

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Doctorado en Ciencias e Ingeniería



Modelo biocomputacional para representar el comportamiento de sistemas biológicos

Presenta:

EDGAR MANZANAREZ OZUNA

Directora:

DRA. DORA LUZ FLORES GUTIÉRREZ

*Con todo el amor que soy capaz de ofrecer y como un esfuerzo por reconocer todo el apoyo
que me han brindado les dedico de forma especial esta tesis a:*

A mi madre, Eloisa

A mi padre, Efren

A mi esposa, Patricia

A mis hijos: Oscar, Christian y Yosadat

Al tío Lupe.

Universidad Autónoma de Baja California

Facultad de Ingeniería, Arquitectura y Diseño

“Modelo biocomputacional para representar el comportamiento de sistemas biológicos”

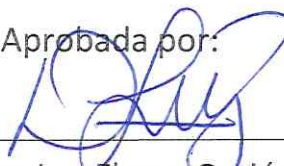
TESIS

que para cubrir parcialmente los requisitos necesarios para obtener el grado de
DOCTOR EN CIENCIAS

Presenta

EDGAR MANZANAREZ OZUNA

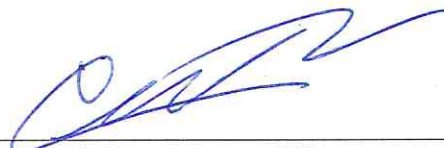
Aprobada por:



Dra. Dora Luz Flores Gutiérrez
Director de Tesis



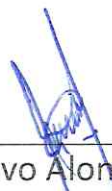
Dr. Everardo Gutiérrez López
Co Director de Tesis



Dr. David Cervantes Vásquez
Miembro del Comité



Dra. Patricia Juárez Camacho
Miembro del Comité



Dr. Gustavo Alonso Hirata Flores
Miembro del Comité

Ensenada, Baja California, mayo de 2018.

Agradecimientos

Para todos y cada uno de aquellos que de alguna forma fueron parte de esta aventura.

Edgar Manzanarez Ozuna

Resumen

La proteína Smad7 es un regulador negativo de la vía de señalización de TGF- β , que está regulada positivamente en pacientes con cáncer de mama. Los miRNA regulan las expresiones de proteínas deteniendo o degradando los mRNA. El propósito de este trabajo es identificar un perfil de miARN que regula la expresión del ARNm que codifica Smad7 en el cáncer de mama utilizando los datos de pacientes con cáncer de mama obtenidos del Proyecto Atlas del Genoma del Cáncer. Desarrollamos un método de búsqueda automático basado en algoritmos genéticos para encontrar un modelo predictivo basado en redes neuronales profundas que se ajuste al conjunto de datos biológicos y aplicar el algoritmo Olden para identificar la importancia relativa de cada miRNA. Se muestra un modelo computacional de regresión no lineal, basado en redes neuronales profundas que predicen la regulación dada por las transcripciones de miARN diana. ARNm que codifica la proteína Smad7 en pacientes con cáncer de mama, con R2 de 0.99 y MSE de 0.00001. Además, el modelo se valida con los resultados en experimentos in vivo e in vitro informados en la literatura. El conjunto de miRNAs hsa-mir-146a, hsa-mir-93, hsa-mir-375, hsa-mir-205, hsa-mir-15a, hsa-mir-21, hsa-mir-20a, hsa-mir-503, hsa-mir-29c, hsa-mir-497, hsa-mir-107, hsa-mir-125a, hsa-mir-200c, hsa-mir-212, hsa-mir-429, hsa-mir-34a, hsa-let-7c, hsa-mir-92b, hsa-mir-33a, hsa-mir-15b, hsa-mir-224, hsa-mir-185 y hsa-mir-10b integran un perfil que regula críticamente la expresión de el ARNm que codifica Smad7 en el cáncer de mama. Aunque la confirmación de los resultados por experimentos de laboratorio no ha ocurrido, los resultados permiten sugerir que el perfil de

