effect when compared to the results obtained when using SimWalk2 with the whole pedigree; and (3) the performance of both programs is qualitatively similar in the missing data scenario. These conclusions are based on the sample distributions of the lod score values and of the estimates of the recombination fraction.

Introduction

Linkage analysis is a strategy that aims to identify genes responsible for certain inherited diseases solely through their chromosomal location within the genome. When extended multigenerational families are available and the family members are unambiguously clinically characterized as affected or unaffected, one of the most popular statistical tools is the parametric logarithm of odds (lod) score method. This method has been implemented using different algorithms in several public domain software packages.

Two commonly used programs that perform multipoint linkage analysis are GENEHUNTER (Kruglyak et al., 1996) and SimWalk2 (Sobel & Lange, 1996). In the context of extended kindred analysis, GENEHUNTER has the advantage that

a considerable number of markers can be analyzed although the number of individuals it can handle is limited. The maximum number of individuals depends on the structure of the family but it allows for approximately 12 non-founders (individuals whose parents are in the pedigree). When the pedigree exceeds the maximum allowed number of individuals, the program discards individuals starting with the unaffected. In order to include as many individuals as possible, one recommendation is to manually split the kindred into smaller pedigrees. In contrast, SimWalk2 has the advantage that it can handle up to 258 individuals as well as a considerable number of markers, although a drawback is a much longer processing time as compared to GENEHUNTER.

When positive but non-conclusive lod score values are found for one or more chromosomal regions with one program, a reliable measure of precision would be desirable for further analysis. One option is to run the same data using a different linkage program to see whether both analyses lead to the same result. However, if the conclusions reached by these programs differ, an assessment has to be made as to which result is more reliable. In this work, we have studied for different disease models parametric linkage analysis using GENEHUNTER and SimWalk2 under various scenarios, including splitting the kindred, discarding individuals (also referred to as 'trimming' method), and the presence of missing data. From now on GENEHUNTER and SimWalk2 will be referred to as GH and SW2, respectively.

The effect of discarding individuals as well as the effect of splitting the kindred into smaller pedigrees has been previously explored by Goedken et al. (2000). From their results it is not possible to reach any conclusions about the real effect of discarding individuals and of splitting the pedigree other than the fact that there is a loss of linkage information in both situations. Our intent was to take this investigation further using simulated data. We have simulated two disease models based on the structure of one real extended pedigree. The first model consists of a major dominant gene with a penetrance of 90%. In the second model, we have simulated a low penetrance (50%) dominant gene. For each of these two models, we have simulated two alternatives: no-linkage and linkage between a qualitative trait and a selected marker.

Aside from the limitations of some programs to analyze extended kindreds, another frequent problem is that we do not typically have access to the genotypes of all the family members. When this happens, we are confronted with a problem of missing data in the analysis. Therefore, we were also interested in exploring the effect of missing data on the behavior of the parametric lod score obtained through GH and SW2.

Methods

Simulated genotyping data sets were generated based on a real extended multigenerational family (Figure 1). The structure and characterization of this family is a preliminary version of a kindred used in a real linkage study (Canizales-Quinteros et al., 2003). The working hypothesis was that the phenotype in this family is due to an autosomal dominant gene with incomplete penetrance (90%). Seventeen polymorphic genetic markers (microsatellites) in chromosome 2 were considered for the analysis.

The kindred comprises five generations with 72 individuals, 11 of which are affected with familial hypercholesterolemia (Figure 1(a)). In this family we have 31 individuals with missing genotype information and 22 individuals with unknown affection status (19 of the latter are among the 31 individuals with missing genotype data, while three have genotype data). We will refer to this pedigree as 'original'.

Two other cases were also considered. Each one of them was produced by changing one particular characteristic in the description given above. For one case, the penetrance was reduced to 50%. Using this penetrance and the same family structure, the affection status was reassigned for all individuals. The resulting pedigree has 8 out of 72 individuals affected as shown in Figure 1(b) and will be referred to as 'low-penetrance'.

