CENTRO DE INVESTIGACIÓN Y ASISTENCIA EN TECNOLOGÍA Y DISEÑO DEL ESTADO DE JALISCO, A. C.

DISEÑO DE UNA HERRAMIENTA AUXILIAR EN EL DIAGNÓSTICO DE CÁNCER DE PULMÓN MEDIANTE LA CUANTIFICACIÓN DE PROTEÍNAS BIOMARCADORAS UTILIZANDO REDES NEURONALES ARTIFICIALES.

TESIS

QUE PARA OBTENER
EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIA Y TECNOLOGÍA
EN LA ESPECIALIDAD DE
BIOTECNOLOGÍA PRODUCTIVA

PRESENTA

I.I.A. JOSÉ MIGUEL FLORES FERNÁNDEZ

GUADALAJARA, JAL. AGOSTO 2011.

CIENCIA Y TECNOLOGIA

Guadalajara, Jalisco a 29 de Julio de 2011

Dr. Guillermo Rodríguez Vilomara
Director de Posgrado
PICYT-CIDESI
Querétaro

Los abajo firmantes miembros del comité tutorial del estudiante **José Miguel Flores Fernández**, una vez leída y revisada la Tesis titulada "**DISEÑO DE UNA HERRAMIENTA AUXILIAR EN EL DIAGNÓSTICO DE CÁNCER DE PULMÓN MEDIANTE LA CUANTIFICACIÓN DE PROTEÍNAS BIOMARCADORAS UTILIZANDO REDES NEURONALES ARTIFICIALES**" aceptamos que la referida tesis revisada y corregida sea presentada por el alumno para aspirar al grado de Maestro en Ciencia y Tecnología en la opción terminal de Biotecnología productiva durante el examen correspondiente.

Y para que así conste firmamos la presente a los veintinueve días del mes de Julio de 2011

Dr. Moisés Martínez Velázquez
Tutor

Dra. Enrique Jaime Herrera López
Tutor en Planta

Dr. Alejandro Ricardo Femat Flores
Asesor

CIENCIA Y TECNOLOGIA

Guadalajara, Jalisco a 17 de Agosto de 2011

Dr. Guillermo Rodríguez Vilomara
Director de Posgrado
PICYT-CIDESI
Querétaro

Los abajo firmantes miembros del Jurado de Examen del estudiante **José Miguel Flores Fernández**, una vez leída y revisada la Tesis titulada "**DISEÑO DE UNA HERRAMIENTA AUXILIAR EN EL DIAGNÓSTICO DE CÁNCER DE PULMÓN MEDIANTE LA CUANTIFICACIÓN DE PROTEÍNAS BIOMARCADORAS UTILIZANDO REDES NEURONALES ARTIFICIALES**" aceptamos que la referida tesis revisada y corregida sea presentada por el alumno para aspirar al grado de Maestro en Ciencia y Tecnología en la opción terminal de Biotecnología productiva durante el examen correspondiente.

Y para que así conste firmamos la presente a los veintinueve días del mes de Julio de 2011

Dra. Ana María Puebla Pérez
Presidente

Dra. Griselda Quiroz Compeán
Secretario

Dra. Erika Nahomy Marino Marmolejo
Vocal

Dr. Rodolfo Hernández Gutiérrez
Vocal

Dr. Moisés Martínez Velázquez
Vocal

AGRADECIMIENTOS

Al Fondo Sectorial de Investigación en Salud y Seguridad Social SSA/IMSS/ISSSTE-CONACYT, por el apoyo otorgado al proyecto 87628: “Análisis de la expresión de biomarcadores seleccionados específicos de cáncer de pulmón, y desarrollo de una prueba de ELISA prototipo para el diagnóstico de la enfermedad”.

Al Consejo Nacional de Ciencia y Tecnología por el apoyo económico otorgado. No. de becario:

Al Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco que me brindó la oportunidad de realizar esta Tesis de Maestría en el área de biotecnología médica y farmacéutica.

A las instituciones de salud que permitieron el reclutamiento de pacientes y la toma de muestras biológicas para realizar el presente trabajo: Centro Oncológico Estatal (COE) del Instituto de Seguridad Social del Estado de México y Municipios (ISSEMyM), ubicado en la ciudad de Toluca, México, a la Unidad Médica de Alta Especialidad (UMAE) del Centro Médico Nacional de Occidente (CMNO) y al Hospital Civil de Guadalajara (HCG), “Fray Antonio Alcalde”, ubicados en el Municipio de Guadalajara, Jalisco.

A quien considero un gran amigo, el Dr. Enrique Jaime Herrera López, quien me dirigió en este trabajo de tesis de maestría, con el cual he convivido y compartido momentos de alegría, de trabajo. Que me ha impulsado a seguir adelante y que con su ayuda y esfuerzo logré concluir esta meta.

Al director del proyecto, el Dr. Moisés Martínez Velázquez, quien hizo posible la realización de este trabajo de investigación, así mismo por haberme aceptado como tesista de maestría.

Al Dr. Héctor Escalona Buendía por haberme asesorado y guiado en lo relacionado a estadística.

Al grupo de neumólogos: Francisco Sánchez Llamas y Luz Audina Mendoza Topete de CMNO; Pablo Juárez Contreras del Hospital General Regional 46 y Marco Antonio Padilla Navarro del Hospital General Regional 110. Así mismo a los médicos oncólogos José Antonio Rojas Calvillo y Paula Anel Cabrera Galeana del COE por su participación en el reclutamiento de pacientes y toma de muestras biológicas.

AGRADECIMIENTOS

A Dios, por ayudarme a concluir este ciclo en la vida, gracias por darme la fuerza para hacer este sueño realidad y por estar conmigo en cada momento de mi vida. Por cada regalo de gracia que me has dado y que inmerecidamente he recibido.

Aquellos que con amor Me brindaron su confianza y calor, Imaginándome grande cuando aún era pequeño, Sabiendo que el tiempo Podría convertir su sueño en realidad. Ahora con la ilusión ya culminada, Desde hoy soy alguien más Respondiendo a sus anhelos y desvelos, Entendiendo hoy su labor y Su infinita dedicación.

Sabiendo que no existirá una forma de agradecer toda una vida de sacrificios y esfuerzos, quiero que mis padres sientan que este objetivo logrado también es suyo y que la fuerza que me ayudó a conseguirlo fue su cariño y apoyo.

A mi Madre que es el ser más maravilloso del mundo. Gracias por tu apoyo moral, tu cariño y comprensión que desde pequeño me has brindado; por guiar mi camino y estar siempre junto a mí en los momentos más difíciles.

A mi Padre porque desde pequeño ha sido para mí un hombre grande y maravilloso el cual siempre he admirado y seguiré admirando. Gracias por haber guiado mi vida con energía.

A mis hermanos Karen Isabel Flores Fernández y Efren Ricardo Flores Fernández por compartir tristezas y alegrías, éxitos y fracasos; por todos los detalles y afectos de cariño y apoyo que me han brindado durante mi vida como estudiante.

Con admiración y respeto para la mejor consejera de la familia, a mi Tía María de los Ángeles Fernández Montero que desde pequeño siempre me aconsejó para ir por el buen camino. Porque gracias a su apoyo y consejo he llegado a realizar una más de mis metas. Tía Te agradezco infinitamente y te prometo seguir siempre adelante como hasta ahora.

A mi Tía Paula Fernández Montero y a mi Tío Luis Manuel Robles quienes con su confianza, cariño y apoyo, sin escatimar esfuerzo alguno, me han aconsejado y guiado para ser una persona de provecho, ayudándome al logro de una meta más.

A uno de mis mejores amigos, el Q. José Abraham Álvarez García con quien aprendí el camino de la investigación, y con quien he compartido muchos logros a lo largo de mi vida profesional.

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El diagnóstico de cáncer de pulmón en etapas tempranas es de primordial importancia para el tratamiento oportuno de pacientes con esta enfermedad. Un biomarcador se utiliza para describir anormalidades o alteraciones fisiológicas en un organismo. Alteraciones en las concentraciones de proteínas biomarcadoras específicas pueden indicar la presencia de una enfermedad como el cáncer. No obstante, ningún biomarcador utilizado clínicamente de manera individual, ha demostrado ser lo suficientemente especifico y sensible para detectar cáncer de pulmón.

En este trabajo se evaluó en conjunto un grupo de biomarcadores con el propósito de incrementar la eficiencia en la detección de cáncer de pulmón. La interpretación de la información provista por estos biomarcadores es una tarea muy compleja y se requiere de potentes herramientas computacionales para el adecuado análisis de datos. Por esta razón, se desarrolló una Red Neuronal Artificial (RNA) para determinar que biomarcadores son relevantes en el diagnóstico de cáncer de pulmón en suero. La RNA se entrenó con catorce proteínas reportadas como posibles biomarcadores de cáncer de pulmón. La RNA clasificó correctamente 133 casos de 150 estudiados. Sin embargo, una prueba de detección con catorce proteínas no es viable para utilizarse como prueba de diagnóstico, debido a los costos asociados a la evaluación de cada proteína. Un análisis de componentes principales permitió reducir el número de biomarcadores con el que se entrenaba la RNA a seis y posteriormente a sólo cuatro, obteniendo los mismos resultados que si se hubieran utilizado todas las proteínas. La RNA entrenada con los biomarcadores Cyfra 21-1, CRP, CEA y CA125 produjo un aumento en la sensibilidad de 18.31%, es decir, 94.5%, en comparación con la del mejor biomarcador, Cyfra 21-1 (sensibilidad 76.19%), a una especificidad del 80%. La RNA mejoró significativamente la sensibilidad de los marcadores biológicos, por lo tanto, las RNAs son una prometedora herramienta auxiliar en el diagnóstico de cáncer de pulmón.

ABSTRACT

The diagnosis of lung cancer in early stages is of paramount importance for treatment of patients with this disease. A biomarker is used to describe physiological abnormalities or changes in the organism. Levels of altered concentration of specific biomarkers could indicate the presence of a disease like cancer. However, clinically no biomarker used individually, has proved to be specific enough to detect lung cancer.

In this work a combined panel of biomarkers was evaluated in order to increase efficiency in the detection of lung cancer. The interpretation of the information provided by these biomarkers is a complex task and requires powerful computational tools for their proper analysis. For this reason, we developed an artificial neural network (ANN) to determine which biomarkers are relevant in the lung cancer diagnosis in serum. The ANN was trained with fourteen reported proteins as potential biomarkers of lung cancer. The ANN correctly classified 133 of 150 cases studied. However, a screening of fourteen proteins is not viable for use as a diagnostic test due the costs of each protein. A principal component analysis allowed reducing the number of biomarkers for training the RNA from 14 to six and finally to only four. The ANN trained with biomarkers Cyfra 21-1, CRP, CEA and CA125 produced an increase in the sensitivity of 18.31%, i.e., 94.5% compared with the best biomarker, Cyfra 21-1 (sensitivity 76.19%) a specificity of 80%. The ANN significantly improved the sensitivity of biomarkers, therefore, ANNs are promising auxiliary tools in the diagnosis of lung cancer.

ORGANIZACIÓN DEL DOCUMENTO

EL presente documento se divide en cinco capítulos y cuatro apéndices organizados de la siguiente manera.

En el Capítulo 1 se presenta una introducción al tema, así como los antecedentes, justificación, objetivos, hipótesis y la descripción del proyecto. En el Capítulo 2 se presentan aspectos fundamentales de las redes neuronales artificiales. En el Capítulo 3 se expone la metodología seguida en la tesis. El Capítulo 4 presenta los resultados y discusiones de las redes neuronales artificiales desarrolladas y la comparación con diversos métodos estadísticos. El capítulo 5 presenta las conclusiones y las perspectivas del estudio. Posteriormente se enlistan las referencias bibliográficas y finalmente se presentan los siguientes apéndices: Apéndice A, Información sobre biomarcadores analizados. Apéndice B, Información de kits de ELISA comerciales. Apéndice C, descripción de algunos métodos estadísticos. Apéndice D, Publicaciones generadas por esta investigación.

Capítulo I

INTRODUCCIÓN

El cáncer es una de las principales causas de muerte a nivel mundial. De los diferentes tipos de cáncer que existen, el cáncer de pulmón ocupa el primer lugar en cuanto a mortalidad en hombres y el segundo en mujeres. El cáncer de pulmón está principalmente asociado al tabaquismo, aunque existen otros factores de riesgo, como: el virus del papiloma humano, la tuberculosis, la combustión con leña y carbón, la contaminación ambiental, la exposición a asbestos y radón, los factores genéticos, y condiciones crónicas como la enfermedad pulmonar obstructiva crónica (EPOC). Los síntomas del cáncer de pulmón se pueden confundir con algunas otras enfermedades pulmonares, esto ocasiona que el diagnóstico sea usualmente tardío. Recientemente se han buscado moléculas (biomarcadores) en el torrente sanguíneo. Cuando alguna de estas moléculas se encuentra alterada en su nivel de concentración en el organismo, se pudiera asociar con alguna enfermedad. Sin embargo, hasta la fecha ningún biomarcador ha demostrado ser lo suficientemente sensible (capacidad de detectar la enfermedad en una prueba diagnóstica), ni específico (capacidad de detectar personas que no presentan la enfermedad en una prueba diagnóstica). Actualmente se está evaluando la posibilidad de utilizar conjuntos de proteínas como biomarcadores específicos de enfermedades; sin embargo, determinar la combinación ideal de proteínas para estos propósitos, es una tarea compleja, y la interpretación de los datos obtenidos es complicada, lo que obstaculiza la aplicación de esta alternativa. La generación y aplicación de herramientas computacionales que faciliten la interpretación de los datos multifactoriales obtenidos en las áreas biomédicas es actualmente un campo de estudio de gran relevancia. Con base en lo anteriormente expuesto, este trabajo de maestría se enfocó al desarrollo y aplicación de herramientas auxiliares para la detección de cáncer de pulmón a partir del análisis de proteínas biomarcadoras, utilizando redes neuronales artificiales.

1.1 ANTECEDENTES

1.1.1 Anatomofisiología del pulmón

Los pulmones son un órgano par rodeado por la pleura. El espacio entre ambos recesos pleurales, se denomina mediastino y contiene órganos importantes como el corazón. Los pulmones tienen una forma semicónica irregular con la base dirigida hacia abajo. El pulmón derecho es más ancho que el izquierdo, aunque un poco más corto. El pulmón izquierdo, en la porción inferior presenta una incisura cardiaca, ya que es donde se encuentra el corazón (Fig. 1).

Los alveolos pulmonares tienen forma redondeada y su diámetro varía de acuerdo a la profundidad de la respiración. Los alveolos son la unidad terminal de la vía aérea y su función es llevar a cabo el intercambio gaseoso. Los alveolos se comunican entre sí por unas aberturas que reciben el nombre de poros de Kohn, que permiten una buena distribución de los gases entre los alvéolos y previenen del colapso por oclusión de la vía aérea pulmonar (Dmitriy y cols. 2002).

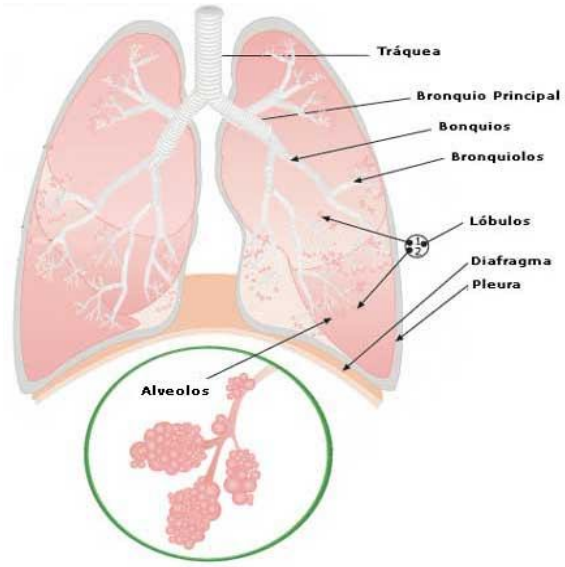

Figura 1. Anatomía del pulmón.

1.1.2 Patologías pulmonares

Las principales patologías de las vías respiratorias son:

El asma bronquial: es una enfermedad crónica recurrente, caracterizada por hiperactividad de las vías respiratorias, con broncoconstricción reversible desencadenada por distintos estímulos exógenos y endógenos (Hechavarría Miyares y Cols, 1999).

La neumonía: se caracteriza por la sustitución del aire de los alveolos por un exudado inflamatorio (neumonía bacteriana) o por la infiltración de las paredes alveolares y espacios intersticiales por células inflamatorias (neumonías víricas o atípicas) o ambas cosas. Sus principales síntomas son: disnea, aumento de frecuencia respiratoria, silbilancias, fiebre, somnolencia y falta de apetito (Bembibre y Lamelo, 2004).

El enfisema: es un aumento permanente de los espacios aéreos distales a los bronquiolos terminales, con destrucción de paredes, sin signos de fibrosis (Saínz-Menéndez, 2006).

Bronquiectasias: Es una dilatación permanente de bronquios y bronquiolos debido a una destrucción del músculo y tejido elástico de sostén, por infecciones necrosantes, facilitadas por procesos congénitos u obstrucción bronquial crónica (Ocampo y Cols, 2008).

Enfermedad pulmonar obstructiva crónica (EPOC): es una enfermedad pulmonar crónica y progresiva caracterizada por una inflamación sistémica, pero predominante en el parénquima pulmonar y las vías aéreas, que causa destrucción alveolar y limitación al flujo aéreo no completamente reversible, que atrapa al aire y produce disnea progresiva y desacondicionamiento muscular periférico, tos y expectoración. Su causa principal es la inhalación crónica de partículas o gases nocivos, como el humo de cigarro, así como la exposición crónica al humo de leña. Se ha reportado que sólo del 15 al 20% de fumadores desarrollan la EPOC. No obstante el porcentaje puede ser mayor, debido a los síntomas clínicos que aparecen en forma tardía, lo que ocasiona que muchos pacientes no sean diagnosticados hasta que la enfermedad está avanzada (Celli, 2004). Lamentablemente el proceso inflamatorio es crónico, progresivo e irreversible.

Se ha estimado que los enfermos de EPOC tienen de 2 a 5 veces mayor probabilidad de desarrollar cáncer de pulmón (Barreiro, 2008). Se ha demostrado la existencia de una relación inversa entre el grado de obstrucción de las vías aéreas y el riesgo de desarrollar

cáncer pulmonar. Los mecanismos por los cuales la EPOC induce un aumento en el riesgo de la aparición de una neoplasia apenas se conocen y están en estudio (Barreiro, 2008).

1.1.3 Cáncer de pulmón

Las células son las unidades de vida más pequeñas del cuerpo humano. Normalmente el cuerpo genera células nuevas a medida que se necesitan para reemplazar a las células envejecidas que mueren (Takkunen, 2009). Este proceso sincronizado, en el tiempo y en el espacio, permite que siempre exista un balance adecuado de células para cada etapa de la vida. Algunas veces, este proceso no resulta ser el esperado y se produce una multiplicación de células de manera descontrolada en el cuerpo, en este caso en el pulmón; es decir, crecen células nuevas que no son necesarias y las células envejecidas no mueren. Estas células que crecen de manera descontrolada pueden formar una masa de tejido anormal que se le denomina neoplasia (Arias y Cols. 2001).

El cáncer de pulmón se produce habitualmente en las paredes internas de los bronquios, y al crecer puede obstruir el paso del aire y alterar la respiración (García-Carlos, 2008). Existen neoplasias benignas y malignas. Las neoplasias benignas son aquellas formadas por células que no se diseminan a otros tejidos o partes del cuerpo y no comprometen la vida de la persona. Las neoplasias malignas o los tumores pueden provocar la muerte de la persona ya que éstos suelen invadir el tejido a su alrededor y diseminarse a otros órganos. Cuando el cáncer se disemina de una parte del cuerpo a otra se le denomina metástasis (NCI, 2007). Las células malignas pueden trasladarse a través de la linfa o de la sangre y llegar a cualquier parte del cuerpo provocando un segundo tumor, que se denomina metastásico (García-Carlos, 2008).

1.1.3.1 Epidemiología

De la población mundial que falleció de cáncer en el año 2010, el 27.6% fue por cáncer de pulmón, seguido por el de próstata en hombres y el de mama en mujeres. Se calcula que de cada 100 pacientes recién diagnosticados con cáncer de pulmón, fallecen 74 (Jemal y Cols. 2010). En México, en el año 2004 por cada 100,000 habitantes, 6.49 morían debido al cáncer de pulmón (Ruíz-Godoy y Cols, 2007). La mayor mortalidad se presenta en los estados de la República Mexicana con serios problemas de contaminación ambiental tales

como: Nuevo León, Baja California, Chihuahua, Estado de México, Distrito Federal y Jalisco (Ruíz-Godoy y Cols, 2007).

1.1.3.2 Etiología

El tabaquismo activo, el tabaquismo pasivo, la EPOC, la contaminación ambiental, la neumonía, la tuberculosis, el virus del papiloma humano (VPH 16 y 18), la combustión de leña y carbón, así como la genética y los hábitos alimenticios, son factores que tienen influencia en el desarrollo del cáncer de pulmón, en las diferentes poblaciones en el mundo (Ying-Chin y Cols, 1997).

Un estudio realizado por Franco Marina y cols. (2006) estimó que el riesgo atribuible al tabaquismo activo fue de 55.76% para hombres y 27% para mujeres. En el caso de tabaquismo pasivo, el riesgo fue de 17% en mujeres y 3.9% en hombres. En este estudio se resalta que aproximadamente una tercera parte de los casos de cáncer de pulmón que aparecen en la Ciudad de México, no son explicados por el tabaquismo activo o pasivo, y este porcentaje es considerablemente mayor en mujeres. El estudio concluye que es necesario identificar contaminantes ambientales posiblemente asociados con el cáncer de pulmón.

1.1.3.3 Clasificación y subclasificación del cáncer de pulmón

El cáncer de pulmón de acuerdo al tipo de células presentes en el tumor se puede clasificar en: cáncer de células pequeñas (SCLC, por sus siglas en Inglés) y de células no pequeñas (NSCLC) (Ruiz Alonso y Cols, 2004). De estos se desprenden las siguientes categorías:

1.- **Carcinoma de células pequeñas y carcinoma combinado de células pequeñas**: Es un tumor muy agresivo, que tiende a diseminarse con facilidad. Se caracteriza por presentar marcadores de diferenciación neuroendócrina, lo que sirve para diferenciarlo de los NSCLC.

2.- **Carcinoma de células escamosas**: Es el más frecuente en los hombres, se deriva del epitelio de los bronquios proximales y por tanto es de localización central. Es el tipo histológico más relacionado con el hábito del tabaco.

3.- **Carcinoma de células grandes**: Es el menos frecuente, pero es el de peor pronóstico. Se presenta como una masa voluminosa de localización periférica, con afectación de hilo pulmonar y áreas asociadas de neumonitis.

4.- **Adenocarcinoma**: Se deriva del epitelio de los alveolos y las glándulas mucosas bronquiales. Es de localización periférica, afectando incluso a la pleura o pared torácica. Predomina en mujeres y en pacientes no fumadores y, aunque también está relacionado con el tabaco, se observa cuando existen cicatrices o patología pulmonar previa (Ruiz Alonso y Cols, 2004).

1.1.3.4 Etapas de cáncer de pulmón

Para determinar en qué etapa de desarrollo se encuentra el cáncer, se requiere evaluar el grado de extensión del tumor desde el sitio en donde se originó, así como el tamaño.

Etapas para cáncer de pulmón de células pequeñas (SCLC):

- Etapa limitada: El cáncer se encuentra sólo en un pulmón y en los ganglios linfáticos cercanos.
- Etapa extensa: El cáncer se ha diseminado fuera del pulmón donde se originó, a otros tejidos del tórax o a otras partes del cuerpo.

Etapas para cáncer de pulmón de células no pequeñas (NSCLC):

- Etapa I: El cáncer se encuentra únicamente en el pulmón y está rodeado por tejido normal.
- Etapa II: El cáncer se ha diseminado a los ganglios linfáticos cercanos.
- Etapa III. El tumor se ha extendido a la pared torácica o al diafragma cerca del pulmón; o se ha diseminado a los ganglios linfáticos en el área que separa los dos pulmones (mediastino); o a los ganglios linfáticos al otro lado del tórax, o a los del cuello.
- Etapa IV: El cáncer se ha diseminado a otras partes del cuerpo (Medicina, 2009).

1.1.3.5 Signos y síntomas de la enfermedad

Los signos y síntomas de la enfermedad que presentan la mayoría de los pacientes son:

- Tos o dolor en el tórax que no desaparece y que puede ir acompañada de expectoración.
- Un silbido en la respiración, falta de aliento.
- Tos o esputos con sangre.
- Ronquera o hinchazón en la cara y el cuello.
- Sensación de falta de aire.

1.1.3.6 Diagnóstico del cáncer de pulmón

Actualmente, el diagnóstico del cáncer de pulmón se basa en la utilización de los siguientes métodos (Tabla 1):

Tabla 1. Técnicas de diagnóstico de cáncer de pulmón.

Diagnóstico	Observaciones
Radiografía de Tórax	No se detecta en etapas tempranas, por lo tanto no reduce mortalidad.
Tomografía Computada (TC)	Del 5 al 15% de las lesiones positivas presentan adenopatías menores a 5 mm de diámetro, que no pueden ser detectadas, por lo tanto puede haber diagnósticos negativos.
Tomografía por emisión de positrones	Presenta falsos positivos por enfermedades granulomatosas.
Resonancia Magnética	Se ha reportado que tiene mayor ventaja que TC, pero no se demostró con una comparación.
Fibrobroncoscopía	Tiene alta sensibilidad. Es invasivo, el tiempo de obtención de resultados es tardado y la enfermedad puede avanzar. Hay sitios anatómicos donde no se puede acceder.
Citología de Esputo	Presenta falsos negativos (sensibilidad de 0.66 y falsos positivos con 0.09).

Además de los métodos descritos en la Tabla 1, desde hace algún tiempo se ha propuesto la utilización de biomarcadores como elementos complementarios que coadyuven en el diagnóstico oportuno de cáncer de pulmón (Fernández Suárez y Cols, 2007).

1.1.4 Biomarcadores en cáncer

El término biomarcador se aplica a toda molécula, sustancia o proceso alterado que es producido por las células tumorales o por el tejido normal circundante, como respuesta a la lesión tumoral, cuya presencia puede ser detectada en el suero u otros líquidos biológicos, y que es susceptible de utilizarse en la detección, diagnóstico, pronóstico, diagnóstico de recidivas o control evolutivo del tumor (Hayes y Cols, 1996). Los marcadores tumorales se comportan como indicadores de la presencia de una neoplasia maligna (Fernández-Suárez y Cols, 2007). La naturaleza del biomarcador es diversa, desde un ácido nucleico, un péptido, o una proteína, hasta procesos como apoptosis, angiogénesis y proliferación, entre otros, medibles por una técnica apropiada. Pueden ser detectables en tejido, plasma sanguíneo, saliva, orina y otros fluidos corporales (Schrohl y Cols, 2003).

