**Universidad Autónoma de Baja California**
Instituto de Ingeniería
Maestría y Doctorado en Ciencias e Ingeniería


# Análisis de microarreglos en Dermatomiositis Juvenil por técnicas de aprendizaje de automático

*Tesis que para obtener el grado de:*

*MAESTRO EN CIENCIAS*

***Presenta***

Jesús Naranjo Avilez

**Director de Tesis:**

Dr. Félix Fernando González Navarro

Mexicali, B. C. Diciembre del 2013

# Reconocimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por apoyarme y brindarme la oportunidad de estudiar un posgrado.

A la Universidad Autónoma de Baja California (UABC) por permitir lograr mi deseo de crecer académicamente dentro de su recinto.

Al Instituto de Ingeniería y todos los académicos que participaron en mi formación por compartir sus enseñanzas y favorecer mi aprendizaje, disipando mis inquietudes.

Al Dr. Félix Fernando González Navarro por su apoyo durante el desarrollo de esta etapa en mi vida educativa tan importante.

A mi familia por respaldarme no solo a lo largo de este proyecto, sino en cada momento importante de mi vida.

A mis compañeros de posgrado por escucharme y apoyarme a lo largo de mi estancia.

A todos, y cada uno, gracias.

# Resumen

La Dermatomiositis juvenil (DMJ) es una enfermedad rara sistémica autoinmune caracterizada por una inflamación crónica del músculo esquelético y la piel, cuya causa aún permanece desconocida. Es una enfermedad de aparición estacional, siendo más frecuente en primavera y verano, coincidiendo con la mayor incidencia de infecciones virales o bacterianas. La ocurrencia es baja, de 1 a 9 casos por cada millón de niños menores de 16 años al año. Si bien es poco frecuente, el 16-20% de todas las Dermatomiositis (DM) se inician en la niñez. La edad de aparición alrededor de los 5-15 años de edad y a los 45-54 años con superioridad en el sexo femenino. Actualmente, el estudio de la DMJ y su análisis se hace desde diferentes perspectivas, como son la biología molecular, genética, bioquímica y social. Sin embargo, el análisis de la DMJ por medio de métodos computacionales, como lo es la inteligencia artificial es prácticamente nulo. Es por ello que en este trabajo de investigación se ataca este problema desde el punto de vista computacional con el objetivo de encontrar perfiles genéticos atípicos en la DMJ. Los resultados arrojaron un perfil o grupo de genes que mostraron consistencia o relevancia desde el punto de vista matemático –i.e. en la tasa de reconocimentos– hacia la DMJ. Por otra parte, mediante una búsqueda exahustiva en la literatura médica de estos genes, se encontró evidencia de una relación positiva de algunos de los genes encontrados con la DMJ.

# Contenido

# Lista de Figuras

# Lista de Tablas

# Capítulo 1

# Introducción

## 1.1 Planteamiento del problema

La Dermatomiositis (DM) es una miopatía inflamatoria idiopática, donde los principales órganos afectados son la piel y el musculo esquelético. Por otra parte; existen además formas exclusivamente cutáneas. A la fecha no sé ha podido identificar con exactitud qué factores detonan la evolución y pronóstico de los pacientes (Blanco *et al.*, 2011).

Una variante de la DM es la Dermatomiositis Juvenil (DMJ), la cual es una rara enfermedad sistémica autoinmune caracterizada por inflamación crónica del musculo esquelético y piel. Hay evidencia que sugiere que existe una asociación con exposiciones ambientales en personas genéticamente susceptibles (Mamyrova *et al.*, 2006). La DMJ comparte varios síntomas con la DM de adultos, como las erupciones cutáneas y la inflamación perivascular celular. Por otra parte se cree que difiere de los adultos de otras maneras, como la mayor frecuencia de calcificación distrófica, ulceración y vasculopatía con la angiogénesis (Mamyrova *et al.*, 2006).

La DM es potencialmente grave, en donde el dermatólogo puede facilitar el diagnóstico y contribuir en la detección de tumores y complicaciones sistemáticas prematuras. La mayoría de los pacientes tienen un buen diagnóstico, aunque requieren periodos de tratamiento prolongados. Los casos más difíciles son los

asociados con neoplasias y cuando existen problemas cardiopulmonares (Blanco *et al.*, 2011).

Como en la mayoría de las enfermedades autoinmunes la causa precisa de la DMJ es desconocida. Su origen es probablemente multifactorial y es considerada una enfermedad no hereditaria, aunque se ha detectado influencia de factores ambientales en su aparición. Se piensa que, como en otras enfermedades autoinmunes, el contacto con algunos microorganismos puede desencadenar una respuesta anormal del sistema inmune. Dado que se desconoce la causa la enfermedad no se puede realizar ninguna recomendación de tipo preventivo (Blanco *et al.*, 2011).

Actualmente, el estudio de la DMJ y su análisis se hace desde diferentes perspectivas, como son la biología molecular, genética, bioquímica y social. Sin embargo, el análisis de la DMJ por medio de métodos computacionales, como lo es la inteligencia artificial es prácticamente nulo, a diferencia de otros problemas médicos como lo es el cáncer.

La Inteligencia Artificial (IA) constituye uno de los campos interdisciplinarios donde convergen muchas ciencias. La aparición de las computadoras y la elaboración de las teorías de la computación, la información y el control, proporcionaron los soportes experimentales y teóricos para la investigación en el área de la IA (Honavar, 2006). Se considera esencial entre sus líneas estratégicas de investigación, aquellas que tienden a analizar y estudiar cierto tipo de problemas médicos, que impacten en el diagnóstico y el tratamiento de diversas enfermedades (Trias Capella, 1993).

Conjuntamente los continuos avances tecnológicos en la Biología Molecular, unidos al desarrollo informático, han aumentado las posibilidades de conocer el funcionamiento de los seres vivos a nivel molecular y celular. Es necesario unificar toda esta información para alcanzar un cuadro completo de la biología de la célula para comprender cómo se alteran distintos procesos en distintas enfermedades.

Por eso, hoy en día es difícil entender la investigación en el área de las enfermedades genéticas humanas sin la Bioinformática (Lourdes, 2009).

La Bioinformática ha ido evolucionando para ocuparse cada vez con mayor profundidad del análisis e interpretación de los distintos tipos de datos. Las bases de datos utilizadas en biología molecular son archivos de datos que provienen de diferentes áreas almacenados de modo eficaz y uniforme y de uso público para la comunidad científica. Existen cientos de bases de datos, por el tipo de información se pueden distinguir: bibliográficas, taxonómicas, de nucleótidos, genómicas, de proteínas, de microarreglos y otras (Lourdes, 2009).

Actualmente la DMJ es un tema abandonado por la comunidad científica de las ciencias de la computación. Al momento, son pocos o nulos los estudios de la DMJ usando técnicas inteligentes en la búsqueda de perfiles genéticos o biomarcadores, es por ellos que la motivación de este trabajo de investigación, es contribuir al avance científico en el entendimiento de la DMJ, que permita en un futuro determinar las causas de esta rara enfermedad.

## 1.2 Objetivo General

Utilizar de modelos computacionales basados en técnicas de aprendizaje de máquina para identificar perfiles genéticos atípicos en la DMJ.

### 1.2.1 Objetivos Específicos

1. Determinar los genes que presentan comportamiento atípico en muestras de pacientes con DMJ.

2. Establecer los algoritmos computacionales de clasificación que mejor respondan al problema separación entre clases en datos de expresión genética de la DMJ.

3. Realizar una asociación entre los perfiles genéticos encontrados y la evidencia médica-biológica existente..

Esta tesis esta organizada de la siguiente manera: el capítulo 1 consiste en una introducción al tema de tesis, el planteamiento del problema y los objetivos planteados; en el capítulo 2 se expone el marco teórico, en el cual se explica la dermatomiositis juvenil, se tocan temas como su definición, factores de riesgo, diagnóstico y tratamiento, expresión genética, tecnología de microarreglos y sus métodos de análisis, técnicas de aprendizaje de máquina, algoritmos de selección de atributos y de clasificación, y métodos de validación de resultados; en el capítulo 3 de materiales y métodos, se da la metodología utilizada para el desarrollo de la investigación, una descripción de datos, las técnicas utilizados y el procedimiento experimental seguido; en el capítulo 4 se muestran los resultados y por último en el capítulo 5 se aprecian las conclusiones y trabajo futuro.

# Capítulo 2

# Marco Teórico

En este capítulo, se expondrán detalles a profundidad de lo que es la DMJ, como su definición, factores de riesgo, diagnóstico y tratamiento y el ambiente de aprendizaje de máquina como soporte general de los experimentos realizados.

## 2.1 La Dermatomiositis Juvenil

### 2.1.1 Definición

La DMJ es una enfermedad rara sistémica autoinmune caracterizada por una inflamación crónica del músculo esquelético y la piel, cuya causa aún permanece desconocida. En estudios recientes –e.g. Mamyrova *et al.* (2006), Blanco *et al.* (2011), Blaszczyk *et al.* (2000), Reed & Mason (2005)– sugieren que existe una relación de la exposición al medio ambiente de personas que son mayormente susceptibles.

Al inicio, la DMJ comparte ciertas características con la DM adulta, como las erupciones cutáneas y la inflamación perivascular celular. Pero la forma juvenil se define como la aparición de la enfermedad antes de los 18 años, y se diferencía de la forma del adulto por el predominio de lesiones vasculopatícas y calcinosis cutis (Kannee, 2010), así como una mayor frecuencia de calcificación distrófica, ulceración y la vasculopatía(Mamyrova *et al.*, 2006).

Su incidencia y prevalencia es difícil de precisar debido a la rareza de la misma, la ocurrencia sería de 1 a 9 casos por cada millón de niños menores de 16 años al año. Si bien es poco frecuente, el 16-20% de todas las DM se inician en la niñez. La edad de aparición es bimodal con un pico a los 5-15 años de edad y otro a los 45-54 años con superioridad en el sexo femenino (Feber, 2005; J. & R., 1995; MP & de Oca F., 2002).

### 2.1.2 Factores de riesgo

Aún no se tiene en claro cuáles son las razones de esta extraña enfermedad. Estudios sugieren una patogenia autoinmune y una predisposición genética asociándose a ciertos antígenos de histocompatibilidad. Cuando se forman anticuerpos, éstos reaccionan frente a los capilares dentro del músculo ocasionando una vasculitis con isquemia y lesión de las células del mismo (Kannee, 2010).

Las lesiones musculares se traducen en la elevación de las enzimas musculares: creatininfosfoquinasa (CPK), lactatodeshidrogenasa (LDH), alanino aminotransferasa (TGP), aspartato aminotransferasa (TGO) y aldolasa. La CPK es el más sensible (J. & R., 1995; MP & de Oca F., 2002; RM & HR., 1989).

Se han encontrado en algunos casos –poco frecuentes en niños– anticuerpos antisintetasa. El cuadro clínico asociado a estos anticuerpos es una miositis más resistente al tratamiento, con enfermedad intersticial pulmonar (50-70%), artritis (90%) y el fenómeno de Raynaud (60%). Otro anticuerpo detectado es el anti-PM/Scl. Éste se encuentra en un síndrome de solapamiento conocido como escleromiositis que asocia síntomas cutáneos de esclerodermia (Feber, 2005).

La DMJ es una enfermedad episódica, de aparición estacional, siendo más frecuente en primavera y verano, coincidiendo con la mayor incidencia de infecciones virales o bacterianas. Se han implicado como desencadenantes de la enfermedad agentes infecciosos como enterovirus (coxsackievirus B), Toxoplasma gondii, estreptococo del grupo A, herpes simple, hepatitis A, influenzae A y B, parainfluenzae 1 y 3, micoplasma, sarampión, citomegalovirus (MP & de Oca F.,

2002; S. & WL., 1975). Se cree que existe una similitud molecular entre la miosina del músculo del estriado y un componente del agente infeccioso. Otro mecanismo propuesto es la participación de componentes bacterianos o virales como superantígenos (MP & de Oca F., 2002), (LM., 2004).

### 2.1.3 Diagnóstico y tratamiento

En 1975 Bohan y Peter establecieron una clasificación de las DM. Se han utilizado clasificaciones nuevas más amplias que incluyen nuevas entidades, como se ve en la siguiente tabla 2.1.

| **Dermatomiositis (DM)** |
|---|
| **Adulto** |
| DM clásica* |
| DM clásica |
| DM clásica asociada a malignidad |
| DM clásica asociada a enfermedad del tejido conectivo |
| DM clínicamente amiopática |
| DM amiopática |
| DM hipomiopática |
| **Juvenil** |
| DM clásica |
| DM clínicamente amiopática |
| DM amiopática |
| DM hipomiopática |
| **Polimiositis (PM)** |
| PM |
| PM asociada a malignidad |
| PM asociada a enfermedad del tejido conectivo |
| Miositis por cuerpos de inclusión |
| Miositis por fármacos |

Tabla 2.1: Clasificación de las miopatías inflamatorias idiopáticas

El término DM clásica se refiere a la presencia clínica de lesiones cutánea y debilidad muscular típicas. La DM amiopática se refiere a los pacientes que solo presentan lesiones cutáneas y no debilidad muscular. La DM amiopática se considera provisional cuando no existe afectación muscular en al menos 6 meses.

Se considera definitiva si el periodo es más de 2 años. Los criterios de exclusión incluyen la ausencia de tratamiento con inmunosupresores u otros fármacos capaces de producir lesiones cutáneas similares a la DM (Blanco *et al.*, 2011).

Para el diagnóstico se utilizan los criterios de Bohan y Peter, que incluyen dificultad para deglutir, dolor, rigidez y debilidad muscular, párpados superiores de color púrpura o violeta, erupción cutánea de color rojo o púrpura y dificultad respiratoria (Blanco *et al.*, 2011). En las imágenes 2.1 y 2.2 se pueden observar algunos síntomas (Blaszczyk *et al.*, 2000).

Figura 2.1: Síntomas de menor masculino con erupción cutánea en pecho y brazos.

Figura 2.2: Síntomas de menor femenina con parpados color púrpura.

La debilidad muscular llega de manera repentina o bien puede aparecer lentamente en semanas o meses, el enfermo presenta dificultad en levantar los brazos por encima de la cabeza, ponerse de pie estando sentado, así como al subir escaleras. Las erupciones aparecen en la cara, nudillos, cuello, hombros, la parte superior del tórax y espalda. Cuando se presenta fiebre es moderada y viene acompañada de malestar general y anorexia (Blaszczyk *et al.*, 2000).

Cuando comienza de un modo agudo y brusco con empeoramiento rápido hay mayor riesgo de muerte durante el primer año que las de comienzo lento que es

más frecuente. En un 10% de los casos se presenta una afectación orofaríngea con disfagia y disfonía (voz nasal), siendo esto último muy peligroso ya que podrían darse atragantamientos por aspiración y bloqueo de las vías aéreas superiores (Mamyrova *et al.*, 2006).

