UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

INSTITUTO DE INGENIERÍA

Selección de Genes en Datos Obtenidos por Microarreglos para la clasificación de la Esclerosis Lateral Amiotrófica

T E S I S

que presenta para obtener el grado de MAESTRO EN CIENCIAS

Edgar Armando Castro Astengo

DIRECTOR DE TESIS:

Dr. Félix Fernando González Navarro

Mexicali, B. C. Enero del 2013

RESUMEN de la Tesis de EDGAR ARMANDO CASTRO ASTENGO, presentada como requisito parcial para la obtención del grado de MAESTRO EN CIENCIAS en CIENCIAS DE LA COMPUTACIÓN. Mexicali, Baja California, México. Enero del 2013.

Selección de Genes en Datos Obtenidos por Microarreglos para la clasificación de la Esclerosis Lateral Amiotrófica

Resumen aprobado por:

_____

Dr. Félix Fernando González

Director de Tesis

La esclerosis lateral amiotrófica (Amyotrophic Lateral Sclerosis, ALS) es una enfermedad degenerativa de las neuronas motoras. La ocurrencia es aproximadamente 5 personas de cada 100,000. Comúnmente se presenta después de los 40 años de edad, aunque existen pocos casos de niños y jóvenes. A la fecha no existe un tratamiento o cura, por lo que el deceso del paciente resulta en fallas respiratorias aproximadamente 3 a 5 años después de los primeros síntomas. Las causas exactas que provocan la enfermedad son desconocidas pero en cerca del 20% de los casos se ha descubierto que la mutación de ciertos genes son responsables. Actualmente el uso del análisis de expresión de genes es una moderna herramienta que ha demostrado ser altamente informativa en la determinación de los genes más influyentes en procesos celulares. Su aplicación en el estudio del ALS es escasa, representando con ello una oportunidad de contribución científica. En el presente trabajo se propone el desarrollo de modelos computacionales basados en técnicas de aprendizaje de máquina para la identificación de las causas genómicas de la enfermedad. De esta manera se podrían crear mejores formas de diagnóstico y/o tratamientos genéticos a fin de desarrollar fármacos para su control.

Palabras clave: esclerosis lateral amiotrófica, Datos de Expresión Genética con Microarreglos, Selección de grupos genéticos, Aprendizaje de Maquina, Clasificación, Remuestreo Bootstrap.

ABSTRACT of the thesis, presented by EDGAR ARMANDO CASTRO ASTENGO, in order to obtain the MASTER IN SCIENCE DEGREE in COMPUTER SCIENCE. Mexicali, Baja California, México. January 2013.

Microarray Gene Subset Selection in Amyotrophic Lateral Sclerosis Classification

Approved by:

_____

Dr. Félix Fernando González

Thesis Supervisor

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease causing a progressive loss of motor neurons. The disease prevalence is 5 per 100,000 people. The onset age is usually at 40 years old, however there are cases of children and young people alike. There is no cure and it leads generally to death resulting from respiratory failure in approximately 3-5 years after the first symptoms. The exact causes of the disease are still unknown, but almost 20% of the cases have shown gene mutations on the patients. The use of gene expression analysis is a powerful tool to discover the most relevant genes in a cellular process. Its application in the ALS research is scarce, therefore it is a great opportunity to contribute to the scientific community. This work proposes the development of computer models based in machine learning techniques to identify the genomic causes of the disease. As a result, it can help to find new ways to diagnose or to develop genetic treatment to control it.

Keywords: Amyotrophic Lateral Sclerosis, Microarray Gene Expression Data, Gene Subset Selection, Machine Learning, Classification, Bootstrap Resampling.

AGRADECIMIENTOS

A mis padres, Armando Castro Flores e Irma Leticia Astengo Ramos por su constante apoyo incondicional a lo largo de mi vida con un inmenso amor que ha ayudado a mi superación personal.

A mi profesor el Dr. Félix Fernando González por su paciencia a lo largo del curso de maestría, por su disposición de ayudar en el desarrollo de este documento, gracias a él fue posible el completar exitosamente el grado de maestría.

A mi hermana Arq. Irma Cristina Castro Astengo y mi gran amigo el Ing. Saúl Delgado Pardo por ayudarme en la redacción de este documento y otros documentos a lo largo de este curso.

Al Conacyt por su ayuda financiera, gracias a esta organización se me facilito en gran manera el que yo pudiera obtener el grado de maestría.

Índice

Índice de figuras

Índice de tablas

Capitulo 1
Introducción

1.1. Planteamiento del problema

ALS es una enfermedad que causa degeneración y perdida de las neuronas motoras superiores (Upper Motor Neuron, UMN) y de las neuronas motoras inferiores (Lower Motor Neuron, LMN), que son las células de los nervios motores de la médula espinal (células del asta anterior), tronco cerebral (nervios craneales selectos) y el cerebro (área motora o corteza cerebral) el cual contrae los músculos que están atados al esqueleto y controlan los músculos involucrados en el movimiento voluntario del cuerpo.

Los síntomas del ALS varían de una persona a otra según el grupo de músculos que son afectados por la enfermedad. Tropezar, dejar caer cosas, fatiga anormal en los brazos y/o piernas, problemas al hablar, dificultad para hablar en voz alta, ataques incontrolables de risa o llanto, calambres musculares y espasmos; son síntomas del ALS. ALS suele iniciar en las manos y causar problemas al vestirse, bañarse, u otras simples tareas. Puede progresar a más de un lado del cuerpo, en general procede a los brazos o las piernas. Si empieza en los pies, el caminar se dificulta. ALS también puede comenzar en la garganta, causando dificultad para tragar. Las personas afectadas con ALS no pierden su capacidad de ver, oír, tocar, oler o saborear. La vejiga y los músculos de los ojos de la persona no se ven afectados, ni la capacidad sexual. La enfermedad no afecta la mente de la persona [4].

La enfermedad es implacablemente progresiva, concluye con la muerte por parálisis respiratoria, la media de supervivencia es de 3.5 años. Hay casos muy raros de estabilización e incluso de regresión. En la mayoría de las comunidades existe una prevalencia de 1 - 9 por cada 100,000 personas. En los Estados Unidos y Europa, los hombres tienen más frecuencia de ser afectados que las mujeres. [2].

Durante años, los expertos han tratado de encontrar factores comunes en las personas que desarrollan ALS, tales como toxinas ambientales, riesgos profesionales, los

lugares de trabajo o de residencia, etcétera. Hasta el día de hoy, la evidencia de las causas que provocan la enfermedad o factores de riesgo han sido frustrantemente inciertos. [5] Sin embargo, estudios epidemiológicos han encontrado factores de riesgo que pueden ser causas de esta enfermedad, por ejemplo, la exposición a los plaguicidas e insecticidas, fumar, etc., inclusive, existe un informe mostrando que el servicio en el ejército puede ser un factor de riesgo. Mientras que la ALS es angustiosamente un trastorno esporádico, entre un 10 - 20% de los casos es debida a la herencia de un rasgo autosómico[1] dominante. [2].

La tecnología de microarreglos ha cambiado drásticamente en como los biólogos analizan el código genético, esta herramienta permite monitorear los niveles de expresión genética de cualquier organismo.

El código genético del ser humano tiene una gran diversidad genética, a pesar de que la tecnología de microarreglos actual es capaz de representar numéricamente casi todo el genoma humano, es necesario técnicas poderosas para poder distinguir los genes involucrados en el progreso de la enfermedad.

Los métodos de aprendizaje de máquina en análisis de expresión de genes, son herramientas muy poderosas que ayudan a identificar genes de interés en muestras de tejido. Es la intensión de este documento el hacer uso de estas técnicas para encontrar nueva evidencia genética vinculada con esta terrible enfermedad.

1.2. Descripción de la tesis

El segundo capítulo explica con un mayor detalle la enfermedad del ALS, como se descubrió, el proceso que sigue en la degradación neuronal, como diagnosticarla y los avances que se han hecho para combatirla. Se le dedicó todo un capitulo por la gran complejidad de la enfermedad, a pesar de todos los avances en la medicina moderna, existe muy poco conocimiento sobre esta enfermedad como para ayudar a los pacientes que la contraen.

[1] Cualquier cromosoma que no sea sexual

El capítulo tres habla sobre el análisis de expresión de genes usando microarreglos. Para comprender como es que funcionan los microarreglos y el entender los resultados que arrojan, es necesario entender la función que tiene el material genético en los organismos. Este capítulo solo presenta los principales descubrimientos que permitieron el avance de este tipo de tecnologías, pero la descripción de la tecnología de microarreglos se hace con mayor detalle, desde cómo se adquiere la información del chip de microarreglos hasta terminar con una matriz numérica representando la expresión genética de una muestra orgánica.

En el cuarto capítulo se habla del aprendizaje de máquina, sus principales paradigmas, se explica la definición de un clasificador y se mencionan los clasificadores que se usarán en la experimentación. También se describen las técnicas de selección de características, donde se habla del concepto de la maldición de la alta dimensión que está adherida a muestras con gran número de características como las proporcionadas por los microarreglos, para combatir este problema se describen las técnicas de bootstrap que son capaces de lidiar exitosamente con la dificultad establecida por la maldición de la alta dimensión.

El quinto capítulo describe los experimentos realizados, se da a conocer los tipos de datos que se utilizaron y cómo se hizo uso de las técnicas descritas en el cuarto capítulo para obtener los resultados incluidos en este documento, todo esto con el propósito de validar la selección de genes al final del capítulo.

En el último capítulo se muestran las conclusiones finales de este documento, también se plantean las posibles opciones que existen para mejorar la investigación en futuros trabajos.

1.3. Objetivos

El objetivo general de este documento es determinar, mediante el uso de métodos de inteligencia artificial, patrones de genes que muestren comportamiento atípico en el ALS, utilizando datos de expresión genética de microarreglos.

Los objetivos particulares son los siguientes:

- Determinar un conjunto de genes a partir de datos de expresión genética que permitan clasificar o diferenciar, con la mayor exactitud posible, muestras de ALS de muestras normales.

- Establecer diferentes escenarios de experimentación para problemas con bajo número de muestras, y de alta dimensión o número de variables.

Capitulo 2

Esclerosis lateral amiotrófica

La palabra “amiotrofia” proviene de raíces griegas, la cual significa “músculos sin nutrimiento” y se refiere a la perdida de la señal de células nerviosas que comúnmente se le envían a las células musculares. “Lateral” significa “hacia los lados” lo cual se refiere a la localización del daño en la espina dorsal. “Esclerosis” significa “endurecimiento” y se refiere a la forma que toma la espina dorsal cuando la enfermedad se encuentra muy avanzada. [5]

La esclerosis lateral amiotrófica es conocida también como “La enfermedad de Lou Gehrig's” en los Estados Unidos, en la mayoría de los países del mundo, es conocida como ALS. En el Reino Unido, Australia, Nueva Zelanda y el Sur de Africa es conocida como “enfermedad de las neuronas motoras” (Motor Neuron Disease, MND). Internacionalmente se le refiere a la enfermedad como ALS/MND y se le refiere como enfermedad neurodegenerativa o neuromuscular [1].

2.1. Características clínicas

Las características clínicas del ALS son consecuencia directa de la pérdida progresiva de las neuronas motoras altas y bajas, también se debe a la denervación y la re-inervación de los músculos. La denervación muscular ocurre cuando las neuronas pierden contacto con los músculos, los nervios al ser desconectados intentan innatamente regenerarse, a esto se le conoce como re-inervación.

Mientras la re-inervación compense la denervación muscular, la debilidad clínica puede pasar por desapercibida. Sin embargo, así como las unidades motoras crecen y al mismo tiempo decrecen, las primeras consecuencias que afectan los músculos pueden causar fatiga más rápidamente que los músculos con unidades motoras normales, consecuentemente, uno de los primeros síntomas del ALS puede ser el incremento de fatiga. Así como los números de las unidades motoras decrecen cada vez más, la re-

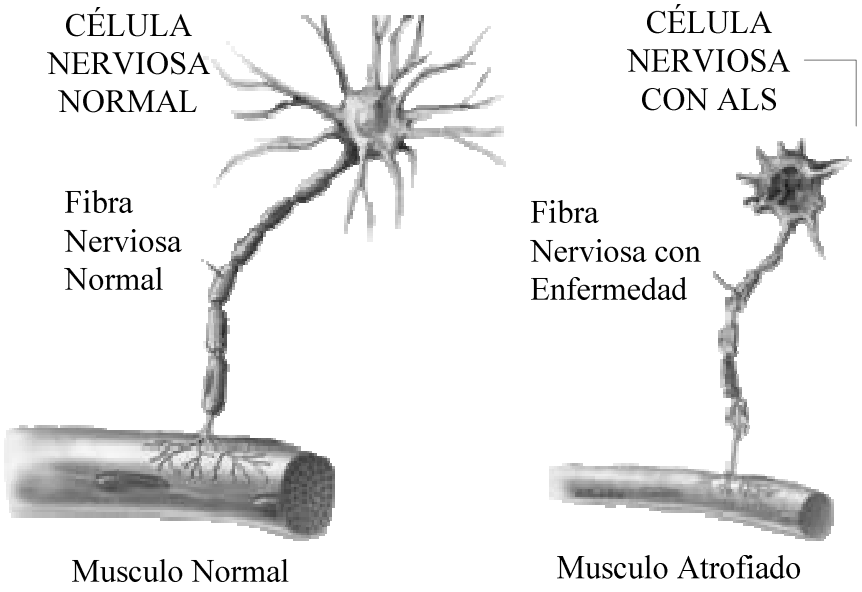

Figura 2.1. Fotografía demostrando el daño causado por ALS [11]

inervación no puede compensar la denervación, la debilidad muscular se vuelve permanente y los músculos afectados gradualmente se atrofian.

Las partes del cuerpo afectadas por los primeros síntomas del ALS dependen de las neuronas motoras que son dañadas primero. Aproximadamente el 75% de los pacientes con la enfermedad empiezan a mostrar debilidad en músculos de las extremidades superiores o inferiores, los pacientes pueden experimentar torpeza, dificultad para realizar tareas sencillas como el abrocharse la camiseta o el escribir. Los músculos afectados pueden desarrollar espasmos, calambres, o rigidez. En el otro 25% de los casos el sitio donde se asienta la enfermedad en el núcleo motor del tronco cerebral y/o en neuronas del córtico bulbar. Estos pacientes experimentan primero dificultad para hablar, arrastran las palabras, dificultad para tragar y perdida de la movilidad de la lengua. La figura 2 muestra las áreas afectadas por la enfermedad.

Sin importar la región del cuerpo que es afectada primero, la debilidad muscular y la atrofia se extiende a otras partes del cuerpo mientras que la enfermedad progresa. Pacientes experimentan dificultad para moverse, comer, formar palabras; la inefectividad

para tragar saliva provoca babeo excesivo; la debilidad en los músculos del aparato respiratorio provoca ineficiencia en la inhalación y exhalación de aire.

En la exploración neurológica, puede existir una combinación de signos en el deterioro de neuronas motoras superiores e inferiores (UMN y LMN); síntomas de la degeneración del UMN incluyen un aumento del tono muscular (espasticidad), reflejos exagerados (hiperreflexia), signos en el área piramidal puede presentarse como respuestas extensas plantares, pérdida de la destreza en el momento de usar fuerza normal. Los síntomas de la neurodegeneración LMN incluyen espasmos y atrofia muscular, disminución o ausencia de reflejos en los tendones (Fasciculaciones). Las fasciculaciones no son

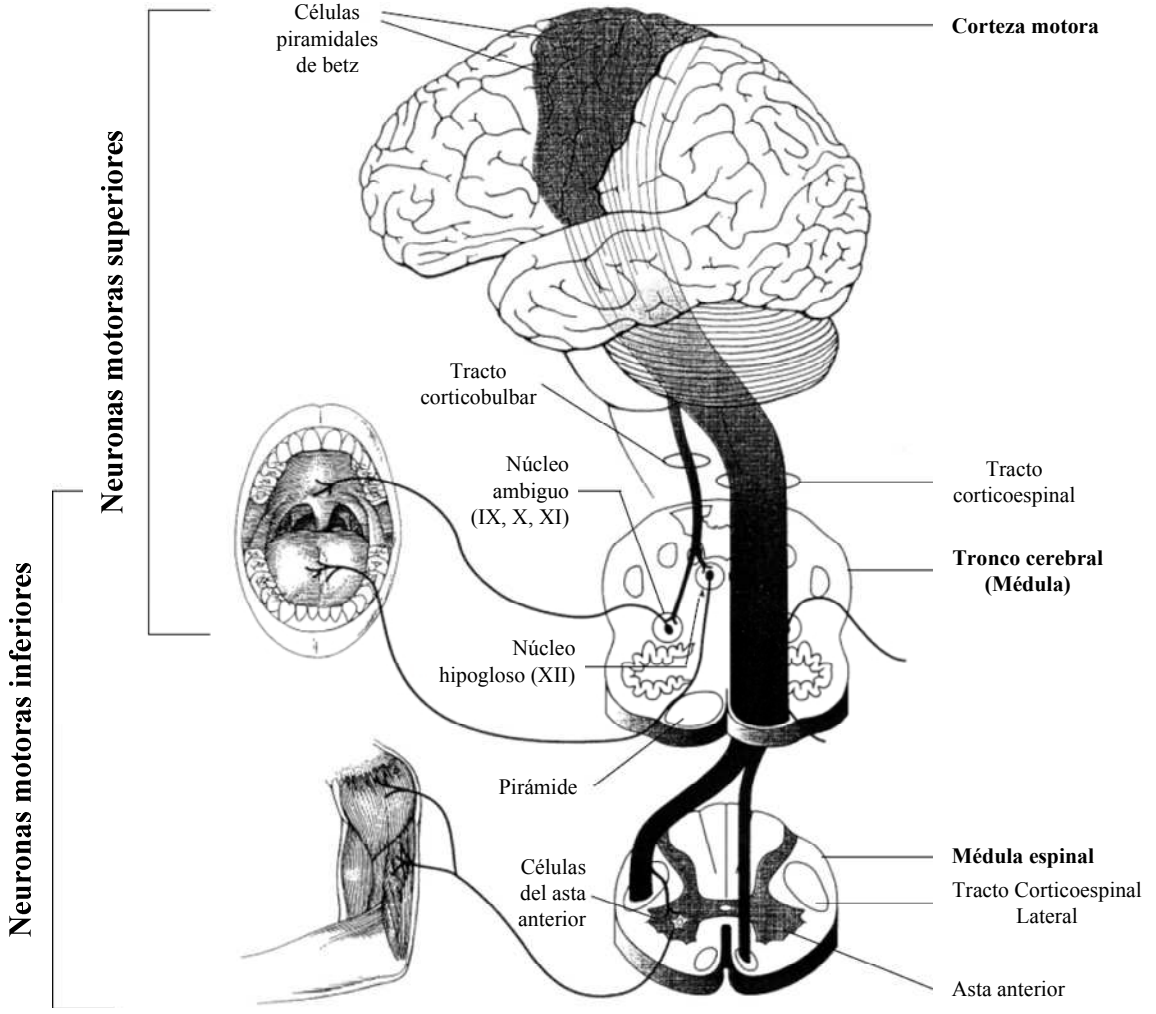

Figura 2.2. Neuronas motoras afectadas por ALS y sus respectivas funciones en el cuerpo humano [3]

específicas para el ALS, pero se pueden encontrar en otras enfermedades del sistema nervioso periférico, a veces se pueden encontrar en individuos sanos, mas sin embargo, la constancia de grandes y rápidas fasciculaciones en muchos sitios es un poderoso indicativo de la enfermedad [6].

Alrededor del 15 - 45% de los pacientes experimentan síntomas emocionales incontrolables, como la risa inapropiada, el llanto o la sonrisa, características comunes del complejo del síndrome pseudobulbar, que se asocia con alteraciones del habla y la ingestión, distintivo en la participación bilateral de vías cortico bulbares [6].

Aunque la secuencia de los síntomas emergentes y la tasa de la progresión de la enfermedad son variables, la mayoría de los pacientes al paso del tiempo desarrollaran discapacidad en todas las tareas motoras, no serán capaz de pararse o caminar, o usar sus manos y brazos. La dificultad para tragar y masticar impide al paciente la habilidad de comer y aumenta el riesgo de asfixia e infecciones pulmonares. La pérdida de peso es una consecuencia del deterioro nutricional y la pérdida de masa muscular. Debido a que la enfermedad por lo general no afecta a las habilidades cognoscitivas, los pacientes están conscientes de la pérdida progresiva de su movilidad y ocasiona que lleguen a ser ansiosos y deprimidos. Por otro lado, la preservación cognoscitiva permite a los pacientes tomar decisiones, las relacionadas con su propio final de vida, por si mismos, incluso después de alcanzar un estado de inmovilidad total (locked-in) [6].

Hasta un 50% de pacientes experimentan dificultades leves con la generación de palabras, la atención, o la toma de decisiones. De un 5 - 10% desarrollan demencia frontotemporal abierta, caracterizada por cambios en la personalidad y el deterioro de las funciones ejecutivas y el lenguaje; esto es más común entre aquellas personas con antecedentes familiares de demencia [6].

A medida que el diafragma y los músculos intercostales se debilitan, la presión inspiratoria disminuye. La mayoría de los pacientes con ALS mueren de fallo respiratorio o neumonía. La muerte generalmente ocurre dentro de los dos a cinco años del diagnóstico

(mediana de 27 meses). Menos del 5% de los pacientes sobreviven más de 10 años. El 6% tienen una forma detenida de la enfermedad. En general, la supervivencia es mejor en pacientes jóvenes y aquellos donde la enfermedad inició en las extremidades. Curiosamente, hay pocas funciones motoras que usualmente no son afectadas por el ALS: el control de los músculos del ojo es de la función conservada, aunque en algunos pacientes con una duración extremadamente larga de la enfermedad (más de 20 años) también pueden perder el control de los ojos. El control de la vejiga e intestinal por lo general se mantienen, aunque como consecuencia de la inmovilidad y cambios en la dieta, problemas intestinales como el estreñimiento pueden requerir de un atención intensiva [6].

2.2. Diagnóstico

No existe forma rápida de diagnosticar ALS, la mejor manera de diagnosticarla es usar el criterio “El Escorial”, también conocido como la prueba “Airlie House”, la cual consiste en seguir la neurodegeneración del paciente usando una serie de estudios médicos.