Los biomarcadores séricos son aquellos detectados en sangre periférica de pacientes con cáncer. La presencia de células malignas en sangre fue descrita desde los años sesenta (Zidman y Cols, 1961) y centenares de estudios en la última década han reportado sustancias metabólicas en sangre, producto del proceso de transformación maligna, la cual incluye: aumento en la proliferación, pérdida de características morfológicas propias de un tejido o indiferenciación y pérdida de la adhesión, dando origen a la metástasis de muchos tipos de cáncer (Xi L y Cols, 2007)**.** Los marcadores tumorales “no-séricos” son aquellos que proceden de fluidos corporales diferentes a sangre, como por ejemplo la orina o los que se estudian a partir de tejido neoplásico de donde se obtiene una gran diversidad de componentes celulares detectables.

El primer marcador de cáncer se utilizó en 1846 y consistía en una inmunoglobulina monoclonal obtenida de la precipitación de una proteína en orina de enfermos con melanoma (Solomon y Cols, 1969). Entre 1963 y 1965 se descubrió la α- feto proteína y el antígeno carcinoembrionario, que son algunos de los biomarcadores más utilizados en la actualidad. Posteriormente en la década de los ochenta, la tecnología de los anticuerpos monoclonales facilitó el descubrimiento de una nueva gama de marcadores tumorales, como las glicoproteínas CA125, CA15.3, CA19.9 y el antígeno prostático específico (Cruz Tapias y Cols, 2008).

Desafortunadamente estos biomarcadores no son específicos de neoplasias, pudiendo encontrarse concentraciones considerables en diversas condiciones fisiológicas o

patológicos no tumorales. Es importante resaltar la importancia de observar alteraciones cuantitativas anormales en las concentraciones de un determinado biomarcador, dado que éstas pueden ser la señal de alarma de una enfermedad en el organismo (Fernández Suárez y Cols, 2007). Diferentes proteínas biomarcadoras se han propuesto para diagnosticar cáncer.

1.1.4.1 Correlación de biomarcadores con el cáncer

Las proteínas biomarcadoras han demostrado ser una potencial herramienta para el diagnóstico de distintos tipos de cáncer. Es plausible que al presentarse una enfermedad, el organismo genere una respuesta que altere más de un biomarcador, aumentando o disminuyendo los niveles basales. Últimamente se ha estado trabajando en la búsqueda y caracterización de biomarcadores tumorales únicos para cáncer (William, 2007); sin embargo, la investigación sobre biomarcadores para el diagnóstico temprano de cáncer de pulmón se ha realizado con poco éxito, esto se ha debido a que ningún marcador identificado ha presentado una adecuada sensibilidad y especificidad. No obstante, se ha demostrado que al analizar varios biomarcadores en combinación aumenta la precisión del diagnóstico (Li Zhong y Cols, 2005). Para el análisis e interpretación de distintos biomarcadores como: proteínas (Patz y Cols. 2007) o autoanticuerpos (Farlow y Cols. 2010) se han utilizando herramientas estadísticas multivariables para extraer la mayor información del análisis. En ocasiones este mecanismo de correlación no es evidente debido a las no linealidades e incertidumbre presentes en los procesos biológicos, y en ocasiones los métodos estadísticos no pueden obtener la información relevante del análisis. Algunas herramientas de la Inteligencia Artificial han sido utilizadas ampliamente en sistemas biológicos y biomédicos para resolver problemas donde se presente incertidumbre o donde haya escaso conocimiento del proceso. En especial las Redes Neuronales Artificiales (RNAs) han demostrado ser una herramienta auxiliar prometedora para el diagnóstico y tratamiento de diferentes enfermedades.

1.1.5 Aplicaciones de las redes neuronales artificiales en cáncer

Las Redes Neuronales Artificiales (RNAs), la lógica difusa y los algoritmos genéticos son herramientas de la inteligencia artificial. Estas herramientas han sido utilizadas en la clasificación y el reconocimiento de patrones e imágenes, optimización y control de procesos, minimización de funciones, predicción del clima y bolsa de valores. Las RNAs son herramientas de software que pueden imitar muy bien las complejas relaciones no

lineales entre los parámetros de un proceso, simplemente aprendiendo a partir de ejemplos sobre un conjunto de datos tomados de un proceso real (Andreas Lübbert y Rimvydas Simutis, 1994). Las RNAs en medicina se han utilizado desde los años noventas. Astion y Wilding (1992) aplicaron una RNA para determinar la malignidad en cáncer de mama, a partir de la edad del paciente y concentración de triglicéridos, colesterol, albúmina y el marcador CA15-3, entre otros. La RNA predijo correctamente el 80% de los casos. Zhou y cols (2002) propusieron un procedimiento automático de diagnóstico patológico denominado “detección basada en un ensamble neuronal” (NED) que utiliza un conjunto de RNAs, para identificar las células de cáncer de pulmón en las imágenes de biopsia por aguja obtenidas de los pacientes diagnosticados con cáncer de pulmón. La NED que propusieron logró no sólo un alto índice de identificación general, sino también una baja tasa de falsos positivos. La tasa de identificación falsa global, de falsos negativos y positivos fue del 11.6%, 2.7% y 4.5%, respectivamente. Kiyan y Yildirim (2003) compararon redes neuronales de base radial (RBF), de regresión generalizada (GRNN) y probabilísticas (PNN) sobre la base de datos de cáncer de mama de Wisconsin, encontrando que las redes neuronales GRNN ayudan a aumentar la exactitud y la objetividad en el diagnóstico de cáncer de mama.

Schneider y cols (2002) incrementaron significativamente la sensibilidad de un panel de marcadores y aumentaron la eficacia diagnóstica en cáncer de pulmón, aplicando lógica difusa. Lancashire y cols (2005) entrenaron una RNA a partir de la identificación de perfiles proteómicos de biomarcadores, con el objeto de identificar el estadio de desarrollo de melanoma (I-IV), de las muestras de los pacientes, donde obtuvieron un 98% de exactitud. Zhang y cols (2007) mediante una RNA combinaron 4 biomarcadores que individualmente no tenían suficiente sensibilidad para la detección de cáncer epitelial de ovario en etapa I. Analizaron los biomarcadores CA125II, CA72-4, CA15-3 y el factor estimulante de macrófagos (M-CSF), los cuales fueron utilizados como entradas en la RNA. La red se entrenó con 127 mujeres sanas, 101 mujeres con condiciones benignas y 90 pacientes con cáncer de ovario epitelial invasivo. Realizaron una prueba para determinar el desempeño de la red con muestras desconocidas (98 mujeres sanas y 52 pacientes con cáncer epitelial de ovario en etapas tempranas). La red neuronal artificial mostró una sensibilidad del 71%, a una especificidad del 98%, logrando incrementar un 25% la sensibilidad del mejor biomarcador, CA125II, que individualmente presentó una sensibilidad del 46%.

Karabatak e Ince (2008) propusieron un sistema que consta de reglas de asociación junto con una RNA, para detectar cáncer de mama, empleando como entrada de la red: el espesor, uniformidad, tamaño, forma, adhesión, núcleos desnudos, nucléolos y mitosis de las células, obteniendo una clasificación correcta del 95.6%. Por otra parte, Fan Zhang y cols. (2009) diagnosticaron cáncer de mama basado en perfiles proteómicos, utilizando RNAs.

Las herramientas de la Inteligencia Artificial han sido utilizadas en el campo biomédico para la detección y diagnóstico de cáncer. Este campo del conocimiento ofrece aún muchas oportunidades de estudio y desarrollo, sobre todo en la interpretación de la información cuando se evalúan grandes grupos de biomarcadores. En el presente trabajo se plantea la aplicación de herramientas de la Inteligencia Artificial, con el apoyo de herramientas estadísticas, para correlacionar la concentración de biomarcadores con el cáncer de pulmón, con el objetivo final de generar una herramienta auxiliar en el diagnóstico de ésta enfermedad.

1.2 PLANTEAMIENTO DEL PROBLEMA

El cáncer de pulmón es una enfermedad heterogénea difícil de diagnosticar oportunamente. Más del 75% de los diagnósticos se realiza en etapas avanzadas de la enfermedad, donde la opción terapéutica es limitada y, sobreviviendo 5 años sólo el 14% de los pacientes, y únicamente el 8%, 10 años en el mejor de los casos (López y Cols. 2005). El diagnóstico del cáncer de pulmón frecuentemente se realiza de forma tardía, debido a que esta enfermedad es confundida comúnmente con diversas afectaciones respiratorias. La mayor oportunidad terapéutica para el tratamiento quirúrgico está en los pacientes diagnosticados en etapas tempranas. Una herramienta potencial para el diagnóstico del cáncer de pulmón está dada por las proteínas biomarcadoras. Desafortunadamente, las pruebas basadas en biomarcadores individuales hasta ahora no han tenido la sensibilidad y especificidad deseada para el diagnóstico. Por esta razón, nuevas investigaciones se están enfocando al análisis combinado de múltiples marcadores. No obstante, la complejidad en la interpretación de los datos ha frenado el uso de este tipo de metodología de diagnóstico. En ocasiones, las herramientas estadísticas multifactoriales no son lo suficientemente robustas para extraer la información de las complejas relaciones e incertidumbres de los proceso biológicos, contrastando con las RNAs, las cuales han demostrado ser una herramienta auxiliar prometedora para la extracción de información incierta en medicina.

1.3 HIPÓTESIS

La utilización de herramientas de la inteligencia artificial, como las redes neuronales artificiales, permitirá incrementar la sensibilidad de las proteínas biomarcadoras séricas, mediante su análisis combinado, potenciando así su capacidad en el diagnóstico de la enfermedad.

1.4 OBJETIVOS

1.4.1 Objetivo general.

Diseñar una herramienta auxiliar para el diagnóstico de cáncer de pulmón mediante la cuantificación de proteínas biomarcadoras, utilizando redes neuronales artificiales.

1.4.2 Objetivos específicos

1.- Elaborar una base de datos con la información clínica más relevante de los pacientes que participen en el estudio.

2.- Determinar la concentración de las proteínas de interés por el método de ELISA.

3.- Entrenar y validar una RNA, correlacionando la concentración de las proteínas biomarcadoras séricas con el diagnóstico de la enfermedad.

4.- Comparar el funcionamiento de la RNA contra algunos métodos estadísticos.

En el próximo capítulo se presentan los conceptos fundamentales de las RNAs.

Capítulo II

FUNDAMENTOS DE REDES NEURONALES ARTIFICIALES

2.1 Introducción

La toma de decisiones es un punto clave en la práctica médica, tanto en el proceso diagnóstico como en el terapéutico. En cualquier situación, estas decisiones deben estar avaladas por criterios de evidencia y experiencia. Para ello, es preciso recabar y procesar adecuadamente la información referente al estado de salud del paciente (historial, exploración o pruebas diagnósticas, etc.) y confrontarla con la evidencia acumulada en el seguimiento de grupos amplios de pacientes en condiciones controladas. La información disponible se concreta en distintos tipos de variables que reflejan la condición del paciente. A partir de estas variables, es necesario aplicar criterios objetivos que permitan extraer una conclusión adecuada acerca de la posible evolución de la enfermedad, la posibilidad de complicaciones, etc. Considerando la complejidad del problema, la evaluación de la información debe hacerse desde una perspectiva multivariable, de manera que se consideren simultáneamente todos los factores implicados y se obtenga una generalización adecuada que permita una clasificación apropiada de nuevos casos (Trujillano y Cols. 2004).

Las técnicas estadísticas multivariables proporcionan una solución a este tipo de problemas. Así, el análisis discriminante puede utilizarse en la obtención de un criterio diagnóstico a partir de los valores de varias variables, mientras que el análisis de supervivencia permite evaluar convenientemente la contribución de diversas variables a la supervivencia en distintas circunstancias de interés médico. Todas estas técnicas han tenido un auge muy importante en su aplicación en medicina. Este auge se justifica, en parte, por su fácil disponibilidad al estar incluidas en casi todos los paquetes estadísticos de uso habitual. Sin embargo, como sucede con cualquier técnica estadística, su utilización debe tener en cuenta las condiciones apropiadas de aplicación, que en general se referirán a la distribución de las variables con las que se trabaja, la independencia entre ellas, etc.

Debido a la dependencia entre las variables consideradas o a efectos no-lineales no incluidos en el modelo, los resultados de la aplicación de la técnica estadística pueden estar alejados de la realidad. En estas situaciones, es posible utilizar modelos más

elaborados que incluyan interacciones entre las variables y efectos no lineales. Sin embargo, cuando el problema contiene un elevado número de variables predictoras, su complejidad implica que se conviertan en un problema difícil de abordar y resolver mediante las técnicas habituales. En este caso, una posible alternativa al empleo de este tipo de análisis basados en técnicas estadísticas, más o menos clásicas, la encontramos en metodologías propias de otros campos científicos, como puede ser la Inteligencia Artificial. En particular, las redes neuronales artificiales son capaces de desarrollar un modelo de predicción que incorpora automáticamente relaciones entre las variables analizadas sin necesidad de incorporarlas explícitamente en el modelo.

2.2 Redes Neuronales Artificiales

Una Red Neuronal Artificial (RNA) es un algoritmo de cálculo que se basa en una analogía del Sistema Nervioso. La idea general consiste en emular la capacidad de aprendizaje del Sistema Nervioso, de manera que la RNA aprenda a identificar un patrón de asociación entre los valores de un conjunto de variables predictoras (entradas) y los estados que se consideran dependientes de dichos valores (salidas) (Trujillano y Cols. 2004). Desde un punto de vista técnico, la RNA consiste en un grupo de unidades de proceso (neuronas) que se asemejan a las neuronas biológicas y que están interconectadas por medio de un entramado de relaciones (pesos) análogas al concepto de conexiones sinápticas en el sistema nervioso. A partir de los nodos de entrada, la señal progresa a través de la red hasta proporcionar una respuesta en forma de nivel de activación de los nodos de salida (Cross y Cols. 1995). Los valores de salida proporcionan una predicción del resultado en función de las variables de entrada. Las neuronas de las RNA funcionan de manera similar a las neuronas biológicas: cuando la suma de señales de entrada es suficientemente alta (en el caso de una neurona diríamos que se acumula suficiente neurotransmisor), la neurona envía una señal a las neuronas con las que mantiene contacto (se genera un potencial de acción). Esta situación se modela matemáticamente como una suma de pesos de todas las señales de llegada al nodo que se compara con un umbral característico. Si el umbral es superado, entonces el nodo se dispara, enviando una señal a otros nodos, que a su vez procesarán esa información junto con la que reciben de nodos adyacentes. Evidentemente, la respuesta de cada nodo dependerá del valor de las interacciones con los nodos precedentes dentro de la estructura de la red. Como en el caso del sistema nervioso, el poder computacional de

una RNA deriva no de la complejidad de cada unidad de proceso sino de la densidad y complejidad de sus interconexiones (Sarle, 1997).

En la Figura 2 se aprecia una RNA de una sola neurona. Las entradas de la red (x_i) son multiplicadas por el peso que tiene la conexión (w_i), donde cada entrada tendrá un peso específico. La suma de las entradas y los pesos respectivos permiten que un conjunto de datos den una salida específica (θ), la que se pasa por $f(\theta)$. La salida de la RNA es evaluada por funciones de activación $f(\theta)$ tipo: escalón, lineales, sigmoidales, tangenciales y gaussianas (ver Fig. 3). El elemento de tendencia desplaza el umbral de decisión de la RNA. La expresión matemática que describe a una RNA está dada por la Ecuación 1.

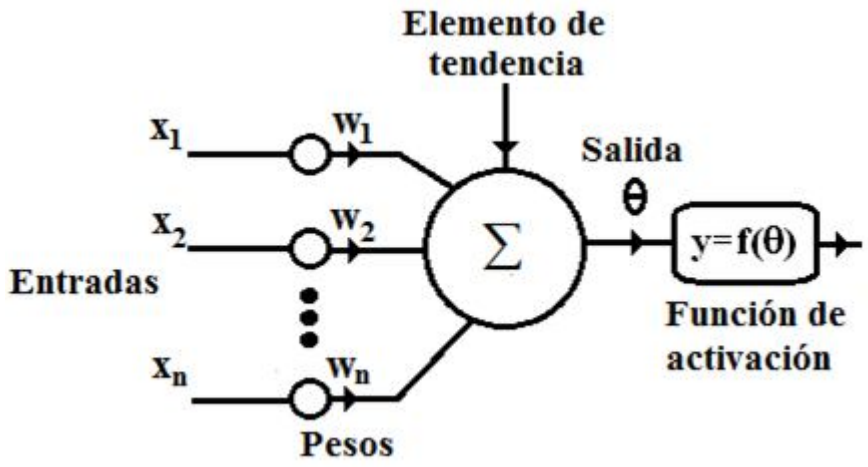

Figura 2. Representación de una Red Neuronal Artificial de una sola neurona.

$$\theta = x_1 w_1 + x_2 w_2 + \text{L}\, x_n w_n = \sum_{i=1}^{n} x_i w_i$$

$$y = f(\theta) \qquad (1)$$

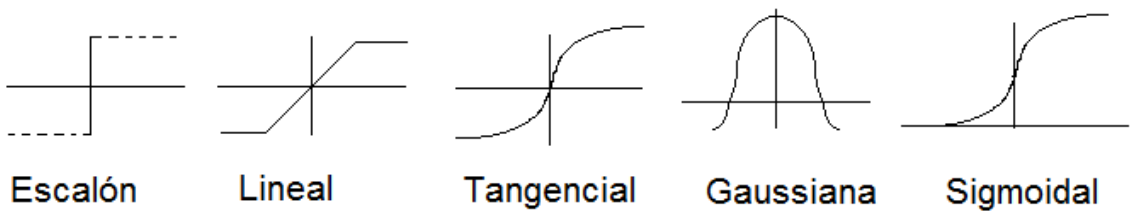

Figura 3. Funciones de activación.

2.3 Tipos de entrenamientos de las RNAs

El entrenamiento de una RNA consiste en la presentación repetida de un conjunto suficientemente amplio de datos, formado por las entradas y los valores correspondientes de las variables a predecir, hasta conseguir que los pesos internos (interacciones entre nodos) conduzcan a resultados óptimos de salida, acercándose lo más posible a los resultados esperados. Existen tres tipos de entrenamientos:

2.3.1 Supervisado

En este método se le muestra a la red qué salida se desea, con base en un conjunto de datos de entrada. Los pesos de las conexiones entre la red se modifican hasta que la diferencia entre la salida de la red y la salida deseada sea mínima.

2.3.2 No supervisado

En este tipo de entrenamiento la red no necesita que se le indique qué salida se desea. Estas redes clasifican los datos de entrada de acuerdo con patrones que ellas mismas desarrollan.

2.3.3 Reforzado

Es un proceso de entrenamiento más lento, ya que a la red se le indica si la salida que generó es correcta mediante una variable que puede tomar el valor de verdadero o falso.

En un contexto médico, el entrenamiento consistiría en presentar a la RNA, de forma iterativa, los valores de distintas variables clínicas (en forma de valores de la capa de entrada) de cada paciente y conseguir que la red sea capaz de predecir el estado final observado en cada paciente (indicados por el estado de las capas de salida de la red) de la manera más precisa posible. En la práctica, el ajuste de los pesos durante el entrenamiento se consigue mediante un proceso iterativo cuya finalidad es minimizar una

función de error que cuantifica la discrepancia entre las predicciones de la red y los valores observados en la muestra. La medida más utilizada para evaluar el error en la predicción es la raíz cuadrada del error cuadrático medio (RME) entre los valores de salida de la RNA y sus valores esperados según los datos disponibles.

2.3.4 RNAs con conexiones hacia adelante (feedforward).

Como su nombre indica, la información se mueve en una sola dirección, desde la entrada hacia la salida, a estas conexiones se las conoce como conexiones hacia adelante o feedforward. Estas redes están organizadas en capas y cada capa agrupa a un conjunto de neuronas que reciben la información de las neuronas de la capa anterior y emiten sus salidas hacia las neuronas de la capa siguiente. Cabe señalar que entre las neuronas de una misma capa no se presenta una conexión. En este tipo de redes debe existir al menos una capa de entrada que está formada por las neuronas que reciben los datos de entrada y una capa de salida, formada por una o más neuronas que emiten la salida de la red. El tiempo de procesamiento en estas redes es rápido por el hecho de que no existen conexiones entre las neuronas de una misma capa. En este tipo de redes están: la Perceptrón multicapa (MLP), las redes Adaline y Madaline, y la red Lineal Adaptativa (Widrow y Hoff, 1960).

2.3.5 Redes con retroalimentación total o parcial.

En este tipo de redes las neuronas pueden enviar estímulos a neuronas de capas anteriores, de su propia capa o a ellas mismas. A estas redes con conexiones hacia atrás se les conoce como feedback o interactiva. Cada neurona puede estar conectada a todas las demás; de este modo, cuando se recibe la información de entrada a la red, cada neurona tendrá que calcular y recalcular su estado varias veces, hasta que todas las neuronas de la red alcancen un estado estable. Un estado estable es aquel en el que no ocurren cambios en la salida de ninguna neurona. (Pearlmutter, 1990). Este tipo de redes son más lentas ya que no se sabe cuánto tiempo tomará en llegar a un estado estable. Entre estas RNAs se encuentran las redes Hopfield, LVQ, SOM, Cognitrón y el Neocognitrón, junto con las redes Boltzman y Cauchy.

2.4 Tipos de RNAs

2.4.1 Perceptrón

Es la red más sencilla que existe, sólo tiene una neurona (ver Figura 4) por lo que su capacidad de procesamiento está muy limitada. La función de activación utilizada es el escalón por lo tanto la salida sólo puede pertenecer a una clase, por ejemplo, en la Figura 4, en el problema de separar un patrón el perceptrón da un 1 si la salida pertenece a 'A' y un -1 si pertenece a 'B'. El perceptrón sólo puede resolver problemas sencillos que son linealmente separables.

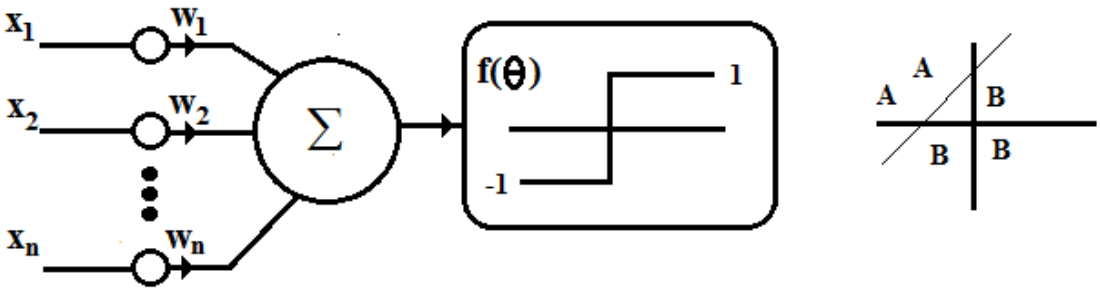

Figura 4. Representación de una Red perceptrón.

2.4.2 El perceptrón multicapa

2.4.2.1 El perceptrón multicapa como ejemplo de RNA de aplicación en medicina

Dentro de las redes supervisadas unidireccionales, la estructura más utilizada es la llamada perceptrón multicapa (MLP). La arquitectura típica de este tipo de red está constituida por varias capas de neuronas con interconexión completa entre ellas. El caso más sencillo en este tipo de red consiste en sólo 2 capas de neuronas, las de entrada y la de salida. De esta manera se puede obtener un modelo adecuado para problemas lineales del tipo de la regresión lineal múltiple. Si se quiere analizar problemas no lineales, es necesario incorporar otras capas de neuronas intermedias u ocultas (ver Figura 5).

Las capas del perceptrón multicapa comprenden:

2.4.2.1.1 Capa de entrada

En esta capa entran todas las variables que alimentan a la red, (X1, X2,...Xn). Esta capa no tiene efecto directo en la red, sólo es de entrada.

2.4.2.1.2 Capas ocultas

Estas capas son llamadas ocultas porque no tienen contacto directo con el exterior. El perceptrón multicapa tiene como mínimo una capa oculta. Generalmente la función de activación de esta capa es no lineal.

2.4.2.1.3 Capa de salida

En esta capa se da la salida que la red neuronal ha procesado. La función de activación en ocasiones es lineal o no lineal.

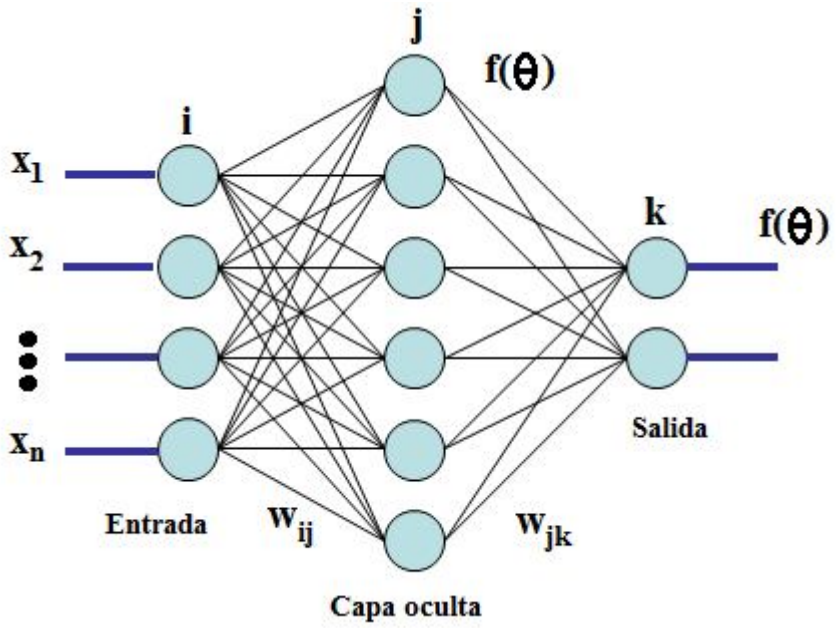

Figura 5. Representación las diferentes capas de una red MLP.

