UNIVERSIDAD NACIONAL DE COLOMBIA

Modelo computacional para la identificación de endofenotipos en pacientes con Artritis Reumatoide utilizando información del Antígeno Leucocitario Humano HLA clase II

Luis Antonio Morales Muñoz

Universidad Nacional de Colombia

Facultad de Medicina

Bogotá D.C, Colombia

2014

Modelo computacional para la identificación de endofenotipos en pacientes con Artritis Reumatoide utilizando información del Antígeno Leucocitario Humano HLA clase II

Luis Antonio Morales Muñoz

Tesis o trabajo de investigación presentado como requisito parcial para optar al título de:

Magister en Ingeniería Biomédica

Director:

Luis Fernando Niño Vásquez, Ph.D.

Grupo de Investigación:

Laboratorio en Sistemas Inteligentes - LISI

Universidad Nacional de Colombia

Facultad de Medicina

Bogotá D.C, Colombia

2014

A Dios a mis padres y hermanas.

Agradecimientos

Agradezco en primera instancia a Dios por estar siempre presente en mi vida y en la de mi familia. A mis padres y mis hermanas que siempre están conmigo.

A mi Director y Codirector de Tesis el profesor Luis Fernando Niño y el Dr. Gerardo Quintana, quienes siempre me orientaron en la realización de esta Tesis. A la profesora Myriam Navarrete por sus enseñanzas y apoyo.

A la Universidad Nacional y a la Fundación Santa Fe por brindarme las herramientas para lograr este objetivo.

A los integrantes de los grupos de investigación LISI (Laboratorio de Sistemas Inteligentes) de la Universidad Nacional y REUMAVANCE (Grupo de Reumatología de la Fundación Santa Fe) por sus oportunos comentarios y por permitirme trabajar con ellos.

Resumen

La Artritis Reumatoide es una enfermedad autoinmune que afecta del 0.4% al 1% de la población humana y existen marcadores genéticos implicados en esta enfermedad. Se han utilizado varios modelos de inteligencia computacional para la clasificación e identificación de endofenotipos (relación entre fenotipo y marcadores genéticos) en pacientes con Artritis Reumatoide y controles sanos a partir de información genética, principalmente el HLA DRB1 (Antígeno Leucocitario Humano), así como la teoría del Epítope Compartido. Esta hace referencia a la asociación entre la AR y el HLA-DRB1, principalmente los alelos que contienen un motivo común de aminoácidos de las secuencias QKRAA, QRRAA o RRRAA en las posiciones 70 a 74 de la cadena del DRB1, los cuales les confieren una susceptibilidad particular de la enfermedad a los individuos.

En este trabajo se logró desarrollar un modelo computacional para clasificación de pacientes de artritis reumatoide y controles utilizando técnicas de estadística aplicada como son las redes neuronales, las redes bayesianas y los métodos de kernel. Como datos de entrada se utilizaron las variables de los pacientes en relación a serología (factor reumatoide y citrulinas), número de articulaciones inflamadas y dolorosas, reactantes de inflamación (velocidad de sedimentación globular y Proteína C Reactiva), rigidez matinal, edad, género, antecedentes de co morbilidades y la información del alelo HLA-DRB1.

Se obtuvieron resultados importantes para el diagnóstico de la enfermedad, su categorización y como potencial aplicación en la medicina personalizada de los individuos afectados por esta enfermedad. A partir de esta información, se diseñaron y probaron varios modelos computacionales para clasificación. Particularmente,. Adicionalmente, se hicieron varios agrupamientos de los datos aspectos relacionados con laóel la algunas Para definir el número de grupos se utilizó el coeficiente de silueta. Finalmente se obtuvo un árbol de distancia entre las secuencias de aminoácidos de los

distintos alelos HLA DRB1 lo cual permitió además visualizar los grupos con cercanía genética.

Los métodos utilizados permiten una mejor estratificación de la enfermedad en relación a la predicción de fenotipos y posibles desenlaces de la enfermedad, así como para la potencial prevención primaria de la enfermedad.

Palabras clave: HLA Antígeno Leucocitario Humano, Epítope Compartido, Inteligencia Computacional, Artritis Reumatoide

Abstract

Rheumatoid arthritis is an autoimmune disease that affects from 0.4% to 1% of the human population and genetic markers have been identified to be involved in the development of this disease. In this work, several computational intelligence models were used for the classification and identification of endophenotypes (relationship between phenotype and genetic markers) in patients with rheumatoid arthritis and healthy controls from genetic information, primarily the HLA-DRB1 (Human Leukocyte Antigen), and the Shared Epitope theory. This refers to association between RA and the HLA-DRB1 alleles, mainly containing common amino acid motif sequences QKRAA, RRRAA, QRRAA at positions 70 to 74 in the DRB1 sequence, which confer them particular disease susceptibility individuals.

A computational model for classification was proposed, particularly, some statistical and computational intelligence techniques such as neural networks, bayesian networks and kernel methods were applied. Input variables used in relation to patients were serology (rheumatoid factor, Anti–citrullinated protein antibodies ACPA), number of swollen and tender joints, inflammation reactants (erythrocyte sedimentation rate and C-reactive protein), morning stiffness, age, gender, history of co morbidities and the information of the HLA-DRB1.

We obtained promising results for disease diagnosis, and its categorization as potential application in personalized medicine for individuals suffering from this disease. From this

information we designed computer models for classification in which neural networks properly classified 137 instances of the 138 instances in total representing 99.2% and 0.72% classified improperly, bayesian networks properly classified 136 instances representing 98.5% and inadequately 2 instances representing 1.44% of the 138 instances. Finally we obtained a tree distance between amino acid sequences of the different alleles HLA DRB1 which allows us to visualize genetic proximity groups.

These applications provide better stratification of the disease in relation to the predicted phenotypes, and the potential for primary prevention of this disease.

Keywords: HLA Human Leukocyte Antigen, Shared Epitope, Computational Intelligence, Rheumatoid Arthritis.

Contenido

Pág.

Resumen ..... **V**

Lista de figuras ..... **X**

Lista de tablas ..... **XI**

Lista de abreviaturas **13**

1. Introducción **14**
1.1 Identificación del problema 15
1.2 Justificación 15
1.3 Objetivo general y objetivos específicos 18
1.3.1 Objetivo general 18
1.3.2 Objetivos específicos 18

2. Sobre la Artritis Reumatoide **19**
2.1 Aspectos Inmunológicos 19
2.1.1 Inmunidad Adaptativa 20
2.1.2 Inmunidad Innata 25
2.2 Factores Genéticos 25
2.3 Factores de Riesgo 27
2.4 Aspectos Clínicos 28
2.4.1 Diagnóstico 28
2.4.2 Tratamiento 30
2.4.3 Drogas Modificadoras de la AR (DMARDs) 31
2.4.4 Corticoesteroides 31
2.4.5 Agentes Biológicos 31

3. Aplicación de métodos informáticos en la clasificación de enfermedades **33**
3.1 Redes Neuronales 34
3.2 Redes Bayesianas 34
3.3 Métodos de Kernel 34
3.4 WEKA 34
3.5 KNIME 35
3.6 JALVIEW 35

4. Metodología para la clasificación de pacientes con artritis reumatoide **36**
4.1 Datos del estudio 36

5. Resultados **41**
5.1 Medidas de desempeño 41
5.2 Modelos Computaciones para clasificación de AR y riesgo de severidad. 42
5.3 Identificación de endofenotipos en pacientes con AR. 59

6. Discusión. **69**

7. Conclusiones y recomendaciones **72**
7.1 Conclusiones 72

7.2 Recomendaciones 73

Participaciones 74

Bibliografía 76

Lista de figuras

Pág.

Figura 2-1: Respuesta inmune en la Artritis Reumatoide. 22
Figura 4-1: Esquema de la metodología. 39
Figura 5-1: Red bayesiana para clasificación utilizando dos variables genéticas..... 43
Figura 5-2: Red neuronal perceptrón multicapa utilizando cinco variables..... 46
Figura 5-3: Red bayesiana utilizando cinco variables. 47
Figura 5-4: Gráfica coeficiente silueta para agrupamiento de pacientes y controles utilizando 5 variables 50
Figura 5-5: Visualización de las variables FR y Anti CCP. El color rojo corresponde al grupo 1 y el color azul al grupo 2. 52
Figura 5-6: Red bayesiana obtenida usando once variables..... 54
Figura 5-7: Relación entre la red bayesiana y la probabilidad de cada una de las variables con el desenlace de artritis reumatoide. 55
Figura 5-8: Coeficiente silueta para agrupamiento de pacientes 57
Figura 5-9: Distribución de los grupos según las variables articulaciones inflamadas y anti CCP. 59
Figura 5-10: Alineamiento de las secuencias de aminoácidos de pacientes con 28 articulaciones comprometidas. 60
Figura 5-11: Alineamiento de las secuencias de aminoácidos de pacientes con 28 articulaciones comprometidas; con el recuadro rojo se resalta el área correspondiente al epítope compartido y con el recuadro negro las secuencias en común. 60
Figura 5-12: Histograma de los alelos HLA DRB1 para el alelo 1. 63
Figura 5-13: Alineamiento de los distintos alelos HLA DRB1. 65
Figura 5-14: Árbol de distancia entre secuencias de aminoácidos de los distintos alelos HLA DRB1..... 66
Figura 6-1: Red bayesiana y sus tablas de probabilidades para la AR 71

Lista de tablas

Pág.

Tabla 2-1: Criterios de clasificación 2010 American College of Rheumatology European League Against Rheumatism. 29
Tabla 4-1: Datos del estudio 37
Tabla 4-2: Variables utilizadas dentro del estudio. 37
Tabla 5-1: Esquema general de la matriz de confusión para los resultados obtenidos ... 42
Tabla 5-2: Matriz de confusión y medidas de desempeño de la red neuronal tipo perceptrón utilizando dos variables genéticas para la clasificación..... 43
Tabla 5-3: Matriz de confusión y medidas de desempeño utilizando la red bayesiana con dos variables genéticas. 44
Tabla 5-4: Matriz de confusión y medidas de desempeño de la red neuronal utilizando cinco variables para clasificación. 46
Tabla 5-5: Matriz de confusión y medidas de desempeño de la red bayesiana utilizando cinco variables..... 48
Tabla 5-6: Análisis estadístico de los datos que conforman el grupo 1 51
Tabla 5-7: Análisis estadístico de los datos que conforman el grupo 2 51
Tabla 5-8: Matriz de confusión y medidas de desempeño de la red neuronal utilizando once variables para clasificación. 53
Tabla 5-9: Matriz de confusión y medidas de desempeño de la red bayesiana utilizando once variables para clasificación. 55
Tabla 5-10: Análisis estadístico de los datos que conforman el grupo 1 57
Tabla 5-11: Análisis estadístico de los datos que conforman el grupo 2 58
Tabla 5-12: Resultados del análisis de alelos HLA DRB1 en pacientes y controles. Los alelos 0101, 0102, 0105, 0401, 0403, 0404, 0405, 0408, 0410, 0428, 0433, 0440, 0442, 1001, 1344, 1402 corresponden al epítope compartido. 62
Tabla 5-13:Frecuencia de alelos en pacientes. 63
Tabla 5-14: Frecuencia de alelos en controles. 63
Tabla 5-15: Correspondencia entre alelos y secuencias de aminoácidos. Las secuencias de aminoácidos en QKRAA, QRRAA y RRRA corresponden al epítope compartido. 64
Tabla 5-16: Grupos de alelos obtenidos a partir del árbol de cercanía genética 67

Tabla 5-17: Tabla de riesgo de AR según alelos, tomada de Classification of HLA–DRB1 alleles according to the third hypervariable region of the DRβchain and their association with RA in French caucasian patients (Data adapted from du Montcel et al.) 68

Lista de abreviaturas

Abreviaturas

Abreviatura	Término
ACPA	Anticuerpos Contra Péptidos Citrulinados
AR	Artritis Reumatoide
EC	Epítope Compartido
FR	Factor Reumatoide
HLA	Human Leukocyte Antigen
HTA	Hipertensión Arterial
PCR	Proteína C Reactiva
FR	Factor Reumatoide

1. Introducción

La Artritis Reumatoide (AR) es una enfermedad autoinmune que está asociada con discapacidad progresiva, complicaciones sistémicas, muerte temprana y altos costos socioeconómicos. La causa es desconocida y su pronóstico es reservado [1].

La AR se caracteriza por la inflamación crónica de las articulaciones y afecta del 0.5 al1% de la población adulta[2]. Las principales articulaciones afectadas son pequeñas articulaciones de manos y pies [3].

El más importante factor de riesgo genético para la AR se encuentra en el HLA (Human Leukocyte Antigen) HLA Antígeno Leucocitario Humano. En particular los alelos que poseen las secuencias de aminoácidos QKRAA, QRRAA o RRRAA en las posiciones 70-74 de la cadena DRB1. Esto es conocido como Epítope Compartido (EC)[1], [4].

El curso clínico de la enfermedad difiere de un individuo a otro y los factores genéticos ligados a HLA DR han surgido como potencial factor pronóstico de esta enfermedad. El EC ha sido asociado con la predisposición y progresión de la enfermedad [1], [2].Los criterios para el diagnóstico de AR están definidos por el Colegio Americano de Reumatología y el EULAR 2010 [5].

Existe predominio en el género femenino con una relación mujer hombre de 3:1[6], la herencia de la artritis reumatoide es poligénica y no sigue un patrón mendeliano. La importancia de encontrar factores genéticos asociados con la artritis reumatoide radica en la contribución a la comprensión de los mecanismos patogénicos de la enfermedad y

su posible aplicación clínica para la identificación de marcadores, diagnóstico y pronóstico. Es posible a través de una herramienta computacional que use técnicas de Sistemas inteligentes identificar endofenotipos como son la severidad de la enfermedad a partir de la tipificación del HLA Clase II en Pacientes con Artritis Reumatoide colombianos[7].

1.1 Identificación del problema

La AR afecta aproximadamente del 0.5 % al 1% de la población humana con predominio femenino en una relación mujer hombre de 3:1[6], la herencia de la artritis reumatoide es poligénica y no sigue un patrón mendeliano. La importancia de encontrar factores genéticos asociados con la artritis reumatoide está en comprender los mecanismos patogénicos de la enfermedad y su posible aplicación clínica para la identificación de marcadores, diagnóstico y pronóstico. Se plantea entonces el problema de cómo identificar endofenotipos (relación entre fenotipo y marcadores genéticos) como son la severidad de la enfermedad o el daño óseo utilizando únicamente la información de las secuencias esto a partir de la búsqueda de genes y polimorfismos de nucleótidos. Lo anterior a partir de las secuencias de nucleótidos del HLA Clase II de pacientes con Artritis Reumatoide colombianos

1.2 Justificación

La Artritis Reumatoide (AR) es una enfermedad crónica inflamatoria autoinmune y multi sistémica, cuyo principal órgano blanco es la membrana sinovial. Es una enfermedad progresiva que incapacita al paciente desde los primeros años de su aparición, asociada con altos costos directos e indirectos, mala calidad de vida y muerte prematura[1].

Infortunadamente es frecuente el retardo en el diagnóstico de la AR, por lo que el objetivo de iniciar tempranamente una terapia óptima no siempre es posible. En las etapas tempranas usualmente el diagnóstico se basa más en una apropiada historia clínica que en exámenes diagnósticos complementarios. Los criterios del Colegio Americano de Reumatología son útiles para clasificar pacientes con poliartritis, sin embargo, es posible que en muchos casos de AR temprana no se cumplan[5].

Un paciente con AR tendrá síntomas de dolor, rigidez y enrojecimiento articular que usualmente empeora en las mañanas o después de un período de inactividad. Al examen físico se puede encontrar inflamación simétrica con dolor a la presión de las pequeñas articulaciones de las manos, los pies y de acuerdo con el grado de actividad de la enfermedad de base, un compromiso de articulaciones mayores con presencia de sinovitis.

El pronóstico de la AR depende de un conjunto de interacciones de factores genéticos, psicosociales, bioquímicos, hormonales y relacionados con el tratamiento. No todos los pacientes con AR requieren terapia intensiva. La evaluación clínica debe estar dirigida para determinar los pacientes con riesgo de enfermedad persistente. Un sistema de evaluación de severidad ha sido propuesto para clasificar a estos pacientes[8]. Es así como la sinovitis por más de seis semanas de las metacarpofalángicas y el compromiso poli articular predice la enfermedad crónica y el daño erosivo. El sexo femenino, el factor reumatoide positivo, los niveles altos de reactantes de fase aguda, han sido asociados a un pobre pronóstico. El factor de pronóstico más consistente en todas las cohortes es el factor reumatoide, el cual es importante para predecir el daño articular y la discapacidad funcional[9].

Los factores de riesgo que se han asociado con un mal pronóstico son: inicio tardío del tratamiento adecuado, severidad de la enfermedad, presentación en edades extremas, comorbilidad asociada, múltiples articulaciones inflamadas, factor reumatoide, proteína C reactiva elevada, VSG (Velocidad de Segmentación Globular) elevada, presencia de nódulos reumatoideos y erosiones tempranas[7].

Se ha encontrado que la presencia del alelo HLADR4 está asociada a la enfermedad progresiva ya que estos pacientes tienen un mayor número de articulaciones inflamadas y mayores alteraciones radiológicas[10].

El tratamiento temprano y la respuesta clínica rápida se encuentran asociados a una alta frecuencia de remisión de la enfermedad desde etapas tempranas, que persiste durante el tiempo. En los esquemas de tratamiento con terapia combinada se logra una inducción

mayor de remisión, que alcanza del 37 a 42%; estos pacientes también logran una mayor reducción en la progresión del daño radiológico[11].

Asociada al Complejo Mayor de Histocompatibilidad clase II, se asocia a la presencia del alelo HLA DRB1, con el dominio conocido como "epítope compartido" (en los aminoácidos del 69 al 74 de la tercera región hipervariable del receptor de célula T(RCT) y la AR[2].