El criterio “El Escorial” presentado por la asociación del ALS requiere la presencia de la siguiente evidencia para realizar un diagnóstico:

1. Signos de la degeneración en las neuronas motoras bajas (LMN) mediante pruebas clínicas como electro-físicas o examinación neuropatológica.

2. Señales de la degeneración de las neuronas motoras altas (UMN) por examinación clínica.

3. Señales del progreso de la neurodegeneración dentro de la misma región o hacia otras regiones del cuerpo.

4. La ausencia de evidencia electro-física de otras enfermedades que pueden explicar los signos de la degeneración en el LMN y UMN.

5. Evidencia en imagines neuronales del proceso de la enfermedad que explique la evidencia del progreso de la enfermedad observado clínicamente y por señales electro-físicas.

Existen 4 categorías en el criterio “El Escorial”, el diagrama de la figura 2.3 presenta el proceso que se debe seguir si se desea diagnosticar la enfermedad exitosamente, a continuación se explicarán las categorías con mayor detalle:

- ALS Definitivo: Se define solamente con pruebas clínicas que muestran la presencia de la enfermedad en la región bulbar y otras regiones de la espina dorsal, las cuales señalan una clara degeneración de la UMN y LMN, estas pruebas deben por lo menos marcar 3 áreas diferentes en la región espinal. Los determinantes importantes para determinar el diagnóstico definitivo del ALS es la ausencia de señales electro-físicas, imágenes neuronales y pruebas de laboratorio indicando señales de la degeneración del UMN y LMN en múltiples regiones.

- ALS Probable: Está definido por señales de neurodegeneración en 2 región, tanto UMN y LMN. Mientras que las regiones pueden ser diferentes, los signos del UMN deben ser en el rostro. Existen múltiples combinaciones donde signos de UMN y LMN pueden estar presentes en un paciente para ser diagnosticarlo con ALS probable.

- Posible ALS: Se define cuando existe evidencia de degeneración en solo una zona, puede que señales de UMN y LMN marcan solamente una región. Existen casos especiales donde se detectan degeneración parcial en la espina dorsal y el diagnóstico es ALS probable.

- Sospecha de ALS: Se manifiesta cuando existe degeneración por lo menos en una región pero no existe suficiente evidencia patológica.

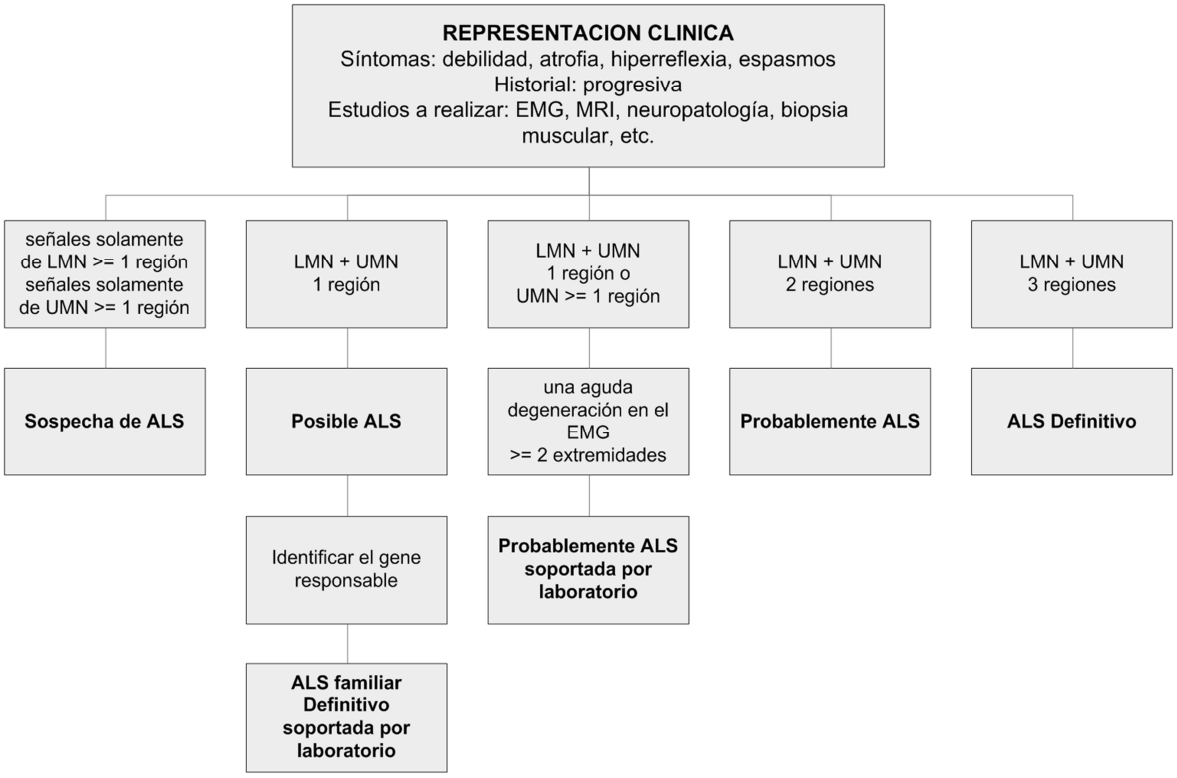

Figura 2.3. Criterio “El Escorial” [3]

2.3. Avances realizados para combatir la enfermedad.

La enfermedad no es muy conocida por la población en general, pero es muy popular dentro de la comunidad médica. El Dr. Jean-Martin Charcot, renombrado neurólogo francés conocido como ‘El padre de la neurología’, publicó la primera descripción clínicas de las características patológicas del ALS en 1874. Basado en sus descubrimientos en autopsias de los músculos y la medula espinal, el nombró la enfermedad [1].

No fue hasta que el popular beisbolista Estadounidense “Lou Gehrig” (ver figura 2.4) contrajo la enfermedad y tuvo que retirarse en 1939 de las grandes ligas y murió después de una rápida degeneración muscular en 1941, que la población de Estados Unidos cobro conciencia de la enfermedad y se le invirtió recursos para combatirla.

Figura 2.4. Lou Gehrig, jugador de los Yankees de Nueva York que presentó la enfermedad del ALS en la década de 1930

Lambert E.H. y Mulder D.W. formularon un criterio para diagnosticar la enfermedad usando electro-diagnósticos (EDX) a finales de la década de 1950, este criterio fue revisado en España y se le incluyeron técnicas modernas y se le renombró al criterio “El Escorial” en honor a la sede de esta revisión.

Durante años, todos los esfuerzos para encontrar información sobre la enfermedad no tuvieron ningún avance, se intentó localizar algún virus responsable, se trató de encontrar algún indicio que indicara que estuviese asociada con la polio; después de extenuantes pruebas y tratamientos experimentales sin éxito. Actualmente la droga Rilutek® ha presentado cierto nivel de éxito, pero por la fuerte controversia acerca de su efectividad y el alto costo de la droga, no todos los médicos la utilizan en sus pacientes.

Según el diario EPMA [6] existen dos principales tipos de ALS, ALS Familiar (Familial ALS, FALS) y ALS Esporádico (Sporadic ALS, SALS).

- ALS Familiar: Aproximadamente un 5 - 10% de los casos de ALS son familiares con un patrón de herencia mendeliana. En la mayoría de los casos

la enfermedad se transmite de forma autosómica dominante. La edad media de los primeros síntomas es 10 a 20 años más temprano en pacientes con FALS que en pacientes con la enfermedad esporádica, y la variabilidad entre las familias es mayor que la variabilidad dentro de las familias. La mayoría de los casos de FLAS pueden ser fácilmente distinguibles de la enfermedad esporádica, mientras que en otros tienen fenotipos únicos.

- ALS Esporádico: Las causas del SALS aún se desconocen. Se considera que es una enfermedad multifactorial, donde la aparición y la progresión de la enfermedad es debida a factores ambientales y genéticos.

El descubrimiento más importante fue la mutación de la enzima SOD1 en 1991, dicha enzima normalmente protege las células contra los daños causados por los procesos celulares, las investigaciones resientes apuntan que una enzima defectuosa SOD1 no parece producir la enfermedad por la pérdida de esta función de protección, sino a una propiedad toxica adquirida por la mutación [13]. Varios marcadores polimórficos fueron examinados

Tabla 2.1. Los genes y sus respectivos vínculos en los casos de ALS [6]

Nombre de la Enfermedad	Localización Genética	Nombre de la Proteína	Símbolo del Gen	Herencia	Incidencia en FALS
ALS1	21q22.1	Cu–Zn Superóxido dismutasa 1	SOD1	Dominante	10–20%
ALS2	2q33	Alsin	ALS2	Recesivo	Raro
ALS3	18q21			Dominante	Raro
ALS4	9q34	Senataxin	SETX	Dominante	Raro
ALS5	15q15.1–21.1			Recesivo	Raro
ALS6	16q12	Fundido en el sarcoma	FUS	Ambos	4%–5%
ALS7	20p13			Dominante	Raro
ALS8	20q13.3	Proteina Vesicula-asociada de la membrana Proteína-asociada a la proteina B	VAPB	Dominante	Raro
ALS9	14q11.2	Angiogenin	ANG	Dominante	Raro
ALS10	1p36.2	TAR-DNA	TARDBP	Dominante	1–4%
ALS?	2p13	Dynactin 1	DCTN1	Dominante	Raro
ALS-FTD1	9q21–22			Dominante	
ALS-FTD2	9p13.2–21.3			Dominante	
ALS-FTD con parkinsonismo	17q21.1	Proteína tau asociada a microtúbulos	MAPT	Dominante	Raro
ALS-X	Xcen			Dominante	Raro

ALS: Amyotrophic Lateral Sclerosis; FTD: Frontotemporal Dementia

para enlazar ALS familiar, los cuales llevaron a la exclusión de muchas regiones del genoma; mas sin embargo, varios resultados fueron moderadamente buenos en los marcadores para el cromosoma 21.

La realización del mapa de haplotipos del genoma humano y el desarrollo de la tecnología avanzada de genotipado ha permitido la realización de estudios de asociación amplia del genotipo (Genotype-Wide Association Studies, GWAS). La tabla 2.1 muestra los genes responsables de algunos casos de ALS que se han descubiertos hasta el día de hoy. [6]

Casi todos los genes descubiertos en FALS, se han encontrado gracias a marcadores químicos, sin embargo, ALS-FTD2, se identificó con técnicas de análisis en expresión de genes usando microarreglos. La demencia frontotemporal (Frontotemporal Dementia, FTD) es un trastorno neurológico, la cual se presenta con un cambio de personalidad y un comportamiento socialmente inadecuado, tiene relativa preservación de las funciones cognoscitivas y de memoria. Eventualmente ocurre un déficit cognoscitivo global y frecuentemente la muerte ocurre por inmovilidad e infección respiratoria. De igual forma el FTD se ha relacionado con mutaciones genéticas. Caroline Vance et al. encontró un grupo de genes relacionados con ambas enfermedades a partir de una muestra de 16 familias que padecían de ALS y FTD. La localización de los genes fue detectada en 9p31.2-21.3. [9]

Se han realizado estudios para determinar genes responsables del SALS, pero todas las investigaciones han resultado inconclusas, los genes encontrados no tienen suficientes fundamentos para definir su participación en la aparición o progreso de la enfermedad. La degeneración y pérdida de neuronas motoras se presenta en 3 estados morfológicos en el transcurso de la enfermedad:

1) Cromatolisis: Se caracteriza por la dispersión de la sustancia Nissl sin la condensación nuclear.
2) Desgaste somatodendrítica: Es caracterizado por un citoplasma homogénico, condensación nuclear, y la preservación del nucléolo.

3) Apoptosis: La contracción neuronal con una extrema condensación del núcleo y el citoplasma es lo que caracteriza este estado.

Los mecanismos de degeneración neuronal, la forma en que son seleccionadas las neuronas afectadas, el proceso de propagación a través del sistema motor, son desconocidos. Pero existe evidencia en común a favor de la apoptosis como la vía en que la enfermedad provoca la muerte celular [10].

2.4. Tratamiento

Aunque no se conoce ningún tratamiento para retrasar ni detener la progresión de la enfermedad, siempre existen investigaciones de medicamentos experimentales a través de pruebas clínicas. Durante los últimos 20 años ha habido numerosas pruebas de drogas experimentales. Hasta la fecha, todas las pruebas clínicas de medicamentos para el tratamiento potencial del ALS han fracasado, con la excepción de riluzol, que fue aprobado por la FDA (Federal Drug Administration) en 1996. Riluzol, llamado también como Rilutek®, está disponible con receta médica en forma de comprimidos. Riluzol es un compuesto anti-glutamato, y es el primer fármaco que demostró tener un efecto terapéutico, pero su beneficio es limitado. Teóricamente riluzol puede inhibir la acción del glutamato, un neurotransmisor excitatorio que potencialmente contribuye a la degeneración de las neuronas motoras. Se reporta que riluzol prolonga la supervivencia de los pacientes con ALS solo unos meses. Sin embargo, los estudios no revelan que riluzol haya tenido algún efecto en la función respiratoria o alguna mejora en la calidad de vida del paciente. Debido a los altos costos y los pocos beneficios que proporciona, muchos médicos no prescriben Rilutek®, a menos de que los pacientes puedan cubrir el costo. A pesar del uso de la droga, la debilidad muscular no se detiene. Además, hay que considerar que el uso óptimo de la ventilación nasal, por lo general puede prolongar aún más la supervivencia del paciente que la administración de Rilutek®. [1].

El análisis en expresión de genes con microarreglos, es la utilización de técnicas de aprendizaje de máquina para examinar los datos obtenidos de un chip de microarreglos oligonucleótidos. Hoy en día estos chips contienen todo el genoma humano, la tecnología

genera enormes cantidades de datos, y un subcampo en bioinformática se ha desarrollado para minar estos datos. En vez de utilizar la investigación tradicional (la cual está limitada por la hipótesis) se usa la investigación exploratoria donde se puede minar los datos para desarrollar nuevas hipótesis [10]. Con todos estos avances se ha abierto una nueva frontera en la búsqueda de genes responsables del ALS que inclusive puede ayudar a descubrir áreas para intervención terapéutica.

Al descubrir nuevos genes responsables de la enfermedad aumenta la eficiencia al diagnosticarla por laboratorio, también, gracias a los avances de la medicina moderna se están haciendo pruebas de drogas personalizadas donde se utilizan tratamientos “sense” y “antisense”. En el caso del ALS se están realizando pruebas para crear medicina que suprima la acción tóxica de la enzima mutante del SOD1 usando el tratamiento “antisense”, pero también se requiere de una enzima sintética SOD1 que realice su tarea de protección original, aquí es donde interviene el tratamiento “sense”.

Capitulo 3

Análisis en expresión de genes usando microarreglos

No podemos hablar de microarreglos sin antes hablar sobre el DNA (Deoxyribonucleic Acid, Ácido Desoxirribonucleico) y el RNA (Ribonucleic Acid, Ácido Ribonucleico); junto con las proteínas, constituyen las 3 macromoléculas de las que están formados todo ser vivo en el planeta.

3.1. Ácido Desoxirribonucleico

La historia del DNA comúnmente parece iniciar en 1944 con Avery, MacLeod y McCarty mostrando que el DNA es el material hereditario en los seres vivos. En solo 10 años de sus experimentos, Watson y Crick descifraron su estructura y en otra década después el código genético fue accesible. Mas sin embargo, la historia del DNA realmente empezó en 1869 con el joven doctor Friedrich Miescher. Acabando sus estudios como médico, Miescher se trasladó a Tübingen para trabajar en el laboratorio de bioquímica Hoppe-Seyler, su propósito era revelar la unidad básica de la vida. Escogiendo leucocitos como su principal material, investigó primero las proteínas de estas células. Durante sus experimentos, noto una substancia con propiedades inesperadas que no concordaban con las proteínas. Miescher obtuvo la primera cruda purificación del DNA. Después de examinar estas propiedades y su composición de la enigmática sustancia, demostró que realmente difería de las proteínas. Dada que la sustancia se encontraba en el núcleo de las células, la llamo “nucleico”, término que todavía permanece hasta el día de hoy en el nombre de Ácido Desoxirribonucleico [14].

En 1888 Wilhelm Waldeyer observó ciertos filamentos en la división celular que nombró cromosomas [18]. En la actualidad, sabemos que los cromosomas son estructuras extremadamente doblegados de DNA y se encuentran en el núcleo de la célula eukaryota. Cada cromosoma contiene una sola línea molecular de DNA asociada con ciertas proteínas. En células procariotas, la mayoría de la información genética reside en un solo círculo molecular de DNA y se encuentra en la región central de la célula [19]. La clasificación de

las células es bastante compleja, pero a grandes rasgos las células procariotas son células simples que constituyen a los seres unicelulares, como las bacterias, mientras que las células eukaryotas son células complejas, como las que forman parte en animales y plantas.

Para entender los mecanismo que hacen posible la vida, tenemos que entender la estructura doble hebra de la molécula de DNA, explicada visualmente en la figura 3.1. Cada monómero en una sola hebra de DNA es un nucleótido que consiste en dos partes: un azúcar (desoxirribosa) con un grupo adherido de fosfatos, y una base, que puede ser una de las siguientes: Adenina (A), Guanina (G), Citosina (C), Timina (T). Cada azúcar está ligado a su siguiente grupo fosfático, creando así una cadena de polímeros compuesta de la repetición azúcar-fosfato como esqueleto, teniendo como protuberancia su respectiva base, cada hebra de la cadena puede tener cualquier orden, siendo que cada enlace está unido al siguiente de la misma manera. En una limitante en las células vivas: el DNA no es sintetizado como una hebra libre aislada, es un molde formado por hebras DNA ya existentes. Las bases que sobresalen de la cadena ya existente adhieren los hilos que van siendo sintetizados en la célula, siguiendo la estricta regla definida por las bases

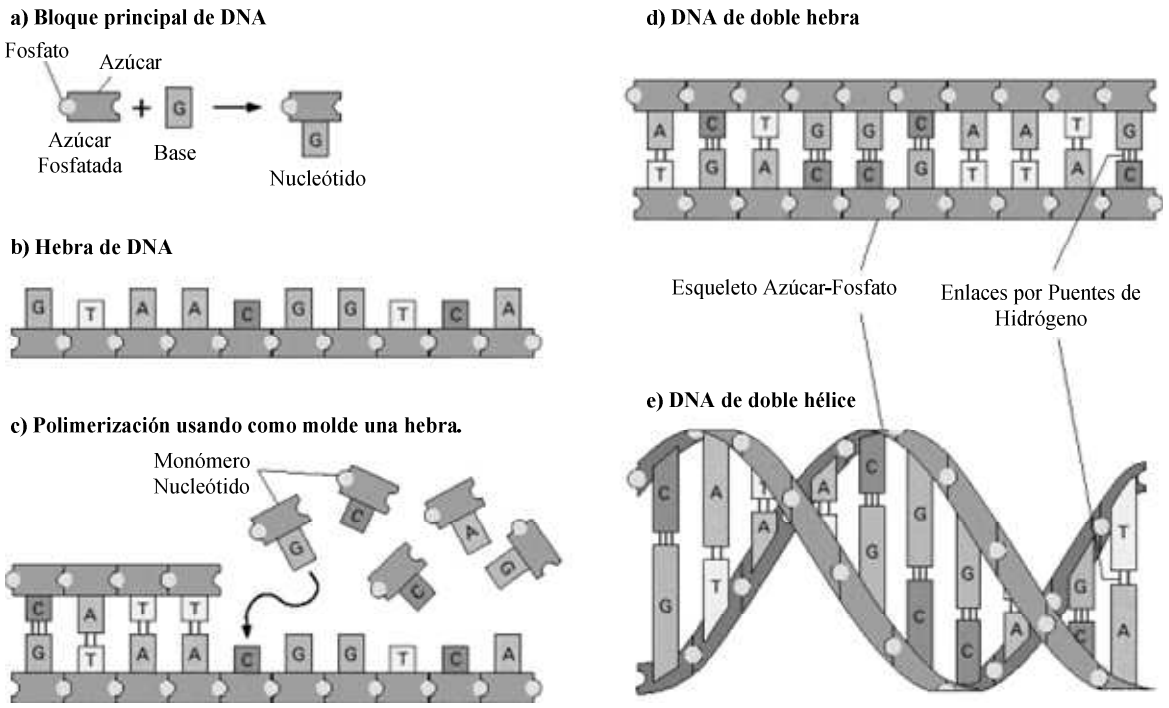

Figura 3.1. Explicación visual de la composición y estructura del DNA, así como la forma en que se replica [16]

complementarias: A se adhiere a T y C se adhiere a G. Estos pares de bases mantienen monómeros frescos en su lugar, por lo tanto, existe un control en la selección de uno de cuatro posibles monómeros que van siendo añadidos a la cadena que se está formando. Al terminar de complementar la secuencia de nucleótidos, ambos hilos se tuercen formando la doble hélice. Las uniones entre los pares de bases es mucho más débil que las uniones del grupo de azucares fosfáticos, esto permite jalar los hilos de la hélice del DNA sin dañar la secuencia de las bases. Así que cada hilo del DNA puede servir como molde para sintetizar nuevo DNA. El proceso de replicación del DNA en diferentes células, puede llegar a ser muy diferente, pero las bases previamente explicadas son universales: el DNA es un molde de polímeros donde la información puede ser almacenada y copiada para mantener la vida a este mundo. [16]

Para mantener la función de almacenamiento de información, el DNA debe hacer más que solo copiarse así mismo en cada división celular (utilizando el mecanismo previamente explicado), también debe expresar su información, ponerla en uso para guiar la síntesis de otras moléculas en la célula. El mecanismo con lo que se logra esto, también es igual en todos los seres vivos, dando lugar a otras 2 clases de macromoléculas de la vida: el RNA y las proteínas. El proceso inicia con la transcripción, donde segmentos de la secuencia del DNA son usados como moldes para guiar la síntesis de moléculas más

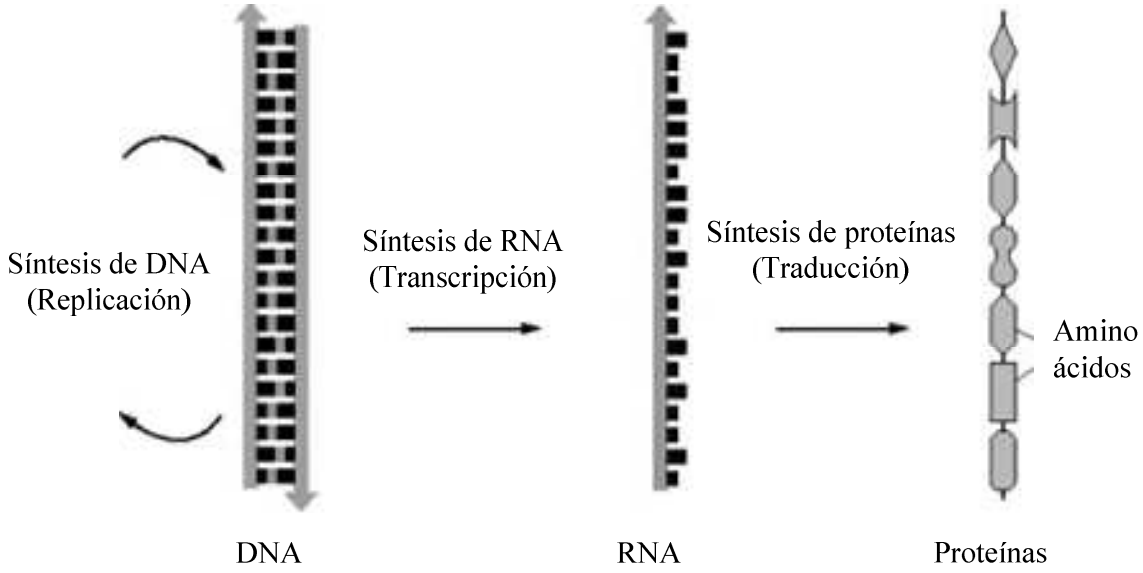

Figura 3.2. Procesos realizados por las macromoléculas de la vida [16]

pequeñas de polímeros relacionados con el DNA, el ácido ribonucleico, conocido también como RNA. Después entra un proceso más complejo llamado traducción, donde muchas de estas moléculas de RNA sirven para sintetizar polímeros con características químicas muy diferentes llamadas proteínas, la figura 3.2 muestra este proceso. [16]

3.2. Ácido Ribonucleico

Severo Ochoa gano el premio nobel en 1959 al comprobar que el RNA se sintetizaba en el núcleo de la célula y era liberado en el citoplasma, ratificando las sugerencias que hizo Caspersson y Brachet en 1939 [21].