2.4.3 Entrenamiento de la RNA multicapa.

El perceptrón multicapa es una RNA que utiliza entrenamiento supervisado; éste debe ser realizado fuera de línea. Un algoritmo de entrenamiento para el perceptrón multicapa es

‘backpropagation’ (propagación hacia atrás). Para entender el algoritmo backpropagation es necesario primero entender cómo se entrena cualquier RNA. Los pasos a seguir para entrenar una RNA son:

1. Inicializar la RNA con pesos aleatorios.

2. Aplicar un vector de entrada a la red y calcular su respectiva salida.

3. Comparar la salida de la red con un vector deseado y determinar el error.

4. Calcular los nuevos pesos de la red.

5. Repetir los pasos del 2 al 4 hasta que el error cumpla una tolerancia establecida.

La forma más simple de calcular el error (paso 3) es por medio de la regla de mínimos cuadrados (LMS, least mean square, ver ecuación (2)), también llamada regla delta ya que trata de minimizar una delta o diferencia entre el valor deseado y el valor real.

$$\xi^2 = \frac{1}{2L} \sum_{k=1}^{L} \xi_k^2 \qquad (2)$$

L = Número de patrones de entrada

ξ = Diferencia entre la salida deseada y la real

Otra forma de representar la regla delta es por medio de la ecuación (3):

$$\mathrm{E_p} = \frac{1}{2} \sum_j \left(t_{pj} - o_{pj} \right)^2 \qquad (3)$$

Esta representación es la que se utiliza para calcular el error aplicado a una red, donde t_{pj} es el vector deseado y *Opj* es el vector de salida. En el aprendizaje se busca disminuir el error entre un vector de salida deseado y el vector de salida real. La regla delta permite reducir este error. Cuando ya no existe diferencia entre el vector deseado y el vector real, es porque la red ha aprendido. Freeman y cols. (1991), explican que el uso de la regla delta para el aprendizaje se vuelve bastante difícil y engorroso, por la manipulación de matrices, por lo que se desarrolló otro tipo de aprendizaje más eficiente, tal como se explica a continuación.

Existe una modificación de la regla delta en la que se usa descenso de gradiente. Esta modificación equivale a dejar que el algoritmo de entrenamiento encuentre por sí mismo el error mínimo. Un gradiente indica la pendiente más pronunciada hacia arriba, la dirección opuesta al gradiente indica hacia donde se encuentra el mínimo. El método de descenso de gradiente minimiza la diferencia de la suma de cuadrados entre el vector deseado y el vector real de salida. Rumelhart y cols. (1986) explican que la derivada del error medida con respecto a cada peso es proporcional al cambio de peso dictado por la regla delta, pero con una constante de proporcionalidad negativa, tal como se ve en la Ecuación (4), donde 'α' es igual a la razón de aprendizaje, 'ε' es el error y 'w' son los pesos. El descenso de gradiente equivale a realizar un descenso por pasos en un espacio de pesos. Es decir, que en cada iteración de entrenamiento el algoritmo desciende a pasos hasta llegar a cumplir con la tolerancia deseada.

$$\Delta W_j = -\alpha \left(\frac{\partial \xi_k}{\partial W_j} \right) \qquad (4)$$

El gradiente del error es:

$$\frac{\partial \xi_k^2}{\partial W_i} = -\xi_k x_{ki} \qquad (5)$$

X = entradas de la red.

Las modificaciones de los pesos son proporcionales al gradiente del error

$$\Delta W_i = - \alpha \left(- \xi_k x_{ki} \right) = \alpha \xi_k x_{ki} = \alpha \left(t - o \right) x_{ki} \qquad (6)$$

En una red que no tenga capas ocultas, la superficie de error tiene la forma que se presenta en la Figura 6. Sólo existe un mínimo en esta superficie, por lo que el descenso de gradiente puede encontrar fácilmente el mínimo.

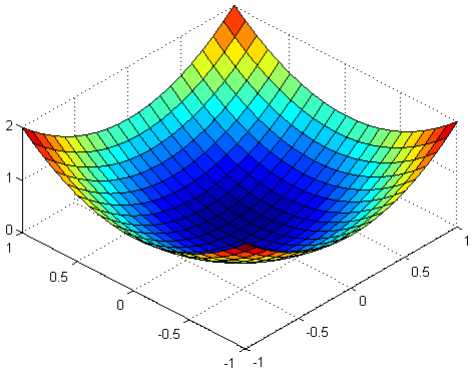

Figura 6. Superficie de error cuando no hay capas ocultas.

La razón de aprendizaje 'α', genera pequeños incrementos en los pesos para calcular el nuevo error, este valor está entre $0<\alpha<1$. α influye en la velocidad en que la red converge. Si la razón de aprendizaje es pequeña la red tendrá que hacer un gran número de iteraciones para encontrar el error requerido, por el contrario si la razón de aprendizaje es muy grande es posible que se esté oscilando alrededor del mínimo global y tal vez la red no encuentre el error especificado.

Cuando se tienen redes con capas ocultas es mucho más difícil encontrar el mínimo global, ya que la superficie de error puede estar llena de mínimos locales, (ver Figura 7) y existe la posibilidad de quedar atrapado en uno de ellos sin poder alcanzar convergencia. Rumelhart y Cols. (1986), también desarrollaron la regla delta generalizada para utilizarse con redes que tuvieran capas ocultas y que su función de activación fuera semilineal, es decir, que fuera una función diferenciable y no decreciente. La Ecuación (7) representa la

salida neta de la RNA y la Ecuación (8) describe la salida de la red pasada por una función de activación.

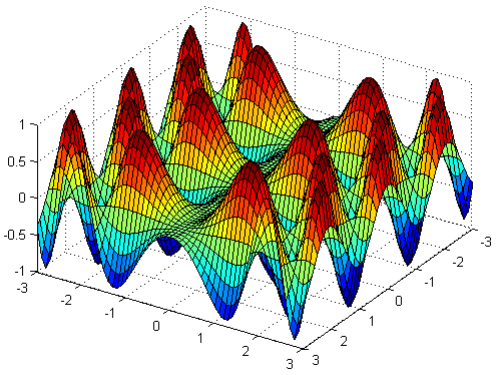

Figura 7. Superficie de error cuando existen capas ocultas.

$$\text{net}_j = \sum_{i=1}^{n} W_{ij} x_i \qquad (7)$$

$$o_j = f_i(\text{net}_j) \qquad (8)$$

W_{ij} = Vector de pesos.

X_i = Entradas de la red

O_j = Salida de la red.

Existen tres ecuaciones (9), (10) y (11) que describen cómo se realiza el aprendizaje en una red con capas ocultas. Estas tres ecuaciones describen la regla delta generalizada o el algoritmo de entrenamiento backpropagation. La Ecuación (9) calcula la salida de la red, a partir de la capa oculta a la capa de salida, recordemos que la capa de entrada no tiene efecto directo sobre las demás capas.

$$\Delta W_{ij} = \eta \delta_j x_i \qquad (9)$$

η = Razón de aprendizaje

δ = (t - o), diferencia entre el vector deseado y el vector real.

X = Vector de entrada

La Ecuación (10) calcula los errores en la capa de salida.

$$\delta_j = (t_j - o_j) f'_j \left(net_j\right) \qquad (10)$$

f ' = Derivada de la función.

La Ecuación (11) calcula los errores para la capa oculta.

$$\delta_j = \sum_k \delta_k W_k f'(net_j) \qquad (11)$$

Las ecuaciones anteriores describen como funciona el algoritmo de entrenamiento backpropagation o propagación hacia atrás. Este algoritmo de entrenamiento primero permite que un vector de entrada pase por las capas de la red, el perceptrón multicapa es una red con conexiones hacia adelante, por lo que el vector de entrada debe recorrer primero la capa de entrada, después la capa oculta y finalmente la capa de salida. Cuando ya se tiene un valor de salida se calculan los nuevos pesos para la capa de salida y la capa oculta, con base en los errores obtenidos. El ajuste de pesos se hace primero en la capa de salida y posteriormente en la capa oculta, esta es la razón por la que este algoritmo recibe el nombre de propagación hacia atrás. Vale la pena mencionar que el flujo de señales de la red es hacia adelante, la señal no fluye hacia atrás, sólo en el algoritmo de entrenamiento hay propagación hacia atrás para ajustar los nuevos pesos.

En una red perceptrón multicapa no es posible calcular directamente el error y los pesos para la capa oculta, ya que esta capa no tiene vectores deseados como la capa de salida, por lo cual primero se debe calcular el error en la capa de salida y posteriormente se calcula el error para la capa oculta. Se puede consultar el artículo sobre backpropagation

escrito por Williams y Cols. (1997), donde explican el entrenamiento con backpropagation de una forma muy simple de entender. Para un análisis más profundo se puede revisar el artículo escrito por Rumelhart y Cols.

Los pasos que se siguen para el entrenamiento en backpropagation son:

- Inicializar los pesos de las conexiones aleatoriamente.
- Dar el vector de entrada a la red, y calcular la salida, (Ecuación 9).
- Calcular el error para la capa de salida, (Ecuación 10).
- Calcular el error para la capa oculta, (Ecuación 11).
- Con base en el error se calculan los nuevos pesos, comenzando por la capa de salida, y luego los pesos de la capa oculta.
- El proceso se repite hasta minimizar el error.

Si la RNA queda atrapada en un mínimo que no permita satisfacer la tolerancia preestablecida, es decir, no llega al error predestinado, se puede volver a entrenar. Con esto se logra que los pesos de inicialización sean distintos y por lo tanto las condiciones iniciales sean diferentes. Otra posible solución es aumentar el número de neuronas en la capa oculta o aumentar el número de capas. Existen otras posibilidades como agregar momento al entrenamiento. El momento consiste en agregar una pequeña fracción del cambio anterior a los nuevos pesos, esto permite mantener la dirección de los pesos en una sola dirección, ver Ecuación (12).

$$\Delta W_{ij}(n+1) = \eta\left(\delta_j o_i\right) + \alpha\,\Delta W_{ij}(n) \qquad (12)$$

Existe una modificación del algoritmo backpropagation, conocido como algoritmo de entrenamiento Levenberg-Marquardt. Este algoritmo al igual que backpropagation buscan encontrar un mínimo preestablecido, pero utilizando diferentes métodos. En general Levenberg-Marquardt es más rápido que los algoritmos anteriores, pero tiene como inconveniente que consume grandes cantidades de memoria. Dado que este algoritmo es poco usado existen pocas referencias sobre su funcionamiento. Una fuente sobre Levenberg-Marquardt se puede encontrar en el manual de redes neurales de MATLAB™

(Neural Networks Toolbox). La ecuación utilizada en MATLAB™ para el ajuste de pesos se puede ver en (13).

$$\Delta W = \left(J^T J + \mu I \right)^{-1} J^T e \qquad (13)$$

ΔW = Incremento en los pesos.

J = Determinante Jacobiano de la derivada de cada error respecto a los pesos.

J^T = Tranpuesta de J.

e = Vector de error.

μ = Valor de una constante que se incrementa o decrementa.

I = Valor de una matriz identidad.

El algoritmo de entrenamiento funciona de la siguiente manera: si el valor de μ es muy grande la expresión (13) se aproxima mucho al método descenso de gradiente, mientras que si μ es pequeño la expresión (13) se convierte en un método Gauss-Newton para encontrar mínimos. Si el error se hace más pequeño, μ se hace más grande, por el contrario si el error se incrementa, μ decrece. Si se quiere obtener más información sobre el algoritmo Levenberg-Marquardt enfocado a la obtención de mínimos, se puede consultar 'Recetas numéricas en C', en el capítulo de modelos no lineales (Press y cols. (1992).

2.5 RNAs utilizadas en este estudio

2.5.1 RNA Feedforward

Este tipo de red se explicó la sección 2.3.4

2.5.2 RNA Pattern Recognition (PR)

También es una red feedforward entrenada con el algoritmo backpropagation, tiene las mismas características que la feedforward descrita anteriormente en cuanto a la función de entrenamiento y de aprendizaje, sólo que esta red emplea una función de salida de transferencia de capa de tansig y presenta funciones graficas (generación de curvas ROC (Receiver Operating Characteristics/Características de operación del receptor) y tabla de porcentaje de confusión) adicionales incluidas en las propiedades de la misma red.

2.5.3 RNA Probabilística

La RNA probabilística pueden ser utilizada para problemas de clasificación. Cuando una entrada es presentada, la primera capa calcula las distancias desde el vector de entrada a los vectores de entrada entrenados y produce un vector cuyos elementos indican que tan cerca está la entrada respecto a la entrada entrenada. La segunda capa suma estas contribuciones para cada clase de entrada y produce con ellas un vector de salida con las probabilidades. Finalmente, una función de transferencia en la salida de la segunda capa del tipo *"compete"* obtiene la máxima de las probabilidades y produce un 1 para esa clase y cero para las demás clases. Así la red clasifica el vector de entradas dentro de una clase específica de las *n* clases. Es una red de solo dos capas donde sólo la primera tiene bias.

2.5.4 RNA Learning Vector Quantization (LVQ)

Esta red es usada para problemas de clasificación. Es una red con aprendizaje supervisado. Los vectores de entrada se dividen en un número de regiones distintas. Cuando se presenta un nuevo vector de entrada, se determina a qué región pertenece (Haykin, 1994). La arquitectura de la red LVQ presenta dos capas con n neuronas de entrada y m de salida. Cada una de las n neuronas de entrada se conecta a las m de salida hacia delante (feedforward). Entre las neuronas de la capa de salida, existen conexiones laterales de inhibición (peso negativo). El valor que se asigne a los pesos de las conexiones feedforward entre las capas de entrada y salida durante el proceso de aprendizaje de la red, va a depender de esta interacción lateral. El aprendizaje es de tipo competitivo. Las neuronas de la salida compiten por activarse y sólo una de ellas permanece activa ante una determinada información de entrada a la red. Esta neurona se denomina prevalente o vencedora, y en función de ella se ajustan los pesos de las

conexiones (Hilera y Martinez, 1995). En esta red ninguna capa presenta bias. El algoritmo de aprendizaje para esta red es learnlv2.

2.6 Características de las RNAs

Existe una gran variedad de RNAs. Cada red tiene características que la hacen diferente de otro tipo de red. En la Tabla 2 se puede observar una clasificación de algunas de las RNAs más utilizadas (Hagan y Cols, 1996). Se pueden observar diferencias entre los distintos tipos de redes, por ejemplo, un perceptrón tiene sólo dos capas, mientras que el perceptrón multicapa puede tener 'n' capas.

Tabla 2. Clasificación de las redes neuronales artificiales.

Tipo de Red neuronal	Topología	Supervisado/No supervisado	Aprendizaje en línea/Fuera de línea	Tiempo de entrenami-ento	Tiempo de ejecución
Perceptrón	2 Capas	Supervisado	Fuera de línea	Regular	Rápido
Adaline	2 Capas	Supervisado	Fuera de línea	Regular	Rápido
Perceptrón Multicapa	n Capas	Supervisado	Fuera de línea	Lento	Rápido
Hopfield	1 Capa	No supervisado	En línea	Rápido	Regular
Kohonen	2 Capas	No supervisado	En línea	Regular	Rápido
Bam	2 Capas	No supervisado	En línea	Rápido	Rápido
Boltzman	1 o 3 Capas	Supervisado	Fuera de línea	Lento	Lento
Base radial	3 Capas	Supervisado	Fuera de línea	Rápido	Rápido

2.7 Requerimientos para un diseño adecuado de RNAs

2.7.1 Validación cruzada

Para elaborar una red que sea eficaz es conveniente dividir los datos en 3 conjuntos, cuidando que cada uno de ellos mantenga la representatividad de la población origen: a) el conjunto de entrenamiento, b) el conjunto de validación, y c) un conjunto de prueba.

El conjunto de entrenamiento se usa para ajustar los pesos durante la fase de entrenamiento, mientras que el conjunto de validación se utiliza para decidir cuándo parar el proceso de entrenamiento. Como criterio general, el entrenamiento debe detenerse cuando el error del conjunto de validación sea mínimo. De esta manera, nos aseguramos que la red es capaz de predecir correctamente los resultados de un conjunto de datos que no forman parte de los ejemplos de entrenamiento. Esta técnica se denomina validación

cruzada. Si continuamos el entrenamiento más allá de este punto, la red empieza a aprender de memoria los datos del conjunto de entrenamiento pero pierde capacidad de generalización. La búsqueda de una generalización óptima, que es la capacidad de la red de proporcionar una respuesta correcta ante patrones que no han sido empleados en su entrenamiento, requiere que se cumplan tres condiciones: a) que la información recogida en las variables sea suficiente, es decir, una selección apropiada de las variables y una buena calidad en la recolección de datos; b) que la función que aprenda la red sea suave, es decir que pequeños cambios en las variables de entrada produzcan pequeños cambios en las variables de salida y c) que el tamaño en la base de datos sea suficiente. De esta manera aseguramos que el conjunto de entrenamiento sea representativo de la población a estudiar. Excepto la segunda condición, el resto de los requisitos son comunes a cualquier técnica multivariante que se emplee. Una vez finalizado el entrenamiento, la RNA entrenada evalúa el conjunto de prueba y produce las correspondientes predicciones con datos que no se han utilizado en el entrenamiento, ni en la validación cruzada. Esta prueba final aporta un resultado independiente acerca de la capacidad de generalización de la red.

2.7.2 Tamaño y arquitectura de la red

La arquitectura de una red viene determinada por el número de capas y nodos que la forman. La complejidad de la red viene determinada por el número de interconexiones que contiene. En general, no es inmediato establecer de forma exacta cuál será la arquitectura ideal para cada aplicación. Así, problemas de discriminación lineal o de regresión logística pueden solucionarse con redes simples. Los problemas surgen al enfrentarse a modelos más complicados. En aplicaciones médicas, un MLP con una única capa oculta puede ser adecuado en muchos casos. Existen algoritmos evolutivos que determinan, de forma automática, esta arquitectura óptima al aumentar o retirar nodos o capas del modelo. En cualquier caso, la arquitectura óptima debe alcanzarse en la práctica, mediante un proceso iterativo, validando la capacidad predictiva de las distintas arquitecturas consideradas. Por otra parte, cuanto más compleja sea una red, mayor número de parámetros o pesos deberemos estimar y, por lo tanto, necesitará mayor número de patrones para ser entrenada de manera adecuada.

En el otro extremo, la utilización de una red demasiado compleja para solucionar un problema sencillo nos conduce a un sobreajuste (overfitting) que dificulta la capacidad de generalización de la red. Como regla general, para reducir el número de parámetros de

una red es conveniente seleccionar apropiadamente las variables de entrada, descartando variables poco informativas. Sin embargo, esta selección no es tan sencilla como en los métodos multivariantes habituales y puede requerir distintas etapas de entrenamiento. Desde un punto de vista más técnico, existen procedimientos propios de la metodología de redes que simplifican la estructura de la red. A modo de ejemplo, algunos de estos métodos consisten en compartir pesos entre varios nodos (weight sharing), realizar un podado de la red (pruning) eliminando los pesos con menor influencia en el resultado del modelo final o aplicar el método de decaimiento de pesos (weight decay), eliminando automáticamente los pesos que tienden a cero.

2.8 Correspondencia entre RNAs y técnicas estadísticas

En el ámbito de la medicina, la utilización de las RNAs se ha desarrollado paralelamente a las técnicas estadísticas. Dependiendo del problema específico estudiado, esta confrontación ha llevado, durante esta última década, a alternar entre el optimismo (Eldar y Cols. 2002) y el pesimismo (Borque y Cols. 2001) en la utilización de las redes en el entorno de los estudios médicos.

La Tabla 3 muestra la correspondencia entre algunas redes y distintos procedimientos estadísticos habitualmente utilizados en medicina. Es interesante apreciar que existen algunos tipos de redes que no poseen una correspondencia concreta con un método estadístico (Sarle 1994). La comparación más frecuentemente analizada en la literatura se realiza entre el tipo de red más empleada (MLP + backpropagation) y la regresión logística múltiple.

Tabla 3. Relación entre redes neuronales artificiales y técnicas estadísticas.

Red neuronal artificial	**Técnica estadística**
Perceptrón simple (nodo umbral)	Análisis discriminante
Perceptrón simple (nodo sigmoideo)	Regresión logística
Adalina	Regresión lineal
Perceptrón multicapa	Regresión no lineal
Aprendizaje no supervisado	Análisis de componentes principales
Mapas de Kohonen	Escalas multidimensionales
LVQ (cuantificación de vectores)	Análisis discriminante
Función de base radial (RBF)	Método de regresión de Kernel

Algunas de las ventajas que presentan las RNAs, es que necesitan menos formalismo estadístico para su desarrollo, logran detectar relaciones no lineales, detectan interacciones entre variables predictoras y tienen múltiples algoritmos de entrenamiento, aunque por otro lado presentan algunas desventajas como que son cajas negras para identificar las interacciones, tienen cierta dificultad de utilización, necesitan mayores recursos computacionales, emplean una metodología poco conocida y su implementación está llena de procesos empíricos. En una revisión que realizó Sargent de 28 aplicaciones distintas, concluye que las RNAs son, en el peor de los casos, equivalentes o en general ligeramente superiores a la regresión logística múltiple al no tener que depender de exigencias rígidas de independencia de las variables o de los supuestos inherentes al modelo logístico (Sargent, 2001).

En una aplicación concreta, las redes pueden interpretar de manera distinta la información contenida en las variables respecto a cómo se interpreta esta información en un procedimiento estadístico. Esto nos obliga a analizar cuidadosamente la distinta contribución de cada variable al modelo final y a interpretar sus interdependencias. A partir de este análisis es posible mejorar los modelos estadísticos (por ejemplo, añadiendo interacciones encontradas entre las variables). De este modo, ambas técnicas pueden colaborar para proporcionar un modelo final adecuado al problema objeto de estudio.

Como principal desventaja, debemos mencionar que proporciona un modelo que es esencialmente una caja negra. La RNA es capaz de predecir resultados, pero no disponemos de una interpretación evidente de los parámetros en los mismos términos en que podemos interpretar los resultados de una técnica estadística.

2.9 Indicaciones prácticas acerca del desarrollo de una RNA

El desarrollo de una RNA necesita planificarse adecuadamente para conseguir una red convenientemente entrenada que alcance una precisión óptima (Schwarzer y Cols 2000). Para facilitar una aproximación práctica al uso de esta metodología, expondremos una serie de pasos para la creación de una RNA.

2.9.1 Paso 1: Una base de datos adecuada

El resultado obtenido con una RNA depende de los datos que se utilizan para su entrenamiento y, por lo tanto, los sesgos de muestreo pueden influir negativamente el resultado.

2.9.2 Paso 2: Conjuntos de entrenamiento, verificación y validación

La partición de la serie de datos en los conjuntos de desarrollo (entrenamiento y validación) y el correspondiente conjunto de prueba, determina que el tamaño muestral sea suficientemente grande. En aquellos casos en que no sea así, se ha propuesto la utilización de técnicas de remuestreo con lo que se consigue tener múltiples conjuntos de entrenamiento que aseguran, dentro de las posibilidades y limitaciones de estas técnicas, un proceso de entrenamiento adecuado que conducirá a una buena generalización (Tourassi, 1997).

2.9.3 Paso 3: Construcción y entrenamiento de la red

Las características de las RNAs determinan que su utilización requiera de programas informáticos adecuados. En este punto, existen múltiples opciones comerciales y de libre distribución. Usando estos programas se comienza a probar qué arquitectura es más conveniente (este proceso es básicamente empírico).

2.9.4 Paso 4: Prueba de la red

Debe comprobarse la capacidad de generalización de la red enfrentándola a datos distintos de los utilizados en su desarrollo (conjunto de entrenamiento y validación).

2.9.5 Paso 5: Evaluación de los resultados (precisión de la red)

En los problemas donde la predicción es una probabilidad, lo indicado es evaluar la discriminación y la calibración de la red (Van, 2000). Para valorar la discriminación (capacidad de distinguir entre dos estados) se pueden emplear tablas de contingencia eligiendo punto de corte, estableciendo porcentajes de correcta clasificación ó analizando las curvas ROC resultantes (especialmente calculando el área bajo la curva ROC) (Burgueño y Cols. 1995).

En resumen las RNAs proporcionan un método general para desarrollar modelos de predicción en medicina. La ventaja principal de esta técnica, si se aplica convenientemente, radica en su capacidad para incorporar interacciones entre las variables sin necesidad de incluirlas a *priori*. Además, su aplicación no queda restringida a un tipo determinado de distribución de los datos. Con ellas, se dispone de herramientas que se fundamentan en el cálculo intensivo y que desafían a los planteamientos

estadísticos convencionales. Estas técnicas proporcionan nuevos puntos de vista que pueden ayudar a obtener herramientas más eficaces en muchas aplicaciones prácticas.

En este capítulo se hizo una introducción sobre algunos aspectos básicos de las RNAs. Se habló sobre el perceptrón, y como se puede mejorar su funcionamiento con el perceptrón multicapa. Se comentó sobre la regla delta generalizada, que es usada para entrenar redes con capas ocultas, y se mostraron las diferencias en la superficie de error entre redes que tienen y no tienen capas ocultas.

Capítulo III

METODOLOGÍA

En este capítulo se presenta la metodología utilizada durante el desarrollo de la tesis. La metodología se divide en cuatro secciones:

1) Procedimiento utilizado para seleccionar los pacientes con cáncer.
2) Técnicas analíticas usadas durante la tesis.
3) Aspectos relacionados con la red neuronal artificial.
4) Técnicas estadísticas utilizadas.

3.1 Procedimiento utilizado para seleccionar los pacientes con cáncer

3.1.1 Diseño del estudio

Se realizó un estudio exploratorio, observacional, abierto, transversal, analítico y prospectivo de casos y controles. Para el estudio únicamente se incluyeron a pacientes mexicanos (de ascendencia mexicana), diagnosticados por primera vez con cáncer de pulmón primario, confirmado por histopatología, vírgenes de tratamiento, cualquier tipo histológico y etapa de avance de la enfermedad, cualquier edad y ambos sexos.

No se incluyeron en el estudio a aquellos pacientes que en el momento de la toma de muestra cursaban por procesos infecciosos activos, que presentaran cáncer de pulmón secundario o que recibieran tratamientos previos contra la enfermedad. Se excluyeron del estudio aquellos pacientes que presentaron un diagnóstico dudoso o no documentado, con resultados de laboratorio o historia clínica incompleta, así como pacientes con muestras de sangre insuficiente, muestras de mala calidad o cuando hubo pérdida de la muestra.

3.1.2 Recolección de muestras

Las muestras se obtuvieron de pacientes que ingresaron a la Unidad Médica de Alta Especialidad (UMAE) del Centro Médico Nacional de Occidente (CMNO), ubicada en el Municipio de Guadalajara, Jalisco y del Centro Oncológico Estatal (COE) del Instituto de Seguridad Social del Estado de México y Municipios (ISSEMyM), en la ciudad de Toluca, México, y que fueron diagnosticados con cáncer de pulmón. Se obtuvo la aprobación del comité de ética correspondiente y todos los participantes dieron su consentimiento

informado por escrito. Las muestras y la información de salud fueron marcadas con códigos para proteger la confidencialidad de los participantes.

El estudio constó de dos grupos: el primero (casos) conformado por 63 pacientes con cáncer de pulmón, se comparó con un segundo grupo (controles) que constó de 29 fumadores activos (que no tuvieran o presentaran antecedentes de enfermedades pulmonares). Los pacientes fueron reclutados en el CMNO, el Hospital Civil de Guadalajara y el ISSEMyM. En este grupo control también se incluyeron a 58 pacientes con Enfermedad Pulmonar Obstructiva Crónica (EPOC), los cuales representaron la condición benigna de inflamación crónica y procedieron del Hospital Civil de Guadalajara (HCG), “Fray Antonio Alcalde”.