miRNAs podría usarse como biomarcadores u objetivos en terapias dirigidas

Abstract

The Smad7 protein is negative regulator of the TGF- β signaling pathway, which is up-regulated in patients with breast cancer. miRNAs regulate proteins expressions by arresting or degrading the mRNAs. The purpose of this work is to identify a miRNAs profile that regulates the expression of the mRNA coding for Smad7 in breast cancer using the data from patients with breast cancer obtained from the Cancer Genome Atlas Project. We develop an automatic search method based on Genetic Algorithms to find a predictive model based on deep neural networks which fit the set of biological data and apply the Olden algorithm to identify the relative importance of each miRNAs. A computational model of non-linear regression is shown, based on deep neural networks that predict the regulation given by the miRNA target transcripts mRNA coding for Smad7 protein in patients with breast cancer, with R2 of 0.99 is shown and MSE of 0.00001. In addition, the model is validated with the results in vivo and in vitro experiments reported in the literature. The set of miRNAs hsa-mir-146a, hsa-mir-93, hsa-mir-375, hsa-mir-205, hsa-mir-15a, hsa-mir-21, hsa-mir-20a, hsa-mir-503, hsa-mir-29c, hsa-mir-497, hsa-mir-107, hsa-mir-125a, hsa-mir-200c, hsa-mir-212, hsa-mir-429, hsa-mir-34a, hsa-let-7c, hsa-mir-92b, hsa-mir-33a, hsa-mir-15b, hsa-mir-224, hsa-mir-185 and hsa-mir-10b integrate a profile that critically regulates the expression of the mRNA coding for Smad7 in breast cancer. Although the confirmation of the results by laboratory experiments has not occurred, the results allow suggesting that miRNAs profile could be used as biomarkers or targets in targeted therapies.

Índice general

1. Introducción	2
1.1. Cáncer	3
1.2. Cáncer de mama	3
1.2.1. El factor de crecimiento transformante $\beta 1$	4
1.2.2. Familias de proteínas SMADs	4
1.2.3. La vía de señalización del factor de crecimiento transformante β . . .	5
1.2.4. Los micro ácido ribonucleicos (miRNAs)	5
1.2.5. Antecedentes	6
1.2.6. Regulación de proteína Smad7 dada por miRNAs	8
1.3. Modelado biocomputacional	9
1.3.1. La complejidad como enfoque de investigación	9
1.3.2. Aprendizaje automático	10
1.3.3. Computación natural	11
1.3.4. Modelo biocomputacional	12
1.3.5. Antecedentes	12

1.4. Planteamiento del problema	12
1.5. Preguntas de investigación	13
1.6. Objetivo general	13
1.6.1. Objetivos específicos	13
1.7. Hipótesis	14
1.8. Justificación	14
1.9. Organización del documento	17
2. Marco teórico	18
2.1. Los seres vivos y su interacción con el entorno	18
2.2. La célula como elemento fundamental de vida	20
2.2.1. Las vías de señalización celular	22
2.2.2. El ciclo celular	22
2.2.3. La apoptosis	23
2.2.4. El dogma central de la biología molecular	24
2.3. Aprendizaje automático	24
2.3.1. Regresión lineal	29
2.4. Preprocesamiento de datos	30
2.4.1. Transformada wavelets	30
2.4.2. Multicolinealidad	32
2.4.3. Escalamiento de datos	32
2.5. Técnicas de computación natural	33

2.5.1. Redes neuronales artificiales	33
2.5.2. Perceptrón multicapas	35
2.5.3. Redes neuronales profundas (DNN)	36
2.5.4. Algoritmos genéticos	37
2.5.5. Importancia relativa Olden	39
3. Metodología	40
3.1. Recopilación y tratamiento de datos	40
3.1.1. Conjunto de datos	40
3.1.2. Preprocesamiento de datos	41
3.2. Diseño del modelo de predicción	43
3.2.1. Método de búsqueda evolutivo	43
3.2.2. Algoritmo genético	44
3.2.3. Redes neuronales profundas	45
4. Resultados y validación del modelo	46
4.1. Evolución del método EMDAV	46
4.2. Modelo predictivo basado en DNN	47
4.3. Importancia relativa de miRNAs con mRNAs-Smad7	48
5. Conclusiones y perspectivas	52

Índice de figuras

1.1. Lazo de regulación negativa TGF- β 1 y Smad7 en la vía de señalización del TGF- β	6
1.2. Regulación de mRNA-Smad7 por miRNAs	8
1.3. Los diez cánceres con mayor incidencia en el mundo en 2012. [1]	15
1.4. Los diez cánceres con mayor incidencia en México en 2012. [1]	16
2.1. Descripción abstracta de una célula	21
2.2. Vía de señalización celular	22
2.3. El ciclo celular	23
2.4. Dogma central de la biología molecular.	24
2.5. Modelo genérico de aprendizaje automático	25
2.6. Capacidad de algoritmo de aprendizaje automático [27]	28
2.7. Relación típica entre la capacidad y el error [27]	28
2.8. Descripción de las neuronas biológicas.	33
2.9. Descripción de las neuronas artificiales y su relación con las neuronas biológicas.	34
2.10. Arquitectura de ANN con múltiples capas y alimentación directa.	36

