UNIVERSIDAD NACIONAL DE COLOMBIA

ANÁLISIS MEDIANTE INTELIGENCIA COMPUTACIONAL DE MARCADORES GENÉTICOS EN PACIENTES COLOMBIANOS CON ENFERMEDAD DE ALZHEIMER

EDGAR ALEXANDER OSPINA GRANADOS

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2014

ANÁLISIS MEDIANTE INTELIGENCIA COMPUTACIONAL DE MARCADORES GENÉTICOS EN PACIENTES COLOMBIANOS CON ENFERMEDAD DE ALZHEIMER

EDGAR ALEXANDER OSPINA GRANADOS

Tesis presentada como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas y Computación

Director:
Luis Fernando Niño Vásquez, Ph.D.

Codirector:
Humberto Arboleda Granados, M.D., M.Sc.

Grupos de Investigación:

Laboratorio de Investigación en Sistemas Inteligentes (LISI) – GRUPO DE NEUROCIENCIAS

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2014

Dedicada a mi esposa y mis padres.

Agradecimientos

A los profesores Luis Fernando Niño Vásquez, Director del Grupo de Investigación LISI, y Humberto Arboleda Granados, Director del Grupo de Neurociencias, quienes durante el desarrollo de este proyecto brindaron el mejor soporte técnico y humano. A los miembros del Laboratorio de Genética y a los compañeros de los grupos de investigación mencionados, quienes con sus aportes ayudaron a estructurar este alcance.

Resumen

Un objetivo principal de la genética humana es comprender la relación entre las variaciones en las secuencias de ADN y la susceptibilidad a ciertas enfermedades. En este trabajo en particular, la información genética se analiza en relación con la enfermedad de Alzheimer (EA), con el fin de mejorar su diagnóstico, prevención y tratamiento. En Colombia, esta enfermedad requiere actualmente una atención especial debido a que su incidencia ha aumentado significativamente en los últimos años. Por esta razón, este trabajo analiza un conjunto de doce marcadores genéticos o polimorfismos de nucleótido simple (SNP) en un grupo de pacientes colombianos con EA a través de un método de inducción constructiva basado en un enfoque de aprendizaje de máquina, denominado reducción de dimensión multifactorial (MDR). Además, se realizaron análisis de epistasia estadística, obteniendo la relación de sinergia entre las variables mediante la ganancia de la información de los genes relacionados con la AD, proporcionando una metodología simple para caracterizar las interacciones en los estudios de asociación genética y determinar los rasgos más importantes que describen el comportamiento de la enfermedad.

Palabras clave: Interacciones entre genes, epistasia estadística, inteligencia computacional, enfermedad de Alzheimer.

Abstract

A main goal of human genetics is to understand the relationship between variations in DNA sequences and the susceptibility to certain illnesses. In this particular work, genetic information is analyzed in relation to the Alzheimer's disease (AD) in order to improve its diagnosis, prevention and treatment. In Colombia, this disease currently requires special attention because its incidence has increased significantly in recent years. Thus, this work analyzes a set of twelve genetic markers or single nucleotide polymorphisms (SNPs) in a set of colombian patients through a constructive induction method based on a machine

learning approach, namely, multifactor dimensionality reduction (MDR). Also, some statistical epistasis analysis is carried out. Particularly, epistasis is obtained based on information gain from AD related genes, providing a simple methodology to characterize interactions in genetic association studies and capturing important traits that describe the behavior of the disease.

Keywords: Gene to gene interactions, statistic epistasis, computational intelligence, Alzheimer's disease.

Contenido

Pág.

Resumen ..... **V**

Lista de figuras ..... **IX**

Lista de tablas ..... **X**

Lista de abreviaturas ..... **XI**

Introducción **1**

1. Enfermedad de Alzheimer **5**
1.1 Genética de la enfermedad 6
1.2 Descripción de los genes de estudio 11
1.2.1 APOE apolipoprotein E 11
1.2.2 CR1 componente del complemento (3b/4b) del receptor 1 12
1.2.3 BIN1 puente integrador 1 12
1.2.4 CLU clusterina 12
1.2.5 SORL1 sortilin-related receptor 13
1.2.6 PICALM proteína de ensamblaje clatrina unión fosfatidilinositol 13
1.3 Obtención del conjunto de datos usado 13

2. Inteligencia computacional en estudios de asociación **17**
2.1 Descripción del problema 19
2.2 Enfoques relacionados 20
2.2.1 Reducción de Dimensionalidad Multifactorial (MDR) 20
2.2.2 Métodos bayesianos 21
2.2.3 Medidas de desempeño 22
2.2.4 Ganancia de información 23
2.2.5 Filtros para selección de atributos 25

3. Metodología propuesta para el análisis de marcadores genéticos de la enfermedad de Alzheimer **28**
3.1 Análisis descriptivo del conjunto de datos 30
3.2 Selección de atributos 32
3.3 Inducción constructiva 35
3.4 Relación de sinergia entre los atributos 37

4. Resultados y discusión **39**
4.1 Atributos seleccionados 42
4.2 Combinaciones asociadas al riesgo mediante inducción constructiva 44

4.3 Variables epistáticas mediante ganancia de información..... 46

5. Conclusiones y recomendaciones 51

A. Anexo: Combinaciones de dos y tres SNP obtenidas mediante inducción constructiva para todas las variables..... 53

Bibliografía 61

Lista de figuras

Pág.

Figura 3-1: Metodología propuesta. 29
Figura 3-2: Composición de la muestra original. 31
Figura 3-3: Algoritmo Relief 33
Figura 3-4: Cálculo de la función $\mathrm{diff}(A,I1,I2)$ para atributos nominales 34
Figura 3-5: Algoritmo ReliefF 34
Figura 3-6: Algoritmo TuRF 35
Figura 3-7: Ejemplo de tablas de contingencias para genotipo *multi-locus*. 36
Figura 4-1: Frecuencias genotípicas para la muestra sin faltantes..... 39
Figura 4-2: Valor promedio de los puntajes obtenidos con los algoritmos RelifF, TuRF y SURF 43
Figura 4-3: Mapa de interacciones para el conjunto de seis variables obtenidas por los filtros ReliefF, TuRF y SURF..... 46
Figura 4-4: Mapa de interacciones para todas las variables 47

Lista de tablas

Pág.
Tabla 1-1: SNP objeto de estudio 15
Tabla 3-1: Variables del proyecto. 31
Tabla 4-1: Equilibrio de Hardy Weinberg calculado sobre las variables de entrada. ... 40
Tabla 4-2: Reglas de asociación encontradas mediante el algoritmo a priori. 41
Tabla 4-3: Posicionamiento obtenido mediante los algoritmos ReliefF, TuRF y SURF 42
Tabla 4-4: Combinaciones de dos SNP obtenidas mediante inducción constructiva para las seis variables obtenidas con los filtros. 44
Tabla 4-5: Combinaciones de tres SNP obtenidas mediante inducción constructiva para las seis variables obtenidas con los filtros. 45

Lista de abreviaturas

Abreviaturas

Abreviatura	Término
ADN	Ácido desoxirribonucleico
EA	Enfermedad de Alzheimer
MDR	Multifactor Dimensionality Reduction
SNP	Polimorfismo de nucleótido simple
APOE	Apolipoproteína E

Introducción

Una meta central de la genética humana es entender la relación entre las variaciones de la secuencia de ADN y la susceptibilidad a cierta enfermedad, para este caso particular en relación con la enfermedad de Alzheimer (EA), con el fin de mejorar el diagnóstico, prevención y tratamiento de la misma. Es acertado suponer que las enfermedades comunes deben ser controladas por mecanismos genéticos más complejos caracterizados por la acción conjunta de varios genes, teniendo cada gen un solo efecto marginal pequeño, tal vez porque la selección natural ha eliminado los genes que producen los efectos más grandes, y en este escenario, se deben analizar conjuntamente grupos de marcadores para el desarrollo de estudios de asociación, mediante la agrupación de marcadores en genotipos *multi-locus* para que la unidad básica del análisis estadístico sea el individuo [1].

Esta interacción entre los diferentes genes al expresar un carácter fenotípico, se denomina epistasia [64] y ha sido reconocida por muchos años como la desviación de las proporciones de segregación mendeliana o alteraciones de aditividad en un modelo estadístico lineal y es probable en parte a mecanismos de estabilización de selección en la evolución de las redes genéticas [24-25]. Particularmente, la epistasia de carácter estadístico se produce a nivel de la población cuando hay variaciones interindividuales en secuencias de ADN, y es difícil de detectar y caracterizar en estudios en seres humanos debido a su inherente no linealidad, que puede derivarse de fenómenos como la heterogeneidad del *locus*, la fenocopia y la dependencia de efectos genotípicos en factores ambientales (es decir, las interacciones gen-ambiente) y genotipos en otros *loci* (es decir, interacciones gene–gene o epistasia) [23]. El modelado de dichas interacciones no lineales requiere de métodos analíticos especiales ya que los enfoques estadísticos paramétricos, como la regresión logística, son útiles en la detección de efectos principales independientes, pero tienen bajo poder de detectar interacciones no lineales. Adicional a esto, los modelos que relacionan las combinaciones de polimorfismos de nucleótido simple, en adelante SNP, con la susceptibilidad de la enfermedad son intrínsecamente

difíciles de interpretar debido a su dimensión [26, 27]; por ejemplo, los análisis de genoma humano presentan dificultades para su análisis e interpretación, sobre todo cuando se incluyen grandes cantidades de SNP y de individuos [41].

La interpretación de la interacción de genes en la enfermedad de Alzheimer es un reto importante, teniendo en cuenta que los índices de las enfermedades de tipo neurodegenerativo y mental se incrementa en Colombia. Aunque el principal factor de riesgo de la enfermedad es la edad [45], de acuerdo con diversos estudios se ha podido identificar que la EA puede ser consecuencia de factores genéticos y/o ambientales, siendo esta premisa demostrada mediante estudios de EA que han divulgado resultados donde se presenta una fuerte asociación con la codificación de polimorfismos del gen y promotor de la apolipoproteína E (APOE) en muestras colombianas. Sin embargo, dichos polimorfismos explican sólo una fracción del riesgo genético asociado con la enfermedad de Alzheimer, y es posible que las variantes en genes adicionales puedan conferir un mayor riesgo para desarrollar la enfermedad, por lo cual es factible que futuras combinaciones de plataformas de genotipado de alto rendimiento, modelos de análisis multivariado y técnicas computacionales puedan llevar a la identificación de otros factores de susceptibilidad genética en los pacientes de la población colombiana [2], incrementando la importancia en el desarrollo de aplicaciones y profundización en las técnicas de análisis anteriormente mencionadas.

Como un problema de carácter nacional, la enfermedad ha sido abordada desde las diferentes áreas del conocimiento, especialmente la medicina, la estadística y la informática. Como soporte a esa iniciativa, el presente trabajo tiene como objetivo analizar mediante inteligencia computacional marcadores genéticos para un conjunto de pacientes colombianos con dicha enfermedad. Para el logro de este fin se ha propuesto diseñar e implementar una metodología que logre extraer conocimiento de la muestra de casos y controles, y efectuar la comparación de los resultados obtenidos frente a los realizados en estudios con poblaciones de otros países específicamente para características de la enfermedad.

El estudio se fundamenta en los proyectos realizados por el Instituto de Genética de la Universidad Nacional de Colombia, para los cuales se conformó una base de datos con pacientes que padecían EA (casos) y personas con características similares sin la enfermedad (controles). A partir de estos datos se han generado interpretaciones de la enfermedad desde el punto de vista clínico, y como complemento a ese avance en este documento el producto final se estructuró atendiendo los requerimientos de analizar exclusivamente las variables genéticas de los individuos con el fin de descubrir patrones que aporten al conocimiento de la enfermedad.

En el primer capítulo se describe la enfermedad de Alzheimer y cómo la genética influye en su detección mediante los marcadores CR1rs3818361, BIN1rs744373, CLUrs2279590, CLUrs11136000, SORL1rs1121830, PICALMrs3851179, GWArs11622883, PVRL2rs6859, TOMM40rs2075650, APOErs440446, APOErs429358 y APOErs7412. En el segundo capítulo se describen los avances en inteligencia computacional enfocados en la obtención de patrones en muestras genéticas como son las redes estadísticas, aprendizaje de máquina supervisado y regresión logística. Posteriormente, se describe la metodología donde se quiere entender la relación entre las variaciones en la secuencia de ADN y la susceptibilidad a ciertas enfermedades, para este caso particular en lo referente a la enfermedad de Alzheimer, para mejorar el diagnóstico, prevención y tratamiento haciendo uso del conjunto de datos de los doce polimorfismos de nucleótido simple en pacientes colombianos a través de una inducción constructiva y aprendizaje de máquina para reducir la dimensionalidad de la muestra (MDR) y un software de representación de redes para describir la epistasia estadística. Finalmente, se describe cómo en el presente estudio se logró representar esta epistasia en relaciones de dos y tres SNP basadas en la ganancia de la información, entropía e información mutua de las variables, proporcionando una metodología simple para caracterizar las interacciones en los estudios de asociación genética y la captura de características importantes que describen el comportamiento de la enfermedad.

1. Enfermedad de Alzheimer

La enfermedad Alzheimer es la principal clase de demencia en el mundo, aunque el impacto médico y social de la enfermedad se proyecta más en los países en desarrollo, sus factores etiológicos específicos en estas poblaciones con diversos antecedentes genéticos y ambientales permanecen inexplorados [4, 28, 29]. Se define como una combinación de déficit cognitivo lentamente progresivo, síntomas psiquiátricos y lesiones macroscópicas y microscópicas del cerebro manifestándose primero como deterioro cognitivo leve amnésico antes de progresar a una demencia, en que el deterioro de la memoria generalmente sigue siendo el déficit cognitivo más prominente. Gran cantidad de estudios clínico-patológicos hospitalarios han demostrado que los efectos producto de los depósitos amiloideos fibrilares, entendidos estructuralmente como filamentos anormales de proteína en placas amiloideas, y los ovillos neurofibrilares compuestos por la proteína Tau [8], son posiblemente los causantes de la enfermedad [6, 10].

Su descubrimiento inicia con observaciones y estudios realizados por Tomlinson y colaboradores [46], quienes indicaron la presencia de lesiones sustanciales en personas ancianas cognitivamente intactas; luego estos resultados fueron complementados por la investigación de Braak y su grupo de trabajo [47], quienes lograron demostrar el aumento gradual de los depósitos amiloideos en el envejecimiento cerebral. Posteriormente, luego del análisis de cerebros afectados por la enfermedad, la proteína TAU fue identificada como el principal constituyente de los ovillos neurofibrilares (agregados de proteína tau hiperfosforilada), y Yanker et. al., descubrieron las propiedades neurotóxicas de la proteína amiloide beta [6].

El estudio riguroso de la enfermedad continuó en la década de los 70 con la identificación de los primeros blancos terapéuticos para el desarrollo de fármacos; las divisiones del lóbulo temporal medial tomaron importancia para la descripción detallada de los patrones de la atrofia asociada con la pérdida progresiva de la memoria. En la década de los 90, se identificaron los primeros genes que confieren riesgo en la aparición temprana de la

enfermedad y en su fase tardía (apoliprotein APO ε4, proteína precursora del amiloide beta. Recientemente, luego de estos descubrimientos se han identificado polimorfismos en otros genes, aparentemente implicados en la separación y procesamiento del amiloide beta, sustentados en los estudios del genoma, y observaciones que posteriormente fueron confirmadas por otros grupos en diversas muestras clínicas [6]. La enfermedad de Alzheimer, es pues, la forma más común de demencia siendo una enfermedad degenerativa incurable y terminal, frecuente en personas mayores de 65 años aunque pueda presentarse mucho antes de esta edad [7].

La enfermedad afecta la mayoría de las personas (hasta el 75% de más de 35 millones que sufren de demencia en todo el mundo), y se cree que su prevalencia puede duplicarse cada 20 años, estimando que en el 2050 aproximadamente 115 millones de personas pueden verse afectadas. El impacto de la enfermedad no solamente radica en las víctimas sino también recaen las personas que cuidan de ellos, por lo que se considera de gran importancia para el sistema de salud de cada país. Los factores de origen son en su mayoría desconocidos aunque existe una creciente evidencia de que aspectos psicosociales, vasculares y trastornos conforman las causas de la enfermedad [7]; y en la actualidad se reconocen los factores genéticos como parte importante de esta demencia, contextualizándola como una serie continua de interacciones entre los diferentes grados de la variable genética versus la influencia ambiental. Se cree que estas variables genéticas intervienen en la tendencia que tienen proteínas específicas de acumularse de forma anormal y depositarse en el cerebro, por lo que la intervención en esta área puede disminuir el riesgo de su desarrollo o al menos retrasar la manifestación clínica [8].

1.1 Genética de la enfermedad

Los descubrimientos en genómica están transformando la ciencia médica, proporcionando nuevos métodos para predecir la ocurrencia o avance de enfermedades que tienen una base genética [48]. Debido a su importancia en los temas de salud pública, existe gran interés en la identificación de genes de susceptibilidad relacionados con enfermedades comunes, que proporcionen información sobre el riesgo debido a su interacción compleja y a la intervención de los factores ambientales; con el soporte de la certeza que ofrecen

las variantes mendelianas altamente penetrantes, donde la presencia o ausencia de mutaciones específicas es claramente asociada con la manifestación futura de determinada enfermedad [9].

Desde la finalización de la secuenciación del genoma humano, se ha evidenciado el avance en la descripción de la variación genética humana; la más común en el genoma humano es el polimorfismo de nucleótido simple (SNP) [1]. Existen en el genoma al menos 10 millones de SNP con frecuencia > 1%, que se cree son responsables de alrededor del 90% de la variación genética humana. Actualmente, hay más de 10 millones SNP en el repositorio público de las variaciones del ADN, dbSNP. Con este completo catálogo de SNP disponibles, los investigadores están desarrollando métodos para ser aplicados en estudios de asociación genética, con el objetivo de identificar variantes de ADN que tienen influencia en la predisposición a enfermedades humanas, por medio de la re-secuenciación de genes o regiones genómicas en varias muestras de la población [11].

Existen algunos ejemplos en los cuales las variantes comunes son constantemente asociadas con un fenotipo común de la enfermedad por ejemplo, CARD15 y la enfermedad de Crohn, y APOE y la enfermedad Alzheimer [1]. Sin embargo, algunos investigadores sostienen que no todas las enfermedades humanas pueden atribuirse a variantes comunes, y que es más probable que varias variantes raras en varios sitios (heterogeneidad alélica y genética) podrían reflejarse de manera más evidente en los fenotipos estudiados. La temática se torna más compleja cuando se analiza el término “variante común” ya que algunos investigadores lo definen como un alelo con > 20% MAF, y otros como un alelo con > 1% MAF. En general, los factores como la historia de la población, la selección natural y la recombinación genética hacen difícil predecir a priori qué regiones del genoma humano tendrán niveles de información aceptable para estudios de asociación genética; más aún cuando, aproximadamente el 70% del riesgo de un individuo determinado para el desarrollo de EA es conferido a través de su repertorio genético [10]. Existen relativamente pocos polimorfismos únicos recientemente descubiertos que alteran el riesgo de EA, como los alelos en CLU, SORL1, PICALM y BIN1; pero su penetrancia (*penetrance*, en inglés), que es la probabilidad condicional de que una persona seleccionada al azar de la población posea la enfermedad, es mucho

más débil, es decir, el efecto de la mutación en el riesgo para el fenotipo de la enfermedad es menos previsible en comparación con las mutaciones en APOE [8].

Para los procesos de análisis de muestras de la enfermedad, se utilizan los casos y controles cuya característica radica en que los sujetos incluidos en la muestra se seleccionan aleatoriamente de una población dada por su condición de enferma o afectada. Los marcadores genéticos de los individuos pertenecientes a los dos grupos, casos y controles, se comparan con el objetivo de que sus diferencias, en algunas regiones estrechas del genoma, puedan ofrecer una explicación causal para el estado de la enfermedad. En este entorno se cuenta con los dos alelos o los tres genotipos de un *locus* y se comparan en los dos grupos, afectados y controles. Si hay una diferencia en las frecuencias entre las dos muestras, hay evidencia de que el marcador está en desequilibrio de ligamiento con el gen afectando la susceptibilidad de la enfermedad [11].