Another case keeps the 90% penetrance. However, the affection status of the individuals with no status information available in the original pedigree was randomly assigned according to Mendelian laws. In Figure 1(c), we show the resulting pedigree with a total of 20 affected individuals. This pedigree will be referred to as 'known-status'.

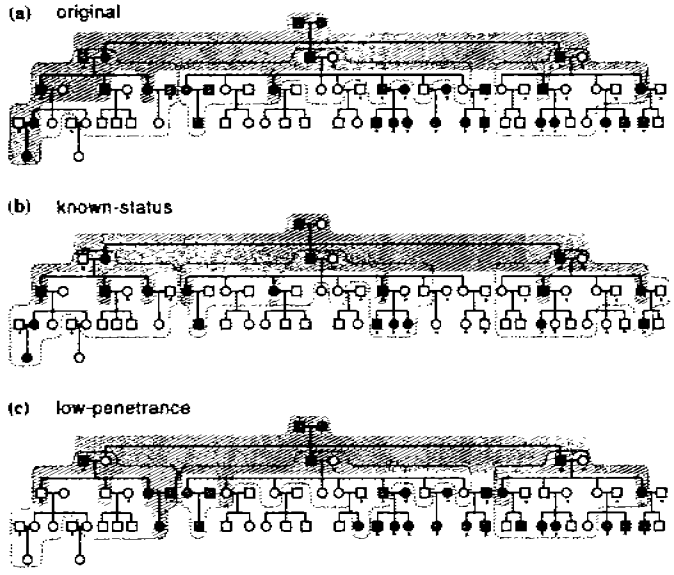

Figure 1. Kindreds studied. Bold symbols correspond to affected individuals. Open symbols correspond to unaffected individuals. Gray symbols correspond to individuals whose affection status is unknown ('status 0'). Individuals with missing genotype information are indicated with asterisks. The shaded area includes the individuals that GH keeps in the analysis. Dashed lines mark the sub-pedigrees used by GH.

For each of the three cases described above, simulations were performed using two models. The first model assumes that the disease gene is unlinked to chromosome 2. The second model assumes that the disease gene is linked to marker 14 (located at 168.33 cM) at a recombination fraction of zero ($\theta = 0$). Both models were developed taking into consideration the affection status of each individual (see Figure 1), and the corresponding penetrance for the disease.

Using the allele frequencies estimated in Mexican population for the 17 markers, random genotypes were assigned to individuals whose parents were not included in the pedigree (founders) under the assumption of Hardy-Weinberg equilibrium. To generate the genotypes for individuals with parents in the pedigree, the number of recombinations, along with their position among the markers was simulated according to a Poisson process as described by Terwilliger, Speer and Ott (1993).

Using this procedure, 100 replicates of the three different pedigrees in Figure 1 were generated for each of the two models. These replicates reflect the hypothetical scenario where the genotype for each individual in the pedigree is known. Based on these replicates, five additional sets of 100 were generated, with a different scenario considered for each set. We therefore generated a total of 6 sets of 100 replicates for each of the two models studied (Table 1). In Set 1 all 72 individuals in the pedigree have known genotypes. In Set 2 the genotypes of 31 out of 72 individuals are assumed unknown as is the case for the real data of the pedigree in Figure 1. In Set 3 the pedigree consists of the individuals included by GH in the analysis. In Set 4 the genotypes of 8 out of included individuals of Set 3 are assumed unknown. In Set 5 the original structure of the kindred is split into three smaller sub-pedigrees of 19, 31 and 20 individuals. In Set 6 we consider the same splitting pattern as in Set 5 but, in this scenario, the genotypes of 6 individuals in the first sub-pedigree, 12 individuals in the second sub-pedigree and 11 individuals in the third sub-pedigree are assumed unknown.