Se ha demostrado que la presencia de este epítope es crucial, no sólo en la aparición de la enfermedad sino en la severidad y progresión de esta[2]; especialmente, existe una relación genética con la enfermedad y, la severidad de los síntomas depende de la presencia de este alelo, y es así como individuos homocigóticos tendrán una enfermedad más agresiva, con erosión articular y participación sistémica[1].

Aunque se ha comprobado que en algunas poblaciones son otras las especificidades del HLA-DR (B1-0101, B1-0114), las que se asocian con la AR, el patrón de agresividad en presencia de estos alelos se comporta de la misma forma que en el caso del epítope compartido[12].

Se ha demostrado la asociación AR y DR B1, en ellos se ha comprobado, que en la posición 71 del epítope, si el aminoácido presente es lisina, cuando se produzca la interacción celular entre estas células T, portadoras del DRB1, las células segregarán factor reumatoide (FR) y la enfermedad será más agresiva; mientras que si el aminoácido presente en la posición 71 es una arginina, el paciente no segregará FR y su enfermedad tendrá un mejor pronóstico[13].

En la Fundación Santa Fe de Bogotá, se cuenta con una serie de pacientes diagnosticados con AR y controles caracterizados según una serie de variables que permitirían clasificarlos entre pacientes y sanos. En este estudio se mostrará como las técnicas de inteligencia computacional brindan una buena herramienta para lograr este objetivo.

1.3 Objetivo general y objetivos específicos

1.3.1 Objetivo general

Desarrollar un modelo computacional para la identificación de endofenotipos en pacientes con Artritis Reumatoide utilizando el HLA clase II.

1.3.2 Objetivos específicos

Análisis estadístico de los datos de los pacientes con Artritis Reumatoide para ser utilizados por los modelos computacionales.

Exploración de modelos computacionales para la clasificación de pacientes con Artritis Reumatoide y determinación de la severidad en los mismos.

Analizar las secuencias de aminoácidos del complejo mayor de Histocompatibilidad DRB1 para la identificación de endofenotipos en pacientes con Artritis Reumatoide severa.

2. Sobre la Artritis Reumatoide

La Artritis Reumatoide (AR) es una enfermedad crónica autoinmune que afecta del 0.5 al 1% de la población en el mundo [2], [4].La AR se caracteriza por la inflamación, hiperplasia a nivel articular, así como la producción de anticuerpos (factor reumatoide FR y anticuerpos contra péptidos citrulinados Anti CCP) esto causa daño óseo y alteraciones cardiovasculares y pulmonares [2], [14].La AR es una enfermedad autoinmune que está asociada con incapacidad progresiva, complicaciones sistémicas, muerte temprana y costos socioeconómicos [1].

2.1 Aspectos Inmunológicos

La causa de la AR es desconocida y de pronóstico reservado, varios agentes infecciosos (por ejemplo, el virus de Epstain-Barr, citomegalovirus, especies de *Proteus* y *E. coli*) y sus productos (por ejemplo, proteínas de choque térmico) han sido involucrados en la patogénesis de la Artritis Reumatoide, aunque los mecanismos de unificación no se han esclarecido, se postula alguna forma de mimetismo molecular [1].

La formación de complejos inmunes durante la infección puede desencadenar la inducción de factor reumatoide, un anticuerpo de alta afinidad contra la fracción cristalizable de la inmunoglobulina IgG (FR), el cual ha servido durante mucho tiempo como marcador diagnóstico de la enfermedad [1], [15].

Recientemente varios estudios han planteado la relación existente entre la periodontitis y la AR. Ambas enfermedades se caracterizan por la inflamación crónica, daño óseo y una

respuesta inmune similar [16]. La AR parece estar asociada con la enfermedad periodontal causada por *Porphyromonas gingivalis* que expresa *PADI4* (Peptidil Arginina Deaminasa Isoforma 4) gen que codifica para enzimas responsables de la conversión de residuos de arginina a residuos de citrulina, lo cual promueve la citrulinación de proteínas [2], [16].

Porphyromonas gingivalis es la única bacteria conocida que expresa la enzima PAD. Aunque no es completamente homologa a la PAD humana, esta enzima es responsable de la post traducción y la conversión de arginina a citrulina, que lleva a la producción de anticuerpos anti-CCP. En humanos PAD existe en 5 isoformas, la isoforma PADI4 es la más asociada a autoinmunidad [17], [18].

Durante el curso de la AR las células del compartimiento sinovial presentan hiperplasia que forma el pannus sinovial que invade y destruye cartílago y hueso [19]. La sinovitis ocurre cuando los leucocitos infiltran el compartimiento sinovial. La acumulación de leucocitos refleja principalmente la migración más que la proliferación local. La migración celular es facilitada por la activación de la micro-vasculatura sinovial, la cual incrementa la expresión de moléculas de adhesión (incluyendo integrinas, selectinas y miembros de la superfamilia de las inmunoglobulinas) y quimioquinas [1].

Una característica clave del proceso de inflamación es la relacionada con las células que producen citoquinas pro-inflamatorias como TNF-α, Interleucina 1 (IL-1), interleucina 6 (IL-6) y otras. Estas interleucinas promueven la destrucción de las articulaciones[20].

2.1.1 Inmunidad Adaptativa

En la AR la presencia de anticuerpos sitúa a la inmunidad adaptativa en el centro de la patogénesis. Sin embargo, a pesar de que las células T son abundantes en el medio sinovial, el papel funcional de estas sigue siendo poco comprendido [1].

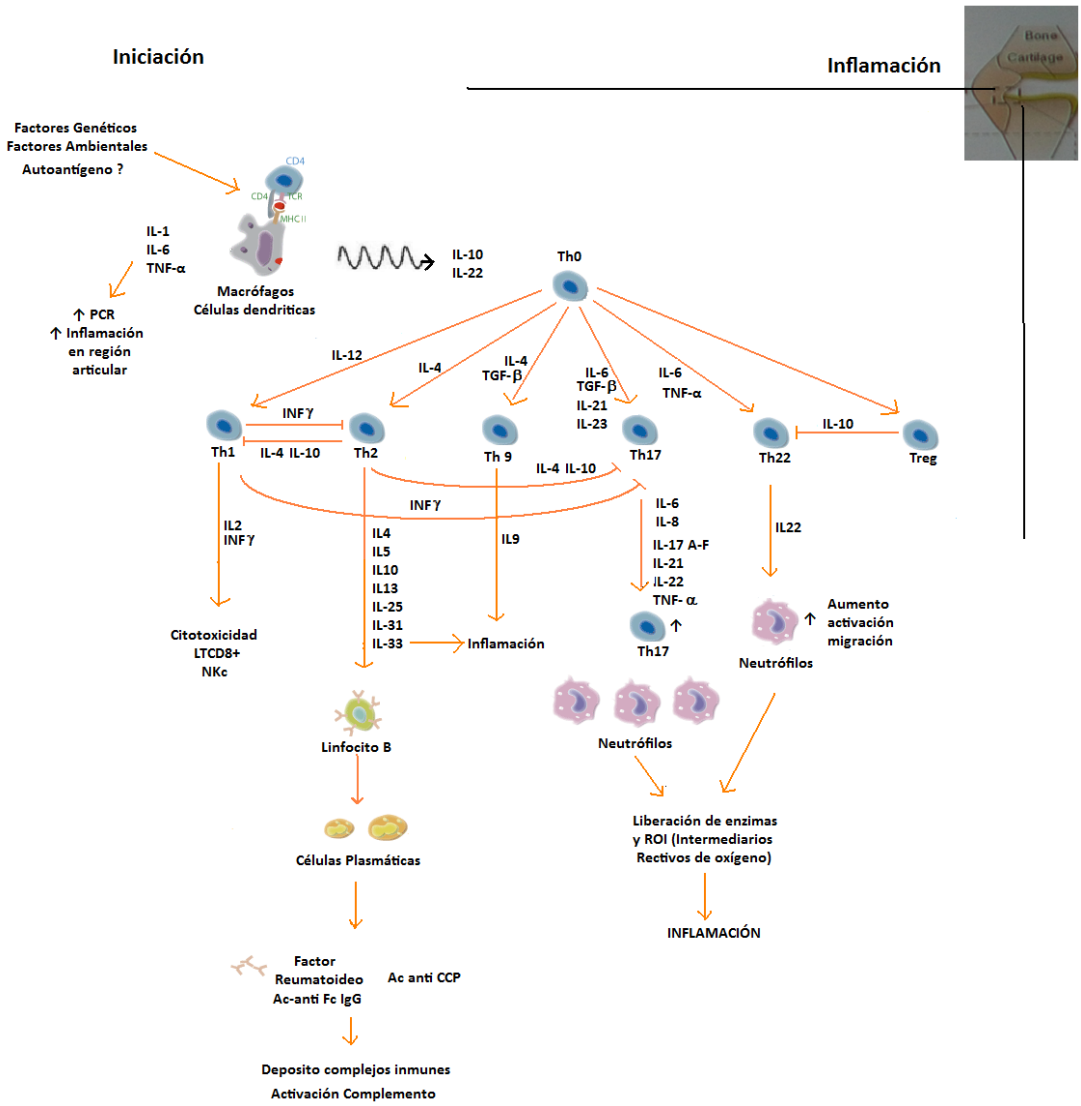
Las células Th1 median la inmunidad dependiente de células, incluyendo la citotoxicidad y respuesta de hipersensibilidad retardada a través de la producción específica de interferon gamma (IFNγ) e interleucina 2 (IL2) demostrando tener un papel importante.

Los linfocitos Th1 contribuyen a la inducción y persistencia de la inflamación que induce el daño tisular [21].

Los linfocitos Th2 se caracterizan por la producción de IL4, IL5 e IL13, que favorece la inmunidad humoral y regulan los Th1. La respuesta Th2 está asociada con IL4/IL13 mediante la producción de IgE por linfocitos B de memoria e IL5 que induce eosinofilia [21].

La Figura 2-1 muestra la interacción entre células dendríticas, células T y células B que generan la respuesta autoinmune; se observa como en la membrana sinovial se presenta el proceso inmune el cual lleva a promover el daño y remodelación del tejido.

Figura 2-1: Respuesta inmune en la Artritis Reumatoide.



La AR es convencionalmente considerada una enfermedad que es mediada por Th, actualmente la atención se ha centrado cada vez más en el papel de Th17, una subpoblación que produce interleucinaIL-17A, IL-17F, IL-21, IL-22 y factor de necrosis tumoral (TNF-α). La interleucina 17A trabaja con TNF-α para promover la activación de fibroblastos y condrocitos [1], [22].

Los Th 17 son células inflamatorias CD4+ que producen IL-17A pero no IFN-γ. Estas células y sus citoquinas secretadas se encuentran elevadas en la sangre periférica de pacientes con AR [23]. En el líquido sinovial, los niveles de Th17 demostraron estar más elevados que en la sangre periférica, sugiriendo su papel en la AR [24].

Los linfocitos Th9 son una subpoblación de linfocitos Th CD4+, estos se diferencian a partir de los linfocitos LTh0 en presencia de TGF-β e IL-4. Los Th9 expresan IL-9 esta ejerce principalmente acción sobre los mastocitos promoviendo su expansión y producción de IL-1β, IL-5, IL-6, IL-9, IL10, IL-13 and TGF-β [25], [26]. La IL-9 es también expresada por Th2, Th 17 y Treg [26].

La IL-9 actúa sobre LTreg en donde promueve funciones de supresión, en LT17 induce su proliferación y en las células presentadoras de antígenos promueve la producción de TGF-β [26].

Los linfocitos Th22 son una subpoblación de linfocitos T helper CD4+ recientemente identificados, los cuales se caracterizan por la producción de IL-22 pero no IL-17 ni IFN-γ. La respuesta pro inflamatoria de Th22 es sinérgicamente dependiente de IL-22 y TNF-α. Esto significa que las células Th22 están probablemente implicadas en la fisiopatología de algunas enfermedades autoinmunes como Lupus Eritematoso Sistémico (LES) y Esclerosis Sistémica (ES) [24].

La citoquina efectora de las células Th22 es IL-22, la cual pertenece a la familia de las IL-10 [27]. La función precisa de la IL-22 no es clara. Las células Th22 juegan un papel importante en la inflamación y en las enfermedades autoinmunes [24], [28].

Las células T reguladores CD4+CD25+ (Treg) juegan un papel importante en la prevención de la autoinmunidad. Las células T CD4+ pueden diferenciarse en distintos fenotipos que secretan citoquinas y promueven el desarrollo de distintas respuestas inmunes (inmunidad mediada por células e inmunidad humoral). Las células Treg incluyen las células T CD4+CD25+ Foxp3+ juegan un papel activo en la prevención y desarrollo de la autoinmunidad, muchos estudios han buscado determinar si la deficiencia en la actividad de Treg puede contribuir al desarrollo de enfermedades autoinmunes como la AR [29].

La respuesta inmune en la AR inicia con la inflamación de la articulación a nivel de la membrana sinovial; esta es una estructura acelular con un recubrimiento formado por

macrófagos y fibroblastos conocidos como sinoviocitos. En la AR, esta membrana se vuelve hiperplásica debido a la entrada y proliferación de éstos últimos. Las células T se acumulan en la membrana sinovial [30].

Los macrófagos y los fibroblastos producen IL-1, TNF-α e IL-8 y muchas otras citoquinas que contribuyen a la inflamación de la membrana sinovial así como a la activación de las células adyacentes que incrementan la inflamación sinovial. Las células B y las células dendriticas forman agregados con las células T y macrófagos tisulares en aproximadamente 20% de los pacientes. Hay producción local de auto anticuerpos que resultan en la formación de inmunocomplejos los cuales se depositan en las articulaciones. Los anticuerpos reconocen antígenos articulares, tales como colágeno tipo II y proteoglicanos, o se unen a la Fc de la IgG normal, como el factor reumatoide [30].

Los linfocitos B sintetizan FR que es un anticuerpo de la clase IgM, IgG o IgA contra la porción Fc de la inmunoglobulina G (IgG), este funciona como un auto anticuerpo contra la IgG formando complejos inmunitarios, que luego se depositan en la articulación y activan la cascada de complemento e inducen los mecanismos de fagocitosis mediada por células de la línea monocito-macrófago y neutrofilos. Esto es importante en la perpetuación de la respuesta inmunopatológica secundaria en la AR [1], [16].

Los auto anticuerpos más frecuentemente encontrados en paciente con AR son anticuerpos que se unen al dominio constante de la IgG e IgM y Anti CCP [31]. El papel de las células B en la patogénesis de la AR va mas allá de la producción de auto anticuerpos, incluye la presentación de auto antígenos y producción de citoquinas pro inflamatorias como IL-6 y TNF-α [1]
La citrulinación consiste en la modificación del aminoácido arginina por citrulina y además realiza la precipitación de complejos inmunes que contienen anti CCP, con potencial pro inflamatorio [32]. La citrulinación consiste en el cambio pos traduccional producido por la PAD sobre residuos de arginina. Esta modificación post-traduccional tiene el potencial de alterar la estructura, antigenicidad y la función de las proteínas [16].

2.1.2 Inmunidad Innata

Una variedad de células efectoras innatas, incluyendo macrófagos, mastocitos y células asesinas naturales, se encuentran en la membrana sinovial, que residen principalmente en el líquido sinovial [1].

Los macrófagos actúan a través de la liberación de TNF-α e IL-1, IL-6, IL-12, IL-15, IL-18 e IL- 23, reactivos intermediarios de oxígeno, nitrógeno, además producen prostanoides y enzimas degradantes de matriz (Matrix metalloproteinases MMPs), realizan fagocitosis y mecanismos de presentación de antígenos[1].

Los macrófagos son activados por receptores tipo Toll (Toll Like Receptors TLRs), por ejemplo, TLR 2/6, 3, 4, y 8, además dominios de oligomerización unidos a nucleótidos receptores tipo (NOD Like Receptors NLRs) que reconocen un rango de patrones moleculares asociados a patógenos (Pathogen Associated Molecular Patterns PAMPs) y patrones moleculares asociados a daño (Damage Associated Molecular Patterns DAMPs) que incluyen bacterias, virus y ligandos endógenos putativos [1].

Las citoquinas de la familia de IL-1 (por ejemplo, IL-1α, IL-1β, IL-18, e IL-33) son ampliamente expresadas en la AR. Ellas promueven la activación de leucocitos, células endoteliales, condrocitos y osteoclastos [1]

Los neutrófilos contribuyen a la sinovitis porque sintetizan prostaglandinas y reactivos intermediarios de oxígeno. La activación de la respuesta inmune innata contribuye a la sinovitis [1]. El aumento en la respuesta inflamatoria, la generación de anticuerpos contra antígenos propios modificados son los factores que contribuyen a la presencia de la AR.

2.2 Factores Genéticos

Las asociaciones genéticas de la AR incluyen el HLA-DR4 y DRB1 [33]. La AR implica una compleja interacción entre factores genéticos y ambientales. La asociación entre la AR y el HLA-DRB1 ha sido confirmado en pacientes que tienen Anti CCP positivo, los pacientes con alelos que contienen un motivo común de aminoácidos con secuencias

QKRAA, QRRAA o RRRAA en el HLA-DRB1, en las posiciones 70 a 74 de la cadena del DRB1 llamado Epítope Compartido (EC) presentan una susceptibilidad particular [1], [2].

Los alelos más conocidos que codifican para el EC incluyen miembros del grupo de alelos HLA-DRB1*04 (por ejemplo *0401, *0404, *0405 y *0408), HLA-DR *0101 o *0102, HLA-DRB1 *1402 y HLA-DRB1 *1001 [34].

Se han identificado más de 20 factores de riesgo genéticos en los últimos 4 años, la mayoría de estos factores han sido localizados cerca a genes implicados en rutas inmunológicas. Estos hallazgos confirman el papel del sistema inmune en la patogénesis de la AR [2].