Para diagnosticar al paciente el médico lleva acabo un examen físico el cual puede comprender pruebas de sangre para medir los niveles de creatinfosfocinasa y aldolasa, ECG, electromiografía, resonancia magnética o biopsia del músculo. El trazado electromiográfico es diagnóstico de la DM pero es una exploración difícil y necesita sedación, por lo que es muy poco práctica en niños. La desventaja de la biopsia muscular es que existirían resultados falsamente negativos. La resonancia magnética con técnicas de supresión de grasa T2 es una prueba muy útil en el diagnóstico de las miositis inflamatorias. Al observarse una mejor imagen, debido a la acumulación de agua en los músculos lo que hace posible el identificar los músculos afectados de los no afectados, siendo posteriormente más fácil de realizar (Mamyrova *et al.*, 2006).

Se da un diagnóstico definitivo cuando se presentan más de tres criterios con erupción cutánea, un diagnóstico probable si se presentan dos criterios con erupción cutánea, un diagnóstico posible cuando se presenta un criterio con erupción cutánea (Blaszczyk *et al.*, 2000). En los últimos años la tendencia es establecer una clasificación según la presencia de algunos anticuerpos. Algunos anticuerpos parecen definir grupos homogéneos de pacientes que presentan características epidemiológicas. Recientemente se ha demostrado que los pacientes con DM asociada a cáncer presentan los anticuerpos anti p-155 y anti p-155/140 (Blaszczyk *et al.*, 2000).

El tratamiento de la DM se apoya en la utilización anticipada y a dosis adecuadas de corticoides sistémicos. Las dosis iniciales deben ser entre moderadas y altas, 1 y 1.5 mg/kg/día de prednisona. Cuando se obtiene una mejoría estable con descenso en los niveles de creatina fosfoquinasa debe iniciarse un descenso lento de los corticoides hasta alcanzar una dosis mínima de mantenimiento y eventualmente la eliminación si la recuperación es total (Blaszczyk *et al.*, 2000).

### 2.1.4 Estudios de expresión genética en la DMJ

En esta sección se presentan algunos estudios de expresión genética que al momento se ha realizado sobre la DMJ.

#### Duración de la inflamación crónica altera la expresión de genes en el músculo de las niñas no tratadas con dermatomiositis juvenil

En este estudio –ver Chen *et al.* (2008)– se evaluó el impacto de la duración de la inflamación crónica en la expresión genética en biopsias de músculo esquelético de niños no tratados con DMJ. Además se buscó identificar los genes y procesos biológicos asociados con la progresión de la enfermedad. Los datos de perfiles de expresión de 16 niñas con síntomas de la DMJ mayor a 2 meses se compararon con tres niñas con síntomas de menos de 2 meses.

Fueron 79 genes expresados diferencialmente entre los grupos con larga o corta duración de la enfermedad no tratada. Los genes implicados en la respuesta inmune y la remodelación vascular se expresaron en un nivel más alto en biopsias musculares de los niños con mayor o igual a 2 meses de los síntomas, mientras que los genes involucrados en las respuestas al estrés y la rotación de proteína se expresa en un nivel inferior. Entre los 79 genes, la expresión de nueve genes mostraron una regresión lineal significativa con la duración de la enfermedad no tratada.

#### Análisis auto antígeno de micro arreglos de sueros de pacientes con DMJ: Asociación con un SLE tipo I firma de interferón

Balboni *et al.* (2007) investigó características de los anticuerpos en niños con DMJ, utilizando micro arreglos de auto antígeno para lograr una mejor clasificación de enfermedades. Se vio el caso de 37 niños con DMJ con resultado negativo para anticuerpos de la miositis, y niños de 10 años con un control emparejado del sexo usando el micro arreglo auto antígeno-1040 asociado con varias enfermedades de tejido conectivo.

El programa informá Significado de Análisis de Micro arreglos (SAM) se utilizó para determinar diferencias en la expresión de auto anticuerpos en muestras con DMJ y de control y en los análisis de subgrupos. SAM no identificó diferencias notables en la expresión de anticuerpos entre DMJ y un control, o un subgrupo de análisis de anticuerpo antinuclear de pacientes positivos frente a los negativos.

**Expresión génica en DQA1 * 501$^{+}$ Los niños con dermatomiositis intratable: un nuevo modelo de la patogénesis**

En esta investigación –ver Tezak *et al.* (2011)– se entrevistaron a los padres o tutores de 375 niños con reciente diagnóstico de DMJ y se encontró que el primer síntoma definitorio son las enfermedades sistémicas. Con el propósito de describir la patogenia de la DMJ, se utilizaron micro arreglos tipo Affymetrix, para analizar los genes expresados en los músculos de los niños no tratados con DMJ. Se presentaron los datos de perfil de expresión de cuatro niños con DMJ.

Se obtuvieron biopsias musculares de cuatro niñas de raza blanca (edad 5-16 años) que cumplían los criterios para la DMJ definitiva, ellas fueron negativas para miositis específicas, así como la miositis asociada. Esta selección se hizo para asegurar que los cuatro de estos niños probablemente compartían las mismas características de la enfermedad. Inmediatamente antes de la anestesia para la biopsia del músculo, se extrajo sangre de cada niño para su posterior análisis de los sueros.

El modelo expuesto en este artículo puede predecir que la mayoría de los auto antígenos representan proteínas inducidas durante perfiles de regeneración o de compensación, por lo que los éstos puede ser una consecuencia de un efecto de conflicto en perfiles de expresión sin una relación directa con el agente causante.

**Expresión de TNF-$\alpha$ por las fibras musculares en las biopsias de niños con DMJ no tratada: la asociación con el alelo TNF-308A**

En este estudio –ver Fedczyna *et al.* (2001)– se investigó la asociación entre la expresión de TNF-$\alpha$ [1] por las fibras musculares, la actividad de la enfermedad, duración de la enfermedad sin tratamiento y el polimorfismo TNF-308.

Los niños no tratados con DMJ que tenían el alelo TNF-308A tuvieron un aumento del número de fibras musculares teñidas TNF que los niños con el alelo TNF-308G. No hubo asociación con la actividad de la enfermedad o la duración de la enfermedad no tratada.

Se especula que la producción de las fibras musculares del TNF-$\alpha$ ofrece un microambiente en el que TNF actúa sinérgicamente con otros mediadores para prolongar el daño de las fibras musculares (Fedczyna *et al.*, 2001).

En la actualidad, el uso de la tecnología de expresión genética con microarreglos no es nuevo. Gran cantidad de enfermedades de diversa índole han sido analizadas por este medio. En la tabla 2.2 se muestran algunos ejemplos y referencias de dichos estudios:

## 2.2 Expresión genética

En esta sección se hablará sobre la tecnología de expresión genética la cual transforma la información codificada por los ácidos nucleicos en las proteínas necesarias para su desarrollo y funcionamiento.

### 2.2.1 Tecnología de microarreglos

El Microarreglo es un pequeño dispositivo genético que permite la exploración genómica con precisión y velocidad sin precedentes en la biología. Son chips de vidrio que contienen decenas de miles de genes que son usados para examinar

[1] Del inglés tumor necrosis factor alpha. Es miembro de un grupo de citocinas que estimulan la fase aguda de la reacción inflamatoria

| Enfermedad | Estudio |
|---|---|
| Esclerosis Lateral Amiotrófica | *Gene Subset Selection from Microarray Data in Amyotrophic Lateral Sclerosis Disease Classification* (Astengo, 2012). |
| Cáncer de Colon | *Cancer Diagnosis with DNA Microarrays (Knudsen, 2006)* |
| Leucemia | *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring* (Golub *et al.*, 1999) |
| Cáncer de Pulmón | *Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lunga Cancer and Mesothelioma* (Gordon *et al.*, 2002) |
| Cáncer de Próstata | *Gene expression correlates of clinical prostate cancer behavior (Singh* et al., *2002)* |
| Cáncer de Mama | *Gene expression profiling predicts clinical outcome of breast cancer (van't Veer* et al., *2002)* |

Tabla 2.2: Ejemplos de estudios genéticos con microarreglos

muestras fluorescentes preparadas por etiquetado de Ácido Ribonucleico Mensajero (mRNA en inglés) de las células, tejidos, y otros recursos biológicos. La expresión de genes es el proceso celular por el cual la información genética pasa del gen de mRNA a proteína (Schena, 2003).

Las proteínas están involucradas en la mayoría de las funciones principales (por ejemplo, creación de anticuerpos, el movimiento muscular, las tareas hormonales, las funciones estructurales y de transporte, la señalización celular) que tienen lugar en prácticamente todas las células. Las proteínas son representadas por secuencias lineales de aminoácidos, por lo general de 100 a 1000. Las proteínas no son ensambladas por ellas mismas, se construyen sobre la base de planos almacenados en el Ácido Desoxirribonucleico (DNA por sus siglas en inglés). Los genes codificados en el ADN se transcriben en un primer actor intermedio llamado ARN pre-mensajero (Knudsen, 2006).

El proceso de síntesis de una proteína a partir de un molde de mRNA se conoce como traducción. Todas las moléculas de ARN consisten en una secuencia de nucleótidos: Adenina (A), Citosina (C), Guanina (G) y Uracilo (U). Hay una correspondencia uno a uno entre las moléculas de de la proteína sintetizada y las moléculas de mRNA. Así, este proceso de creación de proteínas puede ser determinada mediante la cuantificación de la abundancia de mRNA. Los ensayos de microarreglos consisten en la medición de esta cantidad dando una estimación aproximada del nivel de expresión de los genes implicados (Simon *et al.*, 2003).

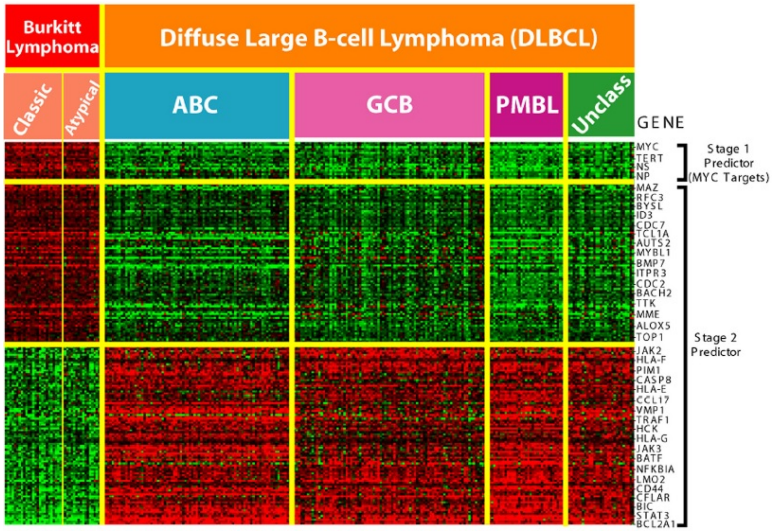


Figura 2.3: Análisis de microarreglo de ADN de linfoma de Burkitt y células grandes B de linfoma difuso que muestra diferencias en los patrones de expresión genética. Los colores indican los niveles de expresión, el verde indica los genes que están sobre-expresados en células normales en comparación con células de linfoma y el rojo indica genes que se sobre-expresan en células de linfoma en comparación con las células normales.

La base fundamental de la tecnología de microarreglos de ADN es el proceso de *hibridación*. Dos cadenas de ADN se hibridizan si son complementarios entre sí. La regla complementaria de Watson-Crick refleja que la Adenina (A) se une a la Timina (T) y la Citosina (C) se une a la Guanina (G). Una o ambas cadenas de los híbridos de ADN pueden ser sustituidas por hibridación del ARN y seguirá

ocurriendo mientras se complementen (Knudsen, 2006).

Si el polinucleótido está etiquetado radiactivamente, la hibridación puede ser visualizada en una película fotográfica que es sensible a la radiación. La cantidad de radiación capturada en la película fotográfica depende de la cantidad de mRNA. La molécula de mRNA es relativamente frágil y fácilmente se puede dividir. Así, los investigadores comúnmente manipulan una forma de ADN que posee las bases complementarias del mRNA, mientras que existen en un estado más estable. Esta forma de ADN, conocido como ADN complementario (cDNA en inglés), se crea directamente a partir del mRNA de la muestra a través de un procedimiento conocido como transcripción inversa, que consiste en la transcripción de secuencias complementarias de base genética de ARN a ADN (McLachlan *et al.*, 2005).

Un microarreglo consiste en una superficie sólida en las que cadenas de polinucleótidos se han unido en posiciones específicas. Estos polinucleótidos fijos son conocidos como sondas. Las moléculas de cDNA marcado se unen por la hibridación de las sondas en el gen de la matriz. Después de un tiempo suficiente para la reacción de hibridación, cada sonda en el microarreglo debe estar sujeta a una cantidad de cDNA etiquetada que es proporcional al nivel de expresión de los genes representados por esa sonda. Esta cantidad es leída iluminando la superficie sólida con luz láser y midiendo la intensidad de la fluorescencia en cada sonda en la matriz con un detector, por lo general un microscopio fluorescente. La salida es un archivo de imagen que luego es procesada por algoritmos de análisis de imagenes para dar una lectura numérica del nivel de expresión (Simon *et al.*, 2003).

Actualmente, hay tres grandes campos que la investigación científica sigue en el uso de conjuntos de datos de microarreglos:

1. La *Comparación de Clases*, cuyo objetivo principal es encontrar los genes que son expresados diferencialmente entre algunas clases. Su valor científico


Figura 2.4: Dos microarreglos Affimetrix. Típicamente, un microarreglo puede almacenar hasta 40,000 genes.

y médico reside en la comprensión de las diferencias biológicas entre enfermedades a partir de la determinación de la relación entre las funciones biológicas conocidas y desconocidas y genes expresados atípicamente. Las técnicas de análisis más utilizadas son aquellas que caen en modelos estadísticos como las pruebas t y las de permutaciones.

2. El *Descubrimiento de Clases* se centra en los métodos para encontrar grupos de genes que muestran ciertos comportamientos comunes o, en general para encontrar patrones en los perfiles de expresión de microarreglos. Esta segunda orientación en el análisis de microarreglos es especialmente etiquetado como no supervisado, ya que no hay información de clase que se utilizada. Como su definición lo implica, el análisis de conglomerados es la herramienta más solicitada (Simon *et al.*, 2003).

3. La *Predicción de Clases* consiste en el uso de varios métodos y técnicas del aprendizaje de máquina. Su objetivo principal es desarrollar predictores multivariados de clase para asignar inteligentemente los casos para especificaciones de clases (su condición biológica). En un caso conjunto de datos típico, existe un gran desequilibrio entre el número de variables o genes y

los casos, esta situación viene a ser un asunto que debe ser considerado como un primer paso a analyzar. Por lo tanto, los métodos de reducción de dimensionalidad entran a escena con el fin de facilitar el desarrollo de modelos predictivos más eficaces.

La regulación de genes provee un selectiva evolución y una reconstrucción celular hasta donde se necesita, y permitiendo a los organismos adaptarse a una gran cantidad en diferentes condiciones ambientales los cuales son expuestos durante su vida.

### 2.2.2 Métodos de Análisis

Uno de los objetivos más importantes en estudios de microarreglos es el identificar los genes que están expresados diferencialmente entre clases preestablecidas. La identificación de genes expresados diferencialmente conducen a una mejor comprensión de las diferencias biológicas entre las clases (Simon *et al.*, 2003).