Multipoint linkage analysis was performed to each replicate in each set using GENEHUNTER v.2-β and/or SimWalk2 v.2.82. Table 1 summarizes the main characteristics of the six different sets and the programs used in each scenario. In order to simplify the description of each set the following notation was assigned in the last

Table 1. Description summary of the sets generated for each of the two models (no linkage and linkage) of the three cases (original, known-status and low-penetrance)

Set[a]	Number of Pedigrees	Individuals[b] (O, KS, LP)	Missing[c] (O, KS, LP)	Program used	Notation used[d]
1	1	(72, 72, 72)	(0, 0, 0)	SW2	WC
2	1	(72, 72, 72)	(31, 31, 31)	SW2	WM
3	1	(18, 16, 16)	(0, 0, 0)	GH/SW2	DC
4	1	(18, 16, 16)	(8, 8, 8)	GH/SW2	DM
5	3	(45, 50, 46)	(0, 0, 0)	GH	SC
6	3	(45, 50, 46)	(14, 20, 15)	GH	SM

[a]For each of the two models (no-linkage and linkage) of the three cases (O: original, KS: known-status and LP: low-penetrance), 100 replicates were generated (Set 1). Based on these 100 replicates, five additional scenarios (Set 2–6) were considered.
[b]Number of individuals considered in each set for the three cases.
[c]Number of individual for whom genotype information is unavailable.
[d]W, S and D refer to the pedigree cases: W, whole; S, split; and D, with discarded individuals (or trimmed pedigrees). C and M refer to the complete and missing genotype information scenarios, respectively.

column: 'W' refers to the whole family, 'D' refers to the family with discarded individuals and 'S' refers to the split family, while 'C' and 'M' stand for complete genotype data and missing information scenario, respectively.

Results

Comparison of the overall results from GENEHUNTER and SimWalk2

When pedigrees of moderate size were analyzed, GH and SW2 produced nearly the same results. In order to perform this comparison, each of the 100 replicates of Set 4 (DM) was analyzed with both programs.

As shown in Figure 2, lod scores obtained from both GH and SW2, for the 17 markers, are in good agreement when assuming no linkage. We observe a similar behavior when linkage is assumed, taking into account only the lod scores obtained for the disease locus (at marker 14). Figures 2(a–c) correspond to the three pedigrees original, known-status, and low-penetrance, respectively.

The estimates for the recombination fraction produced by both programs are shown in Figure 3. Only the missing information scenarios are considered. These estimates were obtained using the distance, in centiMorgans, between the position that produces the maximum lod score and the position of marker 14. The distances in centiMorgans were transformed into recombination fractions using the corresponding map function. The sample distribution in the three cases was found to be similar for both the GH and SW2 programs.

Table 2 records the power of each program to detect linkage considering the three pedigrees studied. The values in Table 2 are the proportion of maximum lod scores greater than or equal to three in the true-linkage model. Dashed lines indicate scenarios that were not considered. In the case of GH it is not possible to use the whole family, and since SW2 allows the analysis of the entire dataset we did not run the split version of the pedigree.

Comparison between the 'splitting' and 'trimming' methods, and the use of the entire pedigree when linkage is present

When using GH, either splitting the kindred into smaller sub-pedigrees or discarding individuals has an adverse effect when compared to the results obtained using SW2 with the whole pedigree. In order to assess this effect, we contrast the results from Set 6 (SM) and Set 4 (DM) with those obtained from Set 2 (WM) for the original, low-penetrance and known-status pedigrees.

The number of individuals included in each pedigree is shown in Table 1. Figure 4 shows the comparative behavior of the maximum lod scores in the linkage model for the following scenarios: SW2 with the whole pedigree (WM), GH partitioning the pedigree into three sub-pedigrees (SM) and GH with discarded individuals (DM), for the three pedigrees studied. The difference between the