El diagnóstico histopatológico para los pacientes con cáncer de pulmón primario fue establecido de acuerdo a la clasificación de tumores de pulmón de la Organización Mundial de la Salud y la Asociación Internacional para el Estudio de Cáncer de Pulmón. Los pacientes con EPOC, fueron diagnosticados de acuerdo con la Iniciativa Global para la Enfermedad Pulmonar Obstructiva Crónica, que define a la EPOC por una relación FEV1/FVC post broncodilatador <0.70. Tanto para los pacientes con cáncer de pulmón como para los pacientes con EPOC, se obtuvieron los historiales clínicos, los exámenes físicos, los valores espirométricos (únicamente para pacientes con EPOC), estudios histopatológicos (únicamente para pacientes con cáncer de pulmón), los valores de gas en sangre arterial, electrocardiogramas y radiografías de tórax. Las muestras e información clínica así como los diferentes exámenes de cada paciente fueron etiquetados con identificadores únicos para proteger la confidencialidad de los pacientes.

3.1.3 Obtención de sueros

De cada participante en el estudio se recolectaron entre 8 y 10 ml de sangre periférica, obtenida de la vena cubital, utilizando tubos separadores de suero (BD Vacutainer®, Franklin Lakes NJ USA). Dichas muestras se procesaron inmediatamente sometiéndolas a centrifugación a una velocidad de 3000 rpm a temperatura ambiente durante 10 minutos. El suero separado fue alicuotado y almacenado a -70°C para su posterior análisis.

3.2 Descripción de técnicas analíticas

3.2.1 Cuantificación de la concentración de biomarcadores séricos

La concentración de las proteínas Apolipoproteína A-I, α1-Antitripsina, Activador de plasminógeno tipo urocinasa, Antígeno carcinoembrionario, Antígeno de cáncer 125, Enolasa específica de neuronas, Fragmentos de citoqueratina 19, Haptoglobina, Metaloproteinasa de matriz 1, Metaloproteinasa de matriz 9, Proteína C reactiva, Proteína de unión a retinol, Transferrina y YKL40 fue determinada en las muestras de suero, utilizando kits de ELISA comerciales y de acuerdo con el protocolo de cada kit: MMP-1 y MMP-9 (R&D Systems; Minneapolis, MN), CA 125, CEA y NSE (ALPCO Diagnostics; Salem, NH), uPA, HPT, TF, AT, CRP, APOAI and RBP (ASSAYPRO; St. Charles, MO), Cyfra 21-1 (DRG Instruments GmbH; Germany), y YKL-40 (Quidel Corporation; San Diego, CA). La absorbancia especificada por los protocolos de los kits fue medida en un espectrofotómetro de microplaca (Bio-Rad Laboratories; Philadelphia, PA). Para más detalle ver apéndice A y B donde se presentan las principales características de cada biomarcador.

3.2.2 Principio general de la técnica de ELISA

El ensayo inmunoabsorbente ligado a enzimas (ELISA: Enzyme-Linked Immunosorbent Assay) se basa en la detección de un antígeno, teniendo previamente adherido un anticuerpo a una superficie. Tras la formación del complejo antígeno-anticuerpo se agrega una enzima, la cual se adhiere al complejo, y ésta al reconocer a su sustrato produce un color observable que puede cuantificarse mediante un espectrofotómetro (Moreno, 2008). Existen dos tipos de ELISA, el directo y el indirecto; el primero ayuda a detectar la presencia de antígeno, mientras que el segundo es empleado para la detección de anticuerpos. Se emplearon ELISAS directos para cuantificar la concentración de antígeno, utilizando dos variantes, ELISA competitivo y no competitivo.

Un ELISA no competitivo hace reaccionar la muestra que contiene el antígeno, con el anticuerpo que está previamente adherido a una superficie; posteriormente dicha reacción se enfrenta a un segundo anticuerpo unido a una enzima. Una vez terminada la reacción se añade el sustrato de la enzima, desarrollándose un color que es cuantificable y es proporcional a la concentración de la muestra. En un ELISA competitivo, el antígeno de la muestra va a competir con el conjugado (antígeno biotinilado), por un número limitado de sitios del anticuerpo, previamente adherido a la superficie. Al añadir el sustrato, esté será

reconocido por la enzima y se desarrollará color. En este tipo de ensayo habrá ausencia de color en una muestra positiva debido a que el sustrato no encontrará a la enzima porque el conjugado ha sido desplazado por el antígeno presente en la muestra. A manera de ejemplo se presentan a continuación dos protocolos: el primero de ellos empleado para determinar la concentración del biomarcador HPT mediante un kit de ELISA competitivo y el segundo para determinar la concentración del biomarcador MMP-9 con un ELISA no competitivo.

3.2.3 Desarrollo de un ELISA competitivo

1. Añadir 25 μl del estándar y/o 25 μl de muestra a cada pozo pretratado con un anticuerpo policlonal específico contra HPT, e inmediatamente añadir 25 μl de la HPT biotinilada a cada pozo. Mantener en agitación suave e incubar durante 60 min.

2. Lavar 5 veces cada pozo con 200 μl de buffer de lavado.

3. Añadir 50 μl del conjugado estreptavidina-peroxidasa a cada pozo e incubar durante 30 min.

4. Lavar como se indica en el paso número 2.

5. Añadir 50 μl de cromógeno-sustrato (tetrametilbenzidina-peróxido de hidrógeno) a cada pozo e incubar durante 10 min en agitación suave o hasta que se desarrolle la densidad óptima del color azul.

6. Añadir 50 μl de la solución de paro (ácido clorhídrico 0.5N) a cada pozo. El color cambiará de azul a amarillo.

7. Leer la densidad óptica a una longitud de onda de 450nm inmediatamente después de agregar la solución de paro.

3.2.4 Desarrollo de un ELISA no competitivo

1. Añadir 100 μl del diluyente RD1-34 (buffer a base de proteína con agentes conservadores) a cada pozo pre-tratado con un anticuerpo monoclonal específico contra MMP9.

2. Añadir 100 μl de estándar y/o 100 ml de muestra a cada pozo e incubar durante 2 horas en agitación constante a 500±50rpm.

3 Lavar 5 veces cada pozo con 200 μl de buffer de lavado.

4 Añadir 200 μl del conjugado enzimático de MMP-9 (anticuerpo policlonal contra MMP9 marcado con peroxidasa) a cada pozo e incubar durante 1 hora con agitación constante.

5. Lavar como se indica en el paso número 3.

6. Añadir 200 μl de solución cromógeno sustrato (tetrametilbenzidina-peróxido de hidrógeno) a cada pozo e incubar durante 30 min protegiendo de la luz.

7. Añadir 50 μl de la solución de paro (ácido sulfúrico 2N) a cada pozo. El color cambiará de azul a amarillo.

8. Leer la densidad óptica a una longitud de onda de 450nm inmediatamente después de agregar la solución de paro.

3.2.5 Cálculo de la concentración

Los resultados fueron convertidos a partir de la media de la absorbancia de los duplicados de cada muestra, después de restar los valores de absorbancia del pozo con concentración 0.
Las proteínas recombinantes de origen humano MMP1, MMP9, CA125, CEA, NSE, Cyfra 21-1, YKL40, uPA, HPT, TF, CRP, APOAI, RBP o AT fueron usadas como estándares para cada prueba. La curva estándar se preparó simultáneamente con la medición de las muestras. Para la determinación de la concentración se utilizó el programa MasterPlex ReaderFit 101^{TM} que es capaz de generar una regresión logística de cinco parámetros. En el caso de las muestras que se diluyeron para su análisis, se consideró el factor de dilución para realizar el cálculo de la concentración.

3.3 Entrenamiento de la Red Neuronal Artificial

La RNA fue entrenada con las concentraciones de los biomarcadores seleccionados para el estudio. Los biomarcadores fueron utilizados como entradas de la red. Las dos neuronas de salida de la red corresponden con: 1) presencia de cáncer y 2) sin presencia de cáncer (grupo control), (ver Figura 8). El software utilizado para el diseño de la red fue MATLAB™ y se utilizó el toolbox de redes neuronales.

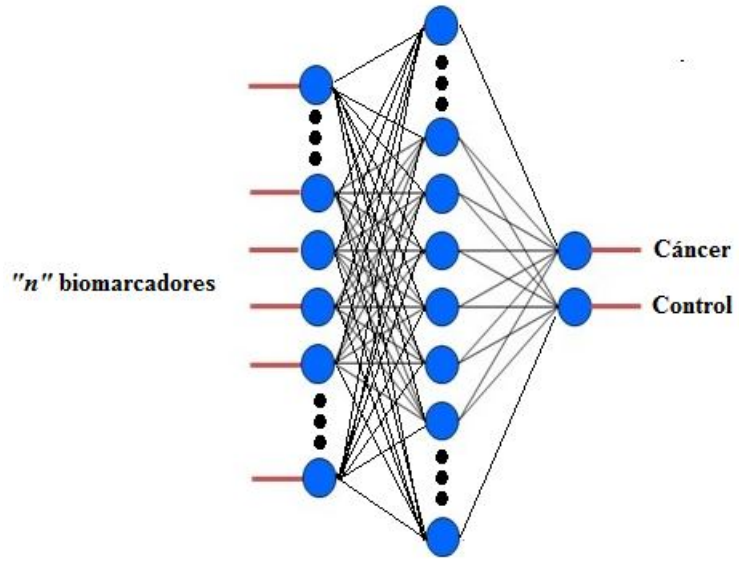

Figura 8. RNA utilizada para clasificar pacientes con cáncer y controles.

3.3.1 Pre-procesamiento de datos

Los valores de concentración de los biomarcadores fueron pre-procesados numéricamente antes de ser utilizados como entradas para el entrenamiento de las RNAs. Se normalizaron los valores con respecto a una media igual a 0 y con una desviación estándar de 1, con la función “mapstd”. Posteriormente los datos normalizados fueron aleatorizados antes de ser utilizados como entradas para la RNA. El conjunto de datos se dividió utilizando la función "divideind” de la siguiente manera: 60% para la fase de entrenamiento, 20% para la fase de validación y el 20% restante para la fase de prueba de la red (los datos empleados para la prueba fueron muestras con diagnóstico desconocido y que no habían sido utilizados en el entrenamiento o validación de la RNA). Para cada entrenamiento se utilizaron los mismos vectores. Para evitar el sobre entrenamiento se empleó una validación cruzada igual a 10 veces.

3.3.2 Tipos de RNAs entrenadas

Para determinar cuál RNA era la apropiada para el conjunto de datos de concentración de biomarcadores, se probaron cuatro tipos de RNAs: 1) Feedforward (FF), 2) Learning

Vector Quantization (LVQ), 3) Pattern Recognition (PR) y 4) una red probabilística. Se probaron diferentes configuraciones, arquitectura, funciones de activación de las redes escogidas. La RNA que presentó la mejor capacidad para clasificar correctamente el mayor número posible de datos, fue escogida para utilizarse en el estudio. En este caso fue la de Pattern Recognition.

3.3.3 Arquitectura de la RNA

Se describe la arquitectura de la RNA Pattern Recognition. Para la capa oculta de la RNA se utilizó una función de activación tangencial y para la capa de salida una función de activación lineal. Se probaron diferentes configuraciones de neuronas en la capa oculta, 4, 6, 8 y 10. También se probaron dos capas ocultas con el siguiente número de neuronas 2-6, 6-3 y 5-5. La arquitectura óptima de la RNA se eligió con base en el error mínimo cuadrático de entrenamiento.

3.3.4 Algoritmos de entrenamiento

Se probaron los siguientes algoritmos de entrenamiento contenidos en el Toolbox de redes neuronales: Levenberg-Marquardt (LM), BFGS Quasi-Newton (BFG), Resilient Backpropagation (RP), Scaled Conjugate Gradient (SCG), Conjugate Gradient with Powell/Beale Restarts (CGB), Fletcher-Powell Conjugate Gradient (CGF), Polak-Ribiére Conjugate Gradient (CGP), One Step Secant (OSS) y Variable Learning Rate Backpropagation (GDX). EL algoritmo más adecuado se eligió de acuerdo al error mínimo cuadrático de entrenamiento.

3.3.5 Características de la RNA

La función empleada para la activación de las neuronas en la capa oculta fue "tansig", mientras que para las neuronas pertenecientes a la capa de salida fue "purelin", en un rango de 0 a 1. Los valores de los pesos y del bias de la RNA fueron determinados por la función de aprendizaje "learngdm". El criterio que se usó para detener la fase de entrenamiento en cada RNA fue cuando el error medio de la raíz cuadrada fue inferior a 0.09 o cuando la tasa correcta de clasificación fuera igual o mayor al 90%. Los valores de los biomarcadores estuvieron directamente involucrados en la modificación de los pesos de conexión en el modelo de la RNA durante el entrenamiento.

3.4 Comparación de la capacidad de detección de la RNA contra métodos estadísticos

La sensibilidad de la RNA con la mejor combinación de biomarcadores, se comparó con el mejor biomarcador por medio de las curvas ROC, a una especificidad con valor clínico relevante. Un valor $p<0,05$ fue considerado significativo.

3.4.1 Análisis estadístico univariado

3.4.1.1 Cumplimiento de supuestos estadísticos

Se sometieron los datos de concentración de cada biomarcador a diferentes pruebas para ver si cumplían con los supuestos estadísticos para poder ser analizados mediante estadística inferencial. Se realizó la prueba de Shapiro-Wilks para determinar la normalidad de los datos. Para determinar la homogeneidad de varianzas se uso la prueba de Bartlett. Los datos con un valor de $p<0.05$ no cumplen con el supuesto de normalidad o de homogeneidad de varianzas.

3.4.1.2 Comparación de grupos

Se analizó si existían diferencias estadísticas significativas de concentración de los biomarcadores, entre sexo, edad, índice tabáquico y tipo de exposición; así mismo se realizó el análisis estadístico del grupo de cáncer de pulmón con respecto al grupo control para determinar qué proteínas son significativamente diferentes, con respecto a su concentración. Este análisis se realizó mediante la prueba U de Mann-Whitney, que es la versión no paramétrica de la habitual prueba t de Student. Un valor de $p<0.05$ fue considerado significativo. El análisis estadístico se realizó con el software SigmaStat 8.0®.

3.4.1.3 Curvas ROC (Receiver Operating Characteristics)

Para describir y comparar el valor diagnóstico de cada biomarcador, se desarrollaron curvas ROC. Estas curvas correlacionan la proporción de verdaderos positivos y falsos positivos en una prueba diagnóstica. El área bajo la curva indica la probabilidad de clasificar correctamente un par de individuos. El punto de corte y la sensibilidad fueron calculados a una especificidad del 80% en la curva ROC de cada biomarcador. El software empleado para desarrollar las curvas ROC fue SigmaPlot™ 10.0.

3.4.2 Análisis estadístico multivariado

Se realizaron análisis estadísticos multivariados para comparar el funcionamiento de la RNA.

3.4.2.1 Análisis de Componentes Principales

Los biomarcadores que presentaron diferencia estadística entre los grupos de estudio, fueron considerados para realizar un análisis de componentes principales (PCA). Con esta prueba se determinó qué biomarcadores tienen mayor influencia para describir al grupo de cáncer de pulmón y al grupo control. El PCA se realizó con el software Simca-P+ 12.0^{TM}. Los mejores biomarcadores fueron utilizados para el entrenamiento de las RNAs.

3.4.2.2 Análisis Discriminante

Se generó un análisis discriminante con el programa Statgraphics plus 5.1^{TM} con el fin de determinar cuántas muestras es posible separar en el grupo de cáncer de pulmón con respecto al grupo control. Se realizó el análisis discriminante con selección hacia atrás, el cual elige de acuerdo a su algoritmo qué variables tienen mayor importancia para separar los grupos de estudio. Un valor $p<0.05$ se consideró significativo en las funciones discriminantes.

3.4.2.3 Árbol de Clasificación y Regresión

Se realizó un árbol de clasificación y regresión (CART). Este método se basa en la segmentación binaria de datos. Se dividieron los datos en dos partes, la primera consistió de 49 pacientes con cáncer de pulmón y 71 controles para la creación del árbol. El criterio que se empleó para detener la clasificación fue cuando el error alcanzó a un mínimo del 7%. Posteriormente se realizó una fase de prueba con 14 pacientes de cáncer de pulmón y 16 sujetos control con el objeto de verificar el potencial de clasificación cuando se le presentan nuevas muestras. En cada división los datos son divididos en dos grupos mutuamente excluyentes. Se utilizó el programa STATISTICA 8.0^{TM} para realizar el CART, usando el algoritmo Gini splitting con una validación cruzada igual a 10, lo cual favoreció que en cada división no permitiera dividir los nodos con cinco o menos observaciones.

3.4.2.4 Combinación de biomarcadores mediante curvas ROC

Se combinaron los valores de concentración de cada biomarcador con respecto a su media muestral y a su área bajo la curva generada por su misma curva ROC. Se realizaron todas las combinaciones posibles para los diferentes biomarcadores mediante la siguiente ecuación:

Concentración Combinada=Biomarcador X_1 + β_1 Biomarcador X_2+ β_2 Biomarcador X_3 + β_n Biomarcador X_n

Donde: $\beta_1 = \left[\frac{Biomarcador\ \overline{X_1}}{Biomarcador\ \overline{X_2}}\right]\left[\frac{ABC\ Biomarcador\ X_2}{ABC\ Biomarcador\ X_1}\right]$

$\beta_2 = \left[\frac{Biomarcador\ \overline{X_1}}{Biomarcador\ \overline{X_3}}\right]\left[\frac{ABC\ Biomarcador\ X_3}{ABC\ Biomarcador\ X_1}\right]$

$\beta_n = \left[\frac{Biomarcador\ \overline{X_1}}{Biomarcador\ \overline{X_n}}\right]\left[\frac{ABC\ Biomarcador\ X_n}{ABC\ Biomarcador\ X_1}\right]$

En este capítulo se ha descrito la metodología seguida para la selección de los pacientes, técnicas analíticas utilizadas, el entrenamiento de la RNA y algunos métodos estadísticos. En el próximo capítulo se describirán y discutirán los resultados de este estudio. El apéndice C describe con mayor detalle algunos de los métodos estadísticos utilizados.

Capítulo IV

RESULTADOS Y DISCUSIÓN

4.1 Perfil demográfico de los participantes en el estudio

Se incluyeron dos grupos en el estudio, el primero estuvo conformado por 63 pacientes consecutivos (37 hombres, 26 mujeres) con diagnóstico reciente de cáncer de pulmón, el cual se comparó con un segundo grupo (control), conformado por 58 pacientes (40 hombres, 18 mujeres) con enfermedad pulmonar obstructiva crónica (EPOC), que representa la condición benigna de inflamación crónica, y 29 fumadores activos sin antecedentes de enfermedades pulmonares (17 hombres, 12 mujeres). Se obtuvo la aprobación del comité de ética correspondiente y todos los participantes dieron su consentimiento informado por escrito. En la Figura 9 se presenta la proporción de muestras que se tiene en los diferentes grupos de estudio. Las muestras y la información de salud fueron marcadas con códigos para proteger la confidencialidad de los participantes.

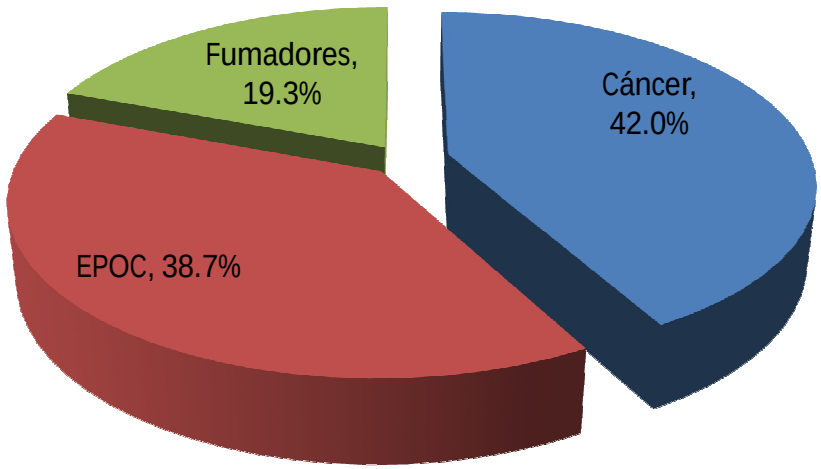

Figura 9. Proporción de muestras en los grupos de estudio.

En la Tabla 4 se resume la información clínica de los participantes en este estudio, en la cual se puede resaltar que los grupos son homogéneos en cuanto a edad ya que no presentan una diferencia significativa entre ellos. Con respecto a la frecuencia del sexo, para el grupo de cáncer de pulmón se tiene mayor porcentaje en el sexo masculino con un 58.7%, mientras que para el sexo femenino es del 41.3%, esta relación es similar en el grupo control, con un 64.3% y 34.7%, respectivamente.

Tabla 4. Perfiles demográficos y clínicos de pacientes y controles.

Demográfico	Controles (n=86)	Cáncer de pulmón (n=63)	p-value
Edad (años)	61 ± 15.2	64 ± 13.9	0.267
Rango (edad)	19-92	31-89	-
Femenino	30	26	-
Masculino	57	37	-
Fumador activo	65	50	-
Fumador pasivo	0	2	-
Cocinó con leña	2	6	-
Tabaco/Leña	9	2	
Causa desconocida		3	
Índice tabáquico	30 ± 27.6	36 ± 30.4	0.412
NSCLC	-	50	-
SCLC	-	9	-
N.D.	-	4	-
III	-	8	-
IV	-	23	-

N.D. = No disponible

Ambos grupos de estudio tienen antecedentes de haber fumado y cocinado con leña. Se destaca el consumo de tabaco, debido a que éste es el principal factor de riesgo para desarrollar cáncer de pulmón. El índice tabáquico (paquetes/año) nos indica la intensidad con la que fuma el individuo. En la Tabla 4 se puede ver que el grupo de cáncer de pulmón y el grupo control son homogéneos en esta variable, al no presentar una diferencia significativa.

Se sabe que el tipo histológico más frecuente de cáncer de pulmón es NSCLC. En la Figura 10 se puede notar que en este estudio el tipo histológico NSCLC es el más frecuente con un 79.4%, mientras que el tipo histológico de SCLC fue de sólo un 20.6%.

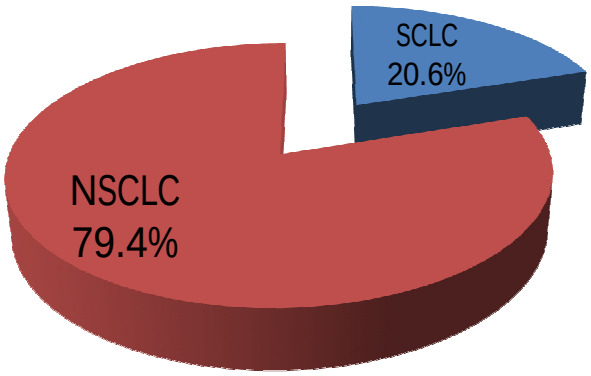

Figura 10. Proporción de tipos histológicos en el grupo de cáncer de pulmón.

En el estudio se eligieron 14 diferentes proteínas (ver Tabla 5) que han sido reportadas en diferentes trabajos de investigación donde se han asociado al cáncer de pulmón en distintas poblaciones del mundo. Sin embargo no hay un estudio de estas proteínas en población mexicana y estas proteínas hasta ahora no habían sido evaluadas en conjunto.

Tabla 5. Lista de proteínas biomarcadoras propuestas para el diagnóstico de cáncer de pulmón.

Apolipoproteina A-I
α1-Antitripsina
Activador de plasminógeno tipo urocinasa
Antígeno carcinoembrionario
Antígeno de cáncer 125
Enolasa específica de neuronas
Fragmentos de citoqueratina 19
Haptoglobina
Metaloproteinasa de matriz 1
Metaloproteinasa de matriz 9
Proteína C reactiva
Proteína de unión a retinol
Transferrina
YKL40

4.2 Capacidad diagnóstica de biomarcadores individuales

La medición de la concentración de los biomarcadores MMP1, MMP-9, uPA, TF, AT, HPT, CA125, CEA, Cyfra 21-1, NSE, APOAI, RBP, CRP y YKL40, en muestras de suero de ambos grupos de estudio se realizó utilizando kits de ELISA disponibles comercialmente y de acuerdo a las instrucciones de los fabricantes.

La prueba estadística U de Mann-Whitney permitió evaluar diferencias significativas en la concentración de biomarcadores de los pacientes de cáncer de pulmón con respecto al grupo control. Trece de las catorce proteínas evaluadas presentaron diferencias estadísticamente significativas con un *p*-value <0.05. El grupo de cáncer de pulmón mostró tener una mayor concentración en diez proteínas; MMP1, MMP9, AT, HPT, CA125, CEA, Cyfra 21-1, NSE, CRP y YKL40 sobre el grupo control (ver Tabla 6), mientras que las proteínas TF, APOAI y RBP se encontraron disminuidas en el grupo de cáncer de pulmón con respecto al grupo control. La proteína uPA mostró ser estadísticamente igual en ambos grupos de estudio.

Tabla 6. Diferencias estadísticas entre grupos de estudio.

	Control		**Cáncer de pulmón**		
Proteína	**Mediana**	**Cuartil (25-75%)**	**Mediana**	**Cuartil (25-75%)**	**p-value**
MMP1	17.48	6.56-23.61	26.48*	18.18-34.1	<0.0001
MMP9	551.81	369.46-747.24	900.38*	470.02-1299.27	<0.0001
uPA	1.49	0.7-4.71	1.8	0.33-5.12	0.7159
TF	5123.76*	1639.0-11371.4	1585	817.2-6307.86	<0.0001
AT	1211.61	621.19-1691.84	2091.36*	913.06-2571.92	<0.0001
HPT	2362.71	1230.2-3392	4722.2*	2887.2-7581.4	<0.0001
CA125	4.71	1.5-11-43	25.31*	14.82-72.5	0.0000
CEA	1.52	0.71-2.23	2.82*	1.13-18.93	0.0002
Cyfra 21-1	0.79	0.79-1.12	4.2*	1.28-9.48	0.0000
NSE	10.36	5.84-19.59	12.69*	8.71-23.63	0.0202
APOAI	8671.1*	3450.06-11865.2	6254.24	1847.53-10395.8	0.0293
RBP	36.26*	25.67-314.48	244.32	21.34-299.2	0.0371
CRP	7150	2813.3-16532.8	24833.9*	19350-32000.00	<0.0001
YKL40	171.88	110.21-326.23	353.79*	238.51-591.74	<0.0001

*Grupo que mostró diferencias significativas p-value < 0.05.

Se realizó un análisis estadístico para cada biomarcador con respecto a edad, sexo, índice tabáquico y tipo de exposición. Las catorce proteínas evaluadas en cuanto a sexo e

índice tabáquico son estadísticamente iguales, mientras que únicamente el biomarcador YKL40 se encontró en mayor concentración cuando los pacientes estuvieron expuestos tanto al humo de leña como al de tabaco (ver Tabla 7).