En los últimos años, el estudio de los marcadores genéticos ha sido revolucionado por el éxito de las asociaciones encontradas en el genoma. La mayoría de estos estudios ha utilizado una estrategia de análisis de *locus* simple, en el que cada variante se prueba individualmente para la asociación con un fenotipo específico. Sin embargo, una razón por la que a menudo se cita la falta de éxito en estos estudios genéticos de enfermedades complejas es la existencia de interacciones entre los *loci*. Si un factor genético funciona principalmente a través de un mecanismo complejo que involucra múltiples genes distintos y, posiblemente, factores ambientales, el efecto podría perderse si el gen es examinado aisladamente sin tener en cuenta sus interacciones potenciales con estos otros factores desconocidos [12].

Para tener mayor claridad sobre los conceptos es necesario describir algunos términos básicos [14].

Heterocigótico. Un individuo es heterocigótico en una ubicación del gen si tiene 2 alelos diferentes uno en el cromosoma materno y otro en el paterno.

Homocigótico. Un individuo es homocigótico en una ubicación del gen si tiene 2 alelos idénticos en esa ubicación.

Acoplamiento. La tendencia de genes u otras secuencias de ADN en *loci* específicos a heredarse juntos como consecuencia de su proximidad física en un solo cromosoma.

Desequilibrio de Ligamiento. Una medida de asociación entre alelos en diferentes *loci*.

***Locus/loci*.** El sitio o sitios en un cromosoma en el que se encuentra el gen para un rasgo particular o un gen en el que se encuentra un SNP particular.

Mutación. Una variante rara en un gen, que ocurre en el 1% de una población.

Polimorfismo. Existencia de dos o más variantes de un gen, que ocurren en una población con al menos 1% frecuencia de la variante menos común.

SNP. Abreviatura de polimorfismo de nucleótido simple, un solo par de bases cambia en la secuencia de ADN en un determinado punto en comparación con la secuencia común o natural (*wild type,* en inglés). Los científicos han catalogado más de 12 millones de SNP a la fecha. Algunos de ellos pueden cambiar la secuencia de aminoácidos de la proteína resultante. Otros SNP están en áreas del cromosoma que directamente no codifican proteínas, pero aún pueden influir en la función de la célula a través de otros medios, como el control de la cantidad de proteína que la célula construye. Dado el gran número de SNP, su nomenclatura puede ser confusa, pero el sistema más común utiliza un número con el prefijo "rs" por ejemplo, rs1228756. Las diferentes formas o variantes que puede llevar un polimorfismo particular se llaman alelos. Hay varias ventajas de usar datos de SNP. En primer lugar, los SNP cambian de manera poco probable con el tiempo, es decir, el patrón de SNP de un paciente suele ser el mismo desde su nacimiento o en cualquier fase de su vida. En segundo lugar, se pueden extraer de cualquier tejido en el cuerpo por lo que es relativamente fácil en comparación con los microarreglos [15].

En los estudios de casos y controles se compara la frecuencia de alelos SNP en dos grupos bien definidos: casos, que han sido diagnosticados con la enfermedad bajo estudio, y controles, que se sabe que no son afectados o que han sido seleccionados al azar de la población. Una mayor frecuencia de un SNP encontrado en los casos comparada con la encontrada en los controles indica que la presencia del alelo SNP puede aumentar el riesgo de enfermedad. El principal problema en estudios de casos y controles es asegurar a un buen ajuste entre la información genética de casos y controles, para que cualquier diferencia genética entre ellos esté relacionada con la enfermedad bajo estudio y no a un sesgo en la muestra [16].

Epistasia. La definición estadística de epistasia fue dada en [21] como las desviaciones de efectos aditivos en un modelo estadístico lineal. Este concepto se ha generalizado como la interacción entre genes, y se ha convertido en un tema popular en genética humana en los últimos diez años [22]. Existe una creencia creciente de que la susceptibilidad a enfermedades comunes puede ser regida por la posible interacción entre múltiples variantes genéticas, y ésta es impulsada en gran parte por la idea de que redes bioquímicas y procesos de regulación que involucran varios genes, tienen un extremo funcional que puede estar influenciado por la presencia simultánea de múltiples variantes en dichos genes [42,69]. Además de esta importancia teórica, la epistasia ha demostrado funcionalmente que juega un papel importante en las enfermedades humanas, como es el caso de la enfermedad de Hirschsprung [72], la esquizofrenia [73], la artritis reumatoide [74], el cáncer de vejiga [78], la tuberculosis [79], el glaucoma [80], la fibrilación auricular [81] y la enfermedad de Alzheimer en los genes GAB2 y APOE [82].

La presencia de epistasia es una característica relevante en el proceso, ya que, si el efecto de un lugar geométrico es alterado o enmascarado por efectos en otro lugar geométrico, la capacidad estadística para detectar el primer lugar geométrico se reduce y la identificación de los efectos conjuntos en los dos lugares geométricos se confunde por su interacción. Si están involucrados más de dos lugares geométricos, es probable que se complique aún más por la posibilidad de interacciones complejas entre algunos o todos los lugares [22].

1.2 Descripción de los genes de estudio

El estudio se realizó sobre la muestra de estudio procedente del grupo de investigación en el área genética de la Universidad Nacional de Colombia. Esta muestra está conformada por pacientes que padecen EA (denominados casos) y personas de características similares desprovistas de la enfermedad en mención (denominados controles). Las variables o atributos corresponden a marcadores genéticos que han sido identificados en la literatura como posibles causantes de la enfermedad, por esta razón, se incluyeron como objeto de estudios clínicos. Asimismo, partiendo de que la intención es lograr predecir el riesgo de presentar la enfermedad, se incluye también la clase enfermo/sano como variable del análisis. A continuación se describirán los genes de estudio a los cuales pertenecen los marcadores seleccionados.

1.2.1 APOE apolipoprotein E

El APOE se encuentra ubicado en el cromosoma 1 3597bp 9 en un grupo con APOC1 y APOC2. Es una proteína principal de chylomicron que se une a un receptor específico sobre las células del hígado y las células periféricas y es esencial para el catabolismo normal de los componentes de lipoproteínas ricas en triglicéridos. Está ubicado en la región 19q13.2 y consta de cuatro exones y tres intrones en 3597bp. La proteína presenta tres isoformas: APOE ε4, representada por el marcador rs429358, se encuentra asociada al desarrollo de patologías como la EA; APOE ε3 representada por el marcador rs440446; y APOE ε2 representada por el marcador rs7412 que es considerada un factor protector para la EA (por razones históricas, no hay ningún ε1). De los alelos de APOE ε3 es el más frecuente en las poblaciones blancas (78%) y representa el alelo común, ε2 y ε4 son alelos variantes [14].

Fisiológicamente, la proteína APOE lleva una forma de colesterol y se une al receptor de APOE en la superficie de las células para que el colesterol sea metabolizado. De las 3 isoformas de la proteína que se derivan de los 3 alelos correspondientes de APOE, la isoforma ε2 ha disminuido la fuerza de enlace, o afinidad al receptor de APOE. Las proteínas resultantes de los alelos ε3 y ε4, tienen una mayor afinidad en comparación con ε2. Estas isoformas son definidas por cambios en 2 puntos en el ADN, que conducen a

cambios en los aminoácidos de posiciones 112 y 158 que componen la proteína APOE. El alelo ε4 es un fuerte factor de riesgo para EA esporádica y tardía; el grado de riesgo varía dependiendo de si el individuo lleva uno o dos alelos ε4. Estos descubrimientos de nuevos genes de susceptibilidad en la última década han aumentado el conocimiento de posibles determinantes genéticos de EA, pero ninguno de ellos ha sido replicado con tanta frecuencia ni tiene como notable un efecto de riesgo como APOE [14].

1.2.2 CR1 componente del complemento (3b/4b) del receptor 1

Este gen es un miembro de los receptores de la familia de activación del complemento (RCA) y se encuentra en la región “grupo RCA” del cromosoma 1. El gen codifica una glicoproteína de membrana tipo I en los eritrocitos, leucocitos, podocitos glomerulares y células dendríticas foliculares esplénicas. La disminución en la expresión de esta proteína y/o las mutaciones en sus genes se han asociado con los carcinomas de vesícula biliar, lupus eritematoso sistémico y sarcoidosis. Mutaciones en este gene también se han asociado con una reducción en la protección contra la malaria severa. En estudios recientes se ha logrado identificar su relación con procesos neurodegenerativos.

1.2.3 BIN1 puente integrador 1

Se encuentra ubicado en el cromosoma 2. Este gen codifica varias isoformas de una proteína adaptadora núcleo-citoplasmática, uno de los cuales fue identificado inicialmente como una proteína interactuando con las características de un supresor tumoral. Algunas isoformas que se expresan en el sistema nervioso central pueden estar implicadas en la endocitosis de vesículas sinápticas. Estudios en ratones sugieren que este gen juega un papel importante en el desarrollo del músculo cardiaco. Cuenta con 19 exones y la función de BIN1 respecto a la EA, aún no está definida, a pesar de que los estudios de asociación con enfermedad se han replicado en las poblaciones africana y europea.

1.2.4 CLU clusterina

Se encuentra en el cromosoma 8 en la región 8p21-p12 y se encuentra organizado en nueve exones. La proteína codificada por este gen puede bajo ciertas condiciones de

estrés encontrarse en el citosol de la célula. Se ha sugerido que participa en varios eventos biológicos básicos, tales como la muerte celular, la progresión del tumor, y trastornos neurodegenerativos.

1.2.5 SORL1 sortilin-related receptor

Este gen codifica una proteína de mosaico que pertenece a por lo menos dos familias: la proteína vacuolar de clasificación 10 (VPS10) y la familia de receptores (LDLR) de lipoproteína de baja densidad. La proteína codificada probablemente juega un papel en la endocitosis y puede haber una asociación entre la expresión de este *locus* y la enfermedad de Alzheimer. Se encuentra en el cromosoma 11, en la región 11q23.2-q24.2 y codifica para una glicoproteína expresada en muchas neuronas del córtex, hipocampo, cerebelo y cordón espinal.

1.2.6 PICALM proteína de ensamblaje clatrina unión fosfatidilinositol

Este gen codifica una proteína de ensamble de clatrina, que incorpora complejo 2 (AP2) de la proteína clatrina y lo adapta a las membranas. La proteína puede ser necesaria para determinar la cantidad de membrana que debe reciclarse. Está implicado en la endocitosis mediada por clatrina de AP2-dependiente en la unión neuromuscular. Ayuda a establecer la polaridad y control del crecimiento de axones y dendritas en neuronas hipocampales embrionarias. Los polimorfismos de este gen están asociados con el riesgo de enfermedad de Alzheimer, por su intervención en el crecimiento de las neuronas.

1.3 Obtención del conjunto de datos usado

En Colombia; entre los estudios que se han desarrollado en torno a la EA, se destaca principalmente el realizado en la de la Universidad Nacional por el Grupo de Neurociencias, quienes han trabajado con el gen APOE y la EA en una muestra restringida de la población colombiana caucásico-mestiza. Sin embargo, estas variaciones sólo explican una fracción de los riesgos genéticos asociados con EA y polimorfismos en otros genes pueden conferir riesgo adicional para desarrollar EA [83-84].

La base de datos AlzGene [85] lista cada día nuevas variantes genéticas con asociación significativa, estas asociaciones se basan en asociaciones alélicas y meta-análisis de todos los datos disponibles publicados en el tema, incluyendo estudios de gran impacto como los GWAS. Los marcadores en un GWAS consisten de SNP que son elegidos con base en su capacidad para cubrir variaciones comunes en el genoma humano, en estos también pueden incluirse los cálculos de supresiones o multiplicaciones de ciertos segmentos del cromosoma de longitud variable (*copy-number variants*, en inglés). Desde el 2007 se cuenta con este tipo de estudios que ha servido como base para realizar diferentes estudios de tipo caso-control en búsqueda de réplicas de asociación en diferentes marcadores.

El conjunto de datos objeto del presente trabajo es producto de estudios de asociación realizados por el Grupo de Neurociencias de la Universidad Nacional de Colombia en los que se han presentado genes implicados con mecanismos biológicos que tratan de explicar la aparición y desarrollo de la enfermedad. El grupo cuenta con cerca de 600 muestras entre pacientes con diagnóstico de demencia (casos) y personas sin antecedentes propios ni familiares de enfermedades neurodegenerativas y/o psiquiátricas (controles). Esta información se ha obtenido mediante muestras de sangre de pacientes y controles, y ha sido tomada con previo consentimiento informado. Las historias clínicas de los pacientes, controles y familiares permanecen de manera confidencial en las instalaciones en el Instituto de Genética de la Universidad Nacional. Las muestras de sangre tomadas son codificadas y almacenadas a -70 º C hasta su utilización. En estas muestras se han analizado SNP en los genes CLU, PICALM, CR1, APOE, TOMM40, PVRL2, SORL1, BIN, los cuales fueron escogidos después de una revisión bibliografía de estudios de genoma completo (GWAS) y réplicas que hayan reportado asociación significativa con la EA en regiones distintas a Colombia. Además, se tuvieron en cuenta estudios funcionales que demostraban potencial asociación funcional de los genes candidatos con vías metabólicas involucradas en el desarrollo de la patología [84]. Los correspondientes SNP se presentan a continuación en la Tabla 1.

Tabla 1-1: SNP objeto de estudio

GEN	SNP	CROMOSOMA	LOCALIZACIÓN	ALELOS	MAF	OR 95% (CI)
CLU	rs11136000	8	Intrón	C/T	T	0.86 (0.82-0.90)
PICALM	rs3851179	11	5'	A/G	A	0.87 (0.83-0.91)
CLU	rs2279590	8	Intrón	A/G	A	0.86 (0.81,0,90)
BIN1	rs744373	2	---	C/T	C	1.17 (1.13,1.20)
TOMM40	rs2075650	19	Intrón	A/G	G	2.79 (2.38,3.27)
PVRL2	rs6859	19	3' UTR	A/G	A	1.50 (1.39,1.62)
APOE	rs440446	19	Intrón	C/G	C	0.57(0.50,0.65)
SORL1	rs11218304	11	Intrón	A/G	G	1.13 (1.02,1.25)
CRI	rs3818361	1	Intrón	C/T	T	1.19 (1.11-1.26)
GWA 14 q32.13	rs11622883	14	---	A/T	A	0.84 (0.69,1.06)

Para la obtención de los genotipos de cada instancia caso/control se desarrolló por parte del grupo de Neurociencias una metodología que incluyó las siguientes actividades [83]:

- Aislamiento genómico. Se evalúa la calidad del ADN genómico extraído y conservado a 4°C de pacientes y controles por medio de electroforesis en geles de agarosa 0.8%, teñidos con SYBR® Safe DNA Gel y visualizados con luz U.V, para esto se usó 5µl de muestra de DNA y 5µl de buffer de carga (azul de bromofenol 2x).

- Diseño de primers. Se diseñaron tres cebadores (*primers*, en inglés) para cada una de las variaciones de SNP, fueron utilizados para la amplificación por Multiplex polymerase chain reaction (multiplex PCR) y su diseño se hizo mediante el programa MPprimer (http://biocompute.bmi.ac.cn/MPprimer/) y Primer 3 (http://frodo.wi.mit.edu/).

- Multiplex PCR. Las multiplex PCR fueron estandarizadas de acuerdo con su temperatura de fusión (*melting temperatura*, en inglés) y concentración de cloruro de magnesio. Fueron realizadas en el grupo de pacientes y controles. Las verificaciones de amplificación se realizaron en geles de agarosa al 1.5%, teñidos con SYBR® Safe DNA Gel y visualizado con rayos ultra violeta.

- Técnica de identificación de los SNP (SNaPshot). Los procesos de identificación de SNP se realizaron con el secuenciador automático ABI PRISM 3500 y el sistema *Multiplex* SNaPshot®

2. Inteligencia computacional en estudios de asociación

En la actualidad los métodos de inteligencia computacional han adquirido importancia en la solución de problemas en temas relacionados con las ciencias exactas, que por su grado de complejidad, impiden a las metodologías estadísticas o matemáticas ofrecer una solución rápida y fácil de implementar, e incluso en algunos casos no pueden brindar soluciones a problemas particulares.

Con todos los proyectos de secuenciación de genoma realizados, los datos genéticos tales como secuencias de ADN, proteínas, estructuras moleculares, entre otros, crecen exponencialmente. Las técnicas de inteligencia computacional que combinan elementos de aprendizaje y minería de datos se encuentran muy bien adaptadas para muchos de estos problemas que se presentan en biología, ya que tienen capacidad para manejar grandes volúmenes de datos de la vida real con ruido, ambigüedad, falta de datos y procesamiento de información cambiante. En computación biológica a menudo se requiere buscar algunas características o patrones en los extensos conjuntos de datos que normalmente se caracterizan por su alta dimensión y porque las muestras cuentan con pocos individuos [5]. Es por esto que se hace necesario el desarrollo de enfoques más eficientes para análisis de patrones que los métodos tradicionales, y la inteligencia computacional es una importante herramienta para este fin.

La Inteligencia Computacional combina elementos de aprendizaje, adaptación, evolución y lógica para crear programas que pueden por sí mismos lograr un objetivo específico. Asimismo, tiene la capacidad para aprender o para afrontar situaciones nuevas que podrían asociarse con atributos propios de la razón como lo son la generalización, el descubrimiento, la asociación y la abstracción. Autores como Bezdek [49], Marks [50] y Pedrycz [51] han definido la inteligencia computacional de diferentes maneras según los

desarrollos y acontecimientos luego del surgimiento de esta nueva disciplina, teniendo en cuenta que ésta es emergente y no asociada sólo a un número limitado de temas; más bien tiene la visión de expandirse en diversas direcciones y emerger con otras disciplinas existentes. La tendencia reciente está enfocada en integrar diferentes componentes para tomar ventaja de las características complementarias y desarrollar un sistema sinérgico.

Dentro de este contexto es pertinente mencionar el concepto de aprendizaje de máquina que consiste en programar el computador para optimizar un criterio de rendimiento utilizando datos de ejemplo o experiencia pasada. Generalmente se cuenta con un modelo definido y se implementa el aprendizaje ejecutando un programa informático que optimice los parámetros del modelo con datos de entrenamiento o experiencia pasada. El modelo puede ser predictivo para conocer comportamientos con datos futuros, descriptivo para obtener conocimiento de los datos o ambos, utiliza la teoría estadística en la construcción de modelos matemáticos, ya que la tarea principal es hacer inferencia de una muestra [30].

Asimismo, dentro del conjunto de objetivos que persigue la inteligencia computacional pueden mencionarse como principales el agrupamiento, que ubica a los individuos en clases de características similares; la asociación entendida como la tarea de descubrir las reglas para la cuantificación de la relación entre dos o más atributos; y la clasificación donde existe un número de clases y una muestra de individuos con varias características y el objetivo es asignar a cada individuo una de las clases de acuerdo con sus características, proceso en el cual la aplicación principal podría ser la obtención de una regla de clasificación para la predicción de instancias nuevas.

De acuerdo con lo anterior, la identificación de patrones asociados con la EA puede enfocarse hacia las técnicas de inteligencia computacional; por tal razón, a continuación se presenta el contexto del problema a resolver en el presente trabajo, y posteriormente se describen las alternativas utilizadas en este campo.

2.1 Descripción del problema

La EA es actualmente una de las enfermedades que presenta un riesgo elevado y variable dentro de las poblaciones humanas, impactando en la sociedad y en los sistemas de salud pública, siendo el tipo más común de demencia. Particularmente en Colombia existe la prevalencia en cuanto a enfermedades de este tipo, debido a que la tasa de supervivencia de la población adulta mayor está en crecimiento [84]. Por tal motivo, el interés por estudiar las diferentes características de ésta patología es cada vez mayor.

Tal como se mencionó en el capítulo anterior, existen numerosos estudios que demuestran que el padecimiento de esta enfermedad está relacionado con el comportamiento de algunos genes [85], lo que convierte a los estudios genéticos en un frente importante para la caracterización de la EA. Es allí donde surgen conceptos como el de epistasia, que permite describir las interacciones entre genes y facilita el entendimiento de las condiciones que propician la aparición o no de la enfermedad.

Sin embargo, dado que la EA es un ejemplo de enfermedad compleja, estas interacciones son difíciles de detectar y caracterizar usando métodos estadísticos paramétricos tradicionales, tales como la regresión lineal o logística (86).

Según el contexto anterior, para lograr analizar el conjunto de datos objetivo se deben identificar las interacciones y reglas de asociación de dichos atributos con el fenotipo. El conjunto de datos está conformado por 393 casos y 236 controles, cada uno con 12 SNP o atributos que representan tres tipos de genotipos (homocigotos de mayor frecuencia, homocigotos de menor frecuencia y heterocigotos); además para este conjunto de datos existe un porcentaje de datos faltantes. Las técnicas utilizadas para alcanzar este objetivo deben mostrar a los profesionales en ciencias de la salud, la relevancia y aporte de información de cada atributo, el riesgo que éstos pueden conferir, y las sinergias más relevantes entre ellos, que en dado caso podría aumentar la posibilidad de padecer la enfermedad.