2.11. Operadores genéticos. Proceso de cruzamiento en un punto (a) y proceso de mutación en un punto (b).	38
3.1. La lógica de la simulación como método de investigación.	41
3.2. Método evolutivo del conjunto de datos,arquitectura y validación (EMDAV).	43
4.1. Modelo predictivo basado en DNN	48

Índice de tablas

4.1. Los mejores modelos predictivos basados en DNN	47
4.2. Conjunto de datos evolucionado del modelo predictivo	48
4.3. Importancia relativa de miRNAS que hacen blanco en mRNA-Smad7	51

Capítulo 1

Introducción

Los seres vivos se integran en una serie de niveles de organización, cada uno de los cuales constituyen los cimientos del siguiente nivel. Partiendo desde el nivel básico integrado por los bio-elementos hasta el nivel más alto integrado por el conjunto de seres vivos y componentes inertes que comprenden el planeta tierra, llamado biosfera [2]. Llevan a cabo funciones de reproducción, crecimiento y desarrollo, homeóstasis y organización.

La reproducción es la producción de descendientes, el crecimiento es el aumento de la cantidad de materia viva en un organismo, el desarrollo es la serie de cambios que un organismo experimenta para alcanzar su forma final de adulto, la homeóstasis consiste en mantener las condiciones apropiadas para el funcionamiento interno del organismo compensando los cambios en su entorno mediante el intercambio regulado de materia y energía y la organización establece que todas las funciones o procesos vitales son fundamentales para la existencia e interactúan en conjunto para crear un sistema vivo, único y ordenado [3].

Los organismos vivos están formados por células que son unidades pequeñas rodeadas de una membrana que contiene una solución acuosa concentrada de sustancias químicas, denominadas unidades fundamentales de la vida y cuentan con procesos y mecanismos para

comunicarse, nutrirse y reproducirse [4, 5].

1.1. Cáncer

En tejidos normales existe un balance entre la generación de células nuevas mediante división celular y la pérdida de células a través de la muerte. Las células viejas son dañadas a través del tiempo y son eliminadas. Sin embargo, en el cáncer este proceso ordenado pierde el control. A medida que las células se hacen cada vez más anormales las células viejas o dañadas sobreviven cuando deberían morir y se generan células nuevas cuando no son necesarias.

Estas células adicionales pueden dividirse sin interrupción y pueden formar masas que se llaman tumores que se pueden extender o invadir a los tejidos cercanos. Además, al crecer estos tumores, algunas células cancerosas pueden desprenderse y moverse a lugares distantes del cuerpo por medio del sistema circulatorio o del sistema linfático y formar nuevos tumores lejos del tumor original [6, 7].

1.2. Cáncer de mama

El cáncer de mama hace referencia a un tumor maligno que se ha desarrollado a partir de células mamarias. Generalmente, se origina en las células de las glándulas productoras de leche, en los conductos que transportan la leche hasta el pezón y con menos frecuencia, en los tejidos conjuntivos grasos y fibrosos de la mama. Con el paso del tiempo, las células cancerosas pueden invadir el tejido mamario sano circundante y llegar a los ganglios linfáticos (pequeños órganos que eliminan sustancias extrañas del cuerpo) de las axilas y obtienen una puerta de acceso hacia otras partes del cuerpo [8].

El cáncer de mama siempre se origina por una anomalía genética. No obstante, solo de

un 5 % a 10 % de los casos son producto de una anomalía heredada de la madre o el padre. En cambio, del 85 % a 90 % de los casos de cáncer de mama tienen su origen en anomalías genéticas vinculadas al proceso de envejecimiento y el "desgaste natural" de la vida [8].

1.2.1. El factor de crecimiento transformante $\beta 1$

La superfamilia de factores de crecimiento del TGF- β es un grupo que comprende una gran cantidad de proteínas capaces de regular varios procesos celulares incluyendo proliferación, diferenciación, migración, invasión y angiogénesis. Estas proteínas ejercen sus efectos en todas las células y su patrón de expresión es temporal y específico del tejido [9].

La familia del TGF β se compone de una citocina con el mismo nombre que presenta tres isoformas: TGF- $\beta 1$, TGF- $\beta 2$ y TGF- $\beta 3$. Son estructuralmente similares y entre sus funciones destacan la inducción del arresto del ciclo celular en células hematopoyéticas y epiteliales, el control de la proliferación y la diferenciación de células mesenquimales, el proceso de transición epitelio-mesénquima (EMT), la cicatrización, la producción de matriz extracelular y la inmunosupresión [9].

El factor de crecimiento transformante beta 1 (TGF- $\beta 1$) se encuentra sobreexpresado en el cáncer de mama [10] y regula estos procesos mediante la iniciación de la vía de señalización celular TGF- β , la cual induce la expresión génica y procesos celulares como respuesta a estímulos en la parte exterior de la célula [9, 11].