Otros alelos de riesgo en AR con Anti CCP positivo están implicados en la regulación inmune, incluyendo factor nuclear κB (NF- κB) dependiente de señalización (por ejemplo TRAF1-C5 y c-REL) y estimulación de células T, activación y diferenciación funcional (por ejemplo PTPN22 y CTLA4). Por otra parte, las interacciones gen a gen incrementan el riesgo de la enfermedad, como el descrito entre HLA-DRB1 y PTPN22 [1].

En la población Japonesa se ha identificado el gen PADI4 como el segundo factor de riesgo de la AR [2]. En Colombia, el estudio que examinó la relación entre HLA y la AR, mostró que la frecuencia tanto del HLA-DR4 como del DR1 fue del 23% en 69 pacientes, sin embargo, este estudio no incluyo población control. En un pequeño grupo de pacientes afro colombianos de Quibdó, pero que incluyó a toda la población afectada, no se observó asociación de alelos del HLA-DRB1 con la AR. También se reportó en una población antioqueña la asociación del SE QRRAA en el HLA-DRB1*04, en particular el DRB1*0404, con la enfermedad. Este alelo corresponde al factor de riesgo más importante y uniforme para desarrollar artritis reumatoide en latinoamericanos [4].
Algunos alelos HLA-DRB1 están descritos como factores de riesgo para RA Anti CCP positivo. Para RA Anti CCP negativo la situación es claramente diferente, así HLA-DR3 predispone a AR Anti CCP negativo. Recientemente se ha sugerido que polimorfismo en el gen del receptor del neuropéptido S puede estar implicado en AR Anti CCP negativo [2]. Factores genéticos de protección contra la AR están predominantemente asociados con HLA-DRB1*13:01. Estos alelos son más frecuentes en controles sanos comparados

con pacientes con AR. La protección contra la AR está conferida por la secuencia DERAA en las posiciones 70-74 del alelo HLA-DRB1, la cual se encuentra en la misma posición de los alelos HLA-DRB1 del EC. En varios estudios también se propone una asociación principalmente con el ácido aspártico (D) en la posición 70 (D70) y para AR Anti CCP negativa no existen alelos HLA-DRB1 asociados. Por otro lado, el análisis de genotipos específicos del EC revelan que genotipos HLA-DRB1*0101/*0401/*0404 fueron los principales predictores de mortalidad por infarto cardiaco [35].

2.3 Factores de Riesgo

El riesgo de desarrollar AR es atribuible a factores genéticos en un 50% [31]. Además, muchos factores han sido asociados con el incremento del riesgo de desarrollar AR. Pero el tabaquismo es el único factor de riesgo ambiental que ha sido extensamente estudiado y ampliamente aceptado. Especialmente en individuos que poseen una o dos copias del EC HLA-DRB1[2], [36].

Investigaciones epidemiológicas han demostrado una fuerte asociación entre el consumo de cigarrillo y el riesgo de AR, especialmente en personas que tienen predisposición genética [33], [37].

La asociación de RA Anti CCP positivo y tabaquismo puede ser explicada por la observación de proteínas citrulinadas en los pulmones de los fumadores [2]. La edad avanzada, historia familiar de la enfermedad, género femenino también están asociados con un riesgo mayor de presentar la enfermedad [33].

La alta prevalencia de la AR en las mujeres ha permitido identificar el posible papel hormonal en la susceptibilidad de la enfermedad, aunque no se han encontrado diferencias en los niveles de hormonas en las mujeres con AR y controles sanos [36].

Varios agentes infecciosos han sido estudiados e implicados en el desarrollo de la AR. En los últimos años la bacteria *Porphyromonas gingivalis* ha sido implicado como un factor para el desarrollo de AR*P. gingivalis* es el principal agente causante de periodontitis, una enfermedad que es más frecuente en pacientes con AR que en la

población sana[18]. Esta bacteria expresa la enzima peptidil arginina-deiminasa (PAD) responsable de citrulinización de proteínas y produce inflamación crónica, caracterizado por la presencia de citoquinas pro inflamatorias y TNF [36], [38].

La exposición a sílice es un factor de riesgo bien definido para AR. El sílice está presente en la minería, construcción, cerámica y vidrio así como en la agricultura [36].

2.4 Aspectos Clínicos

2.4.1 Diagnóstico

La AR es una enfermedad autoinmune caracterizada por inflamación y daño a nivel de articulaciones que conlleva a discapacidad y mortalidad prematura [5].

Los pacientes con AR presentan dolor y rigidez a nivel de articulaciones de muñecas, articulaciones interfalángicas y metacarpofalángicas son las principalmente involucradas. La rigidez matinal de articulaciones que duren más de 1 hora sugiere un origen inflamatorio [33].

El diagnóstico de la AR es a menudo un desafío, esto debido al amplio espectro de manifestaciones clínicas y los cambios progresivos de la enfermedad con el tiempo, y, quizás, la más importante, la falta de una prueba de oro clínica o de laboratorio para definir la presencia o ausencia de enfermedad [39][40].

El Factor Reumatoide (FR) no es específico para la AR y puede estar presente en pacientes con otras enfermedades, como la hepatitis C, y en personas mayores sanas. El Anticuerpo anti-péptido citrulinado es más específico para la AR [41]
Los niveles de PCR (Proteína C Reactiva) y Velocidad de Sedimentación Globular (VSG) se aumentan frecuentemente en la AR activa, además estos reactantes de fase aguda son parte de los nuevos criterios de clasificación de AR [33].

En el año 2010 el Colegio Americano de Reumatología y la Liga Europea contra el Reumatismo colaboraron para la creación de los nuevos criterios de clasificación de la AR, los cuales se presentan en la Tabla 2-1[5][5][5][5].

Tabla 2-1: Criterios de clasificación 2010 American College of Rheumatology European League Against Rheumatism.

Población objetivo: ¿A quiénes se les debe hacer la prueba? Pacientes: **1. Que tengan al menos 1 articulación con sinovitis clínica definida.** **2. En los que la sinovitis no puede ser explicada por otras enfermedades.** **Un puntaje mayor de 6 sobre 10 es necesario para clasificar a un paciente con AR**	
A. Articulaciones comprometidas:	
1 articulación grande	0
2 a 10 articulaciones grandes	1
1–3 articulaciones pequeñas (con o sin compromiso de articulaciones grandes)	2
4–10 articulaciones pequeñas (con o sin compromiso de articulaciones grandes)	3
>10 articulaciones (por lo menos una articulación pequeña)	5
B. Serología (al menos un resultado de la prueba es necesario para clasificación)	
Factor Reumatoide negativo y anti CCP negativo	0
Factor Reumatoide positivo bajo y anti CCP positivo bajo	2
Factor Reumatoide positivo alto y anti CCP positivo alto	3
C. Reactantes de fase aguda (al menos un resultado es necesario para clasificación)	
PCR normal y VSG normal	0
PCR anormal o VSG anormal	1
D. Duración de los síntomas	
< 6 semanas	0
>= 6 semanas	1

Las radiografías de manos y pies se deben realizar para evaluar el daño óseo principalmente los cambios erosivos, lo cual puede evidenciar una enfermedad más

agresiva [33]. Dentro del diagnóstico diferencial están: el LES (Lupus Eritematoso Sistémico), la Esclerosis Sistémica y la Artritis Psoriásica [33].

La AR temprana es aquella en la cual la duración de la enfermedad en menor a 6 meses [42]. Aunque la AR es considerada como una enfermedad que afecta a las articulaciones, esta es también una enfermedad sistémica capaz de afectar múltiples órganos y producir alteraciones cardiacas y pulmonares [33].

2.4.2 Tratamiento

El objetivo del tratamiento de AR es lograr la remisión de la enfermedad o disminuir la actividad de la enfermedad [42][43]. La intervención terapéutica temprana ha mejorado los resultados en el tratamiento reduciendo el daño en las articulaciones y disminuyendo la discapacidad [5][44].

El tratamiento temprano de los pacientes con AR trae beneficios durante el curso de la enfermedad. El manejo actual de la AR busca la remisión de la enfermedad para los pacientes [45]. El lograr la remisión de la enfermedad mejora la calidad de vida en el manejo de la AR temprana [42]. La PCR y VSG se usan para seguimiento de la actividad de la enfermedad y la respuesta a la medicación [11].

El objetivo del tratamiento incluye disminuir el dolor y la inflamación, prevenir la deformidad (por ejemplo, desviación cubital) y el daño óseo (tales como erosiones), mejorar la calidad de vida y prevenir las manifestaciones extra articulares [5], [33].

Existe evidencia que demuestra el beneficio del tratamiento agresivo en pacientes con AR temprana lo que soporta el cambio en las estrategias de tratamiento para lograr la remisión de la enfermedad [42][46].La literatura describe una “ventana de oportunidad” para prevenir el daño permanente en paciente con artritis reumatoide temprana [43]. La evidencia sugiere que la remisión es más probable en pacientes con AR temprana que en pacientes con AR antigua [42].

2.4.3 Drogas Modificadoras de la AR (DMARDs)

Las drogas modificadoras de la AR (DMARDs) pueden ser biológicas y no biológicas [47]. El metotrexate es recomendado como la primera línea de tratamiento en pacientes con AR activa [33].Una buena respuesta, que incluye la remisión puede ser lograda en algunos pacientes tratados con DMARDs [42].

En la última década, el uso de DMARDs, en particular el Metotrexate , ha mejorado considerablemente el éxito en el manejo de la AR [5]. El metotrexate está contraindicado en pacientes con enfermedad hepática, como la hepatitis C y en pacientes con insuficiencia renal [33].

La sulfasalazina o hidroxicloroquina puede ser utilizada como monoterapia en pacientes con una baja actividad de la enfermedad [33]. La terapia combinada con 2 o más DMARDs es más efectiva que la monoterapia, sin embargo, los efectos adversos pueden ser mayores [33].

2.4.4 Corticoesteroides

Dentro del tratamiento para la AR están los corticoides los cuales pueden ser administrados por vía oral, intramuscular o intraarticular para manejo del dolor y la inflamación [33].

Los glucocorticoides tales como la prednisona son utilizados con las DMARDs en el tratamiento de la AR para lograr la remisión y detener la progresión de la enfermedad. Estos son utilizados para el control de la inflamación [42].

2.4.5 Agentes Biológicos

Los agentes biológicos, lo más reciente en tratamientos para la AR, han logrado la remisión de la enfermedad en un porcentaje significativo de los pacientes [42].

El primer agente biológico fue el antagonista contra el Factor de Necrosis Tumoral (TNF); a este grupo de agentes biológicos también pertenecen etanercept, infliximab, adalimumab, golimumab y certolizumab [42].
Los agentes biológicos incluyen los anticuerpos monoclonales y los receptores recombinantes para bloquear las citoquinas que promueven la cascada de inflamación responsable de los síntomas de la AR. Si la AR no está bien controlada con un DMARD no biológico, un DMARD biológico puede ser usado, los inhibidores TNF con la primera línea de tratamiento son los más estudiados [33].

Los DMARDs biológicos son recomendados únicamente en pacientes quienes presentan una enfermedad altamente activa durante 3 a 6 meses o una actividad elevada durante menos de 3 meses y características de mal pronóstico [42].

Otros agentes biológicos aprovechan para el tratamiento de la AR la inhibición de diferentes rutas de inflamación dentro los cuales están: abatacept (inhibe la activación de linfocitos T), rituximab (causa depleción de linfocitos B) y tocilizumab (inhibe el receptor de interleucina 6) [42].

3. Aplicación de métodos informáticos en la clasificación de enfermedades

El análisis de asociación genética se basa en correlaciones estadísticas causa a efecto entre dos variables [48]. Normalmente, dicha asociación causa y efecto no es necesaria en el análisis genético ya que los genes no son siempre la causa y los fenotipos son siempre el efecto. Sin embargo, entre los fenotipos y biomarcadores, la asignación de la causa y el efecto cobra sentido y la inferencia causal puede ser útil, es así como la relación causal entre dos biomarcadores anti CCP y factor reumatoide ha sido establecida en la AR [49].

Genome Wide Association Studies (GWAS) debe su popularidad a la expectativa de lograr un gran impacto en el diagnóstico, pronóstico y tratamiento de la enfermedad por el descubrimiento de la genética subyacente a fenotipos clínicos [12]. El análisis de datos GWAS hasta el momento consta de una amplia confianza en los métodos que hacen hincapié en la contribución de SNPs a la enfermedad en particular su asociación estadística con fenotipos. Métodos multi-variables, sin embargo, pueden extraer más información considerando las asociaciones de múltiples SNPs simultáneamente [40][48].

En el estudio [40] a partir de los datos de GWAS encuentran seis SNPs, la mayoría procedentes del locus MHC. Utilizando estos SNPs desarrollaron dos modelos predictivos que pueden clasificar los casos y controles con una precisión de 0,81, que se verificaron con los datos de prueba independientes de la misma cohorte.

3.1 Redes Neuronales

Las redes neuronales artificiales son sistemas basados en computación paralela masiva que se utilizan para realiza tareas de control y de clasificación, entre otros. Una red neuronal biológica puede ser modelada matemáticamente por un grafo dirigido con nodos (neuronas) interconectadas. Los nodos artificiales calculan funciones simples cuyos argumentos son las sumas de las entradas al nodo; algunos trabajos sobre redes neuronales utilizan funciones de nodo que utiliza sólo valores binarios[50].

3.2 Redes Bayesianas

El enfoque bayesiano proporciona una forma consistente para hacer inferencia mediante la integración de la evidencia de los datos con conocimientos previos del problema. El enfoque bayesiano ofrece herramientas eficaces para evitar el sobre ajuste incluso con modelos muy complejos y además facilita la estimación de los intervalos de confianza de los resultados. Una de las ventajas de las redes bayesianas puede explicar el modelo que se genera caso contrario a las redes neuronales [51].

3.3 Métodos de Kernel

Los métodos de aprendizaje que emplean kernels son métodos de aprendizaje y de estimación de un núcleo de las funciones definidas en el dominio de datos. Trabajar en estos espacios tiene la ventaja de facilitar la construcción y análisis de algoritmos de aprendizaje. Los métodos de kernel permiten transformar datos de cualquier naturaleza, mediante una función kernel, a un espacio que está dotado de producto interno [52]

3.4 WEKA

Weka es un software de código abierto. El sistema fue desarrollado en la Universidad de Waikato en Nueva Zelanda. Weka representa un entorno para análisis de datos. El software está disponible gratuitamente en la página de internet de Weka, está escrito en Java, un lenguaje orientado a objetos. Weka proporciona implementaciones de los más avanzados algoritmos de minería de datos y algoritmos de aprendizaje de máquina.

Además contiene módulos de pre procesamiento de datos, clasificación, agrupamiento y extracción de reglas de asociación [53].

3.5 KNIME

KNIME es un entorno de trabajo fácil de usar para todo el proceso de análisis, acceso y transformación de datos, además permite análisis predictivo, visualización y presentación de informes. La plataforma ofrece más de 1000 módulos para procesamiento y análisis de datos. [54].

3.6 JALVIEW

Jalview es un visor de alineamiento de secuencias, editor y herramienta para análisis múltiple. Jalview está diseñado para ser independiente de la plataforma (se puede ejecutar en Mac, MS Windows, Linux y cualquier otra plataforma que soporte Java), capaz de editar y analizar grandes alineaciones (miles de secuencias) con un mínimo de requerimientos, y capaz de mostrar múltiples vistas integradas de la alineación y otros datos. Jalview permite leer y escribir muchos formatos de secuencia comunes incluyendo FASTA, Clustal, MSF (GCG) y PIR [55].

4. Metodología para la clasificación de pacientes con artritis reumatoide

En este capítulo se hace una descripción general de la metodología propuesta para realizar el análisis de los datos de pacientes con artritis reumatoide y controles.

4.1 Datos del estudio

Para la selección del conjunto de datos a ser utilizado por los modelos computacionales se utilizó información de pacientes con AR y controles sanos, con previo consentimiento informado por parte de los pacientes de la Fundación Santa Fe de Bogotá. Al inicio del estudio se contaba con 135 pacientes y 140 controles sanos, para un total de 275 datos. Al realizar el pre-procesamiento de los datos, varios de los registros fueron eliminados por presentar información faltante o inconsistente. Finalmente se logró tener una muestra completa con 138 datos que correspondían a 63 pacientes y 75 controles.

Los datos utilizados en este estudio corresponden a información clínica, serológica y genética de 138 personas, los cuales corresponden a pacientes con AR y controles. Su distribución se puede observar en la Tabla 4-1 y la descripción de las variables se presenta en la Tabla 4-2.

Tabla 4-1: Datos del estudio

Número total de Pacientes	Número de controles sanos	Número total
63	75	138

Las variables de los datos de la muestra se presentan en la Tabla 4-2.

Tabla 4-2: Variables utilizadas dentro del estudio.

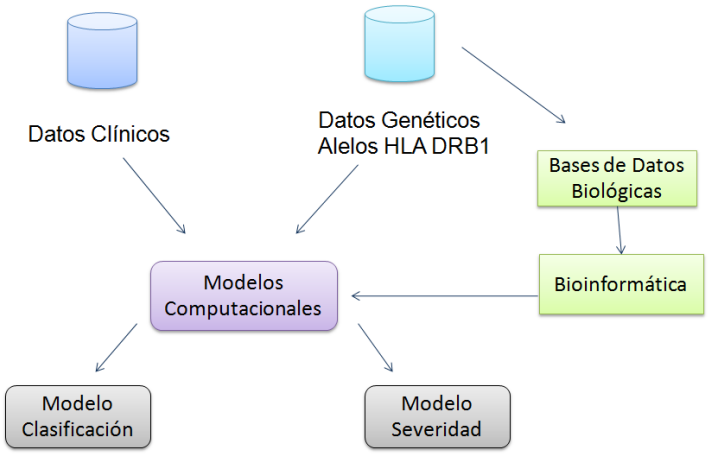
Variables	Representación	Tipo Información
Edad	Representa la edad de cada persona expresada con un número entero.	Clínica
Género	El género se representa de la siguiente manera hombre con 1 y mujer con 0.	Clínica
Tabaquismo	El consumo de cigarrillo se representa se la siguiente manera 0 = nunca, 1 = 1 a 5 paquetes/año, 2 = 6 a 10 paquetes año, 3 = más de 10 paquetes año	Clínica
Antecedentes hipertensión	NO=0, SI=1.	Clínica
Antecedente dislipidemia	NO=0, SI=1	Clínica
Número de articulaciones inflamadas	Representa el número de articulaciones comprometidas, un número entero.	Clínica
Número de articulaciones dolorosas	Representa el número de articulaciones dolorosas, un número entero	Clínica
Rigidez matinal	Rigidez matinal mayor a una hora=1, menor de 1 hora=2	Clínica
PCR	Proteína C Reactiva, un número real.	Serológica
FR	Factor Reumatoide, un número real	Serológica
AntiCCP	Anticuerpo contra péptido citrulinado, un número real	Serológica
HLADRB1uno	Representa el alelo uno del individuo. Cadena de caracteres	Genética
HLADRB1dos	Representa el alelo dos del individuo. Cadena de caracteres	Genética

La muestra seleccionada tiene información de pacientes con artritis reumatoide y controles quienes están tipificados en el HLA DRB1 para cada uno de sus alelos; además se cuenta con información clínica y serológica de estos pacientes. En la Tabla 4-3 se presenta el análisis estadístico de las variables.