Como se puede ver en la Tabla 8 la única proteína que presentó diferencia significativa fue CA125, presentándose en mayor concentración en los pacientes con una edad menor a 50 años. Sin embargo como se puede notar prácticamente no existen diferencias significativas, lo cual demuestra que las concentraciones de los biomarcadores en las diferentes edades, género, índice tabáquico y tipo de exposición son homogéneas.

Tabla 7. Relación entre información clínica y patológica (tipo de exposición e índice tabáquico) con la concentración de proteínas circulantes en pacientes con cáncer de pulmón.

	Tipo de exposición				Índice tabáquico		
Proteína	**Tabaco**	**Leña**	**T/L**	**p-value**	**< 30**	**> 30**	**p-value**
MMP-1	29.76	16.77	19.58	0.988	18.78	26.41	0.356
MMP-9	887.47	948.02	1477.57	0.387	1017.11	638.88	0.479
uPA	1.62	4.18	0.00	0.875	0.61	0.91	0.252
TF	1376.10	4813.23	2536.80	0.852	2250.80	3876.40	0.626
AT	2147.72	1342.28	1887.43	0.604	1716.40	1891.12	0.891
HPT	5061.04	2882.30	3221.54	0.988	4668.70	3176.80	0.356
CA-125	24.72	69.86	89.51	0.277	24.18	24.72	0.427
CEA	3.25	9.69	1.91	0.233	2.25	3.56	0.270
CYFRA 21-1	3.73	9.46	9.02	0.798	1.44	4.45	0.148
NSE	13.87	10.01	17.35	0.490	21.09	14.23	0.399
APO A-I	6988.77	1595.71	6995.68	0.398	2638.18	3080.63	0.924
RBP	262.36	19.40	288.08	0.578	19.93	25.67	0.770
CRP	24845.89	23314.85	26580.27	0.759	23446.43	22405.80	0.197
YKL-40	345.19	358.32	870.76*	0.042	359.23	318.91	0.318

*Grupo que mostró diferencias significativas p-value < 0.05.
T/L = Tabaco/Leña

Tabla 8. Relación entre información clínica y patológica (edad y sexo) con la concentración de proteínas circulantes en pacientes con cáncer de pulmón.

	Edad				Sexo		
Proteína	**30-50**	**50-70**	**70-90**	**p-value**	**Masculino**	**Femenino**	**p-value**
MMP-1	20.11	29.92	23.47	0.779	24.53	30.23	0.296
MMP-9	1299.27	812.10	819.49	0.320	853.65	919.57	0.838
uPA	1.55	2.00	1.00	0.521	1.86	1.65	0.565
TF	2536.80	1090.80	1703.50	0.961	1943.50	1380.20	0.469
AT	1740.96	2091.36	2184.36	0.729	1964.32	2160.88	0.980
HPT	4722.20	3579.90	5061.04	0.779	3721.00	6117.80	0.296
CA-125	56.38*	21.32	25.18	0.024	35.22	21.69	0.214
CEA	2.82	1.50	5.86	0.090	5.25	1.65	0.793
CYFRA 21-1	1.64	7.49	2.91	0.312	4.68	2.45	0.759
NSE	17.35	13.12	12.56	0.662	13.91	12.64	0.381
APO A-I	8604.14	5437.96	4706.46	0.635	5521.66	6733.24	0.792
RBP	288.08	28.05	28.14	0.376	25.67	261.28	0.745
CRP	26269.28	24313.62	24391.37	0.922	21221.64	25479.84	0.090
YKL-40	317.51	278.44	362.95	0.674	312.86	361.69	0.077

*Grupo que mostró diferencias significativas p-value < 0.05.

4.3 Sensibilidad y especificidad de los biomarcadores individuales (curvas ROC)

Se determinó la capacidad de diagnóstico para cada biomarcador mediante curvas ROC (Receiver Operating Characteristics). En la Figura 11 se pueden observar las curvas ROC generadas para las proteínas CRP, Cyfra 21-1 y CA-125 que fueron las que presentaron un área bajo la curva mayor a 0.8 y mostraron diferencia significativa en el grupo de cáncer de pulmón con respecto al grupo control. Las áreas bajo la curva (ABC) para CRP, Cyfra 21-1 y CA-125 fueron de 0.83, 0.85 y 0.85 respectivamente.

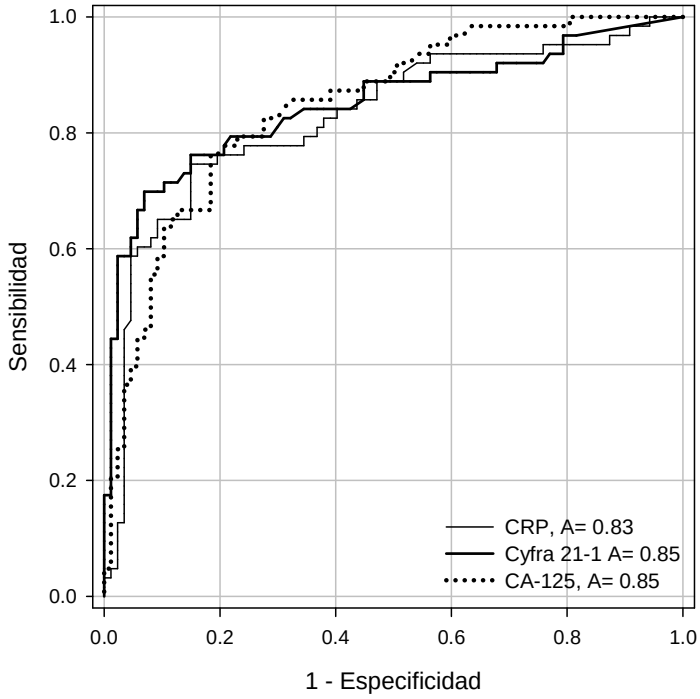

Figura 11. Curvas ROC para proteínas que presentaron una diferencia estadística significativa para el grupo de cáncer de pulmón con respecto al grupo control con un ABC > 0.8.

Las proteínas NSE, CEA, YKL-40, MMP-9, AT, MMP-1 y HPT presentaron un ABC de 0.61, 0.68, 0.72, 0.72, 0.74, 0.74 y 0.77 respectivamente (ver Figura12). Las proteínas TF, RBP y APOAI, que estadísticamente se encuentran en menor concentración en el grupo con cáncer de pulmón con respecto al grupo control, presentaron un ABC ≤ 0.5 (ver Figura 13).

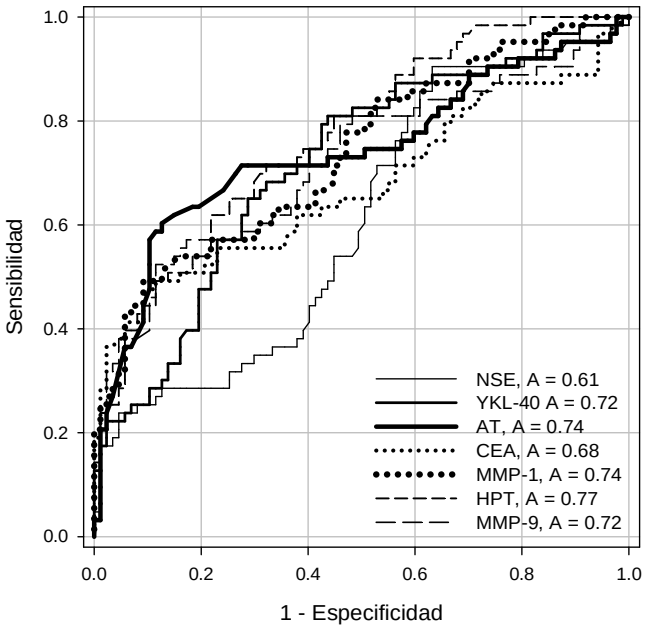

Figura 12. Curvas ROC para proteínas que presentaron una diferencia estadística significativa para el grupo de cáncer de pulmón con respecto al grupo control con un ABC entre 0.6-0.8.

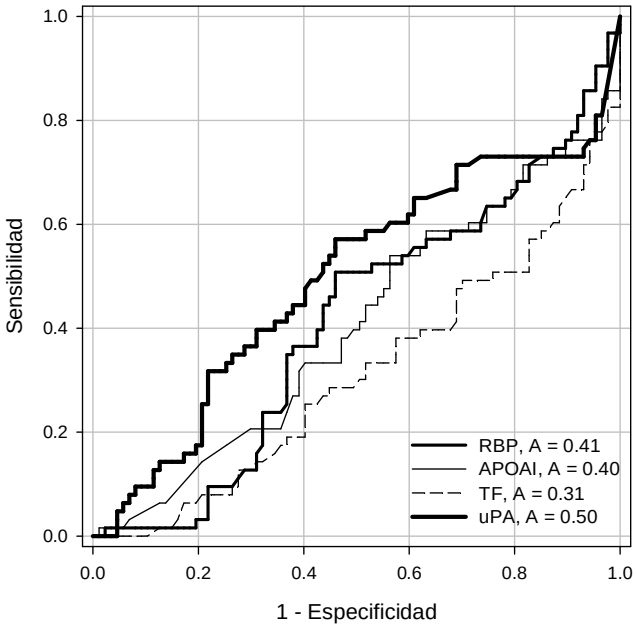

Figura 13. Curvas ROC para proteínas que se presentaron en menor concentración en el grupo de cáncer de pulmón con respecto al grupo control, o que no presentaron diferencia estadística significativa. ABC ≤ 0.5.

Para calcular la sensibilidad de cada uno de los biomarcadores se estableció una especificidad fija del 80% y se calculó la sensibilidad. Se encontró que las proteínas que tienen mayor sensibilidad son: Cyfra 21-1, CRP y CA-125 con un 76.19%, seguidos por la AT con un 63.49%, MMP9 un 57.14%, MMP1 y HPT un 53.97%, seguidos por CEA con un 50.79%, y finalmente YKL40, NSE, uPA, APOAI, TF y RBP con una sensibilidad del 47.62%, 28.57%, 17.46%, 14.29%, 6.35% y 3.18% respectivamente (ver Figura 14).

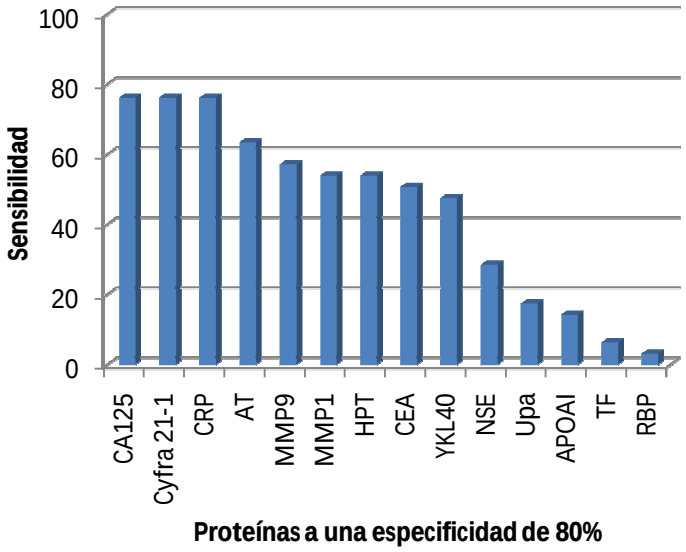

Figura 14. Sensibilidad de cada biomarcador evaluado a un 80% de especificidad.

Los resultados obtenidos en las curvas ROC concuerdan con los resultados del análisis estadístico realizado con la prueba no paramétrica U de Mann-Whitney. Las mismas 10 proteínas que presentan diferencias significativas para el grupo de cáncer de pulmón con respecto al grupo control presentan un área bajo la curva comprendido entre 0.7 y 0.9, con lo cual de acuerdo a la escala propuesta por Swets son útiles para algunos propósitos de diagnóstico clínico (Swets, 1988).

De acuerdo a este análisis la proteína que tiene mayor sensibilidad es Cyfra 21-1, con un valor de 76.19%, a una especificidad del 80% y con un ABC = 0.85. Una vez que se determinó el valor diagnóstico para cada proteína individual, se procedió a trabajarlas en conjunto empleando como herramientas las redes neuronales artificiales.

4.4 Resultados de la Red Neuronal Artificial

4.4.1 Entrenamiento de la RNA con las 14 proteínas biomarcadoras

Se probaron cuatro diferentes tipos de RNAs: Feedforward (FF), Probabilística, Learning Vector Quantization (LVQ) y Pattern Recognition (PR). La RNA que mostró tener un mejor desempeño en cuanto a la clasificación correcta de los participantes en el estudio fue la RNA PR con un 88.7%, seguida por la FF, la probabilística y la LVQ (ver Tabla 9).

Tabla 9. Porcentaje de clasificación para las distintas Redes Neuronales entrenadas.

Tipo de RNA	% clasificación
Feedforward (FF)	78.67 (118 de 150)
Probabilística	66.60 (100 de 150)
Learning Vector Quantization (LVQ)	57.33 (86 de 150)
Pattern Recognition (PR)	88.70 (133 de 150)

La red PR fue entrenada utilizando las catorce proteínas biomarcadoras: MMP-1, MMP-9, uPA, TF, AT, HPT, CA-125, CEA, Cyfra 21-1, NSE, APOA1, RBP, CRP y YKL-40 (ver Figura 15).

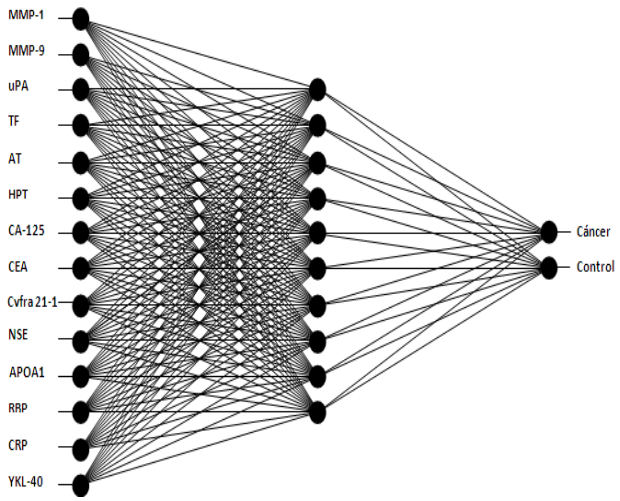

Figura 15. Arquitectura de la RNA con 10 neuronas en la capa oculta entrenada con catorce biomarcadores.

El óptimo número de neuronas y capas de la RNA se determinó con base en el error mínimo cuadrático. Múltiples RNA's fueron entrenadas con distintos números de neuronas y capas ocultas. A medida que se incrementa el número de neuronas, el error de clasificación disminuye. Sin embargo, existe un número óptimo de neuronas donde después de ese valor el error crece al seguir aumentando el número de neuronas (ver Figura 16).

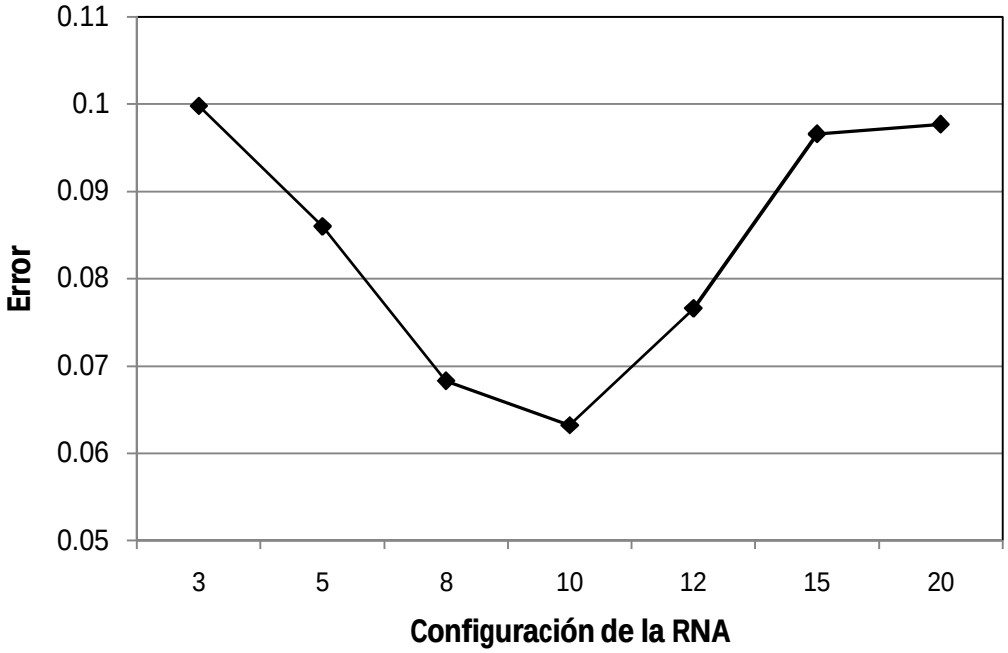

Figura 16. Determinación de la arquitectura para la RNA con una capa oculta entrenada con catorce proteínas biomarcadoras.

La mejor configuración para la RNA entrenada con 14 biomarcadores se obtuvo con 10 neuronas en la capa oculta (ver Figura 15) debido a que con esta configuración se obtuvo el error mínimo de 0.0632. Las configuraciones probadas con dos capas ocultas mostraron una tendencia constante en cuanto al desempeño de la RNA (ver Figura 17)

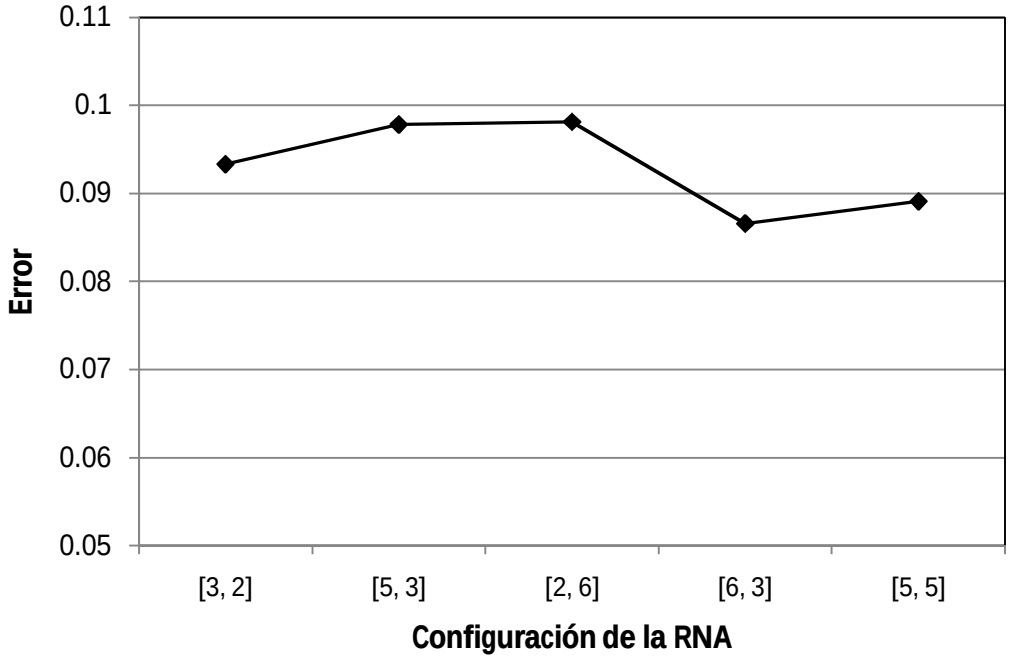

Figura 17. Determinación de la arquitectura para la RNA con doble capa oculta entrenada con catorce proteínas biomarcadoras.

Usando la RNA de PR con 10 neuronas en la capa oculta y entrenada con los 14 biomarcadores fue posible clasificar correctamente 133 de los 150 sujetos usados en el universo de pacientes (ver Figura 15). La tasa de clasificación está representada por un 92.2%, 96.6% y 76.7% en las fases de entrenamiento, validación y prueba respectivamente.

La red entrenada y probada fue capaz de extraer la información relevante de las 14 proteínas biomarcadoras para clasificar adecuadamente los controles y los pacientes con cáncer. No obstante, usar las 14 proteínas biomarcadoras para detectar cáncer de pulmón no es la mejor opción, principalmente por el alto costo de los kits para determinar las proteínas biomarcadoras y el tiempo que llevaría realizar dichas pruebas. Sería deseable y necesario utilizar un menor número de proteínas para realizar el diagnóstico de cáncer de pulmón con la RNA.

4.4.2 Reducción del número de proteínas biomarcadoras para entrenar la RNA.

Para reducir el número de proteínas se llevó a cabo un análisis de componentes principales (PCA por sus siglas en inglés) sobre el conjunto de biomarcadores que mostraron ser estadísticamente significativos en el grupo de cáncer de pulmón con respecto al grupo control. El PCA, por ser una técnica estadística multivariable y no paramétrica permitió extraer la información más relevante de las proteínas. El PCA se realizó con la intención de determinar aquellos biomarcadores que mejor representan a la mayoría de los casos de cáncer de pulmón. Como se puede ver en la Figura 18, al graficar los grupos de estudio con respecto al valor del componente que tomaron, se observa una marcada diferencia entre el grupo de pacientes con cáncer de pulmón y el grupo control.

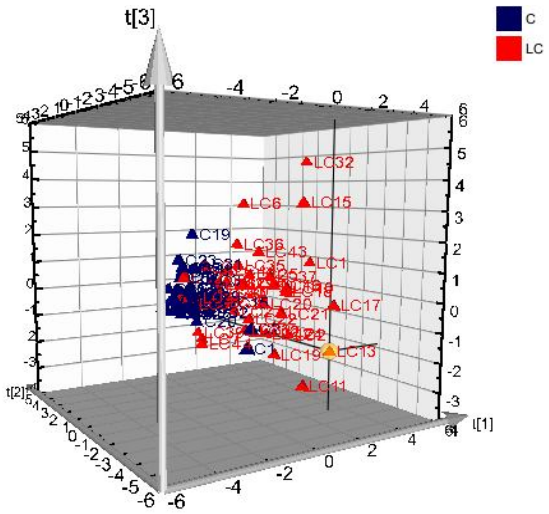

Figura 18. Distribución del grupo con cáncer de pulmón (LC) y grupo control (C) con respecto a sus componentes principales.

En la Figura 19 se muestran los pesos de los coeficientes de cada una de las proteínas biomarcadoras, de acuerdo a sus componentes principales. Como se puede observar, las 10 proteínas que se incluyeron en el análisis tienen una marcada influencia sobre el grupo de cáncer de pulmón. Sin embargo el conjunto de biomarcadores que describen de manera más adecuada al grupo de cáncer de pulmón son CEA, CA125, MMP9, Cyfra 21-

1, YKL40 y CRP, al tener un mayor valor en sus pesos sobre el primer y segundo componente. Esto implica que el vector inicial de biomarcadores se ha logrado reducir de catorce a seis proteínas.

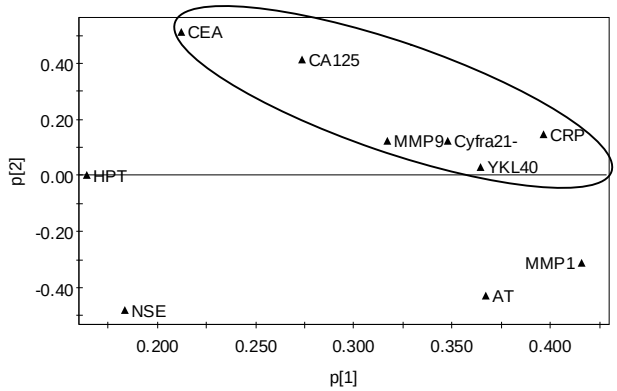

Figura 19. Análisis de Componentes Principales para las 10 proteínas biomarcadoras que mostraron ser significativas para el grupo de cáncer de pulmón.

Se entrenó una segunda red que llamaremos RNA_2, con los 6 biomarcadores que tuvieron mayor peso en el análisis PCA. Se investigó si la disminución del número de biomarcadores tenía efectos negativos sobre la correcta clasificación de la red (ver Figura 20).

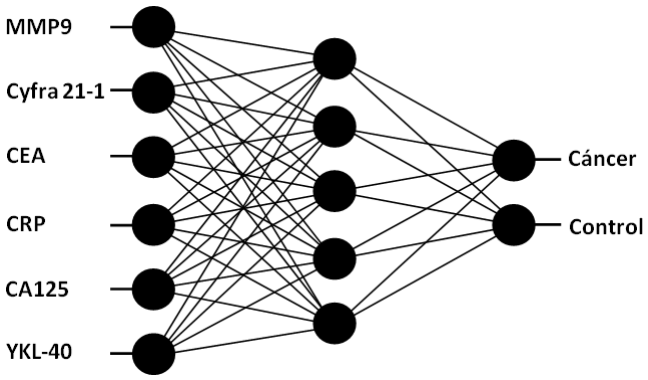

Figura 20. Arquitectura de la RNA con 5 neuronas en la capa oculta, entrenada con seis biomarcadores.

Se determinó la mejor arquitectura de la RNA_2 de acuerdo al error mínimo de entrenamiento, eligiendo como la mejor configuración, cinco neuronas en la capa oculta (ver Figura 21). Con esta nueva RNA_2 fue posible clasificar correctamente 133 de 150 sujetos, con una tasa de clasificación correcta del 93.3%, 96.6% y 89.7% para las fases de entrenamiento, validación y prueba respectivamente.

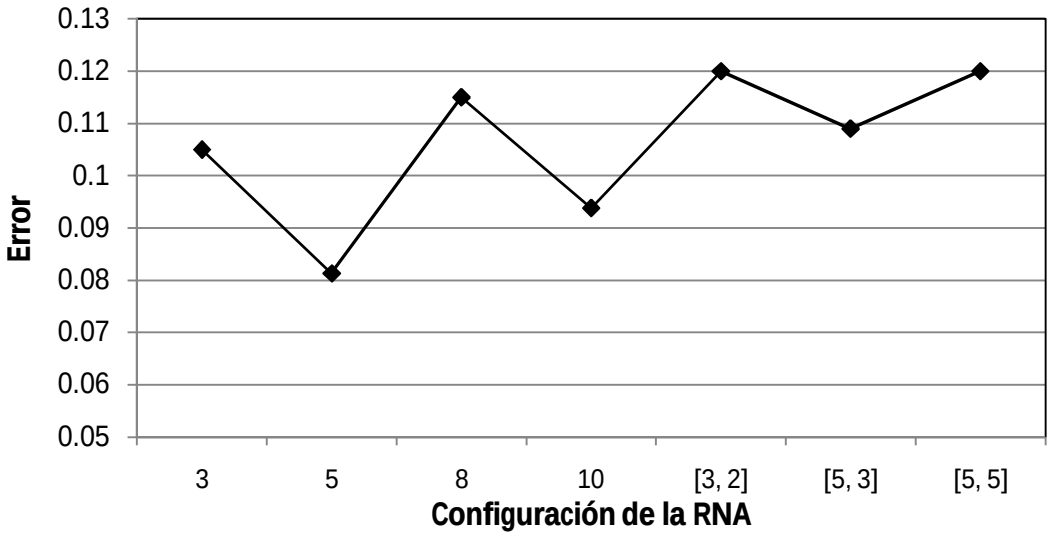

Figura 21. Determinación de la arquitectura para la RNA entrenada con seis proteínas biomarcadoras.