El RNA está formado con azucares ligeramente diferentes a aquellos que forman el DNA, usa ribosa en vez de desoxirribosa, y una de sus cuatro bases, es ligeramente diferente, utiliza Uracilo (U) en vez de Timina (T), las otras 3 bases son las mismas, y los cuatro pares bases, y todas las cuatro bases puede complementarse en pares exactamente como en el DNA, solo hay que recordar que en esta ocasión U se adhiere a A. Durante la transcripción, monómeros de RNA se alinean y son seleccionados para la polimerización usando un hilo de DNA como molde, de la misma forma en que los monómeros de DNA son replicados. El resultado por lo tanto, es una molécula de polímeros que tienen una secuencia de nucleótidos representando fielmente la parte de información genética de la célula. A pesar de estar escrita con un alfabeto un poco diferente, usando monómeros de RNA en vez de monómeros de DNA. [16]

3.3. Proteínas

Los monómeros de las proteínas, los aminoácidos, son algo diferentes al DNA y el RNA, existen 20 tipos en vez de solo cuatro. Cada aminoácido es construido alrededor del mismo núcleo estructural, de esta manera puede ser ligado a otros grupos de aminoácidos, adherido a este núcleo existe un grupo que otorga a cada aminoácido un carácter químico distintivo. Cada molécula proteínica, o polipéptido, es creada al unir aminoácidos en una secuencia particular la cual tiene una precisa forma tridimensional donde existen sitios reactivos en su superficie. Estos aminoácidos son polímeros por lo que pueden adherirse a moléculas con una gran especificidad y actuar como encimas para catalizar reacciones

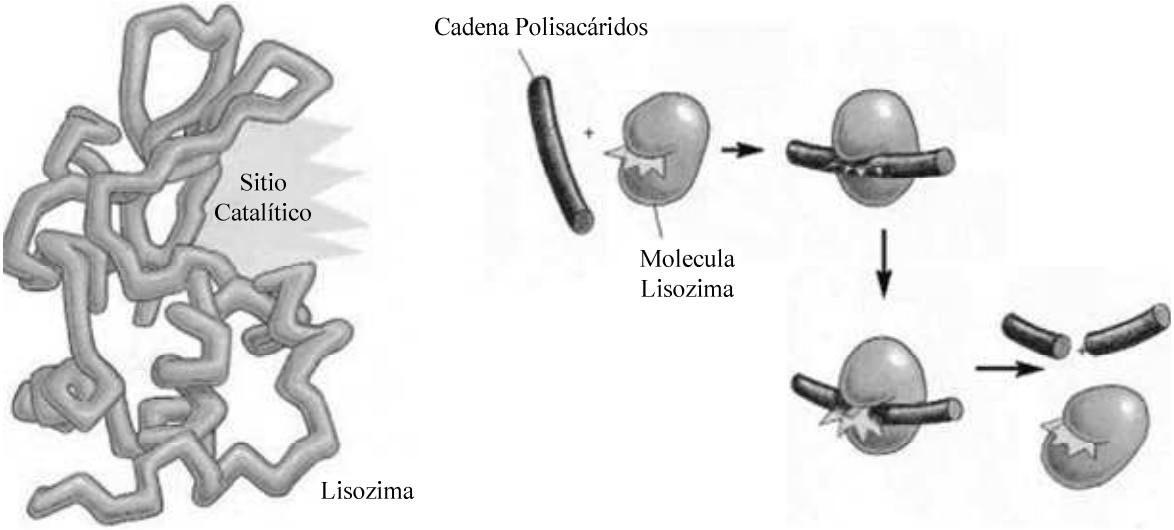

Figura 3.3. Reacción química de una molécula proteínica [16]

donde enlaces covalentes son creados y destruidos. De esta manera pueden dirigir la gran mayoría de los procesos químicos en la célula. Las proteínas pueden ser anfitrionas de otras funciones, así como mantener estructuras celulares, generar movimientos, detección de señales; y por lo tanto cada molécula proteínica puede desempeñar funciones específicas de acuerdo a su propia secuencia genética de aminoácidos. Pero sobre todas las cosas las proteínas son moléculas que pueden poner la información genética de la célula en acción. [16]

Como regla, las moléculas de DNA son muy grandes, contienen las especificaciones de miles de proteínas. Segmentos de toda la secuencia de DNA son transcritos en diferentes moléculas de RNA, con cada segmento se codifica una diferente proteína. Un gen es definido como un segmento de la secuencia de DNA que le corresponde a una sola proteína (o un solo catalítico o la estructura de una molécula de RNA que no produce una proteína) [16].

3.4. Transcripción inversa

Existe también el proceso de transcripción inversa, la cual es observada en retrovirus, como influenza, neumonía, polio, rabia, etc. Por qué los virus no pueden crecer o reproducirse por su propia cuenta no se les considera que estén vivos. Para sobrevivir, los virus deben infectar una célula huésped y tomar control de su maquinaria interna para

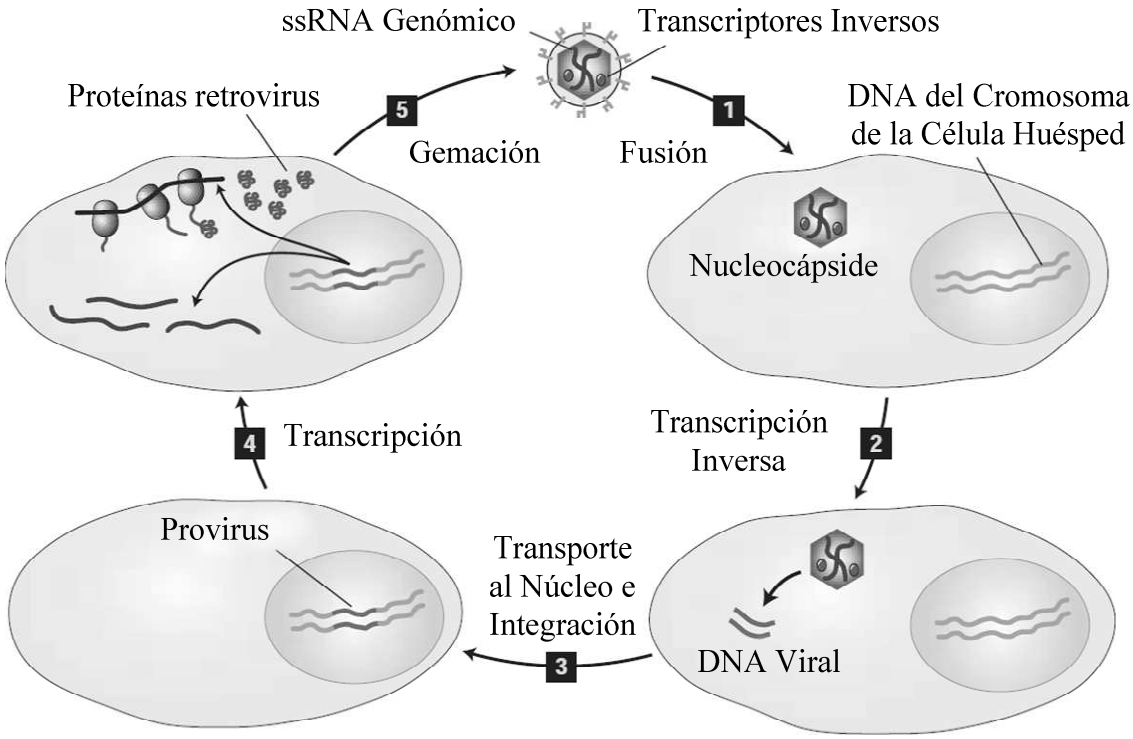

Figura 3.4. Ciclo de vida de un retrovirus [19]

sintetizar proteínas virulentas [19]. Los virus actúan como un molde para la formación de la molécula de DNA, el flujo es inverso al de la transcripción común de DNA a RNA. En el ciclo de vida de un retrovirus, una encima viral se transcribe inversamente para copiar el RNA virulento en un solo hilo de DNA complementario; la misma enzima cataliza la síntesis del hilo de DNA complementario, dando como resultado DNA de doble hilo, que es integrado en los cromosomas de la célula infectada. Al final la integración de DNA, llamado provirus, es transcrito a la célula usando la misma maquinaria para producir RNA, el cual es traducido en proteínas virulentas o es empaquetado dentro de proteínas con una envoltura vírica para luego ser liberados fuera de la membrana celular. Muchos de los retrovirus no matan a las células huéspedes, células infectadas pueden replicarse y producir más células con el virus integrado en su DNA. Se ha descubierto que algunos virus contienen material genético canceroso, nombrados oncogenes. [16]

3.5. Expresión genética

En todas las células, la expresión individual de genes es regulada, en vez de manufacturar todo el posible repertorio de proteínas todo el tiempo, las células ajustan la

frecuencia en que realizan la transcripción y la traducción de genes independientes acorde con sus necesidades. Pedazos de DNA regulatorio son intercalados junto con los segmentos que codifican las proteínas, y estas regiones del DNA al adherirse a moléculas proteínicas especiales, controlan la frecuencia local de la transcripción. Existen otras partes del DNA que no codifican nada, algunas sirven como un "punto y aparte" genético, donde definen cuando la información de una proteína inicia y donde acaba. La cantidad y la organización del DNA regulatorio y las áreas que no codifican pueden variar mucho de un tipo de organismo a otro, pero la estrategia básica es universal. De esta manera, la genética de las células, tomando en cuenta la información completa de la secuencia del DNA, dictamina no solo la naturaleza de las proteínas de la célula, sino que también cuando y donde deben ser creadas. [16]

En el proyecto para revelar el genoma humano, se ha descubierto que el DNA de todos los individuos es 99% idéntico. Mas sin embargo, esas pequeñas diferencias en la secuencia del DNA pueden generar grandes cambios en la expresión genética, frecuentemente se tienen importantes implicaciones biológicas. Los mismos genes pueden estar altamente expresados en una persona, pero tener una expresión mínima o nula en otra. Estas diferencias pueden ser por pequeñas mutaciones o la influencia del ambiente en el que vive el organismo. Por ejemplo, una persona puede tener pigmento más obscuro en la piel por que el gen de melanina esta sobre expresado por un incremento en la exposición a los rayos ultravioleta del sol [20].

Cada célula en el cuerpo humano tiene la misma secuencia de DNA. Esto se debe a que cada célula en el cuerpo de un organismo multicelular proviene de la división de una sola célula original, esta resulta de la fertilización de células reproductivas de los padres. No obstante, cada gen está expresado en diferente manera. Existen genes muy importantes que están activos en todas las células, como los que se necesitan para extraer energía de las moléculas de comida, pero hay otros genes que solo se expresan en grupos celulares específicos, como en la piel, estas células son capaces de producir proteínas que protegen a la piel de los rayos solares [20].

En 1977 James C. Alwine et. al. creó la técnica de “Northern Blot Analysis” para medir la expresión genética basándose en el proceso de hibridación genética, el problema es que solo se podía medir un gen o unos cuantos genes al mismo tiempo, por lo que el proceso era muy lento y tedioso, siendo que existen una gran cantidad de genes en seres vivos multicelulares, esta técnica solo se usó para analizar genéticamente cierto tipos de bacterias [17].

El proceso de hibridación genética es la forma más simple de formar pares con las bases del DNA permitiendo que las hebras independientes de DNA vuelvan a formar la doble hélice, la cual puede ser replicada en un tubo de ensayo. Este proceso ocurre cuando una colisión aleatoria se yuxtapone con sus nucleótidos complementarios permitiendo la

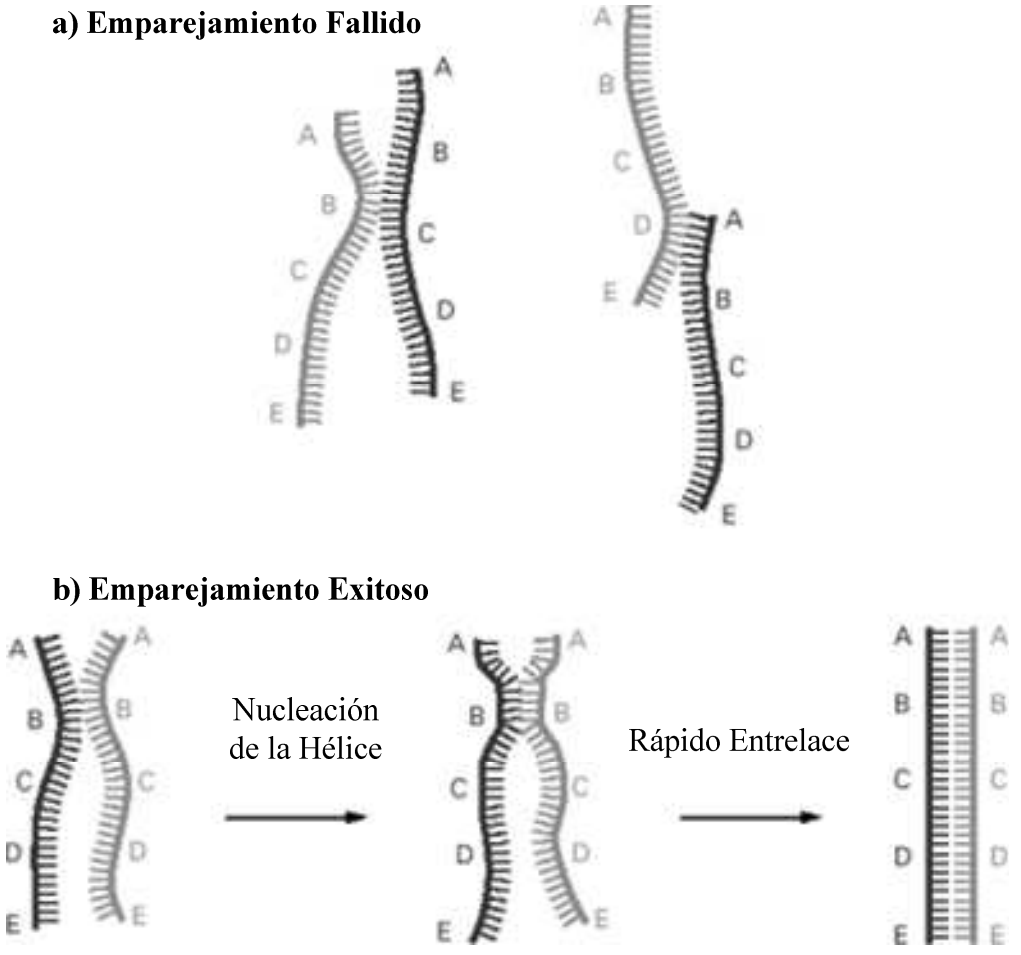

Figura 3.5. Hibridación genética [16]

formación de una pequeña parte de la doble hélice, este paso es relativamente lento, pero una vez formando esta unión es seguida de un rápido entrelace, entre más pares de bases de unen, la interacción se vuelve más rápida, como lo muestra le figura 3.5. [16]

3.6. Microarreglos

Basándose de la técnica “Northern Blot Analysis” y el uso de la misma tecnología con la que se manufacturan microprocesadores fue posible la creación de microarreglos. La tecnología de microarreglos se ha convertido en una de las herramientas indispensables para muchos biólogos que necesitan monitorear los niveles de expresión genética de algún organismo dado, esto se logra midiendo la cantidad de copias de RNA que un gen produce. Cuando un gen está activo ayuda a producir proteínas, las cuales pueden ser importantes para cierto tipo de células. El estudio de la expresión genética puede ayudar a entender funciones celulares, lo que puede ayudar a tratar enfermedades en una forma mucho más específica que la medicina tradicional ha logrado hasta el día de hoy. Los científicos usan microarreglos para encontrar un vínculo entre un gen y las enfermedades, así como desarrollar drogas para tratar a la enfermedad y crear pruebas médicas para diagnosticar enfermedades. Las investigaciones de la ALS en los últimos años han dado a conocer que el problema es genómico, obviamente esta técnica ayudará a contribuir a la comunidad científica con nuevo conocimiento.

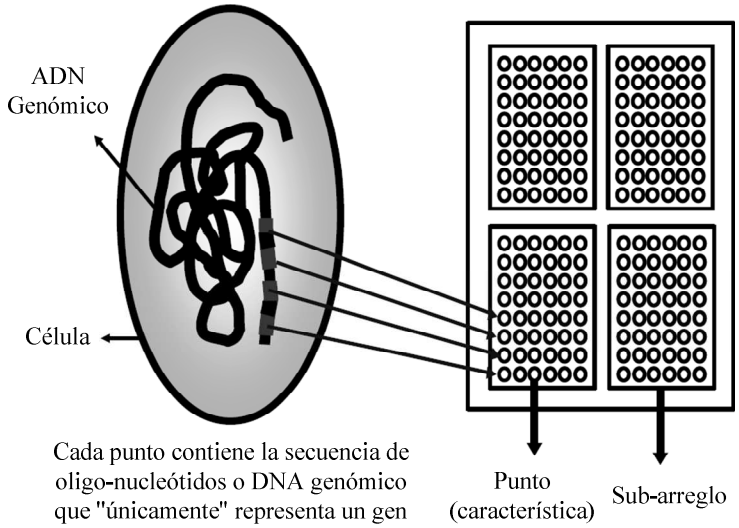

Figura 3.6. Chip de DNA típico [7]

Un microarreglo es típicamente una placa de vidrio en el cual las moléculas de DNA se fijan de una manera ordenada en lugares específicos llamados puntos (o características). Un microarreglo puede contener miles de puntos y cada punto puede contener varios millones de copias moleculares de DNA idénticas que corresponden únicamente a un gen. La figura 3.6 muestra abstractamente un chip de DNA de microarreglos.

El DNA en un punto puede ser DNA genómico o un corto tramo de cadenas de oligo-nucleótidos que corresponden a un gen. Los puntos están impresos por un robot en una placa de vidrio o son sintetizados por el proceso de fotolitografía.