Variables	Promedio	Desviación Estándar	Rango
Edad	64.26	16.3	20 - 91
HTA	NO = 110 instancias SI = 18 instancias		
Dislipidemia	NO = 119 SI = 19		
PCR	1.79	2.02	0 – 11.6
FR	31.33	44.19	0 - 349
Anti CCP	55.6	71.8	0.1 – 257
Articulaciones inflamadas	10.32	12.27	0 – 28
Articulaciones Dolorosas	10.61	12.41	0 – 28
Alelo 1	0403 = 25 más frecuente 1401 = 1 menos frecuente		
Alelo 2	1501 = 17 más frecuente 0401 = 1 menos frecuente		

La metodología general utilizada en el trabajo se representa por el esquema de la Figura4-1.

Figura 4-1: Esquema de la metodología.



Primero se realiza la búsqueda y selección de la información clínica y genética principalmente el HLA DRB1 de los pacientes y controles. Con toda la información recolectada se realiza un pre procesamiento de los datos, principalmente se eliminan registros los cuales presenta información inconsistente o información nula. También se realiza el análisis estadístico de dicha información. Finalmente, se selecciona una muestra de 138 individuos que corresponde a 63 pacientes y 75 controles. De los cuales se selecciona el 80% de los datos para entrenamiento y 20% de los datos para prueba de manera estratificada.

Este proceso se inicia con la búsqueda de las secuencias de aminoácidos para cada uno de los alelos correspondientes a los datos del estudio. Esto se realizó mediante la búsqueda de los aminoácidos correspondientes a cada alelo en bases de datos biológicas que contienen esta información. Se usó principalmente la base de datos IMGT/HLA (*ImMunoGeneTics Project*). Posteriormente, se realiza un alineamiento entre las secuencias obtenidas., lo cual se realizó utilizando una matriz tipo Blosum.

Una vez seleccionados los datos clínicos y genéticos de los pacientes y controles, se realiza el modelado e implementación de los modelos computaciones basados en técnicas de inteligencia computacional como redes neuronales tipo perceptrón, perceptrón multicapa, redes bayesianas y métodos de agrupamiento como.

ño de los modelos se utilizaron las y Rapidminer se usaronpara analizar.; con Jalview

Finalmente para realizar el proceso de clasificación, se diseñaron una red neuronal tipo perceptrón multicapa con una capa oculta, se optimizan los parámetros para la selección del mejor modelo en el caso de la red neuronal el parámetro a optimizar es la tasa de aprendizaje, también se diseño una red bayesiana la cual es de tipo Naive Bayes.

Para la agrupación de los datos entre pacientes y controles se utilizó *k-means* y como criterio de selección del número de grupos se utilizó el coeficiente de silueta. Se trabajó en los modelos con distinto número de variables, diseñándose primero modelos computaciones que únicamente utilizan las variables genéticas correspondientes a los alelos HLA DRB1, después se decide aumentar a los modelos otras variables de tipo clínico y serológico.

5. Resultados

5.1 Medidas de desempeño

Para determinar el desempeño de la clasificación, se utilizan algunas medidas. Aquí se reportarán la sensibilidad, especificidad, precisión y exactitud, que se definen a continuación. Estas medidas se calculan con base en los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP), falsos negativos (FN), número de muestras positivas (P) y número de muestras negativas (N)[56]. Las ecuaciones correspondientes se presentan a continuación:

$$Sensibilidad = \frac{VP}{P} \quad ; \; Especificidad = \frac{VN}{N}$$

$$Precisión = \frac{VP}{VP + FP} \quad ; \; Exactitud = \frac{VP + VN}{P + N}$$

La matriz de confusión que se utilizará para presentar los resultados de las medidas de desempeño de cada modelo propuesto será la siguiente presentada en la Tabla 5-1:

Tabla 5-1: Esquema general de la matriz de confusión para los resultados obtenidos

	AR Real	**NO AR Real**	
Clase AR predicha por el modelo	VP	FP	PRECISION
Clase NO AR predicha por el modelo	FN	VN	
	SENSIBILIDAD	ESPECIFICIDAD	EXACTITUD

5.2 Modelos Computaciones para clasificación de AR y riesgo de severidad.

El primer escenario que se plantea en el trabajo es usar varios clasificadores utilizando únicamente dos variables, las cuales corresponden a las variables genéticas, es decir, los alelos del HLA DRB1.

El primer modelo computacional que se diseñó fue una red neuronal tipo perceptrón multicapa con una capa oculta, se usó una tasa de entrenamiento de 0.7. Este perceptrón recibe como entrada los valores de los dos alelos HLA DRB1,los cuales corresponden a datos nominales, que representan alelos como 0101, 0301, 0404 etc. A las entradas se les realizó previamente un proceso de binarización. La salida de la red neuronal, como clasificador, determina dos clases: paciente o control.

La red neuronal tipo perceptrón utilizando únicamente dos variables genéticas, lo que significada que se utilizaron 2 variables de entrada y como parámetros una tasa de aprendizaje de 0.7, la cual fue optimizada y una validación tipo Split, a partir de esto se obtuvo un porcentaje de clasificación de 85.71% que corresponde a 24 instancias correctamente clasificadas de los 28 datos de prueba que corresponde al 20% del total

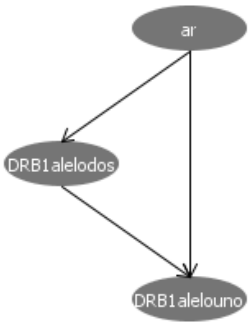
de los datos. En la Tabla 5-2 se presenta la matriz de confusión y los valores de las medidas de desempeño del clasificador.

Tabla 5-2: Matriz de confusión y medidas de desempeño de la red neuronal tipo perceptrón utilizando dos variables genéticas para la clasificación.

	AR Real	**NO AR Real**	
Clase AR predicha por RN Perceptrón	10	1	90.91%
Clase NO AR predicha por RN Perceptrón	3	14	
	76.92%	93.33%	85.71%

El segundo modelo que se implementó fue una red bayesiana usando dos variables genéticas. La red bayesiana fue desarrollada utilizando un estimador simple y un algoritmo de búsqueda tipo TAN (*Tree Augmented Naive Bayes*).La red bayesiana que se obtuvo se puede observar en la Figura5-1.

Figura 5-1: Red bayesiana para clasificación utilizando dos variables genéticas.



Con la red bayesiana se obtuvieron los siguientes resultados: se clasificaron adecuadamente 19 instancias, que corresponde a un 67.85%.La Tabla 5-3 presenta la

matriz de confusión y las medidas de desempeño de la red bayesiana. La relación entre el alelo 1 y el alelo 2 representa la probabilidad de que los alelos sean iguales lo cual sería un homocigoto o por el contrario un heterocigoto, además representa la probabilidad que los dos alelos hagan parte de la teoría del epítope compartido los cuales son más frecuentes en pacientes con AR.

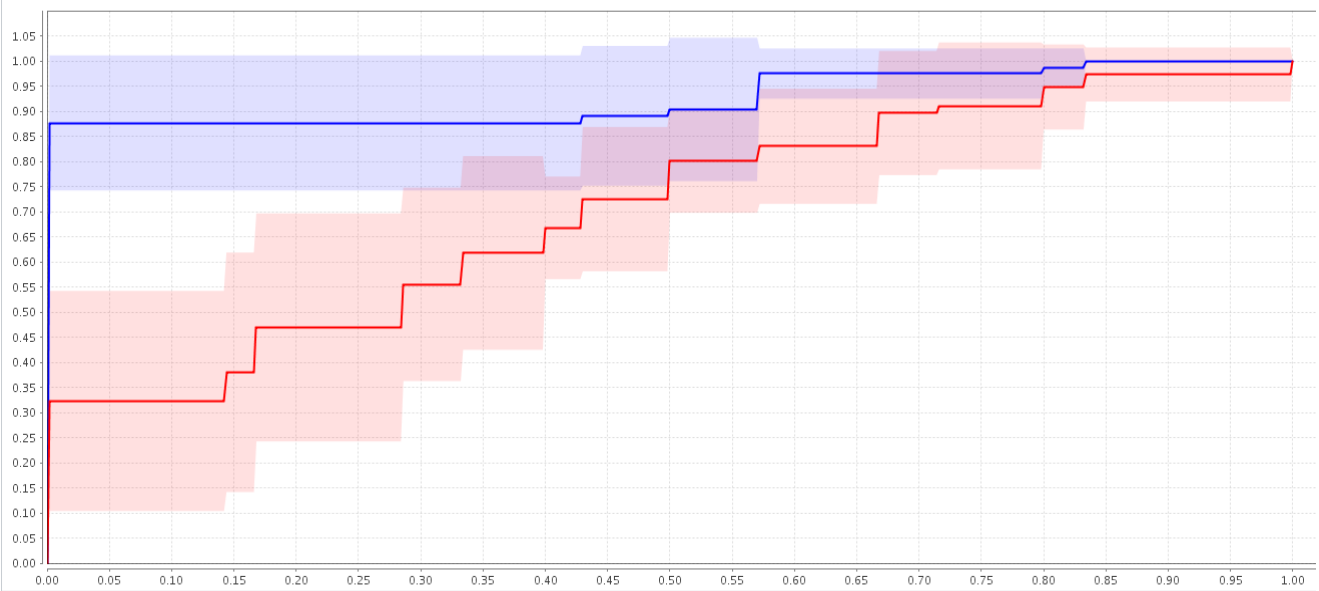
Tabla 5-3: Matriz de confusión y medidas de desempeño utilizando la red bayesiana con dos variables genéticas.

	AR Real	**NO AR Real**	
Clase AR predicha por la Red Bayesiana	7	7	50%
Clase NO AR predicha por la Red Bayesiana	2	12	
	77.77%	63.1%	67.85%

En la Figura 5-2 se muestran las curvas ROC para los dos clasificadores red neuronal y red bayesiana utilizando dos variables. La curva de color azul corresponde a la red neuronal tipo perceptrón multicapa y la curva roja corresponde a la red bayesiana.

El área bajo la curva permite comparar diferentes tipos de clasificadores, específicamente, el mejor clasificador se determina teniendo en cuenta la mayor área bajo la curva ROC y la forma de dicha curva. La mejor curva es la que está más cerca a los valores deseados de falsos positivos y de verdaderos positivos. De acuerdo con lo anterior se observa en la gráfica que el clasificador neuronal representado en color azul tiene un mejor desempeño con respecto al clasificador neuronal representado en color rojo.

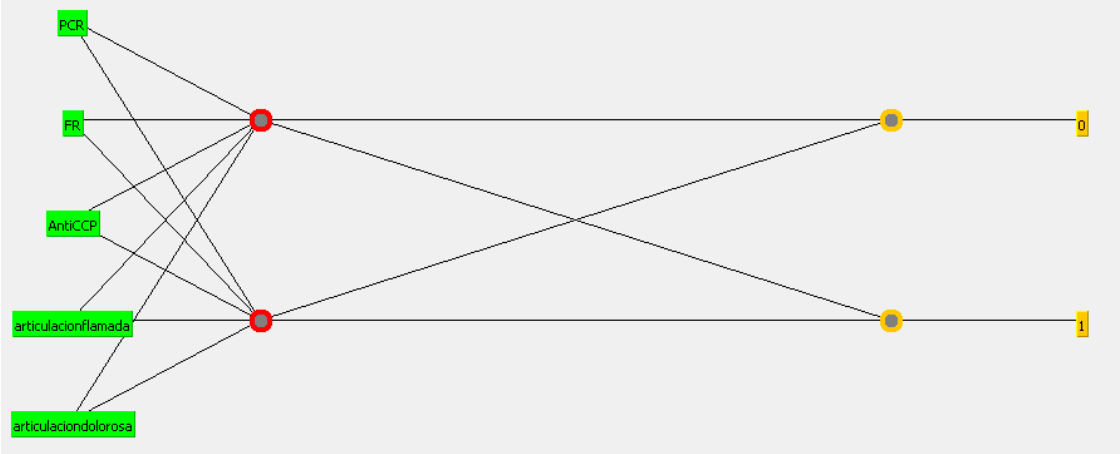
Figura 5-2: Curvas ROC de los clasificadores la curva roja representa la red bayesiana y la curva azul representa la red neuronal



El segundo escenario que se plantea es utilizar los criterios del Colegio Americano de Reumatología para clasificación de la artritis reumatoide, para lo cual se usaron las cinco variables implicadas en estos criterios: articulaciones inflamadas, articulaciones dolorosas, factor reumatoide, proteína C reactiva y anticuerpos contra péptido citrulinado, entre otras.

Como pre procesamiento de los datos se normalizan todas las variables, se utiliza una red neuronal tipo perceptrón multicapa con 1 capa oculta que contiene 7 neuronas, una tasa de aprendizaje de 0.2,con un aprendizaje supervisado mediante corrección del error, se optimizaron parámetros, se utilizó el 80% de los datos para entrenamiento y 20% de los datos para prueba. En la Figura 5-3 se puede observar la red neuronal que se obtuvo.

Figura 5-3: Red neuronal perceptrón multicapa utilizando cinco variables.



La matriz de confusión y las medidas de desempeño utilizando la red neuronal con 5 variables se presentan en la Tabla 5-4. La cual muestra que el modelo presenta una exactitud de 89.28%, una precisión de 85.71%, con una sensibilidad de 92.3% y una especificidad de 86.66%.

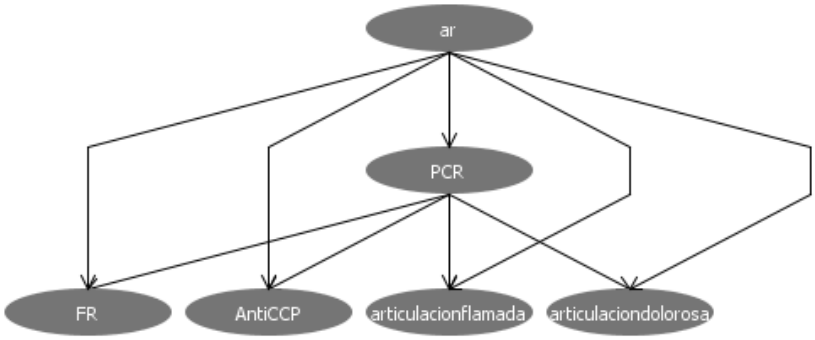
Tabla 5-4: Matriz de confusión y medidas de desempeño de la red neuronal utilizando cinco variables para clasificación.

	AR Real	**NO AR Real**	
Clase AR predicha por RN Perceptrón	12	2	85.71%
Clase NO AR predicha por RN Perceptrón	1	13	
	92.3%	86.66%	89.28%

Después se implementa una red bayesiana utilizando las cinco variables propuesta por el Colegio Americano de Reumatología, la red bayesiana fue desarrollada utilizando un

estimador simple y un algoritmo de búsqueda TAN (*Tree Augmented Naive Bayes*), la red se muestra en la Figura5-4.

Figura 5-4: Red bayesiana utilizando cinco variables.



El diseño de la red bayesiana muestra la relación de la variable PCR con las otras variables FR, AntiCCP, articulaciones dolorosas y articulaciones inflamadas. De esta manera, la probabilidad de determinar AR o NO AR se da a partir de la tabla de probabilidad de PCR, la cual está directamente relacionada con las otras variables.

La matriz de confusión y los valores de las medidas de desempeño de la red bayesiana utilizando 5 variables se pueden observar en la Tabla 5-5.

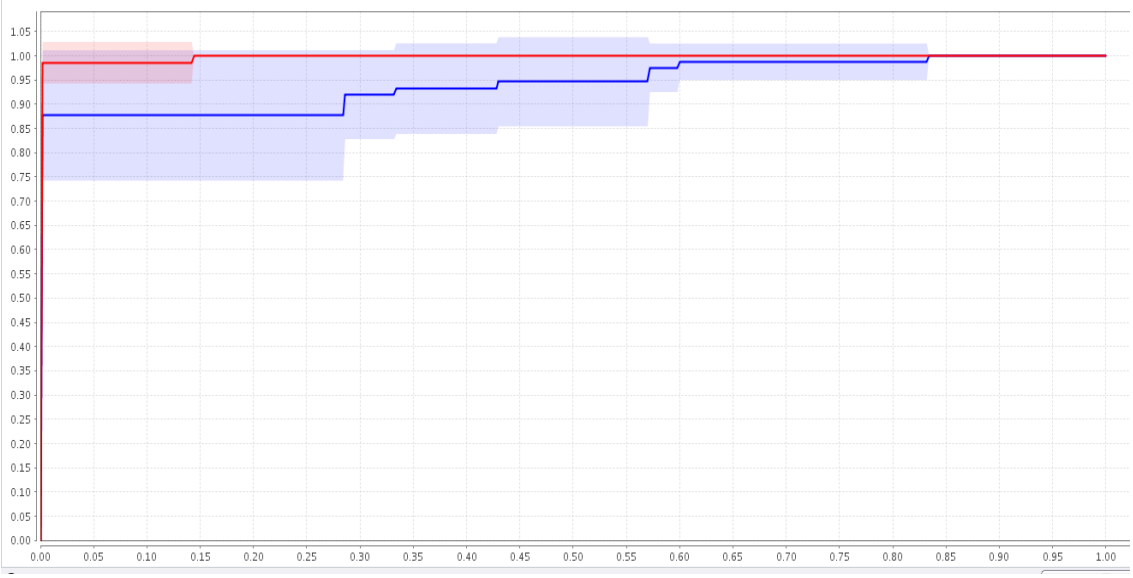
Tabla 5-5: Matriz de confusión y medidas de desempeño de la red bayesiana utilizando cinco variables.