Comenzamos considerando el caso de dos clases, en cuyo caso deseamos descubrir si un gen tiene mayor expresión en una clase, en comparación con el otro. Los análisis de datos de expresión suelen utilizar niveles logarítmicamente transformados; utilizando logaritmos base 2 para la mayoría de los análisis. Supongamos que se tiene disponible $J_1$ muestras de la clase 1 y $J_2$ de la clase 2. Si consideramos la posibilidad de un gen particular, con las expresiones de la clase 1 dados por $x_{11}, x_{12}, \ldots, x_1 j_1$, y en la clase 2 dada por $x_{21}, x_{22}, \ldots, x_2 j_2$, resumiendo las expresiones promedio de las clases 1 y 2 por los medios de los valores de la clase 1 y clase 2, $\bar{x}_1$ y $\bar{x}_2$. Si $\bar{x}_1$ es mayor (menor) que $\bar{x}_2$, esto sugeriría que el gen se expresa más (menos) en la clase 1 que en la clase 2.

Hay varias cosas a tener en cuenta acerca de este tipo de comparación. En primer lugar, esta comparación tiene sentido incluso si cada valor de la expresión es relativa a un patrón de referencia (como sería generalmente el caso para las microarreglos de cDNA), a condición de que el mismo estándar de referencia se utilizara para las muestras de ambas clases. En segundo lugar, suponiendo que el procesamiento de los microarreglos se realiza de la misma manera para ambas

clases de muestras, no necesita preocuparse demasiado acerca de las fuentes de sesgo sistemático que están presentes en los microarreglos de ambas clases; estos sesgos tienden a cancelarse cuando se comparan las expresiones entre las clases (Simon *et al.*, 2003).

### Prueba-*t*

Si hay una gran población de muestras de clase 1 y de clase 2, y las distribuciones de las expresiones de la clase 1 y clase 2 son las mismas en la población. Esto se le conoce como la hipótesis nula (Simon *et al.*, 2003). La teoría estadística permite estimar la probabilidad de que podría haber una diferencia tan grande como la observada. Existen diferentes métodos para hacer esto. El más utilizado consiste es el estadístico-t:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{J_1} + \frac{2}{J_1})}}, \tag{2.1}$$

donde:

$$s_p^2 = \frac{(J_1 - 1)s_1^2 + (J_2 - 1)s_2^2}{J_1 + J_2 - 2} \tag{2.2}$$

y para $i$=1, 2,

$$s_i^2 = \frac{1}{J_i - 1}\sum_{j=1}^{J_i}(x_{ij} - \bar{x}_i)^2. \tag{2.3}$$

El estimador de la varianza $s_p^2$ estima la variabilidad intra-clase (agrupados) de las expresiones de genes, su raíz cuadrada es conocida como la desviación típica en clase. Una alternativa del estadístico-$t$ utiliza $\sqrt{\frac{s_1^2}{J_1} + \frac{s_2^2}{J_2}}$ en el denominador. Esta alternativa del estadístico-$t$ se conoce como el estadistico-$t$ mediante la fórmula de varianzas separadas, mientras que al primero (formula 2.1) se conoce la fórmula de varianza conjunta.

### Prueba de permutación 76

El cálculo del valor de $p$ del estadístico-$t$ asume implícitamente que la distribución estadística del numerador es aproximadamente normal. Debido a que la distribución normal puede ser completamente caracterizada por sus parámetros de

desviación estándar y media, las pruebas-t son ejemplos de lo que se conoce como pruebas paramétricas. La aproximación de la distribución normal quizá no sea buena si uno está interesado en muy pequeños valores de $p$. Un enfoque alternativo para la estimación de los valores de $p$ que no dependen de la distribución normal aproximada es a través de las pruebas de permutación (Simon *et al.*, 2003).

El valor de $p$ de la prueba-$t$ de permutación se estima:

$$p - valor = \frac{1 + \#depermutacionesaleatoriasdonde|t*| \geqslant |t|}{1 + \#depermutacionesaleatorias} \tag{2.4}$$

Con muestras pequeñas y en algunos casos especiales, se puede enumerar todo:

$$\binom{J_1 + J_2}{J_1} \equiv \frac{(J_1 + J_2)!}{(J_1!J_2!)} \tag{2.5}$$

### Prueba de la suma de rangos de Wilcoxon

Una prueba de permutación alternativa a la prueba-$t$ de permutación es la prueba de la suma de rangos de Wilcoxon. Los niveles originales $J_1 + J_2$ de expresión son reemplazados por sus filas. Es decir, al valor de x más grande del conjunto combinado de valores $J_1 + J_2$ se le asigna el valor 1; el segundo más grande, el valor 2, y este ranking continua hasta que se da el valor más pequeño $J_1 + J_2$.

A continuación se calcula la suma de los rangos para únicamente esas observaciones en la clase 1. Esta suma se compara con los valores presentados especialmente construidos para calcular el valor de $p$ (valores grandes o pequeños de la suma indican las diferencias de clase). Las ventajas de la prueba de suma de rangos de Wilcoxon sobre la prueba-t de permutación son que las permutaciones no tienen que ser construidas (debido a que el valor de $p$ depende sólo de la suma de rangos de $J_1$ y $J_2$) y, en la medida en que se utilizan las filas, la prueba de la suma será relativamente insensible a los valores extremadamente grandes o pequeños.

Los valores de $p$ de pruebas $t$ paramétricas por lo general son pequeñas, incluso con un pequeño número de muestras, pero estas pruebas no son muy eficaces en la detección de las diferencias de clase reales debido a que las estimaciones de la variabilidad dentro de la clase son pobres.

### Más de dos clases

Para algunas aplicaciones, no habrá más de dos clases de muestras para comparar. Por ejemplo, los especímenes pueden estar disponibles a partir de tres diferentes subtipos histológicos de cáncer. En particular, la hipótesis nula es la distribución de la expresión génetica es la misma para todas las clases. La hipótesis alternativa es que al menos una de las clases tiene una distribución de la expresión genética es diferente de las otras clases. La estadística análoga al estadístico-$t$ adecuado con más de 2 clases es el estadístico-$F$:

$$F = \frac{[J_1(\bar{x}_1 - \bar{x})^2 + J_2(\bar{x}_2 - \bar{x})^2 + \ldots + J_i(\bar{x}_i - \bar{x})^2]/(I-1)}{s_p^2}, \tag{2.6}$$

donde:

$$s_p^2 = \frac{1}{J_1 + J_2 + \ldots + J_I - I} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (\bar{x}_{ij} - \bar{x}_i)^2 \tag{2.7}$$

y

$$\bar{x} = \frac{1}{J_1 + J_2 + \ldots + J_I} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \bar{x}_{ij} \tag{2.8}$$

El estadístico-$F$ se convierte en un valor de $p$, que es aproximadamente correcto siempre que las medias dentro de la clase $\bar{x}_i$ es aproximadamente una distribución normal. Para I = 2, el valor de $p$ obtenido del estadístico-$F$ es idéntico al valor de $p$ obtenido de la varianza conjunta del estadístico-$t$. Análogamente a la prueba-$t$ de permutación, se calcula el valor de $p$ para la prueba de permutación-$F$: Las etiquetas de clase se permutan al azar para que $J_1$ de las muestras se etiquetan temporalmente clase 1, $J_2$ de los muestras restantes se etiquetan temporalmente como clase 2, y así sucesivamente. El numerador del estadístico-$F$ se calcula

sobre esta base de datos temporal. Entonces fórmulas análogas anteriores se utilizan para calcular el valor de $p$. Alternativamente, las filas de los valores de x se pueden utilizar en el análisis de permutación, produciendo el test de Kruskal-Wallis (Simon *et al.*, 2003).

## 2.3 Métodos de aprendizaje de máquina

El aprendizaje de máquina es programar computadoras para optimizar el criterio de desempeño usando datos de ejemplo o experiencias del pasado. Definimos un modelo para algunos parámetros, y el aprendizaje es la ejecución del programa en la computadora para optimizar esos parámetros usando datos de entrenamiento. El modelo podría ser predictivo para hacer pronósticos a futuro, o descriptivo para obtener conocimiento de los datos o ambos.

El aprendizaje de máquina utiliza teorías de la estadística para construir modelos matemáticos, porque la teoría central está haciendo deducción a partir de una muestra, el rol de la computación es doble ya que primero se debe entrenar; para eso necesitamos algoritmos eficientes para resolver el problema de optimización, también para almacenar y procesar la gran cantidad de datos que tenemos generalmente, y segundo una vez que el modelo es aprendido, su representación y solución algorítmica por deducción deben ser eficientes. En algunas aplicaciones, la eficiencia o deducción del algoritmo, el espacio y complejidad del tiempo computacional puede ser tan importante como lo es su capacidad de predicción (Alpaydin, 2004).

El aprendizaje de máquina tiene como objetivo generar expresiones de clasificación lo suficientemente simples para ser entendidas fácilmente por el ser humano. Deben imitar el razonamiento humano lo suficiente como para proporcionar información sobre el proceso de decisión. Al igual que los enfoques estadísticos, el conocimiento de fondo puede ser explotado en el desarrollo, pero la operación se asume sin intervención humana (Michie *et al.*, 1994).

### 2.3.1 Algoritmos de selección de atributos

El rápido avance en la tecnología de microarreglos de expresión de genes ha proporcionado a científicos la oportunidad de observar las complicadas relaciones entre varios genes en un genoma por medición simultánea de los niveles de expresión de decenas de miles de genes en experimentos masivos. El análisis de gran dimensionalidad de datos genómicos presenta oportunidades y retos para las áreas de minería de datos como el gene clustering, descubrimiento de clases y clasificación. Un típico conjunto de datos de un microarreglo cuenta con tres características, la primera una alta dimensionalidad (cientos de miles) de genes, la segunda es una muy baja cantidad de muestras (decenas) y tercera hay abundancia de redundancia entre genes (Liu & Motoda, 2008).

Se define la selección de atributos como el proceso de buscar un subconjunto de características del conjunto original formando patrones de un conjunto de datos, de acuerdo al criterio definido de la selección de atributos (criterios de bondad de atributos). Se considera a la técnica de selección de atributos como el proceso de encontrar el mejor subconjunto de características $X_{opt}$ del conjunto original de patrones de características, de acuerdo al criterio de bondad de atributos definido $J_{feature}$ $(X_{feature\_subset})$(Kurgan, 2007).

La teoría del aprendizaje computacional sugiere que la búsqueda en el espacio de soluciones es exponencialmente larga. Con muestras muy limitadas y un largo conjunto de genes, guían a muchas hipótesis estadísticas relevantes que son igualmente válidas interpretando el mismo conjunto. Por lo tanto, seleccionar una pequeña cantidad de genes discriminantes es esencial para una clasificación exitosa (Liu & Motoda, 2008).

Desde un punto de vista práctico, la selección de un pequeño subconjunto de genes discriminantes a menudo ayuda a identificar genes que son relevantes para la causa o consecuencia de una enfermedad o pueden ser usados como biomarcadores para diagnósticos de enfermedades, midiendo la toxicología de una medicina. Un conjunto de genes compacto es deseable para los biólogos por el elevado gasto

asociado con el seguimiento de validación biológica o clínica de genes seleccionados.

Son 2 grandes enfoques, los utilizados como métodos de selección de atributos, los métodos wrapper y filter –ver Figura 2.5. El primero de ellos, conocidos como métodos de lazo cerrado (Closed loop methods) o Wrapper, usan la exactitud de precisión de un predeterminado algoritmo de aprendizaje para determinar la bondad del subconjunto seleccionado (Liu & Motoda, 2008). El alto costo computacional al manejar una gran cantidad de atributos, y el subconjunto seleccionado es dependiente del algoritmo de aprendizaje (Kurgan, 2007).

El otro gran enfoque, integra los llamados métodos de lazo abierto (open loop methods) o también llamados métodos Filter –ver Figura 2.6–, son su mayoría son basados en selección de atributos a través del uso de un criterio de separabilidad entre clases. Estos métodos no consideran el efecto de los atributos seleccionados en el rendimiento del proceso entero del algoritmo (clasificador).

Los métodos Filter se basan en características generales de los datos de entrenamiento para elegir características. Los métodos tradicionales en la selección de genes están dentro de este segundo enfoque, y a menudo evalúan genes en aislamiento sin considerar la correlación gen a gen. Estos métodos son computacionalmente eficientes ya que su complejidad es de naturaleza lineal O(N). Sin embargo, no son capaces de remover genes redundantes (Taylor & Francis Group, 2008).

Estos métodos seleccionan, por ejemplo, aquellas características por las cuales el resultado del conjunto reducido tiene un máximo de separabilidad entre clases, usualmente definido de entre clases y covarianza entre clases (o matriz scatter) y sus combinaciones. La ignorancia del efecto del atributo seleccionado en el rendimiento del predictor es un punto débil de los métodos de lazo abierto. Sin embargo, estos métodos son computacionalmente menos pesados.

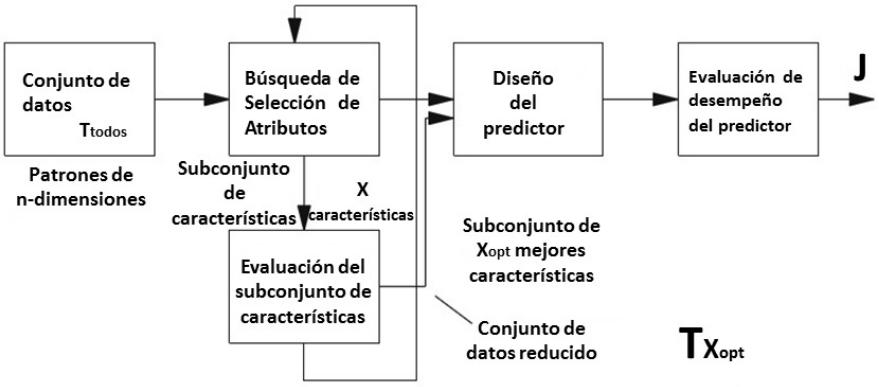


Figura 2.5: métodos de lazo abierto

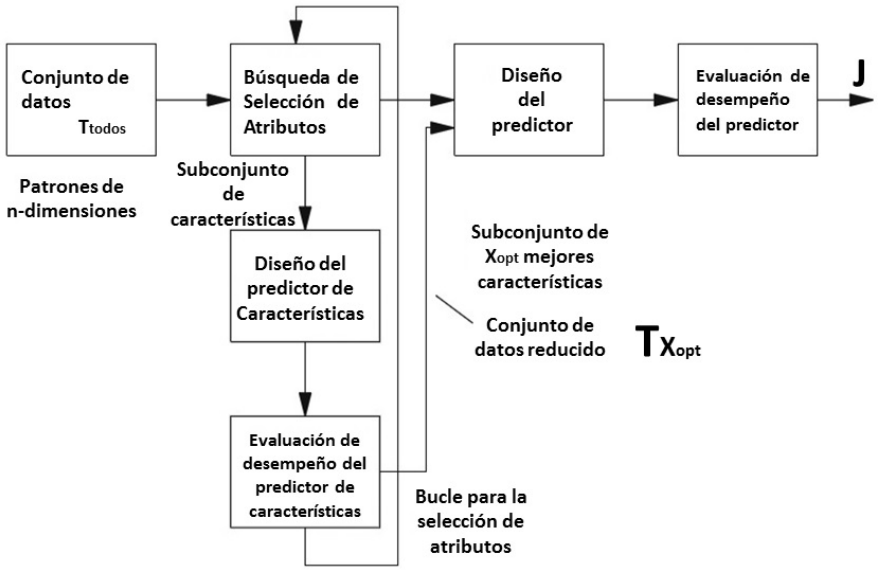


Figura 2.6: métodos de lazo cerrado

### Sequential Floating Feature Selection

Uno de los mejores algoritmos disponibles en la literatura de selección de atributos es el Sequential Floating Feature Selection (SFFS) (Pudil *et al.*, 1994) consiste en aplicar después de cada paso hacia adelante un numero de pasos hacia atrás hasta que el resultado del subconjunto sea mejor al previamente evaluado. En consecuencia, no hay retrocesos en absoluto si el resultado no se puede mejorar. Este algoritmo permite un retroceso auto-controlado por lo que llega a encontrar bue-

nas soluciones ajustando el equilibrio entre avances y retrocesos dinámicamente. En cierto modo, se computa sólo lo que necesita sin ningún ajuste de parámetros.