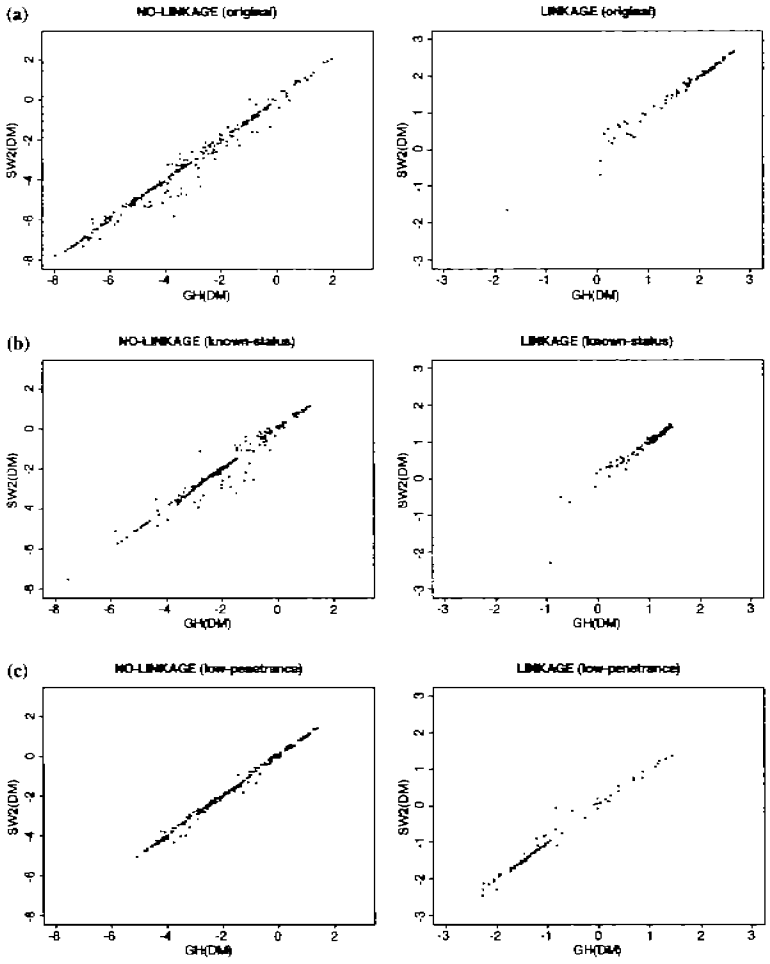

Figure 2. Comparison of the lod scores obtained through GH and SW2 for the three moderate-size pedigrees (a–c) under the missing genotype information scenario. Results for the 17 markers were used in the NO-LINKAGE model, whereas only the results for marker 14 were used in the LINKAGE model.

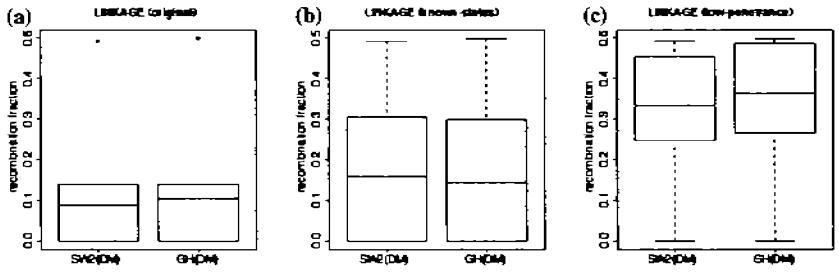

Figure 3. Comparison of the sample distribution of the estimates of the recombination fraction obtained through GH and SW2 for the three moderate-size pedigrees (a–c) in the LINKAGE model, under the missing genotype information scenario.

Table 2. In the linkage model, the proportion of maximum lod scores greater than or equal to 3

	W[a]		S[a]		D[a]	
	C[b]	M[b]	C[b]	M[b]	C[b]	M[b]
SW2						
Original	0.83	0.75	–	–	0	0
Known-status	0.82	0.43	–	–	0	0
Low-penetrance	0.02	0.01	–	–	0	0
GH						
Original	–	–	0.61	0.48	0	0
Known-status	–	–	0.65	0.14	0	0
Low-penetrance	–	–	0	0	0	0

[a]W, S and D refer to the whole, split and trimmed pedigrees, respectively.
[b]C and M refer to the complete and missing genotype information scenarios, respectively.
Dashes indicate scenarios that were not analyzed.

boxes in Figures 4(a, b) represents the negative effect of splitting and trimming the kindred compared to the use of the entire data set (when the penetrance is 90%). From these results, it is clear that in these two cases splitting the kindred is a better alternative than discarding individuals.