	AR Real	**NO AR Real**	
Clase AR predicha por Red Bayesiana	12	1	92.3%
Clase NO AR predicha por Red Bayesiana	1	14	
	92.3%	93.33%	92.85%

En la Figura 5-5 se muestran las curvas ROC con los dos clasificadores red neuronal y red bayesiana utilizando dos variables, el color rojo corresponde a la red bayesiana tipo perceptrón multicapa y la curva azul corresponde a la red neuronal.

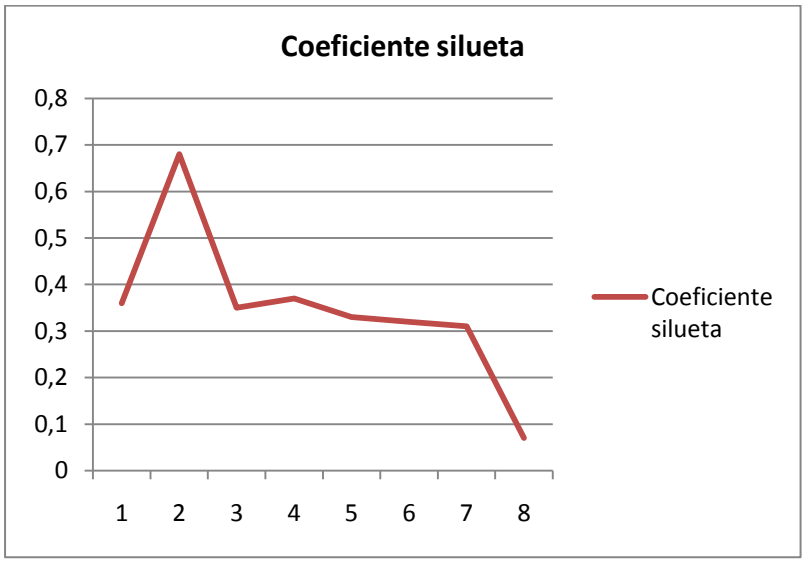
En la Figura 5-5 se observa que, según la curva ROC, el clasificador bayesiano, representado en color rojo, tiene un mejor desempeño que el clasificador neuronal, representado en color azul.

Figura 5-5: Curvas ROC de los clasificadores. La curva roja corresponde a la red bayesiana y la curva azul a la red neuronal



También se utilizó *k-means* como método de agrupación, para los datos considerando cinco variables según los criterios del Colegio Americano de Reumatología: PCR, FR, anti CCP, número de articulaciones dolorosas y número de articulaciones inflamadas. Se utilizó el coeficiente silueta para determinar el mejor *k* y utilizando una distancia euclidiana. El la Figura 5-6 se muestra la gráfica que presenta el coeficiente silueta con respecto a cada *k* posible, donde se muestra que el mejor coeficiente se obtiene con un *k* igual a 2.

Figura 5-6: Gráfica coeficiente silueta para agrupamiento de pacientes y controles utilizando 5 variables



A partir de los 138 datos se realizó el agrupamiento utilizando *k-means* sobre las variables normalizadas. Los dos grupos se conformaron de la siguiente manera: el primer grupo con 56 instancias y al segundo grupo con 82 instancias. El análisis estadístico de los dos grupos obtenidos se puede observar en las Tablas 5-6 y 5-7.

En el grupo 1 se observa que el valor del promedio del FR es de 63.54 lo cual es un valor positivo con respecto a los valores normales para dicho examen de laboratorio y además es mayor al obtenido en el grupo 2 que fue de 4.2. La desviación estándar del anti CCP del grupo 1 es de 76.05 con respecto a 40.28 del grupo 2, con lo cual se puede decir que los grupos no son homogéneos como se esperaba.

Los promedios del número de articulaciones inflamadas y dolorosas del grupo 1 fueron23.71 y 25, respectivamente. En contraste, los promedios del grupo 2 fueron 1.183 y 0.792. Lo anterior sugiere que los individuos del grupo 1 tiene un compromiso articular mayor con respecto al grupo 2.

Tabla 5-6: Análisis estadístico de los datos que conforman el grupo 1

Grupo 1	PCR	FR	Anti CCP	Articulaciones Inflamadas	Articulaciones Dolorosas
Promedio	2,34	63,54	109,11	23,71	25,00
Máximo	11,60	349,00	257,60	28,00	28,00
Mínimo	0,00	19,00	3,00	4,00	14,00
Desviación estándar	2,93	52,47	76,05	6,23	4,31

En el grupo 2 el promedio de los datos para la variable PCR fue de 0.11, el cual corresponde a un valor cualitativo negativo para dicho examen de laboratorio, siendo el máximo de 2.4 en este grupo. Por otro lado, en el grupo 1 el promedio fue de 2.34 con un valor máximo de 11.6.

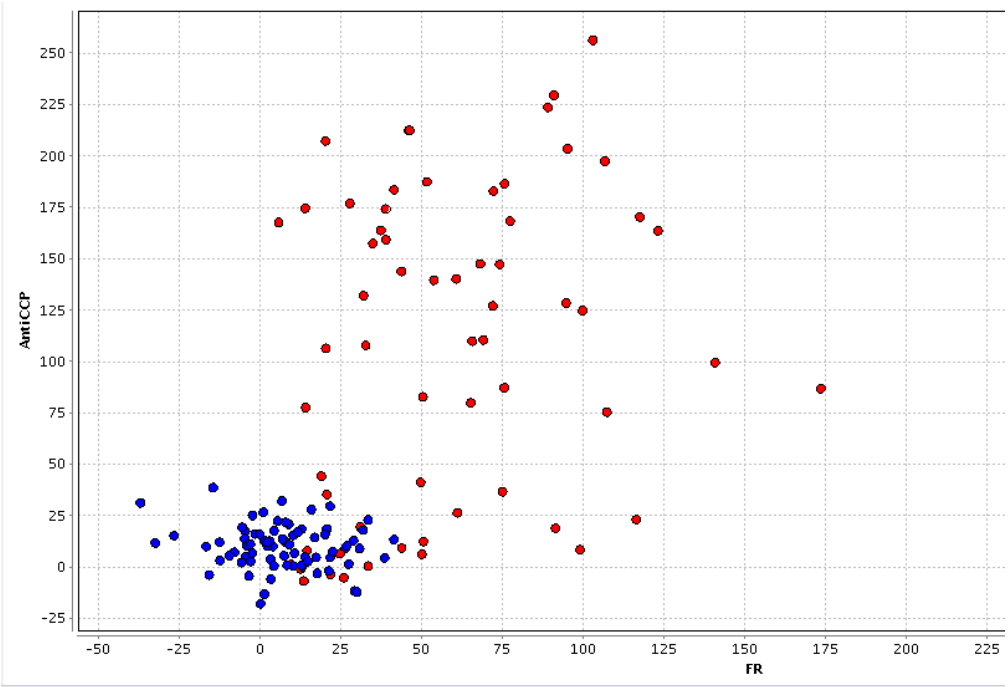
Tabla 5-7: Análisis estadístico de los datos que conforman el grupo 2

Grupo 2	PCR	FR	Anti CCP	Articulaciones Inflamadas	Articulaciones Dolorosas
Promedio	0,11	4,2	10,13	1,183	0,792
Máximo	2,4	92	213,4	28	12
Mínimo	0	0	0	0	0
Desviación estándar	0,43	16,15	40,28	4,45	2,71

Los valores promedio de anti CCP varían considerablemente en los dos grupos, siendo el promedio del grupo 1 de 109.11 y del grupo 2 de 10.13. Lo que indica, según lo reportado en la literatura, una predisposición mayor a presentar la enfermedad. Al examinar estos resultados se puede decir que los individuos del grupo 1 presentan una actividad mayor de la enfermedad con respecto a los individuos del grupo 2.

En la Figura 5-7 se grafican las variables FR y anti CCP representándose con color rojo el grupo 1 y con color azul el grupo 2de los dos grupos conformados. En el gráfico se puede observar que los individuos del grupo 1 toman valores mayores en la variables anti CCP con respecto a los valores que toman los individuos del grupo 2, lo cual corrobora los resultados reportados previamente en la literatura[57].

Figura 5-7: Visualización de las variables FR y Anti CCP. El color rojo corresponde al grupo 1 y el color azul al grupo 2.



También se diseñaron e implementaron varios modelos utilizando 11 variables: dos variables genéticas correspondientes a los dos alelos del HLA DRB1; las variables serológicas proteína C Reactiva, factor reumatoide y anticuerpos contra péptidos

citrulinados y seis variables clínicas: edad, antecedente de tabaquismo, antecedente de hipertensión arterial, antecedente de dislipidemia, número de articulaciones dolorosas y número de articulaciones inflamadas.

La red neuronal diseñada tuvo 7 neuronas para una capa oculta. Se aplicó filtro de nominal a binario para las variables de entrada, con una taza de aprendizaje de 0.3 y momento 0.2. La red neuronal se entrenó a través del método de retropropagación del error. Por otro lado, para la red bayesiana se utilizó un estimador simple y el algoritmo de búsqueda tipo TAN. Finalmente, se realizó agrupación con *k-means* usando una distancia euclidiana y para calcular el mejor *k* se utilizó el coeficiente de silueta.

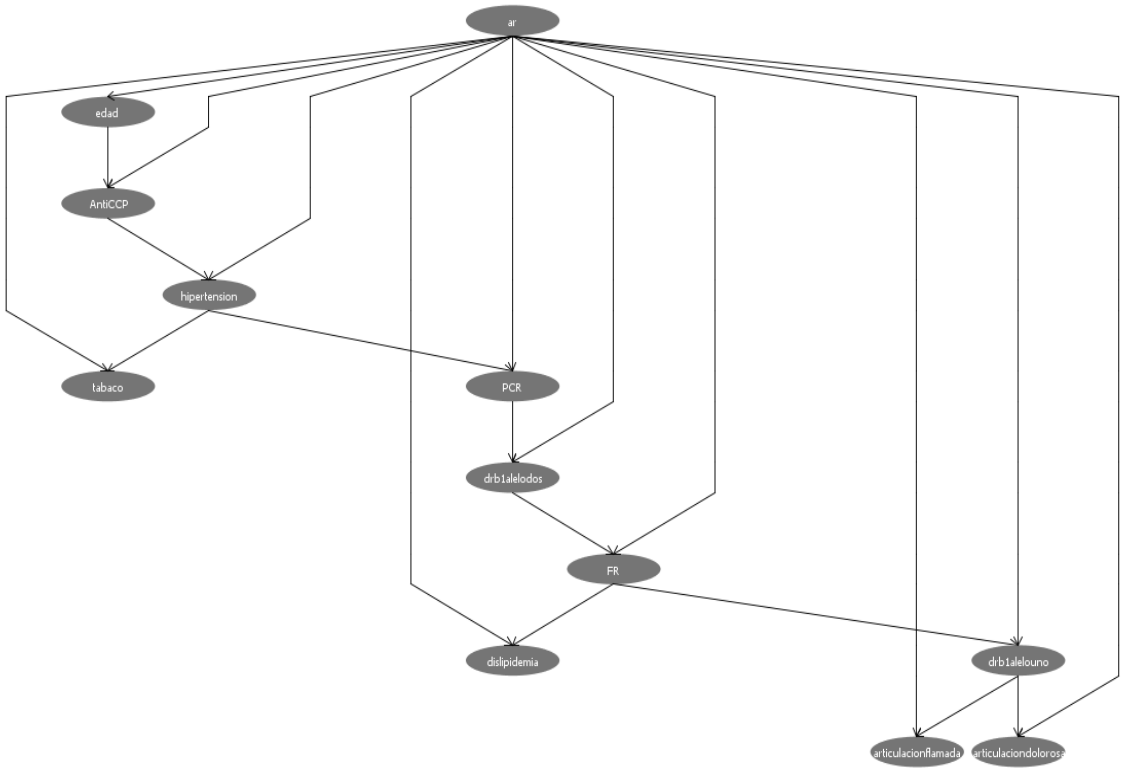
La red neuronal tipo perceptrón multicapa utilizando once variables obtuvo una exactitud de 89.28% que corresponde a 25 instancias correctamente clasificadas de 28 instancias de prueba. La correspondiente matriz de confusión y las medidas de desempeño se presentan en la Tabla 5-8.

Tabla 5-8: Matriz de confusión y medidas de desempeño de la red neuronal utilizando once variables para clasificación.

	AR Real	**NO AR Real**	
Clase AR predicha por RN Perceptrón	11	1	91.6%
Clase NO AR predicha por RN Perceptrón	2	14	
	84.61%	93.33%	89.28%

En la Figura5-8 se observa la red bayesiana implementada utilizando once variables para clasificación entre pacientes y controles, donde se pueden observar las dependencias probabilísticas entre las variables utilizadas.

Figura 5-8: Red bayesiana obtenida usando once variables.



En la red bayesiana observamos las relaciones que se presentan entre las distintas variables genéticas, clínicas y serológicas, principalmente la dependencia que se observa entre la variable alelo 1 y las variables clínicas articulaciones inflamadas y articulaciones dolorosas, esto nos demuestra la relación existen entre la genética de la enfermedad y sus manifestaciones clínicas.

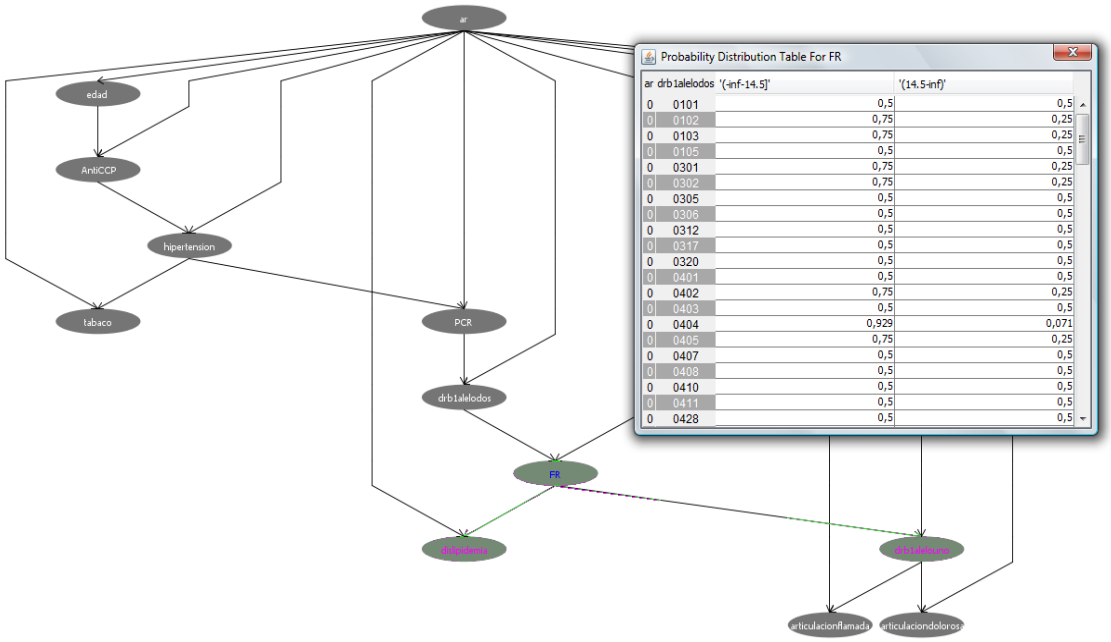
Con la red bayesiana se obtuvieron los siguientes resultados: de las 28 instancias de prueba se clasificaron adecuadamente 25, que corresponde a un 92.85%, e incorrectamente 3 instancias, que corresponde a un 7.15%, con una sensibilidad de 92.3% y una especificidad de 93.33%. La matriz de confusión y las medidas de desempeño se presentan en la Tabla 5-9.

Tabla 5-9: Matriz de confusión y medidas de desempeño de la red bayesiana utilizando once variables para clasificación.

	AR Real	**NO AR Real**	
Clase AR predicha por Red Bayesiana	11	1	92.3%
Clase NO AR predicha por Red Bayesiana	2	14	
	92.3%	93.33%	92.85%

Los nodos en la red bayesiana se relacionan mediante probabilidades de que un suceso ocurra, por ejemplo, en este caso para calcular la probabilidad de que una persona sufra de artritis reumatoide .En la Figura5-9 se muestra la tabla de probabilidad de que una persona presente artritis reumatoide dado que tiene antecedentes de hipertensión arterial y antecedentes de tabaquismo.