2.2 Enfoques relacionados

La selección de variables que puedan llevar a predecir condiciones, en este caso particular, conocer que predisposición tiene una persona de padecer EA, es un problema que ha sido tratado por la inteligencia computacional y el aprendizaje de máquina. Son técnicas que no se ajustan a un modelo único predeterminado, intentan ser eficientes computacionalmente y evitan los problemas de sobreajuste mediante validación cruzada. Generalmente, estos métodos están referidos a algoritmos que tienen la capacidad de aprender de la experiencia y para el presente estudio el enfoque está dirigido al aprendizaje de máquina supervisado en el cual la variable de salida binaria, caso/control, guía el proceso de aprendizaje. El objetivo principal es predecir esta variable a partir de atributos de entrada dados (en esta caso un conjunto de SNP). Seguidamente, se describen algunas de las técnicas alternativas a los métodos estadísticos, en las cuales se resalta la capacidad de detectar interacciones no lineales en conjuntos de datos con alta dimensión y que han sido aplicadas en genética.

2.2.1 Reducción de Dimensionalidad Multifactorial (MDR)

MDR es un método de aprendizaje de máquina diseñado específicamente para identificar combinaciones de variaciones genéticas que interactúan y están asociadas con el incremento del riesgo a padecer enfermedades humanas comunes, complejas y multifactoriales [56,57,61,62]. En MDR no se calculan parámetros (método no paramétrico) y no asume algún modelo genético. El objetivo de MDR es encontrar una combinación de atributos asociados con el fenotipo (caso/control) tratando de reducir al mínimo el número de individuos mal clasificados. Ubica todos los genotipos en grupos de alto o bajo riesgo, pasando con esto los atributos a una única dimensión correspondiente al factor de riesgo. Este proceso donde un nuevo atributo (factor de riesgo) está definido como una función de otros dos o más atributos se denomina inducción constructiva [60,65,66]. A este nuevo atributo se le evalúa la habilidad de clasificar y predecir el estatus de la enfermedad si es alto riesgo es probable que sea un caso; de lo contrario será un control. MDR forma una hipótesis contando la frecuencia de diversas combinaciones de genes dentro de la muestra de entrenamiento, lo cual es análogo al funcionamiento de un clasificador bayesiano.

El algoritmo MDR [56] ayuda a identificar el fenómeno de epistasia [57,58,64], y desarrolla esta tarea específicamente mediante búsqueda exhaustiva junto con un clasificador para identificar la combinación óptima de los polimorfismos que se traduce en patrones de predicción de la enfermedad. Otros algoritmos de inducción constructiva, primero seleccionan combinaciones interesantes de polimorfismos mediante medidas de entropía [59] para evaluar y visualizar la ganancia de la información asociada, teniendo en cuenta las interacciones de atributo. Una vez seleccionados los SNP de mayor relevancia son utilizados para construir nuevos atributos con MDR y estos a su vez pueden ser evaluados usando cualquier método de aprendizaje de máquina. El objetivo final es crear o descubrir una representación que facilite la detección de interacciones no lineales o no aditivas entre los atributos tales que la predicción de la variable de clase se mejora sobre el de la representación original de los datos. La combinación de SNP que conforman los nuevos atributos se logra dando un umbral T definido como la relación entre el número de casos y el número de controles observados para un SNP particular. Una combinación de genotipos *multi-locus* es considerada de alto riesgo si $T \geq 1$, en caso contrario se considera de bajo riesgo.

2.2.2 Métodos bayesianos

Los Métodos Bayesianos son una clase de métodos estadísticos que tienen algunas propiedades atractivas para resolver los problemas de aprendizaje de máquina, particularmente cuando el proceso a modelar tiene aspectos inciertos o aleatorios. La salida de un análisis bayesiano es una distribución de probabilidad sobre una cantidad de interés. El uso de la probabilidad para representar la incertidumbre no tiene carácter obligatorio, pero es inevitable si se quiere respetar el sentido común al hacer inferencias coherentes y racionales, convirtiéndola en una manera satisfactoria de cuantificar la incertidumbre. Por ejemplo, en [52] se muestra que si los valores numéricos se utilizan para representar los grados de creencia, entonces un simple conjunto de axiomas que codifican propiedades de sentido común de este tipo de creencias; únicamente conduce a un conjunto de reglas para manipular los grados de creencia que son equivalentes a la suma y producto de reglas de probabilidad. Esto proporcionó la primera prueba de que la teoría de la probabilidad podría ser considerada como una extensión de la lógica booleana

para situaciones de incertidumbre [53]. El teorema de Bayes se utiliza para convertir una probabilidad a priori en una probabilidad posterior al incorporar la evidencia proporcionada por los datos observados. A manera de ejemplo se pueden captar suposiciones acerca de w, antes de observar los datos, en forma de distribución de probabilidad a priori $p(w)$. El efecto de los datos observados $D = \{t1,\ldots, tn\}$, se expresa a través de la probabilidad condicional $p(D|w)$, entonces el teorema de Bayes podrá representarse de la siguiente forma que muestra la evaluación de incertidumbre en w después de observada D en forma de probabilidad posterior $p(w|D)$:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} \quad (1)$$

Donde $p(w|D)$ es la probabilidad condicional de w dado D, $p(D|w)$ es la probabilidad condicional de D dado w, $p(w)$ y $p(D)$ son las probabilidades de w y D respectivamente. El valor $p(w|D)$ de la parte izquierda del teorema de Bayes es calculado para los datos observados del conjunto D y puede entenderse como una función del vector de parámetros w, en cuyo caso se denomina función de verosimilitud. Esta expresa qué tan probable es el conjunto de datos observados para diferentes configuraciones del vector de parámetros w. Se debe aclarar que la verosimilitud no es una distribución de probabilidad sobre w y su integral con respecto a w no es necesariamente es igual a uno. El denominador es la constante de normalización la cual garantiza que la distribución de probabilidad posterior $p(w|D)$ es una densidad de probabilidad válida y su integral sí es igual a uno.

2.2.3 Medidas de desempeño

Para determinar el desempeño de la clasificación, a menudo se utilizan las medidas de exactitud, sensibilidad, especificidad y precisión. Cada una de estas medidas es una función del porcentaje de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). La precisión se define como $(TP+TN)/(TP+TN+FP+FN)$; la sensibilidad como $TP/(TP+FN)$; la especificidad como $TN/(TN+FP)$; y la precisión como $TP/(TP+FP)$. Las anteriores medidas generalmente se calculan después de aplicar validación cruzada en k grupos, que consiste en dividir el conjunto de

entrenamiento en k subconjuntos de igual tamaño. Luego con cada subconjunto se prueba de manera iterativa empleando el clasificador entrenado sobre k-1 conjuntos [43].

2.2.4 Ganancia de información

Las medidas de cantidad de información han surgido como una herramienta muy útil para cuantificar las interacciones sinérgicas entre varios atributos genéticos. Estas medidas se basan en la entropía de Shannon, que cuantifica la cantidad de información, o la incertidumbre de una variable aleatoria [76]. Considerando los atributos genéticos y rasgos fenotípicos como variables aleatorias, se puede utilizar la medida de información teórica basada en entropía para cuantificar la información compartida entre un gen y un rasgo (efecto principal). Así como, la ganancia de información adicional sobre un rasgo se puede obtener de la combinación de múltiples genes, es decir, el efecto sinérgico o epistasia.

Esta ganancia se puede explicar como la cantidad de información necesaria promedio para describir una variable aleatoria. Para una variable discreta X con alfabeto X y $\mathrm{p(x)}$ su densidad de probabilidad, la entropía $\mathrm{H(X)}$ se define como:

$$\mathrm{H(X)} = -\sum_{x \in X} \mathrm{p(x)} \log \mathrm{p}\,(\mathrm{x}) \tag{2}$$

Donde x corresponde a las instancias y $\mathrm{H(X)}$ es la sumatoria de las entropías de cada instancia $(\mathrm{h(x)})$.

La entropía conjunta de dos variables discretas aleatorias X y Y con una distribución conjunta p se define como:

$$\mathrm{H}(X,Y) = -\sum_{x \in X} \sum_{\mathrm{y} \in \mathrm{Y}} \mathrm{p(x,y)}\ \mathrm{logp}\ (\mathrm{x},\mathrm{y}) \tag{3}$$

La entropía condicional de X dado el conocimiento Y puede obtenerse por la regla de la cadena como:

$$\mathrm{H(X|Y)} = \mathrm{H(X,Y)} - \mathrm{H(Y)} \tag{4}$$

La dependencia entre dos variables aleatorias puede describirse utilizando información mutua. Esta es una medida de la cantidad de información que contiene una variable aleatoria sobre la otra, o puede entenderse como la reducción de la incertidumbre de una variable aleatoria dado el conocimiento de otra. En el contexto de los estudios de asociación genética, la información mutua puede ser muy útil para calcular cuánto de un estatus fenotípico se explica por las variaciones genotípicas. Considerando un atributo genético $G1$ y la clase fenotípica C, por ejemplo, caso o control, dos variables aleatorias discretas, etc. La información mutua $I(G1;C)$ mide la reducción de la incertidumbre de la clase C debido al conocimiento sobre el genotipo de $G1$, definido como:

$$I(G_1; C) = H(C) - H(C|G_1) \tag{5}$$

Intuitivamente, $I(G1;C)$ puede ser usada como una medida del efecto principal del atributo genético $G1$ en la clase de información mutua C. También se puede extender para medir el efecto de interacción epistática entre dos atributos. Dados $G1$ y $G2$, la información mutua puede describirse como:

$$I(G_1, G_2; C) = H(C) - H(C|G_1, G_2) \tag{6}$$

Esto puede explicar el estado de la clase fenotípica dados $G1$ y $G2$ juntos. Restando los efectos principales de $G1$ y $G2$ de su efecto conjunto $I(G1,G2;C)$ se obtiene la ganancia de información mutua:

$$IG(G_1; G_2; C) = I(G_1, G_2; C) - I(G_1; C) - I(G_2; C) \tag{7}$$

La ganancia de información $IG(G1,G2,C)$ es la ganancia de información mutua de saber tanto $G1$ como $G2$ con respecto a la clase C. Un valor positivo de la $IG(G1,G2,C)$ indica sinergia entre $G1$ y $G2$, mientras que un valor negativo indica redundancia o correlación entre ellas.

No existe una definición formal ampliamente aceptada de ganancia de información que incluya más de dos atributos genéticos. Sin embargo, en [77] se propone una definición incluyendo interacciones de tres atributos utilizando medidas de ganancia de información basadas en entropía, de la siguiente forma:

$$\mathrm{IG}_{\mathrm{alternativa}}(\mathrm{G}_1;\mathrm{G}_2;\mathrm{G}_3;\mathrm{C}) = \mathrm{I}(\mathrm{G}_1,\mathrm{G}_2,\mathrm{G}_3;\mathrm{C}) - \mathrm{IG}(\mathrm{G}_1;\mathrm{G}_2;\mathrm{C}) - \mathrm{IG}(\mathrm{G}_1;\mathrm{G}_3;\mathrm{C}) - \mathrm{IG}(\mathrm{G}_2;\mathrm{G}_3;\mathrm{C}) - \mathrm{I}(\mathrm{G}_1;\mathrm{C}) - \mathrm{I}(\mathrm{G}_2;\mathrm{C}) - \mathrm{I}(\mathrm{G}_3;\mathrm{C}) \quad (8)$$

Donde $\mathrm{IG}_{\mathrm{alternativa}}$ es la ganancia de información de las variables $\mathrm{G1}$, $\mathrm{G2}$ y $\mathrm{G3}$ con respecto a la clase C.

Por otra parte, existen estudios tales como el GWAS donde se intentan establecer los efectos genéticos sobre ciertas enfermedades consideradas complejas mediante estudios de asociación y genotipificación; proyectos de bases de datos como AlzGene o KEGG (*Kyoto Encyclopedia of Genes and Genomes*) que ponen a disposición genomas de redes biológicas y herramientas de procesamiento; el proyecto HapMap con el cual se pretende descubrir los genes involucrados con la aparición de ciertas enfermedades y detectar los posibles fármacos para su tratamiento, esto basado en mapas de haplotipos (combinación de alelos que se encuentran en sitios cercanos dentro del mismo cromosoma).

2.2.5 Filtros para selección de atributos

El objetivo de los filtros es seleccionar los atributos o SNP de mayor interés del conjunto de los posibles candidatos. Esto puede lograrse usando cualquier método de filtrado como una prueba de χ^2 de independencia, ReliefF [54] o medidas basadas en la entropía de IG [55]. Estos filtros han proporcionado una manera de determinar el aumento de información sobre una clase (por ejemplo, el estado caso-control). Con estas medidas puede calcularse el beneficio de considerar dos o más atributos como una unidad.

Hay varias medidas para estimar la calidad de los atributos. Si el objetivo es una variable discreta, como es el caso del problema de clasificación, la calidad del atributo se puede calcular mediante aumento de la información [74], índice de Gini [75], la distancia medida [76], ReliefF [67], Tuning ReliefF [68] e incluso los estadísticos C2 y G.

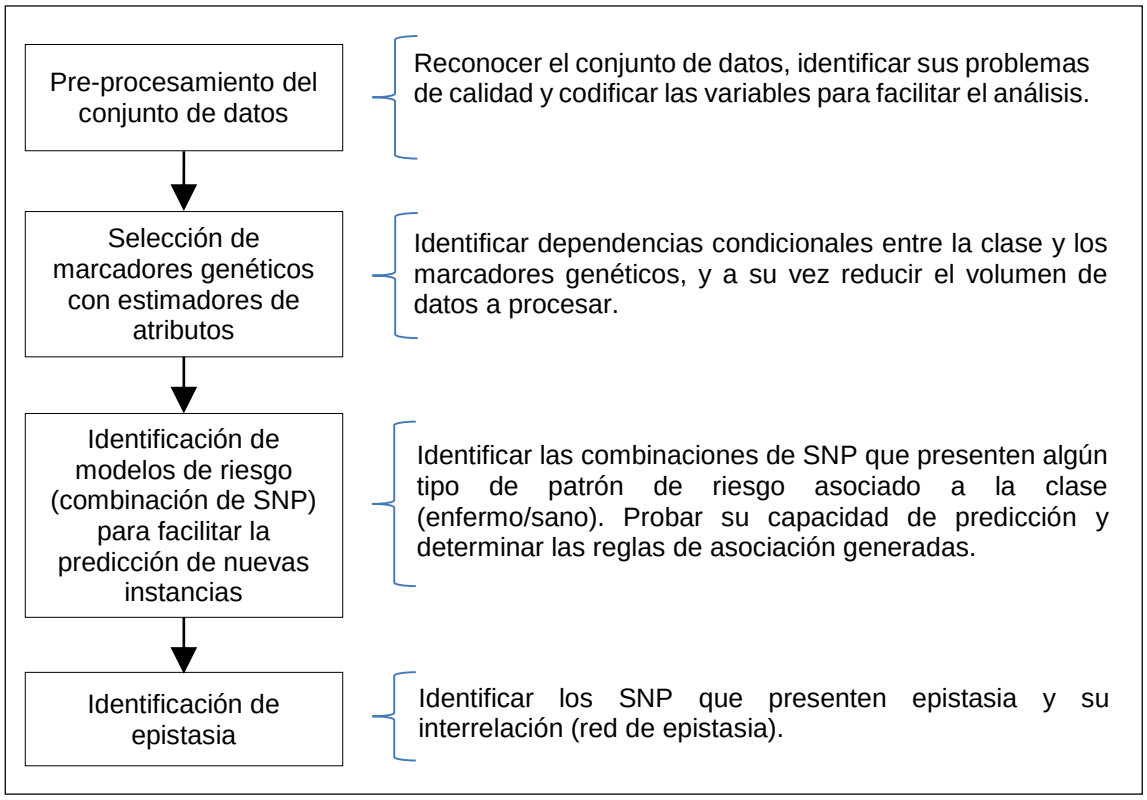
La mayoría de las medidas heurísticas para estimar la calidad de los atributos suponen la independencia condicional de los atributos y por lo tanto son menos apropiadas en problemas que implican gran interacción entre las características. Los algoritmos Relief,

por ejemplo, no hacen esta suposición. Éstos son eficientes y pueden estimar correctamente la calidad de los atributos en problemas con fuertes dependencias entre los mismos, son en la actualidad estimadores de características de uso general y han sido utilizados con éxito en una variedad de entornos, por ejemplo, para guiar la inducción constructiva en problemas de aprendizaje y como método de ponderación de atributos [77].

3. Metodología propuesta para el análisis de marcadores genéticos de la enfermedad de Alzheimer

En este capítulo se describe en detalle la metodología propuesta para el análisis de marcadores genéticos de la enfermedad de Alzheimer. Inicialmente se analiza el conjunto de datos con el fin de identificar la información faltante y codificar las posibles variaciones en los SNP para facilitar el procesamiento; luego se seleccionan los atributos o SNP más representativos, mediante algoritmos que cuantifiquen su aporte de información al proceso de clasificación. Posteriormente, se construyen nuevos atributos a partir de los seleccionados en el paso anterior, utilizando inducción constructiva MDR. Se realizaron análisis de epistasia estadística, obteniendo la relación de sinergia entre las variables mediante la ganancia de la información de los genes relacionados con la AD, proporcionando una metodología simple para caracterizar las interacciones en los estudios de asociación genética y determinar los rasgos más importantes que describen el comportamiento de la enfermedad. La metodología utilizada se describe a continuación (ver Figura 3-1).

Figura 3-1: Metodología propuesta.



- Eliminar las instancias que presentan datos faltantes y codificar los posibles genotipos de cada SNP (heterocigotos, homocigotos de mayor y menor frecuencia).
- Seleccionar los atributos que aportan mayor información al proceso de clasificación, mediante la sumatoria de los *rankings* obtenidos con los algoritmos ReliefF, TuReliefF y SURF.
- Identificar las combinaciones de atributos asociadas al riesgo de la EA: Se crean nuevos atributos mediante inducción constructiva, los cuales son combinaciones genotipos en dos o tres SNP que están asociados con un riesgo alto o bajo de padecer la enfermedad.
    - Para un orden de interacción M (modelo de M SNP), se seleccionaron M atributos o SNP del conjunto de datos.

- Se construyó una tabla de contingencia usando los M SNP y se calcularon las relaciones caso-control para cada combinación de genotipos (genotipos *multi-locus*).
- Siendo T la relación de casos/controles de todo el conjunto de datos, para cada genotipo *multi-locus*, si la relación caso/control excede T se consideró este genotipo como “Alto Riesgo”, en caso contrario, se consideró como “Bajo Riesgo”.
- El nuevo atributo formado como una combinación de M SNP está compuesto por los genotipos catalogados como alto o bajo riesgo.

- Utilizar un clasificador probabilístico para modelar la relación entre los atributos construidos y la clase caso-control. Esto se realiza para todas las posibles combinaciones de M SNP lo cual identifica el mejor modelo con relación a la media aritmética de la sensibilidad y especificidad, denominada precisión balanceada:

$$\text{Precisión balanceada} = (\text{sensibilidad} + \text{especificidad})/2 = 1/2(\text{TP}/(\text{TP}+\text{FN})+\text{TN}/(\text{TN}+\text{FP})) \quad (9)$$

  Dentro de este proceso se implementa la validación cruzada con 10 grupos.

- Identificar las variables epistáticas: Se explora la existencia de epistasia en el conjunto de datos, para lo cual se evalúan las relaciones de sinergia entre las variables por medio del concepto de ganancia de información.