In the low-penetrance case, analysis of the whole pedigree with SW2, healthy non-penetrant individuals reduce the lod score value. This is shown in Figure 4(c) where the lod score values obtained using SW2 are lower than those produced by GH in the splitting and trimming scenarios. This is due to the number of individuals considered in the analysis.

Figure 5 shows the estimates for the recombination fraction obtained for the same scenarios as in Figure 4. It is possible to observe that using the entire pedigree produced more accurate estimates in the original and low-penetrance pedigrees, whereas in the known-status pedigree the estimates produced by SW2 using the whole data set are similar to those produced by GH using a split pedigree. The power of each program is shown in Table 2.

Comparison using complete and missing genotype information when linkage is present

To assess the performance of each program when some of the data are missing we compared the scenarios of complete versus incomplete information. The comparison was made for each pedigree studied (original, known-status and low-penetrance). The number of missing individuals in each case is indicated in Table 1.

For SW2, the contrast was made between the results obtained from Sets 1 (WC) and 2 (WM). Figures 6 (1a and c) show that missing genotype information does not have an impact in the original and low-penetrance pedigrees. However, when all the missing individuals have known status, missing information plays an important role as shown in Figure 6(1b). A similar behavior can be observed in the sample distribution of the estimates of the recombination fraction (Figure 7(1)).

For GH, the comparison was made between results for Sets 3 (DC) and 4 (DM). Missing genotype data produces a consistent underestimation of the lod score (Figure 6(2)). The distribution of the estimates of the recombination fraction shows that in the original and known-status pedigrees (Figures 7 (2a and b)) missing data has the effect of producing a less accurate estimate, whereas in the low-penetrance case the results are similar (Figure 7(2c)).

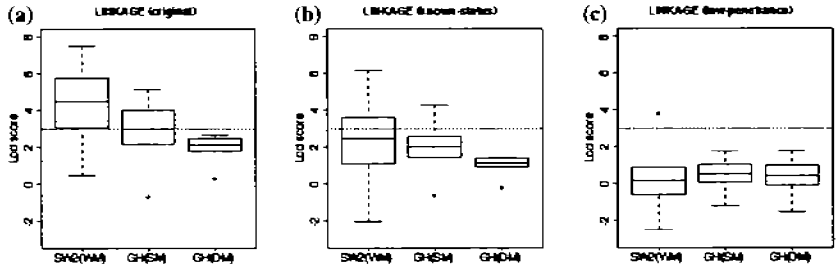

Figure 4. Comparison of the maximum lod scores for the three pedigrees (a–c), obtained through SW2 using the entire pedigree and those obtained through GH using the splitting and trimming methods in the LINKAGE model, under the missing genotype information scenario. The dashed line corresponds to a lod score value of three.

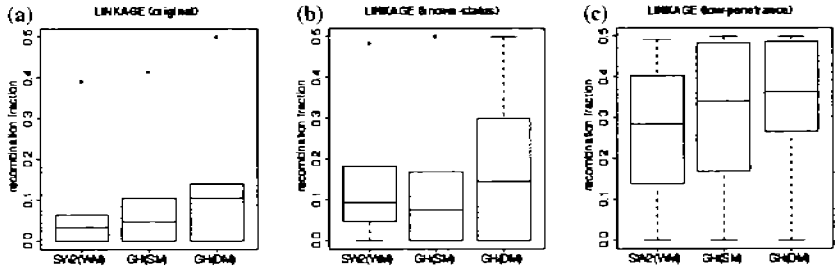

Figure 5. Comparison of the sample distribution of the estimates of the recombination fraction, for the three pedigrees (a–c), obtained through SW2 using the entire pedigree and those obtained through GH using the splitting and trimming methods in the LINKAGE model, under the missing genotype information scenario.