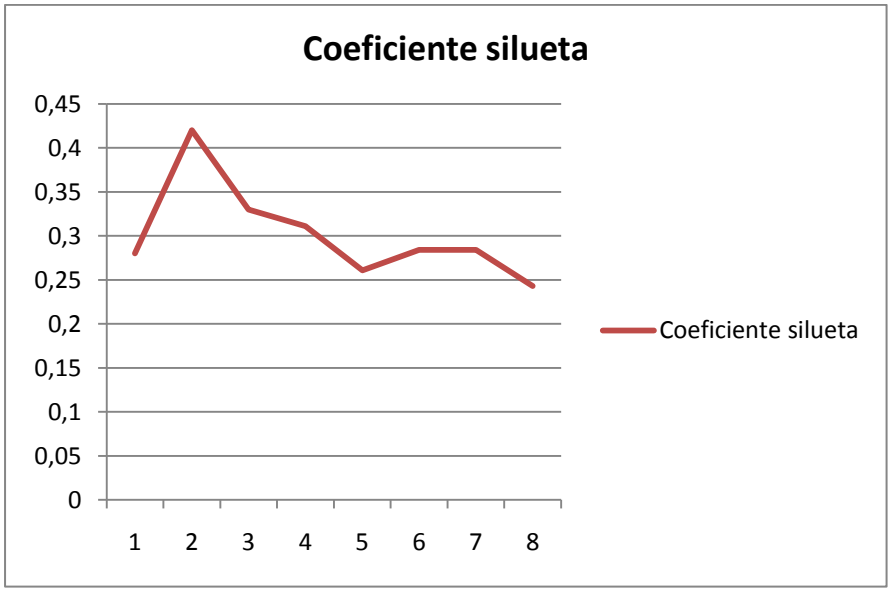
Figura 5-9: Relación entre la red bayesiana y la probabilidad de cada una de las variables con el desenlace de artritis reumatoide.



En el grafico se observan las tablas de probabilidad obtenidas a partir de la red bayesiana para calcular la probabilidad de AR, en el ejemplo se observa la probabilidad por alelo según valores del factor reumatoide.

También se realizó una agrupación con los datos de los 63 pacientes. Esto hace parte del estudio de clasificación de severidad de la enfermedad. Se utilizó el valor del coeficiente silueta para determinar el mejor número de grupos. La Figura 5-10 muestra los valores del coeficiente silueta para varios valores de k. El mejor valor del coeficiente silueta se obtiene para $k = 2$.

Figura 5-10: Coeficiente silueta para agrupamiento de pacientes



Para esto se utilizó *k-means* sobre las variables normalizadas para identificar entre los pacientes aspectos relacionados con la severidad de la enfermedad. Se hizo una agrupación en dos grupos, los cuales quedaron conformados de la siguiente manera: el primer grupo con 40 individuos y el segundo grupo con23 individuos. El análisis estadístico de los dos grupos obtenidos se puede observar en las Tablas 5-10 y 5-11.

Tabla 5-10: Análisis estadístico de los datos que conforman el grupo 1

Grupo 1	PCR	FR	Anti CCP	Articulaciones Inflamadas	Articulaciones Dolorosas
Promedio	1,82	66,53	32,13	23,26	24,08
Máximo	7,8	349	225,4	28	28
Mínimo	0	19	3	4	4
Desviación estándar	2,18	75,79	49,95	7,88	7,07

Tabla 5-11: Análisis estadístico de los datos que conforman el grupo 2

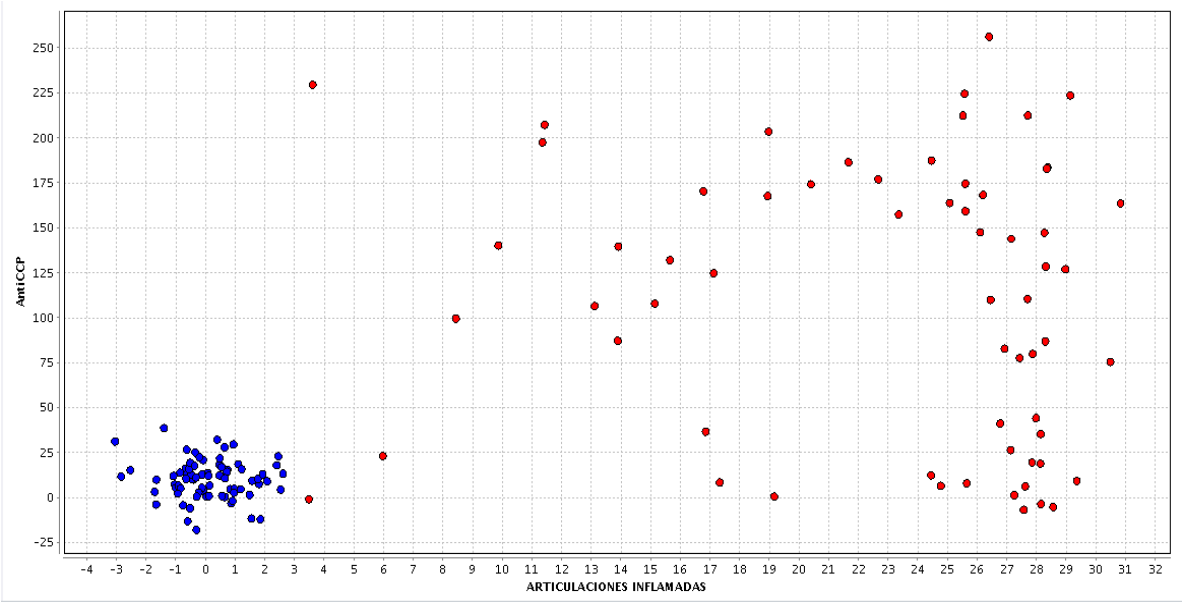
Grupo 2	PCR	FR	Anti CCP	Articulaciones Inflamadas	Articulaciones Dolorosas
Promedio	2,47	59,33	155,05	22,25	22,78
Máximo	11,6	123	257,6	28	28
Mínimo	0,1	19	68,3	4	9
Desviación estándar	3,08	28,25	46,49	6,66	6,15

Al comparar los dos grupos, se identificó que el promedio de articulaciones inflamadas y dolorosas entre los dos grupos fue similar, siendo para el grupo 1 de 23.26 y 24.08 con respecto al grupo 2 que fue de 22.25 y 22.78, respectivamente. Por lo cual, estas dos variables no son discriminantes. Por el contrario, el promedio de la variable anti CCP del grupo 1 fue 32.13 y del grupo 2 fue 155.05, lo que puede sugerir una enfermedad más activa en el grupo 2 con respecto al grupo 1. También, al analizar las desviaciones estándar del anti CCP se puede inferir que esta variable puede llegar a ser discriminante.

Los valores para la variable FR fueron: para el grupo 1 un máximo de 349 y un mínimo de 19 con una desviación estándar de 75.79 y para el grupo 2 un máximo de 123 y un mínimo de 19 con una desviación estándar de 28.25, todos estos valores son positivos con relación a los valores normales paraesta prueba de laboratorio. Lo que concluye que los individuos pertenecientes a estos grupos presentan una actividad de la enfermedad, de forma similarel promedio de articulaciones inflamadas y dolorosas, los cuales fuero,para el grupo 1, el promedio de articulaciones inflamadas y dolorosas fue de 23.26 y 24,08, y para el grupo 2 de 22,25 y 22,78 respectivamente lo cual muestra también actividad de la enfermedad. En la Figura 5-11 se puede observar el grafico que representa el número de articulaciones inflamadas con respecto a los valores de anti CCP. Se pueden observar claramente para estas dos variables los dos grupos. Las

muestras del grupo azul están densamente agrupadas en torno a la media de Anti CCP. Mientras que los datos del grupo rojo presentan una dispersión significativamente mayor.

Figura 5-11: Distribución de los grupos según las variables articulaciones inflamadas y anti CCP.

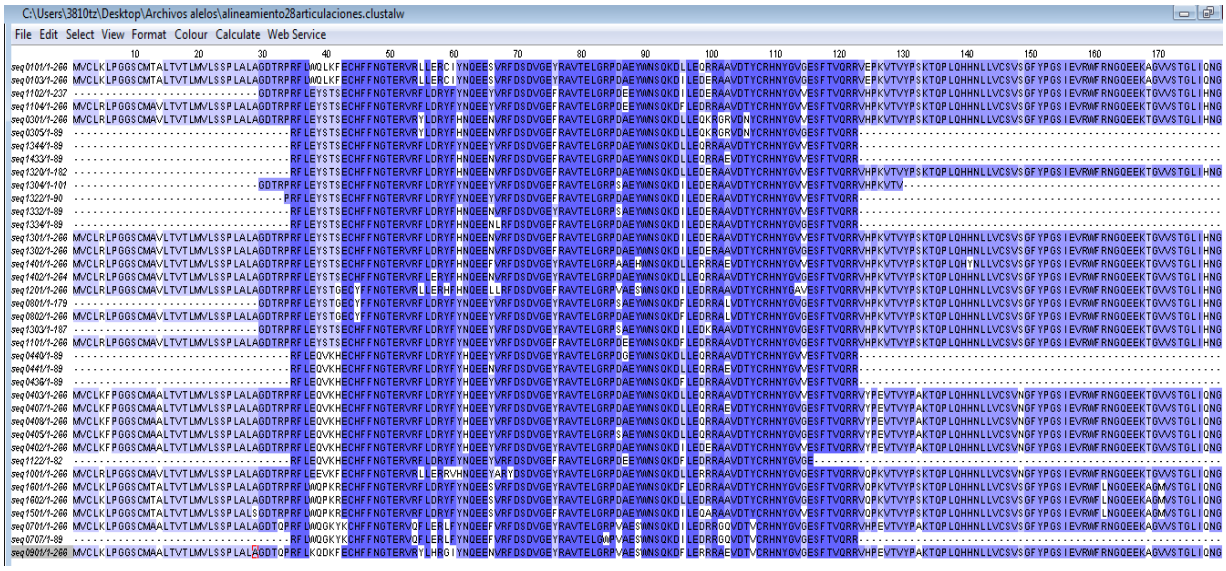


5.3 Identificación de endofenotipos en pacientes con AR.

Por otro lado, para la identificación de endofenotipos se realizó el alineamiento de las secuencias de aminoácidos de los distintos alelos correspondiente a pacientes con un desenlace en común. Esto se hizo usando clustalw [58], Para este caso se seleccionaron los alelos de los pacientes con 28 articulaciones comprometidas como criterio de severidad y se realizó el alineamiento de las secuencias correspondientes calculando una matriz Blosum62. Como resultado se encontraron secuencias comunes, además de las secuencias de la teoría del Epítope compartido.

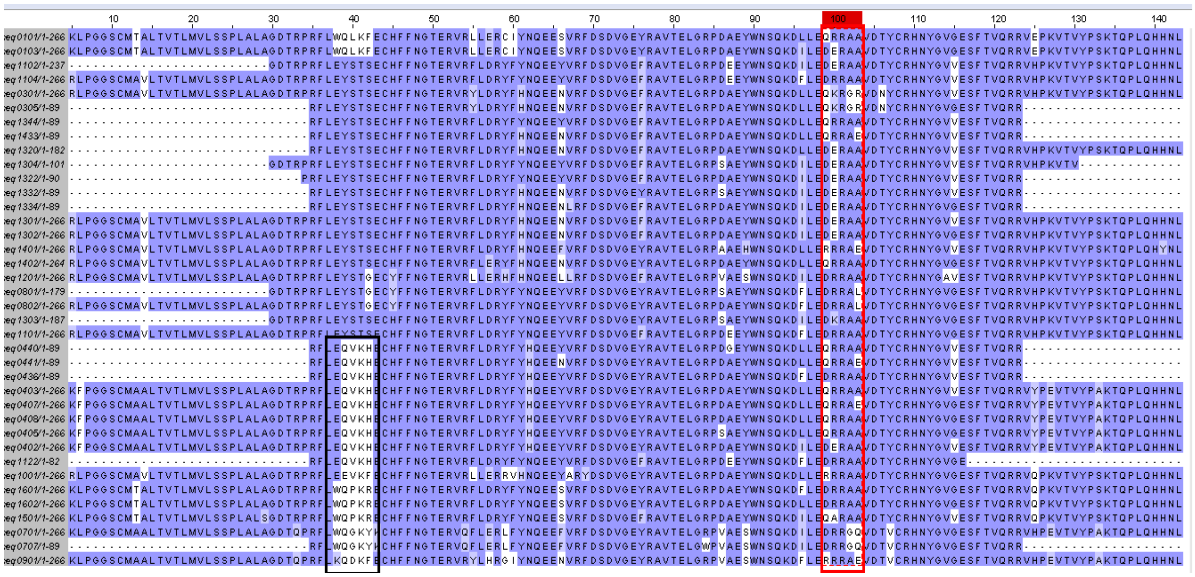
Se realizó primero la identificación de la identidad entre las secuencias de aminoácidos que corresponden a pacientes con 28 articulaciones comprometidas; el resultado se puede ver en la Figura 5-12.

Figura 5-12: Alineamiento de las secuencias de aminoácidos de pacientes con 28 articulaciones comprometidas.



En la Figura 5-13 se observa que las secuencias de aminoácidos QVKH y WQPKR son comunes en pacientes que tienen 28 articulaciones comprometidas, lo cual sugiere que están asociadas con este desenlace. En el recuadro rojo se resalta el área correspondiente al epítope compartido. En este estudio se logró determinar que, además de las secuencias de aminoácidos comunes entre pacientes con AR, también existen otras secuencias de aminoácidos comunes entre pacientes, las cuales pueden ser candidatas para la posible identificación de desenlaces de la enfermedad.

Figura 5-13: Alineamiento de las secuencias de aminoácidos de pacientes con 28 articulaciones comprometidas; con el recuadro rojo se resalta el área correspondiente al epítope compartido y con el recuadro negro las secuencias en común.



Se realizó un análisis de los datos tipificados en el HLA DRB1 y la información clínica de los pacientes con artritis reumatoide y de los controles. Esto se realizó a partir de la hipótesis del epítope compartido para esto se calculó la frecuencia de cada alelo entre los pacientes y los controles. En la Tabla 5-12 se muestran los resultados obtenidos. De las 138 personas que participaron en el estudio se identificaron 59 alelos diferentes para el HLA DRB1, de los cuales 8 son homocigotos y pertenecen a 4 pacientes (0101, 0103, 0441, 0701) y 4 controles (0701, 1101, 0802, 1303).

Tabla 5-12: Resultados del análisis de alelos HLA DRB1 en pacientes y controles. Los alelos 0101, 0102, 0105, 0401, 0403, 0404, 0405, 0408, 0410, 0428, 0433, 0440, 0442, 1001, 1344, 1402 corresponden al epítope compartido.

Alelo	Frecuencia alelo en Pacientes	Frecuencia alelo en Controles	Alelo	Frecuencia alelo en Pacientes	Frecuencia alelo en Controles	Alelo	Frecuencia alelo en Pacientes	Frecuencia alelo en Controles
0101	***13***	***7***	***0428***	***0***	***0***	1102	1	2
0102	***2***	***4***	0431	0	0	1104	0	2
0103	1	1	***0433***	***1***	***0***	1122	1	0
0105	***1***	***0***	0436	0	0	1201	1	1
0301	5	6	***0440***	***1***	***0***	1202	0	1
0302	0	1	0441	3	0	1301	3	4
0305	1	0	***0442***	***1***	***0***	1302	2	5
0306	1	0	0448	1	0	1303	1	3
0312	0	0	0501	0	1	1304	1	0
0317	1	0	0701	11	19	1322	1	0
0320	0	0	0707	1	0	1332	0	0
0401	***1***	***2***	0801	2	3	1334	0	0
0402	1	2	0802	4	10	***1344***	***2***	***0***
0403	***3***	***23***	0803	0	1	1401	5	4
0404	***5***	***6***	0816	1	0	***1402***	***4***	***5***
0405	***8***	***1***	0825	1	0	1501	9	14
0407	8	1	0901	3	3	1502	1	0
0408	***1***	***0***	0902	1	0	1601	3	6
0410	***1***	***0***	***1001***	***2***	***4***	1602	3	0
0411	0	0	1101	2	8			

La Figura 5-14 muestra el histograma para los alelos HLA DRB1. Los alelos más frecuentes en los datos correspondientes a pacientes se presentan en la Tabla 5-13:

Figura 5-14: Histograma de los alelos HLA DRB1 para el alelo 1.

Tabla 5-13:Frecuencia de alelos en pacientes.

Alelo	Frecuencia en pacientes	Porcentaje
***0101**	13	10.3%
***0701**	11	8.7%
***1501**	9	7.1%
***0405**	8	6.3%
***0407**	8	6.3%

Los alelos más frecuentes en los controles sanos se presentan en la Tabla 5-14:

Tabla 5-14: Frecuencia de alelos en controles.

Alelo	Frecuencia en controles	Porcentaje
***0403**	23	15.3%
***0701**	19	12.6%
***1501**	14	9.3%
***0802**	10	6.6%
***1101**	8	3.3%

Las secuencias QKRAA, QRRAA y RRRAA que corresponde a los aminoácidos de los alelos del epítope compartido, se repitieron 98 veces en los datos analizados, estas se relacionan con el epítope compartido. Esto significa que estas secuencias hacen parte de las reportadas en la literatura como epítope compartido[2].

A partir de esta información, para cada uno de los alelos en estudio se encontraron las secuencias que se muestran en la Tabla 5-15.

Tabla 5-15: Correspondencia entre alelos y secuencias de aminoácidos. Las secuencias de aminoácidos en QKRAA, QRRAA y RRRA corresponden al epítope compartido.

Secuencia de aminoácidos	Alelos	Pacientes	Controles sanos	Total
QKRAA	0401, 0433	2	2	4
QRRAA	0101, 0102, 0105, 0403, 0404, 0405, 0408, 0410, 0428, 0440, 0442, 1344, 1402.	42	46	88
RRRAA	1001	2	4	6
DERAA	0103, 0402, 1102, 1301, 1302, 1304, 1322, 1332, 1334.	10	14	24
DRRGQ	0701	11	19	30
QARAA	1501	9	14	23
DRRAL	0802	4	10	14
QKRGR	0301	5	6	11
QRRAE	0407	8	1	9
DRRAA	1101	2	8	10

Se realizó un alineamiento entre las distintas secuencias de aminoácidos de los alelos. El resultado de este alineamiento se muestra en la Figura 5-15.