3.1 Análisis descriptivo del conjunto de datos

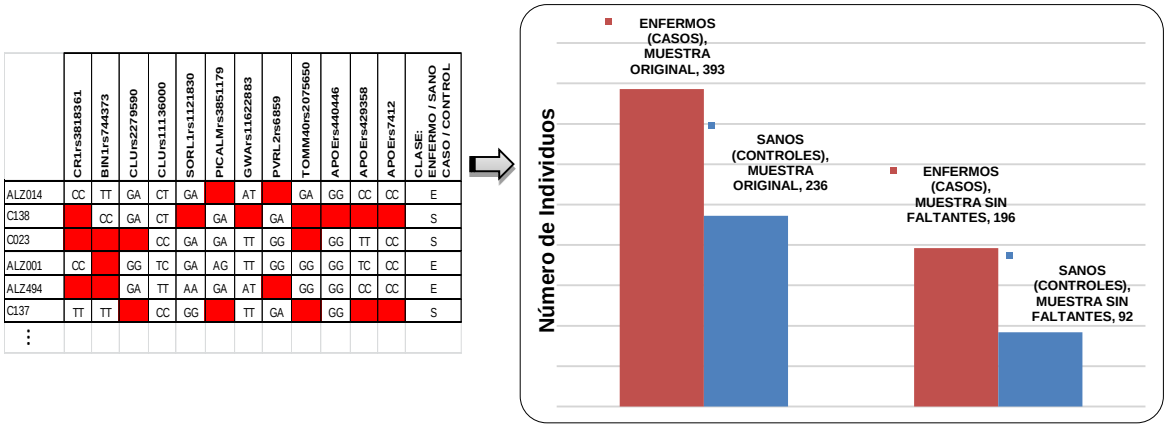
El conjunto de datos inicial está constituido por 393 casos y 236 controles, como se mencionó en el capítulo uno, provenientes de los resultados del proceso de secuenciación que identificó los genotipos correspondientes a cada SNP objeto de estudio, para un total de doce (12) atributos o marcadores genéticos y una clase enfermo/sano (caso/control) como se aprecia en la Tabla 3-1.

Tabla 3-1: Variables del proyecto.

Gen	SNP	Variable	Genotipos		
			0	1	2
APOE	rs7412	APO12	TT	CC	CT
APOE	rs440446	APO46	CC	GG	CG
APOE	rs429358	APO58	CC	TT	CT
BIN1	rs744373	BIN	CC	TT	CT
CLU	rs11136000	CLU00	TT	CC	CT
CLU	rs2279590	CLU90	AA	GG	AG
CR1	rs3818361	CR1	TT	CC	CT
GWA_14q32.13	rs11622883	GWA	AA	TT	AT
PICALM	rs3851179	PIC	AA	GG	AG
PVRL2	rs6859	PVR	AA	GG	AG
SORL1	rs1121830	SOR	GG	AA	AG
TOMM40	rs2075650	TOM	GG	AA	AG
CLASE: ENFERMO (E) o SANO (S)					

Sin embargo, en el proceso de secuenciación existe un grado de error atribuido a factores propios de los instrumentos de medición, equipos, reactivos o factores humanos, causante de la pérdida de información. Como consecuencia, la muestra presenta datos faltantes o no disponibles. Para encontrar la relación entre los marcadores y la enfermedad se necesita la información en todos los atributos de un individuo, por lo cual se descartaron los pacientes y controles con datos faltantes Se obtuvo entonces el conjunto de datos final, objeto de estudio, compuesto por el 50% de los casos (196 casos) y el 39% de los controles (92 controles) de la muestra original entregada por el Instituto de Genética como se presenta en la Figura 3-2.

Figura 3-2: Composición de la muestra original.



Para la codificación de los genotipos correspondientes a los marcadores, se tuvo como base el tipo de herencia co-dominante, la cual puede aplicarse a enfermedades donde influyen múltiples factores como los ambientales o simplemente las mutaciones en varios genes. Para esto se calcularon las frecuencias genotípicas de cada SNP en los casos y controles (E/S) mediante la relación del número de individuos que presentaban el genotipo entre el total de individuos, esto para cada marcador en cada clase (enfermo y sano). Posteriormente, se le asignó a cada genotipo la etiqueta 2 para los heterocigotos, 1 para los homocigotos de mayor frecuencia y 0 para los homocigotos de menor frecuencia, como se en la columna “genotipos” de la Tabla 3-1.

Se realizó la prueba de equilibrio de Hardy Weinberg, HW [14], que sostiene que la composición genética de una población permanece en equilibrio mientras no actúe la selección natural ni algún otro factor como las mutaciones, es decir, la herencia mendeliana por sí misma no genera cambio evolutivo.

Para establecer de manera preliminar las reglas de asociación que presenta el conjunto de datos, se implementó el algoritmo *a priori* el cual logra encontrar dichas reglas con suficiente soporte y confianza y está basado en dos pasos principales: (1) la búsqueda de conjuntos de elementos frecuentes, los que tienen suficiente soporte, y (2) convirtiéndolos en reglas con suficiente confianza mediante la partición de los ítems en dos, como ítems en el antecedente e ítems en el consecuente. El primer criterio de asociación del algoritmo a priori es la precisión o confianza, dada por el porcentaje de veces de las instancias que cumplen el antecedente y que también cumplen el consecuente, y el soporte que está dado por el número de instancias sobre las que es aplicable la regla; en otras palabras, la proporción de pacientes en el conjunto de datos que contiene dicho conjunto de características, es decir, los valores que toman las variables para un paciente.

3.2 Selección de atributos

La etapa de selección de atributos permite estimar la calidad de los mismos, y en caso de querer escoger un subconjunto de variables, tratando de mantener la mayor cantidad de información posible para describir al conjunto de datos. Para este fin se utilizaron filtros

que seleccionaban las variables más relevantes en términos de su aporte de información, basados en las variantes del algoritmo Relief, particularmente, ReleifF, TuRF (Tuned ReliefF) y SURF [54,67,68] que fueron originados a partir del algoritmo introducido en [67]. Relief trata de estimar de manera heurística la calidad de atributos de acuerdo a qué tan bien se distinguen sus valores entre las instancias que están cerca. Este algoritmo es capaz de identificar atributos funcionales en conjuntos de datos que incluyen interacciones con otros atributos que influyen en su dependencia con relación a la clase (fenotipo).

Para este fin dada de manera aleatoria una instancia R_i el algoritmo busca sus dos vecinos más cercanos, uno de la misma clase llamado el acierto más cercano H, y otro de la clase diferente llamado el fallo más cercano M. Luego actualiza la estimación de calidad $\mathrm{W[A]}$ para todos los atributos A dependiendo de sus valores de R_i, M y H. Si las instancias R_i y H tienen diferentes valores de los atributos A, entonces el atributo A separa dos instancias de la misma clase, lo cual no es deseable en tanto que reduce la estimación de calidad $\mathrm{W[A]}$. El algoritmo Relief se muestra a continuación en la Figura 3-3.

Figura 3-3: Algoritmo Relief

1. set all weights $\mathrm{W[A]}:=0.0$;
2. for $\mathrm{i} := 1$ to m do begin
3. randomly select an instance R_i;
4. find nearest hit H and nearest miss M;
5. for $\mathrm{A} := 1$ to a do
6. $\mathrm{W[A]}:= \mathrm{W[A]} - \mathrm{diff(A}, R_i, \mathrm{H)/m} + \mathrm{diff(A}, R_i, \mathrm{M)/m}$;
7. end;

La función $\mathrm{diff(A,I1,I2)}$ calcula la diferencia entre los valores del atributo A para dos instancias I_1 e I_2. Para atributos nominales este valor se calcula como se muestra en la Figura 3-4.

Figura 3-4: Cálculo de la función $\mathrm{diff(A,I1,I2)}$ para atributos nominales

$$\mathrm{diff}\,(\mathrm{A},I_1,I_2) = \begin{cases} 0;\ \mathrm{value}\,(\mathrm{A},I_1) = \mathrm{value}\,(\mathrm{A},I_2) \\ 1;\ \mathrm{otherwise} \end{cases}$$

Se define la distancia entre dos individuos como el número de SNP con genotipos diferentes. La distancia total se calcula simplemente como la suma de distancias de todos los atributos.

ReleifF [54] no está limitado solo a problemas de dos clases, es más robusto y puede trabajar con datos faltantes y ruidosos. De la misma forma que Relief selecciona una instancia R_i (línea 3 en la Figura 3-4), entonces busca k de sus vecinos más cercanos de la misma clase, llamados los aciertos más cercanos H_j (línea 4 en la Figura 3-6) y también k vecinos más cercanos de clases diferentes llamados fallos más cercanos $M_j(\mathrm{C})$ (líneas 5 y 6 en la Figura 3-5).

Figura 3-5: Algoritmo ReliefF

1. set all weights $\mathrm{W[A]:=0.0}$;
2. for $\mathrm{i := 1}$ to m do begin
3. randomly select an instance R_i;
4. find k nearest hits H_j;
5. for each class $\mathrm{C} \neq$ class $(R_{i,})$ do
6. from class C find k nearest misses $M_j(\mathrm{C})$;
7. for $\mathrm{A := 1}$ to a do
8. $\mathrm{W[A]:= W[A] - \sum_{j=1}^{k} diff\,(A,R_i,H_j) \,/\, (m * k) +}$
9. $\sum_{\mathrm{C}\neq\mathrm{class}(\mathrm{R_i})} [\frac{\mathrm{P(c)}}{1-\mathrm{P(class\,(R_i))}} \sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{diff\,(A,R_i,M_j(C))}] \,/\, (\mathrm{m} * \mathrm{k})$;
10. end;

La fórmula de actualización es similar a la de Relief, excepto que en esta se promedia la contribución de todos los aciertos y los fallos. Además la contribución para cada clase de fallos se multiplica por la probabilidad previa de cada clase $\mathrm{P(C)}$, donde C corresponde a la clase, estimada para el conjunto de entrenamiento, garantizando que la suma de

probabilidades de los fallos es igual a 1. Como la probabilidad de los aciertos se excluye de esta suma, se divide cada peso de probabilidad en un factor de $1\text{-P}(\text{Clase }(\text{Ri}))$.

Por otra parte, TuRF (Tuned ReliefF) o ReliefF sintonizado sistemáticamente elimina los atributos de baja calidad y calcula los valores de ReliefF, si los atributos restantes pueden ser re-estimados como se muestra en la Figura 3-6.

Figura 3-6: Algoritmo TuRF

1. let a be the number of attributes
2. for $\text{I} = 1$ to n do
3. estimate ReliefF
4. sort attributes
5. remove worst n/a attributes
6. end for
7. return last ReliefF estimate for each attribute.

Según el estudio en [68], TuRF es significativamente mejor que ReliefF y aunque los algoritmos de la familia Relief no son perfectos, son apropiados para estudios de análisis genético y son capaces de identificar interacciones no lineales entre genes en presencia de SNP con ruido.

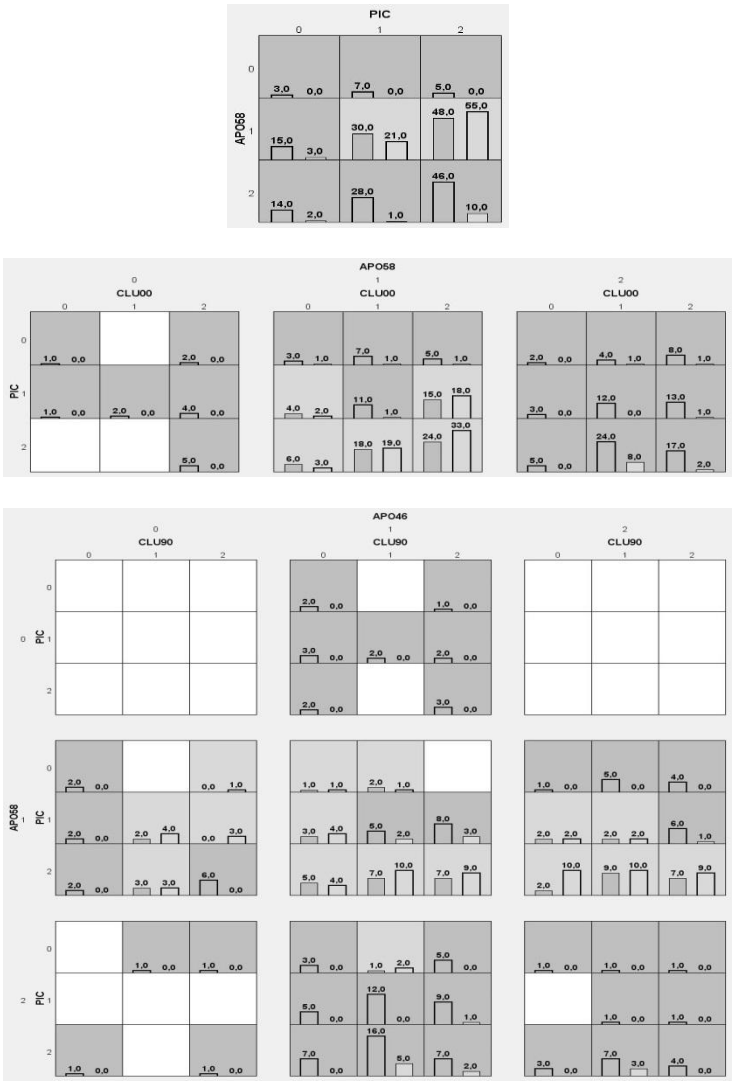
Finalmente, *Spatially Uniform ReliefF* (SURF), al igual que ReliefF, ajusta los pesos para todos los SNP mediante el uso de información de los vecinos. Mientras que ReliefF utiliza un número fijo de estos, SURF usa todos los vecinos dentro de una distancia fija que puede ser entendida como un umbral de similitud, sin aumentar el nivel de complejidad del algoritmo.

3.3 Inducción constructiva

En seguida se realizó un proceso basado en el principio de inducción constructiva seguido de la selección de una combinación óptima de genotipos *multi-locus* asociados con alto o bajo riesgo de padecer la enfermedad. Mediante tablas de contingencia se calcula la razón caso/control para cada genotipo *multi-locus* y en caso de superar el umbral 2,1304 (total

de casos entre total de controles de todo el conjunto de datos), el correspondiente genotipo *multi-locus* es considerado de alto riesgo, en caso contrario se cataloga como bajo riesgo. Cuando todos los atributos han sido etiquetados como alto o bajo riesgo, se generan nuevos atributos binarios como se muestra en la Figura 3-7.

Figura 3-7: Ejemplo de tablas de contingencias para genotipo *multi-locus*.



Las celdas resaltadas con gris oscuro corresponden a la etiqueta alto riesgo y las resaltadas con gris claro son las combinaciones de bajo riesgo. Se presentan como ejemplo los genotipos *multi-locus* de 2, 3 y 4 combinaciones: PIC-APO58, CLU00-PIC-

APO58 y CLU90-PIC-APO46-APO58. El conteo de los casos se ubica en la parte izquierda de la celda y los controles en la parte derecha. La combinación APO58=0-PIC=0, parte superior de la figura, se incluye en el grupo de alto riesgo ya que existen 3 casos frente a 0 controles y para la indeterminación 3/0 se considera como alto riesgo. En el caso opuesto, bajo riesgo, está la celda resaltada con gris claro APO58=1-pic=2 con 48 casos y 55 controles, la relación 48 sobre 55 es menor que el umbral 2,1304 y se considera como bajo riesgo. Esta tabla de contingencia se repite para todas las posibles combinaciones desde dos hasta tres genotipos *multi-locus*.

Posteriormente se utilizó un clasificador bayesiano con validación cruzada de 10 grupos para modelar la relación entre los atributos obtenidos con inducción constructiva y la clase enfermo o sano (caso o control), el cual evalúa mediante precisión balanceada todas las posibles interacciones. El mejor modelo de todos los posibles, es el que obtiene la mayor precisión y la mayor consistencia en la validación cruzada.

3.4 Relación de sinergia entre los atributos

Para obtener los modelos que presentan el efecto de interacción no lineal entre múltiples factores genéticos entre genes debe confirmarse la relación de sinergia entre los SNP que componen el modelo. Actualmente, sólo existen técnicas que pueden identificar relaciones con máximo tres SNP mediante entropía y ganancia de información. Para tal fin se utilizó ViSEN [74], una herramienta basada en teoría de información que puede visualizar la epistasia mediante los valores de efectos principales, interacción entre dos y tres SNP generando una gráfica sencilla que muestra estos efectos de orden superior. Adicional a esto puede cuantificar los efectos antes mencionados utilizando ganancia de información, con lo que puede priorizar las parejas o trío conformados de acuerdo con las interacciones.

Las redes permiten una representación estructurada de una colección de variables y sus relaciones, que proporciona un marco adecuado para el estudio de epistasia. El gráfico que se obtiene con esta técnica, consiste en un conjunto de vértices que corresponden a SNP y bordes que muestran la interacción entre los SNP representados por cada vértice. El tamaño de un vértice indica su efecto principal, y el ancho de borde indica la intensidad

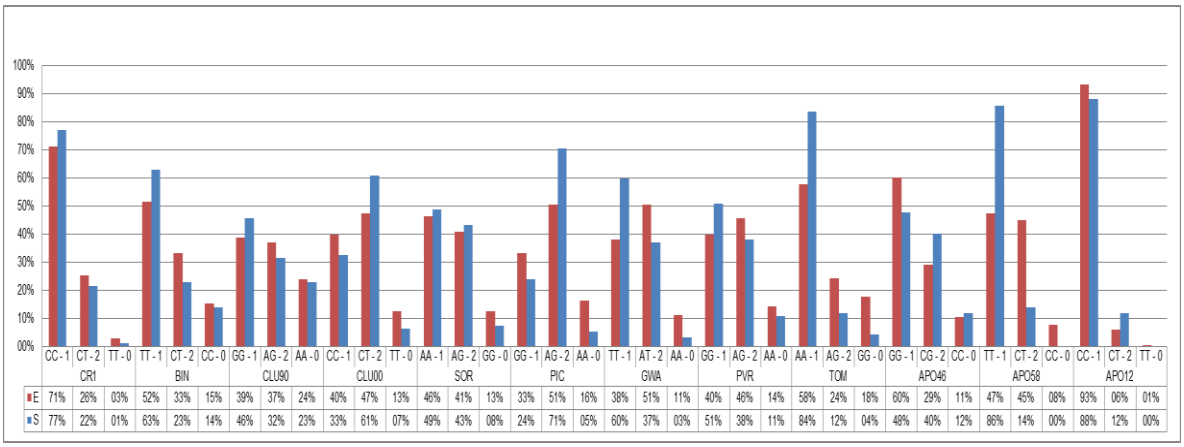
de la sinergia entre los pares o tríos. La red está diseñada para agregar gradualmente las interacciones de pares, clasificados por su fuerza (la ganancia de información entre pares), y los vértices son SNP. Cada círculo es un SNP con su nombre y su fuerza de efecto principal. Cada borde representa una epistasia en pares con sus respectivas fuerzas o en un triángulo en el caso de los tríos.

Estas relaciones están basadas en teoría de la información y entropía que pueden explicarse como la cantidad de información requerida para describir una variable aleatoria y la dependencia entre dos variables aleatorias puede representarse utilizando información mutua, cuyo significado consiste en la cantidad de información que contiene una variable sobre otra, o puede ser entendido como una reducción de la incertidumbre de una variable aleatoria dado el conocimiento de otra. En el contexto de los estudios de asociación genética, la información mutua puede ser muy útil para calcular cuánto de un estatus fenotípico se explica por las variaciones genotípicas [76]. Un valor positivo de la ganancia de información indica sinergia, un valor negativo indica redundancia o correlación. La sinergia puede definirse como las interacciones epistáticas de dos o tres atributos genéticos, según la ecuación (7), donde IG es la ganancia de la información, I es la información mutua, $G1$, $G2$ y $G3$ son los atributos genéticos y C es la clase fenotípica.

4. Resultados y discusión

La primera etapa planteada en la metodología consistía en el pre-procesamiento del conjunto de datos, en donde se excluyeron las instancias que tenían datos faltantes. Se codificaron los genotipos con el número 2 para los heterocigotos, 1 para los homocigotos de mayor frecuencia y 0 para los homocigotos de menor frecuencia, como se mencionó en el capítulo anterior. Mediante conteo se calcularon las frecuencias genotípicas de la muestra final compuesta por 12 atributos, la clase enfermo o sano (E o S), 196 casos y 92 controles (ver Figura 4-1).

Figura 4-1: Frecuencias genotípicas para la muestra sin faltantes



Se puede observar que existen diferencias marcadas en las frecuencias de la clase para los SNP CLU00 y PIC en el heterocigoto, GWA y TOM en el homocigoto de mayor frecuencia, y BIN y APO58 en el homocigoto de mayor frecuencia y en el heterocigoto; este último el que presenta la mayor diferencia, ya que en estudios anteriores se ha determinado que este gen tiene relación directa con la EA, para el cual las instancias que poseen el genotipo TT tienen menor riesgo de padecer la enfermedad que las tienen el genotipo CT. Lo anterior confirma lo determinado en [112] donde se identificó al alelo C como indicador de riesgo o alelo asociado a la EA.

Al verificar el equilibrio de Hardy Weinberg, se obtuvieron los resultados mostrados en la Tabla 3-2 para cada SNP, en donde un valor de X^2 *test P value* < 0.05 significa que el marcador no se encuentra en equilibrio (ver Tabla 4-1).

Tabla 4-1: Equilibrio de Hardy Weinberg calculado sobre las variables de entrada.