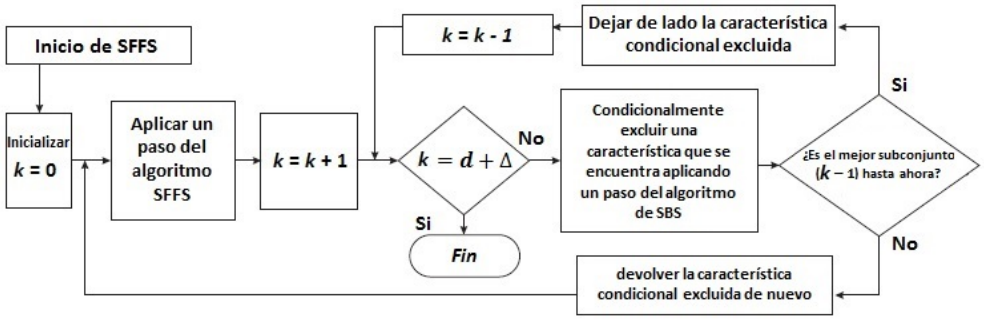


Figura 2.7: Diagrama de flujo del algoritmo SFFS.

Este algoritmo funciona de la siguiente manera, supongamos que $k$ características ya han sido seleccionadas del conjunto de medidas $Y = \{y_j | j = 1, 2, \ldots, D\}$, para formar el conjunto $X_k$ con el criterio correspondiente a la función $J(X_i)$ para todos los subconjuntos precedentes de tamaño $i = 1, 2, \ldots, k-1$. Este algoritmo consiste en tres pasos principales:

1. El primer paso es la inclusión, donde se selecciona la característica $X_{k+1}$ del conjunto disponible, $Y - X_k$, para formar el conjunto de características $X_{k+1}$. La más significativa $X_{k+1}$ con respecto al conjunto $X_k$ es añadida a $X_k$. Por lo tanto:

$$X_{k+1} = X_k + x_{k+1} \tag{2.9}$$

2. El segundo paso es la exclusión condicional, en la cual se busca la característica menos significativa en el conjunto $X_{k+1}$. Si $x_{k+1}$ es la característica menos significativa en el conjunto $X_{k+1}$.

$$J(X_{k+1} - xk + 1) \geq J(X_{k+1} - xj), \forall j = 1, 2, \ldots, k \tag{2.10}$$

   Entonces $k = k + 1$ y regresamos al primer paso, pero si $x_r, 1 \geq r \geq k$, es la característica menos importante del conjunto $X_{k+1}$,

$$X'_k = X_{k+1} - x_r. \tag{2.11}$$

Teniendo en cuenta que ahora $J(X'_k) > J(X_k)$ si k=2, entonces $X_k = X'_k$ y $J(X_k) > J(X'_k)$ y regresamos al paso 1, de otra forma avanzamos al siguiente paso.

3. El tercer paso es la continuación de exclusión condicional. Encontramos la característica menos importante $x_s$ en el conjunto $X'_k$. Si $J(X'_k - x_s) \leq J(X_{k-1})$ entonces $X_k, J(X_k) = (X'_k)$ y regresamos al primer paso. Si $J(X'_k - x_s) > J(X_{k-1})$ entonces excluimos $x_s$ de $X'_k$ para formar un nuevo y reducido conjunto $X'_{k-1}$.

$$X'_{k-1} = x'_s \tag{2.12}$$

Y $k$=$k$-1. Ahora si $k$=2, entonces $X_k = X'_k$ y $J(X_k) = J(X'_k)$ y regresamos al primer paso, de otro modo se repite el tercer paso. El algoritmo comienza con $k$=0 y $X_0 = 0$, y el método SFFS se utiliza hasta que se obtiene un conjunto de características de cardinalidad $k$. Entonces el algoritmo continúa con el primer paso.

### 2.3.2 Algoritmos de clasificación

En muchas ocasiones, uno se interesa en asignar objetos a clases en la base de medición hechas en esos objetos. Esos problemas tienen dos grandes vertientes: la discriminación en grupos (o llamado clustering o aprendizaje no supervizado) y la discriminación en clases (o llamado aprendizaje supervisado). En el aprendizaje no supervisado las clases son desconocidas *a priori* y necesitan ser descubiertas por los datos. Esto envuelve un número estimado de clases (clusters) y asigna objetos a esas clases.

En contraste, en el aprendizaje supervisado (también conocido como clasificación, análisis discriminante, predictor de clase y reconocimiento de patrones supervisado), las clases están predefinidas y la tarea es entender la base para la clasificación de un conjunto de objetos etiquetados (conjunto de entrenamiento). Esta información es usada para construir un clasificador que luego será aplicado para predecir la clase de futuras observaciones sin etiquetar. En varias situaciones, estos dos problemas están relacionados, porque las clases que han sido

descubiertas por un aprendizaje no supervisado son usadas a menudo en un momento posterior en una configuración de aprendizaje supervisado (Lele *et al.*, 2003).

La clasificación es una predicción o problema de aprendizaje en el cual la variable a ser pronosticada asume una $K$ predefinida y valores desordenados, $\{c_1, c_2, \ldots, c_K\}$, arbitrariamente reetiquetado por los enteros $\{1, 2, \ldots, K\}$ o $\{0, 1, 2, \ldots, K-1\}$, y algunas ocasiones {-1, 1} en clasificación binaria. Los $K$ valores corresponden para $K$ predefinidas clases (por ejemplo; cáncer de mama, dermatomiositis juvenil). Asociado con cada uno de los objetos son: una respuesta o variable dependiente (etiqueta de clase), Y $\in \{1, 2, \ldots, K\}$ y un conjunto de G mediciones que forman el vector de características o el vector de predicción de variables, $\mathbf{X} = (X1, \ldots, XG)$. El vector de características $\mathbf{X}$ pertenece a un espacio de características $\mathbf{X}$. Esta labor es para clasificar un objeto a alguna de las $K$ clases sobre la base de una medición observada $\mathbf{X}$=x.

El clasificador o predictor para $K$ clases es un mapeo de $C$ de $\chi$ en $\{1, 2, \ldots, K\}$. $C$: $\chi \rightarrow \{1, 2, \ldots, K\}$, donde $C$(x) denota la clase predicha para un vector característico x. Eso es, un clasificador $C$ corresponde a una partición de un espacio característico $\chi$ dentro de $K$.

Los clasificadores son construidos o entrenados de experiencias pasadas (de observaciones que son conocidas y pertenecen a ciertas clases), como observaciones comprendidas del conjunto de aprendizaje, $\pounds = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Un clasificador construido de un conjunto de aprendizaje $\pounds$ es denotado por $C(\cdot; \pounds)$, cuando el conjunto de aprendizaje es visto como una colección de variables aleatorias, el resultado del clasificador es también conocido como variable aleatoria.

En la clasificación de datos de expresión genética en mediciones de cáncer, las características corresponden para las mediciones de expresión de diferentes genes y las clases corresponden a diferentes tipos de tumores (tumores con buen

diagnóstico vs tumores con mal prognosis).

Existen tres principales tipos de problemas estadísticos asociados con la clasificación de tumores:

1. *Identificación de una nueva clase de tumor utilizando la expresión de genes (aprendizaje no supervisado).*

2. *Clasificación de lesiones malignas en clases conocidas (aprendizaje supervisado).*

3. *Identificación de genes marcadores que caracterizan las diferentes clases de tumor (selección de atributos).*

A continuación procederemos a explicar los principales algoritmos de clasificación usados a lo largo de esta investigación.

### Algoritmo Naive Bayes

El Clasificador Bayesiano se aplica a las tareas de aprendizaje donde cada instancia está descrita por un conjunto de valores de atributos y donde la función objetivo toma cualquier valor de un conjunto finito (Mitchell, 1997).

En la improbable situación en que las densidades condicionales de clase $p_k(x) = p(x|Y = k)$ y el conocimiento previo de clase $\pi_k$ son conocidos, el teorema de Bayes es utilizado para expresar la probabilidad posterior $p(k|\mathtt{x})$ de la clase $k$ dado el vector de características x como:

$$p(k|\mathtt{x}) = \frac{\pi_k p_k(\mathtt{x})}{\sum_l \pi_l p_l(\mathtt{x})}. \tag{2.13}$$

El clasificador Naive Bayes predice la clase de una observación de x por la probabilidad posterior más alta:

$$\mathcal{C}_B(\mathtt{x}) = \mathtt{argmax}_k \ p(k|\mathtt{x}). \tag{2.14}$$

La probabilidad de clase posterior refleja la confianza en la predicción de observaciones individuales; entre más cercanos están a uno, mayor su confianza.

La regla de Bayes minimiza la función de riesgo o el rango de clasificación errónea bajo una función de pérdida simétrica – riesgo de bayes. Para una función de perdida general L, la regla de clasificación la cual minimiza la función de riesgo es:

$$\mathcal{C}_B(\mathbf{x}) = \sum_{h=1}^{K} \mathtt{L}(h,l)p(h|\mathbf{x}). \tag{2.15}$$

Una diferencia interesante entre el método de naive bayes y el aprendizaje de otros métodos es que no hay búsqueda explícita a través del espacio de posibles hipótesis. En su lugar, la hipótesis se forma sin necesidad de buscar, simplemente contando la frecuencia de las diversas combinaciones de datos dentro de los ejemplos de entrenamiento (Mitchell, 1997).

### Funciones discriminantes lineales y cuadráticos

Estos algoritmos construyen reglas discriminantes desde la base del enfoque bayesiano o regla de bayes. Cuando las características tienen una distribución gaussiana dentro de cada clase, hacemos P($\mathbf{X}$|Y=k)$\sim$N($\mu_k$,$\Sigma_k$), con $\mu_k$=($\mu_{k1}$,. . .,$\mu_{kG}$) y $\Sigma_k$ denotando respectivamente el valor esperado y la matriz de covarianza del vector de características en la clase $k$. Así, la regla de bayes se traduce a:

$$\mathcal{C}(\mathbf{x}) = argmax_k\{(\mathbf{x}-\mu_k)\Sigma_k^{-1}(\mathbf{x}-\mu_k)' + \log|\Sigma_k| - 2\log\pi_k\}. \tag{2.16}$$

De esta manera, la regla de clasificación es asignar la clase $k$ asociada a la $\mathcal{C}(\mathbf{x})$ más alta. La discriminación cuadrática es similar a la discriminación lineal, pero el límite entre dos regiones de discriminación ahora permite ser una superficie cuadrática (Clarke *et al.*, 1979).

En la regla del discriminante cuadrático, conocido como Análisis Discriminante Cuadrático (QDA por sus siglas en inglés), la cantidad principal $(\mathbf{x}-\mu_k)\Sigma_k^{-1}(\mathbf{x}-\mu_k)'$ es conocida como la distancia cuadrada de Mahalanobis de una observación x a la media de la clase $k$ del vector $\mu_k$.

Cuando las densidades de clase tienen la misma matriz de covarianza $\Sigma_k = \Sigma$, la regla discriminante está basada en el cuadrado de la distancia Mahalanobis y es lineal en x, esto se traduce en la regla del discriminante lineal (LDA) definida como $\mathcal{C}(\mathtt{x}) = argmax_k\{(\mathbf{x} - \mu_k)\Sigma^{-1}(\mathbf{x} - \mu_k)' = argmax_k(\mu_k\Sigma^{-1}\mu_k' - 2x\Sigma^{-1}\mu_k')$

**Clasificador basado en vecinos cercanos**

Los métodos de vecinos cercanos están basados en una función de distancia por observaciones en pares, como es la distancia euclidea o uno menos de la correlación de Pearson en sus vectores característicos. Para clasificar una nueva observación sobre la base del conjunto de aprendizaje, la regla del $k$ vecinos cercano básico ($k$-NN) procede como sigue: encontrar las $k$ observación más cercana en el conjunto de aprendizaje y predecir la clase por la mayoría de votos.

Los clasificadores $k$-vecinos cercanos sugieren una simple estimación de la clase por probabilidades posteriores, $d_i(\mathtt{x}) = d(\mathtt{x}, \mathtt{x}_i)$ denotando la distancia de observación de $\mathtt{x}_i$ entre el conjunto de aprendizaje y el caso de pruebas x, y al denotar las distancias ordenadas por $d_{(1)}(\mathtt{x}) \leq \ldots \leq d_{(n)}(\mathtt{x})$. Entonces, la clase es estimada de acuerdo a la probabilidad posterior $\hat{p}(l|\mathtt{x})$ para el estándar del clasificador $k$-vecinos cercanos son dados por la fracción de la clase $l$ entre los vecinos más cercanos de los casos de prueba.

Aunque los clasificadores con $k$=1 son a menudo muy exitosos, el número de $k$ vecinos puede tener un gran impacto en el rendimiento del clasificador y debería ser elegido cuidadosamente. La manera común de seleccionar el número de vecinos es por medio de validación cruzada. Cada observación en el conjunto de entrenamiento es tratado como si su clase fuera desconocida. Esta distancia para todos los otros datos del conjunto de entrenamiento (menos así mismo) son medidos, y son clasificados por la regla del vecino más cercano. La clasificación para cada observación del conjunto de entrenamiento es entonces comparada con la verdad para producir la tasa de error de la validación cruzada. Esto se realiza por el número de $k's (k \in \{1, 3, 5, 7\})$ y la $k$ para cual la tasa de error de la validación cruzada es la más pequeña.

### 2.3.3 Métodos de validación

Una manera de eliminar la tasa de error en tareas de clasificación es a través de la técnica de o validación cruzada o cross validation (CV por sus siglas en inglés) y es discutida en un contexto más amplio por (Stone, 1974) y (Geisser, 1975). Este consiste en dividir el conjunto de datos en $K$ partes iguales. Para la $k-th$ parte, el algoritmo es entrenado en usando los datos restantes de las $K-1$ particiones, y se calcula el error de clasificación en los datos de la $k-th$ partición. Este proceso se repite para $k = 1, 2, \ldots K$. De esta manera el error de CV estará dado por:

$$CV = \frac{1}{n}\sum_{i=1}^{N} L(y_i, f^{k(i)}(x_i)) \tag{2.17}$$

Cuando hacemos K=N, es decir el número de particiones es igual al número de datos, la técnica de CV degenera en lo que se conoce como leave-one-out (LOO) descrita por (Lachenbruch & Mickey, 1968). Sin embargo, esta vertiente requiere esfuerzo computacional en exceso. Los valores recomendados en la literatura son de $K$=5 o 10 (McLachlan *et al.*, 2005).