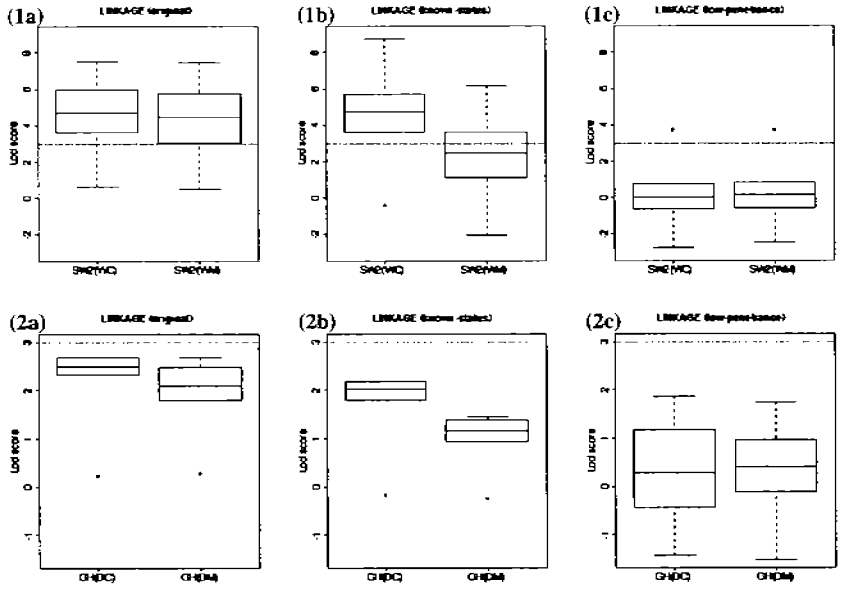

Figure 6. Comparison of the maximum lod scores for the three pedigrees (a–c) obtained through SW2 using the entire pedigree under the complete and missing genotype information scenarios (group 1). Figures 2(a–c) correspond to the obtained maximum lod scores for the three pedigrees obtained through GH using the trimmed pedigree under the complete and missing genotype information scenarios. The dashed line corresponds to a lod score value of three.

Discussion

Although there are several linkage programs available in public web sites, GH is one of the most frequently used, mainly because of its user-friendly interface and rapid operation. However, its usefulness is limited to relatively small pedigrees. In contrast, SW2 allows the use of larger kindreds at the expense of easy and fast operation. SW2 requires a more complex input file format; nevertheless, one can overcome this requirement by means of the MEGA2 (Manipulation Environment for Genetic Analysis) software (Mukhopadhyay et al., 1999). MEGA2 transforms a set of input files (very similar to those required by GH) into alternative formats, including that required for SW2.

During the course of an actual hypercholesterolemia research project, when real chromosome 2 marker data were analyzed (see Figure 1(a)), the