1.2.2. Familias de proteínas SMADs

Las Smads (mothers against decapentaplegic) son los sustratos del receptor β tipo I (T β RI) y tienen un papel clave en la transducción de la señal desde los receptores a los genes blanco en el núcleo [9]. Basados en sus características estructurales y funcionales se dividen en tres

categorías: R-Smads (Smad2 y Smad3), que son reguladas y activadas por los receptores y son las primeras en transducir el mensaje intracelularmente después de ser activadas por $TR\beta I$; Smads cooperadoras (Smad4), se asocian con la R-Smads para formar el complejo SMAD; I-Smads: (Smad6 y Smad7), que son antagonistas de la vía [11]

1.2.3. La vía de señalización del factor de crecimiento transformante β

La vía de señalización del TGF- β inicia con la unión del ligando TGF- $\beta 1$ activado al receptor II ($T\beta RII$) el cual recluta al ($T\beta RI$) para formar el complejo receptor dependiente del ligando activado, y activa mediante fosforilación a $T\beta RI$, el $T\beta RI$ activado, fosforila a las proteínas R-Smads que en conjunto con las Smad4 integran el complejo SMAD que se interna al núcleo y en asociación con proteínas de unión al DNA actúan como factores de transcripción que regulan la expresión de genes objetivo [9].

La proteína Smad7 interrumpe el proceso de transducción de la vía de señalización del TGF- β mediante la degradación proteosomal del complejo receptor dependiente del ligando activado, previniendo la fosforilación de las R-Smads, evitando la formación de complejos SMAD funcionales y bloqueando la unión del complejo SMAD al DNA [9].

Lo anterior, establece un lazo de regulación negativa entre el TGF- $\beta 1$ como promotor y Smad7 como inhibidor de la vía de señalización del TGF- β , Ver Figura 1.1.

1.2.4. Los micro ácido ribonucleicos (miRNAs)

Por otra parte, los micro ácido ribonucleicos (miRNAs) son pequeños RNAs no codificantes de longitud entre 21-25 nucleótidos implicados en la regulación de la división celular, desarrollo, oncogénesis, apoptosis, entre otros procesos mediante el bloqueo de la traducción

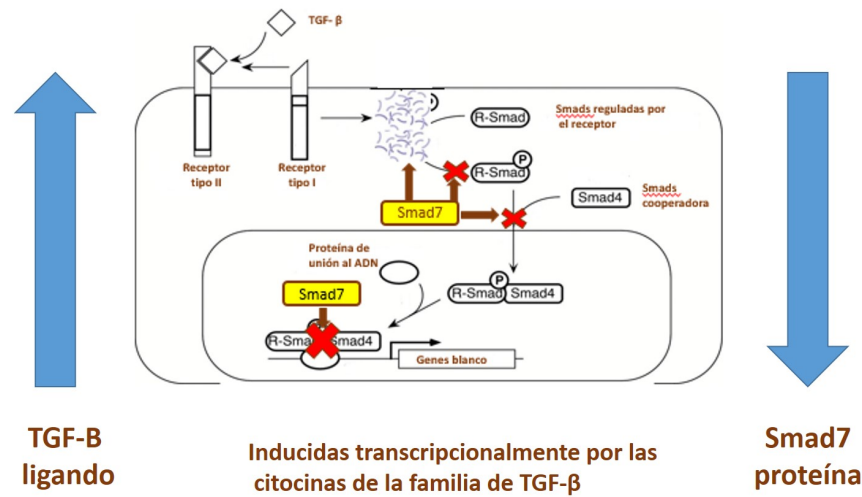


Figura 1.1: Lazo de regulación negativa TGF- β 1 y Smad7 en la vía de señalización del TGF- β .

de las proteínas y la degradación de transcritos de ácido ribonucleicos mensajeros (mRNA). Los miRNAs se transcriben mediante RNA polimerasa II como parte de transcritos primarios poliadenilados y protegidos (pri-miRNA) que pueden ser de codificación proteica o no codificante. El transcrito primario se escinde (segmenta) mediante la enzima Drosha ribonucleasa III para producir un miRNA precursor de bucle de tallo de aproximadamente 70 nt (pre-miRNA), que se escinde (segmenta) adicionalmente mediante la ribonucleasa dicer citoplásmica para generar el miRNA maduro y la estrella de miRNA antisentido (miRNA*). El miRNA maduro se incorpora a un complejo de silenciamiento inducido por RNA (RISC), que reconoce los mRNAs objetivos a través del emparejamiento de bases imperfecto o perfecto con el miRNA [12].