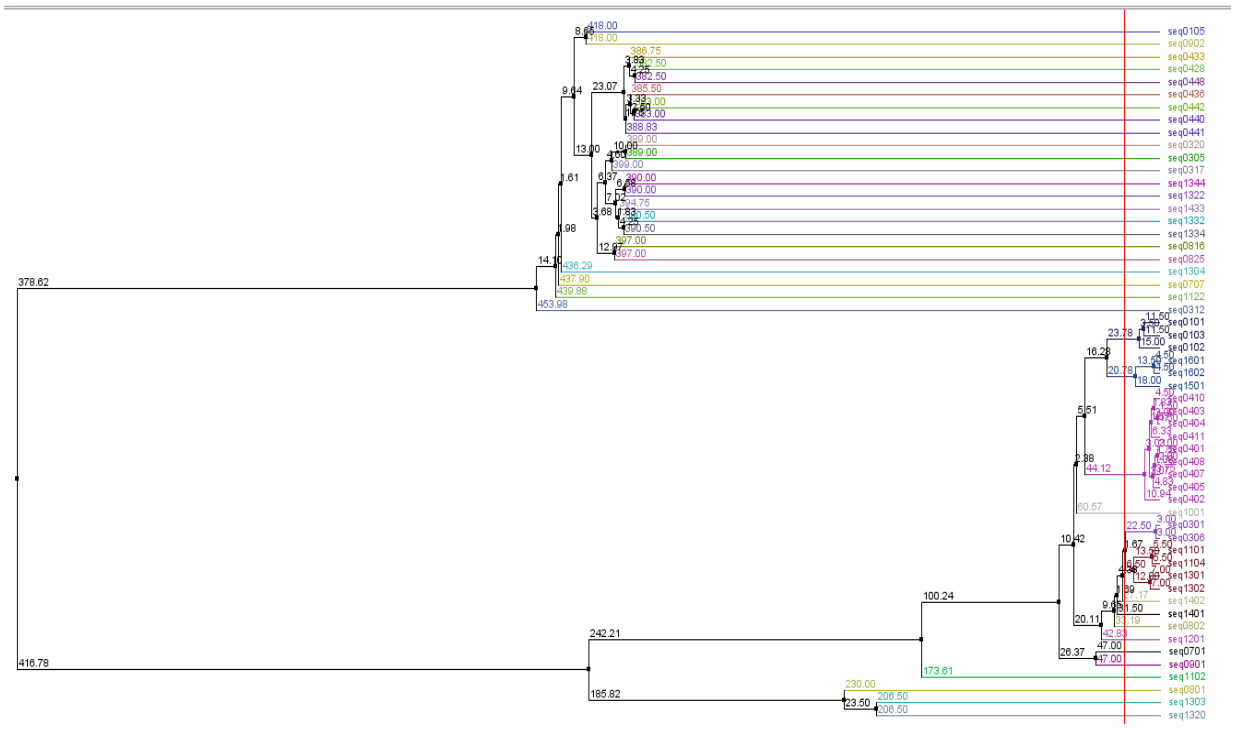
Con los datos utilizados en el presente estudio se confirma la hipótesis del epítope compartido en los pacientes colombianos con AR, esto a partir de que los alelos mas frecuentes y que corresponden a las secuencias QKRAA, QRRAA y RRRAA están presentes en los pacientes con AR, se logra identifica el alelo 0101 implicado en la AR es el más frecuente en pacientes con AR.

Figura 5-15: Alineamiento de los distintos alelos HLA DRB1.

```
seq0101    YNQEESVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQRRAAVDTYCRHNYGVGESFTV 120
seq0105    YNQEESVRFDSDVREYRAVTELGRPDAEYWNSQKDLLEQRRAAVDTYCRHNYGVGESFTV 86
seq0102    YNQEESVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQRRAAVDTYCRHNYGAVESFTV 120
seq0103    YNQEESVRFDSDVGEYRAVTELGRPDAEYWNSQKDILEDERAAVDTYCRHNYGVGESFTV 120
seq0410    YHQEEYVRFDSDVGEYRAVTELGRPSAEYWNSQKDLLEQRRAAVDTYCRHNYGVVESFTV 120
seq0411    YHQEEYVRFDSDVGEYRAVTELGRPSAEYWNSQKDLLEQRRAEVDTYCRHNYGVVESFTV 120
seq0401    YHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQKRAAVDTYCRHNYGVGESFTV 120
seq0433    YHQEEYVHFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQKRAAVDTYCRHNYGVGESFTV 86
seq0408    YHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQRRAAVDTYCRHNYGVGESFTV 120
seq0428    YHQEEYVRFDSDVGEFRAVTELGRPSAEYWNSQKDLLEQRRAAVDTYCRHNYGVGESFTV 86
seq0448    YHQEEYVRFDSDVGEYRAVTELGRPSAEYWNSQKDLLEQRRSAVDTYCRHNYGVGESFTV 86
seq0405    YHQEEYVRFDSDVGEYRAVTELGRPSAEYWNSQKDLLEQRRAAVDTYCRHNYGVGESFTV 120
seq0407    YHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQRRAEVDTYCRHNYGVGESFTV 120
seq0436    YHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDFLEDRRAAVDTYCRHNYGVVESFTV 86
seq1122    YNQEEYVRFDSDVGEFRAVTELGRPDEEYWNSQKDFLEDRRAAVDTYCRHNYGVGE---- 82
seq0442    YHQEEYVRFDSDVGEFRAVTELGRPDAEYWNSQKDLLEQRRAAVDTYCRHNYGVVESFTV 86
seq0440    YHQEEYVRFDSDVGEYRAVTELGRPDGEYWNSQKDLLEQRRAAVDTYCRHNYGVVESFTV 86
seq0441    YHQEENVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQRRAEVDTYCRHNYGVVESFTV 86
seq0403    YHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQRRAAVDTYCRHNYGVVESFTV 120
seq0404    YHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQRRAAVDTYCRHNYGVVESFTV 120
seq0402    YHQEEYVRFDSDVGEYRAVTELGRPDAEYWNSQKDILEDERAAVDTYCRHNYGVVESFTV 120
seq0312    HNQEENVRFDSDVGEFRAVTELGRPSAEYWNSQKDLLEQKRGRVDNYCRHNYGVVE---- 80
seq0320    HNQEENVRFDSDVGEFRAVTELGRPDAEYWNSQKDLLEQKRGRVDNYCRHNYGAVESFTV 86
seq0305    HNQEENVRFDSDVGEFRAVTELGRPDAEYWNSQKDLLEQKRGRVDNYCRHNYGVGESFTV 86
seq0301    HNQEENVRFDSDVGEFRAVTELGRPDAEYWNSQKDLLEQKRGRVDNYCRHNYGVVESFTV 120
seq0306    HNQEENVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQKRGRVDNYCRHNYGVVESFTV 120
seq0302    HNQEENVRFDSDVGEYRAVTELGRPDAEYWNSQKDLLEQKRGRVDNYCRHNYGVGESFTV 91
```

Finalmente, se creó un árbol entre los alelos de los pacientes con AR por el método de distancia promedio (*average distance*) usando Blosum62 [59] con el fin de buscar cercanía genética entre los alelos para posteriormente asociarlos con lo reportado en la literatura. En la Figura 5-16se presenta el árbol encontrado, el cual fue obtenido usando *Jalview*.

Figura 5-16: Árbol de distancia entre secuencias de aminoácidos de los distintos alelos HLA DRB1.



En la Figura 5-16 se observan los grupos que se forman a partir de información de los pacientes. Allí se identifican algunos grupos que verifican lo que ha sido reportado en la literatura sobre criterios de severidad. Particularmente se forman grupos con riesgo alto que corresponde a los alelos 0401 y 1303, riesgo intermedio 0101, 0102, 0404, 0405, 0408, 1001, 1402 y riesgo bajo que corresponde a los alelos 0103, 0402, 03, 0403, 0407, 0411, 07, 08, 09, 1101, 1102, 1103, 1104, 12, 1301, 1302, 1323, 1401, 1404, 15, 16.

En la Tabla 5-16 se presenta los grupos obtenidos por cercanía genética a partir de la secuencia de aminoácidos correspondiente a cada alelo los cuales se pueden comparar con resultados previos obtenidos en otros estudios como el que se presenta en la Tabla 5-17.

Tabla 5-16: Grupos de alelos obtenidos a partir del árbol de cercanía genética

Grupo	Alelos
1	0105
2	0902
3	0433, 0707
4	0428, 0442, 0305
5	0448, 0440, 0441
6	0436
7	0320
8	0317
9	1344
10	1320, 1322, 1334
11	1433
12	1303, 1304, 1332
13	0816
14	0825
15	0101, 0102, 0103,
16	1501, 1601, 1602

Tabla 5-17: Tabla de riesgo de AR según alelos, tomada de Classification of HLA–DRB1 alleles according to the third hypervariable region of the DR β chain and their association with RA in French caucasian patients (Data adapted from du Montcel et al.)

	DRβ–chain amino acid (shared epitope)								
	HLA–DRB1						*Disease*		
allele[¶]	67	68	69	70	71	72	73	74	Association[#]
S_{3P}									**Int risk**
DRB1*0101	L	L	E	Q	R	R	A	A	+
DRB1*0102	–	–	–	–	–	–	–	–	+
DRB1*0404	–	–	–	–	–	–	–	–	+
DRB1*0405	–	–	–	–	–	–	–	–	+
DRB1*0408	–	–	–	–	–	–	–	–	+
DRB1*1001	–	–	–	R	–	–	–	–	+
DRB1*1402	–	–	–	–	–	–	–	–	+
S_{3D}									**Low risk**
DRB1*1101	–	F	–	D	–	–	–	–	–
DRB1*1104	–	F	–	D	–	–	–	–	–
DRB1*12	–	I	–	D	–	–	–	–	–
DRB1*16	–	F	–	D	–	–	–	–	–
S_1									**Low risk**
DRB1*0103	–	I	–	D	E	–	–	–	–
DRB1*0402	–	I	–	D	E	–	–	–	–
DRB1*1102	–	I	–	D	E	–	–	–	–
DRB1*1103	–	F	–	D	E	–	–	–	–
DRB1*1301	–	I	–	D	E	–	–	–	–
DRB1*1302	–	I	–	D	E	–	–	–	–
DRB1*1323	–	I	–	D	E	–	–	–	–
DRB1*15	I	–	–	–	A	–	–	–	–
S_2									**High risk**
DRB1*0401	–	–	–	–	K	–	–	–	+
DRB1*1303	–	I	–	D	K	–	–	–	+
X									**Low risk**
DRB1*03	–	–	–	–	K	–	G	R	+/–
DRB1*0403	–	–	–	–	–	–	–	E	–
DRB1*0407	–	–	–	–	–	–	–	E	–
DRB1*0411	–	–	–	–	–	–	–	E	–
DRB1*07	–	I	–	D	–	–	G	Q	–
DRB1*08	–	F	–	D	–	–	–	L	–
DRB1*09	–	F	–	R	–	–	–	E	–
DRB1*1401	–	–	–	R	–	–	–	–	–
DRB1*1404	–	–	–	R	–	–	–	E	

6. Discusión.

Mediante técnicas computaciones, como las redes neuronales, utilizando dos variables genéticas se logró clasificar a pacientes con AR y controles sanos, logrando clasificar adecuadamente el 85.71% de los datos. Por otro lado, una red bayesiana logró clasificar de manera adecuada el 67.85% de los datos.

Utilizando los criterios del Colegio Americano de Reumatología, es decir cinco variables, se logró mejorar el desempeño de los modelos. En el caso de la red neuronal se logró clasificar adecuadamente el 89.28% de los datos. Con la red bayesiana con cinco variables se clasificó adecuadamente el 92.85% de los datos. Por otro lado, la red neuronal utilizando 11 variables clasificó adecuadamente el 89.28% y la red bayesiana clasificó adecuadamente el 92.85% de los datos.

Los resultados obtenidos con *k-means* con el grupo de 138 individuos permitieron determinar que el grupo 1 presentaba una susceptibilidad mayor a presentar AR dada por los resultados de las variables analizadas.

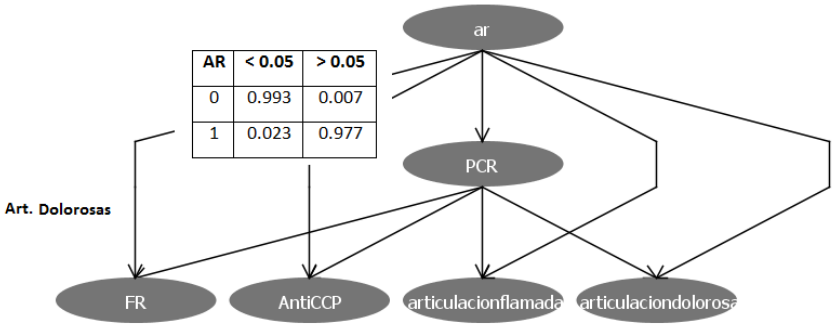
Posteriormente, se realizó un nuevo agrupamiento con pacientes para determinar severidad de la enfermedad. Al utilizar el coeficiente silueta se determino que le mejor agrupamiento se logra con dos grupos. Al comparar los dos grupos, se identificó que el promedio de articulaciones inflamadas y dolorosas entre los dos grupos fue similar, por lo cual, estas dos variables no son discriminantes. En contraste, la variable anti CCP resultó discriminante para la condición de severidad de la enfermedad. Por otro lado, los valores para la variable FR son cualitativamente positivos, lo cual solo indica que los individuos

pertenecientes a estos grupos presentan una actividad de la enfermedad. Algo similar se puede observar para las articulaciones inflamadas y dolorosas.

Adicionalmente, en la red bayesiana se identificaron relaciones significativas entre las variables, tales como la relación entre el número de articulaciones inflamadas y dolorosas con las variables serológicas y las variables genéticas. Este hallazgo está acorde con los resultados obtenidos en [49], donde analizaron las relaciones existente entre los fenotipos y biomarcadores en un estudio de AR. Por lo cual, en esta tesis a través de los métodos computaciones mencionados se encontraron relaciones de dependencia entre estas variables, principalmente usando los modelos bayesianos. Las tablas de los valores de las probabilidades correspondientes indican la posibilidad de que una persona presente artritis reumatoide.

En la Figura 6-1se observan las tablas de probabilidad y las relaciones de dependencia existente entre las distintas variables, dado es el caso de relación de dependencia existente entre anti CCP, PCR y AR. Otra relación que se observa es la dependencia directa entre las articulaciones inflamadas y las articulaciones dolorosas. También se evidencia que existe una probabilidad mayor de presentar AR en los individuos que presentan una PCR mayor de 0.05, FR mayor de 9.5 y un anti CCP mayor de 1.5.

Figura 6-1: Red bayesiana y sus tablas de probabilidades para la AR



AR	< 0.05	> 0.05
0	0.993	0.007
1	0.023	0.977

FR

AR	PCR	< 9.5	> 9.5
0	<0.05	0.993	0.007
0	>0.05	0.05	0.5
1	<0.05	0.25	0.75
1	>0.05	0.008	0.992

Anti CCP

AR	PCR	< 1.5	> 1.5
0	<0.05	0.993	0.007
0	>0.05	0.5	0.5
1	<0.05	0.25	0.75
1	>0.05	0.008	0.992

Articulaciones Inflamadas

AR	PCR	< 2	> 2
0	<0.05	0.993	0.007
0	>0.05	0.5	0.5
1	<0.05	0.25	0.75
1	>0.05	0.008	0.992

Articulaciones Dolorosas

AR	PCR	< 2	> 2
0	<0.05	0.993	0.007
0	>0.05	0.5	0.5
1	<0.05	0.25	0.75
1	>0.05	0.008	0.992

Por otro lado, en este estudio se identificaron 59 diferentes alelos del HLA DRB1 en la muestra estudiada de los 138 individuos entre pacientes y controles, siendo el alelo más frecuente en pacientes el *0101, el cual está implicado en la teoría del epítope compartido, lo cual confirma lo reportado en la literatura[2].

Finalmente, mediante el alineamiento de las secuencias de aminoácidos del HLA DRB1 se identificaron secuencias comunes en los pacientes con un grado de severidad determinado por tener 28 articulaciones dolorosas o inflamadas. Se encontraron las posibles secuencias comunes QVKR Y WQPKR entre estos pacientes.

7. Conclusiones y recomendaciones

7.1 Conclusiones

En este trabajo se realizó el análisis de un conjunto de datos de pacientes con artritis reumatoide aplicando métodos de inteligencia computacional con el fin de tratar de identificar endofenotipos para la enfermedad. Los datos incluían información genética, clínica y serológica. También se trataron de obtener marcadores para determinar la severidad de la enfermedad en los pacientes.

Se identificaron las relaciones de dependencia entre algunas de las variables a través de una red bayesiana. Específicamente se estableció que las variables anti CCP y el factor reumatoide están implicadas de manera significativa en la enfermedad. También, se identificaron endofenotipos relacionados con la severidad de la enfermedad en pacientes; particularmente se identificaron secuencias de aminoácidos comunes entre pacientes con el mismo desenlace.

Se concluyó que utilizando técnicas de inteligencia computacional y bioinformática fue posible desarrollar modelos computacionales que utilizando información genética y clínica logran clasificar entre pacientes y controles sanos, así como también clasificar según criterios de severidad a los pacientes.

7.2 Recomendaciones

Se sugiere realizar análisis similares al llevado a cabo en este trabajo sobre conjuntos con un mayor número de datos para confirmar los resultados de este estudio y para tratar de identificar otros endofenotipos para la artritis reumatoide.

Por otro lado, se sugiere aplicar otras técnicas de inteligencia computacional, tales como máquinas de soporte vectorial, técnicas como *boosting* y *bagging*, para clasificar y diferenciar entre los pacientes y controles sanos.

Se recomienda también trabajar en la búsqueda de otras secuencias comunes entre los distintos pacientes con grados de severidad similares, realizando un estudio similar al de esta tesis, incluyendo otras variables como niveles elevados del Factor Reumatoide, niveles elevados de anticuerpos contra Péptidos Citrulinados y HAQ elevados.