#### Estimación del la tasa de error por el valor 0.632 Bootstrap

Los métodos bootstrap son herramientas para la construcción de estimados del error estándar e intervalos de confianza de determinados estadísticos, cuando los tamaños de muestra son pequeñas, y la distribuciuón del estadístico analizado es desconocida. Asumiendo que tenemos un conjunto de datos $S = \{(x_i, y_i)\}_{i=1\div p}$, se extraen *muestras bootstrap* $S_1^*, \ldots, S_B^*$ de tamaño $p$ muestreando $S$ con reemplazo y se procede a construir el modelo de clasificaión en cada una de las $S_b^*$. El estadístico de inters $\hat{\theta} = \theta(S)$ es estimado de la forma tradicional, e.g.:

$$\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\theta_b^*; \qquad Var(\theta^*) = \frac{1}{B-1}\sum_{b=1}^{B}(\theta_b^* - \overline{\theta}^*)^2 \tag{2.18}$$

estas últimas expresiones son la media y la variable de la distribución bootstrap de $\hat{\theta}$, y $\theta_b^* = \theta(S_b^*)$. Una estimación para el sezgo de $\hat{\theta}$ es obtenida como $\overline{\theta}^* - \hat{\theta}$.

El estimado leave-one-out (LOO por sus siglas en inglés) del error de predicción, para tareas de clasificación:

$$Err^* = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|S^{-i}|}\sum_{b\in S^{-i}}\mathbb{I}[y_i \neq f_b^*(x_i)] \tag{2.19}$$

donde $\mathbb{I}(z)$ es 1 cuando $z$ es verdad y 0 en caso contrario, $S^{-i}$ es el conjunto de índices de las muestras bootstrap que no contienen la observación $x_i$ y $f_b^*$ es el model desarrollado en $S_b^*$. Similarmente, el error de re-substitución es estimado como:

$$\overline{e}^* = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|S^{i}|}\sum_{b\in S^{i}}\mathbb{I}[y_i \neq f_b^*(x_i)] \tag{2.20}$$

donde $S^i = \{1 \div B\} \setminus S^{-i}$. En ambos casos, si el índice es un conjunto vacuo, el término es ignorado de la fórmula y ésta es renormalizaca. De esta manera el estimado 0.632 bootstrap es calculado como:

$$Err^{(0.632)} = 0.368\ \overline{e}^* + 0.632\ Err^*. \tag{2.21}$$

# Capítulo 3

# Materiales y Métodos

En este capítulo explicaremos la metodología seguida para desarrollar la investigación; se ofrecerá una descripción de los datos de expresión genética utilizados; algoritmos de aprendizaje de máquina; el hardware empleado para el desarrollo y ejecución de los algoritmos; y el procedimiento experimental que fue seguido.

## 3.1 Metodología

La Metodología utilizada es la propuesta por (Duda & Hart, 2001) y es el Ciclo de Diseño de un Sistema de Reconocimiento de Patrones. En la Figura 3.1 se presenta de forma gráfica.


Figura 3.1: Metodología (Duda & Hart, 2001)

1. La primera etapa, de recolección de datos, consta de reunir los conjuntos de datos para realizar todo el estudio de reconocimiento de patrones. En nuestro caso, los datos se obtuvieron del repositorio EMBL-EBI del Instituto Europeo de Bioinformática NCBI (2007).

2. La selección de características, previamente explicado, es el punto central del trabajo de tesis.

3. Elegir modelo, que consiste de seleccionar el mejor modelo i.e. el que mejor separe los datos de sus clases componente, fue fijado desde un inicio. Esto es los clasificadores que fueron utilizados poseen 2 características importantes:

   (a) Son rápidos.

   (b) No requieren parámetros de entrenamiento.

   Para la tarea que nos ocupa, fueron seleccionados los algoritmos de Naive Bayes, 1-vecino cercano, 3-vecinos cercanos, el discriminante lineal y el discriminante cuadrático.

4. El proceso de entrenamiento, el cual depende del clasificador mismo.

5. Finalmente, evaluar el clasificador, consta de evalúar la certeza estadística de la capacidad de generalización del modelo seleccionado. En el caso que ocupa esta tesis, el método de Bootstrap y el error 0.632B fue seleccionado para ello, en virtud de que el conjunto de datos analizado, es de tamaño pequeño y de distribución desconocida.

## 3.2 Descripción de los datos

La base de datos utilizada en esta tesis fue obtenida del repositorio EMBL-EBI del Instituto Europeo de Bioinformática NCBI (2007). El experimento *E-GEOD-3307* analiza un rango de enfermedades musculares para la expresión de genes con fines comparativos. Un total de 121 muestras de 11 patologías (además de varias muestras sanas) integran los datos: la miopatía tetrapléjica aguda, la dermatomiositis juvenil, la esclerosis lateral amiotrófica, la paraplejia espástica, la distrofia muscular fascioscapulohumeral, la distrofia muscular de Emery Dreifuss, la distrofia muscular de Becker, la distrofia muscular de Duchenne, la calpaína 3, la disferlina y la FKRP usando el chip de Affimetrix U133A y U133B.

Estas son enfermedades con incidencia extremadamente baja en la población general. La DMJ, el grupo objetivo en este trabajo, consiste en 18 muestras de DMJ y 21 muestras sanas descritas por 22283 genes o características.

## 3.3 Métodos y técnicas utilizadas

Los datos bajo estudio tienen una distribución de clases de la siguiente manera: el 46% con DMJ y 54% sin la enfermedad además, tienen una gran cantidad de atributos (22283) pero pocas muestras de pacientes (39). Es por ello que se recurre a métodos de remuestreo bootstrap. El total de muestras bootstrap generadas con este método fue de $B = 100$.

Se implementaron los algoritmos de test de Welch (Welch, 1947) y el Score de Fisher filtros, con el propósito de facilitar la selección de atributos:

El test de Welch esta definido de la siguiente manera:

$$t_{w_j} = \frac{|\overline{X}_j^{(+)} - \overline{X}_j^{(-)}|}{\sqrt{\frac{S_j^{2+}}{N^+} + \frac{S_j^{2-}}{N^-}}} \tag{3.1}$$

Y el Score de Fisher se define como:

$$f_{w_j} = \frac{|\overline{X}_j^{(+)} - \overline{X}_j^{(-)}|^2}{S_j^{2+} + S_j^{2-}} \tag{3.2}$$

Ambos calculan un valor o score de importancia de cada gen, es decir, valores altos indican una gran posibilidad de separabilidad de clases –ver Figura 3.2. Contrariamente, un valor pequeño indicaría que un algoritmo de clasificación tendría pocas posibilidades de lograr altas tasas de reconocimiento–ver Figura 3.3. Estos dos filtros permiten "guiar" el proceso de reducción de la dimensionalidad donde los atributos o genes que posean valores o score altos son los que serán primeramente seleccionados por el algoritmo de selección de atributos.

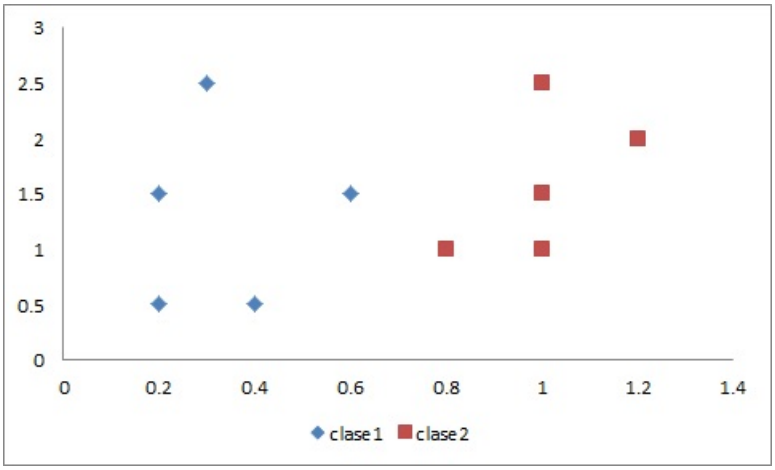


Figura 3.2: Separabilidad de clases para un atributo con alto valor o score de importancia.

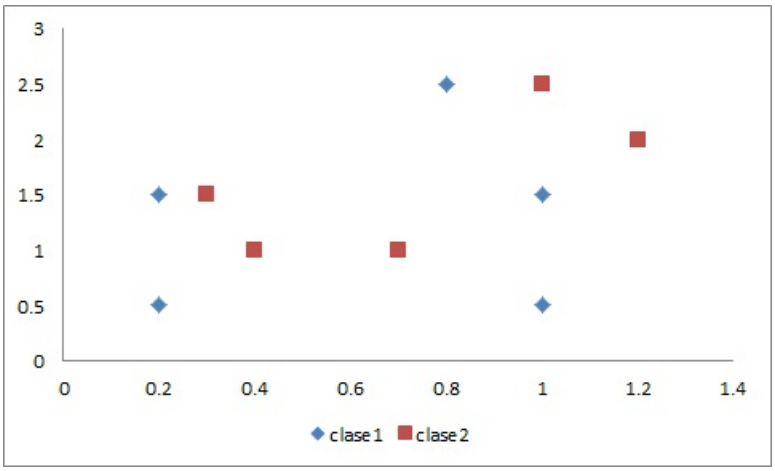


Figura 3.3: Separabilidad de clases para un atributo con bajo valor o score de importancia.

Al ordenar los resultados de los filtros de manera descendente, se procedió a seleccionar los primeros 100 genes que tienen una mayor separabilidad entre clases, y se introdujeron al algoritmo de Sequential Forward Feature Selection (SFFS). Este proceso se repitió para cada una de las $B = 100$ muestras bootstrap. Esta estrategia repetitiva permitió variar las condiciones del problema –i.e. simular la disponibilidad de 100 muestras diferentes.

Es conocido en la literatura científica que a condiciones diferentes para un mismo conjunto de datos, los subconjuntos finales en un proceso de selección de atributos, serán diferentes. Sin embargo, es posible que un determinado número de genes aparezca repetitivamente en los subconjuntos finales. Esto representó la base para establecer un perfil genético de la DMJ a partir de un grupo de genes que, sin importar las condiciones, aparecerá en los resultados finales.

Puesto que se seleccionó el algoritmo SFFS, este usa un algoritmo de clasificación como función criterio para la selección de los mejores atributos. Como se mencionó anteriormente los siguientes clasificadores fueron los utilizados:

1. Naive Bayes
2. Discriminante Lineal
3. Discriminante Cuadrático
4. 1 Vecino Cercano
5. 3 Vecinos Cercanos

Se utilizó el simulador matemático MATLAB R2011a para desarrollar las rutinas de selección de atributos, así como los algoritmos para el cáculo de los filtros. Además, se utilizó el toolbox para reconocimiento de patrones llamado prtools (V. 4.1.6), el cual cuenta con varios clasificadores ya implementados. Todos los experimentos fueron ejecutados en laptop AMD Athlon x2 Dual-Core 2.10 GHz, arquitectura 64 bits con 4 GB. de ram.

## 3.4 Procedimiento experimental

El procedimiento experimental, ilustrado en la Figura 3.4, consistió de lo siguiente: en primer lugar se aplicó el algoritmo SFFS a cada muestra de bootstrap. Para cada muestra bootstrap se calcula el error de bootstrap –ver Ecuación 2.21.

Una vez terminado de aplicar el procedimiento para el total de las muestras $B = 100$, se obtuvo una distribución de frecuencias a partir de los genes que

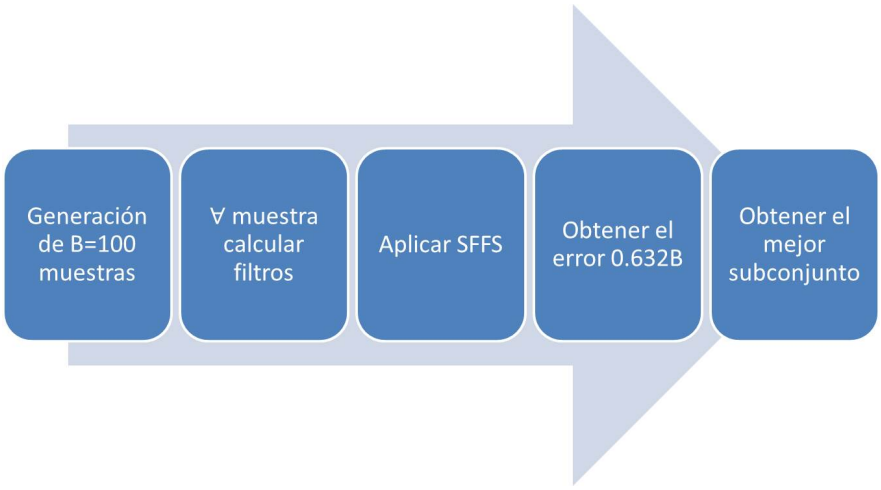


Figura 3.4: Procedimiento Experimental.

fueron seleccionados en los 100 subconjuntos de atributos o genes finales. Posteriormente, se hizo una extensa investigación documental de los primeros genes en la literatura médica sobre su relación con la DMJ.

# Capítulo 4

# Resultados

En este capítulo se presentarán los resultados experimentales que fueron logrados mediante la aplicación de los métodos y algoritmos descritos en el capítulo anterior.

## 4.1 Resultados Experimentales

En las Figuras 4.1 a 4.5 se muestran los diagramas de frecuencias de genes que mayormente aparecen en el proceso de selección de atributos. Hay que recordar que para cada muestra Bootstrap, el proceso de selección de atributos se lleva a cabo. Puesto que las muestras Bootstrap son hasta cierto punto diferentes, los resultados finales, i.e. el subconjunto final, serán diferentes. Sin embargo, se presume la existencia de genes que aparecerán con mayor frecuencia que el resto, aduciendo una posible relación matemática consistente con la enfermedad, i.e. en términos de clasificación.

En la gráfica de frecuencias del clasificador Naive Bayes y filtro Fisher Score (Figura 4.1 izquierda), son 3 los genes que mayormente aparecen seleccionados por el algoritmo SFFS de las 100 muestras Bootstrap: el gen *CHRNA4* con un total de 51 veces; seguido del gen *ALPI* con 42; y en tercer lugar el gen *C8orf39* con 27 ocasiones. La tasa de reconocimientos alcanzada es de 89.6%. Para el filtro Test de Welch (Figura 4.1 derecha) el gen *HLA-B* es visible en 52 subconjuntos; seguido del gen *STAT*1 con 43 veces; el tercer gen ADAR aparece en 32

subconjuntos. La tasa de reconocimientos alcanzada es de un sobresaliente 98.9%.