Los microarreglos pueden ser utilizados para medir la expresión de genes de muchas maneras, pero una de las aplicaciones más populares es para comparar la expresión de un conjunto de genes de una célula mantenida en una condición particular (condición A) para el mismo conjunto de genes de una célula de referencia mantenida en condiciones normales (condición B). Inicialmente, se extrae el RNA de las células. A continuación, las moléculas de RNA en el extracto son transcritas inversamente en cRNA utilizando una enzima de trascripción inversa y nucleótidos marcados con diferentes tintes fluorescentes. Por ejemplo, a partir del cRNA de células cultivadas en el estado A pueden ser etiquetadas

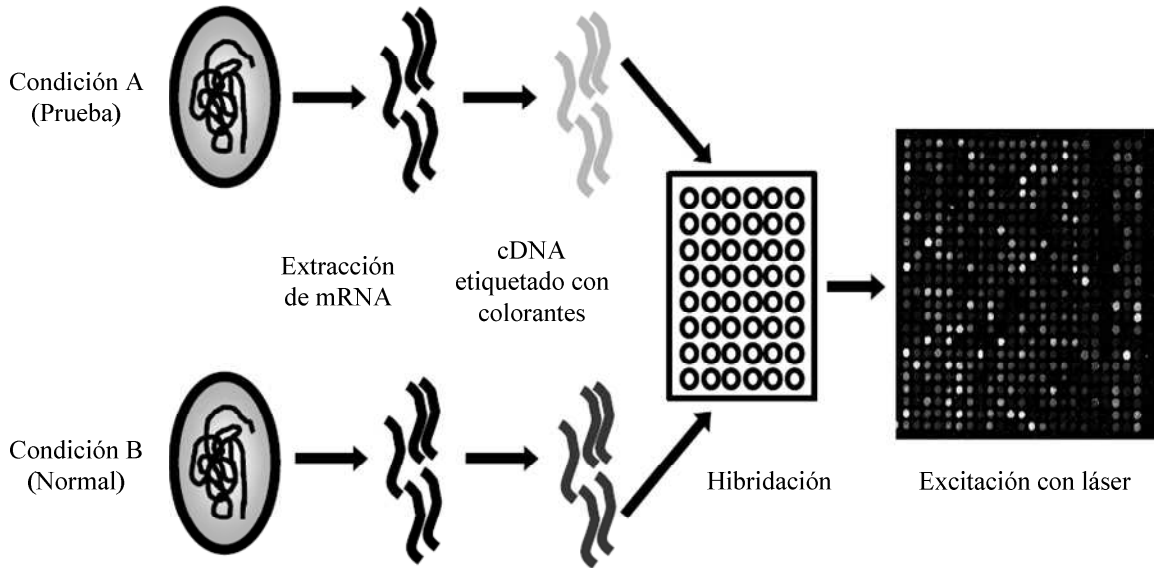

Figura 3.7. Proceso de microarreglos para la obtención de datos [7]

con un tinte de color rojo y las células cultivadas en la condición B con un tinte verde. Una vez que las muestras son marcadas con diferentes etiquetados, se usa la hibridación en la misma placa de vidrio. En este punto, cualquier secuencia de RNA en la muestra se combina con puntos específicos en la placa de vidrio que contiene su secuencia complementaria. La cantidad de RNA unido a un punto será directamente proporcional al número inicial de moléculas de RNA presentes en ese gen en ambas muestras. Tras el paso de hibridación, los puntos del microarreglo con la hibridación son excitados por un láser y escaneado en longitudes de onda adecuados para la detección de los colorantes rojo y verde. La cantidad de fluorescencia emitida a la excitación determina la cantidad de ácido nucleico. Por ejemplo, si el cDNA de la condición A para un determinado gen se encuentra en mayor abundancia que el de la condición B, el punto será de color rojo. En condiciones inversas, el punto sería de color verde. Si el gen se expresaba de la misma manera en ambas condiciones, el punto será de color amarillo, y si el gen no se expresa en ninguna condición, el punto sería negro. Por lo tanto, el resultado final de la fase experimental es una imagen de microarreglos, en el que cada punto corresponde a un gen con un valor de fluorescencia asociadas al nivel de su expresión relativa de ese gen. Este proceso se describe en el diagrama de flujo de la Figura 3.7. [7]

La intensidad de las imágenes son escaneadas por un detector de alta resolución espacial, de manera que cada punto de la prueba está representado por pixeles. Con el fin de

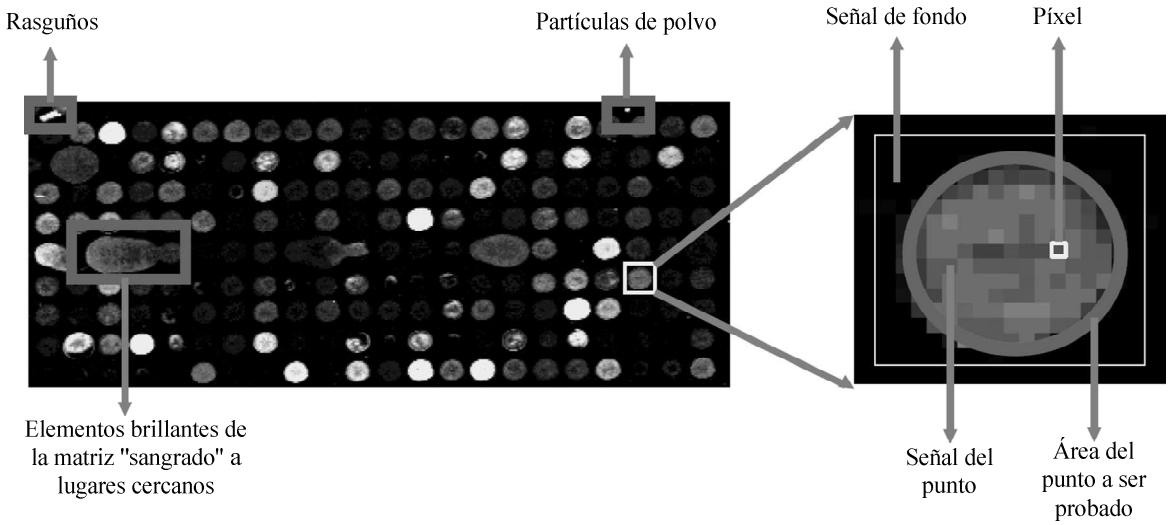

Figura 3.8. Áreas de interés en una imagen de microarreglo [7]

obtener un único valor de intensidad general para cada prueba, los pixeles correspondientes deben ser identificados (segmentación), y las intensidades deben ser concisas (cuantificación). Además de todos estos procesos, se pueden calcular también otros parámetros, como la estimación de la intensidad del "fondo local" (donde no existe ningún punto), o la calidad de la medición de los puntos. La figura 3.8 muestra un ejemplo de este proceso [8].

3.7. Proporción de expresión

Como se ha mostrado en las secciones anteriores, el nivel relativo de expresión genética puede ser medido por la cantidad de luz roja o verde emitida en el proceso de excitación. La métrica más común para relacionar esta información es el uso de la "proporción de expresión", la cual es denotada por la variable T_k y definida por la fórmula 3.1.

$$T_k = {R_k}/{G_k} \tag{3.1}$$

Por cada gene *k* en el microarreglo, R_k representa la intensidad métrica del punto o característica para la muestra de prueba y G_k representa la intensidad métrica del punto en la muestra de referencia. La intensidad métrica de cada gen puede ser representada como el valor total de la intensidad o sustrayendo el valor medio del fondo local. Si elegimos el valor medio de los pixeles, entonces la media de la "proporción de expresión" de un punto está dado por la fórmula 3.2.

$$T_{medio} = \frac{R_{medio}^{punto} - R_{medio}^{fondo}}{G_{medio}^{punto} - G_{medio}^{fondo}} \tag{3.2}$$

Donde R_{medio}^{punto} y R_{medio}^{fondo} son la intensidad media de los valores del punto y el fondo local respectivamente, para la muestra de prueba.

3.7.1. Transformaciones de la proporción de expresión.

La “proporción de expresión” es una forma relevante para representar las diferencias en una manera muy intuitiva. Por ejemplo los genes que no difieren en su nivel de expresión tendrían un valor de uno. Mas sin embargo, esta representación no es muy buena cuando se tiene que representar una regulación hacia arriba o hacia abajo. Por ejemplo, un gen que está regulado hacia arriba por un factor de 4 tendría una “proporción de expresión” de 4 (R/G = 4G/G = 4); pero en el caso de que un gen este regulado hacia abajo con un factor de 4, la “proporción de expresión” se convierte en 0.25 (R/G = R/4R = ¼). Por lo que una regulación hacia arriba puede adquirir un valor entre uno y el infinito, mientras que una regulación hacia abajo solo puede ser entre cero y uno. Para eliminar esta inconsistencia pueden usar la transformación reciproca o la transformación logarítmica.

3.7.1.1. Transformación reciproca o inversa

Este tipo de transformación convierte la “proporción de expresión” con un cambio de doblaje, formulas 3.3, donde los genes expresados con valores menores a uno se representan por su reciproco multiplicado por menos uno, en caso contrario, se deja el valor original. Con este cambio se puede representar los valores regularizados hacia arriba o abajo son representados de la misma manera.

$$Cambio\ de\ doblaje = \begin{cases} T_k, & x \geq 1 \\ -\dfrac{1}{T_k}, & x < 1 \end{cases} \tag{3.3}$$

$$Ejemplo: \begin{cases} 4, & T_k = 4 \\ -4, & T_k = 0.25 \end{cases}$$

El único problema con este método es que existe un espacio discontinuo entre -1 y 1 que puede convertirse en un problema en ciertos análisis matemáticos.

3.7.1.2. Transformación logarítmica

Un mejor procedimiento en la transformación es tomar los valores del logaritmo base 2 en la “proporción de expresión”. Esto tiene la ventaja de que trata las diferencias

Tabla 3.1. Ejemplo de transformación logarítmica

Proporción de Expresión	Transformación Logarítmica
1	0
4	2
0.25	-2

entre la regulación hacia arriba y hacia abajo igualmente, y no pierde continuidad. Ver el ejemplo de la tabla 3.1.

3.7.2. Normalización de la proporción de expresión

Existe un problema al usar la métrica de "proporción de expresión" y sus transformaciones, a pesar de que si puede ser útil para revelar patrones en los datos, remueven toda la información de los niveles absolutos de la expresión genética, por ejemplo, si un gen tiene una proporción de R/G igual a 400/100 y otro 4/1, tienen exactamente la misma "proporción de expresión" de 4, esto ocasiona problemas cuando se intenta identificar acertadamente la regulación de los genes. Una forma de confrontar el problema es normalizando los datos, se puede usar la normalización total de la intensidad o la normalización usando el centro medio logarítmico.

Existen muchos otros métodos de normalización, como regresión lineal, la proporción estadística de Chen y normalización por regresión local (lowess normalization). Terminando este paso se obtiene una matriz numérica de la expresión genética dada por la fotografía del microarreglo, permitiendo iniciar la investigación del perfil genético de las muestras tomadas.

3.7.2.1. Normalización total de la intensidad

Este método asume que la cantidad total de RNA en dos muestras es la misma, y que el número de moléculas de RNA en la hibridación son iguales; el factor de normalización N_{total} puede ser calculado usando la fórmula 3.4.

$$N_{total} = \frac{\sum_{k=1}^{N_g} R_k}{\sum_{k=1}^{N_g} G_k} \quad (3.4)$$

Donde N_g son genes que se sabe de antemano que no van a cambiar en las diferentes muestras. Las intensidades pueden ser escaladas, ver ecuación 3.5.

$$T'_k = \frac{T_k}{N_{total}} \quad (3.5)$$

Aplicando la transformación logarítmica en la ecuación 3.5.

$$log_2(T'_k) = log_2(T_k) - log_2(N_{total}) \quad (3.6)$$

3.7.2.2. Normalización usando el centro medio logarítmico

Este método también necesita un grupo de genes N_g, y asume que la media logarítmica base 2 debe ser igual a 0 para cada N_g. El factor de normalización N_{mk} se calcula usando la fórmula 3.7.

$$N_{mk} = \frac{\sum_{k=1}^{N_g} log_2\left(\frac{R_k}{G_k}\right)}{N_g} \quad (3.7)$$

En la ecuación 3.8, las intensidades ahora pueden ser escaladas.

$$T'_k = \frac{T_k}{2^{N_{mk}}} \quad (3.8)$$

Utilizando la transformación logarítmica, se obtiene la ecuación 3.10.

$$log_2(T'_k) = log_2(T_k) - log_2(2^{N_{mk}}) = log_2(T_k) - N_{mk} \quad (3.9)$$

Esto ajusta a que la "proporción de expresión" para el grupo genético N_g sea igual a cero.

3.8. Problemas en la adquisición de muestras

La obtención de muestras de microarreglos en pacientes diagnosticados con la enfermedad presenta ciertas complicaciones que impiden el progreso en la investigación de las causas de la enfermedad. A continuación se presenta una lista con las principales complicaciones en la adquisición de muestras:

- Desaparición del material de investigación, las neuronas motoras
- Lugares inaccesibles de neuronas motoras importantes
- Variación de la topográfica regional patológica dentro del sistema nervioso
- Baja proporción entre señales patológicas y ruido no patológico
- Dificultades para aislar las células afectadas
- Complicaciones bioquímicas de los tejidos por la necrólisis y el procesamiento de los tejidos

El ALS fundamentalmente se propaga en forma focal, puede iniciar en cualquier parte del sistema nervioso pero de ahí se propaga de forma contagiosa a las neuronas vecinas, hasta que la enfermedad invade el sistema respiratorio provocando la muerte del paciente. Las muestras de tejido obtenido en la autopsia pueden llegar a conservar el RNA intacto, pero solo si se sabe dónde buscar [10].

Los actuales chips de microarreglos contienen tantas características que pueden sondear todo el genoma humano, por lo que una muestra de sangre puede llegar a representar el estado de expresión genética con una gran precisión, dando una buena alternativa al análisis de neuronas en autopsias.

Capitulo 4
Aprendizaje de Maquina

El programar una computadora para optimizar el criterio de rendimiento, usando como ejemplo un conjunto de datos o experiencia pasada, es aprendizaje de máquina. Al tener un modelo definido por ciertos parámetros, el aprendizaje, es la ejecución del programa de computadora para optimizar los parámetros del modelo usando los datos de entrenamiento o la experiencia pasada. El modelo puede ser predictivo, donde los resultados intentan predecir el futuro, o descriptivo, al ganar conocimiento de los datos, o puede manejar ambos conceptos. El aprendizaje de maquina usa la teoría de estadísticas en la construcción matemática de modelos, porque la tarea más importante es hacer inferencias basándose de una muestra. El papel que tiene la ciencia computacional es, primero, el entrenamiento, es necesario algoritmos eficientes para resolver el problema de optimización, así como el guardar y procesar grandes cantidades de datos que comúnmente se generan; segundo, una vez que el modelo haya aprendido, su representación y la solución algorítmica para inferir debe ser eficiente también. En ciertas aplicaciones, la eficiencia del aprendizaje o la inferencia algorítmica, puede ser tan importante como su precisión predictiva. [22]

4.1. Principales paradigmas

Existen cuatro principales paradigmas de aprendizaje de máquina, aprendizaje no supervisado, aprendizaje supervisado, aprendizaje por esfuerzo, aprendizaje semi-supervisado.

4.1.1. Aprendizaje no supervisado.

El paradigma comúnmente es asociado a encontrar agrupamientos, involucra un proceso automático que revela estructuras en los datos y no requiere ninguna supervisión. Dados un grupo de datos con N dimensiones $X = \{x_1, x_2, \dots, x_N\}$, donde cada x_k es caracterizado por un grupo de atributos, se intenta determinar la estructura de X, identificar y describir grupos que se presentan dentro de los datos. [28]

4.1.2. Aprendizaje Supervisado.

En este paradigma los datos son proporcionados con una forma de etiquetas. Si las etiquetas son discretas el modelo resuelve un problema de clasificación, por cada punto x_k existe una etiqueta de clase ω_k, el número de valores que toma k pueden ser un pequeño número de enteros $\omega_k \in \{1,2,\dots c\}$, donde c es el número de clases posibles. El objetivo de este modelo es construir un "clasificador", función Φ, el cual pueda generar la etiqueta de clase como salida, $\Phi(x_k) = \omega_k$. Cuando las etiquetas son variables auxiliares con valores continuos, el modelo resuelve un problema de regresión, el objetivo es construir un "regresor" el cual debe minimizar el error aproximado, en otras palabras, considerando el grupo de datos formados por entradas y salidas $(x_k, y_k), k = 1, 2, \dots, N$, donde $y_k \in R$, el modelo de regresión provee el mapeo de la función $F(x)$ tal que para cada x_k obtengamos como resultado $F(x_k) \approx y_k$. La calidad del modelo depende en la naturaleza de los datos, incluyendo su dispersión, y su forma funcional. [28]

4.1.3. Aprendizaje reforzado.

Este paradigma está posicionado en medio del aprendizaje supervisado y el no supervisado. En el aprendizaje supervisado las posibles clases están disponibles en los datos desde el principio, mientras que en el aprendizaje no supervisado no existe alguna guía para asignar patrones, tienen que ser descubiertos. En el aprendizaje reforzado se ofrece información menos detallada que en el aprendizaje supervisado, esta información sirve como guía, por ejemplo dadas c clases, la señal reforzada $r(\omega)$ tendría una naturaleza binaria como lo muestra las fórmulas 4.1. [28]

$$r(\omega) = \begin{cases} 1, si\ la\ etiqueta\ de\ clase\ es\ par(\omega_2, \omega_4, \dots) \\ -1, en\ otro\ caso \end{cases} \quad (4.1)$$

Al simplificar las clases en pares e impares, el clasificador no sabe qué tipo de clase en específico está siendo procesada, pero si es capaz de reconocer 2 superclases, pares e impares. Esto es mejor opción que remover todas las etiquetas he intentar un aprendizaje no supervisado.

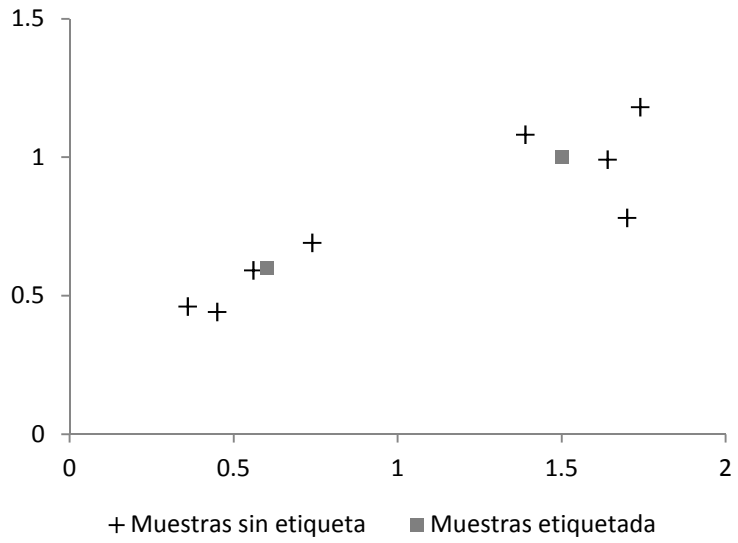

Figura 4.1. Visualización del aprendizaje semi-supervisado

4.1.4. Aprendizaje semi-supervisado

Cuando no todos los datos están etiquetados este paradigma es una mejor opción que el aprendizaje reforzado. Los patrones que muestran los datos etiquetados forman "anclas" que ayudan a navegar el proceso en determinar agrupamientos. El espacio de la búsqueda de estructuras viables en los datos es reducido, simplificado y concentrado en el desarrollo de todo el proceso, en la figura 4.1 se visualiza como dos agrupaciones de muestras no etiquetadas reciben la etiqueta de las muestras con etiqueta más cercanas. [28]

4.2. Clasificadores

Como se explicó anteriormente, el aprendizaje supervisado depende de un clasificador cuando los datos presentan etiquetas discretas. Un clasificador es un algoritmo que discrimina entre clases de patrones. Dependiendo del número de clases el problema puede ser un discrimínante entre 2 clases o entre muchas clases. El diseño del clasificador depende mucho del carácter de los datos, el número de clases, el algoritmo de aprendizaje y los procedimientos de validación. Hay que recordar que el desarrollo de un clasificador da como resultado el mapa de (Φ), donde $\Phi: X \rightarrow \{\omega_1, \omega_2, \dots, \omega_c\}$, este mapa vincula cualquier patrón de x en X a una etiqueta o clase. En la práctica, tanto el mapeo lineal y no lineal requiere de cuidadosa evaluación. La precisión de un clasificador o el error de

clasificación se refiere al porcentaje de aciertos o fallos que obtuvo el clasificador al discriminar un grupo limitado de datos.

$$Precisión = \frac{cantidad\ de\ aciertos}{número\ total\ de\ datos} \tag{4.2}$$

$$Error\ de\ clasificación = \frac{cantidad\ de\ fallos}{número\ total\ de\ datos} \tag{4.3}$$

El construir un clasificador exige el uso prudente de los datos para alcanzar un balance entre la precisión y la habilidad de generalización, este objetivo provoca un acomodo de los datos en grupos de entrenamiento y prueba. Un clasificador que está siendo entrenado y evaluado con el mismo grupo de datos produce resultados sumamente optimistas que provocan que en la práctica el clasificador tenga un comportamiento catastrófico. Para alcanzar cierto nivel de confianza en el clasificador, el desarrollador debe utilizar los datos disponibles prudentemente, comúnmente los datos se dividen en los siguientes grupos:

- Datos de entrenamiento
- Datos de validación
- Datos de prueba

Cada uno de estos subgrupos juega un diferente papel en el diseño. El grupo de entrenamiento es esencial para completar el entrenamiento del clasificador, en particular, todas las actividades de optimización son guiadas por el desempeño reportado por este grupo. El grupo de validación ayuda al clasificador al mantener una óptima estructura, aun cuando el desempeño con el entrenamiento aumente, los datos de validación pueden reportar que la precisión este disminuyendo, indicando que la estructura del clasificador puede ser excesiva. También se pueden usar los datos de validación para monitorear el proceso de aprendizaje, aún más cuando el algoritmo depende de varias iteraciones para arrojar un resultado. Finalmente el grupo de prueba es usado para evaluar el diseño del

clasificador, es la forma de observar el comportamiento del clasificador con datos nunca antes vistos, su poder de generalización. [28]

En general, existen diferencias en los reportes de desempeño del mismo clasificador usando los grupos de entrenamiento y de prueba. Comúnmente cuando el clasificador se desempeña excelentemente con los datos de entrenamiento, los resultados con el grupo de prueba son pobres, este efecto se le conoce como “memorización”. En estos casos, se le denomina también al clasificador que cuenta con gran habilidad de aproximación pero tiene poco poder de generalización. No puede moverse fácilmente más allá de los datos con los cuales fue entrenado y fracasa al procesar datos nunca antes vistos (grupo de prueba). Este es el dilema de aproximación-generalización, el cual no debe considerarse trivial y requiere gran atención en la fase de desarrollo.

Existen un sin número de clasificadores, pero esta fuera del alcance de este documento el enumerarlos todos, más sin embargo en la experimentación se utilizaron principalmente el discriminante lineal, discriminante cuadrático, el vecino más cercano y el clasificador bayesiano ingenuo.

4.2.1. Discriminante lineal y cuadrático

El análisis de un discriminante lineal (LDC, Linear Discriminant Classifier) consiste en buscar una combinación lineal en las variables o atributos en los datos seleccionados, la cual provee una mejor separación dentro de las clases. Estas diferentes combinaciones se le llaman funciones discriminantes, ver figura 4.2. [23]

El discriminante cuadrático (QDC, Quadratic Discriminant Classifier) busca el mismo objetivo que el discriminante lineal, solo que da como resultado una función de orden cuadrático, la gran mayoría del tiempo tiene más poder de generalización y aproximación que su contraparte lineal, pero por su complejidad, es mucho más difícil de encontrar la función discriminante, en la figura 4.3 se muestra que usando los mismos datos el QDC tuvo mejor poder de clasificación.