Al comparar los resultados de la RNA_2, con la RNA original de 10 neuronas en la capa oculta y entrenada con 14 biomarcadores de entrada, se puede discutir que la RNA_2 logró incrementar la tasa de clasificación en la fase de prueba un 13%.

Considerando que el desempeño de la RNA_2 es aceptable y comparable con la primera, se puede concluir que la reducción en el número de proteínas biomarcadoras de entrada en la RNA_2 no afectó el rendimiento de la misma. Esto es importante ya que se ha reducido el número de proteínas biomarcadoras en un 58% y por lo tanto los costos asociados se verían disminuidos. La red ha tenido mejor desempeño probablemente porque se ha disminuido la redundancia de información y ruido que presentaban las proteínas eliminadas por el PCA.

4.4.3 Optimización de las proteínas biomarcadoras utilizadas para entrenar la RNA.

Finalmente, con el propósito de reducir todavía más el número de proteínas utilizadas para entrenar la red,se probaron las posibles combinaciones de proteínas utilizadas en la RNA_2: CEA, CA125, MMP9, Cyfra 21-1, YKL40 y CRP. La mejor combinación estuvo dada por los cuatro biomarcadores, Cyfra 21-1, CEA, CA-125 y CRP. Esta nueva red llamada RNA_3 fue entrenada con las cuatro proteínas escogidas (ver Figura 22), clasificando correctamente 134 de los 150 sujetos de estudio. La RNA_3 tuvo un rendimiento similar al de la RNA_1 y RNA_2. Sorprendentemente, la tasa de clasificación fue mejorada, puesto que un paciente adicional fue clasificado correctamente. La tasa de clasificación de los datos utilizados fue del 88.9%, 96.6% y 93.1% para las fases de entrenamiento, validación y prueba, respectivamente. Las combinaciones de los biomarcadores como vectores de entrada para el entrenamiento de la RNA_3 permitieron reducir nuevamente el vector de 6 a 4 proteínas, así como incrementar la tasa de clasificación en la fase de prueba en un 3.4% y un 16.4% sobre las RNA_2 y RNA_1, respectivamente. Con la RNA_3 se ha reducido el número de proteínas biomarcadoras en un 71.4 % y por lo tanto los costos asociados se verían disminuidos.

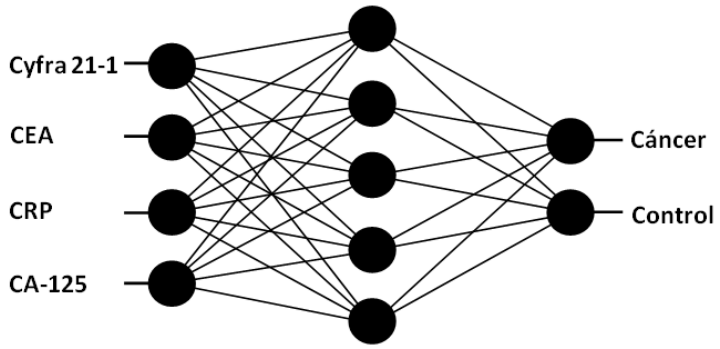

Figura 22. Arquitectura de la RNA con 5 neuronas en la capa oculta, entrenada con cuatro biomarcadores.

Se realizó la curva ROC correspondiente de la salida de la RNA_3 en función de los datos de prueba (muestras desconocidas para la RNA) y se comparó con la curva ROC de la mejor proteína individual que en este caso fue Cyfra 21-1 (que presentó una sensibilidad del 76.19%, al 80% de especificidad y un ABC = 0.85). La curva ROC generada de la RNA_3 presentó un 94.5% de sensibilidad, a la misma especificidad que Cyfra 21-1 (80%),

y un ABC de 0.964 (ver Figura 23), lo que indica que la RNA_3 permitió incrementar la sensibilidad un 18.31%, con respecto al mejor marcador individual.

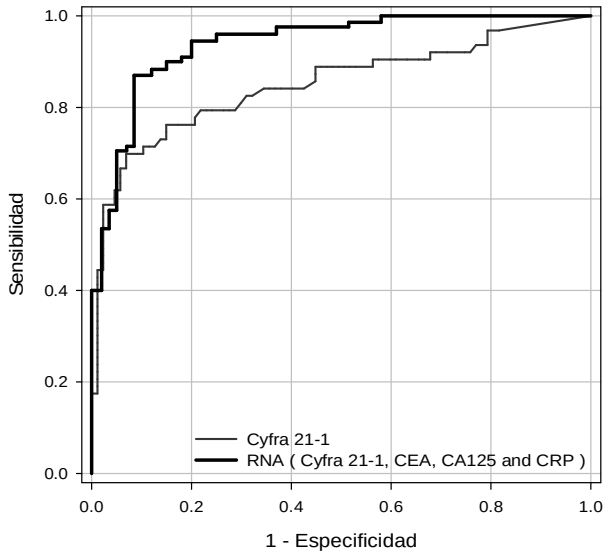

Figura 23. Comparación del desempeño de la RNA entrenada con cuatro biomarcadores y el mejor biomarcador individual, Cyfra 21-1.

En la literatura se han reportado diversos paneles de biomarcadores para detectar cáncer de pulmón y de mama, así como para clasificar las etapas de melanoma. En estos trabajos se reporta el empleo de diversas herramientas estadísticas y de la inteligencia artificial. En la Tabla 10 se pueden apreciar algunos de los paneles propuestos.

Tabla 10. Paneles de biomarcadores propuestos en otros trabajos para detección de cáncer.

Autor	Panel	Especificidad (%)	Sensibilidad (%)	Uso
Schnider y cols, 2002	Cyfra 21-1, NSE y CRP	95	92	Diagnóstico cáncer de pulmón
Leidinger y cols, 2010	macroarreglo (1827 péptidos)	85.7	97.8	Diagnóstico cáncer de pulmón
Lancashire y cols, 2005	27 péptidos	95.3	94.9	Estadio melanoma
Zhang y cols, 2009	5 péptidos	82.5	82.5	Diagnóstico cáncer de mama
Patz y cols, 2007	CEA, RBP, SCC y AT	84.7	89.3	Diagnóstico cáncer de pulmón
Farlow y cols, 2010	TNFa, Cyfra 21-1, MMP-2, IL-1ra, MCP, sE-selectina	85.5	84.8	Diagnóstico cáncer de pulmón
Farlow y cols, 2010	IMPDH, PGAM, HSP70-9B, Ubiquilina, A1 y A2	91.1	94.8	Diagnóstico cáncer de pulmón

SCC = Antígeno de células escamosas, TNFα = Factor de necrosis tumoral, MMP-2 =Metaloproteinasa de matriz 2, IL-1ra = antagonista del receptor de Interleucina 1, MCP = proteína quimiotáctica de monocitos 1, IMPDH = inosina-5-monofosfato deshidrogenasa, PGAM = fosfoglicerato mutasa, HSP70-9B = Proteína de choque térmico 70-9B , A1 =Anexina 1 y A2 = Anexina 2.

Las proteínas Cyfra 21-1 y CRP en esta tesis demostraron ser biomarcadores confiables que se pueden asociar al cáncer de pulmón, además coinciden con las reportadas por Schneider y cols (2002). Estos autores aplicaron lógica difusa como herramienta para mejorar la eficiencia diagnóstica de marcadores tumorales en cáncer de pulmón. Utilizaron un panel de tres biomarcadores: Cyfra 21-1, NSE, y CRP que fueron medidos en 175 pacientes con cáncer de pulmón del tipo histológico NSCLC y diferentes etapas, comparando los resultados con 120 sujetos control (27 con EPOC, 65 con neumoconiosis, y 11 personas con diferentes enfermedades pulmonares inflamatorias). Con la lógica difusa lograron incrementar la sensibilidad aproximadamente un 20% con respecto al mejor biomarcador, Cyfra 21-1, que presentó un 72% de sensibilidad. Esta herramienta de la inteligencia artificial permitió obtener una sensibilidad del 92%, mientras que nuestra RNA_3 presentó una sensibilidad del 94.5%.

En otro estudio, Leidinger y cols (2010) desarrollaron una prueba basada en máquinas de soporte vectorial (Support Vector Machines, SVM) para distinguir entre pacientes con cáncer de pulmón y sujetos control, usando un arreglo de antígenos peptídicos obteniendo una elevada tasa de clasificación (97.8% de sensibilidad, auna especificidad del 85.7%). Aunque con esta metodología lograron una sensibilidad y especificidad elevada, no probaron su modelo de clasificación con nuevas muestras para validarla. Esta

prueba se basó en un macroarreglo que consistió de 1,827 clonas de antígenos peptídicos, la cual representaría un costo muy elevado si se considera en la práctica clínica rutinaria.

Las RNAs en cáncer de pulmón se han utilizado para analizar la morfología de los tumores y clasificar el tipo histopatológico (Zhou y cols. 2002), sin embargo, hasta el momento no existen reportes publicados donde se empleen las RNAs para diagnóstico de cáncer de pulmón. Las RNAs se han aplicado en otros tipos de cáncer para determinar la etapa clínica del melanoma, así como para diferenciar pacientes con cáncer de mama, de pacientes sanos.

En un estudio realizado por Lancashire en el 2005, utilizaron RNAs para discriminar estadios clínicos en melanoma. Como entradas de la RNA emplearon los perfiles proteómicos de SELDI (Surface Enhanced Laser/Desorption Ionization) MS (Mass Spectrometry). La arquitectura de la RNA consistió de 2 neuronas en la capa oculta y una en la capa de salida. Sus salidas fueron codificadas como 1=etapa I y 2= etapa IV de melanoma. El mejor modelo resultante fue con 27 biomarcadores de entrada para la RNA, después de este número no observaron una mejoría significativa. Para comparar la capacidad de la RNA en la discriminación entre estadio IV y I de melanoma, generaron una curva ROC que presentó una sensibilidad y especificidad del 94.9% y 95.3%, respectivamente, donde la sensibilidad del modelo fue la capacidad de clasificar correctamente la etapa IV de melanoma, mientras que la especificidad fue el porcentaje de muestras en etapa I clasificadas correctamente. Con el trabajo desarrollado por estos autores se puede confirmar que combinando múltiples biomarcadores mediante las RNAs es posible incrementar tanto la sensibilidad como la especificidad para una prueba diagnóstica.

En un trabajo similar publicado en el 2009, Zhang y cols, desarrollaron un panel que consta de cinco biomarcadores, basados en perfiles proteómicos. Con el uso de RNAs y curvas ROC logran determinar un conjunto de biomarcadores para diferenciar pacientes con cáncer de mama con un 82.5% de especificidad y sensibilidad. En este estudio, adicionalmente al panel demostraron que usar dos neuronas en la capa de salida de la RNA permite obtener mejores resultados. En la RNA_3 que nosotros empleamos obtuvimos de igual manera una mejor clasificación cuando se tienen dos neuronas en la capa de salida, esto quizás sea debido a la capacidad de la capa oculta de tomar ventaja para

propagar el error efectivamente a través de las dos neuronas en la capa de salida de la RNA.

En resumen 14 proteínas fueron evaluadas en conjunto para incrementar la eficiencia de detección para cáncer de pulmón. Tres RNAs fueron usadas y mejoradas para clasificar con precisión pacientes con cáncer de pulmón y controles. El análisis PCA y las diferentes combinaciones en las entradas de la RNA permitieron reducir el número de biomarcadores. La RNA entrenada con sólo 4 proteínas fue capaz de clasificar correctamente 134 de los 150 sujetos involucrados en el estudio. Con esta RNA se incrementó la sensibilidad en un 18.31% comparado con el mejor biomarcador Cyfra 21-1.

En la siguiente sección se comparará el resultado de la RNA_3 con algunas técnicas estadísticas multivariadas.

4.5 Comparación de la RNA con diferentes métodos estadísticos.

Las herramientas estadísticas multivariables comúnmente se utilizan para el análisis de datos en el área clínica, por tal razón hacemos una comparación entre los resultados de la RNA_3 con algunas técnicas estadísticas de uso común como el análisis discriminante y métodos estadísticos avanzados como la combinación de curvas ROC, y el árbol de regresión y clasificación.

4.5.1 Análisis discriminante

Se generó un análisis discriminante con el objeto de determinar el número de muestras que era posible separar del grupo de cáncer de pulmón con respecto del grupo control. El análisis discriminante eligió a las proteínas biomarcadoras AT, MMP-9, HPT, Cyfra 21-1, YKL40, CEA, TF y RBP como las proteínas necesarias para separar un mayor número de pacientes con cáncer de pulmón del grupo control (ver Tabla 11).

Tabla 11 Pesos de los coeficientes de las proteínas utilizadas en el estudio.

Proteína	Coeficiente
AT	0.656425
MMP9	0.430185
HPT	0.359983
Cyfra 21-1	0.337619
YKL40	0.329232
CEA	0.285891
TF	-0.312352
RBP	-0.708668

Para este análisis se obtuvo un *p*-value < 0.0001 en lambda de Wilks (indicador de discriminación útil entre grupos). En la Figura 24 se puede observar la separación grafica de acuerdo al análisis, el cual elige ocho proteínas de las 14 evaluadas. El análisis discriminante clasificó correctamente un 76.2% del grupo de cáncer de pulmón y un 93.1% del grupo control.

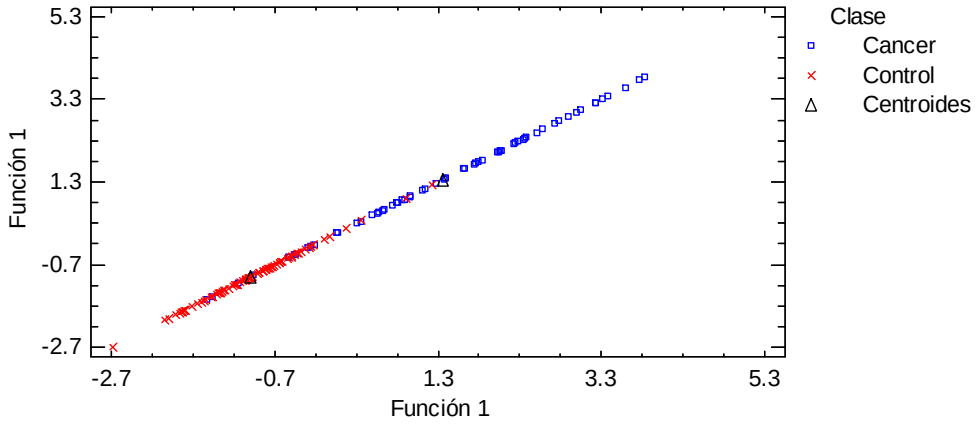

Figura 24. Separación del grupo de cáncer de pulmón y grupo control de acuerdo al análisis discriminante.

Comparando la RNA_3 con r el análisis discriminante se puede notar que la RNA tiene una tasa global de clasificación del 90.31% para el grupo de cáncer de pulmón y de 91.95% para grupo control. Además, el análisis discriminante para lograr una separación entre los grupos de estudio necesita ocho proteínas biomarcadoras lo cual representa un mayor costo económico asociado. Por otro lado, empleando la RNA_3 es posible predecir nuevos sujetos con una elevada sensibilidad.

4.5.2 Combinación de curvas ROC

Otra metodología con la que se comparó el desempeño de la RNA_3 fue la combinación de curvas ROC. Este método fue empleado por Tamura Masaya (2004), donde combinó 3 biomarcadores (Factor de crecimiento endotelial vascular (VEGF)-C, MMP-9, y VEFG), para diagnosticar metástasis en nódulos linfáticos, en pacientes con cáncer de pulmón de células no pequeñas. El método que utilizó Tamura (2004), se describió en la sección 3.4.2.4 de esta tesis.

Se realizaron las combinaciones de las curvas ROC de las proteínas biomarcadoras de acuerdo a las mismas proteínas que se emplearon como vectores de entrada para el entrenamiento de las RNAs. Combinaciones con 14, 6 y 4 biomarcadores fueron realizadas de acuerdo a la ecuación descrita anteriormente (sección 3.4.2.4). La combinación de 14, 6 y 4 biomarcadores mediante esta metodología permitió obtener una sensibilidad del 84.13%, 88.9% y 88.9% respectivamente a una especificidad del 80%. La mejor combinación mediante esta metodología fue con cuatro biomarcadores (Cyfra 21-1, CRP, CA125 y CEA). Como se puede observar en la Figura 25, con esta metodología se logró una sensibilidad del 88.9% al 80% de especificidad con un ABC = 0.91 (de acuerdo a la escala de Swets indica una exactitud alta). Con esta combinación se incrementó un 12.71% la sensibilidad con respecto al mejor biomarcador Cyfra 21.1 que presentó un 76.19% de sensibilidad a la misma especificidad. La RNA entrenada con estos mismos cuatro biomarcadores presentó 5.6% mayor sensibilidad comparada con la combinación de biomarcadores mediante esta metodología.

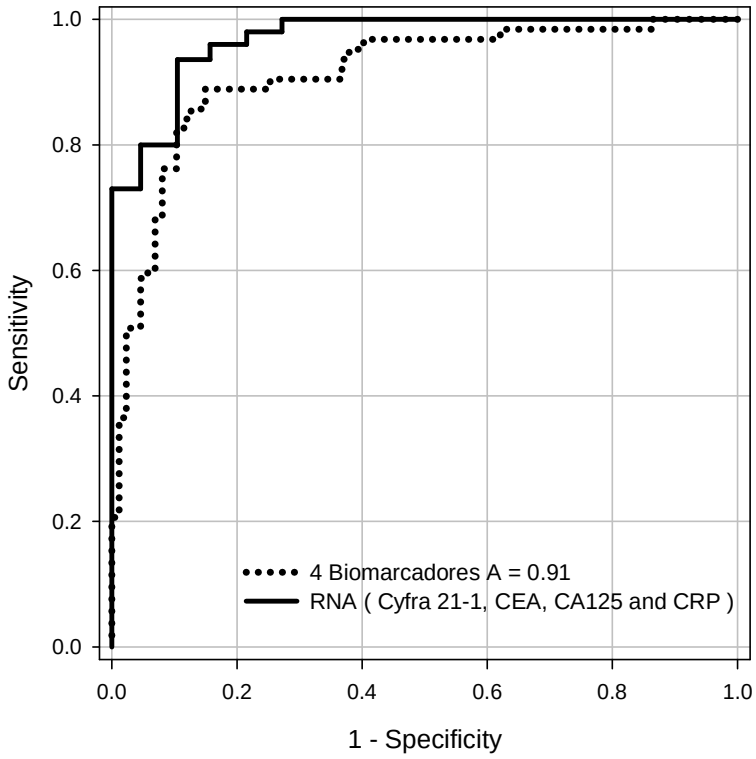

Figura 25. Comparación de la combinación de 4 proteínas biomarcadoras mediante la RNA y la metodología empleada por Tamura (2004).

4.5.3 Árbol de regresión y clasificación

Por otra parte, una metodología comúnmente empleada para proponer paneles de biomarcadores para diagnosticar diferentes patologías son los árboles de regresión y clasificación (CART, por sus siglas en inglés). Se realizó un CART con el objeto de clasificar cada sujeto a su grupo correspondiente tomando en cuenta sus datos de concentración de los biomarcadores. Este método se basa en la segmentación binaria donde el árbol se construye dividiendo repetidamente los datos. Para crear el árbol se tomaron los mismos vectores de datos que se emplearon para entrenar y validar la RNA, y posteriormente se hizo una prueba con el mismo vector con el que se probó la RNA. En la fase de creación del árbol (ver Figura 26) se obtuvo una muy buena clasificación ya que en el tercer y onceavo nodo terminal, 2 y 1 sujetos control fueron clasificados erróneamente como cáncer de pulmón respectivamente, mientras que por otro lado en el

doceavo nodo terminal sólo un paciente de cáncer de pulmón fue clasificado erróneamente como control (ver Tabla 12).

Tabla 12. Resumen de la estructura del Árbol de Clasificación y Regresión.

Nodo	Rama Izquierda	Rama derecha	n en Controles	n en Cáncer	Predicción	Constante de división	Proteína
1	2	3	71	49	Control	-2.7	CYFRA 21-1
2	4	5	69	16	Control	-17822.0	CRP
3			2*	33	Cáncer		
4	6	7	60	3	Control	-1.4	CYFRA 21-1
5	8	9	9	13	Cáncer	-14.1	CA125
6			53	0	Control		
7	10	11	7	3	Control	-3883.5	HPT
8	12	13	9	4	Control	-28.3	MMP1
9			0	9	Cáncer		
10			6	0	Control		
11			1*	3	Cáncer		
12			9	1*	Control		
13			0	3	Cáncer		

* Sujetos clasificados erróneamente

El CART presentó una sensibilidad del 98% a una especificidad del 96%, lo cual es demasiado alto y de gran utilidad para pruebas diagnósticas; sin embargo, cuando se realizó la fase de prueba con muestras desconocidas para el CART la sensibilidad y especificidad se vieron disminuidas en un 34% y 9%, respectivamente. De acuerdo al CART se emplearían cinco biomarcadores (Cyfra 21-1, CRP, CA-125, HPT y MMP1) para diferenciar los grupos, sin embargo, no es lo suficientemente sensible ni específica esta prueba, ya que presentó una sensibilidad del 64% y una especificidad de 87%. Comparando la RNA con el árbol de clasificación, con este método se necesita una proteína biomarcadora más que con la RNA_3 y la sensibilidad y especificidad son muy bajas con respecto a la RNA_3 que en la fase de prueba presentó un 92% de sensibilidad al 87% de especificidad.

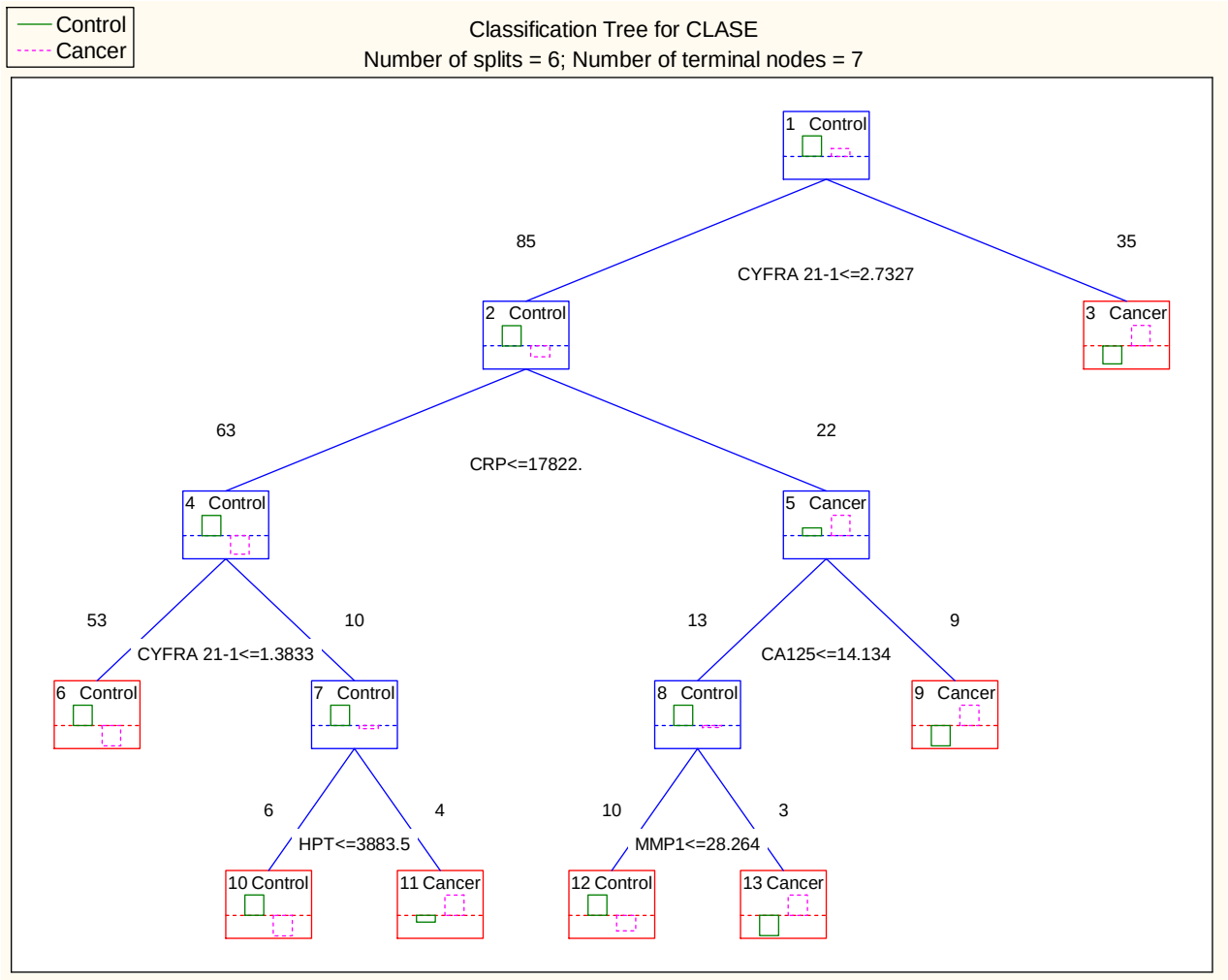

Figura 26. Árbol de clasificación y regresión que clasifica sujetos con cáncer de pulmón y controles.

Nuestros resultados de la RNA_3 tienen una mejor sensibilidad y especificidad comparada con otros trabajos donde han propuesto algunos paneles para detectar cáncer de pulmón. Patz y cols. (2007), propusieron un panel de biomarcadores para el diagnóstico de cáncer de pulmón basado en un árbol de clasificación y regresión empleando seis biomarcadores (TF, HPT, RBP, AT, CEA y Antígeno de células escamosas (SCC)). En este estudio, emplearon 50 pacientes con cáncer de pulmón y 50 sujetos control para crear el modelo de detección de cáncer de pulmón. Con su modelo final proponen emplear solo cuatro biomarcadores CEA, RBP, SCC y AT con el cual lograron obtener una sensibilidad del 89.3% con una especificidad del 84.7%, sin embargo cuando realizaron una fase de

prueba con muestras nuevas obtuvieron una sensibilidad y especificidad del 77.8% y 75.4%.

A diferencia del CART, la RNA_3 sigue presentando una mejor sensibilidad y especificidad aun cuando se trata de clasificar nuevas muestras, esto se puede atribuir a que las RNAs pueden asociar modelos no lineales como los procesos biológicos.