Se recomienda en un trabajo futuro desarrollar una aplicación Web que permita automatizar la metodología de análisis aplicada en esta investigación.

Participaciones

Presentación oral, Universidad de los Andes - Fundación Santa Fe. Encuentro de Investigadores en Ingeniería Biomédica Fundación Santa Fe de Bogotá. Ingeniería Biomédica Universidad de los Andes. Agosto 2012. Bogotá Colombia.

Presentación oral, 1st International Symposium of Health Informatics in Latin America and the Caribbean. August 14, 2013, Cancun, Mexico.

Presentación Poster, 2013 International Conference on Intelligent Biology and Medicine (ICIBM). August 11-13, 2013, Nashville, TN, USA.

Presentación Oral en XIV Congreso Colombiano de Reumatología. 15 – 18 de agosto 2013. Medellín Antioquia.

Aceptación Presentación oral VIII Congreso Colombiano de Computación 8CCC.21 – 23 agosto 2013. Armenia Quindío.

Aceptación Paper AMT-BHI 2013.The 2013 International Conference on Active Technology and Brain and Health Informatics Maebashi Japan.

Publicaciones

L. M. Muñoz, L. F. Niño, and G. Q. Ms, “Modelo Computacional para la identificación de endofenotipos y clasificación de pacientes con Artritis Reumatoide a partir de datos genéticos y clinicos utilizando técnicas de Inteligencia Computacional,” pp. 1–6, 2013.

G. Quintana and L. A. Morales, L. Niño. “Immunological Computation: An overview. Chapter. APLA & LACA 2013. 14th International Congress on Antiphospholipid Antibodies & 4th Latin American Congress on Autoimmunity. 18 – 21 September, 2013. Rio de Janeiro, Brazil.

Modelo computacional para la identificación de endofenotipos en pacientes Colombianos con Artritis Reumatoide. Articulo aceptado en Revista Colombiana de Reumatología Elsevier Pendiente publicación.

Premios

Tercer puesto en XIV Congreso Colombiano de Reumatología. 15 – 18 de agosto 2013. Medellín Antioquia.

Bibliografía

[1] I. B. McInnes and G. Schett, “The pathogenesis of rheumatoid arthritis.,” *The New England journal of medicine*, vol. 365, no. 23, pp. 2205–19, Dec. 2011.

[2] M. Bax, J. van Heemst, T. W. J. Huizinga, and R. E. M. Toes, “Genetics of rheumatoid arthritis: what have we learned?,” *Immunogenetics*, vol. 63, no. 8, pp. 459–66, Aug. 2011.

[3] J.-M. Anaya, P. a. Correa, R. D. Mantilla, and M. Arcos-Burgos, “TAP, HLA-DQB1, and HLA-DRB1 polymorphism in Colombian patients with primary Sjögren’s syndrome,” *Seminars in Arthritis and Rheumatism*, vol. 31, no. 6, pp. 396–405, Jun. 2002.

[4] A. M. Delgado-vega, J. Martín, J. Granados, and J. M. Anaya, “Epidemiología genética de la artritis reumatoide : ¿ qué esperar de América Latina ?,” pp. 562–585, 2006.

[5] D. Aletaha, T. Neogi, A. J. Silman, J. Funovits, D. T. Felson, C. O. Bingham, N. S. Birnbaum, G. R. Burmester, V. P. Bykerk, M. D. Cohen, B. Combe, K. H. Costenbader, M. Dougados, P. Emery, G. Ferraccioli, J. M. W. Hazes, K. Hobbs, T. W. J. Huizinga, A. Kavanaugh, J. Kay, T. K. Kvien, T. Laing, P. Mease, H. a Ménard, L. W. Moreland, R. L. Naden, T. Pincus, J. S. Smolen, E. Stanislawska-Biernat, D. Symmons, P. P. Tak, K. S. Upchurch, J. Vencovský, F. Wolfe, and G. Hawker, “2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative.,” *Arthritis and rheumatism*, vol. 62, no. 9, pp. 2569–81, Sep. 2010.

[6] D. Jawaheer, R. F. Lum, P. K. Gregersen, and L. a Criswell, “Influence of male sex on disease phenotype in familial rheumatoid arthritis.,” *Arthritis and rheumatism*, vol. 54, no. 10, pp. 3087–94, Oct. 2006.

[7] I. C. Scott, S. D. Seegobin, S. Steer, R. Tan, P. Forabosco, A. Hinks, S. Eyre, A. W. Morgan, A. G. Wilson, L. J. Hocking, P. Wordsworth, A. Barton, J. Worthington, A. P. Cope, and C. M. Lewis, “Predicting the risk of rheumatoid arthritis and its age of onset through modelling genetic risk variants with smoking.,” *PLoS genetics*, vol. 9, no. 9, p. e1003808, Jan. 2013.

[8] W. a Bardwell, P. M. Nicassio, M. H. Weisman, R. Gevirtz, and D. Bazzo, “Rheumatoid Arthritis Severity Scale: a brief, physician-completed scale not confounded by patient self-report of psychological functioning.,” *Rheumatology (Oxford, England)*, vol. 41, no. 1, pp. 38–45, Jan. 2002.

[9] C. V. Caballero, J. Rozenboim, E. Afanador, C. Venegas, F. Rocha, M. Carpio, and L. Alonso, “Utilidad de un cuestionario de factores pronósticos en la evaluación de pacientes con artritis reumatoide,” vol. 12, no. 4, pp. 301–311, 2005.

[10] D. L. Mattey, A. B. Hassell, M. J. Plant, N. T. Cheung, P. T. Dawes, P. W. Jones, W. Thomson, K. V Poulton, A. H. Hajeer, and W. E. R. Ollier, “The influence of HLA-DRB1 alleles encoding the DERAA amino acid motif on radiological outcome in rheumatoid arthritis,” pp. 1221–1227, 1999.

[11] R. F. van Vollenhoven, “Treatment of rheumatoid arthritis: state of the art 2009.,” *Nature reviews. Rheumatology*, vol. 5, no. 10, pp. 531–41, Oct. 2009.

[12] L. Massardo, N. Gareca, M. a Cartes, V. Cervilla, a González, and S. Jacobelli, “The presence of the HLA-DRB1 shared epitope correlates with erosive disease in Chilean patients with rheumatoid arthritis.,” *Rheumatology (Oxford, England)*, vol. 41, no. 2, pp. 153–6, Mar. 2002.

[13] V. T. Lima, A. María, T. Lima, M. Victoria, H. Cuéllar, and R. A. Relova, “Inmunopatogenia de la artritis reumatoidea. conceptos actuales,” vol. 14, no. 5, pp. 429–433, 1998.

[14] W. J. van Venrooij, A. J. W. Zendman, and G. J. M. Pruijn, “Autoantibodies to citrullinated antigens in (early) rheumatoid arthritis.,” *Autoimmunity reviews*, vol. 6, no. 1, pp. 37–41, Nov. 2006.

[15] H. Kim, C. Berek, and D. R. Forschungszentrum, “Review B cells in rheumatoid arthritis,” 2000.

[16] N. Wegner, K. Lundberg, A. Kinloch, B. Fisher, V. Malmström, M. Feldmann, and P. J. Venables, “Autoimmunity to specific citrullinated proteins gives the first clues to the etiology of rheumatoid arthritis.,” *Immunological reviews*, vol. 233, no. 1, pp. 34–54, Jan. 2010.

[17] J. G. Routsias, J. D. Goules, A. Goules, G. Charalampakis, and D. Pikazis, “Autopathogenic correlation of periodontitis and rheumatoid arthritis.,” *Rheumatology (Oxford, England)*, vol. 50, no. 7, pp. 1189–93, Jul. 2011.

[18] J. Detert, N. Pischon, G. R. Burmester, and F. Buttgereit, “The association between rheumatoid arthritis and periodontal disease,” 2010.

[19] A. Gizinski and R. Morgan, “NIH Public Access,” vol. 36, no. 2, pp. 311–323, 2011.

[20] H. Xing, P. D. McDonagh, J. Bienkowska, T. Cashorali, K. Runge, R. E. Miller, D. Decaprio, B. Church, R. Roubenoff, I. G. Khalil, and J. Carulli, “Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis.,” *PLoS computational biology*, vol. 7, no. 3, p. e1001105, Mar. 2011.

[21] J. a G. van Roon and J. W. J. Bijlsma, “Th2 mediated regulation in RA and the spondyloarthropathies.,” *Annals of the rheumatic diseases*, vol. 61, no. 11, pp. 951–4, Nov. 2002.

[22] E. Lubberts, “Th17 cytokines and arthritis.,” *Seminars in immunopathology*, vol. 32, no. 1, pp. 43–53, Mar. 2010.

[23] A. K. Piccoli, A. P. Alegretti, L. Schneider, P. S. Lora, and R. M. Xavier, “Expression of complement regulatory CD46 in rheumatoid arthritis,” vol. 51, no. 5, 2011.

[24] L. Zhang, J.-M. Li, X.-G. Liu, D.-X. Ma, N.-W. Hu, Y.-G. Li, W. Li, Y. Hu, S. Yu, X. Qu, M.-X. Yang, A.-L. Feng, and G.-H. Wang, “Elevated Th22 cells correlated with Th17 cells in patients with rheumatoid arthritis.,” *Journal of clinical immunology*, vol. 31, no. 4, pp. 606–14, Aug. 2011.

[25] M. Stassen, E. Schmitt, and T. Bopp, “From interleukin-9 to T helper 9 cells.,” *Annals of the New York Academy of Sciences*, vol. 1247, no. 2011, pp. 56–68, Jan. 2012.

[26] R. J. Noelle and E. C. Nowak, “Cellular sources and immune functions of interleukin-9.,” *Nature reviews. Immunology*, vol. 10, no. 10, pp. 683–7, Oct. 2010.

[27] K. Wolk, E. Witte, K. Witte, K. Warszawska, and R. Sabat, “Biology of interleukin-22.,” *Seminars in immunopathology*, vol. 32, no. 1, pp. 17–31, Mar. 2010.

[28] N. Zhang, H.-F. Pan, and D.-Q. Ye, “Th22 in inflammatory and autoimmune disease: prospects for therapeutic intervention.,” *Molecular and cellular biochemistry*, vol. 353, no. 1–2, pp. 41–6, Jul. 2011.

[29] A. L. Rankin and A. J. Caton, “autoimmune arthritis,” vol. 233, pp. 97–111, 2010.

[30] S. E. Sweeney and G. S. Firestein, “Rheumatoid arthritis: regulation of synovial inflammation,” *The International Journal of Biochemistry & Cell Biology*, vol. 36, no. 3, pp. 372–378, Mar. 2004.

[31] M. J. de Smit, E. Brouwer, a Vissink, and a J. van Winkelhoff, “Rheumatoid arthritis and periodontitis; a possible link via citrullination.,” *Anaerobe*, vol. 17, no. 4, pp. 196–200, Aug. 2011.

[32] F. a H. Cooles and J. D. Isaacs, “Pathophysiology of rheumatoid arthritis.,” *Current opinion in rheumatology*, vol. 23, no. 3, pp. 233–40, May 2011.

[33] A. M. Wasserman, “Diagnosis and management of rheumatoid arthritis.,” *American family physician*, vol. 84, no. 11, pp. 1245–52, Dec. 2011.

[34] J. Holoshitz, “The rheumatoid arthritis HLA-DRB1 shared epitope.,” *Current opinion in rheumatology*, vol. 22, no. 3, pp. 293–8, May 2010.

[35] D. L. Mattey, W. Thomson, W. E. R. Ollier, M. Batley, P. G. Davies, a K. Gough, J. Devlin, P. Prouse, D. W. James, P. L. Williams, J. Dixey, J. Winfield, N. L. Cox, G. Koduri, and a Young, “Association of DRB1 shared epitope genotypes with early mortality in rheumatoid arthritis: results of eighteen years of followup from the early rheumatoid arthritis study.,” *Arthritis and rheumatism*, vol. 56, no. 5, pp. 1408–16, May 2007.

[36] V. Ruiz-Esquide and R. Sanmartí, “Tobacco and Other Environmental Risk Factors in Rheumatoid Arthritis,” *Reumatología Clínica (English Edition)*, vol. 8, no. 6, pp. 342–350, Nov. 2012.

[37] E. W. Karlson, S.-C. Chang, J. Cui, L. B. Chibnik, P. a Fraser, I. De Vivo, and K. H. Costenbader, “Gene-environment interaction between HLA-DRB1 shared epitope and heavy cigarette smoking in predicting incident rheumatoid arthritis.,” *Annals of the rheumatic diseases*, vol. 69, no. 1, pp. 54–60, Jan. 2010.

[38] J.-M. Berthelot and B. Le Goff, “Rheumatoid arthritis and periodontal disease.,” *Joint, bone, spine : revue du rhumatisme*, vol. 77, no. 6, pp. 537–41, Dec. 2010.

[39] J. Sokolove and V. Strand, “Rheumatoid Arthritis Classification Criteria,” vol. 68, no. 3, pp. 232–238, 2010.

[40] A. V Alekseyenko, N. I. Lytkin, J. Ai, B. Ding, L. Padyukov, C. F. Aliferis, and A. Statnikov, “Causal graph-based analysis of genome-wide association data in rheumatoid arthritis.,” *Biology direct*, vol. 6, no. 1, p. 25, Jan. 2011.

[41] K. Nishimura, D. Sugiyama, Y. Kogata, G. Tsuji, and T. Nakazawa, “Annals of Internal Medicine Review Meta-analysis : Diagnostic Accuracy of Anti – Cyclic Citrullinated Peptide Antibody and Rheumatoid Factor for Rheumatoid Arthritis,” 2007.

[42] B. H. Resman-targoff and M. P. Cicero, “Aggressive Treatment of early rheumatoid Arthritis : recognizing the Window of Opportunity and Treating to Target goals © Managed Care & Healthcare Communications , LLC,” vol. 16, no. 9, pp. 249–258, 2010.

[43] J. S. Smolen and D. Aletaha, “What should be our treatment goal in rheumatoid arthritis today?,” *Clinical and experimental rheumatology*, vol. 24, no. 6 Suppl 43, p. S–7–13, 2006.

[44] M. a Karsdal, T. Woodworth, K. Henriksen, W. P. Maksymowych, H. Genant, P. Vergnaud, C. Christiansen, T. Schubert, P. Qvist, G. Schett, A. Platt, and A.-C. Bay-Jensen, “Biochemical markers of ongoing joint damage in rheumatoid arthritis--current and future applications, limitations and opportunities.,” *Arthritis research & therapy*, vol. 13, no. 2, p. 215, Jan. 2011.

[45] K. Raza, “The Michael Mason prize: early rheumatoid arthritis--the window narrows.,” *Rheumatology (Oxford, England)*, vol. 49, no. 3, pp. 406–10, Mar. 2010.

[46] S. Bosello, A. L. Fedele, G. Peluso, E. Gremese, B. Tolusso, and G. Ferraccioli, “Very early rheumatoid arthritis is the major predictor of major outcomes: clinical ACR remission and radiographic non-progression.,” *Annals of the rheumatic diseases*, vol. 70, no. 7, pp. 1292–5, Jul. 2011.

[47] V. Rantalaiho, M. Korpela, P. Hannonen, H. Kautiainen, S. Järvenpää, M. Leirisalo-Repo, M. Hakala, K. Puolakka, H. Julkunen, R. Luosujärvi, and T. Möttönen, “The good initial response to therapy with a combination of traditional disease-modifying antirheumatic drugs is sustained over time: the eleven-year results of the Finnish rheumatoid arthritis combination therapy trial.,” *Arthritis and rheumatism*, vol. 60, no. 5, pp. 1222–31, May 2009.

[48] D. Nikovski, “for Medical Diagnosis from Incomplete and Partially Correct Statistics,” vol. 12, no. 4, pp. 509–516, 2000.

[49] W. Li, M. Wang, P. Irigoyen, and P. K. Gregersen, “Inferring causal relationships among intermediate phenotypes and biomarkers : a case study of rheumatoid arthritis,” vol. 22, no. 12, pp. 1503–1507, 2006.

[50] H. Hassoun, “Book reviews.,” *American journal of veterinary research*, vol. 75, no. 1, p. 4, Jan. 2014.

[51] C. Fan, F. Gao, S. Sun, and F. Cui, “Bayesian Neural Networks and Its Application,” *2008 Fourth International Conference on Natural Computation*, pp. 446–450, 2008.

[52] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, Jun. 2008.

[53] I. Weka, G. N. U. General, P. License, N. Zealand, W. Environment, and K. Analysis, “Introduction to Weka- A Toolkit for Machine Learning.”

[54] K. Q. Guide, W. Screen, N. Status, E. Flow, A. Nodes, C. Nodes, C. Nodes, E. Nodes, Y. O. Voyage, K. Workbench, U. Guide, A. Views, W. Projects, F. Nodes,

N. Repository, N. Description, K. Gui, M. Key, W. Editor, N. Options, E. All, O. View, O. O. View, C. All, E. Custom, N. Name, C. Knime, and G. Page, “KNIME Quickstart Guide,” pp. 1–27.

[55] D. Martin, J. Procter, A. Waterhouse, S. Shehata, and G. Barton, “A manual and introductory tutorial,” no. January, 2013.

[56] F. Wikipedia, “Binary classification,” pp. 1–5, 2014.

[57] P. A. Correa, G. J. Tobón, G. Citera, J. Cadena, E. Schneeberger, J. F. Camargo, J. A. Maldonado-cocco, and J. M. Anaya, “Anticuerpos anti-CCP en artritis reumatoidea : relación con características clínicas , citocinas Th1 / Th2 y HLA-DRB1,” 2004.

[58] M. a Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. a McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, a Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, “Clustal W and Clustal X version 2.0.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 21, pp. 2947–8, Nov. 2007.

[59] S. R. Eddy, “Where did the BLOSUM62 alignment score matrix come from?,” *Nature biotechnology*, vol. 22, no. 8, pp. 1035–6, Aug. 2004.