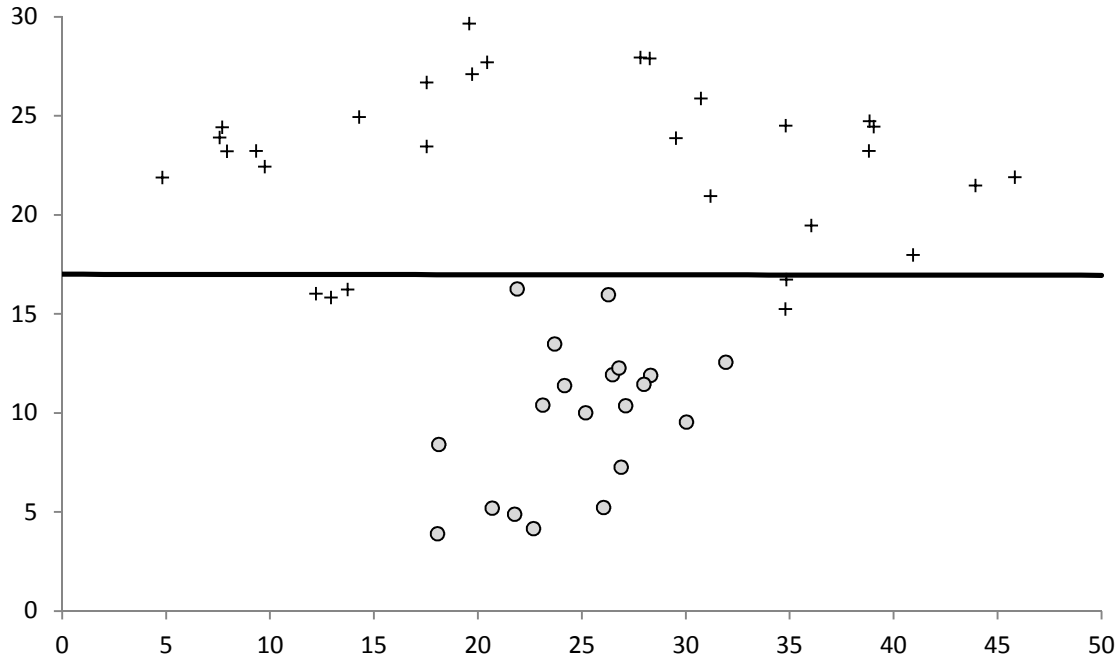

Figura 4.2. Ejemplo de un discriminante lineal

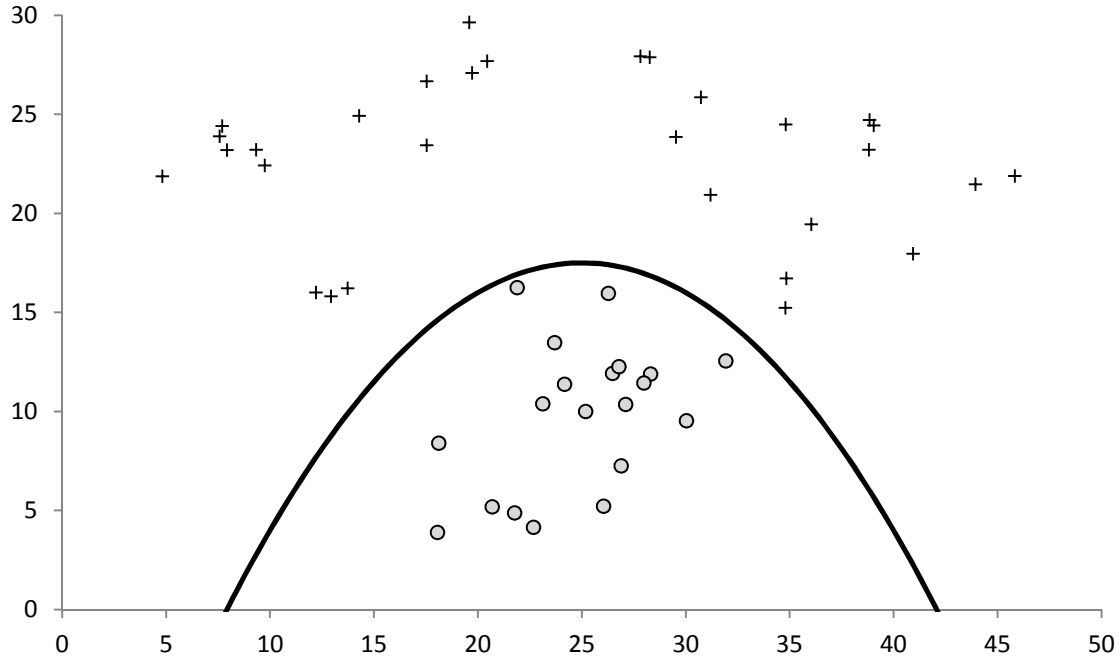

Figura 4.3. Ejemplo de un discriminante cuadrático

4.2.2. El vecino más cercano

Este discrimínate funciona buscando en un grupo de entrenamiento k número de puntos más cercanos al punto muestra y le asigna la clase más común de ese agrupamiento. En general, un clasificador kNN (k Nearest Neighbor) tiene las siguientes propiedades:

- Buena habilidad de rechazo de ruido
- Es fácil de trazar analíticamente.
- Cuando $k = 1$ y se cuenta con un grupo de entrenamiento grande, el error no excede el doble del clasificador bayesiano.
- Es fácil de implementar
- Puede ser integrado en un sistema de base de datos y se puede utilizar los índices como forma de acceso fácil a los datos.

El clasificador kNN es muy popular por las características antes mencionadas y se puede encontrar en muchas aplicaciones, mas sin embargo, el algoritmo necesita calcular todas las distancias entre el grupo de entrenamiento y la muestra, además, ocupa un mecanismo para elegir los k puntos más cercanos y procesar la etiqueta que se le asignará a la muestra. Este proceso impacta negativamente la escalabilidad del algoritmo, es por esto que las nuevas investigaciones proponen usar métodos y técnicas de rápido acceso para computar similitudes, las cuales ya existen en sistemas de bases de datos, para reducir el costo de la búsqueda de una complexidad lineal a logarítmica. Aun así, el costo de búsqueda del clasificador kNN aumenta considerablemente con valores mayores de k. [24]

Dependiendo de la los datos de entrenamiento el valor de k puede variar, en otras palabras, existen casos, que una k pequeña puede ser suficiente para una clasificación exitosa, mientras que en otros casos se necesita una k más grande para determinar adecuadamente la clase de la muestra. Por lo tanto el valor de k puede llegar a ser significativo, esto introduce un nuevo proceso para obtener el valor adecuado de k. [24]

Existen varias formas de calcular la distancia de un punto multidimensional a otro, pero esta fuera del alcance de este documento el detallar cada método, el clasificador kNN

de esta investigación solo utilizo la distancia euclidiana, la cual se calcula usando la fórmula 4.4.

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad (4.4)$$

Donde la distancia $d(p,q)$ es equivalente a la distancia $d(q,p)$ y n es el número de dimensiones que constituyen cada punto.

4.2.3. Clasificador bayesiano ingenuo

Este es un clasificador probabilístico, el cual regresa la probabilidad que tiene cierta etiqueta y dado que se presentan los datos x, la probabilidad a posteriori $p(y|x)$. El clasificador es una variante del clasificador bayesiano, en donde existen dos maneras para obtener $p(y|x)$, la primera es aprender la función que calcula la probabilidad a posteriori de la clase $p(y|x)$ directamente, ha esto se le llama modelo discriminante, la alternativa es aprender la densidad condicional $p(x|y)$ para cada posible valor de y, aprender la probabilidad a priori $p(y)$, y entonces aplicar la regla de bayes para computar la probabilidad a posteriori $p(y|x)$ [25]

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_{i=1}^{C} p(x|y_i)p(y_i)} \qquad (4.5)$$

La sumatoria del dominador en la fórmula 4.5 se le conoce como probabilidad marginal $p(x)$ o evidencia, y es obtenida gracias a la ley de la probabilidad total. C es el número total de clases o etiquetas, en resumen la formula ve que tan frecuente es encontrar datos x con la etiqueta y.

La variante en el clasificador bayesiano ingenuo es remover $p(x)$ considerando que existe independencia en los datos x, a pesar de que sí exista dependencia.

$$p(y|x) = p(x|y)p(y) \qquad (4.6)$$

Esto facilita mucho el cálculo de la probabilidad $p(y|x)$ cuando x es un vector muy grande de atributos, si tenemos la densidad condicional $p(x|y)$ la ecuación queda como una multiplicadora de la probabilidad de cada uno de los atributos del vector x.

$$p(y|x) = p(y)\prod\prod_{i=1}^{n} p(x_i|y) \tag{4.7}$$

Donde x_i son los atributos del vector x, mientras que n es el número total de atributos en el vector.

4.3. Selección de características

La selección de características, conocida también como selección de subgrupos, es el proceso comúnmente usado en métodos de aprendizaje de máquina, donde un subgrupo de características son seleccionadas de los datos para ser usados en un algoritmo de aprendizaje. El mejor subgrupo contiene el menor número de dimensiones posibles que son capaces de contribuir mucho mejor a la precisión del algoritmo, las dimensiones restantes son descartadas siendo que no son relevantes. Este es un paso importante en el pre-procesamiento, y es una forma de evadir la maldición de la alta dimensión, la otra forma es usando la extracción de características, este último método esta fuera del ámbito del presente documento. [12]

Generalmente la maldición de la alta dimensión es la expresión de todos los fenómenos que aparecen en datos con altas dimensiones, y muy seguido es la consecuencia más fortuita en el comportamiento y desempeño de los algoritmos de aprendizaje. Las herramientas del análisis de datos están basadas en el principio de aprender al inferir conocimiento o información de una muestra de datos disponibles para el aprendizaje. Obviamente, los modelos de aprendizaje construidos son solo validos en el rango o volumen del espacio donde los datos de aprendizaje están disponibles. Cualquiera que sea el modelo o clase de modelo, la generalización de datos que son muy diferentes de todos los puntos de aprendizaje es imposible, en otras palabras, la generalización relevante es posible usando la interpolación pero no de la extrapolación. Uno de los ingredientes

principales en el desarrollo exitoso de algoritmos de aprendizaje es tener suficientes datos para aprender, de esta manera llenan el espacio o partes del espacio donde el modelo debe ser válido. Es mucho más fácil de comprender este concepto con el siguiente ejemplo, si el modelo mantiene todas las demás restricciones sin cambiar, el número de datos para el aprendizaje crece exponencialmente con la dimensión, si 10 datos son razonables para un modelo con una sola dimensión, 100 son aparentemente necesarios para un modelo de 2 dimensiones, 1000 para 3 dimensiones. Este incremento exponencial es la principal consecuencia a lo que le llamamos la maldición de la alta dimensión [21]. La matriz numérica que da como resultado el proceso de microarreglos tiene tantos atributos que siempre hay que tener mucho cuidado con la forma en que se está lidiando con el problema de la alta dimensión.

Era muy raro que las publicaciones de 1997 que contenían material relevante en selección de variables y características usaran dominios de más de 40 características. La situación ha cambiado considerablemente en los últimos años, muchas de las publicaciones exploran dominios con cientos, inclusive hasta miles de características o variables, nuevas técnicas son propuestas para manejar estas tareas desafiantes, las cuales involucran tareas con características irrelevantes y comúnmente con pocos datos de entrenamiento. [13]

Existen muchos métodos para seleccionar características, pero esta fuera del alcance de este documento el enumerar y explicar cada uno de esos métodos, la experimentación realizada se enfocó principalmente en la selección hacia adelante y la selección hacia atrás, en la figura 4.4 se muestra donde se encuentran este tipo de selecciones respecto a otros tipos de métodos.

Los métodos llamados “completo” son cualquier método que prueba todos los posibles subgrupos de los datos multidimensionales y escoge el mejor para resolver el problema de clasificación, en otras palabras, este método tiene la mayor precisión posible, idealmente es el mejor, pero cuando existen demasiadas características, el tiempo de ejecución provoca que este método sea imposible de ejecutarse.

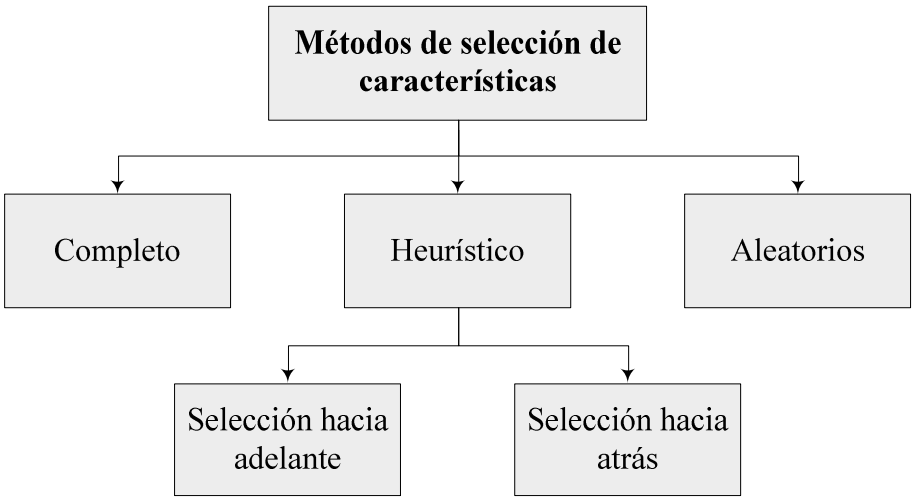

Figura 4.4. Métodos usados en este documento para la selección de características

Los métodos de selección aleatoria de subgrupos, intenta resolver el problema del tiempo que tienen los métodos completos a expensas de que el subgrupo seleccionado puede no ser el óptimo, pero es una solución aceptable en la mayoría de los casos. El principal objetivo de estos métodos es escoger el mejor grupo en una lista limitada de subgrupos aleatorios, dicha lista es tan larga como es posible, manteniendo en mente el costo del proceso computacional en la ejecución de cada subgrupo. Entre más larga la lista es, la probabilidad de que el grupo resultante sea el óptimo aumenta, pero va a tomar más tiempo en ejecutarse.

Los métodos de selección heurística de subgrupos, crean subgrupos mediante suposiciones educadas, dependiendo como las características se desempeñaron en el pasado, tienen más probabilidad de ser seleccionadas de nuevo en el próximo grupo de prueba. El desempeño de los grupos se mide comúnmente mediante la precisión de un clasificador o el error de clasificación. Existen dos principales enfoques para estos métodos:

- Selección hacia adelante: El subgrupo inicia sin variables y se agregan una por una en cada, paso añadiendo la característica que aumenta la precisión

de la clasificación, esto se lleva a cabo hasta que la precisión no aumenta de manera significativa.

- Selección hacia atrás: El subgrupo inicia con todas las posibles variables y remueve una por una, en cada paso se remueve la variable que ayuda a aumentar la precisión de la clasificación hasta que al remover cualquier variable compromete la precisión en vez de mejorarla.

4.3.1. Jerarquía de Variables.

Existen muchos algoritmos de selección de variables que incluyen la jerarquía de variables como mecanismo principal o auxiliar en la selección, por su simplicidad, escalabilidad, y el gran éxito empírico. La jerarquía de variables no necesariamente es usada como base en modelos de predicción, es comúnmente usado como un método de filtrado, es un paso de procesamiento, independiente de las decisiones del predictor en el modelo. Mas sin embargo, dentro de cierta independencia o suposiciones ortogonales, puede ser optimo con respecto al predictor. Por ejemplo, el uso del criterio de Fisher para la jerarquía de variables en un problema de clasificación donde la matriz de covarianza es diagonal es óptimo para un clasificador haciendo uso del discriminante linear de Ficher. Inclusive cuando la jerarquía de variables no es óptima, es preferible a otros métodos de selección de grupos de variables, por su escalabilidad computacional y estadística. Computacionalmente, por su eficiencia, siendo que solo se requiere procesar la calificación de cada variable y ordenarlas según los resultados obtenidos. Estadísticamente, es robusto al prevenir el sobreajuste, porque introduce preferencias pero puede tener muy poca varianza. [13]

Si consideramos un grupo m de ejemplos $\{x_k, y_k\}(k = 1 \dots m)$ considerando n como las variables de entrada $x_{k,i}(i = 1 \dots n)$ y una variable de salida y_k. La jerarquía de variables puede usar una función calificativa $S(i)$ donde procesa los valores de $x_{k,i}$ y$y_k(k = 1 \dots m)$. Por convención, se asume que una calificación alta indica una variable valiosa por lo que se deben ordenar de manera descendente usando $S(i)$. Para usar la jerarquía de variables y de ahí construir predictores, se deben de agrupar subgrupos incorporando progresivamente más y más variables definidas con menor relevancia. [13]

Hay que tener cuidado al seleccionar características, puede que características que sean catalogadas como redundantes o con poco poder de clasificación, sean capaces de ayudar al clasificador al ser agrupadas con otras características consideradas igualmente redundantes. Existen varios métodos de selección de características que son muy sensibles a pequeñas perturbaciones en las condiciones del experimento, por ejemplo, si los datos tienen variables redundantes, si existen varios subgrupos con poder predictivo idéntico, dependiendo de las condiciones iniciales del algoritmo, si se agregan o remueven ciertas muestras, o la adición de ruido. Hay ciertas aplicaciones donde tener diferentes subgrupos de características para la clasificación es algo deseable, pero comúnmente es algo que se debe evitar, por lo general la varianza es un síntoma de un mal modelo, el cual es incapaz de generalizar adecuadamente, otra señal de un modelo inestable es que los resultados no sean reproducibles y finalmente que ningún subgrupo sea capaz de discriminar adecuadamente la mayoría de los casos, pobre poder de precisión en la clasificación.

Existen varios métodos para estabilizar la selección de características, pero en el uso de microarreglos el más recomendado es el método de bootstrap. Puesto que es capaz de ayudar con la maldición de la alta dimensión que presentan los datos obtenidos de microarreglos.

Como se mencionó anteriormente, la matriz numérica obtenida de un microarreglo tiene miles de características, entre más complejo el chip de microarreglos, mas genes pueden ser evaluados al mismo tiempo, actualmente el rango de características en un chip de microarreglos está entre 20,000 a 60,000, también hay que denotar que el proceso necesario para conseguir muestras de microarreglos de buena calidad es demasiado costoso como para tener un lote lo suficientemente grande para usar todas las características en el modelo y obtener resultados aceptables.

4.4. Bootstrap

Es un método de remuestreo para inferencia estadística. Es comúnmente usado para estimar los intervalos de confianza, pero también puede ser usado para estimar el sesgo y variancia de un estimador o para calibrar las pruebas de una hipótesis. [26]

Dado un grupo de observaciones independientes e idénticamente distribuidas, $X_i(i = 1 \dots n)$; un parámetro que puede ser definido como una función, θ=T(x), de variables en la población, y una función estática con las mismas observaciones, $\hat{\theta}$=T(x), el método bootstrap estima la distribución del remuestreo, $F_\theta(x)$, de esa función. Los datos son usados como un estimador de la función de distribución acumulada, $F_x(x)$, de los valores de la población. Las muestras bootstrap son obtenidas repetidamente de la población estimada. La funciones evaluada para cada muestra bootstrap, dado un grupo de valores bootstrap, $\{\theta_i^B\}, i = 1 \dots m$. La distribución empírica de esos valores bootstrap, $\hat{F}_b(x)$, estima la distribución de las muestras teóricas, $F_\theta(x)$. [26]

La distribución bootstrap, $\hat{F}_b(x)$, es usada para estimar el sesgo, estimar el error estándar o construir intervalos de confianza para intereses estadísticos. Las estimaciones bootstrap del sesgo, B_b, y el error estándar, S_b, son estimaciones empíricas calculadas de valores bootstrap m, ver formula 4.8 y 4.9.

$$B_b = \frac{\sum_{i=1}^{m}\left(\hat{\theta}_i^B - \hat{\theta}\right)}{m} \tag{4.8}$$

$$S_b = \sqrt{\sum_{i=1}^{m}\left(\left(\hat{\theta}_i^B - \overline{\hat{\theta}^B}\right)^2/(m-1)\right)} \tag{4.9}$$

El percentile del intervalo de confianza que este método usa es de $\propto/2$ y $1-\propto/2$ cuantiles de $\hat{F}_b(x)$ con un nivel de intervalo de confianza de $1-\propto$ para el parámetro.

Cada muestra bootstrap es una simple muestra aleatoria del lote original de muestras seleccionada con remplazo, por esta misma razón, algunas de las muestras originales van a estar repetidas más de una vez en el lote de muestras bootstrap, el cual tiene un tamaño *n*. [26]

Las muestras bootstrap originalmente se usaban para entrenar el clasificador y el lote de muestras original como partición de prueba, después la precisión otorgada por cada muestra bootstrap es promediada para obtener la estimación bootstrap, conocido también como estimación de resubstitución $\hat{e}_n^{RSB}$. Este estimador es conocido por subestimar el error de predicción al usar los mismos datos tanto para construir como evaluar las reglas de predicción, el problema de sobreajuste se origina cuando el número de dimensiones es mayor que el número de muestras, un síntoma claro de la maldición de la alta dimensión, es posible seleccionar un numero de características para construir un modelo que se ajuste a los datos perfectamente pero no es muy útil al momento de predecir futuras observaciones. Existen muchas variantes, en este documento fue usada la de Leave-One-Out Bootstrap (LOOB) y 0.632+ Bootstrap. [27]

El procedimiento que sigue el LOOB genera un total de B muestras bootstrap de tamaño n. Cada espécimen que no fue seleccionado para formar la muestra bootstrap es usado como parte de la partición de prueba. De esta manera este método evita el usar como prueba observaciones usadas para la construcción del modelo, como resultado otorga una estimación con mucho menos variabilidad que variantes más primitivas del bootstrap, pero hay que considerar que cada muestra bootstrap contiene 0.632 observaciones distintas que conformaban el lote original de muestras, esto comúnmente es inadecuado para representar la distribución de los datos originales cuando el tamaño de las muestras bootstrap n es muy pequeña, provocando que la estimación bootstrap este sobreestimando el verdadero error de predicción. [27]

La variante 0.632+ Bootstrap fue propuesta por Efron y Tibshirani en orden de reducir el sesgo de la variante de LOOB. La fórmula 4.10 muestra cómo obtener del estimador $\hat{e}_n^{0.632+}$.

$$\hat{e}_n^{0.632+} = w\hat{e}_n^{LOOB} + (1-w)\hat{e}_n^{RSB} \quad (4.10)$$

Donde w tiene valores entre 0 y 1, $\hat{e}_n^{RSB}$ es la estimación de la resubstitución y $\hat{e}_n^{LOOB}$ es la estimación proporcionada por el LOOB. El valor que toma w es de 0.632, es la razón del nombre de esta variante. Cuando el error de la resubstitución es cero, la estimación bootstrap se convierte en $0.632\hat{e}_n^{LOOB}$, esto da como resultado una baja sistemática en el sesgo cuando no existen diferencias entre las clases. El método tiene como propósito principal el evitar el problema mencionado incrementando el peso de w. [27]

4.5. Análisis de Componentes Principales

Esta técnica es utilizada para identificar patrones en un grupo de datos, y expresar los datos en cierta forma que se destacan las similitudes y diferencias. En datos de altas dimensiones es imposible graficarlos en su forma natural, pero con el análisis de componentes principales (Principal Components Analysis, PCA) es posible obtener una representación gráfica al reducir el número de dimensiones manteniendo la mayoría de la información multidimensional, esto se lleva acabo al hacer uso de un procedimiento estadístico donde se le aplica la descomposición eigen a la matriz de covarianza obtenida de los datos multidimensionales. Las gráficas resultantes se le conocen como proyecciones o sombras de los datos multidimensionales [30].