MARCADOR	GENOTIPO	No. de Individuos	X^2 test P value
CR1	CC	211	0,678935505
	CT	70	
	TT	7	
BIN	TT	159	1,10063E-06
	CT	86	
	CC	43	
CLU90	GG	118	4,73565E-06
	AG	102	
	AA	68	
CLU00	CC	108	0,052246353
	CT	149	
	TT	31	
SOR	AA	136	0,479110332
	AG	120	
	GG	32	
PIC	GG	87	0,003099677
	AG	164	
	AA	37	
GWA	TT	130	0,268530644
	AT	133	
	AA	25	
PVR	GG	125	0,447328742
	AG	125	
	AA	38	
TOM	AA	190	1,56763E-13
	AG	59	
	GG	39	
APO46	GG	162	0,002229644
	CG	94	
	CC	32	
APO58	TT	172	0,972364642
	CT	101	
	CC	15	
APO12	CC	264	0,516173616
	CT	23	
	TT	1	

En los SNP CLU90, TOM, PIC, APO46 y BIN no se encuentran en equilibrio de Hardy Weinberg. Esto podría deberse a la aleatoriedad de la muestra o a diferencias étnicas y

demográficas de cada población. Para garantizar que sea de las mismas características debería realizarse con un conjunto pareado de casos y controles, lo cual aumentaría la certeza en el resultado.

Como parte del proceso inicial, se llevó a cabo la generación de reglas de asociación, la cual se desarrolló en dos pasos principales. Se encontró el conjunto de ítems más frecuentes en el conjunto de datos, en este caso los ítems son los valores que pueden tomar las variables para los datos de la muestra, es decir los genotipos. Seguidamente se formaron las reglas partiendo de estos conjuntos frecuentes de ítems considerando un valor de confianza mínima. Las medidas de evaluación para la generación de las reglas se basaron en un nivel de confianza superior a 0.9. Las reglas encontradas se muestran en la Tabla 4-2.

Tabla 4-2: Reglas de asociación encontradas mediante el algoritmo a priori.

REGLA	CONFIANZA
SI BIN = 2 Y APO58 = 2 ENTONCES CLASE = E	0,94
SI TOM = 0 Y APO46 = 1 ENTONCES CLASE = E	0,94
SI TOM = 0 Y APO46 = 1 Y APO12 = 1 ENTONCES CLASE = E	0,94
SI TOM = 2 Y APO58 = 1 ENTONCES CLASE = E	0,93
SI TOM = 2 Y APO58 = 2 Y APO12 = 1 ENTONCES CLASE = E	0,93
SI CR1 = 1 Y CLU00 = 2 Y APO58 = 2 ENTONCES CLASE = E	0,91
SI CLU90 = 2 Y APO58 = 2 ENTONCES CLASE = E	0,91

Se encontraron asociaciones con los SNP BIN, APO58, TOM, APO46, CR1, CLU00 y CLU 90 con valores de confianza entre 0.91 y 0.94. En estas se evidenció la presencia del gen APOE con sus SNP APO58, APO56 y APO12 asociados a la clase enfermo en todas las reglas encontradas. Lo anterior confirma lo establecido en [113] sobre la relación entre dichos alelos y la enfermedad.

Por otro lado, se implementó un clasificador bayesiano para comprobar la capacidad de clasificación el conjunto de datos. Se utilizó validación cruzada con 10 grupos y el resultado obtenido fue: se clasificaron correctamente 199 instancias, es decir el 69% de los datos, y

89 instancias fueron clasificadas de manera incorrecta. Este clasificador bayesiano se implementó en el software Weka [94].

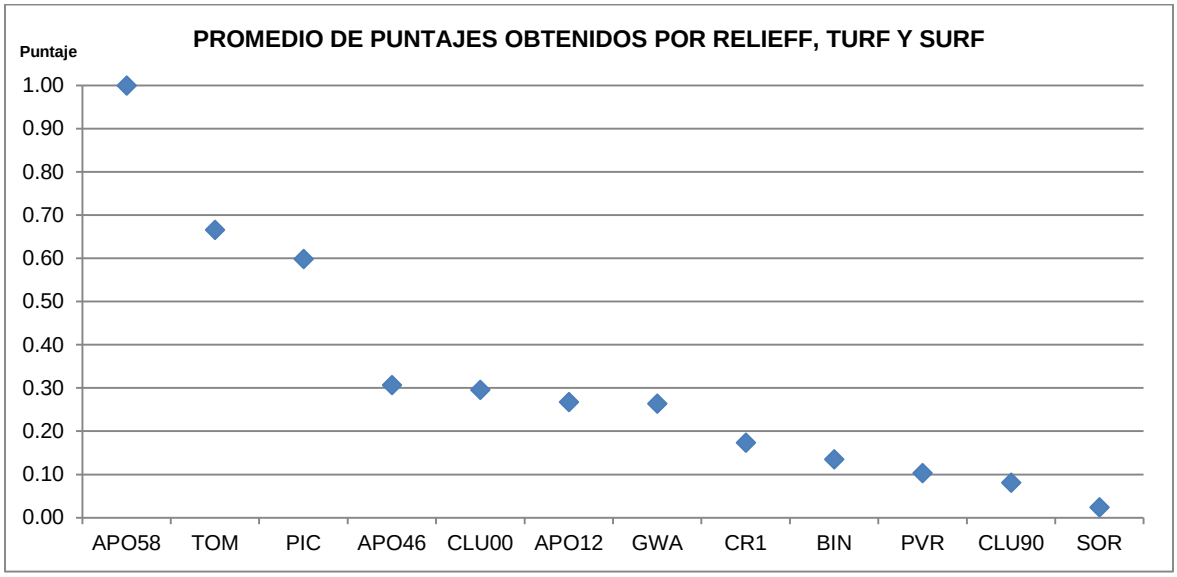
4.1 Atributos seleccionados

En el proceso de selección de atributos se obtuvieron los rankings para las variables aplicando los algoritmos ReliefF, TuRF y SURF (ver sección 3.1.1). Para el cálculo de estos valores, para cada atributo se normalizó el puntaje arrojado por cada algoritmo y se calculó el promedio. Los valores obtenidos se presentan en la columna “PROMEDIO” de la Tabla 4-3 y se muestran gráficamente en la Figura 4-2. Se usaron 10 vecinos cercanos para ReliefF y TuRF, mientras que SURF utilizó todos los vecinos por debajo de un umbral establecido.

Tabla 4-3: Posicionamiento obtenido mediante los algoritmos ReliefF, TuRF y SURF

Atributo	ReliefF	TuRF	SURF	PROMEDIO
APO58	1.000000	1.000000	1.000000	1.000000
TOM	0.738494	0.762461	0.497285	0.666080
PIC	0.594139	0.869738	0.332952	0.598943
APO46	0.282425	0.505754	0.132324	0.306834
CLU00	0.154814	0.574720	0.159188	0.296241
APO12	0.209209	0.241379	0.352672	0.267753
GWA	0.355648	0.072806	0.364390	0.264281
CR1	0.179915	0.011498	0.329523	0.173645
BIN	0.165274	0.095790	0.144613	0.135226
PVR	0.108789	0.122604	0.078594	0.103329
CLU90	0.085773	0.157087	0.000000	0.080953
SOR	0.000000	0.000000	0.072592	0.024197

Figura 4-2: Valor promedio de los puntajes obtenidos con los algoritmos RelifF, TuRF y SURF



Después de aplicar los filtros se utilizó una heurística para seleccionar el número de variables, que consistió en ordenar los valores del puntaje para las variables de cada filtro y calcular la diferencia entre cada par, desde el mayor al menor hasta llegar a un puntaje menor o igual a un umbral predeterminado. En este caso se consideró un umbral de 0,02. Con base en lo anterior, se seleccionaron entonces los atributos APO58, TOM, PIC, APO46, CLU00 y APO12.

En la gráfica se observa que los SNP del gen APOE que son APO58, APO46 y APO12, obtuvieron puntajes de 3, 0,92 y 0,8 respectivamente, lo que les permitió ser incluidos en la selección. Esto confirma la relevante influencia del gen APOE en el riesgo de padecer EA.

4.2 Combinaciones asociadas al riesgo mediante inducción constructiva

Con los atributos antes mencionados se procedió a la fase de identificación de patrones de las variables por medio de inducción constructiva para identificar cuál combinación de SNP estaba asociada al riesgo de padecer la enfermedad. Se buscaron combinaciones de dos y tres SNP teniendo en cuenta la cantidad de atributos y el número de instancias comparadas con los trabajos de [111] en los cuales la muestra estaba compuesta por 321 casos (pacientes con tuberculosis), 347 controles y un total de 19 SNP. En dicho trabajo los autores recomendaron explorar máximo tres SNP por combinación. Las combinaciones de dos y tres SNP fueron obtenidas utilizando el software MDR [20] y se muestran en la Tabla 4-4 y 4-5 respectivamente.

Tabla 4-4: Combinaciones de dos SNP obtenidas mediante inducción constructiva para las seis variables obtenidas con los filtros.

Modelos de dos SNP	Precisión balanceada	Posicionamiento en la validación cruzada
PIC-APO58	0,714064	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
APO58-APO12	0,694654	[2, 2, 2, 5, 2, 2, 3, 3, 2, 2]
CLU00-APO58	0,677684	[3, 3, 3, 2, 3, 3, 2, 2, 3, 3]
TOM-APO58	0,692103	[4, 4, 4, 3, 4, 4, 4, 4, 4, 4]
APO46-APO58	0,692103	[5, 5, 5, 4, 5, 5, 5, 5, 5, 5]
PIC-TOM	0,636868	[6, 7, 6, 6, 6, 6, 6, 6, 6, 6]
TOM-APO12	0,651952	[9, 8, 7, 8, 7, 8, 7, 8, 7, 9]
PIC-APO46	0,619898	[7, 6, 10, 10, 10, 7, 8, 7, 10, 7]
CLU00-TOM	0,605923	[8, 11, 8, 7, 8, 10, 9, 10, 8, 10]
CLU00-PIC	0,634095	[11, 9, 11, 9, 11, 9, 10, 9, 11, 8]
TOM-APO46	0,62256	[10, 10, 9, 11, 9, 11, 11, 11, 9, 11]
PIC-APO12	0,595941	[12, 12, 12, 12, 12, 12, 12, 12, 12, 12]
CLU00-APO46	0,532941	[13, 15, 13, 13, 13, 14, 13, 13, 13, 13]
CLU00-APO12	0,578305	[14, 14, 15, 14, 14, 13, 15, 15, 15, 15]
APO46-APO12	0,534383	[15, 13, 14, 15, 15, 15, 14, 14, 14, 14]

Tabla 4-5: Combinaciones de tres SNP obtenidas mediante inducción constructiva para las seis variables obtenidas con los filtros.

Modelos de tres SNP	Precisión balanceada	Posicionamiento en la validación cruzada
CLU00-PIC-APO58	0,704747	[1, 2, 2, 1, 1, 1, 2, 1, 3, 1]
PIC-TOM-APO58	0,708962	[2, 1, 1, 2, 2, 2, 1, 3, 1, 2]
PIC-APO46-APO58	0,690994	[3, 3, 4, 4, 3, 3, 3, 2, 2, 3]
PIC-APO58-APO12	0,716615	[4, 5, 5, 5, 4, 4, 4, 4, 4, 5]
CLU00-PIC-TOM	0,680568	[5, 6, 7, 3, 5, 9, 6, 5, 11, 4]
CLU00-TOM-APO58	0,673913	[6, 7, 3, 6, 6, 5, 5, 7, 8, 6]
CLU00-APO46-APO58	0,640639	[7, 9, 6, 8, 7, 7, 7, 6, 6, 7]
TOM-APO46-APO58	0,687001	[9, 8, 9, 7, 9, 6, 8, 9, 5, 9]
CLU00-APO58-APO12	0,666482	[8, 10, 10, 9, 8, 8, 9, 10, 7, 10]
TOM-APO58-APO12	0,686335	[10, 13, 11, 10, 11, 10, 10, 11, 9, 13]
APO46-APO58-APO12	0,683784	[12, 11, 12, 12, 12, 11, 11, 12, 10, 14]
PIC-TOM-APO46	0,651176	[11, 4, 13, 11, 13, 12, 12, 8, 12, 8]
PIC-TOM-APO12	0,670364	[13, 12, 8, 13, 10, 13, 13, 13, 13, 12]
CLU00-PIC-APO46	0,597493	[14, 14, 15, 14, 15, 14, 14, 14, 16, 11]
CLU00-TOM-APO12	0,645519	[16, 17, 14, 15, 14, 16, 16, 16, 14, 17]
TOM-APO46-APO12	0,621229	[17, 16, 16, 16, 16, 17, 17, 17, 15, 18]
PIC-APO46-APO12	0,598713	[15, 15, 18, 18, 18, 15, 15, 15, 18, 15]
CLU00-TOM-APO46	0,575311	[18, 19, 17, 17, 17, 19, 18, 18, 17, 19]
CLU00-PIC-APO12	0,608252	[19, 18, 19, 19, 19, 18, 19, 19, 19, 16]
CLU00-APO46-APO12	0,531943	[20, 20, 20, 20, 20, 20, 20, 20, 20, 20]

Con el fin de seleccionar los atributos, durante la validación cruzada, las diferentes combinaciones se ordenan en un listado de mayor a menor según su precisión balanceada. Específicamente, en cada iteración de la validación cruzada se determina la importancia de cada atributo a través de su posición en dicha clasificación. De acuerdo con lo anterior, la columna “posicionamiento en la validación cruzada” en la Tabla 4-4 y 4-5 indica el lugar que ocuparon las combinaciones en las iteraciones de la validación cruzada (ver sección 3 y [114]). Por consiguiente, el objetivo es identificar los atributos que aparecen en las primeras posiciones un gran número de veces. Según lo anterior, fueron escogidas las combinaciones PIC-APO58 y CLU-PIC-APO58. Esta relación entre PIC-APO58 y CLU-PIC-APO58 como factores de riesgo de la EA está acorde con lo reportado en el estudio en [94].

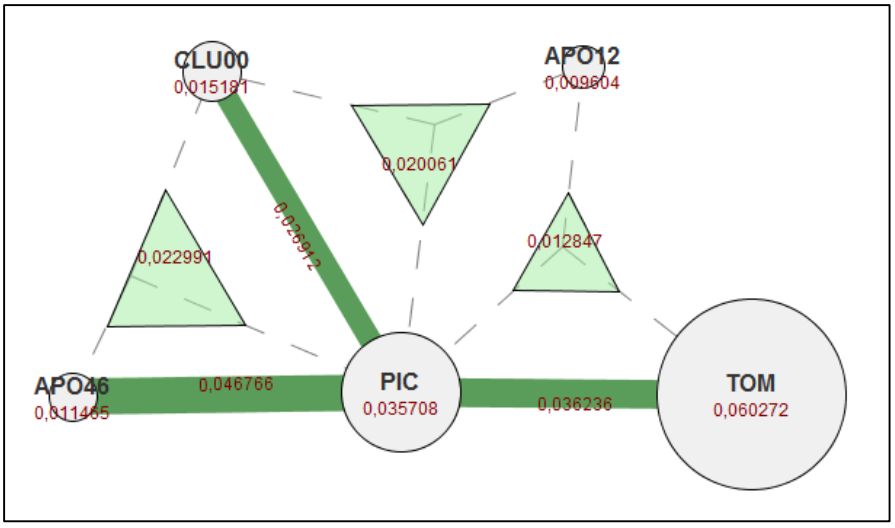
Con el objetivo de identificar la posible pérdida de información, por el hecho de trabajar con seis de las doce variables en los cálculos anteriores, se analizaron todas las

combinaciones para el conjunto de datos con las doce variables (ver Anexo 1). Usando los mismos parámetros, en este caso, los resultados obtenidos son los mismos que cuando se usaron seis variables.

4.3 Variables epistáticas mediante ganancia de información

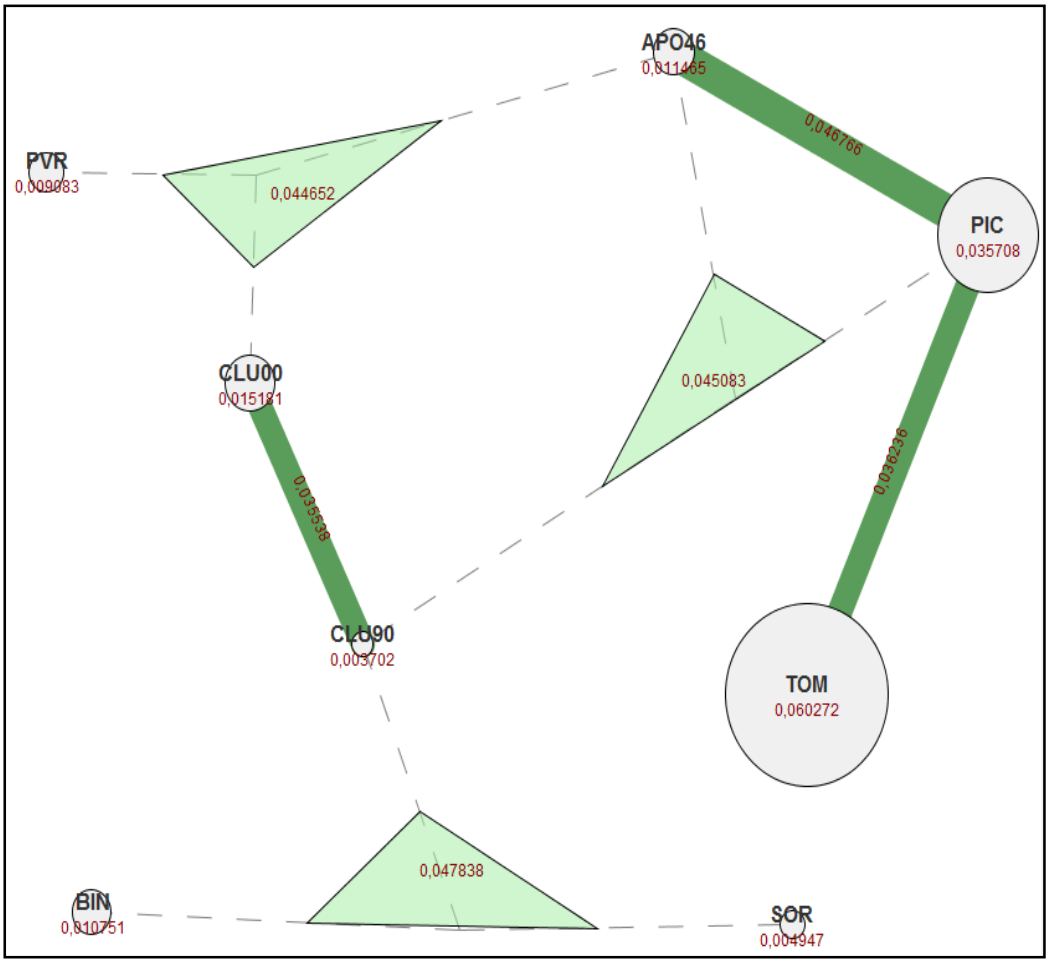
El paso final consiste en verificar si las combinaciones seleccionadas por inducción constructiva corresponden a SNP epistáticos mediante los conceptos de entropía y ganancia de información. Esto se hizo considerando solo seis variables y también con todas las variables utilizando el software ViSEN [74]. Se cuantificaron los efectos principales de cada SNP y las sinergias, entendidas como epistasia, de los modelos de dos y tres atributos para todas las combinaciones posibles. Se seleccionaron los pares y tríos de atributos cuya fuerza de epistasia era mayor a un umbral de 0,01 para tres SNP y 0,025 para dos SNP. Con esto se pudieron obtener los mapas de las interacciones fuertes por parejas y tríos que muestran la estructura de vecindad de cada atributo para el conjunto considerando seis variables (ver figura 4-3) y todas las variables (ver figura 4-4).

Figura 4-3: Mapa de interacciones para el conjunto de seis variables obtenidas por los filtros ReliefF, TuRF y SURF



Durante la selección de combinaciones epistáticas en la muestra con todas las variables, se modificaron los umbrales de selección a 0,044 para tres SNP y 0,35 para 2 SNP, esto en razón a se incrementaron las combinaciones que sobrepasaban el umbral de 0,01 y para efectos de la representación en el mapa se seleccionaron las superiores a éste nuevo umbral.

Figura 4-4: Mapa de interacciones para todas las variables



Con la identificación de las sinergias de ambos conjuntos de datos pueden observarse las combinaciones conformadas por tres SNP, que para seis variables son CLU00-PIC-APO46, CLU00-PIC-APO12 y TOM-PIC-APO12; y para todas las variables son CLU90-BIN-SOR, CLU90-PIC-APO46 y PVR-CLU00-APO46. Las conformadas por dos SNP son

PIC-APO46, TOM-PIC y CLU00-PIC para seis variables; y PIC-APO46, TOM-PIC y CLU00-CLU90.