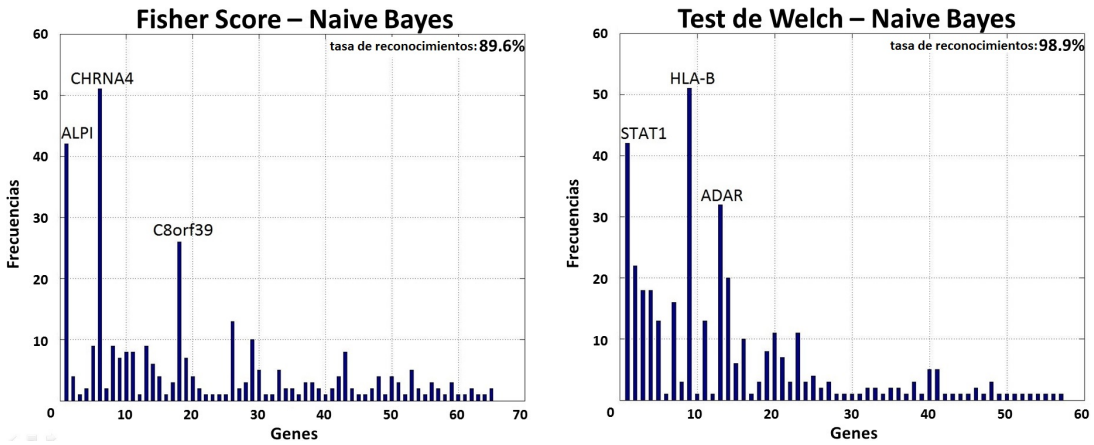


Figura 4.1: Distribución de genes utilizando el clasificador Naive Bayes + los filtros de Fisher Score (izquierda) y Test de Welch (derecha).

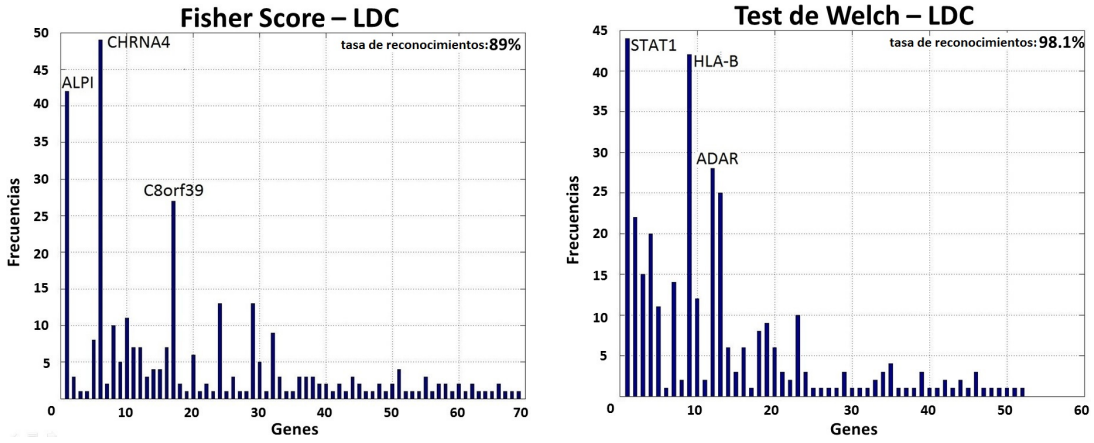


Figura 4.2: Distribución de genes utilizando el clasificador LDC + los filtros de Fisher Score (izquierda) y Test de Welch (derecha).

En la Figura 4.2 se muestran los resultados de la combinación del algoritmo SFFS con el discriminante lineal LDC como medida de rendimiento. El lado izquierdo de esta gráfica (usando score de Fisher) son tres los genes dominantes en las 100 muestras Bootstrap: *CHRNA4* con 48 apariciones; el gen *ALPI* con 44; y el gen *C8orf39* encontrado en 27 subconjuntos. Para el filtro Test de Welch

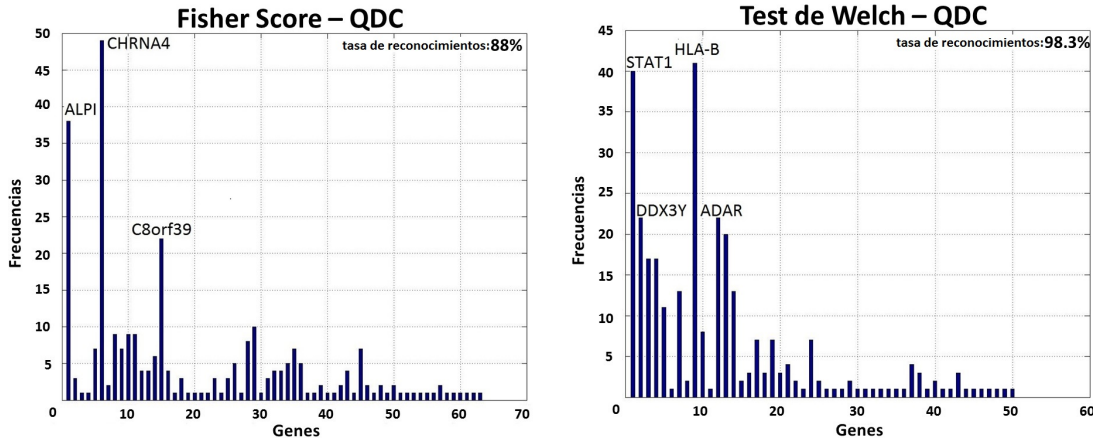


Figura 4.3: Distribución de genes utilizando el clasificador QDC + los filtros de Fisher Score (izquierda) y Test de Welch (derecha).

(Figura 4.2 derecha) sobresalen: el gen *STAT1* con 44 veces siendo seleccionado; el gen *HLA-B* con 42; y el gen *ADAR* con 28. Las tasas de reconocimiento son similares a los resultados descritos en el anterior párrafo, de 89% y 98.1% respectivamente.

En la gráfica de frecuencias 4.3 del clasificador QDC y el filtro Fisher Score (lado izquierdo) los genes de mayor importancia por su aparición en más subconjuntos son como sigue: el gen *CHRNA4* con 49 participaciones en subconjuntos; seguido del gen *ALPI* con 38; y el tercer gen *C8orf39* con un total de 22. Para el filtro Test de Welch (lado derecho) aparece el gen *HLA-B* con 41 repeticiones; seguido del gen *STAT1* con 40; y finalmente el gen *ADAR* con 22. La tasa de reconocimientos alcanzada es muy similar a los valores anteriormente obtenidos por las otras combinaciones SFSS + clasificador + filtro, con 88% y 98.3% respectivamente.

Para el clasificador 1NN –ver Figura 4.4, los resultados para el filtro de Score de Fisher son los siguientes: el gen *CHRNA4* es seleccionado en 47 veces; seguido del gen *ALPI* con 40; y el tercer gen con mayor frecuencia es el *C8orf39* con 26 selecciones. La tasa de reconocimientos alcanzada es de 90.2%, ligeramente mayor al resto de combinaciones para el mismo filtro. Para el filtro Test de Welch:

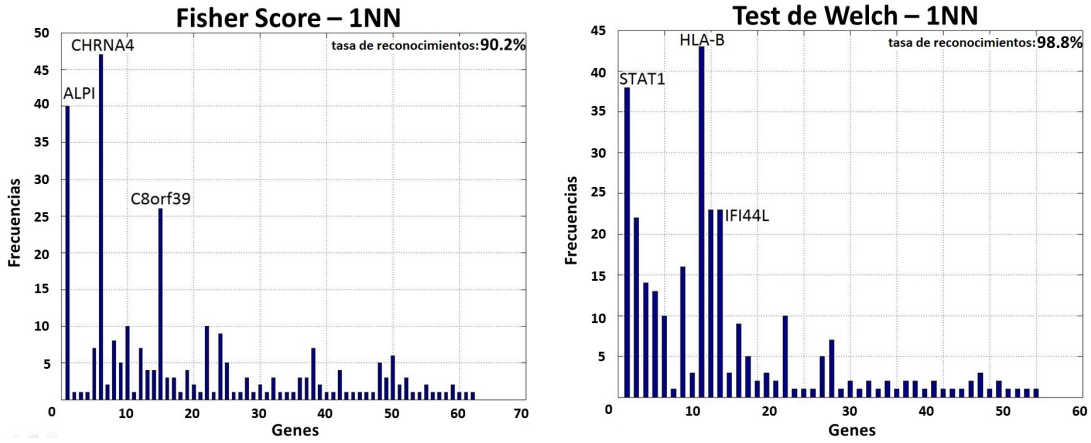


Figura 4.4: Distribución de genes utilizando el clasificador 1NN + los filtros de Fisher Score (izquierda) y Test de Welch (derecha).

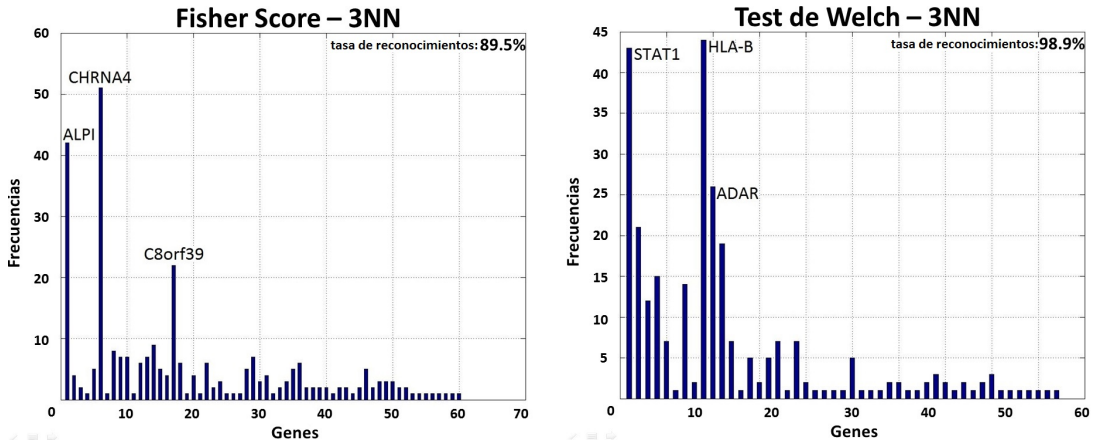


Figura 4.5: Distribución de genes utilizando el clasificador 3NN + los filtros de Fisher Score (izquierda) y Test de Welch (derecha).

el gen *HLA-B* lidera la distrubución de frecuencias con 43; el gen *STAT1* se ubica en segundo lugar con 38; y el gen *FI44L* que aparece en 23 ocasiones con una tasa de reconocimientos de 98.8%.

Finalmente, la respuesta del algoritmo 3NN –Figura 4.5– muestra un comportamiento similar al resto de los experimentos. Para el filtro de Score de Fisher, se aprecian tres genes dominantes: el primer gen *CHRNA4* registra 51 selecciones; el segundo gen *ALPI* con 42; y el tercer gen *C8orf39* tiene una frecuencia igual

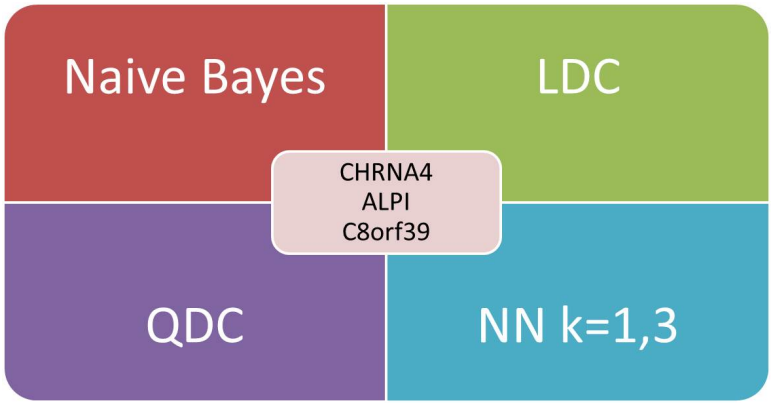


Figura 4.6: Genes con mayor frecuencia de aparición en relación a todo el experimento para el Filtro de Score de Fisher y con los algoritmos de Naive Bayes, LCD, QDC, 1NN y 3NN.

a 22; con una tasa de reconocimientos de 89.5%. Para el filtro Test de Welch los resultados son: el gen *HLA-B* con 44; el gen *STAT1* con 43, el tercer gen *ADAR* es seleccionado en 26 ocasiones y la tasa de reconocimientos es de 98.9%.

### 4.1.1 Discusión de resultados

En la Figura 4.6 podemos apreciar de una forma gráfica los genes que más veces fueron elegidos para el Filtro de Score de Fisher. Se puede notar que los clasificadores seleccionan repetitivamente tres genes principalmente. A pesar de que se utilizaron muestras Bootstrap para generar condiciones de experimentación diferentes, los resultados muestran que existe un potencial perfil genético o grupo de genes que, en la tarea de clasificar o distinguir muestras musculares con DMJ de muestras musculares sanas, logra resultados prometedores.

Por otra parte, en la Figura 4.7 vemos que el perfil genético por clasificador utilizando el filtro de Test de Welch, muestra un ligero aumento en la diversidad de genes seleccionados. En el semicírculo de color verde de la Figura antes referenciada muestra los genes que aparecieron en más ocasiones para el clasificador Naive Bayes, LDC y 3NN. En el azul se aprecia lo logrado por el clasificador de 1NN y en el morado lo respectivo al QDC. Sin embargo, existe un grupo de

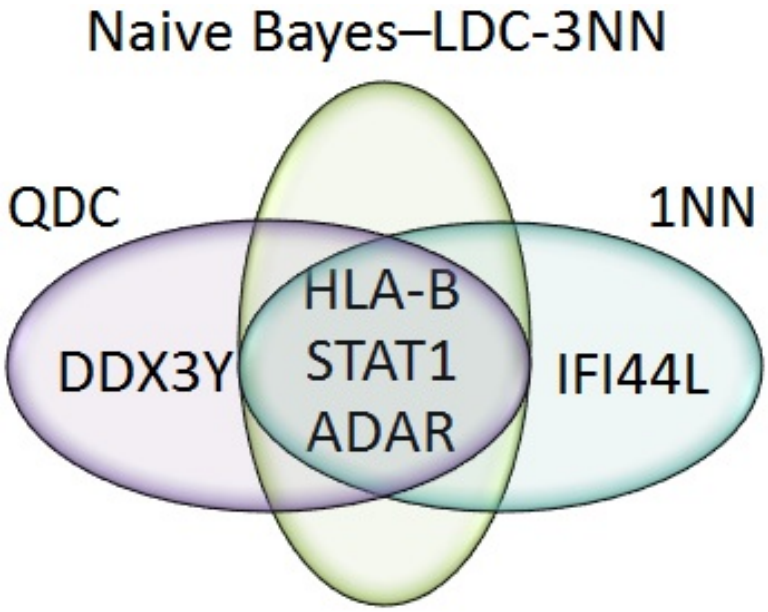


Figura 4.7: Genes con mayor frecuencia de aparición en relación a todo el experimento para el Filtro de Test de Welch y con los algoritmos de Naive Bayes, LCD, QDC, 1NN y 3NN.

tres genes que son comúnes a todas las opciones de experimentación. Y de nueva cuenta, la tasa de reconocimiento es notoriamente alta, alrededor del 98% de exactitud.