El PCA se utilizó en este documento solo para la representación de resultados, pero no solo se puede usar para crear proyecciones, esta técnica puede ser aplicada en métodos de reconocimiento de patrones. En esta investigación no fue posible aplicar este método en la selección de genes por la limitantes del poder computacional, la matriz de covarianza que arroja de datos obtenidos por microarreglos es tan grande que fácilmente puede ocupar un gigabyte de memoria RAM cada muestra, esta cifra rebasa el limite tanto de hardware como de software que fue disponible para los experimentos.

Capitulo 5
Experimentación

5.1. Primer experimento

5.1.1. Datos

Este experimento se efectuara con los datos obtenidos de 242 personas en el experimento “E-GEOD-3307” de la base de datos proporcionados por EMBL-EBI, los cuales son microarreglos de muestras de pacientes con enfermedades musculares, las cuales son: miopatía aguda tetrapléjica, dermatomiositis juvenil, esclerosis lateral amiotrófica, paraplejia espástica, distrofia muscular de facioscapulohumeral, distrofia muscular de Emery-Dreifuss, distrofia muscular de Becker, distrofia muscular de Duchenne, calpaína 3, disferlina, FKRP. De los cuales se utilizarán 9 muestras pertenecientes al ALS

5.1.2. Procedimiento

Primero se separaron del total de muestras, los datos relacionados con los microarreglos de pacientes con ALS y los pacientes normales. Dando como resultado una matriz numérica 27 x 22284 elementos, donde la última columna es la etiqueta de clase, 0 para pacientes normales y 1 para pacientes con ALS.

Como siguiente paso, se creó un lote de 2000 muestras bootstrap 0.632+ y se utilizó la selección hacia adelante con jerarquía de variables. Los clasificadores que se utilizaron fueron el discriminante lineal de Fisher, discriminante cuadrático (QDC), bayes ingenuo y el vecino más cercano con parámetros de $k = 1$ y $k = 3$. El algoritmo utilizado para el QDC fue el que proporciona la librería de matlab PRTOOLS (http://prtools.org).

La jerarquía de variables se obtuvo mediante el criterio de Fisher, la calificación que se le otorgó a cada característica es el promedio del criterio de Fisher de todas las muestras bootstrap.

5.1.3. Resultados

Las gráficas de desempeño muestran 2 líneas; el desempeño del clasificador está regido por el eje de la izquierda, el cual muestra la precisión que adquiere el clasificador con la acumulación de genes seleccionados hasta ese punto. La línea de calificación de Fisher representa la relevancia que el gen obtuvo y está regido por el eje de la derecha el cual esta normalizado para que use valores de 0 a 100, de esta manera la gráfica puede ser más legible.

El algoritmo de selección de características con cada clasificador arrojo una selección de más de 15 genes, pero con los primeros genes más significativos es posible mostrar el patrón que sigue el clasificador en el progreso de la selección.

Como se puede observar en la Figura 5.1 y Figura 5.2, el desempeño del clasificador cuadrático sobrepasa el desempeño del clasificador lineal de Fisher, mas sin embargo, la precisión máxima que otorgó cada clasificador no es satisfactoria para continuar usándolos, solo se incluyeron para tener un punto de referencia y denotar que el problema que estos clasificadores intenta resolver no sigue un orden lineal ni cuadrático, por esta misma razón la tabla de los genes seleccionados usando estos dos clasificadores no se mostrarán en este documento.

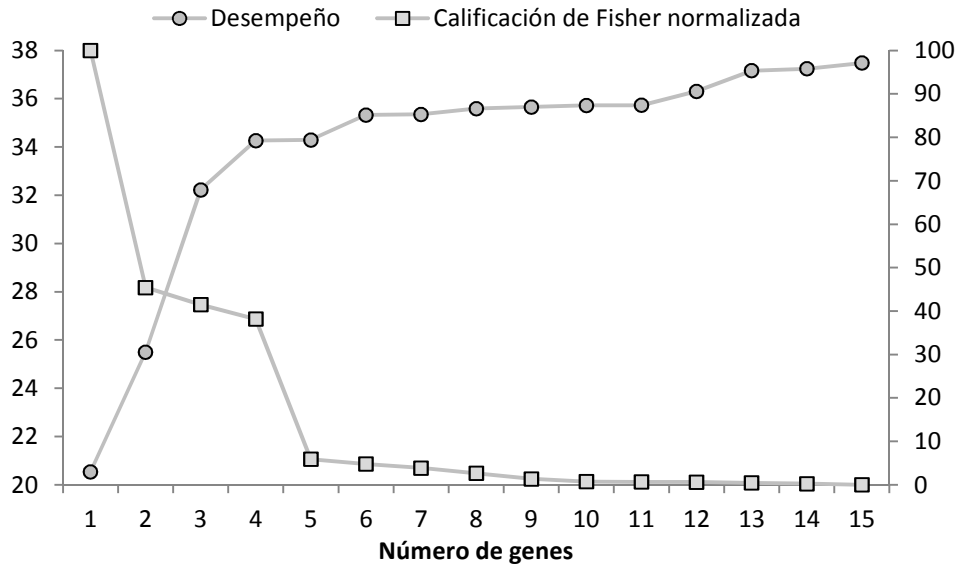

Figura 5.1. Desempeño del clasificador lineal de fisher

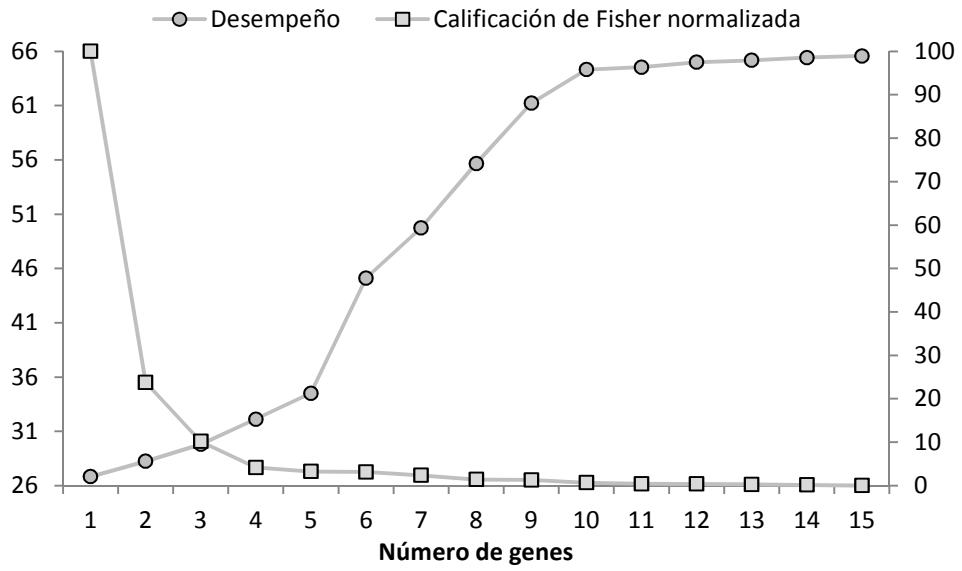

Figura 5.2. Desempeño del clasificador cuadratico

Los clasificadores que otorgaron mejor resultados fueron el de vecinos cercanos (y el bayesiano ingenuo, los cuales se tomaron los primeros 10 genes más significativos y se proyectaron los datos para ver la capacidad discriminante de esos genes utilizando PCA. Las figura 5.3 a la figura 5.8 muestran dichos resultados.

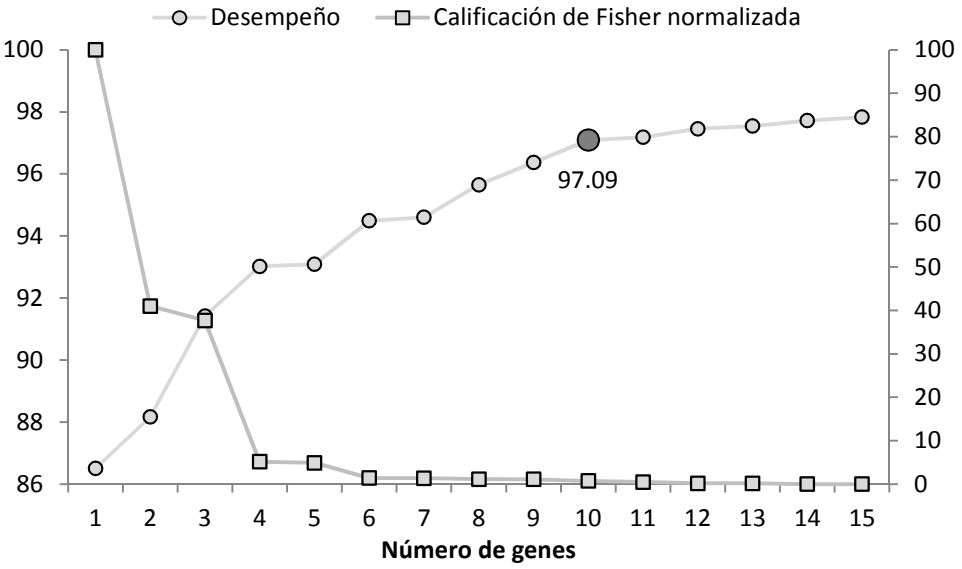

Figura 5.3. Desempeño del clasificador del vecino mas cercano con parametro $\boldsymbol{k = 1}$

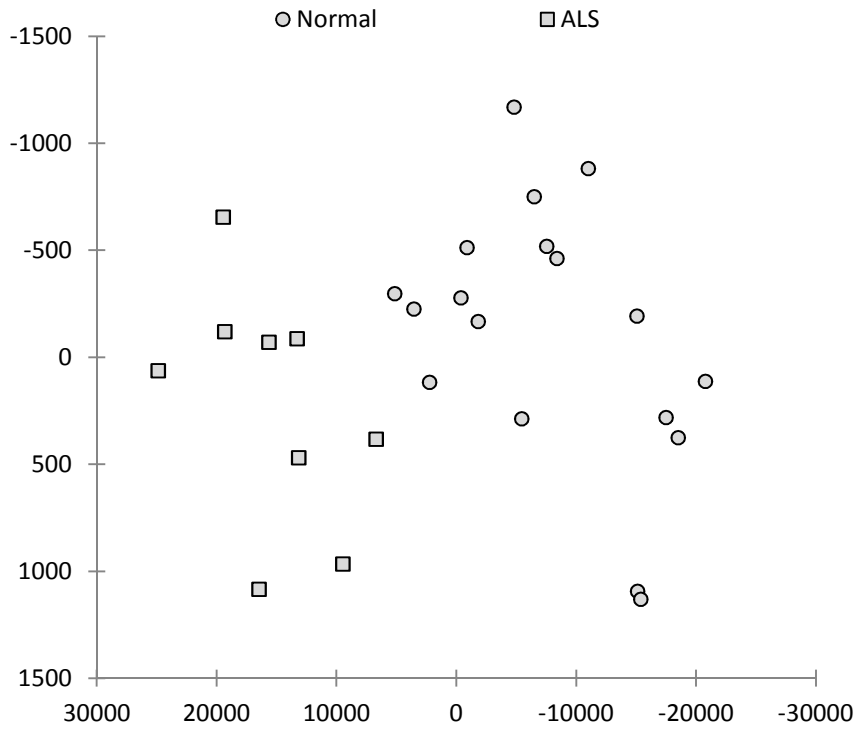

Figura 5.4. Proyeccion de los puntos muestrales usando PCA y los 10 genes mas significados otorgados por el clasificador de vecinos cercanos con parametro de $\boldsymbol{k = 1}$

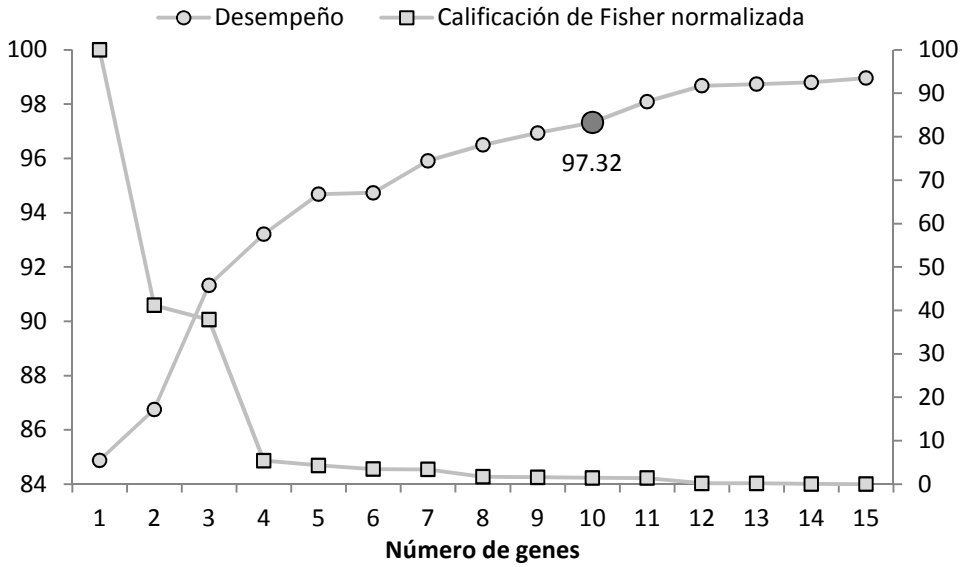

Figura 5.5. Desempeño del clasificador del vecino mas cercano con parametro $\boldsymbol{k = 3}$

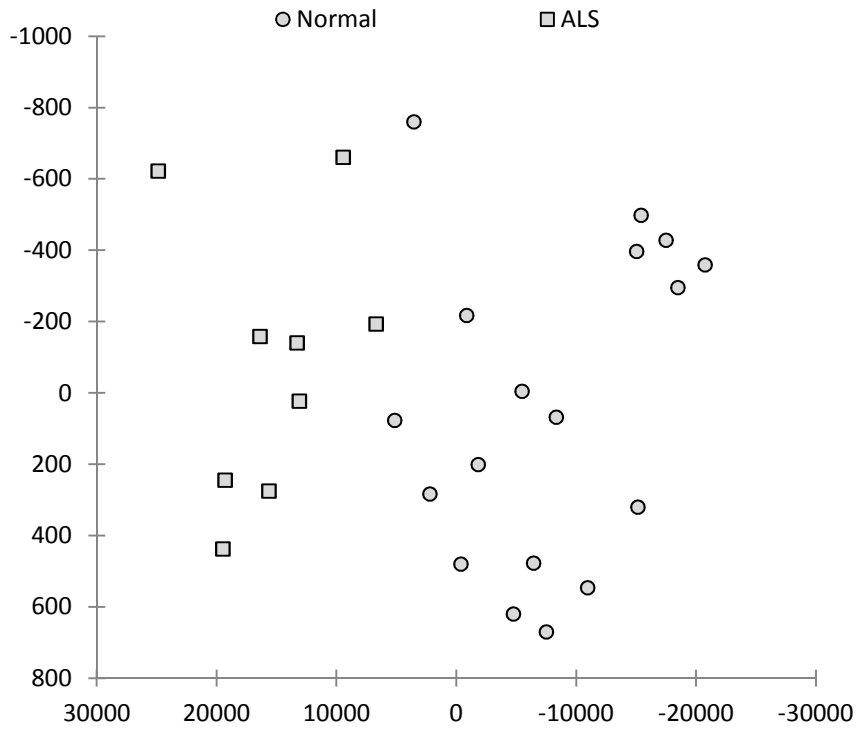

Figura 5.6. Proyeccion de los puntos muestrales usando PCA y los 10 genes mas significados otorgados por el clasificador de vecinos cercanos con parametro de $\boldsymbol{k = 3}$

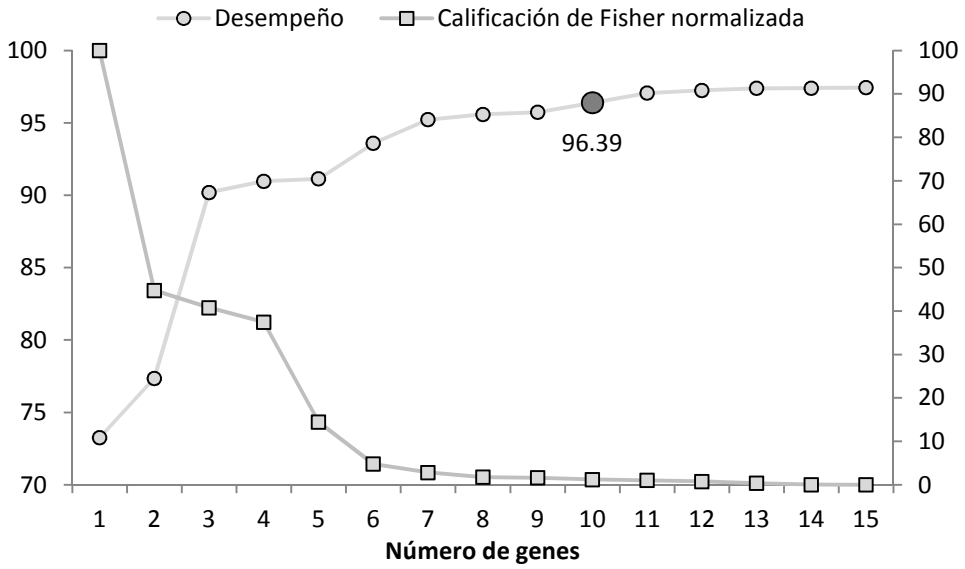

Figura 5.7. Desempeño del clasificador bayesiano ingenuo

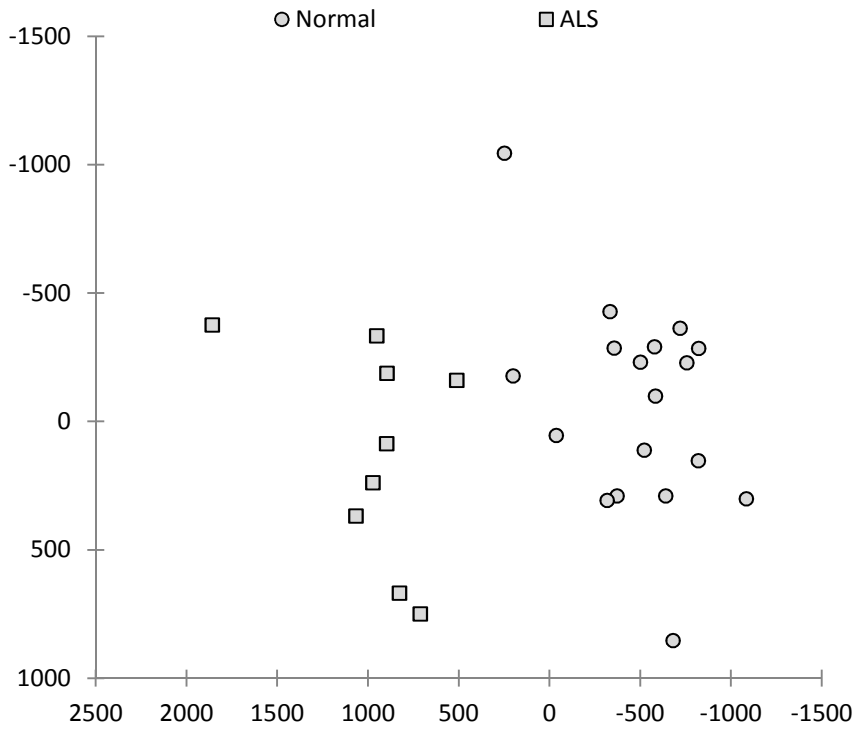

Figura 5.8. Proyeccion de los puntos muestrales usando PCA y los 10 genes mas significados otorgados por el clasificador bayesiano ingenuo

En la tabla 5.1 se muestran los genes más significativos que fueron otorgados por cada clasificador, los números de los genes es el índice de la columna en la tabla original otorgada por los microarreglos, cada uno de estos números corresponde a un gen en la base de datos del genoma humano.

Tabla 5.1. Los 10 genes más significativos otorgados por los mejores clasificadores

Clasificador	**Desempeño promedio**	**Intervalo de confianza**	**Genes seleccionados**
1NN	97.09	(95.6%, 99.1%)	18215, 9769, 17801, 2652, 21752, 1002, 3866, 8912, 3205, 9381
3NN	97.32	(95.7%,99.2%)	18215, 9769, 17801, 2652, 1125, 14070, 2698, 1002, 21925, 8912
Bayesiano Ingenuo	96.36	(95.2%, 98,97%)	18215, 20451, 9769, 17801, 6170, 2652, 14070, 9543, 7246, 10487

5.1.4. Conclusiones

Los resultados otorgados por el clasificador bayesiano ingenuo y el de vecinos cercanos son muy buenos, pero a pesar de que estos resultados son replicables para el mismo lote bootstrap, al generar un lote bootstrap distinto solo algunos genes vuelven a ser seleccionados, indicando que los resultados son consistentes. Se debe realizar otro tipo de experimentación para afirmar la relevancia que tienen los genes seleccionados con el ALS.

5.2. Segundo experimento

5.2.1. Datos

Los datos son los mismos que el primer experimento, experimento “E-GEOD-3307” de la base de datos proporcionados por EMBL-EBI.

5.2.2. Procedimiento

En este experimento se creó un lote de 1000 muestras bootstrap estratificadas, la estratificación consiste en que el mismo número de muestras de cada clase en el lote original este también en cada muestra bootstrap.