En relación a las combinaciones PIC-APO58 y CLU00-PIC-APO58 encontradas mediante inducción constructiva y las encontradas en el mapa de epistasia se puede destacar que los SNP PIC y CLU00 están asociados a la EA con un valor de sinergia o epistasia de 0,027. El SNP APO58 a pesar de ser conocido por su relación con la EA, no figura en la detección de epistasia y esto probablemente se debe a que existe redundancia en información de éste con respecto a las demás variables y a la clase, por esto la metodología (ganancia de información) no se refleja sinergia. Cabe resaltar, el método basado en ganancia de información no muestra la sinergia. Lo cual no significa que este SNP no pueda actuar de manera epistática con otros. Esto debería determinarse en un análisis más profundo.

El SNP TOM aparece con el mayor valor de los efecto principales (0,60272). Esto está acorde con el reciente descubrimiento de su importancia en la aparición de EA, ya que según resultados obtenidos éste afecta las áreas del cerebro vulnerables a EA, por procesos de muerte celular programada (apoptosis) descendente que renuncian a la agregación del amiloide beta extracelular [95,97,98,99].

En cuanto a la relación de TOM y APOE en la combinación TOM-PIC-APO12, esta podría estar relacionada con la zona de desequilibrio de ligamiento que contiene los genes TOMM40, APOE y APOC1. En un estudio filogenético de los polimorfismos de esta zona de desequilibrio [100], con una muestra de pacientes con EA de origen tardío y controles de edad comparable, se determinó que TOMM40 está asociado a la edad de inicio de la EA de origen tardío. Adicional a esto, los genes TOMM40 y APOE comparten la función de codificar una proteína mitocondrial [101], sugiriendo que la susceptibilidad con la EA asociada a esta región es debida a la contribución de los dos genes al mecanismo de la enfermedad [102]. Por último, un reciente estudio de casos y controles centrado en varios SNP ha encontrado una asociación entre TOMM40 y EA comparando individuos con y sin la enfermedad. En dicho estudio se demostró una relación altamente significativa entre el SNP rs2075650 y EA, que se tornaba más fuerte en un haplotipo formado por TOMM40

rs2075650, rs11556505 y APOE rs429358 [103], lo cual soporta los hallazgos de un efecto sinérgico entre TOMM40 y APOE [104].

Con respecto a las combinaciones conformadas por CLU, PIC y APOE (CLU00-PIC-APO46, CLU00-PIC-APO12, CLU90-PIC-APO46, PIC-APO46 y CLU00-PIC) podría describirse la relación en comparación con los siguientes estudios que se comenta a continuación.

Para el gen CLU, existen trabajos divulgados que sugieren que este gen y APOE pueden tener efectos aditivos en la deposición del amiloide beta [105]. Asimismo, este gen en conjunto con PICALM participan en otros procesos que no están relacionados con el amiloide beta ni con el procesamiento de TAU, por lo que el papel de estas proteínas en el cerebro pueden revelar evidencia de mecanismos adicionales de la enfermedad que van más allá de su acumulación. De hecho, existen varios estudios que vinculan estos genes con el metabolismo lipídico y las vías inflamatorias [106]. Finalmente, también a CLU se le ha identificado susceptibilidad a EA en relación a las funciones con el sistema inmune, lo que sugiere un posible papel para el sistema inmune en el riesgo para EA [107].

Por otra parte, en el análisis de epistasia para todas las variables se evidencia la aparición de los SNP BIN, SOR y PVR, los cuales no fueron seleccionados por los filtros pero que al momento de analizar su intervención con otros SNP presentan sinergia en los modelos CLU90-BIN-SOR y PVR-CLU00-APO46. Para el caso de BIN esto puede comprobarse mediante los estudios de asociación realizados por [108], los cuales indican que BIN1 puede afectar el riesgo a EA mediante la alteración de las membranas neuronales y la intervención en los procesos de formación de vasos sinápticos. Además, este gen ocasiona cambios en el resultado de la función del gen PICALM de perturbación en la sinapsis, lo que posiblemente altera los ciclos de la vesícula sináptica alterando el riesgo para la EA [109].

La intervención de SOR en el riesgo de la EA puede explicarse mediante la investigación realizada por [110], en donde el análisis de 50 marcadores de SORL1 ubicados al final del gen confirmó que un haplotipo de los SNP rs668387, rs689021 y rs641120 está asociado

con niveles reducidos de fluido cerebro-espinal (CSF) y deterioro cognitivo leve. Por lo tanto, estos tres SNP han sido confirmados como los más importantes marcadores de riesgo AD dentro del gen SORL1 en muestras caucásicas entre 11.592 casos y 17.048 controles.

5. Conclusiones y recomendaciones

En el presente trabajo, se implementó una metodología capaz de analizar y representar las interacciones de marcadores genéticos mediante el uso de algoritmos de selección de atributos, inducción constructiva, entropía y ganancia de información, en una muestra de casos y controles para la enfermedad de Alzheimer de la población colombiana. La metodología permitió identificar las interacciones más fuertes entre combinaciones de dos y tres SNP que se asocian al riesgo de padecer EA, obteniendo un mapa de las relaciones entre genes que aporta información importante sobre esta enfermedad.

Mediante los resultados obtenidos en este estudio, se pudieron confirmar relaciones reportadas en trabajos anteriores con muestras de diferentes poblaciones, utilizando técnicas computacionales.

Los resultados de la investigación determinaron, para el conjunto de datos, patrones relevantes de epistasia en los genes SORL1, BIN y CLU90; PIC, APO46 y CLU90. Esto coincide con lo reportado en otros estudios sobre la EA. A través de inducción constructiva, se identificó una fuerte relación entre los SNP PIC y APO58 con el riesgo a padecer la enfermedad. Por otro lado, en las reglas de asociación obtenidas a través del algoritmo a priori es importante resaltar la participación de APO58, lo cual confirma su importancia, tal como se ha indicado en numerosas publicaciones sobre estudios de asociación en EA.

Debido al complejo funcionamiento de los genes, la epistasia es fundamental para determinar el riesgo de una enfermedad en un individuo, teniendo en cuenta que desde el punto de vista estadístico es más frecuente encontrar estudios de asociación basados en efectos principales, inclusive los métodos con enfoque puramente estadístico no logran detectar interacciones combinadas. Lo anterior, sumado con el conocimiento e inclusión de otro tipo de variables como las clínicas y ambientales aumenta la posibilidad de comprender el comportamiento genético de las enfermedades humanas. Por esta razón,

en este trabajo se dio especial importancia al estudio de la epistasia con una metodología que puede aplicarse cuando se disponga de datos de varios SNP y que ha sido utilizada en diferentes enfermedades.

Se logró identificar que la aplicación de los filtros para selección de atributos fue un paso importante para el proceso de inducción constructiva ya que no afectó la obtención de combinaciones asociadas al riesgo por el hecho de reducir el número de variables a seis. Sin embargo, tuvo repercusión en el cálculo de las combinaciones epistáticas porque causó la omisión de relaciones importantes que tenían un valor de sinergia más alto, conformadas por variables que no habían sido seleccionadas por los filtros

Es importante realizar estudios con otras muestras de pacientes que incluyan un número mayor de datos con el objetivo de confirmar los resultados encontrados y tratar de descubrir nuevas relaciones que ayuden a tener más información sobre la EA.

Para estudios futuros se sugiere ampliar el número de variables que puedan brindar mayor información sobre la EA, como lo son las variables clínicas y ambientales, para aumentar el número de relaciones y entender mejor el comportamiento de la enfermedad.

Se sugiere un análisis más profundo de diferentes algoritmos de selección de atributos con el fin de lograr mayor precisión en la selección de combinaciones epistáticas.

A. Anexo: Combinaciones de dos y tres SNP obtenidas mediante inducción constructiva para todas las variables

Modelos de dos SNP	Precisión balanceada	*Ranking* en la validación cruzada
PIC,APO58	0,714064	[1, 1, 2, 1, 1, 1, 2, 1, 1, 1]
GWA,APO58	0,706411	[2, 2, 1, 2, 2, 2, 1, 2, 2, 2]
APO58,APO12	0,694654	[3, 3, 4, 11, 3, 3, 5, 5, 3, 3]
CR1,APO58	0,670031	[4, 4, 3, 4, 4, 4, 6, 3, 4, 4]
CLU00,APO58	0,677684	[7, 5, 8, 3, 7, 5, 3, 4, 7, 5]
BIN,APO58	0,692103	[5, 6, 6, 5, 5, 6, 7, 6, 5, 6]
CLU90,APO58	0,692103	[6, 7, 7, 6, 6, 7, 8, 7, 6, 7]
SOR,APO58	0,67835	[8, 8, 5, 7, 8, 8, 4, 8, 8, 8]
PVR,APO58	0,692103	[9, 9, 9, 8, 9, 9, 9, 9, 9, 9]
TOM,APO58	0,692103	[10, 10, 10, 9, 10, 10, 10, 10, 10, 10]
APO46,APO58	0,692103	[11, 11, 11, 10, 11, 11, 11, 11, 11, 11]
PIC,TOM	0,636868	[12, 14, 13, 13, 12, 12, 13, 12, 12, 12]
GWA,TOM	0,596717	[13, 12, 12, 12, 14, 16, 12, 13, 14, 17]
TOM,APO12	0,651952	[17, 15, 14, 15, 13, 14, 14, 15, 13, 16]
PIC,APO46	0,619898	[14, 13, 22, 18, 22, 13, 15, 14, 22, 13]
SOR,TOM	0,633319	[15, 20, 15, 19, 15, 19, 16, 17, 15, 22]
CLU00,TOM	0,605923	[16, 27, 16, 14, 16, 18, 19, 20, 17, 19]
PIC,GWA	0,625998	[18, 16, 23, 16, 24, 15, 17, 16, 24, 15]
CR1,TOM	0,624778	[20, 23, 17, 20, 17, 21, 24, 18, 18, 23]
CLU00,PIC	0,634095	[24, 21, 25, 17, 23, 17, 21, 19, 25, 14]
TOM,APO46	0,62256	[19, 24, 21, 24, 21, 22, 25, 23, 19, 24]
BIN,TOM	0,59339	[21, 25, 18, 25, 18, 23, 18, 22, 20, 25]
CLU90,TOM	0,57165	[22, 26, 19, 23, 19, 20, 23, 24, 16, 18]
PVR,TOM	0,630213	[23, 28, 20, 26, 20, 24, 29, 25, 21, 28]
GWA,APO12	0,622005	[28, 18, 27, 21, 27, 26, 20, 21, 29, 34]
SOR,GWA	0,574534	[29, 17, 24, 22, 26, 25, 30, 34, 31, 31]

CR1,GWA	0,610138	[25, 22, 29, 27, 28, 31, 22, 27, 40, 36]
CR1,PIC	0,564885	[33, 34, 26, 29, 35, 27, 27, 30, 23, 21]
CLU00,GWA	0,578305	[26, 32, 30, 28, 25, 29, 28, 29, 42, 32]
BIN,PIC	0,575421	[34, 35, 34, 34, 29, 30, 34, 26, 26, 20]
GWA,PVR	0,564885	[32, 19, 28, 33, 33, 28, 33, 37, 38, 42]
GWA,APO46	0,581411	[27, 29, 33, 30, 30, 38, 31, 35, 43, 39]
BIN,GWA	0,579414	[31, 30, 31, 31, 32, 33, 26, 28, 47, 40]
CLU90,GWA	0,588731	[30, 31, 32, 32, 31, 39, 32, 36, 49, 41]
PIC,APO12	0,595941	[36, 36, 35, 35, 37, 32, 35, 32, 32, 29]
PVR,APO46	0,59583	[35, 33, 36, 40, 40, 35, 40, 33, 27, 35]
PIC,PVR	0,572981	[37, 38, 40, 36, 50, 34, 39, 41, 28, 27]
CLU90,PIC	0,572649	[38, 40, 38, 37, 49, 36, 38, 40, 36, 26]
SOR,PIC	0,55823	[39, 37, 39, 38, 39, 37, 36, 39, 35, 30]
SOR,APO46	0,574867	[41, 39, 41, 41, 34, 45, 47, 54, 34, 43]
BIN,CLU00	0,528727	[53, 45, 46, 44, 45, 42, 37, 31, 41, 33]
CR1,APO46	0,532165	[40, 57, 42, 46, 43, 58, 45, 50, 30, 51]
BIN,APO46	0,554126	[43, 54, 37, 50, 48, 60, 44, 44, 37, 45]
CLU90,APO46	0,511424	[44, 51, 44, 42, 47, 53, 52, 38, 33, 44]
CLU00,SOR	0,555457	[56, 46, 48, 45, 36, 41, 46, 53, 57, 46]
CLU00,APO46	0,532941	[47, 56, 43, 39, 38, 55, 48, 43, 44, 47]
BIN,PVR	0,528949	[50, 43, 45, 52, 54, 46, 43, 52, 50, 38]
BIN,CLU90	0,55102	[55, 52, 53, 53, 61, 47, 54, 42, 39, 37]
PVR,APO12	0,573425	[51, 41, 49, 54, 55, 40, 55, 57, 48, 59]
SOR,PVR	0,539042	[49, 48, 54, 55, 46, 43, 51, 51, 55, 53]
CLU00,APO12	0,578305	[48, 53, 52, 43, 42, 44, 58, 58, 56, 54]
APO46,APO12	0,534383	[54, 50, 47, 51, 52, 62, 53, 49, 45, 52]
BIN,SOR	0,540705	[58, 49, 55, 57, 57, 54, 42, 48, 52, 50]
CR1,PVR	0,540927	[45, 44, 50, 58, 60, 48, 57, 61, 46, 48]
CLU90,PVR	0,500998	[46, 42, 59, 59, 56, 49, 49, 47, 51, 49]
CR1,BIN	0,55224	[60, 55, 56, 56, 58, 50, 41, 45, 53, 56]
CLU00,PVR	0,508097	[42, 47, 51, 49, 51, 52, 59, 64, 54, 57]
CR1,CLU00	0,546917	[52, 58, 58, 47, 44, 51, 62, 59, 61, 58]
CLU90,CLU00	0,52551	[57, 59, 57, 48, 41, 56, 63, 60, 60, 55]
BIN,APO12	0,522183	[63, 60, 60, 60, 59, 59, 50, 46, 59, 60]
CLU90,SOR	0,484139	[61, 61, 63, 62, 53, 57, 64, 63, 63, 62]
CR1,CLU90	0,488021	[62, 64, 62, 61, 64, 64, 65, 55, 58, 63]
CR1,SOR	0,512422	[59, 62, 61, 64, 62, 61, 61, 65, 64, 64]
CLU90,APO12	0,468722	[65, 65, 66, 63, 63, 63, 60, 56, 62, 61]

SOR,APO12	0,499335	[66, 63, 64, 65, 65, 65, 56, 62, 66, 65]
CR1,APO12	0,475599	[64, 66, 65, 66, 66, 66, 66, 66, 65, 66]

Modelos de tre SNP	Precisión balanceada	*Ranking* en la validación cruzada
CLU00,PIC,APO58	0,704747	[1, 2, 3, 1, 1, 1, 3, 1, 4, 1]
PIC,TOM,APO58	0,708962	[2, 1, 1, 3, 2, 2, 2, 3, 1, 3]
PIC,APO46,APO58	0,690994	[3, 3, 10, 9, 3, 3, 4, 2, 2, 5]
BIN,PIC,APO58	0,70508	[4, 5, 9, 8, 6, 4, 5, 6, 3, 2]
PIC,GWA,APO58	0,677573	[6, 9, 7, 2, 5, 5, 6, 4, 8, 6]
SOR,GWA,APO58	0,671584	[9, 4, 2, 4, 7, 6, 7, 13, 5, 12]
PIC,APO58,APO12	0,716615	[10, 14, 17, 17, 8, 7, 14, 5, 9, 11]
SOR,PIC,APO58	0,708629	[11, 10, 14, 13, 9, 8, 16, 15, 13, 8]
CR1,GWA,APO58	0,69421	[7, 15, 6, 10, 21, 16, 1, 7, 21, 14]
CLU00,GWA,APO58	0,669255	[5, 7, 5, 7, 4, 26, 9, 8, 26, 13]
GWA,TOM,APO58	0,680901	[28, 8, 4, 5, 14, 18, 13, 14, 12, 16]
GWA,PVR,APO58	0,69421	[20, 6, 11, 14, 17, 11, 8, 10, 14, 21]
CLU90,PIC,APO58	0,689774	[12, 16, 26, 28, 11, 10, 22, 16, 17, 4]
CR1,PIC,APO58	0,657498	[18, 18, 19, 15, 12, 9, 17, 9, 22, 10]
PIC,PVR,APO58	0,670918	[16, 19, 31, 24, 16, 12, 32, 12, 10, 7]
CLU90,TOM,APO58	0,695985	[27, 22, 12, 18, 13, 19, 10, 19, 43, 18]
CR1,TOM,APO58	0,69177	[21, 24, 15, 19, 18, 15, 18, 18, 31, 19]
CR1,PVR,APO58	0,682453	[8, 33, 23, 12, 46, 21, 24, 11, 6, 26]
CLU90,CLU00,APO58	0,701642	[19, 38, 42, 22, 22, 17, 28, 26, 11, 15]
CLU90,GWA,APO58	0,668367	[25, 13, 16, 23, 23, 24, 21, 27, 19, 30]
SOR,PVR,APO58	0,677019	[24, 28, 29, 36, 15, 13, 29, 36, 7, 24]
CR1,SOR,APO58	0,673802	[14, 26, 18, 11, 38, 14, 34, 30, 23, 23]
SOR,APO46,APO58	0,705745	[31, 31, 25, 33, 19, 38, 19, 21, 15, 34]
BIN,GWA,APO58	0,666482	[13, 25, 28, 27, 31, 23, 27, 32, 18, 37]
CLU00,PVR,APO58	0,682121	[26, 20, 36, 21, 10, 25, 37, 20, 20, 38]
CLU00,PIC,TOM	0,680568	[29, 17, 27, 6, 24, 45, 30, 23, 56, 9]
CLU00,TOM,APO58	0,673913	[38, 35, 8, 20, 33, 31, 23, 25, 40, 25]
GWA,APO58,APO12	0,708962	[32, 34, 21, 37, 26, 30, 25, 35, 36, 41]
CR1,APO46,APO58	0,676686	[15, 40, 34, 34, 43, 20, 11, 29, 24, 44]
CR1,BIN,APO58	0,674468	[44, 48, 22, 32, 41, 27, 12, 22, 16, 40]
CR1,CLU90,APO58	0,663265	[22, 50, 13, 30, 47, 28, 20, 34, 28, 45]
CLU90,PVR,APO58	0,707631	[17, 36, 46, 42, 42, 29, 35, 37, 32, 28]
CLU90,APO46,APO58	0,688886	[36, 37, 35, 39, 29, 33, 36, 33, 25, 22]