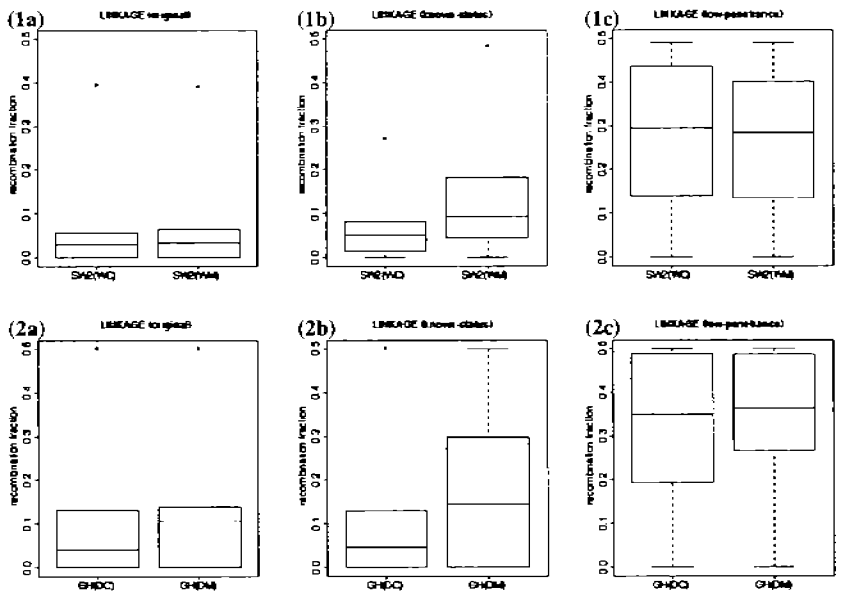

Figure 7. Comparison of the sample distribution of the estimates of the recombination fraction for the three pedigrees (a–c) obtained through SW2 using the entire pedigree under the complete and missing genotype information scenarios (group 1). Figures 2(a–c) correspond to sample distribution of the estimates of the recombination fraction for the three pedigrees obtained through GH using the trimmed pedigree under the complete and missing genotype information scenarios.

results from GH and SW2 led to opposite conclusions. A positive lod score of 1.2 was obtained from GH while SW2 yielded a negative lod score of −4.9. It was these results, along with the limitation of GH concerning the number of individuals that motivated us to carry out the present study.

Goedken et al. (2000) previously studied the effect of discarding individuals or manually splitting a pedigree. They compare the results obtained by the programs GH and VITESSE (O'Connell & Weeks, 1995), using six pedigrees with real marker data with a range of 17–39 individuals, and between 3 and 13 affected individuals per family. From their analysis based on one data set for six families they conclude that there is a loss of pedigree information when GH discards individuals. With respect to splitting a pedigree they state that 'to split the pedigrees can result in a greater loss of information than the trimming method'. However, our results show that for larger pedigrees this statement does not necessarily apply.

In the present study we have reached three main conclusions: (1) for pedigrees of a moderate size, GENEHUNTER and SimWalk2 produce nearly the same results; (2) when using GH, either splitting the kindred into smaller sub-pedigrees or discarding individuals has an adverse effect when compared to the results obtained when using SW2 with the whole pedigree; and (3) the performance of both programs is qualitatively similar in the missing data scenario. These conclusions are based on the sample distributions of the lod scores and of the estimates of the recombination fraction.

When comparing GH and SW2, we have observed few instances where the results differ from one program to another. These results are the points that do not fit exactly into the diagonal straight line in Figure 1. In approximately 95% of the cases the difference between results obtained from each program did not exceed ± 0.5. Therefore for moderate-size pedigrees GH is a suitable alternative.

Regarding the effect of discarding individuals or splitting a kindred, no general quantitative conclusions about the degree of underestimation in the value of the maximum lod score can be reached when linkage is present. The degree of underestimation will depend on the structure and the size of

each pedigree. Although the results suggest that splitting a pedigree is a better option than discarding individuals, caution must be exercised since splitting a kindred still implies loss of information relative to the analysis of the whole pedigree. Specifically, when we analyze more than one pedigree with GH we assumed they are independent. This assumption is of course not longer valid when an extended kindred is split into smaller pedigrees.

When the underlying disease gene has a high penetrance, all the individuals in the pedigree, either affected or unaffected, provide useful linkage information and should not be discarded in any case. Figures 4(a and b) show the benefit of analyzing the whole pedigree with SW2, rather than resorting to either the 'splitting' or 'trimming' methods using GH. However, when the underlying disease gene has low penetrance, from Figure 4(c) it seems that using the whole pedigree may turn out to be a disadvantage, since healthy non-penetrant individuals that might be discarded in the splitting and trimming scenarios are including in the analysis. An alternative in order to avoid this situation is to assign unknown status to all unaffected individuals ('affecteds-only' analysis) on account of the low penetrance, keeping them in the analysis (this last suggestion further discussed below). Therefore, in situations where the number of individuals exceeds that allowed by GH, SimWalk2 is the better option. Note that there are other programs that can handle larger pedigrees. Comparisons among some of them are drawn in Wijsman (2003).