No solo se redujo el número de muestras bootstrap, sino que también se utilizó la calificación de Fisher para filtrar gran parte de los genes, solo se utilizaron los 500 genes más significativos que la calificación de Fisher señaló, esto se debe a que en este experimento, la selección hacia adelante y hacia atrás no van a utilizar la calificación de Fisher como jerarquía de variables, por lo cual estos algoritmos consumen mucho más recursos computacionales que los algoritmos del primer experimento.

Solo se utilizaron el clasificador bayesiano ingenuo y el vecino más cercano con parámetros k=1 y k=3, siendo que el clasificador lineal y cuadrático se desempeñaron pobremente en el experimento anterior.

5.2.3. Resultados

El clasificador de vecinos cercanos arrojó los mismos genes en la selección hacia adelante usando los parámetros $k = 1$ y $k = 3$ como se exhibe en la tabla 5.2, es por esta razón que solo se presentó una sola proyección en la figura 5.9.

Todas las proyecciones están generadas mediante el mismo método de PCA al igual que el primer experimento. No fue posible mostrar las gráficas de desempeño porque ya no se está usando la calificación de Fisher como guía en la selección de genes, los genes son elegidos solamente por la precisión promedio que arroja el clasificador al evaluar las muestras bootstrap.

Tabla 5.2. Los genes seleccionados por los diferentes algoritmos

Clasificador	Desempeño promedio	Intervalo de confianza	Genes seleccionados
Selección hacia adelante, 1NN y 3NN	100%	(99.8%, 100%)	3581, 13641
Selección hacia adelante, Bayesiano Ingenuo	100%	(99.8%, 100%)	19413, 13192, 21725, 7476, 8671, 14242, 11959, 3250
Selección hacia atrás, 1NN	100%	(99.9%, 100%)	693, 2767, 12689, 716, 5450
Selección hacia atrás, 3NN	100%	(99.9%, 100%)	17778, 12180, 2767, 1391, 17679, 12689, 716, 5340, 3932, 8690, 17684, 13492, 609, 9488, 13624, 5450
Selección hacia atrás, Bayesiano Ingenuo	100%	(99.8%, 100%)	21386, 21627, 12812, 3581, 18793, 12193, 13192, 2767, 12380, 11889, 855, 7997, 17679, 12749, 5340, 8690, 2196, 18987, 1346, 817, 6267, 13492, 609, 9488

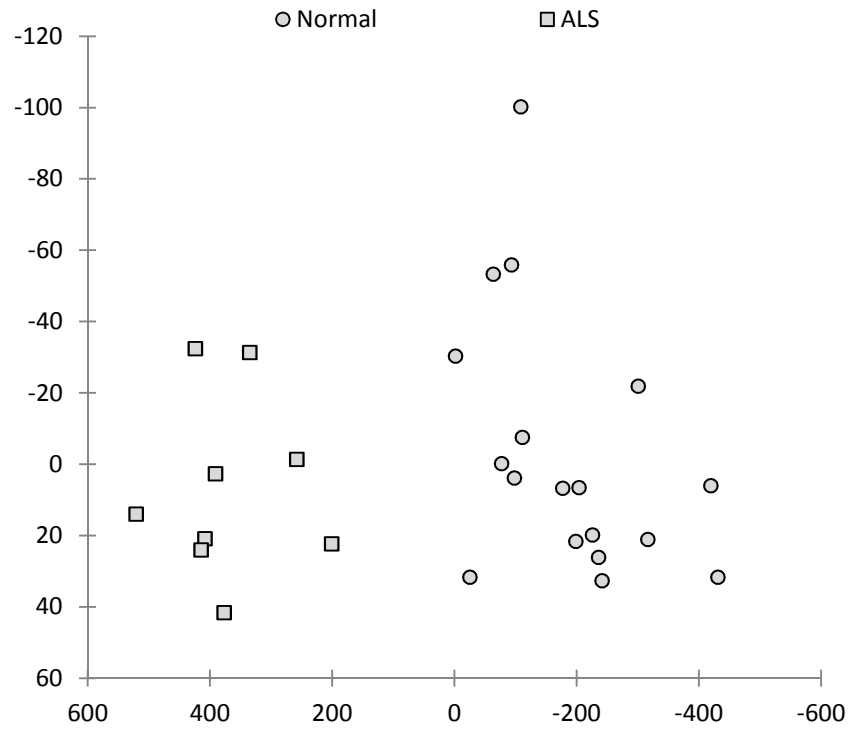

Figura 5.9. Proyeccion de los puntos muestrales usando PCA y los genes obtenidos en la selección hacia adelante con el clasificador de vecinos cercanos

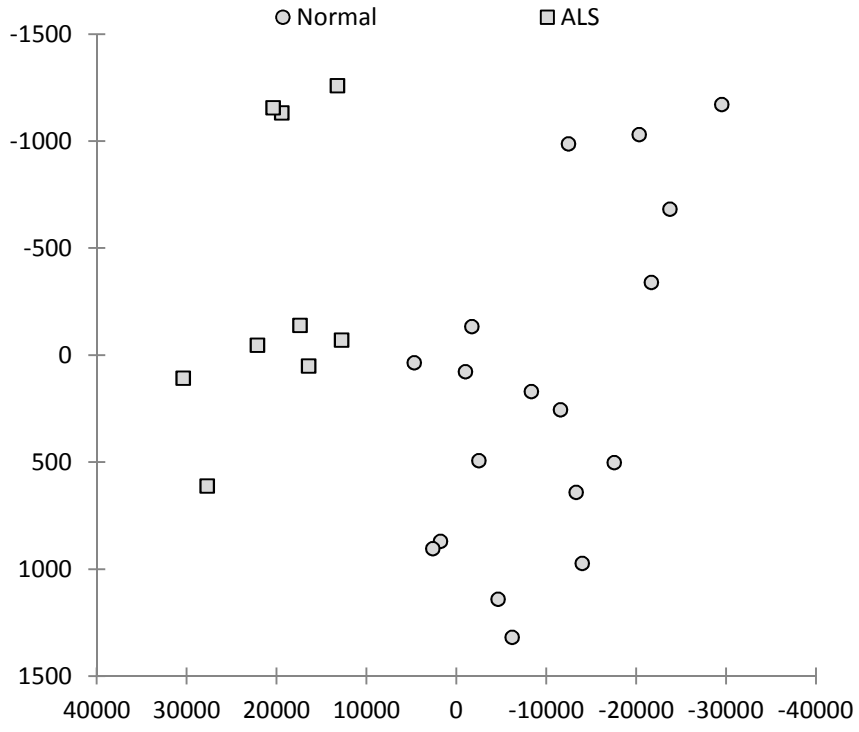

Figura 5.10. Proyeccion de los puntos muestrales usando PCA y los genes obtenidos en la selección hacia adelante con el clasificador bayesiano ingenuo

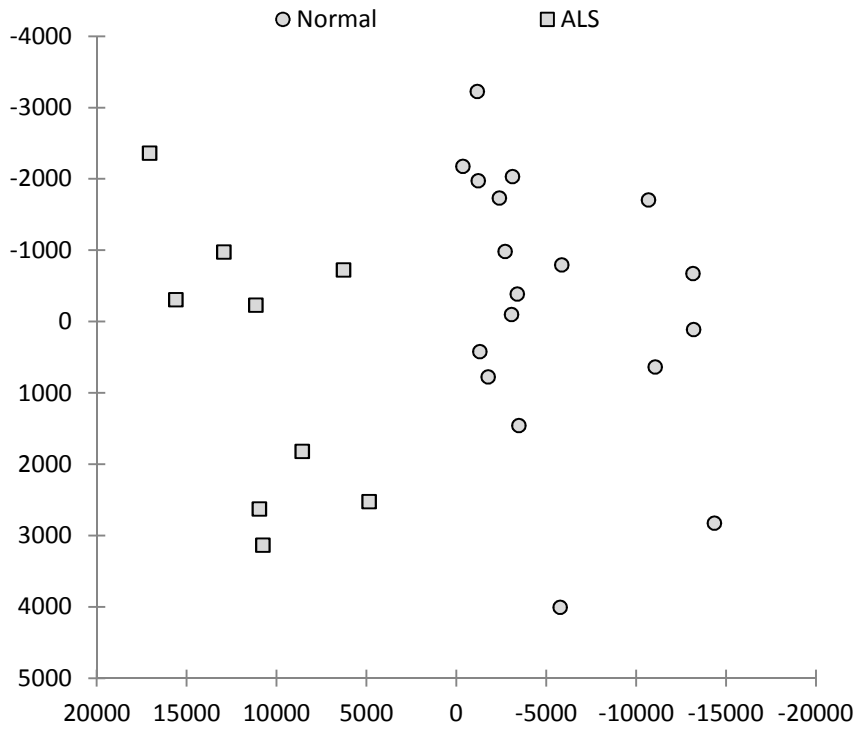

Figura 5.11. Proyeccion de los puntos muestrales usando PCA y los genes obtenidos en la selección hacia atras con el clasificador de vecinos cercanos con parametro $\boldsymbol{k = 1}$

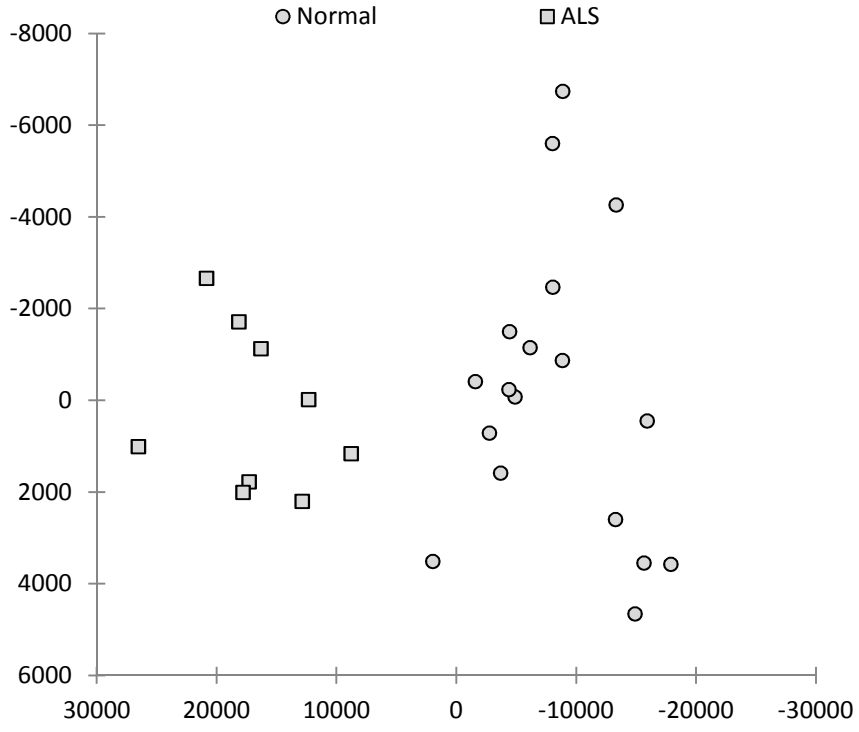

Figura 5.12. Proyeccion de los puntos muestrales usando PCA y los genes obtenidos en la selección hacia atras con el clasificador de vecinos cercanos con parametro $\boldsymbol{k = 3}$

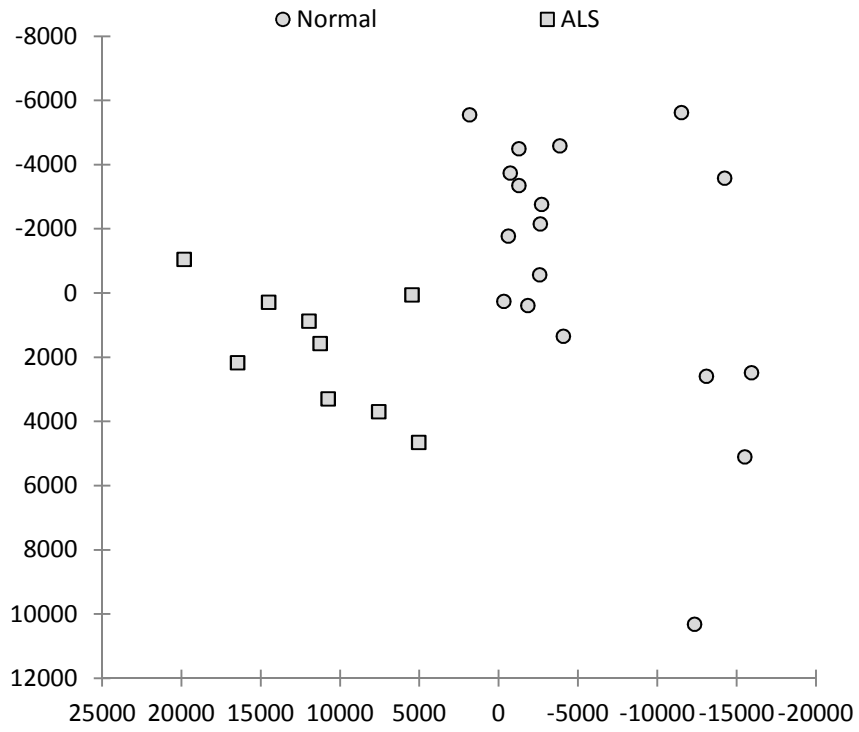

Figura 5.13. Proyeccion de los puntos muestrales usando PCA y los genes obtenidos en la selección hacia atras con el clasificador bayesiano ingenuo

5.2.4. Conclusiones

A simple vista este experimento es mucho más exitoso que al anterior, pero el tener 100% en el desempeño es señal de sobreajuste, otro dato a mencionar es que existen genes que aparecen en la selección de diferentes algoritmos, pero ninguno de estos genes apareció en el experimento pasado. Gracias a las proyecciones, no se puede negar el poder discriminante de ambos experimentos, pero para poder decir que cierto grupo de genes está involucrado en la enfermedad del ALS se necesita mucho más datos y experimentación.

5.3. Tercer experimento

5.3.1. Datos

Proporcionado por la base de datos EMBL-EBI, el nuevo microarreglo proviene del experimento E-TABM-940. Contiene 85 muestras, de las cuales, 28 son de personas saludables y 57 de pacientes con ALS. Contienen 54675 características.

Se estima que el genoma humano está compuesto por 20500 genes, los genes que un chip de microarreglos diseñado para el genoma humano por lo general tiene más características que el número total de genes en una persona, esto se debe a que el microarreglo debe sondear todos los posibles genes que pueden presentarse, por lo que entre más características puede registrar un microarreglo, se adquiere mayor detalle de la estructura genética. En este experimento no solo se tiene un número mayor de muestras, sino que también se duplicó el número de posibles genes a seleccionar.

Los archivos del experimento son de formato CEL, se utilizó el programa AltAnalyze para obtener la matriz numérica equivalente al primer experimento. Como se mencionó en capítulos anteriores, existen diferentes categorías del ALS, en este experimento existen 4 muestras que contienen la etiqueta de ALS probable, pero fueron incluidas en el grupo de ALS definitivo para simplificar los algoritmos de selección.

5.3.2. Procedimiento

Se creó un lote de 2000 muestras bootstrap 0.632+ estratificadas y solo se utilizó la selección hacia adelante. Los clasificadores que se utilizaron fueron el discriminante lineal de Fisher, discriminante cuadrático, bayes ingenuo y el vecino más cercano con parámetros de $k = 1$ y $k = 3$. El algoritmo utilizado para el QDC fue el que proporciona la librería de matlab PRTOOLS (http://prtools.org).

El experimento es muy parecido al primero, pero en vez de usar el lote bootstrap promediando los resultados de cada característica, cada muestra bootstrap se utilizó independientemente, como si fuese la muestra original. El resultado es un histograma marcando la frecuencia que obtuvo cada gen de ser seleccionado en las diferentes muestras.

Se usó la calificación de Fisher como filtro para usar solamente 2000 genes más significativos en los algoritmos de selección, también se usó la calificación de Fisher para ayudar a seleccionar genes según su relevancia. Tener un lote de 85 muestras en vez de 27, es una gran mejora, pero aun así, son muy pocas muestras para ser utilizadas en microarreglos directamente, y sin la ventaja de poder promediar los resultados entre

diferentes muestras bootstrap para la selección de genes, los clasificadores arrojan varios genes con el mismo poder discriminante, por lo que la selección es asistida con la calificación de Fisher, de esta manera se puede escoger el gen con la calificación más alta en un grupo de genes con el mismo poder discriminante.

5.3.3. Resultados

Al igual que en el primer experimento el clasificador lineal y cuadrático no se desempeñaron satisfactoriamente, ver figura 5.14 y figura 5.15. Los genes seleccionados por el clasificador lineal y cuadrático tienen tan poca frecuencia que no se mostraron en la tabla 5.3. Dicha tabla muestra en gris 3 genes que fueron seleccionados en los 3 clasificadores restantes, los cuales son {2222, 10394, 26573}, también existen genes que aparecen por lo menos en 2 selecciones, los cuales son {35088, 23877, 45203, 24146}, la proyección de estos 7 genes se muestra en la figura 5.22.

Tabla 5.3. Los genes seleccionados por los diferentes algoritmos

#	Bayesiano Ingenuo	Vecino Cercano, $k = 1$	Vecino Cercano, $k = 3$
1	2222	22717	2222
2	48194	2222	10394
3	10394	10394	22717
4	2075	26573	23877
5	24146	47939	26573
6	17826	29117	35088
7	29580	34014	24146
8	35088	23877	43865
9	45203	50310	45203
10	26573	45203	29117

Las casillas en gris son los genes que aparecieron en la selección de diferentes algoritmos

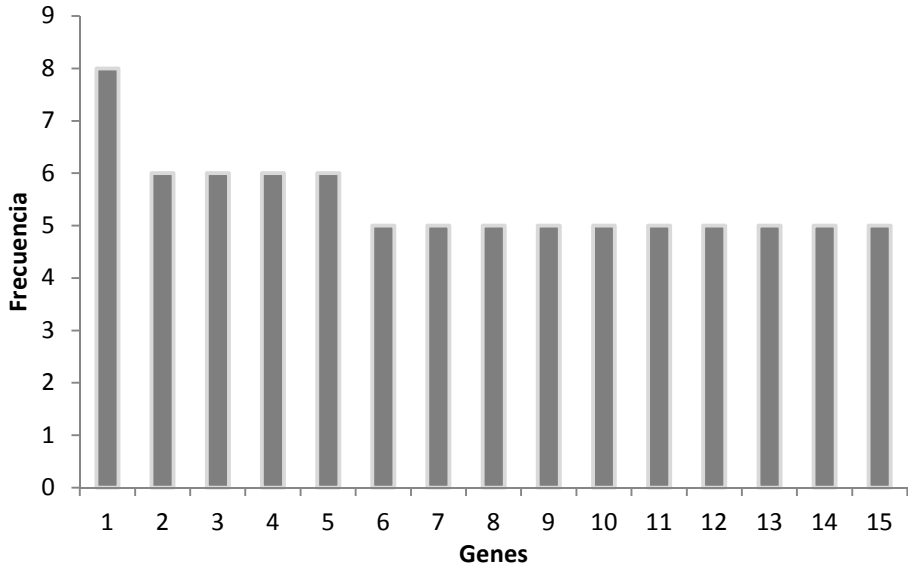

Figura 5.14. Histograma de los 15 genes con mayor frecuencia usando del clasificador lineal

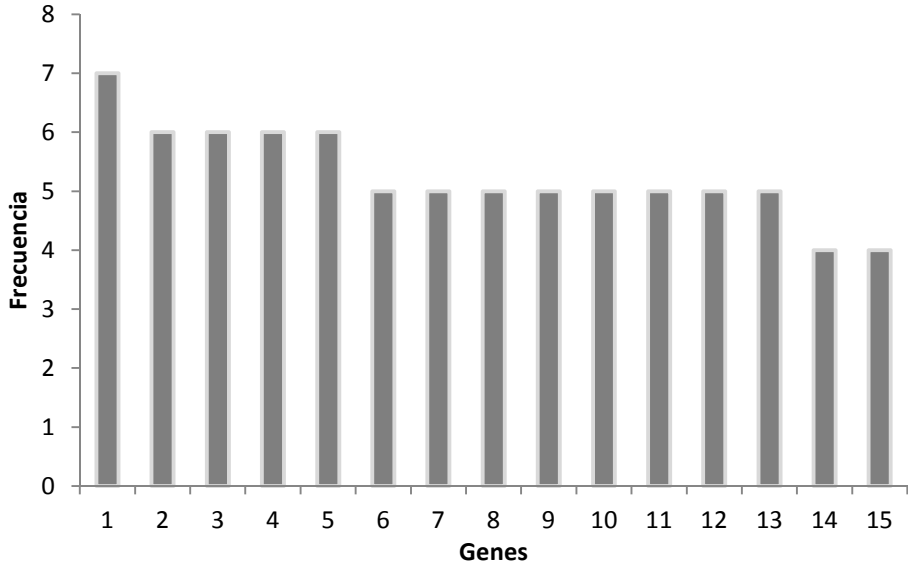

Figura 5.15. Histograma de los 15 genes con mayor frecuencia usando del clasificador cuadratico

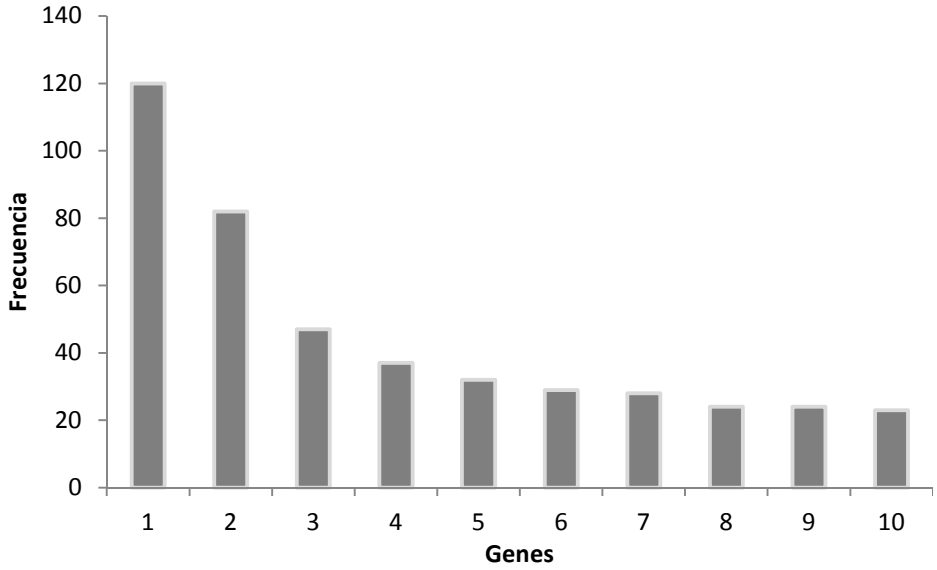

Figura 5.16. Histograma de los 10 genes con mayor frecuencia usando del clasificador de vecinos cercanos con parametro $\boldsymbol{k = 1}$