En otro estudio reciente, se seleccionaron 47 biomarcadores como candidatos para poder proponer un nuevo panel que permitiera diferenciar entre pacientes en etapas tempranas de NSCLC y sujetos control. Este grupo de investigación aplicó primeramente un análisis estadístico para diferenciar qué biomarcadores tienen capacidad diagnóstica y posteriormente generaron un CART con el cual 6 biomarcadores: Factor de necrosis tumoral α, Cyfra 21-1, antagonista del receptor de Interleucina 1, MMP2, proteína quimiotáctica de monocitos 1 y sE-selectina logran una sensibilidad del 99% y una especificidad del 95% en la fase de entrenamiento. Sin embargo, cuando prueban el CART con nuevas muestras, la sensibilidad y especificidad (84.8% y 85.5%, respectivamente) se ven disminuidas en un 14.2 y 9.5%. Comparando los resultados obtenidos en nuestro estudio, la RNA presentó un incremento de la especificidad y sensibilidad en un 1.5% y 7.2%, respectivamente (Farlow y cols. 2010). En otro trabajo, donde emplearon 196 pacientes con cáncer de pulmón en diferentes estadios de la enfermedad y 79 sujetos control, propusieron un panel de biomarcadores basado en la misma metodología del CART que consiste de seis autoanticuerpos, logrando obtener una sensibilidad del 94.8% y una especificidad del 91.1%, sin embargo no realizan una fase de prueba la cual corroboraría el correcto funcionamiento de esta metodología, y por lo visto en los estudios anteriores mencionados, es posible que la especificidad y sensibilidad no se mantengan, sino que disminuyan debido a la metodología empleada.

Con la discusión de estos resultados damos por finalizada la tesis. En el próximo capítulo se redactarán las conclusiones y posibles perspectivas.

Capítulo V

CONCLUSIONES

En este capítulo se presentan las conclusiones sobre el trabajo realizado, así como las perspectivas

Con respecto al objetivo:

1) Elaborar una base de datos con la información clínica más relevante de los pacientes que participen en el estudio.

- De la participación de las instituciones públicas (CMNO e ISSEMyM) fue posible generar una base de datos compuesta por 63 pacientes confirmados de cáncer sin tratamiento y 87 del grupo de controles.

- De la base de datos reunida se confirma que la enfermedad se presenta en mayor frecuencia en el sexo masculino que en el femenino.

- Así mismo se confirmó que el tipo histológico más frecuente es el cáncer de pulmón de células no pequeñas, presentándose en un 79.4%.

Con respecto al objetivo:

2) Determinar la concentración de las proteínas de interés por el método de ELISA.

- Se determinó la concentración de las catorce proteínas evaluadas, en las que diez presentaron mayor concentración en el grupo de cáncer de pulmón que en el grupo control.

- Los biomarcadores, APOAI, RBP, TF y uPA no demostraron ser potenciales en el diagnóstico de cáncer de pulmón.

Con respecto al objetivo:

3) Entrenar y validar una RNA, correlacionando la concentración de las proteínas biomarcadoras séricas con el diagnóstico de la enfermedad.

- Las tres RNAs fueron capaces de clasificar correctamente los pacientes de cáncer de pulmón y los controles. No obstante, los costos asociados con la evaluación de proteínas obligaron a reducir el número de biomarcadores usados por la red.

- El Análisis de Componentes Principales permitió reducir el vector de entrada para el entrenamiento de la Red Neuronal Artificial, de 14 a sólo seis proteínas biomarcadoras: CEA, CA125, MMP9, Cyfra 21-1, YKL40 y CRP)

- Combinar los diferentes biomarcadores y emplearlos como vectores de entrada para la Red Neuronal Artificial permitió proponer un panel de solo cuatro biomarcadores lo que representa una reducción en los costos, en el caso de desarrollar una prueba diagnóstica basada en estos biomarcadores.

  La RNA_3 de reconocimiento de patrones permitió incrementar la sensibilidad en un 18.31% con respecto al mejor biomarcador Cyfra 21-1 que por sí solo presentó una sensibilidad del 76.19%, logrando un mejor desempeño en la detección de la enfermedad.

- La RNA utilizada fue capaz de clasificar correctamente 88.9%, 96.6% y 93.1% para entrenamiento, validación y prueba, respectivamente.

Con respecto al objetivo:

4) Comparar el funcionamiento de la RNA contra algunos métodos estadísticos.

- La RNA_3 tiene mejor capacidad para detectar pacientes con cáncer de pulmón, en comparación con las diferentes herramientas estadísticas.

- La RNA_3 entrenada presentó un 5.5% de mayor sensibilidad con respecto a la combinación empleada para combinar las curvas ROC.

- El análisis discriminante no es capaz de discriminar entre grupos con un número pequeño de proteínas biomarcadoras ya que se deben usar ocho, además no tiene un alto porcentaje de clasificación.

- La RNA_3 presentó un 28% más de sensibilidad en la fase de prueba con respecto al árbol de clasificación y regresión.

Conclusión general

- Estos hallazgos preliminares muestran el potencial que tienen las RNAs en el campo de la validación de biomarcadores para su uso en la oncología clínica.

- La RNA mejoró significativamente la sensibilidad de los marcadores biológicos, por lo tanto, la RNA ofrece una prometedora herramienta auxiliar en el diagnóstico de cáncer de pulmón.

5. 1 PERSPECTIVAS

- Es de esperarse que al incrementar el tamaño de la muestra podrá mejorarse la clasificación de la red; es decir, habrá nuevos casos que ampliarán y reforzarán el universo de posibilidades que presenta la enfermedad.

- Evidentemente existen muchas arquitecturas y tipos de RNAs, es por lo tanto necesario probar otros tipos de redes que pudieran mejorar la capacidad de clasificación de la red.

- Sería interesante utilizar como complemento otras herramientas de la Inteligencia Artificial como Lógica Difusa para incrementar la capacidad de clasificación de la red.

- Una vez que la red dé un resultado con mayor robustez, sería conveniente desarrollar un dispositivo como método auxiliar para su uso en clínicas y como un método que pudiera diagnosticar oportunamente el cáncer de pulmón.

REFERENCIAS BIBLIOGRÁFICAS

Andreas, L. y Rimvydas, S. (1994) Using measurement data in bioprocess modelling and control. Trends Biotechnol. Vol 12. pp. 304-311.

Arias del Castillo, A., Fernández, A. y cols. (2001) Neoplasia de pulmón. Comportamiento epidemiológico. Rev. Cubana Oncol. 17(2):101-4.

Astion, M. y Wilding P. (1992) Application of neural networks to the interpretation of laboratory data in cancer diagnosis. Clin. Chem. 38(1):34-38.

Atkinson, J. y Senior R. (2003) Matrix metalloproteinase-9 in lung remodeling. Am J Respir Cell Mol. Biol. 28(1):12–24.

Akinkugbe, F., Ette, S. y cols. (1999) Iron deficiency anaemia in Nigerian infants. Afr J Med Med Sci. 28(1-2):25-9.

Averbukh, Z., Berman, S., y cols. (2004) Loss of captopril-bound Fe by end-stage renal failure patients during hemodialysis. J Nephrol. 17(1):101-6.

Barreiro, E. (2008) EPOC y cáncer de pulmón. Arch. Bronconeumol. 44(8):399-401.

Bembibre, L. y Lamelo, F. (2004) Neumonía adquirida en la comunidad. Guías clínicas 2004; 4(37).

Besser, D. y cols. (1996) Signal transduction and the u-PA/u-PAR system. Fibrinolysis 10:215-237.

Birkedal-Hansen, H. y cols. (1993) Matrix Metalloproteinases: A Review. Crit. Rev. Oral Biol. Med. 4(2):197-250.

Bjorklund, B. (1980) On the nature and clinical use of tissue polypeptide antigen (TPA). Tumor Diagnostik 1:9 20.

Bonnin, J. (1984) Immunohistochemystry of the central nervous system tumors. Its contributions to neurosurgical diagnosis.J Neurosurg, 60:121-133.

Borque, A., Sanz, G., y cols. (2001) The use of neural networks and logistic regression analysis for predicting pathological stage in men undergoing radical prostatectomy: a population based study. J Urol. 166:16728.

Burgueño, M., García-Bastos, J. y cols. (1995) Las curvas ROC en la evaluación de las pruebas diagnósticas. Med Clin (Barc). 104:661-70.

Camey, D., Marangos, P. y cols. (1982) Neuron-specific enolase: a marker for disease extent and response to therapy of small-cell lung cancer. Lancet 1:583-585.

Celli, B., MacNee, W. y cols. (2004) Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. Eur Resp J. 23:932-946.

Cross, B., Harrison, R. y cols. (1995) Introduction to neural networks. Lancet 346:1075-9.

Cruz, P., Villegas, V. y cols. (2008) Fundamento biológico y aplicación clínica de los marcadores tumorales séricos. Rev. Cienc. Salud. Bogotá, Colombia. 6 (2):85-98.

Dano, K., Behrendt, N. y cols. (1994) The urokinase receptor, protein structure and role in plasminogen activation and cancer invasion. Fibrinolysis. 1:189–203.

Davidson, W. y Thompson, T. (2007). The Structure of Apolipoprotein A-I in High Density Lipoproteins. J Biol Chem. Vol. 282(31), 22249–22253.

DiRusso, S., Sullivan, T. y cols. (2000). An artificial neural network as a model for prediction of survival in trauma patients: validation for a regional trauma area. J Trauma; 49:212-23.

Diez, M., Cerdan, F. y cols. (1991) Evaluation of serum CA125 as a tumor marker in non-small cell lung cancer. Cancer 67:150-154.

Dmitriy, A., Yablonskiy, A. y cols. (2002) Quantitative *in vivo* assessment of lung microstructure at the alveolar level with hyperpolarized ^{3}He diffusion MRI. J Appl Physiol. 99(5):3111-3116.

Don, D. Sin, S. y cols. (2006) Progression of Airway Dysplasia and C-Reactive Protein in Smokers at High Risk of Lung Cancer. Am J Respir Crit Care Med. 173: 535–539.

Engstrom, G. y cols. (2004) Arterioscler. Thromb Vasc Biol. 24(8):1498-502.

Eldar, S. Siegelmann, H. y cols. (2002) Conversion of laparoscopic cholecystectomy to open cholecystectomy in acute cholecystitis: artificial neural networks improve the prediction of conversion. World J Surg. 26:79-85.

Fernández, A., Martínez, A. y cols. (2007) Marcadores tumorales serológicos. Química Clínica 26(2) 77-85.

Franco-Marina, F., Villalba-Caloca, J. y cols. (2006) Role of active and passive smoking on lung cancer etiology in Mexico City. Salud Publica Mex.48: supl I:S75-S82.

James, A., Freeman, y cols. (1991, Redes neuronales, algoritmos aplicaciones y técnicas de programación. Capítulo 2, Adaline y Madaline, editorial Addison-Wesley / Diaz de santos, E.U.A.

García, C. (2008) Hospital General Yagüe de Burgos. Grupo Español de Cáncer de Pulmón. En línea:http://www.elmundo.es/elmundosalud/especiales/cancer/pulmon.html

Gress, M. y cols. (1995) Expression and *in-situ* localization of genes coding for extracellular matrix proteins and extracellular matrix degrading proteases in pancreatic cancer. Int. J. Cancer. 62(4):407-413.

Hasui, Y. y cols. (1992) The content of urokinase-type plasminogen activator antigen as a prognostic factor in urinary bladder cancer. Int. J. Cancer. 50(6):871-873.

Hagan, M., Demuth, H. y cols. (1996). Neural Network Design. Boston, MA: PWS Publishing.

Hayes, F., Bast, C. y cols. (1996) Tumor marker utility gradingsystem: A framework to evaluate clinical utility of tumor markers. J Natl Cancer Inst. 88:1456-1466.

Hechavarría J., Blanco A. y cols. (1999). Algunas consideraciones sobre asma ocupacional. Rev Cubana Med 1999;38(3):188-93.

Hauck, W., Hauptmann, A. y cols. (2004) Alpha-1-antitrypsin levels and genetic variation of the alpha-1-antitrypsin gene in Peyronie's disease. Eur Urol. 46(5): 623-8; discussion 628.

Hong, Y., Chia, S. y cols. (2000) Urinary protein excretion in Type 2 diabetes with complications. J Diabetes Complicat. 14(5):259-65.

Iwata, H. y cols. (1996) Production of matrix metalloproteinases and tissue inhibitors of metalloproteinases in human breast carcinomas. Jpn J Cancer Res 87(6):602-611.

Janeway, A., Travers, P. y cols. (2004) Mark. Immunobiology. Garland Publishing 6th Ed New York and London.

Karabatak, M., y Ince, C. (2008) An expert system for detection of breast cancer based on association rules. Expert Syst Appl. Doi:10.1016/j.eswa.2008.02.064.

Wahab, J.y cols. (2005) Heterozygosity for alpha1-antitrypsin deficiency as a cofactor in the development of chronic liver disease. Ned Tijdschr Geneeskd. 10;149(37):2057-61.

Kobayashi, H., Ooi, H. y cols. (2007) Serum CA125 level before the development of ovarian cancer. Int J Gynecol Obstet 99, 95–99.

Kucharz, J. y cols. (2000) Acute-phase proteins in patients with systemic sclerosis. Clin Rheumatol 19(2):165-166.

Lancashire, L., Ugurel, S. y cols. (2005) Utilizing artificial neural networks to elucidate serum biomarker patterns which discriminate between clinical stages in melanoma. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology CIBCB 05. Nov. 14-15:1-6.

López, M., Valencia, J. y cols. (2005). Comparación de la Tomografía Axial Computarizada con el estudio anatomopatológico en el estadiaje ganglionar del cáncer de pulmón. Oncología.28(7):243-350.

Nagase, H. (1996) Matrix metalloproteinases in Zinc Metalloproteases in Health and Disease Hooper. N.M., ed. Taylor & Francis, Bristol, PA, pp. 153 - 204.

Niloff, J., Klug, T. y cols. (1984) Elevation of serum CA 125 in carcinomas of the fallopian tube, endometrium, and endocervix. Am J Obstet Gynecol 148:1057-1058.

Press, W., Teukolsky, S. y cols. (1992) Numerical recipes in C, the art of scientific computing. Second edition, Cambridge Univerity press. Chapter 15, Modelling of data, nonlinear models, pp. 683-687.

Rumelhart, D., Hinton, G. y cols. (1986) Learning internal representations by error propagation. Parallel Distributed Processing: Explorations in the microstructures of cognition, vol 1, Cambridge, MA: MIT, pp. 318-362.

Sarle, W. (1994). Neural networks and statistical models. Proceedings of the 19th annual SAS Users Group International Conference, Cary, NC (SAS Institute),1538-50.

Schwarzer, G., Vach, W. y cols. (2000) On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Statist Med; 19:541-61.

Sargent, D. (2001) Comparison of artificial neural networks with other statistical approaches. Results from medical data sets. Cancer. 91:1636-42.

Steven, E., Nissen, M. y cols. (2003) Effect of recombinant ApoA-I Milano on coronary atherosclerosis in patients with acute coronary syndromes. JAMA, 290:2292-2300.

Tourassi, G. y Floyd C. (1997) The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. Med Decis Making; 17:186-92.

Leidinger, P., Andreas, K. y cols. (2010) Identification of lung cancer with high sensitivity and specificity by blood testing. Respiratory Research,11:18.

Li, Z., Giovanna E. y cols. (2005). Using Protein Microarray as a Diagnostic Assay for Non–Small Cell Lung Cancer. AJRCCM 172: 1308–1314.

Closs, E., Lyons, J. y cols. (1993) Characterization of the third member of the MCAT family of cationic amino acid transporters. Identification of a domain that determines the transport properties of the MCAT proteins. J. Biol. Chem. 268:20796-20800.

Malinda, K., Ponce, L. y cols. (1999). GP38k, a protein synthesized by vascular smooth muscle cells, stimulates directional migration of human umbilical endothelial cells. Exp. Cel. Res. 250.

Matrisian, L. (1992) The matrix-degrading metalloproteinases. BioEssays 14(7):455-463.

Matuszek, M. y cols. (2003) Haptoglobin elutes from human atherosclerotic coronary arteries a potential marker of arterial pathology. Atherosclerosis 168(2):389-396.

Mellerik, D., Osborn, M. y cols. (1990) On the nature of serological tissue polypeptide antigen (TPA): monoclonal keratin 8, 18, 19 antibodies react differently with TPA prepared from human cultured carcinoma cells and TPA in human serum. Oncogene. 5:100-117.

Mitra, A., Wahed, M. y cols. (2002) Urinary retinol excretion in children with acute watery diarrhoea. J Health Popul Nutr. 2002 Mar;20(1):12-7.

Montaner, J. (2001) El papel de las metaloproteinasas de matriz den la fase aguda del ictus isquémico. Tesis Doctoral. Hospital universitario Vall d'Hebron. Universidad Autonoma de Barcelona. p.p. 18.

Nisman, B., Barak, V. y cols. (2002) Evaluation of urine CYFRA 21-1 for the detection of primary and recurrent bladder carcinoma. Cancer. 2002 Jun 1;94(11):2914-22.

Noy, N. (2000) Retinoid-binding proteins: mediators of retinoid action. Biochem J. 2000 Jun 15;348 Pt 3:481-95.

Ocampo, M., Salmón, J. y cols. (2008) Bronquiectasias: revisión bibliográfica. Revista de Posgrado del 16 a VIa Cátedra de Medicina. No.182.

Okada, S. y cols. (1996) Arterioscl. Thromb. Vasc. Biol. 16:1269.

Pearlmutter, B. (1990) "Dynamic Recurrent Neural Networks". Technical Report, School of Computer Science, Carnegie Mellon Univ. CMU-CS pag. 88-191.

Picardo, A., Torres, A. y cols. (1994) Quantitative analysis of carcinoembryonic antigen, squamous cell carcinoma antigen, CA 125, and CA50 cytosolic content in non-small cell lung cancer. Cancer 73: 2305-2311.

Qi, W., Liu, X. y cols. (2005) Isoform-specific expression of 14-3-3 proteins in human lung cancer tissues. Int J Cancer 2005, 113:359-363.

Recklies, A., White, C. y cols. (2002) The chitinase 3-like protein human cartilage glycoprotein 39 (HC-gp39) stimulates proliferation of human connective-tissue cells and activates both extracellular signal-regulated kinase- and protein kinase B-mediated signaling pathways. Biochem. J.365:119–126.

Rocha-Pereira, P. y cols. (2004) The inflammatory response in mild and in severe psoriasis. Br J Dermatol. 150(5):917-928.

Ruiz, A., Cabezón P. y cols. (2004) Cáncer de Pulmón. Servicio de Oncología Radioterápeutica. Biocáncer 1.

Ruíz-Godoy, L., Rizo, R y cols. (2007) Mortality due to lung cancer in Mexico. Lung Cancer. 58:184-190.

Saínz, B. (2006) Enfisema pulmonar y bullas de enfisema. Clasificación. Diagnostico. Tratamiento. Rev Cubana Cir 2006;45 (3-4).

Sarle, W. (1997) Neural network FAQ, part 1 of 7: introduction, periodic posting to the Usenet newsgroup comp.ai.neural-nets. En linea: ftp://ftp.sas.com/pub/neural/FAQ.html

Schneider, J., Bitterlich, N. y cols. (2002) Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer. Int J Clin Oncol 7:145–151.

Schrohl, A., Holten-Andersen, M. y cols. (2003) Tumor markers: from laboratory to clinical utility. Mol Cell Proteomics. 2:378-387.

Shetty, S., Gyetko, M. y cols. (2005) Induction of p53 by urokinase in lung epithelial cells. J Biol Chem. 280:28133–28141.

Solomon, A., McLaughlin, C. y cols. (1969) Proteins and light chains of immunoglobulins. J Biol Chem. 244:3393-3404.

Swets, J. (1988) Measuring the accuracy of diagnostic systems. Science. 240:1.285-1.293.

Takkunen, M., Mika, H. y cols. (2009) Podosome-like structures of non-invasive carcinoma cells are replaced in epithelial-mesenchymal transition by actin comet-embedded invadopodia. J Cell Mol Med. 28:1965-6240.

Teramoto, S. (2007) COPD pathogenesis from the viewpoint of risk factors. Intern Med. 46(2):77-9.

Tokunou, M. Niki, T. y cols. (2000) Altered expression of the ERM proteins in lung adenocarcinoma. Lab Invest. 80:1643-1650.

Tüba, K. y Tülay, Y. (2003). Breast cancer diagnosis using statistical neural networks, International XII. Turkish Symposium on Artificial Intelligence and Neural Networks TAINN.

Ueda, Y. y cols. (1996) Matrix metalloproteinase 9 (gelatinase B) is expressed in multinucleated giant cells of human giant cell tumor of bone and is associated with vascular invasion. Am J Pathol 148(2):611-622.

Van, H. (2000). Validation, calibration, revision and combination of prognostic survival models. Statist Med 19:3401-15.

Van V. y cols. (2004) Haptoglobin polymorphisms and iron homeostasis in health and in disease. Clin Chim Acta. 345(1-2):35-42.

Chi-Shing, C. (2007) Potentially useful biomarkers for the diagnosis, treatment and prognosis of lung cancer. Biomed Pharmacother, 61:515-519.

Widrow, B. y Hoff, M. (1960) Adaptive switching circuits. Ire wescon Convention Record. New York IRE, pages 96–104.

Williams, P. y Hoskins, J. (1997) "Heuristic algorithms, backpropagation, a generalized delta rule". Byte, pp. 155-162.

Xi, L., Nicastri, D. (2007) Optimal markers for real-time quantitative reverse transcription PCR detection of circulating tumor cells from melanoma, breast, colon, esophageal, head and neck, and lung cancers. Clin Chem. 53:1206-1215.

Ying-Chin, K., Chien-Hung, L. y cols. (1997) Risk factors for primary lung cancer among non-smoking women in taiwan. Iint J Epidemiol. vol. 26, No.1.

Zhi-Hua, Z., Yuan, J. y cols. (2002) Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles. Artif Intell Med. Vol.24, No.1, pp.25-36.

Zidman, I. (1961) The fate of circulating tumors cells. I. Passage of cells through capillaries. Cancer Res. 21:38-39.

Zúñiga, I., Pacheco, S. (2006) Neuroblastoma: El Cáncer como diagnóstico diferencial del Maltrato Infantil. Revista Pediatría Electrónica. Vol 3, No.2.

APÉNDICE A

Descripción de las proteínas biomarcadoras utilizadas en el estudio

La Enolasa Especifica de Neuronas (NSE) es una proteína citoplasmática de las células neurales (Zúñiga Isabel y Cols, 2006), que en condiciones normales está restringida a las neuronas del sistema nervioso central y periférico (Bonnin y Cols, 1984) y que ha mostrado ser un marcador tumoral sensible, especialmente en cáncer de pulmón de células pequeñas (Camey y Cols, 1982).

El activador de Plasminógeno tipo Urocinasa (uPA) regula la apoptosis de células epiteliales de pulmón (Shetty y Cols, 2005). El uPa es una serina preoteasa muy restringida que corta al zimógeno plasminógeno para formar plasmina, una serina proteasa de amplio espectro, capaz de degradar la mayoría de los principales componentes proteínicos de la matriz extracelular. La unión de uPA a su receptor y la posterior proteólisis pericelular mediada por el uPA, están involucradas en muchos procesos celulares como la migración y la remodelación tisular, en angiogénesis, aterogénesis, metástasis de células tumorales y la ovulación (Okada y Besser 1996). Un nivel elevado de uPA es un marcador de mal pronóstico para el cáncer de mama, cáncer de próstata agresivo, cáncer de vejiga y cáncer gástrico (Hasui y Cols, 1992), encontrándose también en neoplasias pulmonares (Dano y Cols. 1994).

La Proteína C reactiva (CRP) es una proteína de fase aguda, la cual se eleva en respuesta a inflamación, uniéndose a la fosfocolina en los microbios. También en la unión del complemento a células extrañas y dañadas, así como potenciador de la fagocitosis por macrófagos, encontrándose en concentraciones elevadas en individuos asintomáticos con lesiones bronquiales precancerosas (Don D y Cols, 2006).

Las metaloproteinasas de matriz juegan un papel importante en procesos fisiológicos y patológicos incluyendo embriogénesis, remodelación de tejido, cicatrización de heridas, inflamación, artritis y cáncer.

La Metaloproteinasa de Matriz 9 (MMP9) es una enzima proteolítica, conocida también como gelatinasa B. Pertenece a la familia de las metaloproteinasas de matriz (MMPs) que son una familia de endopeptidasas dependientes de zinc encargadas de remodelar la matriz extracelular y que en conjunto pueden degradar todos los constituyentes de la

misma (Montaner Villalonga, 2001). La MMP-9 está implicada en la inflamación, remodelación tisular, cicatrización de heridas, la movilización de la matriz, factores de crecimiento y transformación de las citoquinas (Closs y Cols, 1993). Su expresión se correlaciona con la desmoplasia (depósito de colágeno anormal) que acompaña al cáncer de páncreas (Gress, y Cols, 1995), con la metástasis a los ganglios linfáticos por las células del carcinoma de mama (Iwata y Cols, 1996) y con la invasión de los vasos regionales de los tumores de células gigantes de los huesos (Ueda y Cols, 1996). La MMP-9 puede estar elevada en el líquido crevicular gingival y en la saliva de pacientes con gingivitis y enfermedades periodontales (Birkedal, 1993). Se ha reportado que está proteína es la que más predomina en las enfermedades inflamatorias de las vías respiratorias (Atkinson y Cols, 2003).

La **La Metaloproteinasa de Matriz 1 (**MMP-1) tienen un rol significativo en la degradación de colágeno fibrilar, en el remodelado de la matriz extracelular, caracterizada por la ruptura de la triple hélice del colágeno intersticial en fragmentos ¾ y de ¼. La MMP-1 está implicada en una amplia variedad de procesos biológicos donde ocurre la degradación del colágeno. Estos procesos incluyen artritis reumatoide, osteoartritis, enfermedad periodontal, invasión tumoral, angiogénesis, ulceración corneal, remodelación de tejido, enfermedad inflamatoria intestinal, la aterosclerosis, aneurismas y reestenosis (Nagase, 1996).

El Antígeno de Cáncer 125 (CA 125) es un antígeno de superficie asociado con cáncer epitelial de ovario. En el suero, el CA 125 se asocia con una glicoproteína de peso molecular alto. Los estudios publicados han indicado que elevados niveles séricos de CA 125 se pueden encontrar en personas cáncer de ovario de diferentes tipos histológicos como seroso y no seroso. (Kobayashi y Cols. 2007). La concentración en suero de CA 125 es superior a 35 U/mL en el 60% de las mujeres con cáncer de ovario y mayor del 80% de los individuos con cáncer diseminado del ovario. El CA 125 en suero se eleva un 1% de lo normal en mujeres sanas, 3% de lo normal en mujeres sanas con enfermedades benignas de ovario, y un 6% en personas con afecciones no neoplásicas (incluyendo, pero no limitado al embarazo en el primer trimestre, la menstruación, endometriosis, fibrosis uterina, salpingitis agudas, las enfermedades hepáticas e inflamación del peritoneo, pericardio o pleura). La concentración en suero de CA 125 es útil para distinguir entre una respuesta satisfactoria al tratamiento o una respuesta terapéutica pobre; es decir, la progresión de la enfermedad maligna. Hasta la fecha, el CA 125 es el marcador

más sensible para el cáncer de ovario epitelial residual. CA 125 también puede estar elevado en individuos con carcinoma de pulmón, cuello del útero, las trompas de Falopio y útero y la endometriosis (Niloff y Cols. 1984).