1.2.5. Antecedentes

Con el propósito de entender y estudiar los sistemas biológicos se han desarrollado técnicas de experimentación, entre las que se encuentran: *in vitro* donde los experimentos son desarrollados a partir de muestras biológicas en ambientes controlados fuera de organismos

vivos, *in vivo* los experimentos son inoculados en organismos vivos y completos, *in silico* donde los experimentos son abstracciones de sistemas, modelados en programas de computadora que se ejecutan en tiempo no real, *in virtuo* los experimentos son abstracciones de sistemas, modelados en programas de computadora que se ejecutan en tiempo no real con la capacidad de modificar las variables de comportamiento del modelo en tiempo de ejecución [13].

Existen distintos tipos de tratamientos contra el cáncer de mama, tales como la cirugía, quimioterapia, terapia de radiación, hormonoterapia, terapias dirigidas, entre otras. Algunos de carácter invasivo, sistémicos, locales y demás, que apoyan en la lucha contra esta enfermedad [8].

En particular, las terapias dirigidas son tratamientos que abordan características específicas de las células cancerosas, por ejemplo, una proteína que permite que las células cancerosas se multipliquen rápidamente o de forma anormal. Generalmente, es menos probable que estas terapias dañen las células sanas y normales. Algunas terapias dirigidas consisten en anticuerpos que funcionan de la misma manera que los anticuerpos producidos naturalmente por el sistema inmunológico. A veces, estos tipos de terapias dirigidas se denominan inmunoterapias dirigidas [8]. En los últimos años se han enfocado los esfuerzos en la búsqueda y el uso de fármacos u otras sustancias que bloqueen el crecimiento y diseminación del cáncer interfiriendo con moléculas específicas, involucradas en el crecimiento y progresión tumoral del cáncer de mama [14], entre las que se encuentran por ejemplo, Herceptin (nombre genérico: trastuzumab) ataca los tumores de cáncer de mama positivos para HER2 bloqueando la capacidad de las células de cáncer de recibir las señales químicas que les indican que deben multiplicarse, Ibrance (nombre genérico: palbociclib) es un inhibidor de cinasas dependientes de ciclinas 4/6, impide la división y el crecimiento de las células cancerígenas, entre otros [8].

Existe evidencia reportada sobre la relación de los miRNAs y las proteínas inhibidoras de la vía de señalización de TGF- β . La sobreexpresión de miR-21 o la bajaexpresión de

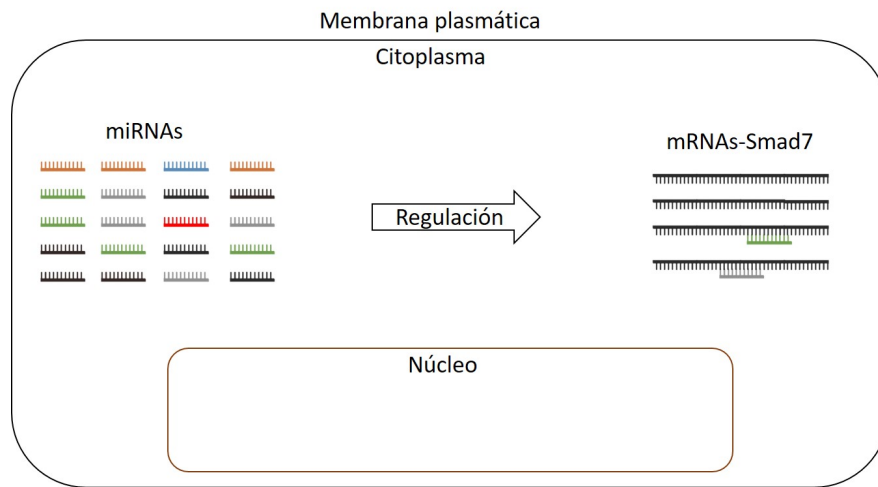


Figura 1.2: Regulación de mRNA-Smad7 por miRNAs

Smad7 promueve la formación de fibroblasto asociado a carcinoma [15]. La sobreexpresión de miR-21 puede inhibir la proliferación de células epiteliales tubulares renales de rata [12]. El miR-21 promueve la proliferación de células de cáncer de mama y la invasión mediante la supresión de Smad7 [16]. En [17] se observó que nueve miRNAs se sobreexpresaron en más de dos veces y siete miRNAs se bajo expresaron en más de dos veces en relación al tejido normal adyacente al tumor. En [18] menciona que la expresión de miR-21, miR-210 y miR-221 tiene un papel significativo en el desarrollo del cáncer de mama triple negativo primario.