GWA,APO46,APO58	0,646628	[23, 23, 30, 25, 30, 34, 26, 38, 33, 43]
CR1,CLU00,APO58	0,6687	[30, 55, 20, 31, 32, 35, 33, 31, 29, 33]
BIN,CLU00,APO58	0,670253	[40, 27, 40, 35, 35, 32, 43, 28, 37, 20]
CLU00,APO46,APO58	0,640639	[42, 46, 24, 40, 34, 41, 31, 24, 35, 27]
BIN,SOR,APO58	0,653394	[37, 45, 33, 45, 36, 37, 15, 41, 27, 48]
TOM,APO46,APO58	0,687001	[47, 44, 41, 38, 48, 39, 38, 40, 34, 46]
BIN,TOM,APO58	0,66748	[34, 53, 44, 43, 20, 40, 47, 50, 38, 47]
BIN,CLU90,APO58	0,664264	[35, 32, 43, 53, 44, 22, 39, 59, 47, 36]
CLU90,SOR,APO58	0,696207	[45, 58, 38, 48, 25, 50, 46, 48, 42, 65]
BIN,PVR,APO58	0,680901	[33, 51, 39, 54, 39, 36, 41, 54, 48, 50]
PVR,APO46,APO58	0,663598	[55, 42, 53, 49, 56, 56, 49, 17, 30, 42]
CLU00,APO58,APO12	0,666482	[43, 52, 51, 46, 45, 42, 52, 49, 39, 54]
CLU00,SOR,APO58	0,648292	[48, 49, 56, 41, 27, 43, 51, 43, 41, 71]
SOR,APO58,APO12	0,688886	[49, 70, 57, 56, 49, 46, 53, 52, 51, 59]
PVR,TOM,APO58	0,673248	[60, 71, 48, 51, 51, 47, 54, 51, 44, 56]
TOM,APO58,APO12	0,686335	[50, 67, 52, 57, 52, 48, 55, 53, 52, 61]
CR1,APO58,APO12	0,658496	[53, 72, 49, 58, 53, 49, 50, 42, 46, 63]
SOR,TOM,APO58	0,650288	[61, 66, 47, 47, 37, 44, 40, 46, 55, 64]
BIN,APO46,APO58	0,661713	[52, 73, 58, 55, 54, 51, 45, 47, 54, 52]
PIC,GWA,TOM	0,640639	[46, 29, 45, 16, 40, 57, 44, 44, 58, 49]
BIN,APO58,APO12	0,689219	[56, 77, 59, 62, 57, 52, 58, 55, 49, 53]
CLU90,APO58,APO12	0,689219	[57, 78, 60, 63, 58, 53, 59, 56, 50, 57]
PVR,APO58,APO12	0,689219	[58, 79, 61, 64, 59, 54, 60, 57, 45, 69]
APO46,APO58,APO12	0,683784	[59, 63, 62, 65, 60, 55, 61, 58, 53, 70]
SOR,GWA,TOM	0,65528	[41, 12, 32, 26, 62, 65, 66, 76, 63, 75]
CLU90,PIC,APO46	0,664264	[39, 30, 72, 60, 66, 59, 56, 45, 57, 35]
PIC,TOM,APO46	0,651176	[54, 11, 67, 61, 65, 58, 67, 39, 64, 32]
CLU00,PIC,GWA	0,666704	[51, 57, 63, 29, 63, 61, 48, 62, 91, 39]
BIN,PIC,TOM	0,636535	[66, 47, 66, 59, 55, 66, 64, 68, 59, 17]
PIC,TOM,APO12	0,670364	[72, 64, 37, 71, 50, 69, 73, 60, 65, 60]
CLU00,GWA,TOM	0,62012	[63, 62, 50, 44, 28, 74, 63, 69, 74, 76]
SOR,PIC,TOM	0,623447	[64, 43, 55, 68, 64, 67, 65, 75, 62, 62]
CLU90,GWA,TOM	0,621451	[68, 56, 69, 50, 61, 68, 42, 70, 60, 85]
CR1,PIC,TOM	0,635648	[71, 60, 76, 66, 67, 60, 78, 63, 61, 55]
CLU90,PIC,TOM	0,634317	[67, 75, 65, 67, 68, 62, 74, 65, 67, 29]
GWA,TOM,APO12	0,612689	[76, 39, 54, 52, 69, 80, 57, 67, 70, 100]
BIN,CLU00,PIC	0,598824	[78, 85, 84, 79, 91, 63, 62, 61, 86, 31]
PIC,PVR,TOM	0,61646	[69, 82, 70, 74, 73, 75, 85, 72, 68, 73]

SOR,PIC,APO46	0,6189	[83, 59, 90, 75, 77, 64, 79, 71, 81, 66]
PIC,GWA,APO46	0,603594	[75, 21, 86, 87, 82, 71, 84, 66, 102, 74]
CLU00,PIC,APO46	0,597493	[73, 74, 85, 76, 85, 70, 75, 73, 98, 58]
BIN,GWA,TOM	0,594499	[62, 68, 78, 69, 71, 83, 68, 80, 75, 96]
CR1,GWA,TOM	0,608252	[79, 54, 68, 70, 70, 89, 69, 78, 71, 116]
BIN,PIC,APO46	0,593944	[86, 76, 97, 93, 83, 73, 76, 77, 78, 67]
GWA,PVR,TOM	0,569987	[80, 41, 64, 73, 75, 85, 80, 64, 87, 107]
SOR,PIC,GWA	0,596828	[90, 61, 71, 77, 102, 76, 86, 93, 111, 68]
CR1,PIC,APO46	0,616016	[92, 69, 75, 98, 112, 72, 87, 74, 100, 77]
CLU90,PVR,TOM	0,597826	[65, 100, 83, 97, 80, 78, 89, 83, 89, 83]
CLU90,SOR,TOM	0,600155	[84, 104, 81, 94, 76, 77, 99, 85, 66, 78]
GWA,TOM,APO46	0,553017	[70, 65, 79, 72, 88, 87, 77, 79, 88, 111]
CLU00,PVR,TOM	0,624113	[74, 106, 74, 83, 72, 81, 107, 104, 83, 98]
SOR,TOM,APO46	0,651508	[87, 97, 77, 92, 84, 109, 96, 95, 76, 97]
BIN,PIC,GWA	0,574867	[81, 91, 107, 81, 105, 79, 71, 81, 107, 84]
CLU90,TOM,APO46	0,55213	[94, 110, 92, 89, 100, 86, 94, 82, 69, 81]
SOR,TOM,APO12	0,641304	[107, 102, 73, 110, 81, 98, 83, 91, 72, 131]
CR1,CLU00,PIC	0,58374	[116, 92, 104, 78, 117, 88, 72, 86, 96, 82]
CLU00,TOM,APO12	0,645519	[109, 107, 82, 86, 79, 95, 109, 98, 77, 117]
BIN,CLU00,TOM	0,553461	[112, 112, 103, 82, 74, 91, 81, 90, 97, 86]
CLU90,CLU00,TOM	0,624778	[96, 129, 110, 85, 92, 90, 118, 102, 82, 93]
SOR,PVR,TOM	0,613132	[91, 116, 89, 117, 87, 82, 92, 105, 95, 110]
CR1,PIC,GWA	0,617236	[108, 98, 108, 112, 130, 103, 70, 97, 108, 80]
CR1,TOM,APO12	0,638199	[129, 108, 80, 105, 86, 104, 120, 87, 84, 129]
TOM,APO46,APO12	0,621229	[110, 99, 88, 107, 90, 111, 111, 100, 85, 118]
BIN,TOM,APO46	0,580634	[77, 115, 100, 106, 98, 99, 98, 88, 92, 104]
CLU90,TOM,APO12	0,586624	[123, 119, 87, 115, 89, 96, 108, 103, 73, 103]
CLU90,PIC,GWA	0,592059	[88, 88, 112, 90, 99, 94, 101, 106, 121, 79]
CLU00,PVR,APO46	0,574645	[85, 113, 114, 99, 96, 105, 123, 101, 80, 109]
PIC,APO46,APO12	0,598713	[99, 86, 127, 129, 120, 84, 106, 84, 118, 94]
CR1,CLU90,TOM	0,620896	[119, 137, 96, 103, 93, 114, 112, 94, 79, 130]
CLU00,SOR,TOM	0,556012	[103, 111, 101, 100, 78, 92, 93, 118, 90, 112]
SOR,GWA,PVR	0,613354	[95, 81, 95, 84, 118, 102, 119, 128, 138, 106]
PIC,PVR,APO46	0,590506	[105, 80, 120, 114, 128, 101, 102, 89, 125, 89]
CLU00,SOR,GWA	0,632542	[120, 83, 98, 80, 107, 93, 134, 153, 143, 99]
BIN,TOM,APO12	0,585071	[124, 122, 91, 120, 97, 124, 88, 115, 101, 126]
BIN,CLU90,PIC	0,566548	[89, 123, 133, 122, 122, 115, 91, 96, 109, 51]
BIN,SOR,TOM	0,588287	[102, 124, 93, 128, 101, 126, 95, 110, 105, 134]

PIC,GWA,APO12	0,624778	[106, 90, 124, 108, 127, 106, 90, 99, 149, 125]
PVR,TOM,APO12	0,62988	[136, 125, 94, 123, 95, 125, 132, 120, 99, 144]
CR1,CLU00,TOM	0,608807	[122, 152, 105, 88, 94, 110, 126, 111, 106, 133]
CR1,PVR,TOM	0,620896	[93, 135, 99, 125, 113, 121, 149, 107, 93, 135]
PVR,TOM,APO46	0,585071	[100, 117, 102, 121, 114, 132, 128, 92, 104, 101]
CLU90,CLU00,GWA	0,608585	[121, 121, 131, 95, 111, 120, 105, 123, 145, 102]
BIN,CLU90,TOM	0,526287	[101, 139, 123, 96, 106, 116, 100, 113, 94, 87]
CLU00,SOR,PIC	0,56311	[126, 103, 118, 113, 104, 117, 125, 131, 137, 88]
BIN,GWA,APO46	0,578194	[82, 127, 117, 131, 109, 107, 110, 125, 110, 141]
CLU90,PVR,APO46	0,59949	[117, 96, 121, 126, 141, 100, 163, 117, 103, 114]
CLU90,CLU00,PIC	0,571096	[134, 136, 139, 109, 131, 97, 104, 122, 130, 72]
CLU00,GWA,PVR	0,604481	[98, 89, 113, 102, 125, 119, 140, 143, 158, 128]
SOR,GWA,APO46	0,616016	[128, 84, 119, 116, 132, 113, 122, 141, 156, 108]
PIC,GWA,PVR	0,579636	[125, 95, 128, 101, 138, 108, 117, 108, 140, 113]
CLU90,GWA,APO46	0,597715	[104, 118, 130, 104, 119, 134, 138, 116, 134, 123]
CLU00,PIC,PVR	0,579082	[113, 133, 138, 91, 142, 122, 113, 134, 123, 90]
CLU90,GWA,PVR	0,569321	[97, 87, 132, 124, 124, 127, 115, 135, 146, 147]
BIN,SOR,GWA	0,569987	[114, 93, 111, 118, 134, 118, 97, 142, 161, 132]
BIN,PVR,TOM	0,554459	[132, 149, 109, 119, 103, 131, 114, 127, 119, 127]
CR1,SOR,TOM	0,611247	[111, 157, 115, 149, 115, 141, 131, 126, 113, 161]
CLU00,TOM,APO46	0,575311	[130, 154, 116, 111, 110, 128, 137, 119, 115, 140]
CR1,TOM,APO46	0,600821	[115, 148, 122, 143, 116, 143, 150, 129, 112, 142]
BIN,SOR,PIC	0,54004	[151, 131, 135, 154, 140, 112, 82, 109, 120, 91]
CR1,BIN,PIC	0,555346	[154, 156, 129, 132, 139, 130, 103, 140, 117, 92]
CR1,BIN,TOM	0,580634	[127, 168, 106, 148, 121, 140, 127, 121, 114, 167]
CLU00,PIC,APO12	0,608252	[142, 153, 145, 139, 135, 123, 147, 139, 151, 105]
BIN,CLU00,GWA	0,540262	[149, 126, 148, 135, 123, 129, 121, 114, 172, 115]
CR1,PIC,PVR	0,582187	[146, 147, 144, 127, 180, 162, 116, 148, 116, 122]
BIN,PVR,APO46	0,576863	[159, 144, 140, 165, 149, 160, 129, 132, 127, 120]
CR1,CLU00,GWA	0,583407	[148, 145, 152, 136, 133, 135, 136, 144, 167, 169]
CLU90,SOR,GWA	0,523957	[147, 94, 125, 133, 137, 136, 155, 149, 180, 151]
CLU00,GWA,APO46	0,554902	[133, 160, 126, 130, 136, 166, 153, 133, 163, 152]
BIN,GWA,PVR	0,530612	[118, 109, 134, 137, 144, 145, 133, 166, 182, 153]
GWA,PVR,APO46	0,535825	[141, 101, 142, 145, 143, 137, 143, 112, 139, 174]
SOR,PVR,APO46	0,56189	[138, 150, 146, 164, 126, 173, 169, 160, 129, 139]
CR1,SOR,GWA	0,619232	[152, 105, 137, 141, 153, 142, 154, 180, 187, 179]
CLU00,GWA,APO12	0,59583	[157, 130, 149, 134, 146, 149, 142, 157, 189, 171]
CR1,BIN,GWA	0,578194	[135, 142, 155, 138, 161, 138, 146, 164, 188, 182]

BIN,PIC,PVR	0,513199	[131, 132, 151, 159, 160, 148, 130, 146, 132, 95]
CR1,GWA,APO12	0,613687	[161, 128, 141, 140, 172, 158, 141, 147, 196, 193]
CLU00,SOR,APO46	0,558119	[171, 155, 136, 152, 108, 164, 173, 186, 150, 148]
SOR,PIC,PVR	0,55091	[162, 159, 171, 151, 169, 159, 164, 155, 141, 124]
BIN,SOR,PVR	0,567436	[165, 166, 165, 181, 155, 154, 124, 136, 148, 156]
BIN,CLU90,PVR	0,518855	[144, 138, 172, 178, 166, 133, 135, 124, 147, 121]
CR1,GWA,APO46	0,562888	[137, 161, 157, 155, 150, 175, 148, 165, 170, 177]
GWA,APO46,APO12	0,583962	[150, 134, 169, 150, 186, 171, 152, 138, 178, 178]
SOR,GWA,APO12	0,562666	[168, 120, 150, 146, 154, 150, 145, 154, 181, 175]
BIN,GWA,APO12	0,575643	[166, 140, 160, 144, 181, 165, 144, 151, 193, 189]
CLU90,GWA,APO12	0,605701	[153, 146, 166, 147, 178, 170, 157, 156, 204, 194]
PVR,APO46,APO12	0,610803	[163, 158, 180, 171, 177, 167, 177, 158, 135, 164]
BIN,CLU90,GWA	0,523292	[140, 141, 158, 162, 151, 174, 158, 152, 159, 155]
CR1,GWA,PVR	0,536158	[143, 114, 147, 142, 176, 139, 151, 171, 169, 196]
BIN,PIC,APO12	0,547027	[158, 177, 164, 170, 168, 155, 139, 159, 153, 160]
PIC,PVR,APO12	0,580634	[160, 165, 177, 161, 200, 144, 172, 176, 155, 143]
CR1,PVR,APO46	0,516193	[139, 162, 170, 153, 165, 161, 165, 172, 122, 149]
BIN,CLU90,SOR	0,529503	[181, 151, 186, 189, 171, 153, 161, 150, 124, 119]
CR1,PIC,APO12	0,552462	[172, 178, 143, 167, 187, 168, 159, 168, 144, 154]
CR1,SOR,PIC	0,53394	[169, 179, 167, 166, 196, 177, 170, 189, 133, 146]
GWA,PVR,APO12	0,563221	[176, 143, 168, 158, 189, 178, 162, 167, 202, 200]
CR1,CLU90,GWA	0,530612	[156, 167, 156, 160, 170, 187, 167, 161, 164, 198]
CLU90,PIC,PVR	0,539042	[145, 175, 183, 172, 195, 176, 171, 182, 175, 136]
CR1,CLU90,APO46	0,5315	[155, 198, 163, 168, 145, 207, 178, 185, 136, 173]
CLU00,SOR,PVR	0,513199	[164, 164, 185, 163, 129, 169, 195, 188, 154, 188]
BIN,CLU90,APO46	0,487245	[174, 186, 178, 184, 157, 179, 179, 137, 128, 157]
BIN,CLU90,APO12	0,561335	[196, 187, 191, 188, 197, 151, 185, 130, 126, 159]
CR1,CLU00,PVR	0,547693	[170, 174, 174, 156, 162, 156, 202, 181, 165, 183]
CR1,CLU90,PIC	0,509982	[178, 183, 154, 169, 193, 183, 160, 175, 142, 145]
CLU90,SOR,PIC	0,517303	[187, 171, 161, 179, 164, 147, 192, 177, 152, 137]
CLU90,SOR,PVR	0,541038	[177, 169, 202, 198, 159, 152, 186, 173, 171, 162]
BIN,SOR,APO46	0,579193	[188, 188, 175, 187, 175, 195, 166, 200, 160, 168]
CLU90,CLU00,APO46	0,522959	[190, 195, 153, 157, 156, 194, 191, 162, 162, 158]
CLU90,SOR,APO46	0,48181	[186, 172, 162, 180, 152, 146, 197, 163, 131, 180]
BIN,CLU00,SOR	0,498114	[201, 163, 190, 174, 158, 163, 156, 183, 173, 165]
CLU90,CLU00,PVR	0,528061	[167, 176, 187, 192, 147, 172, 199, 203, 177, 166]
CLU90,PIC,APO12	0,546695	[180, 190, 184, 177, 205, 157, 183, 184, 197, 138]
BIN,CLU00,APO46	0,475377	[175, 197, 159, 173, 183, 188, 168, 169, 168, 163]

CR1,SOR,APO46	0,518523	[185, 185, 176, 176, 163, 193, 182, 202, 174, 181]
SOR,APO46,APO12	0,539596	[184, 173, 189, 191, 148, 201, 187, 194, 184, 201]
BIN,CLU90,CLU00	0,51575	[194, 200, 205, 194, 184, 184, 180, 145, 185, 150]
SOR,PIC,APO12	0,54736	[182, 194, 192, 185, 207, 189, 174, 187, 206, 184]
CR1,BIN,APO46	0,538598	[183, 210, 181, 183, 191, 211, 188, 192, 176, 191]
CR1,SOR,PVR	0,526287	[179, 199, 199, 209, 173, 181, 198, 197, 179, 197]
CLU90,APO46,APO12	0,517968	[193, 205, 194, 182, 199, 197, 208, 174, 157, 192]
CR1,CLU00,APO46	0,53028	[173, 209, 179, 190, 182, 208, 190, 208, 166, 195]
BIN,CLU00,PVR	0,436114	[197, 182, 198, 197, 190, 185, 175, 170, 191, 172]
CLU90,CLU00,SOR	0,514752	[210, 189, 204, 186, 167, 180, 210, 195, 205, 176]
CLU00,PVR,APO12	0,596162	[189, 184, 197, 199, 179, 190, 212, 215, 190, 210]
CR1,BIN,PVR	0,509871	[195, 181, 182, 201, 211, 192, 184, 205, 200, 170]
BIN,PVR,APO12	0,567214	[207, 180, 195, 200, 206, 196, 193, 207, 201, 190]
BIN,CLU00,APO12	0,545364	[208, 203, 200, 204, 201, 198, 176, 179, 198, 186]
CR1,CLU00,SOR	0,534605	[199, 193, 210, 175, 174, 182, 209, 201, 212, 206]
BIN,APO46,APO12	0,537822	[202, 206, 173, 207, 203, 214, 194, 191, 186, 202]
CR1,CLU90,PVR	0,473602	[191, 170, 196, 206, 204, 186, 201, 178, 194, 208]
SOR,PVR,APO12	0,553793	[198, 196, 201, 205, 185, 191, 200, 198, 211, 211]
CR1,BIN,CLU00	0,497671	[200, 202, 209, 195, 194, 205, 181, 190, 209, 185]
CLU00,APO46,APO12	0,531943	[206, 212, 188, 196, 192, 217, 204, 196, 192, 199]
CR1,APO46,APO12	0,51331	[192, 216, 193, 203, 210, 215, 203, 212, 183, 203]
CR1,BIN,SOR	0,544033	[205, 204, 207, 213, 212, 203, 189, 206, 203, 204]
CR1,CLU90,CLU00	0,499667	[212, 208, 203, 193, 188, 200, 213, 193, 195, 205]
CR1,PVR,APO12	0,552351	[203, 191, 206, 212, 213, 199, 214, 219, 199, 214]
CLU90,CLU00,APO12	0,521739	[214, 211, 212, 202, 202, 206, 215, 211, 210, 207]
CR1,BIN,CLU90	0,450865	[209, 207, 208, 211, 215, 209, 205, 199, 207, 187]
CLU00,SOR,APO12	0,518079	[211, 201, 211, 208, 198, 204, 211, 216, 218, 213]
CLU90,PVR,APO12	0,495453	[204, 192, 213, 215, 214, 202, 207, 209, 208, 209]
BIN,SOR,APO12	0,52429	[218, 213, 215, 217, 216, 213, 196, 214, 213, 212]
CR1,CLU00,APO12	0,520408	[213, 217, 214, 210, 209, 210, 219, 220, 219, 217]
CR1,BIN,APO12	0,527618	[219, 215, 216, 216, 217, 216, 206, 210, 216, 215]
CR1,CLU90,SOR	0,425133	[215, 214, 217, 214, 208, 218, 216, 204, 214, 216]
CLU90,SOR,APO12	0,498004	[217, 218, 218, 218, 218, 212, 218, 213, 217, 218]
CR1,SOR,APO12	0,527063	[216, 219, 219, 219, 219, 219, 217, 218, 220, 220]
CR1,CLU90,APO12	0,469831	[220, 220, 220, 220, 220, 220, 220, 217, 215, 219]

Bibliografía

[1] Crawford, D. C., Akey, D. T., & Nickerson, D. A. (2005). The patterns of natural variation in human genes. *Annu. Rev. Genomics Hum. Genet.*, *6*, 287-312.