La Haptoglobina (HPT) es una proteína de plasma con capacidad de enlazarse a la hemoglobina y a las glicoproteínas de plasma que forman un complejo estable con la hemoglobina para ayudar al reciclaje de hierro del grupo hemo. Es un biomarcador conocido de la hemólisis (Van Vlierberghe, 2004). Niveles elevados de Haptoglobina en plasma se asocian con un aumento de riesgo cardiovascular en hombres obesos (Engstrom y Cols, 2004), con inflamación (Rocha-Pereira y Cols, 2004), aterosclerosis (Matuszek y Cols, 2003), y esclerosis sistémica (Kucharz y Cols, 2000).

La Apolipoproteína AI (APOAI) comprende alrededor del 70% de las lipoproteínas de alta densidad (HDL). La APOAI es un polipéptido de 28 kDa, que carece de glicosilación o puentes disulfuro (Davidson y Thompson, 2007). Entre el 5-10% de APOAI existe en un estado no asociado de lipoproteínas, en el plasma humano. La APOAI parece tener efectos sobre la inhibición de la aterosclerosis, el transporte reverso del colesterol y anti inflamación (Steven y Cols. 2007).

La α1-Antitripsina (AT) es una proteína que protege los pulmones. El hígado usualmente sintetiza la proteína y la libera en el torrente sanguíneo. La AT es un inhibidor de proteasas importante que controla la degradación del tejido. Una reducción en los niveles de AT puede causar un cambio en el metabolismo del colágeno (Hauck y Cols. 2004). La AT inhibe la liberación de elastasa de neutrófilos en los pulmones durante los estados inflamatorios. La deficiencia de AT es una enfermedad genética poco frecuente que puede conducir a enfisema, hepatitis, cirrosis (Kok y Cols. 2005) y a la enfermedad pulmonar obstructiva crónica (EPOC) (Teramoto, 2007).

CYFRA 21-1 es un fragmento de la citoqueratina 19. Aunque se expresa en todos los tejidos corporales, su mayor incidencia está en el pulmón, particularmente en los tejidos de cáncer de pulmón. La principal importancia de Cyfra 21-1 es que es un biomarcador en el diagnóstico diferencial, pronóstico y cuidado posterior de cáncer de pulmón. Además, Cyfra 21-1 ha sido descrito como un marcador tumoral para el seguimiento del cáncer de vejiga (Nisman y Cols. 2002).

La Proteína de Unión a Retinol (RBP) actúa mediante la solubilización y la protección de sus ligandos lábiles en espacios acuosos. También tiene funciones diversas y especificas en la regulación de la disposición, el metabolismo y las actividades de retinoides (Noy y Cols. 2000). La RBP es el portador plasmático específico del retinol, y encargada del transporte de la vitamina del hígado a las células blanco. Bajos niveles en suero de RBP se han observado en casos de diarrea (Mitra y Cols. 2002). Altos niveles de RBP en orina pueden ser un buen indicador de daño renal, así como las complicaciones microvasculares de la diabetes mellitus tipo 2 (Hong y Cols. 2000).

La Transferrina (TF) es una proteína plasmática que transporta el hierro a través de la sangre hacia el hígado, el bazo y la médula ósea. Bajo nivel de transferencia en plasma podría asociarse con anemia y con enfermedad crónica del hígado (Averbukh y Cols. 2004). Por otro lado un alto nivel de transferrina plasmática podría indicar anemia por deficiencia de hierro (Akinkugbe y Cols 2004)

YKL-40 es una proteína que pertenece a la quitinasa de mamíferos; es una glicoproteína unida a heparina de 40 kDa. Comparte homología de secuencia de aminoácidos con quitinasas de animales no mamíferos, pero no muestra actividad de quitinasa. El nombre de YKL-40 se deriva del peso molecular de la proteína y de tres aminoácidos N-terminal (tirosina, lisina y leucina). Su función biológica sigue siendo en gran medida desconocida y es un campo de amplio debate científico. YKL-40 ha mostrado ser un potente factor de crecimiento de tejido conectivo celular (Recklies y Cols. 2002) y un factor de migración potente para las células endoteliales (Malinda 199). En varios estudios realizados se han encontrado niveles importantes de YKL-40 en entornos con inflamación o en la remodelación de la matriz extracelular, incluyendo varios tipos de cáncer, artritis reumatoide, enfermedades inflamatorias del intestino, infecciones bacterianas graves, y la fibrosis del hígado.

APÉNDICE B

Tabla 12. Información de Kits de ELISA comerciales para cuantificación de proteínas.

Biomarcador	Marca	No. Catalogo	Pagina Web
MMP1	R&D Systems; Minneapolis, MN	DMP100	www.rndsystems.com
MMP9	R&D Systems; Minneapolis, MN	DMP900	www.rndsystems.com
uPA	ASSAYPRO; St. Charles, MO	EU1001-1	www.assaypro.com
TF	ASSAYPRO; St. Charles, MO	ET2105-1	www.assaypro.com
AT	ASSAYPRO; St. Charles, MO	EA5001-1	www.assaypro.com
HPT	ASSAYPRO; St. Charles, MO	EH1003-1	www.assaypro.com
CA125	ALPCO Diagnostics; Salem, NH	25-125HU-E01	www.alpco.com
CEA	ALPCO Diagnostics; Salem, NH	25-CEAHU-E01	www.alpco.com
Cyfra 21-1	DRG Instruments GmbH; Germany	EIA-5070	www.drg-diagnostics.de
NSE	ALPCO Diagnostics; Salem, NH	43-NSEHU-E01	www.alpco.com
APOAI	ASSAYPRO; St. Charles, MO	EA5201-1	www.assaypro.com
RBP	ASSAYPRO; St. Charles, MO	ER2005-1	www.assaypro.com
CRP	ASSAYPRO; St. Charles, MO	EC1001-1	www.assaypro.com
YKL40	Quidel Corporation; San Diego, CA	8020	www.quidel.com

APÉNDICE C

Descripción de métodos estadísticos empleados en este estudio

Curvas ROC:

Las curvas características operativas relativas (ROC) representan la razón de verdaderos positivos frente a la razón de falsos positivos en una prueba diagnóstica, es decir permite clasificar cuales de un conjunto de datos pertenecen a un grupo o a otro según se varía el umbral (punto de corte) de discriminación (**Burgueño M, 1993**). El área bajo la curva (ABC) de una curva ROC es una medida global de la exactitud de una prueba diagnóstica, es decir es la probabilidad de clasificar correctamente un par de individuos sano y enfermo, seleccionados al azar de la población, mediante los resultados obtenidos al aplicarles la prueba diagnóstica. Cuando el ABC toma un valor comprendido entre 0.5 no existe diferencia en la distribución de resultados de la prueba entre los grupos enfermo y sano; y con un ABC de 1.0 existe separación perfecta entre las dos distribuciones. Existe una interpretación del ABC: valores entre 0.5 y 0.7 indican baja exactitud, entre 0.7 y 0.9 pueden ser útiles para algunos propósitos y un valor mayor de 0.9 indica exactitud alta (Swets JA, 1988).

Análisis Discriminante

El análisis discriminante es una técnica estadística multivariada cuya finalidad es describir diferencias (si llegasen a existir) entre grupos de objetos sobre los que se observan p variables (variables discriminantes). Se puede considerar como un análisis de regresión donde la variable dependiente es no métrica (categórica) y tiene como categorías la etiqueta de cada uno de los grupos, y las p variables independientes son continuas y éstas determinan a qué grupos pertenecen los objetos.

Con esta técnica se encuentran relaciones lineales entre las variables continuas que mejor discriminen en los grupos dados a los objetos, siendo posible construir una regla de decisión que asigne un objeto nuevo, que no sabemos clasificar previamente, a uno de los grupos prefijados con un cierto grado de riesgo. También permite identificar las variables que son importantes para distinguir entre los grupos.

Árbol de regresión y clasificación

El árbol de regresión y clasificación (CART) es un método multivariado no paramétrico que permite trabajar con todo tipo de variables predictoras (binarias, nominales, ordinales y de intervalo o razón) se basa en la segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. El nodo inicial es llamado nodo raíz o grupo madre y se divide en dos grupos hijos o nodos, luego el procedimiento de partición es aplicado a cada grupo hijo por separado. Las divisiones se seleccionan de modo que la impureza de los nodos sea menor que la del grupo madre y éstas están definidas por un valor de una variable explicativa (Deconinck et al., 2006). Las ramas que conforman el árbol representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición) formando así grupos homogéneos respecto a la variable que se desea discriminar. Las particiones se hacen en forma recursiva hasta que se alcanza un criterio de parada, el método utiliza datos ya conocidos para construir el árbol de decisión, y este árbol se usa para clasificar nuevos datos. El objetivo de este método es discriminar, estimar o predecir la variable Y en función de los predictores X_1……Xp, mediante particiones sucesivas del conjunto de individuos, maximizando una medida de contenido de información respecto a la variable respuesta.

Análisis de Componentes Principales

El Análisis de Componentes Principales (PCA por sus siglas en inglés) es una técnica estadística multivariada cuya finalidad es sintetizar la información, ó reducir la dimensión (número de variables). Para estudiar las relaciones que se presentan entre p variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables no correlacionadas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de componentes principales. Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra y además serán independientes entre sí (coeficiente de correlación igual a 0). Un aspecto clave en el PCA, que no es fácil, es la interpretación de los factores, ya que ésta no viene dada en el análisis, sino que será deducida tras observar la relación de los factores con

las variables iniciales (estudiar tanto el signo como la magnitud de las correlaciones, como también ver que factor o factores están asociados a cierto grupo).

APÉNDICE D

Publicaciones generadas por este trabajo

Artificial Neural Network-Based Serum Biomarkers Analysis Improves Sensitivity in the Diagnosis of Lung Cancer

J.M. Flores[1], E. Herrera[1], G. Leal[1], M.G. González[1], F. Sánchez[2], A. Rojas[3], P.A. Cabrera[3], R. Femat[4] and M. Martínez-Velázquez[1]

[1] Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco/Unidad de Biotecnología Médica y Farmacéutica, Guadalajara, Jalisco, México
[2] Hospital Civil de Guadalajara "Fray Antonio Alcalde"/Servicio de Fisiología Pulmonar e Inhaloterapia, Guadalajara, Jalisco, México
[3] Centro Oncológico Estatal ISSEMyM/Coordinación de Cirugía Oncológica, Toluca, México
[4] Instituto Potosino de Investigación Científica y Tecnológica/División de Matemáticas Aplicadas, San Luis Potosí, México

Abstract—**Lung cancer diagnosis in early stages could be of paramount interest since patients may be treated opportunely decreasing the high death rate caused by this disease. A biomarker may describe abnormalities in the human being and may be correlated with a specific illness. Currently, no single biomarker reported has proved to be sufficiently specific and sensitive for lung cancer, thus the search is an open research issue. In this document a set of fourteen biomarkers were evaluated jointly for lung cancer detection, nevertheless, interpreting the information from these biomarkers is a quite complex task and powerful computational tools are required for proper data analysis. Thus an Artificial Neural Network was trained with a set of lung cancer biomarkers. Principal Component Analysis allowed reducing the biomarkers initial vector from fourteen to seven proteins. The Artificial Neural Network performed satisfactorily classifying correctly 60 out of 64 individuals. ANN trained with seven biomarkers -MMP-1, MMP-9, Cyfra 21-1, CRP, CEA, YKL-40, CA-125- yielded an increase in sensitivity of approximately 20%, i.e., 98.97%, compared with that of the best single biomarker, Cyfra 21-1 (sensitivity 78.9%). The corresponding specificity was 80%. ANN significantly improved the sensitivity of biomarkers, therefore ANN offers a promising auxiliary tool in diagnosis of lung cancer.**

Keywords— **Lung Cancer, Biomarker, Artificial Neural Network, Principal Component Analysis, diagnosis.**

I. Introduction

Lung cancer remains a major health problem worldwide, accounting for up to three million deaths annually [1]. Despite efforts and progress made in diagnosis and treatment of lung cancer patients, overall survival at five years of diagnosis is only 15%, because more than 75% patients present with advanced stage of disease when therapeutic options are limited [2]. These data suggest that if patients could be diagnosed at an earlier stage, while tumor is still small and locally defined, complete surgical resection would increases significantly chances of a cure, reducing the mortality associated with this disease [3]. Considerable efforts have been undertaken to produce an effective screening method for early lung cancer detection. Chest radiograph and low-dose helical computed tomography screening can detect early disease, but also produce some false-positive results, repeated radiation to the chest, with its associated risk of carcinogenesis and unnecessary biopsy or surgery of asymptomatic, benign disease [3].

Biomarkers that discriminate cases of lung cancer among those at high-risk for lung cancer, would enhance diagnostic capabilities, complement imaging studies, and have immediate clinical benefit for lung cancer detection.

Single serum markers have shown only low sensitivity and specificity for the identification of lung cancer patients [4]. However, there is a general agreement that a combination of multiple biomarkers may increase diagnostic sensitivity and specificity over use of individual markers [2-6].

Over the last decades Artificial Intelligence has shown to be a potential tool in medicine. Artificial Neural Networks (ANN), Fuzzy Logic and Genetic Algorithms have been used as an auxiliary tool in diagnosis and prognosis of cancer. Fuzzy logic has been used to increase the sensitivity and specificity of biomarkers for diagnosis [6] and prognosis or lung cancer evolution [7]. On the other hand ANN's have played an important role to detect breast

cancer based on morphological characteristics of the cells [8]. A probabilistic neural network has been used to differentiate morphologically malignant cells and benign cells [9]. Moreover, an ANN was applied to automatically detect pulmonary nodules in computed tomography images of chest [10].

In this study we appraised an assortment of biomarkers which previously demonstrated to have diagnostic or prognostic value for lung cancer, as a first step in the effort to improve the diagnostic efficiency of biomarkers and establish a novel multi-analyte serum test for lung cancer detection through the use of principal component analysis and artificial neural networks.

II. Materials and methods

A. Subjects

19 consecutive patients (10 men, 9 women) with newly diagnosed lung cancer, 27 patients (12 men, 15 women) with Chronic Obstructive Pulmonary Disease (COPD) and 17 current smokers with no history of lung disease (9 men, 8 women) were included in this study. Approval was obtained from the corresponding ethics committee and all participants gave written informed consent. Histological diagnosis of primary lung cancer was established according to the revised classification of lung tumors of World Health Organization and the International Association for Lung Cancer Study. In accordance with current GOLD guidelines, COPD was defined by a post-bronchodilator FEV_1/FVC ratio< 0.70. A history of disease, physical examination, spirometric values, arterial blood gas values, electrocardiograms and chest radiograms were obtained from all COPD patients. Samples and health information were labeled using unique identifiers to protect subject confidentiality.

B. Measurement of Serum Biomarker Concentrations

Ten ml blood was collected from each patient in a serum separator tube and processed immediately by centrifugation at 3000 rpm at room temperature for 10 min. Separated serum was aliquoted, and stored at -70º C for future analysis. Levels of matrix *metalloproteinase-1* (MMP-1), matrix metalloproteinase-9 (MMP-9), urokinase-type plasminogen activator (uPA), transferrin (TF), □1-antitrypsin (AT), haptoglobin (HPT), cancer antigen 125 (CA-125), carcinoembryonic antigen (CEA), cytokeratin 19 fragment (Cyfra 21-1), neuron specific enolase (NSE), apolipoprotein A-I (APO A-I), retinol binding protein (RBP), C reactive protein (CRP) and YKL-40 were measured in serum samples with a solid phase sandwich enzyme-linked immunosorbent assay (ELISA), using commercially available human ELISA assays and in accordance to the kits directions: MMP-1 and MMP-9 (R&D Systems; Minneapolis, MN), CA-125, CEA and NSE (ALPCO Diagnostics; Salem, NH), uPA, HPT, TF, □□, CRP, APOA1 and RBP (ASSAYPRO; St. Charles, MO), Cyfra 21-1 (DRG Instruments GmbH; Germany), and YKL-40 (Quidel Corporation; San Diego, CA). Absorbance specified by kits protocols was measured in a microplate spectrophotometer (Bio-Rad Laboratories; Philadelphia, PA). Human recombinant MMP-1, MMP-9, uPA, TF, □□, HPT, CA-125, CEA, Cyfra 21-1, NSE, APO A-I, RBP, CRP and YKL-40 proteins were used as standards. The standard curve was prepared and measured simultaneously with the test samples.

C. Statistical Analysis

Principal Component Analysis (PCA) was run in order to determine those biomarkers that had the greatest influence describing lung cancer cases. The best biomarkers were used to train the ANN. Receiver Operating Characteristic curves were used to describe and compare the accuracy of diagnosis by the combination of biomarkers and the best single biomarker. SigmaPlot™ 10.0 was used to create the ROC curves and the software used to perform the PCA was Simca-P™.

D. Artificial Neural Network

A Multilayered feedforward ANN with one hidden layer was used. The biomarkers data were normalized to a mean zero value and a standard deviation equal to 1. Data were randomly ordered before they were used as input for the ANN. A tangential and linear activation functions were used in the hidden and output layers, respectively. Distinct configurations for the hidden layer(s) were tested (4, 6, 8 and 10) and two hidden layers (2-6, 6-3, 5-5) to determine the best architecture of the ANN. The algorithm used to train the ANN was Levenberg-Marquardt. The data set was divided as follows: 60% for training, 20% for validation and 20% for testing (samples not used during the training stage). To avoid over-fitting 10-fold cross-validation was used. The software used to design the ANN was MATLAB™.

E. Detection Capability Comparison

The sensitivity of the ANN was compared with the best single biomarker by means of ROC curves at a specificity of relevant clinical value.

III. Results and Discussion

Table 1 summarizes demographic and clinical characteristics of patients and control subjects. These groups are not significantly different regarding to age and smoking index.

A Multilayered feedforward ANN was initially trained with all the biomarkers: MMP-1, MMP-9, uPA, TF, AT, HPT, CA-125, CEA, Cyfra 21-1, NSE, APOA1, RBP, CRP and YKL-40. Distinct hidden layers were tested in order to reach the optimal number of layers and neurons, determining that the best configuration was given with 10 neurons in the hidden layer. Using this ANN it was possible to correctly classify 59 out of 64 patients, which represents a determination coefficient of 97.2%, 91.7% and 83.3% for training, validation and testing phases, respectively.

Table 1 Demographic and clinical profiles of patients and controls.

Demographic	Controls (n=44)	Lung Cancer (n=19)	p-value
Age (years)	61.14 ± 15.93	63.68 ± 13.11	0.633
Range (age)	28-92	40-80	-
Female	23	9	-
Male	21	10	-
Active smoking	38	13	-
Passive smoking	1	2	-
Cooking with wood	5	4	-
Smoking index	23.15 ± 25.37	38.16 ± 32.51	0.255
Adenocarcinoma	-	11	-
Epidermoid	-	3	-
SCLC	-	2	-
Carcinoma	-	1	-
Not available	-	2	-

Although the ANN trained with the fourteen biomarkers had a satisfactory performance it was distant to be the best option for lung cancer diagnosis, mainly due to the high costs involved in the biomarkers kits. It is attractive to obtain an accurate diagnosis test at a lower cost. Therefore is desirable to reduce the number of the biomarkers without detriment on the ANN performance. In order to achieve this task a Principal Component Analysis (PCA) was carried out over the whole set of biomarkers. PCA is a multivariate, non-parametric method for extracting relevant information from confusing data sets. The idea behind the PCA analysis was to determine those biomarkers that best represent the majority of lung cancer cases. In figure 1 the PCA shows a marked difference between control and lung cancer patients. In fact cancer patients are located in the right side quadrants, most of them in the inferior square.

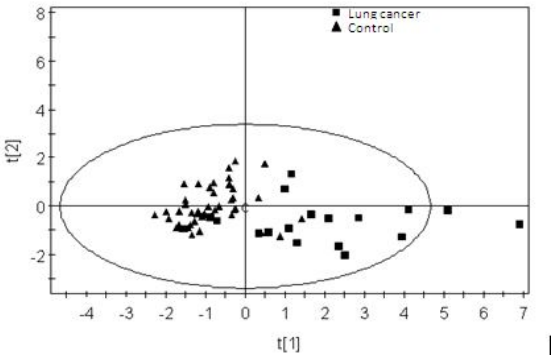

Fig. 1 Distribution of lung cancer and control groups based on Principal Component Analysis

Figure 2 shows the weight of the biomarkers regarding to their principal components. It can be observed that MMP-1, MMP-9, Cyfra-21-1, CRP, CEA, YLK-40 and CA-125 are the proteins which may best describe the cancer group, because these biomarkers have the greatest weights on the first principal component. This means that the initial biomarker vector has been reduced to half, from fourteen to seven biomarkers.

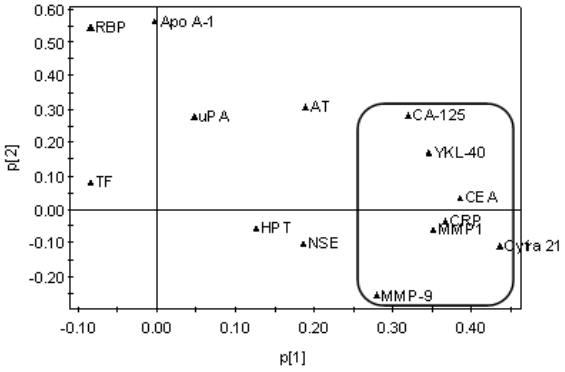

Fig. 2 Principal Component Analysis for the 14 biomarkers

Next, a new ANN was trained with the reduced set of proteins and it was determined if decreasing the biomarkers affect the network performance. A Multilayered feedforward ANN with eight neurons in the hidden layer was chosen as the best configuration in classifying patients (Figure 3).

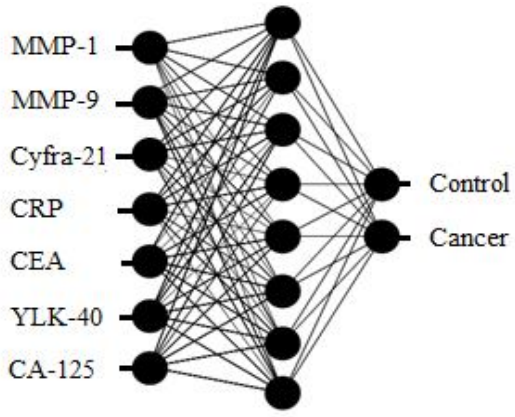

Fig. 3 Artificial Neural Network architecture

The ANN trained with 7 biomarkers (see Fig. 3) classified correctly 60 out of 64 subjects. Thus, the ANN trained with the reduced set of biomarkers performed similar to the former ANN with 14 proteins. Moreover, the classification rate was enhanced since an additional cancer patient was classified correctly. The correct classification rate for patients was 92%, 88% and 86% for training, validation and testing stages, respectively.

Additionally the ROC curves for each single biomarker was determined, resulting Cyfra 21-1 as the best protein since a sensitivity of 78.9% at a specificity of 80% was obtained. The ROC curve of Cyfra 21-1 was compared to ROC curve obtained from the resulting overall performance of the ANN. From combining the markers was possible to increase the sensitivity by 20.07% (sensitivity 98.97%, specificity 80%). The area under the curve for Cyfra 21-1 was 0.853 and for the ANN was 0.915 (See Figure 4).

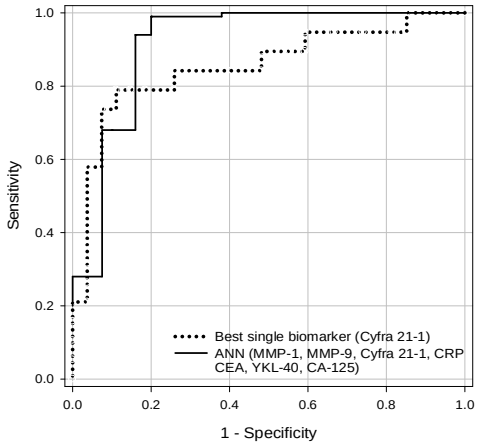

Fig. 4 ROC curves comparing the performance of the best single biomarker, Cyfra 21-1 and the ANN trained with seven biomarkers

IV. Conclusions

Several proteins were evaluated jointly to increase their diagnostic efficacy for lung cancer. An Artificial Neural Network was used as an auxiliary tool to accurately classify lung cancer and healthy patients. Principal Component Analysis allowed reducing the biomarker number to half. The ANN was capable to correctly classify 60 out of 64 subjects. ANN trained with seven biomarkers increased sensitivity by 20.07%, compared with that of the best single biomarker, Cyfra 21-1.

Acknowledgment

This study was supported by CONACYT (Consejo Nacional de Ciencia y Tecnología), Fondo Sectorial de Investigación en Salud y Seguridad Social (FOSISS) 2008-1-87628.

References

1. Cho W (2007) Potentially useful biomarkers for the diagnosis, treatment and prognosis of lung cancer. Biomed Pharmacother 61:515-519
2. Patz E Jr, Campa M, Gottlin E et al. (2007) Panel of serum biomarkers for the diagnosis of lung cancer. J Clin Oncol 25:5578-5583
3. Farlow E, Vercillo M, Coon J et al. (2010) A multi-analyte serum test for the detection of non-small cell lung cancer. Br J Cancer. 103:1221-1228
4. Leidinger P, Keller A, Heisel S et al. (2010) Identification of lung cancer with high sensitivity and specificity by blood testing. Resp Res 11:18
5. Schneider J, Bitterlich N, Velcovsky H et al. (2002) Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer. Int J Clin Oncol 7:145–151
6. Vercillo M, Farlow E, Coon J et al. (2010) A multi-analyte serum test for the early diagnosis of non-small cell lung cancer. J Clin Oncol 28:7s (suppl;abstr 1554)
7. Schneider J, Peltri G, Bitterlich N et al. (2003) Fuzzy logic-based tumor marker profiles improved sensitivity of the detection of progression in small-cell lung cancer patients. Clin Exp Med 2:185–191
8. Karabatak M, Ince M (2009) An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl 36:3465-3469
9. Kıyan T, Yıdırım T (2004) Breast cancer diagnosis using statistical neural networks. J Electric Electron Eng 4:1149-1153
10. Gomathi M, Thangaraj P (2010) A computer aided diagnosis system for detection of lung cancer nodules using extreme learning machine. Int J Eng Sci Tech 2: 5770-5779