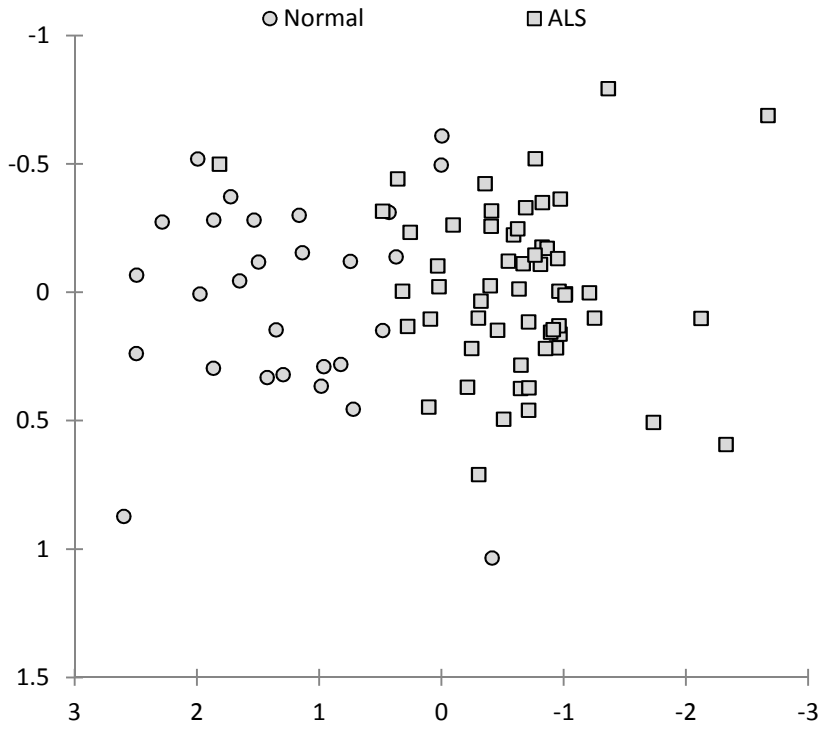

Figura 5.17. Proyeccion de los puntos muestrales usando PCA y los 10 genes con mayor frecuencia otorgados por el clasificador de vecinos cercanos con parametro $\boldsymbol{k = 1}$

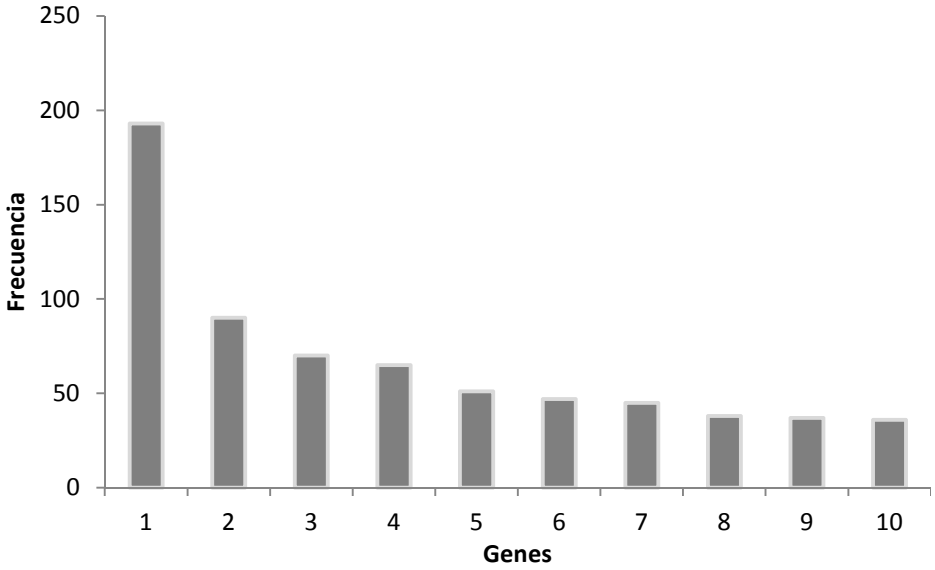

Figura 5.18. Histograma de los 10 genes con mayor frecuencia usando del clasificador de vecinos cercanos con parametro $\boldsymbol{k = 3}$

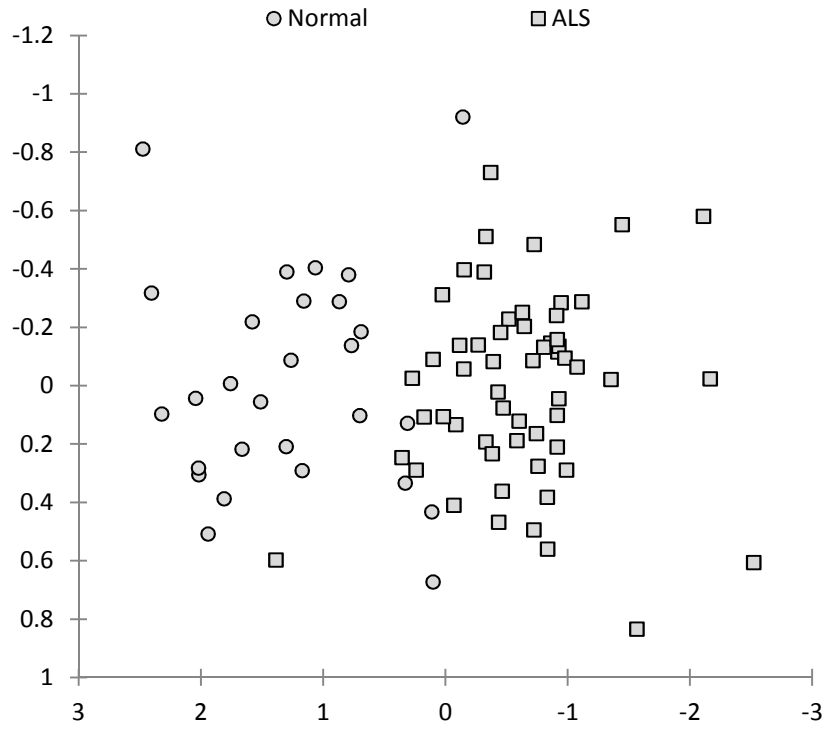

Figura 5.19. Proyeccion de los puntos muestrales usando PCA y los 10 genes con mayor frecuencia otorgados por el clasificador de vecinos cercanos con parametro $\boldsymbol{k = 3}$

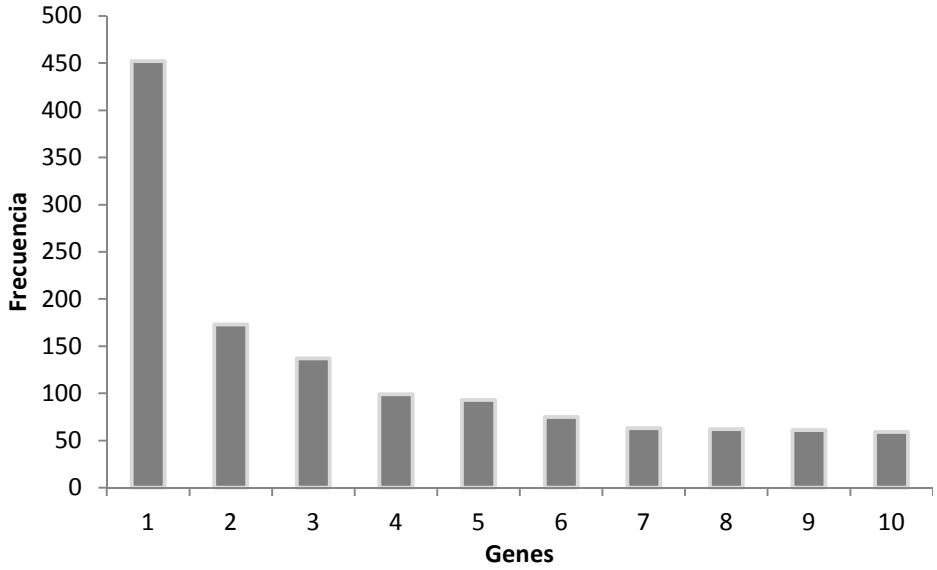

Figura 5.20. Histograma de los 10 genes con mayor frecuencia usando del clasificador bayesiano ingenuo

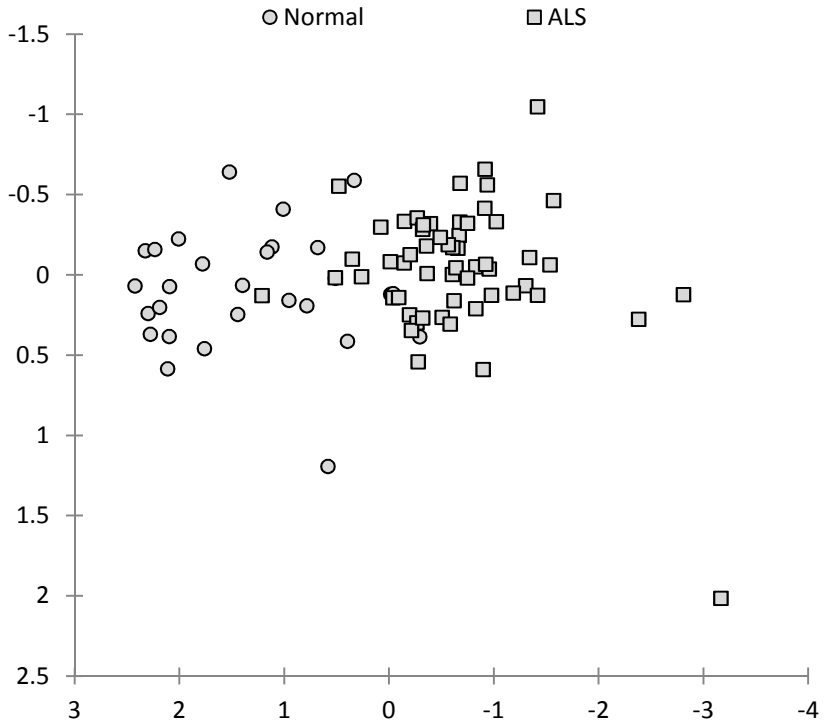

Figura 5.21. Proyeccion de los puntos muestrales usando PCA y los 10 genes con mayor frecuencia otorgados por el clasificador bayesiano ingenuo

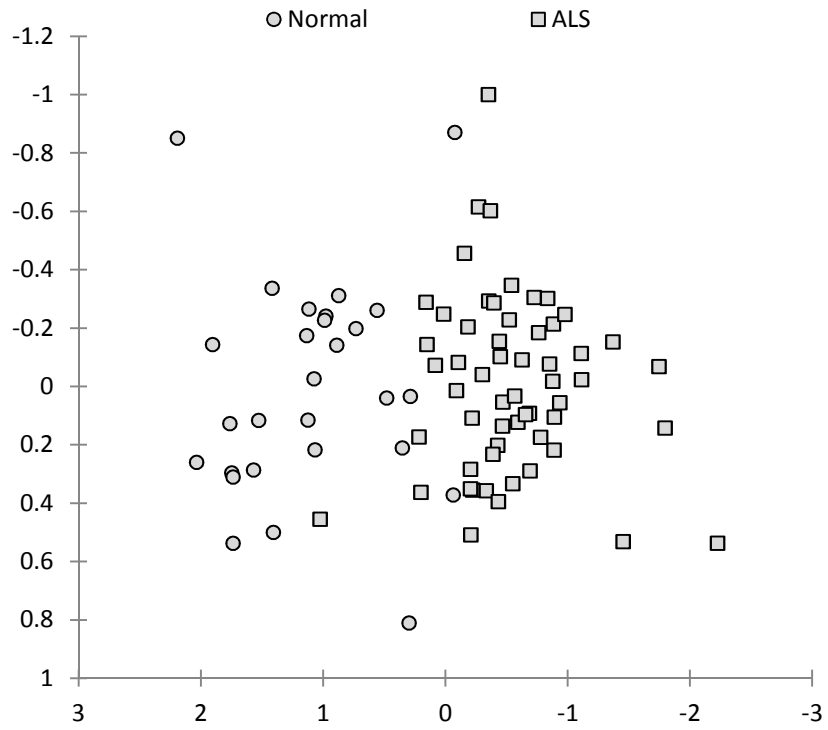

Figura 5.22. Proyeccion de los puntos muestrales usando PCA y los genes repetidos en las selecciones del clasificador bayesiano ingenuo y vecinos cercanos

5.3.4. Conclusiones

En los primeros dos experimentos a pesar de que los genes tenían gran poder discriminativo, era poco probable que volvieran aparecer cuando el lote de muestras bootstrap cambiaba, esto marca un problema en la selección de características.

En el presente experimento cada muestra bootstrap se maneja de forma independiente, simulando así el cambio de lote de muestras, forzando de esta manera la selección de genes más allá del desempeño de los clasificadores, tomando en consideración la frecuencia que tienen los genes de ser elegidos en diferentes particiones muéstrales, esta práctica puede marcar un mayor vínculo con la enfermedad.

Se creó una proyección de los genes que aparecieron en los primeros 10 lugares en diferentes clasificadores, la cual mostró un gran poder discriminatorio, respaldando de esta manera la elección de esos genes.

Tabla 5.4. Nombre y descripción de los genes seleccionados

#	ID	Gen	Descripción
1	2222	241425_at	Nucleoporin like 1 (código de proteínas)
2	10394	242066_at	No existe descripción.
3	26573	244470_at	Proteína del dedo de anular, dominio interactivo LIM
4	35088	212177_at	Serine interactivo PNN / proteína rica en arginine
5	23877	205327_s_at	Receptor A activin, tipo IIA (código de proteínas)
6	45203	230389_at	Formación de enlaces de proteínas 1
7	24146	215857_at	Nicalin (código de proteínas)

Como se mencionó anteriormente, los números que se han estado manejando para identificar los genes, son realmente el índice de la columna del arreglo numérico que representa el microarreglo, el nombre de los genes seleccionados en este experimento se muestra en la tabla 5.4, la descripción fue obtenida gracias a la página web GeneCards, el link es: http://www.genecards.org

Capitulo 6
Conclusiones y Futuros Trabajos

Con el desarrollo de la tecnología de microarreglos, en la actualidad es posible monitorear no solo el total del genoma humano, sino también las mutaciones más comunes, aun así la enfermedad de la esclerosis lateral amiotrófica elude la comprensión de la comunidad científica como para poder combatirla efectivamente. Por desgracia, si una persona la contrae no hay mucho que se pueda hacer al respecto.

Los resultados que muestra este documento son genes que pueden estar vinculados con la enfermedad, sin embargo, es necesaria más experimentación para garantizar su influencia. Como se vio en el progreso de los 3 experimentos, a pesar de que estos genes son capaces de hacer una buena clasificación, su capacidad de generalización puede ser muy pobre, dando como conclusión poca relevancia en el desarrollo de la enfermedad.

Existen experimentos donde las mutaciones genéticas tenían que ser demostradas en diferentes lotes muéstrales para comprobar su validez, puesto que dichos lotes eran muy pequeños para ser válidos por si solos. Por ejemplo, en el descubrimiento de la variante ALS2, en el año del 2001, dos grupos de científicos vincularon una mutación en el cromosoma 2q33 la cual era responsable de la aparición de la enfermedad en familias del norte de África y el medio oriente [29]. Es posible que los genes encontrados en esta investigación coincidan con otra investigación demostrando así su influencia en la enfermedad.

Para futuros trabajos se podrían utilizar diferentes tipos de técnicas para la selección genética, como las máquinas de soporte vectorial, arboles de decisión o clustering. Estas dos últimas técnicas fueron ampliamente recomendadas en los libros que muestran cómo manejar los datos proporcionados por microarreglos. También existen artículos que han demostrado grandiosos resultados al combinar varios métodos en la selección, estas prácticas son particularmente eficientes manejando el ruido en los datos, el cual en datos genómicos es representado por genes irrelevantes.

El 5 de diciembre del 2012 Prize4Life, una organización no lucrativa con la misión de acelerar el descubrimiento de tratamientos para el ALS, junto con el NCRI (Neurological Clinical Research Institute) proporcionaron los fondos necesarios para el lanzamiento de una nueva base de datos de pacientes que presentan la enfermedad, PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials). La visión de PRO-ACT es que la comunidad de ALS y otros investigadores interesados alrededor del mundo, tengan acceso libre a datos que previamente eran inaccesibles, con la finalidad de acelerar el entendimiento de la enfermedad y finalmente el poder combatirla.

Actualmente PRO-ACT contiene 8500 muestras, las cuales pueden hacer una gran diferencia en futuros trabajos, estos registros provienen de compañías farmacéuticas como Sanofi, Novartis, Teva Pharmaceutical Industries y Regeneron Pharmaceuticals, los registros incluyen datos demográficos, antecedentes médicos, calificaciones funcionales y otros datos adquiridos en el diagnóstico de la enfermedad, la liga de internet es: https://nctu.partners.org/ProACT.

Referencias

[1] Artículo: Pamela A. Cazzolli, Understanding Amyotrophic Lateral Sclerosis. ALS Care Project, 2009.

[2] Libro: Harrison, Harrison's Principles of Internal Medicine, McGraw Hill, 2009.

[3] Manual: MD Andrew J. Waclawik; Neurodegenerative Disorders: Amyotrophic Lateral Sclerosis and Inclusion Body Myositis, Neurology board review manual, 2004.

[4] Diario: Ronald Steriti, PhD; Amyotrophic Lateral Sclerosis (ALS), 2002.

[5] Manual: MDA (Muscular Dystrophy Association Inc.), Facts About Amyotrophic Lateral Sclerosis, 2009.

[6] Diario: EPMA Journal (The European Association for Predictive, Preventive and Personalised Medicine), Moving toward a predictive and personalized clinical approach in amyotrophic lateral sclerosis: novel developments and future directions in diagnosis, genetics, pathogenesis and therapies, 2010.

[7] Capítulo 11 de "Computational Genomics": An Introduction to Microarray Data Analysis, Madan Babu, Horizon Press, 2005.

[8] Artículo: Wolfgang Huber, Anja von Heydebreck and Martin Vingron; Analysis of microarray gene expression data, 2003.

[9] Artículo: Caroline Vance, Ammar Al-Chalabi, Deborah Ruddy, Bradley N. Smith, Xun Hu, Jemeen Sreedharan, Teepu Siddique, H. Jurgen Schelhaas, Benno Kusters, Dirk Troost, Frank Baas, Vianney de Jong and Christopher E. Shaw; Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2–21.3, Brain, 2006.

[10] Artículo: John Ravits, MD, FAAN, Patrick Laurie, BS, Brad Stone, PhD; Amyotrophic Lateral Sclerosis Microgenomics, Elsevier Saunders, 2005.

[11] Artículo: SFN (Society for neuroscience); ALS Making a difference today, 2005.

[12] Artículo: Isabelle Guyon, André Elisseeff; An Introduction to Variable and Feature Selection; Journal of Machine Learning Research; 2003.

[13] Página WEB: ALS Association, http://www.alsa.org/research/article.cfm?id=813

[14] Artículo: Ralf Dahm; Friedrich Miescher and the discovery of DNA; Elsevier; 2005

[15] Artículo: Severo Ochoa; Enzymatic synthesis of ribonucleic acid; Nobel Lecture; 1959

[16] Libro: Alberts, Johnson, Lewis, Raff, Roberts, Walter; Molecular Biology of The Cell 4th Edition; Garland Science; NCBI, 2002

[17] Artículo: James C. Alwine, David J. Kemp, George R. Stark; Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes; Biochemestriy; 1977

[18] Artículo: T. Cremer, C. Cremer; Centennial of Wilhelm Waldeyer's introduction of the term "chromosome" in 1888; Cytogenet Cell Genet; 1988

[19] Libro: Lodish, Berk, Matsudaira, Kaiser, Krieger, Scott, Zipursky, Darnell; Molecular Cell Biology 5th Edition; NCBI; 2006

[20] Artículo: Affymetrix; Activity 2 – Student Reading, Structure and Fuction of GeneChip Microarray; Affymetrix Inc.; 2005

[21] Artículo: Michel Verleysen y Damien Francois; The Curse of Dimensionality in Data Mining and Time Series Prediction; UCL, 2005
[22] Libro: Ethem Alpaydin; Introduction to Machine Learning 2nd Edition; The MIT Press, 2010
[23] Artículo: Mostafa MJAHED; Classification of Multi-jet Topologies in e^+ e^- collisions using Multivariate Analysis Methods and Morphological Variables; Ecole Royale de l'Air; 2001
[24] Artículo: Stefanos Ougiaroglou, Alexandros Nanopoulos, Apostolos N. Papadopoulos, Yannis Manolopoulos, Tatjana Welzer-Ruzovec; Adaptive k-Nearest Neighbor Classification Based on a Dynamic Number of Nearest Neighbors; Welzer; 2007
[25] Artículo: Kevin P. Murphy; Naïve Bayes Classifiers; Radical Eye Software; 2006
[26] Artículo: Philip M. Dixon; The Bootstrap; Radical Eye Software; 2001
[27] Artículo: Wenyu Jiang y Richard Simon; A comparison of bootstrap methods and an adjusted bootstrap approach for estimating prediction error in microarray classification; Biometric Research Branch, Division of Cancer Treatment and Diagnosis; 2007
[28] Libro: Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan; Data Mining, A knowledge Discovery Approach; Springer; 2007
[29] Artículo: Kohsuke Kanekura, Yuichi Hashimoto, Takako Niikura, Sadakazu Aiso, Masaaki Matsuoka, Ikuo Nishimoto; Alsin, the product of ALS2 gene, supresses SOD1 mutant neurotoxicity through RhoGEF domain by interacting with SOD1 mutants; The Journal of Biological Chemistry; 2004
[30] Artículo: Lindsay I. Smith; A tutorial on Principal Components Analysis; 2002