1.2.6. Regulación de proteína Smad7 dada por miRNAs

Todo lo antes mencionado sugiere la existencia de regulación de la expresión de la proteína Smad7 dada por parte de los miRNAs sobre el mRNA que codifica para la proteína Smad7 (mRNA-Smad7) descrita en la Figura 1.2, y el posible impacto en el lazo de regulación negativa de la vía de señalización del TGF- β mediada por TGF β 1 y Smad7. Dado que múltiples miRNAs pueden unirse a múltiples transcritos mRNA en forma paralela, para este proceso se observa dinámica no lineal y es factible abordarlo desde el punto de vista de la complejidad.

1.3. Modelado biocomputacional

El Programa de Investigación Intramural (IRP, por sus siglas en inglés) en [19] define el modelado biológico computacional, o biología de sistemas como una simulación computarizada de un sistema biológico utilizado para comprender y predecir las interacciones dentro de ese sistema.

1.3.1. La complejidad como enfoque de investigación

A través de los tiempos se han enfrentado tres tipos de problemas, de dos variables denominados problemas sencillos, los problemas de complejidad desorganizada, caracterizados por billones de variables en el cual el comportamiento del total de sus variables era desconocido sin embargo el sistema concebido como un todo poseía un nivel de orden que permitía analizar sus propiedades, abordados mediante la teoría de probabilidad y la mecánica estadística, y los problemas de complejidad organizada que implican tratar simultáneamente con un número considerable de factores que están interrelacionados en un todo orgánico [20].

Sistemas complejos adaptativos

Los sistemas complejos pueden ser definidos como fenómenos compuestos por una gran cantidad de componentes simples que interactúan entre si y son capaces de intercambiar información con su entorno. Como resultado, bajo algunas circunstancias, pueden emerger nuevos comportamientos. Dichos comportamientos no pueden ser observados al analizar las partes o subsistemas de manera independiente; son el producto de la acción colectiva [21, 22, 23].

Los sistemas complejos adaptativos (CAS, por sus siglas en inglés) son sistemas complejos con características relacionadas a cambios en el tiempo y aprendizaje con base en la

experiencia [21, 22, 23].

Los sistemas biológicos son caracterizados como CAS, dado que son sistemas abiertos organizados en múltiples capas de complejidad creciente integradas por la interrelación, interconexión e interacción, espacio temporal no lineal entre un gran número de componentes diversos de las capas de menor complejidad. Evolucionan, se adaptan y se transforman a través de interacciones dinámicas internas, y externas con entornos inestables y cambiantes que afectan a los subsistemas y componentes tanto en la escala local como global, de los cuales emergen comportamientos complejos [24].

Dado lo anterior, el enfoque de investigación de la complejidad fundamenta que para comprender los sistemas, es necesario partir de los postulados que dictan que el organismo en su conjunto es la entidad fundamental en la existencia biológica, los fenómenos no son el resultado de una cadena causa-efecto en la que el resultado puede ser proporcional y predecible sólo con el conocimiento de las partes, y los sistemas se comportan de manera no lineal. A partir de este conocimiento es posible entender la estructura y el orden emergentes [25].

De igual forma, existe evidencia que en los últimos años la simulación por computadora [22] está ganando terreno como metodología de estudio de los CAS, en relación con los enfoques experimentales y teóricos.

1.3.2. Aprendizaje automático

Tom Mitchell en 1997 [26] establece "Que un programa de computadora aprende de la experiencia E con respecto a una clase de tareas T y una medida de rendimiento P , si el desempeño en tareas en T , medido por P , mejora con la experiencia E ".

En particular, en la clase de tareas de tipo regresión, el programa debe ser capaz de predecir un valor de salida dada una entrada. Para lo cual, el algoritmo de aprendizaje

ajustará una función $f : \mathbb{R}^n \rightarrow \mathbb{R}$ con base en un conjunto de datos integrado por los valores de entrada relacionados con valores de salida, el desempeño es medido en relación a el error entre el estimador y lo estimado, y la bondad del ajuste del modelo a la variable que pretender explicar [27].

1.3.3. Computación natural

El campo de la computación natural se caracteriza por el uso metafórico de conceptos, principios y mecanismos de los sistemas naturales, engloba un conjunto de modelos que tienen como característica común la simulación del modo en que la naturaleza actúa y estudia la forma en que las diversas leyes de la naturaleza producen modificaciones en determinados sistemas. Se divide en tres áreas principales: la computación inspirada en la naturaleza para el desarrollo de nuevas técnicas de resolución de problemas, mediante el desarrollo de algoritmos para la solución de problemas complejos. La simulación y emulación de la naturaleza por medio de la computación que consiste un proceso sintético dirigido a crear patrones, formas, comportamientos y organismos que no necesariamente se asemejan a la vida como la conocemos, sus productos se pueden utilizar para imitar diversos fenómenos naturales, aumentando así nuestra comprensión de la naturaleza y las ideas sobre los modelos de computadora. Computación con materiales naturales corresponde al uso de nuevos materiales naturales para realizar la computación, constituyendo así nuevos paradigmas de computación que vienen a sustituir o complementar las actuales computadoras basadas en silicio [28] .