### 4.1.2 Evidencia Biológica

En esta sección presentaremos evidencia médica sobre la relevancia de los genes encontrados. Esta fue ensamblada mediante una búsqueda exhaustiva de artículos relacionados con cada gen. La información sobre los genes se mostrará la siguiente manera: Símbolo de gen, descripción y su resumen y evidencia médica sobre el gen relacionándolo con la DMJ.

#### ***ALPI* alkaline phosphatase, intestinal**

En un estudio se investigaron biopsias musculares en diez niños con DMJ. Se realizaron biopsias en el vasto externo utilizando anestesia local. La reacción de la fosfatasa ácida mostró grados variables de necrosis, y la reacción de la fosfatasa

alcalina muestra características de moderada a intensa regeneración. Se concluye que la biopsia muscular podría ayudar en el diagnóstico de dermatomiositis infantil, así como en la detección de recurrencias, incluso en casos sin actividad clínica de la enfermedad (Calore *et al.*, 1997).

Cros *et al.* (1980) Analizó la actividad de la fosfatasa alcalina en donde, por métodos estadísticos no paramétricos, particularmente el test de Welch (utilizado en esta tesis) y Student t-test, se buscaron variantes. Se encontró que la reacción de la fosfatasa alcalina identifica una particular anormalidad de fibra muscular humana. La presencia de esas fibras no especifica una inflamación pero probablemente representa una fase particular en la degeneración.

### *STAT1* signal transducer and activator of transcription 1,91kDa

Esta proteína intermediaria en la expresión de una variedad de genes, se piensa que es importante para la viabilidad celular en respuesta a diferentes estímulos celulares y patógenos. En un análisis de inmunohistoquímica utilizando anticuerpos en 57 biopsias de músculos de 10 pacientes con DM y 10 con PM, se analizó la reactividad de STAT1, en donde apareció en la mayoría de las células infiltradas en DM y PM. Los resultados sugieren que en la DM hay una actividad en función local distintiva de citosina que incrementa la expresión del STAT1 en fibras musculares (Illa *et al.*, 1997).

En otro estudio de genes regulados por el sistema IFN-gamma/STAT1 en el músculo humano se encontró que podrían estar involucrados en el proceso de atrofia de las fibras musculares en la DM. La inmuno histoquímica de secciones congeladas de biopsias de pacientes con diferentes enfermedades musculares mostró regulación a la alza de la catepsina L y la calpaína en las fibras musculares perifasciculares en DM. Se concluyó que el aumento de la expresión de las catepsinas L y B en las fibras musculares perifasciculares en DM estan mediadas por IFNgamma/STAT1 y contribuyen a su atrofia (E. *et al.*, 2001).

### *CHRNA4* cholinergic receptor, nicotinic, alpha 4 (neuronal)

Este gen codifica un receptor de acetilcolina nicotínico, el cual pertenece a una súper familia de canales iónicos que juegan un papel en la transmisión rápida de las señales en las sinapsis. Mutaciones en este gen pueden causar epilepsia del lóbulo frontal nocturnas tipo 1 (NCBI, 2007). No se encontró evidencia que ligue este gen a la DMJ en la literatura médica.

### *HLA-B* major histocompatibility complex, class I, B

El gen *HLA-B* provee instrucciones para fabricar una proteúna que juega un papel importante en el sistema inmune. Pertenece a la familia de genes llamados *human leukocyte antigen complex* (HLA por sus siglas en inglés). Estos complejos ayudan al sistema inmune a distinguir las proteínas del cuerpo humano de proteínas hechas por invasores al cuerpo humano como virus y bacterias (NCBI, 2007). Los *HLA* están asociado a diversas enfermedades autoinmunes como artritis reumatoide, lupus sistémico y algunos tipos particulares de cancer (Kumar *et al.*, 2007).

### *C8orf39* DRC1 Dynein Regulatory Complex Subunit 1 Homolog (Chlamydomonas)

La proteina codificada por este gen es un componente clave en la regulación de complejo nexin-dynein y escencial para la integridad del N-DRC (GENECARDS, 2013). No se encontró evidencia de investigaciones particulares de este gen en la literatura médica.

# Capítulo 5

# Conclusiones

## 5.1 Conclusión

A la luz de los resultados obtenidos durante el desarrollo de esta investigación, podemos concluir lo siguiente:

1. En este trabajo de tesis se presentó un perfil genético de la DMJ, enfermedad sistémica autoinmune de muy baja incidencia en la población en general. Al momento de la presentación de esta investigación, las causas que la originan son desconocidas. Por este motivo, los resultados expuestos en este documento vienen a ser una contribución para la literatura médica sobre posibles potenciales biomarcadores –i.e. genes significativos en la distincion entre muestras de tejido con DMJ y normal– con fines terapeúticos o de tratamiento.

2. Se usaron métodos de aprendizaje de máquina en la construcción del perfil genético mediante la aplicación de un algoritmo de selección de atributos, cuyo empleo en la literatura computacional es amplio y razonablemente aceptado. Se utilizaron técnicas de filtrado de atributos, en virtud de que el conjunto de datos analizado posee una gran cantidad de ellos, haciendo su anaális técnicamente difícil si no se recurre a este tipo de técnicas.

3. Debido a la baja cantidad de muestras, se recurrió a un método estadistico de remuestreo para deteterminar con certeza las medidas de rendimiento o de bondad de los subconjuntos de genes –i.e. tasa de reconocimientos– encontrados en la selección de atributos. El esquema *embebido* de selección de atributos dentro del remuestreo permitió explorar el conjunto de datos con condiciones diferentes para tareas de selección de atributos y clasificación, es decir, con diferentes datos.

4. A pesar de la *variación* en los conjuntos de datos, un grupo reducido de genes apareció de manera recurrente en las 100 variantes de datos. Este hecho nos permite afirmar con un grado aceptable de certeza que éstos tienen una incidencia directa o son relevantes, desde un punto matenático en la DMJ. Sin embargo, esto deberá tomarse con precaución, ya que es la comunidad médica o gente especializada la que deberá corroborar esta aseveración, mediante estudios de naturaleza médica-biológica.

5. La robustez de la metodología presentada y el procedimiento experimental desarrollado, permiten variar diversas etapas del proceso mediante el uso de diferentes algoritmo y/o técnicas, por ejemplo la elección de otro algoritmo de selección de atributos, como una selección hacie enfrente simple, selección hacia atras simple, un algoritmo geético, etc.; el uso de otros algoritmos de clasificación, como una regresión logística, máquinas de vectores de soporte, una red neuronal; incrementar considerablemente el tamaño de la muestras bootstrap; diferentes métodos de filtrado, como el Score de Golub, Score de Pearson, etc.

6. Los genes HLA-B, STAT1, DDX3Y, IFI44L y ADAR, son los genes de mayor incidencia, además que se encontró en la literatura médica evidencia biológica y por lo tanto, los genes atípicos en la DMJ que se proponen en esta tesis.

7. La evidencia médica encontrada muestra una relación positiva de algunos de los genes encontrados en este estudio con la DMJ. Para otros genes, no fue posible identificar estudios particulares, sin embargo, este trabajo de

investigación representa una posible vía para que investigadores especializados presten atención a estos genes poco atendidos y pueda contribuirse en el entendimiento de la dinámica de esta enfermedad.

## 5.2 Trabajo Futuro

Como trabajo futuro se plantean las siguientes lineas de investigación:

1. Puesto que la evaluación de muestras bootstrap es independiente entre ellas. La metodología y el procedimiento experimental puede ser implementado con técnicas de cómputo paralelo, lo cual vendrá a incrementar considerablemente el análisis de este tipo de problemas.

2. Es posible ampliar el estudio hacia otras enfermedades neuromotoras y que están contenidas en el conjunto de datos analizado como son la Miopatía Tetrapléjica Aguda, la Asclerosis Lateral Amiotrófica, la Paraplejia Espástica, la Distrofia Muscular Fascioscapulohumeral, la Distrofia Muscular de Emery Dreifuss, la Distrofia Muscular de Becker o la Distrofia Muscular de Duchenne. Estas enfermedades comparten la caractarística de que sus causas también son desconocidas a la fecha. Además de que las contribuciones de estudio desde el punto de vista expuesto en esta tesis sobre ellas es nulo o escaso.

3. Es posible utilizar algoritmos de selección de atributos de naturaleza estocástica como el recocido simulado o algoritmos genéticos, cuya búsqueda en el espacio de soluciones –i.e. subconjuntos de atributos– ha sido comprobada en la literatura que arroja subconjuntos de mejor calidad –i.e. mayores tasas de reconocimiento –ver Gonzalez & Belanche (2008).

# Anexos A

# Lista de genes de mayor frecuencia

En la siguiente tabla se muestran los genes con mayor frecuencia inmediata inferior al perfil genético resultante del estudio de investigación.

Tabla A.1: Genes con mayor frecuencia

| **Probe-set** | **Gene Symbol - Gene Description** | **Frecuencia** |
|---|---|---|
| *201909_at* | *RPS4Y1 - ribosomal protein S4, Y-linked 1* | *16* |
| *211788_s_at* | *TREX2 - three prime repair exonuclease 2* | *11* |
| *208812_x_at* | *HLA-C - major histocompatibility complex, class I, C* | *14* |
| *214022_s_at* | *IFITM1 - interferon induced transmembrane protein 1 (9-27)* | *12* |
| *220847_x_at* | *ZNF221 - zinc finger protein 221* | *16* |
| *219452_at* | *DPEP2 - dipeptidase 2* | *4* |
| *215826_x_at* | *ZNF835 - zinc finger protein 835* | *5* |
| *205309_at* | *SMPDL3B - sphingomyelin phosphodiesterase, acid-like 3B* | *5* |
| *216632_at* | *Hs.655301 - Neuron navigator 3* | *4* |
| *206522_at* | *MGAM - maltase-glucoamylase (alpha-glucosidase)* | *4* |
| *206899_at* | *NTSR2 - neurotensin receptor 2* | *4* |
| *213967_at* | *RALYL - RALY RNA binding protein-like* | *4* |
| *220919_s_at* | *WDR96 - WD repeat domain 96* | *3* |
| *207288_at* | *ASMTL - antisense RNA 1* | *5* |

Tabla A.1: Genes con mayor frecuencia

| Probe-set | Gene Symbol - Gene Description | Frecuencia |
|---|---|---|
| | *(non-protein coding))* | |
| *214165_s_at* | *HS6ST1 - heparan sulfate 6-O-sulfotransferase 1)* | *4* |
| *205314_x_at* | *Hs.461117 Syntrophin, beta 2 (dystrophin-associated protein A1, 59kDa, basic component 2)* | *3* |
| *220377_at* | *FAM30A - family with sequence similarity 30, member A* | *4* |
| *207607_at* | *ASCL2 - achaete-scute complex homolog 2 (Drosophila)* | *2* |
| *220487_at* | *SNTG2 - syntrophin, gamma 2* | *3* |
| *221909_at* | *RNFT2 - ring finger protein, transmembrane 2* | *3* |
| *207190_at* | *ZZEF1 - zinc finger, ZZ-type with EF-hand domain 1* | *2* |
| *214576_at* | *KRT36 - keratin 36* | *3* |
| *207514_s_at* | *GNAT1 - guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1* | *2* |
| *211146_at* | *Hs.696583 - Ig kappa light chain variable region (VkII-A23)* | *2* |
| *216572_at* | *FOXL1 - forkhead box L1* | *2* |
| *206647_at* | *HBZ - hemoglobin, zeta* | *2* |
| *216346_at* | *SEC14L3 - SEC14-like 3* | *2* |
| *221353_at* | *OR3A1 - olfactory receptor, family 3, subfamily A, member 1* | *2* |
| *210922_at* | *BC000772 - Homo sapiens cDNA clone IMAGE:3506689, partial cds.* | *2* |
| *207918_s_at* | *TSPY1-3-4-6-8-10 - testis specific protein, Y-linked* | *2* |
| *206535_at* | *SLC2A2 - solute carrier family 2, (facilitated glucose transporter), member 2* | *2* |
| *214127_s_at* | *Hs.111801 - Serrate RNA effector molecule homolog (Arabidopsis)* | *2* |
| *222086_s_at* | *Hs.29764 - Wingless-type MMTV integration site family, member 6* | *3* |
| *211083_s_at* | *MAP3K13 - mitogen-activated protein* | *3* |

Tabla A.1: Genes con mayor frecuencia

| Probe-set | Gene Symbol - Gene Description | Frecuencia |
|---|---|---|
| | *kinase kinase kinase 13* | |
| *221456_at* | *TAS2R3 - taste receptor, type 2, member 3* | *1* |
| *217366_at* | *CTNNA1P1 - catenin (cadherin-associated protein), alpha 1 pseudogene 1* | *2* |
| *204338_s_at* | *RGS4 - regulator of G-protein signaling 4* | *1* |
| *211314_at* | *CACNA1G - calcium channel, voltage-dependent, T type, alpha 1G subunit* | *3* |
| *221721_s_at* | *LZTS1 - leucine zipper, putative tumor suppressor 1* | *3* |
| *216646_at* | *DSCC1 - defective in sister chromatid cohesion 1 homolog (S. cerevisiae)* | *1* |
| *211053_at* | *KCNG1 - potassium voltage-gated channel, subfamily G, member 1* | *3* |
| *215259_s_at* | *CADM4 - cell adhesion molecule 4* | *2* |
| *220804_s_at* | *TP73 - tumor protein p73* | *1* |
| *216257_at* | *SERPINB13 - serpin peptidase inhibitor, clade B (ovalbumin), member 13* | *2* |
| *214580_x_at* | *KRT6A - keratin 6A* | *1* |
| *220057_at* | *XAGE1B - X antigen family, member 1B XAGE1C - X antigen family, member 1C* | *1* |
| *210445_at* | *FABP6 - fatty acid binding protein 6, ileal* | *1* |
| *216099_at* | *HTR7P1 - 5-hydroxytryptamine (serotonin) receptor 7 pseudogene 1* | *1* |
| *210122_at* | *PRM2 - protamine 2* | *1* |
| *210451_at* | *PKLR - pyruvate kinase, liver and RBC* | *1* |
| *216707_at* | *Hs.675517 - MRNA; cDNA DKFZp761L0812 (from clone DKFZp761L0812); partial cds* | *1* |
| *221417_x_at* | *S1PR5 - sphingosine-1-phosphate receptor 5* | *2* |
| *216717_at* | *Hs.435815 - Family with sequence similarity 48, member A* | *1* |
| *209855_s_at* | *KLK2 - kallikrein-related peptidase 2* | *1* |
| *215872_at* | *Hs.551210 - Chromosome 14 open reading frame 56* | *1* |
| *210400_at* | *GRIN2C - glutamate receptor, ionotropic, N-methyl D-aspartate 2C* | *1* |