[2] Forero, D. A., Benítez, B., Arboleda, G., Yunis, J. J., Pardo, R., & Arboleda, H. (2006). Analysis of functional polymorphisms in three synaptic plasticity-related genes (BDNF, COMT AND UCHL1) in Alzheimer's disease in Colombia.*Neuroscience research*, *55*(3), 334-341.

[3] Sieh, W., Yu, C. E., Bird, T. D., Schellenberg, G. D., & Wijsman, E. M. (2007). Accounting for linkage disequilibrium among markers in linkage analysis: impact of haplotype frequency estimation and molecular haplotypes for a gene in a candidate region for Alzheimer's disease. *Human heredity*, *63*(1), 26-34.

[4] Forero, D. A., Arboleda, G., Yunis, J. J., Pardo, R., & Arboleda, H. (2006). Association study of polymorphisms in LRP1, tau and 5-HTT genes and Alzheimer's disease in a sample of Colombian patients. *Journal of neural transmission*, *113*(9), 1253-1262.

[5] Maulik, U., Bandyopadhyay, S., & Wang, J. T. (2011). *Computational Intelligence and Pattern Analysis in Biology Informatics* (Vol. 20). Wiley. com.

[6] Lazarczyk, M. J., Hof, P. R., Bouras, C., & Giannakopoulos, P. (2012). Preclinical Alzheimer disease: identification of cases at risk among cognitively intact older individuals. *BMC medicine*, *10*(1), 127.

[7] Povova, J., Ambroz, P., Bar, M., Pavukova, V., Sery, O., Tomaskova, H., & Janout, V. (2012). Epidemiological of and risk factors for Alzheimer's disease: A review. *Biomedical Papers*, *156*(2), 108-114.

[8] Paulson, H. L., & Igo, I. (2011, November). Genetics of dementia. In *Seminars in neurology* (Vol. 31, No. 5, p. 449). NIH Public Access.

[9] Roberts, J. S., Christensen, K. D., & Green, R. C. (2011). Using Alzheimer's disease as a model for genetic risk disclosure: implications for personal genomics. *Clinical genetics*, *80*(5), 407-414.

[10] Nelson, P. T., Head, E., Schmitt, F. A., Davis, P. R., Neltner, J. H., Jicha, G. A., ... & Scheff, S. W. (2011). Alzheimer's disease is not "brain aging": neuropathological, genetic, and epidemiological human studies. *Acta neuropathologica*, *121*(5), 571-587.

[11] Montana, G. (2006). Statistical methods in genetics. *Briefings in Bioinformatics*, *7*(3), 297-308.

[12] Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, *10*(6), 392-404.

[13] Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic mapping in human disease.*science*,*322* (5903), 881-888.

[14] Attia, J., Ioannidis, J. P., Thakkinstian, A., McEvoy, M., Scott, R. J., Minelli, C., & Guyatt, G. (2009). How to use an article about genetic association.*JAMA: the journal of the American Medical Association*, *301*(2), 191-197.

[15] Uhmn, S., Kim, D. H., Kim, J., Cho, S. W., & Cheong, J. Y. (2007, October). Chronic Hepatitis Classification Using SNP Data and Data Mining Techniques. In *Frontiers in the Convergence of Bioscience and Information Technologies, 2007. FBIT 2007* (pp. 81-86). IEEE.

[16] Lewis, C. M. (2002). Genetic association studies: design, analysis and interpretation. *Briefings in bioinformatics*, *3*(2), 146-153.

[17] Engelbrecht, A. P. (2007). *Computational intelligence: an introduction*. Wiley.com.

[18] Combarros, O., Cortina-Borja, M., Smith, A. D., & Lehmann, D. J. (2009). Epistasis in sporadic Alzheimer's disease. *Neurobiology of aging*, *30*(9), 1333-1349.

[19] Oh, S., Lee, J., Kwon, M. S., Weir, B., Ha, K., & Park, T. (2012). A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. *BMC bioinformatics*, *13*(Suppl 9), S5.

[20] Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N., & White, B. C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of theoretical biology*, *241*(2), 252-261.

[21] Fisher, R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edin., 52, 399–433.

[22] Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human molecular genetics, 11(20), 2463-2468.

[23] Gibson, G., Wagner, G., 2000. Canalization in evolutionary genetics: a stabilizing theory? BioEssays 22, 372–380.

[24] Phillips, P.C., 1998. The language of gene interaction. Genetics 149, 1167–1171.

[25] Proulx, S.R., Phillips, P.C., 2005. The opportunity for canalization and the evolution of genetic networks. Am. Nat. 165, 147–162.

[26] Moore, J.H., Williams, S.W., 2002. New strategies for identifying gene–gene interactions in hypertension. Ann. Med. 34, 88–95.

[27] Moore, J.H., Williams, S.W., 2005. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays 27, 637–646.

[28] Wimo A, Winblad B, Aguero-Torres H, von Strauss E (2003) The magnitude of dementia occurrence in the world. Alzheimer Dis Assoc Disord 17: 63–67.

[29] Warwick Daw E, Payami H, Nemens EJ, Nochlin D, Bird TD, Schellenberg GD, Wijsman EM (2000) The number of trait loci in late-onset Alzheimer disease. Am J Hum Genet 66: 196–204.

[30] Alpaydin, Ethem. Introduction to machine learning. The MIT Press, 2004.

[31] Larose, Daniel T. Discovering knowledge in data: an introduction to data mining. Wiley-Interscience, 2005.

[32] Pal, Sankar K., and Pabitra Mitra. Pattern recognition algorithms for data mining. Chapman and Hall/CRC, 2004.

[33] E.R. Martin, M.D. Ritchie, L. Hahn, S. Kang, J.H. Moore, A novel method to identify gene–gene effects in nuclear families: the MDR-PDT, Genet. Epidemiol. 30 (2006) 111–123.

[34] A.S. Andrew, H.H. Nelson, K.T. Kelsey, J.H. Moore, A.C. Meng, D.P. Casella, T.D. Tosteson, A.R. Schned, M. Karagas, Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNP, smoking and bladder cancer susceptibility, Carcinogenesis 27 (2006) 1030–1037.

[35] D. Brassat, A.A.Motsinger, S.J. Caillier, H.A. Erlich, K.Walker, L.L. Steiner, B.A. Cree, L.F. Barcellos, M.A. Pericak-Vance, S. Schmidt, S. Gregory, S.L. Hauser, J.L. Haines, J.R. Oksenberg, M.D. Ritchie, Multifactor dimensionality reduction reveals gene–gene interactions associated with multiple sclerosis susceptibility in African Americans, Genes Immun. 7 (2006) 310–315.

[36] S. Qin, X. Zhao, Y. Pan, J. Liu, G. Feng, J. Fu, J. Bao, L. He, An association study of the N-methyl-D-aspartate receptor subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray, Eur. J. Hum. Gene 13 (2005) 807–814.

[37] Arboleda, G., Yunis, J., Pardo, R., Gómez, C., Hedmont, D., Arango, G. and Arboleda, H. Apolipoprotein E genotyping in a sample of Colombian patients with Alzheimer's disease Neuroscience letters, Elsevier, 2001, Vol. 305(2), pp. 135-138.

[38] Camelo, D., Arboleda, G., Yunis, J., Pardo, R., Arango, G., Solano, E., López, L., Hedmont, D. and Arboleda, H. Angiotensin-converting enzyme and alpha-2-macroglobulin gene polymorphisms are not associated with Alzheimer's disease in Colombian patients Journal of the neurological sciences, Elsevier, 2004, Vol. 218(1), pp. 47-51.

[39] Forero, D., Casadesus, G., Perry, G. and Arboleda, H. Synaptic dysfunction and oxidative stress in Alzheimer's disease: emerging mechanisms Journal of cellular and molecular medicine, Wiley Online Library, 2006, Vol. 10(3), pp. 796-805.

[40] Forero, D., Pinzón, J., Arboleda, G., Yunis, J., Alvarez, C., Cataño, N. and Arboleda, H. Analysis of common polymorphisms in angiotensin-converting enzyme and apolipoprotein E genes and human longevity in Colombia Archives of medical research, Elsevier, 2006, Vol. 37(7), pp. 890-894.

[41] Bushlin, I., Petralia, R., Wu, F., Harel, A., Mughal, M., Mattson, M. and Yao, P. Clathrin assembly protein AP180 and CALM differentially control axogenesis and dendrite outgrowth in embryonic hippocampal neurons The Journal of Neuroscience, Society for Neuroscience, 2008, Vol. 28(41), pp. 10257-10271.

[42] Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 2003, 56:73-82.

[43] Kohavi, R. & others A study of cross-validation and bootstrap for accuracy estimation and model selection International joint Conference on artificial intelligence, 1995, 14, 1137-1145.

[44] Alzheimer Research Forum.{En línea} {15 marzo de 2013} disponible en: (http://www.alzgene.org/).

[45] Pradilla, A.G., Vesga, A.B., Leon-Sarmiento, F.E. National neuroepidemiological study in Colombia (EPINEURO). Rev. Panam. Salud. Publica 2003; (14): 104–111.

[46] Tomlinson BE, Blessed G, Roth M: Observations on the brains of nondemented old people. J Neurol Sci 1968, 7:331-356.

[47] Braak H, Braak E: Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol 1991, 82:239-259.

[48] Burke W, Psaty BM. Personalized medicine in the era of genomics. JAMA. 2007 298:1682–1684.[PubMed: 17925520].

[49] J. C. Bezdek (1994), What is Computational Intelligence? Computational Intelligence Imitating Life, J. M., Zurada, R. J. Marks, and C. J. Robinson (Eds.), IEEE Press, NY, pp. 1–12.

[50] R. J., Marks (1993), Intelligence: Computational versus Artificial, IEEE Trans. Neural Networks, 4: 737–739.

[51] W. Pedrycz and F. Gomide (1998), An Introduction to Fuzzy Sets: Analysis and Design, MIT Press, MA.

[52] Cox, R. T. (1946). Probability, frequency and reasonable expectation. American Journal of Physics 14(1), 1–13.

[53] Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.

[54] Robnik-Siknja, M., Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. 53, 23–69.

[55] Jakulin, A., Bratko, I., 2003. Analyzing attribute interactions. Lect. Notes Artif. Intell. 2838, 229–240.

[56] Hahn, L.W., Ritchie, M.D., Moore, J.H., 2003. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. Bioinformatics 19, 376–382.

[57] Ritchie, M.D., Hahn, L.W., Moore, J.H., 2003a. Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, phenocopy, and genetic heterogeneity. Genet. Epidemiol. 24, 150–157.

[58] Ritchie, M.D., White, B.C., Parker, J.S., Hahn, L.W., Moore, J.H., 2003b. Optimization of neural network architecture using genetic programming improves the detection and modeling of gene–gene interactions in studies of human diseases. BMC Bioinform. 4, 28.

[59] Jakulin, A., Bratko, I., Smrke, D., Demsar, J., Zupan, B., 2003. Attribute interactions in medical data analysis. Lect. Notes Artif. Intell. 2780, 229–238.

[60] Michalski, R.S., 1983. A theory and methodology of inductive learning. Artif. Intell. 20, 111–161.

[61] Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H., 2001. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69, 138–147.

[62] Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H., 2001. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69, 138–147.

[63] Moore, J.H., 2004. Computational analysis of gene–gene interactions in common human diseases using multifactor dimensionality reduction. Expert. Rev. Mol. Diagn. 4, 795–803.

[64] Moore, J.H., 2005. A global view of epistasis. Nat. Genet. 37, 13–14.

[65] Wnek, J., Michalski, R.S., 1994. Hypothesis-driven constructive induction in AQ17-HCI: a method and experiments. Mach. Learn. 14, 139–168.

[66] Bloedorn, E., Michalski, R.S., 1998. Data-driven constructive induction. IEEE Intell. Syst. 13, 30–37.

[67] Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Machine Learning: Proceedings of the AAAI'92 (1992)

[68] Moore, J. H., & White, B. C. (2007). Tuning ReliefF for genome-wide genetic analysis. In Evolutionary computation, machine learning and data mining in bioinformatics (pp. 166-175). Springer Berlin Heidelberg.

[69] Templeton AR: Epistasis and Complex Traits. In Epistasis and the Evolutionary Process Edited by: J W, B BIII and M W. New York, Oxford University Press; 2007:41-57.

[70] Greene, C. S., Penrod, N. M., Kiralis, J., & Moore, J. H. (2009). Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. BioData mining, 2(1), 1-9.

[71] Carrasquillo MM, McCallion AS, Puffenberger EG, Kashuk CS, Nouri N, Chakravarti A: Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. Nat Genet 2002, 32:237-244.

[72] Nicodemus, K. K., Callicott, J. H., Higier, R. G., Luna, A., Nixon, D. C., Lipska, B. K., ... & Weinberger, D. R. (2010). Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. Human genetics, 127(4), 441-452.

[73] Kallberg H, Padyukov L, Plenge RM, Ronnelid J, Gregersen PK, van der Helm-van Mil AH et al. Gene-gene and gene environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. Am J Hum Genet 2007; 80: 867–875.

[74] Hu, T., Chen, Y., Kiralis, J. W., & Moore, J. H. (2013). ViSEN: Methodology and Software for Visualization of Statistical Epistasis Networks. Genetic epidemiology.

[75] Cover TM, Thomas JA. Elements of information theory, 2nd edn. New York, NY: Wiley, 2006.

[76] Hu, T., Chen, Y., Kiralis, J. W., Collins, R. L., Wejse, C., Sirugo, G., ... & Moore, J. H. (2013). An information-gain approach to detecting three-way epistatic interactions in genetic association studies. Journal of the American Medical Informatics Association, 20(4), 630-636.

[77] Andrew, A. S., Nelson, H. N., Kelsey, K. T., Moore, J. H., Meng, A., Casella, D. P., Tosterson, T. D., Schned, A. R. & Karagas, M. R. (2006) Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking, and bladder cancer susceptibility. Carcinogenesis 27, 1030–1037.

[78] De Wit E, van der Merwe L, van Helden PD, Hoal EG: Gene-gene interaction between tuberculosis candidate genes in a South African population. Mamm Genome 2011, 22(1–2):100–110.

[79] Thorleifsson G, Magnusson KP, Sulem P, Walters GB, Gudbjartsson DF, et al. (2007) Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. Science 317: 1397–1400.

[80] Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, et al. (2007) Variants conferring risk of atrial fibrillation on chromosome 4q25. Nature 448: 353–357.

[81] Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, et al. (2007) GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron 54: 713–720.

[82] Arboleda, G. H., J. J. Yunis, et al. (2001). "Apolipoprotein E genotyping in a sample of Colombian patients with Alzheimer's disease." Neurosci Lett 305(2): 135-138.

[83] Ortega, J.C. (2013). Estudio de diez polimorfismos -SNPs- en pacientes con enfermedad de Alzheimer (EA) en una muestra colombiana. Aproximación a genotipos haploides. Universidad Nacional de Colombia. Tesis

[84] Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nature genetics, 39(1), 17-23.

[85] Bertram, L., & Tanzi, R. E. (2009). Genome-wide association studies in Alzheimer's disease. Human molecular genetics, 18(R2), R137-R145.

[86] Velez, D. R. et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol. 31, 306–315 (2007).

[87] McKinney, B. A., Reif, D. M., Ritchie, M. D. & Moore, J. H. Machine learning for detecting gene–gene interactions: a review. Appl. Bioinformatics 5, 77–88 (2006).

[88] Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genet. 37, 413–417 (2005).

[89] Chapman, J. & Clayton, D. Detecting association using epistatic information. Genet. Epidemiol. 31, 894–909 (2007).

[90] Bureau A, Dupuis J, Falls K, et al. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 2005 Feb; 28 (2): 171-82.

[91] Breiman L. Random forests. Mach Learn 2001; 45 (1): 5-32.

[92] Breiman L, Friedman JH, Olshen RA, et al. Classification and regression trees. Belmont (CA): Wadsworth International Group, 1984.

[93] Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[94] Amouyel, P. et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat. Genet. Advance online publication, doi:10.1038/ng.439 (6 September 2009).

[95] A. D. Roses, M. W. Lutz, H. Amrine-Madsen et al., "A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease," Pharmacogenomics Journal, vol. 10, no. 5, pp. 375–384, 2010.

[96] M. Mancuso, V. Calsolaro, D. Orsucci et al., "Mitochondria, cognitive impairment, and Alzheimer's disease," International Journal of Alzheimer's disease, vol. 2009, Article ID 951548, 8 pages, 2009.

[97] L. Devi, B.M. Prabhu, D. F. Galati, N. G. Avadhani, and H. K. Anandatheerthavarada, “Accumulation of amyloid precursor protein in the mitochondrial import channels of human Alzheimer’s disease brain is associated with mitochondrial dysfunction,” Journal of Neuroscience, vol. 26, no. 35, pp. 9057–9068, 2006.

[98] R. H. Swerdlow, “Brain aging, Alzheimer’s disease, and mitochondria,” Biochimica et Biophysica Acta, vol. 1812, no. 12, pp. 1630–1639, 2011.

[99] S. J. Baloyannis, “Mitochondria are related to synaptic pathology in Alzheimer’s disease,” International Journal of Alzheimer’s disease, vol. 2011, Article ID 305395, 7 pages, 2011.

[100] Roses AD. Alzheimer diseases: a model of gene mutations and susceptibility polymorphisms for complex psychiatric diseases. Am J Med Genet 1998;81:49–57.

[101] Ferencz, B., Karlsson, S., & Kalpouzos, G. (2012). Promising genetic biomarkers of preclinical Alzheimer's disease: the influence of APOE and TOMM40 on brain integrity. International Journal of Alzheimer's disease, 2012.

[102] J. L. Stein, X. Hua, J. H. Morra et al., “Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer’s disease,” NeuroImage, vol. 51, no. 2, pp. 542–554, 2010.

[103] L. Shen, S. Kim, S. L. Risacher et al., “Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort,” NeuroImage, vol. 53, no. 3, pp. 1051–1063, 2010.

[104] A. D. Roses, M. W. Lutz, H. Amrine-Madsen et al., “A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer’s disease,” Pharmacogenomics Journal, vol. 10, no. 5, pp. 375–384, 2010.

[105] DeMattos RB, Cirrito JR, Parsadanian M, May PC, O’Dell MA, et al. (2004) ApoE and clusterin cooperatively suppress Abeta levels and deposition: evidence that ApoE regulates extracellular Abeta metabolism in vivo. Neuron 41: 193–202.

[106] McLaughlin L, Zhu G, Mistry M, Ley-Ebert C, Stuart WD, et al. (2000) Apolipoprotein J/clusterin limits the severity of murine autoimmune myocarditis. J Clin Invest 106: 1105–1113.

[107] Kauwe, J. S., Cruchaga, C., Karch, C. M., Sadler, B., Lee, M., Mayo, K., ... & Goate, A. M. (2011). Fine mapping of genetic variants in BIN1, CLU, CR1 and PICALM for association with cerebrospinal fluid biomarkers for Alzheimer's disease. PloS one, 6(2), e15918.

[108] Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, et al. (2010) Genome-wide analysis of genetic loci associated with Alzheimer disease. Jama 303: 1832–1840.

[109] Masliah E, Mallory M, Alford M, DeTeresa R, Hansen LA, et al. (2001) Altered expression of synaptic proteins occurs early during progression of Alzheimer’s disease. Neurology 56: 127–129.

[110] Guo, L. H., Westerteicher, C., Wang, X. H., Kratzer, M., Tsolakidou, A., Jiang, M., ... & Perneczky, R. (2012). SORL1 genetic variants and cerebrospinal fluid biomarkers of Alzheimer’s disease. European archives of psychiatry and clinical neuroscience, 262(6), 529-534.

[111] Collins, R. L., Hu, T., Wejse, C., Sirugo, G., Williams, S. M., & Moore, J. H. (2013). Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis. BioData mining, 6(4).

[112] Huang, Y. (2010). "Mechanisms linking apolipoprotein E isoforms with cardiovascular and neurological diseases." Curr Opin Lipidol 21(4): 337-345.

[113] Mahley, R. W., K. H. Weisgraber, et al. (2006). "Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease." Proc Natl Acad Sci U S A 103(15): 5644-5651.

[114] Gui, J., Andrew, A. S., Andrews, P., Nelson, H. M., Kelsey, K. T., Karagas, M. R., & Moore, J. H. (2011). A robust multifactor dimensionality reduction method for detecting gene–gene interactions with application to the genetic analysis of bladder cancer susceptibility. Annals of human genetics, 75(1), 20-28.