En el área de la computación inspirada en la naturaleza se encuentran las técnicas de computación evolutiva, redes neuronales artificiales, inteligencia de enjambre, sistema inmune artificial, sistemas basados en agentes, entre otras y permiten abordar a los CAS bajo los postulados del enfoque de investigación de la complejidad.

1.3.4. Modelo biocomputacional

Con base en todo lo anterior, en esta investigación se considera al modelo biocomputacional como un programa de computadora que resuelve problemas de complejidad organizada de los sistemas biológicos, aprende en forma automática e incorpora técnicas de computación inspiradas en la naturaleza.

1.3.5. Antecedentes

A través del tiempo, se han generado modelos computacionales para experimentación *in silico* e *in virtuo* [29], basados en el enfoque teórico, mediante técnicas tales como: ecuaciones diferenciales ordinarias, ecuaciones diferenciales estocásticas, formalismos *s-system*, ecuaciones diferenciales parciales, redes de Petri, autómatas celulares, modelos basados en agentes, cálculo pi, entre otras. Cada una de las técnicas aporta ventajas y desventajas en casos específicos de sistemas biológicos [30].

1.4. Planteamiento del problema

A partir de lo anterior, se identifica la necesidad de generar modelos biocomputacionales predictivos con el propósito de representar la relación de regulación dada por los miRNAs sobre el mRNA-Smad7 en el cáncer de mama e identificar blancos con capacidad terapéutica que permitan proponer el desarrollo de terapias dirigidas para la cura de esta malignidad.

1.5. Preguntas de investigación

Dada la relevancia de la problemática planteada es preciso instaurar ejes rectores que definan el aporte y el alcance de esta investigación, para lo cual se generan las siguientes preguntas de investigación.

¿Existe un conjunto reducido de miRNAs que contribuyen en gran medida en la relación de regulación dada por los miRNAs sobre el mRNA-Smad7 en el cáncer de mama?

¿Cuáles miRNAs son estadísticamente los más significativos en la regulación del mRNA-Smad7 y representan potencial terapéutico para el desarrollo de terapias dirigidas para el cáncer de mama?

¿Un modelo biocomputacional es capaz de ajustar un función con base a un conjunto de datos sobre la relación de regulación del mRNA-Smad7 dada por los miRNA en el cáncer de mama?

1.6. Objetivo general

Con el propósito de dar respuesta a las preguntas de investigación se establece como objetivo general el desarrollo de un modelo biocomputacional para predecir la relación de regulación dada por los miRNAs sobre el mRNA-Smad7 en el cáncer de mama.

1.6.1. Objetivos específicos

1. Caracterizar la relación de regulación dada por los miRNAs sobre el mRNA-Smad7 en el cáncer de mama.
2. Obtener el conjunto de datos de interacción y expresión de miRNAs y mRNA-Smad7.

3. Determinación de técnicas de computación natural como base para el desarrollo del modelo biocomputacional.
4. Diseño del modelo biocomputacional que represente la relación de regulación dada por los miRNAs sobre el mRNA-Smad7 en el cáncer de mama.
5. Validación del modelo con base a resultados reportados en la literatura.

1.7. Hipótesis

El modelo biocomputacional ajusta al conjunto de datos sobre la relación de regulación del mRNA-Smad7 dada por los miRNA en el cáncer de mama y es capaz de determinar el conjunto de miRNAs que contribuyen estadísticamente en gran medida a la regulación post transcripcional del mRNA-Smad7 en cáncer de mama.

1.8. Justificación

El cáncer de mama es el más común entre las mujeres en todo el mundo, pues representa el 16 % de todos los cánceres femeninos. Se estima que en 2004 murieron 519 000 mujeres por cáncer de mama y, aunque este cáncer está considerado como una enfermedad del mundo desarrollado, la mayoría (69 %) de las defunciones por esa causa se registran en los países en desarrollo [31]

En 2012 el cáncer de mama continua siendo el más común entre las mujeres de todo el mundo, representa el 18.43 % de todos los cánceres excepto el no-melanoma de piel. A nivel mundial en 2012 se estima la tasa anual de incidencia en 47.8 por cada 100 000 mujeres, y la tasa de mortalidad anual en 14.9 por cada 100 000 mujeres [1]. Tal como se observa en la Figura 1.3