Regarding the effect of missing data, Figures 6(2a–c) show the distribution of the maximum lod score obtained from GH, Set 3 (DC) and Set 4 (DM) for the three pedigrees studied. The graphs obtained from SW2 for Set 3 (DC) and Set 4 (DM) were not included since they thoroughly overlap those obtained from GH. Figures 6(1a and c) suggest a minor effect of missing data on the maximum lod score when linkage is present for the original and low-penetrance pedigrees. Particularly, in the case of the original pedigree the power to detect linkage due to missing information goes from 0.83 to 0.75 (Table 2). However, this minor effect cannot hold for the known-status pedigree where the power to detect linkage in the complete information scenario is 0.82 in contrast to 0.43 from the missing information scenario. Recall that the difference between original and known-status pedigrees is that in the former 19 out of 31 individuals with missing information also have unknown affection status, whereas in the latter the affection status of all the individuals in the pedigree is known. Therefore the different effect of missing data in these two pedigrees is due to the affection status of the missing individuals. From the theoretical point of view, this phenomenon does have an explanation in terms of the components included in the likelihood function used in the lod score method. The power to detect linkage in the complete genotype information scenario is 0.83 and 0.82 for the original and known-status pedigree, respectively. These results may suggest that individuals with unknown affection status but known genotype information provide the same contribution to the linkage analysis as the affected or unaffected individuals with known genotype information. This last observation supports the suggestion to assign an unknown status to all unaffected individuals in order to avoid the effect of healthy non-penetrant individuals in a low-penetrance scenario. In other words, unknown-status individuals do not harm the conclusions; on the contrary, they may help in this particular circumstance (see Figures 6(1a–c)).

We have performed an extensive and thorough comparison of the performance of the GH and SW2 software in the context of an extended multigenerational kindred and a qualitative trait locus. We believe this work will help linkage software users in their task of choosing the most suitable program according to their needs and the available data.

Acknowledgments

We thank Salvador Curiel and Luis Javier Álvarez for providing informatics support. This work was supported by grant IN217501 from DGAPA, Universidad Nacional Autónoma de México; grant 30774-M from Consejo Nacional de Ciencia y Tecnología, Mexico; and from Fundación Miguel Alemán, Mexico. Romero-Hidalgo S. was supported by a PhD. fellowship from Consejo Nacional de Ciencia y Tecnología, Mexico.

References

Canizales-Quinteros, S., C.A. Aguilar-Salinas, E. Reyes-Rodríguez, L. Riba, M. Rodríguez-Torres, S. Ramírez-Jiménez,

A. Huertas-Vázquez, V. Fragoso-Ontiveros, A. Zentella-Dehesa, J.L. Ventura-Gallegos, G. Vega-Hernández, A. López- Estrada, M. Aurón-Gómez, F. Gómez-Pérez, J. Rull, N.J. Cox, G.I. Bell & M.T. Tusie-Luna, 2003. Locus on chromosome 6p linked to elevated HDL cholesterol serum levels and to protection against premature atherosclerosis in a kindred with familial hypercholesterolemia. Circ. Res. 92(5): 569–576.

Goedken, R., E. Ludington, R. Crowe, A.J. Fyer, S.E. Hodge, J.A. Knowles, V.J. Vieland & M.M. Weissman, 2000. Drawbacks of GENEHUNTER for larger pedigrees: application to panic disorder. Am. J. Med. Genet. 96: 781–783.

Kruglyak, L., M.J. Daly, M.P. Reeve-Daly & E.S. Lander, 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. 58: 1347–1363.

Mukhopadhyay, N., L. Almasy, M. Schroeder, W.P. Mulvihill & D.E. Weeks, 1999. MEGA2, a data-handling program for facilitating genetic linkage and association analyses. Am. J. Hum. Genet. 65: A436.

O'Connell, J.R. & D.E. Weeks, 1995. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat. Genet. 11(4): 402–408.

Sobel, E. & K. Lange, 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. 58: 1323–1327.

Terwilliger, J.D., M. Speer & J. Ott, 1993. Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet. Epidemiol. 10: 217–224.

Wijsman, E.M., 2003. Summary of Group 8: Development and Extension of Linkage Methods. Genet Epidemiol. 25(Supl. 1): S64–S71.