Tabla A.1: Genes con mayor frecuencia

| **Probe-set** | **Gene Symbol - Gene Description** | **Frecuencia** |
|---|---|---|
| *215347_at* | *AK001127 - Homo sapiens cDNA FLJ10265 fis, clone HEMBB1001014* | *1* |
| *210562_at* | *GREB1 - growth regulation by estrogen in breast cancer 1* | *1* |
| *211401_s_at* | *FGFR2 - fibroblast growth factor receptor 2* | *1* |
| *204936_at* | *MAP4K2 - mitogen-activated protein kinase kinase kinase kinase 2* | *1* |
| *221403_s_at* | *INSL6 - insulin-like 6* | *1* |
| *214497_s_at* | *NHLH2 - nescient helix loop helix 2* | *1* |
| *220932_at* | *NM_024073* | *1* |
| *208533_at* | *SOX1 - SRY (sex determining region Y) -box 1* | *1* |
| *220821_at* | *GALR1 - galanin receptor 1* | *1* |
| *219835_at* | *PRDM8 - PR domain containing 8* | *1* |
| *215320_at* | *FAM75C2 - family with sequence similarity 75, member C2* | *1* |
| *220772_at* | *BPESC1 - blepharophimosis, epicanthus inversus and ptosis, candidate 1 (non-protein coding)* | *1* |
| *206523_at* | *CYTH3 - cytohesin 3* | *1* |
| *214530_x_at* | *EPB41 - erythrocyte membrane protein band 4.1 (elliptocytosis 1, RH-linked)* | *1* |
| *211179_at* | *RUNX1 - runt-related transcription factor 1* | *1* |
| *209976_s_at* | *CYP2E1 - cytochrome P450, family 2, subfamily E, polypeptide 1* | *1* |
| *210439_at* | *ICOS - inducible T-cell co-stimulator* | *1* |
| *221411_at* | *HOXD12 - homeobox D12* | *1* |
| *37020_at* | *CRP - C-reactive protein, pentraxin-related* | *1* |
| *214571_at* | *FGF3 - fibroblast growth factor 3* | *1* |
| *217889_s_at* | *CYBRD1 - cytochrome b reductase 1* | *1* |
| *207811_at* | *KRT12 - keratin 12* | *1* |
| *206700_s_at* | *KDM5D - lysine (K)-specific demethylase 5D* | *6* |
| *203595_s_at* | *IFIT5 - interferon-induced protein with tetratricopeptide repeats 5* | *5* |
| *203596_s_at* | *IFIT5 - interferon-induced protein with tetratricopeptide repeats 5* | *5* |

Tabla A.1: Genes con mayor frecuencia

| **Probe-set** | **Gene Symbol - Gene Description** | **Frecuencia** |
|---|---|---|
| *40273_at* | *SPHK2 - sphingosine kinase 2* | *3* |
| *203882_at* | *IRF9 - interferon regulatory factor 9* | *3* |
| *216237_s_at* | *MCM5 - minichromosome maintenance complex component 5* | *3* |
| *201306_s_at* | *ANP32B - acidic (leucine-rich) nuclear phosphoprotein 32 family, member B* | *3* |
| *220212_s_at* | *THADA - thyroid adenoma associated* | *3* |
| *202687_s_at* | *TNFSF10 - tumor necrosis factor (ligand) superfamily, member 10* | *3* |
| *32099_at* | *SAFB2 - scaffold attachment factor B2* | *2* |
| *202269_x_at* | *GBP1 - guanylate binding protein 1, interferon-inducible* | *2* |
| *203153_at* | *IFIT1 - interferon-induced protein with tetratricopeptide repeats 1* | *4* |
| *213294_at* | *EIF2AK2 - eukaryotic translation initiation factor 2-alpha kinase 2* | *2* |
| *201330_at* | *RARS - arginyl-tRNA synthetase* | *2* |
| *213011_s_at* | *TPI1 - triosephosphate isomerase 1* | *3* |
| *202086_at* | *MX1 - myxovirus (influenza virus) resistance 1 interferon-inducible protein p78 (mouse)* | *3* |
| *202153_s_at* | *NUP62 - nucleoporin 62kDa* | *2* |
| *204409_s_at* | *EIF1AY - eukaryotic translation initiation factor 1A, Y-linked* | *2* |
| *203229_s_at* | *CLK2 - CDC-like kinase 2* | *1* |
| *209486_at* | *UTP3 - small subunit (SSU) processome component, homolog (S. cerevisiae)* | *2* |
| *202863_at* | *SP100 - nuclear antigen* | *2* |
| *203281_s_at* | *UBA7 - ubiquitin-like modifier activating enzyme 7* | *2* |
| *210720_s_at* | *NECAB3 - N-terminal EF-hand calcium binding protein 3* | *1* |
| *220847_x_at* | *ZNF221 - zinc finger protein 221* | *1* |
| *214329_x_at* | *TNFSF10 - tumor necrosis factor (ligand) superfamily, member 10* | *1* |
| *205021_s_at* | *FOXN3 - forkhead box N3* | *1* |
| *217789_at* | *SNX6 - sorting nexin 6* | *1* |

Tabla A.1: Genes con mayor frecuencia

| **Probe-set** | **Gene Symbol - Gene Description** | **Frecuencia** |
|---|---|---|
| *209040_s_at* | *PSMB8 - proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional peptidase 7)* | *1* |
| *219229_at* | *SLCO3A1 - solute carrier organic anion transporter family, member 3A1* | *1* |
| *205483_s_at* | *ISG15 - ubiquitin-like modifier* | *1* |
| *205350_at* | *CRABP1 - cellular retinoic acid binding protein 1* | *1* |
| *202488_s_at* | *FXYD3 - domain containing ion transport regulator 3* | *1* |
| *210046_s_at* | *IDH2 - isocitrate dehydrogenase 2 (NADP+), mitochondrial* | *1* |
| *206782_s_at* | *DNAJC4 - DnaJ (Hsp40) homolog, subfamily C, member 4* | *1* |
| *209848_s_at* | *PMEL - premelanosome protein* | *1* |
| *203566_s_at* | *AGL - amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase* | *1* |
| *213399_x_at* | *RPN2 - ribophorin II* | *1* |
| *213293_s_at* | *TRIM22 - tripartite motif containing 22* | *1* |
| *212203_x_at* | *IFITM3 - interferon induced transmembrane protein 3* | *1* |
| *204410_at* | *EIF1AY - eukaryotic translation initiation factor 1A, Y-linked* | *1* |
| *216292_at* | *AK024455* | *1* |
| *210950_s_at* | *FDFT1 - farnesyl-diphosphate farnesyltransferase 1* | *1* |
| *213039_at* | *ARHGEF18 - Rho/Rac guanine nucleotide exchange factor (GEF) 18* | *1* |
| *207132_x_at* | *PFDN5 - prefoldin subunit 5* | *1* |
| *214453_s_at* | *IFI44 - interferon-induced protein 44* | *1* |
| *218927_s_at* | *CHST12 - carbohydrate (chondroitin 4) sulfotransferase 12* | *1* |
| *217329_x_at* | *COX7B - cytochrome c oxidase subunit VIIb* | *1* |
| *202349_at* | *TOR1A - torsin family 1, member A (torsin A)* | *1* |
| *201641_at* | *BST2 - bone marrow stromal cell antigen 2* | *1* |
| *200822_x_at* | *TPI1 - triosephosphate isomerase 1* | *1* |

Tabla A.1: Genes con mayor frecuencia

| **Probe-set** | **Gene Symbol - Gene Description** | **Frecuencia** |
|---|---|---|
| *208966_x_at* | *IFI16 - interferon, gamma-inducible protein 16* | *1* |
| *53720_at* | *C19orf66 - chromosome 19 open reading frame 66* | *1* |
| *220847_x_at* | *ZNF221 - zinc finger protein 221* | *1* |
| *215061_at* | *ENSG00000258539* | *7* |

# Referencias


Alpaydin, E. (2004). *Introduction to machine learning*. The MIT Press. 21

Astengo, E.A.C. (2012). *Gene Subset Selection from Microarray Data in Amyotrophic Lateral Sclerosis Disease Classification*. Ph.D. thesis, Instituto de ingenieria UABC. 13

Balboni, I., Patel, P., Limb, C., Luera, N., G.Morgan & Utz, P. (2007). Autoantigen microarray analysis of sera from juvenile dermatomyositis patients: Association with an sle-like type i interferon signature. **1**, 1. 10

Blanco, C., Izquierdo, A., Domínguez, C., Fernández, S. & Ochaita, L. (2011). Dermatomiositis. *estudio y seguimiento de 20 pacientes*, **1**, 2–5. 1, 2, 5, 8

Blaszczyk, M., Jablonska, S., J.P., C., Collison, C., Sinal, S., Feldman, B. & Reichlin, M. (2000). Protocolos diagnósticos y terapéuticos en dermatología pediátrica. *Conectivopatías: Dermatomiositis*, **1**, 113–119. 5, 8, 9

Calore, E.E., Cavaliere, M.J. & Perez, N.M. (1997). Muscle pathology in juvenile dermatomyositis. *São Paulo Medicai Journal/RPM*, 1555–1559. 45

Chen, Y.W., Shi, R., Geraci, N., Shrestha, S., Dressman, H.G. & Pachman, L.M. (2008). Duration of chronic inflammation alters gene expression in muscle from untreated girls with juvenile dermatomyositis. **1**, 3–7. 10

Clarke, W.R., Lachenbruch, P.A. & Broffitt, B. (1979). How nonnormality affects the quadratic discriminant function. *Communications in Statistics-Theory and Methods*, **8**, 1285–1301. 29

Cros, D., Pearson, C. & Verity, M.A. (1980). Polymyositis-dermatomyositis diagnostic and prognostic significance of muscle alkaline phosphatase. *American Association of Pathologists*, **101, No.1**, 159–176. 45

Duda, R. & Hart, P. (2001). *Pattern Recognition and Scene Analysis*. John Wiley and Sons. v, 33

E., G., de Andrés I. & I., I. (2001). Cathepsins are upregulated by ifn-gamma/stat1 in human muscle culture: a possible active factor in dermatomyositis. *Department of Neurology, Hospital de la Santa Creu i Sant Pau, Universitat Autonoma Barcelona, Spain*, 847–855. 45

Feber (2005). Conectivopatías: Dermatomiositis, in http://www.aeped.es/protocolos/dermatolog\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{\OT1\i\global\mathchardef\accent@spacefactor\spacefactor}\accent19\OT1\i\egroup\spacefactor\accent@spacefactora. 6

Fedczyna, T., Lutz, J. & Pachman, L. (2001). Expression of tnfalpha by muscle fibers in biopsies from children with untreated juvenile dermatomyositis: association with the tnfalpha-308a allele. **1**, 1. 12

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320–328. 31

GENECARDS (2013). Weizmann institute of science. In http://www.genecards.org/. 46

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, **286**, 531–537. 13

Gonzalez, F. & Belanche, L. (2008). Gene subset selection in microarray data using entropic filtering for cancer classification. *Expert Systems*, **99**, 99–99. 49

Gordon, G.J., Jensen, R.V., Hsiao, L.L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J. & Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, **62**, 4963–4967. 13

Honavar, V. (2006). *Articial Intelligence: An Overview*. Wiley-Interscience. 2

Illa, I., Gallardo, E., Gimeno, R., Serrano, C., Ferrer, I. & Juarez, C. (1997). Signal transducer and activator of transcription 1 in human muscle. *Amnerican Jounal of Pathology*, **Vol. 151**, 81–88. 45

J., C. & R., P. (1995). Juvenile dermatomiositis. *Pediatric Rheumatology*, **3 ed. Philadelphia: Saunders.**, 323–358. 6

Kannee, C.E. (2010). Capitulo 68: Dermatomiositis juvenil, in http://piel-l.org/libreria/item/891. 5, 6

Knudsen, S. (2006). *Cancer diagnostics with DNA microarrays*. Wiley-Liss. 13, 15

Kumar, V., Abbas, A.K., Fausto, N. & Mitchell, R. (2007). *Robbins Basic Pathology*. Saunders Elsevier, 8th edn. 46

Kurgan, K.J.C..W.P..R.W.S..L.A. (2007). *Data Mining A Knowledge Discovery*. Springer Science+Business Media, LLC. 22, 23

Lachenbruch, P.A. & Mickey, M.R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11. 31

Lele, S., Richtsmeier, J., Babu, G., E.Feigelson, Crowley, J., Benedetti, S.G.J., Fairclough, D.L., Pronzato, L., Wynn, H., Zhigljavsky, A., Basford, K., Tukey, J., Waterman, M., Gilks, W., Richardson, S., Spiegelhalter, D., Speed, T., Bailer, A. & Piergorsch, W. (2003). *Interdisciplinary statistics statical analysis of gene expression microarray data*. Chapman & Hall/CRC. 27

Liu, H. & Motoda, H. (2008). *Computational Methods of Feature Selection*. Chapman & Hall/CRC. 22, 23

LM., P. (2004). Dermatomiositis juvenil. en: Behrman re. *Elsevier*, 813–816. 7

Lourdes, E. (2009). Aplicación de herramientas bioinformáticas en el estudio de las enfermedades genéticas humanas. *Curso Mitolab-Ciberer*, 1. 3

Mamyrova, G., O'Hanlon, T.P., Monroe, J.B., Carrick, D.M., Malley, J.D., Adams, S., Reed, A.M., Shamim, E.A., James-Newton, L., Miller, F.W. & Rider, L.G. (2006). Immunogenetic risk and protective factors for juvenile dermatomyositis in caucasians. *for the Childhood Myositis Heterogeneity Collaborative Study Group*, **1**, 1–5. 1, 5, 9

McLachlan, G., Do, K.A. & Ambroise, C. (2005). *Analyzing microarray gene expression data*, vol. 422. Wiley-Interscience. 15, 31

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (1994). Machine learning, neural and statistical classification. 21

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Higher Education. 28, 29

MP, P. & de Oca F., M. (2002). Dermatomiositis juvenil. *Acta Pediatr Esp*, 583–595. 6, 7

NCBI (2007). National center of biothecnology information. In `http://www.ncbi.nlm.nih.gov/`. 33, 34, 46

Pudil, P., Novovičová, J. & Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters*, **15**, 1119–1125. 24

Reed, A.M. & Mason, T. (2005). Recent advances in juvenile dermatomyositis. **1**, 1–4. 5

RM, S. & HR., M. (1989). Estudios microvasculares seriados en la dermartomiositis infantil. *Pediatrics (ed. esp)*, 99–103. 6

S., P. & WL., N. (1975). Fatal pulmonary involvement in dermatomyositis. *Am J Dis Child*, 723–726. 7

Schena, M. (2003). *Microarray Analysis*. John Wiley & Sons. 13

Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W. & Zhao, Y. (2003). *Design and Analysis of DNA Microarray Investigations*. Springer-Verlag. 14, 15, 16, 17, 18, 19, 21

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. & Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209. 13

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 111–147. 31

Taylor & Francis Group, L. (2008). *Computational Methods of Feature Selection*. Chapman & Hall/CRC. 23

Tezak, Z., Hoffman, E.P., Lutz, J.L., Fedczyna, T.O., Stephan, D., Bremer, E.G., Krasnoselsja-Riz, I., Kumar, A. & Pachman, L.M. (2011). Gene expression profiling in dqa1*0501$^{+}$ children with untreated dermatomyositis: A novel model of pathogenesis. **1**, 1. 11

Trias Capella, R. (1993). Inteligencia artificial en medicina: Estado actual y perspectivas. *Medicina clínica*, **100**, 45–46. 2

van't Veer, L.J., Dai, H., Van De Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T. *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, **415**, 530–536. 13

Welch, B.L. (1947). The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, **34**, 28–35. 35