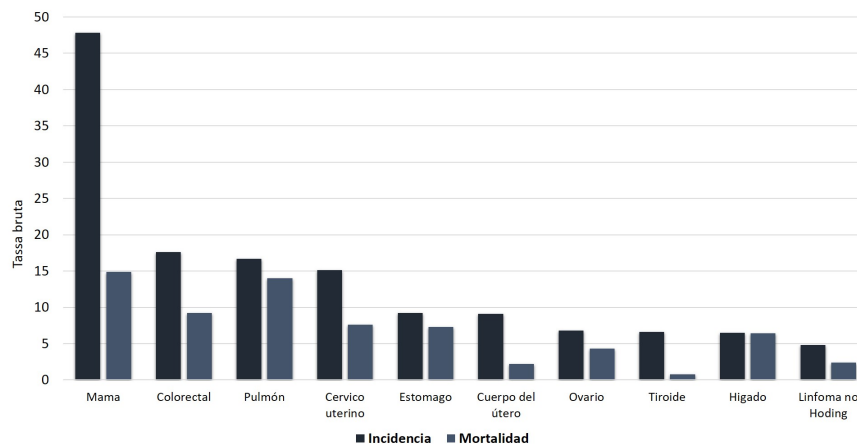


Figura 1.3: Los diez cánceres con mayor incidencia en el mundo en 2012. [1]

La incidencia varía mucho en todo el mundo, con tasas normalizadas por edad de hasta 99.4 por 100 000 en América del Norte. Europa oriental, América del Sur, África austral y Asia occidental presentan incidencias moderadas, pero en aumento. La incidencia más baja se da en la mayoría de los países africanos, pero también en ellos se observa un incremento de la incidencia de cáncer de mama [31].

Las tasas de supervivencia del cáncer mamario varían mucho en todo el mundo, desde el 80 % o más en América del Norte, Suecia y Japón, pasando por un 60 % aproximadamente en los países de ingresos medios, hasta cifras inferiores al 40 % en los países de ingresos bajos [31].

Las tasas bajas de supervivencia observadas en los países poco desarrollados pueden explicarse principalmente por la falta de programas de detección precoz que hace que un alto porcentaje de mujeres acudan al médico con la enfermedad ya muy avanzada, pero también por la falta de servicios adecuados de diagnóstico y tratamiento [31].

De forma correspondiente con los datos a nivel mundial, en México, en 2012 el cáncer de mama continua siendo el más común entre las mujeres del país, representa el 18.17 % de todos los cánceres excepto el no-melanoma de piel. La tasa anual estimada de incidencia en México en el 2012 es de 34.7 por cada 100 000 mujeres, y la tasa de mortalidad anual en el

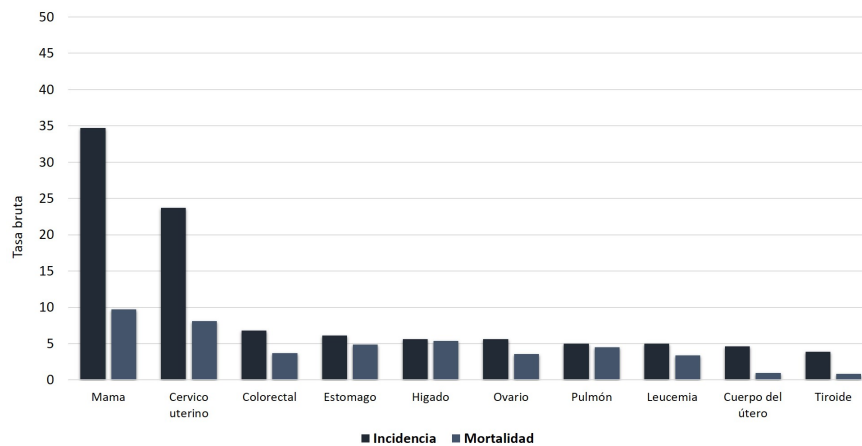


Figura 1.4: Los diez cánceres con mayor incidencia en México en 2012. [1]

2012 en 14.9 por cada 100 000 mujeres [1]. Tal como se observa en la Figura 1.4

De manera general, en México las tasas de mortalidad observadas por tumor maligno de mama han tenido una tendencia a la alza en el periodo de 2011 a 2016; en las mujeres hay una mayor concentración de muertes por esta causa en comparación con los varones. Es en 2016 donde se ubica la tasa más alta para ellas, con 16 defunciones por cada 100 000 mujeres de 20 años y más. Las tasas de mortalidad por esta enfermedad entre los varones son muy bajas; sin embargo, estos datos evidencian la presencia de este padecimiento también en ellos [32].

Entender el incremento en la progresión del cáncer de mama y los efectos de tratamientos a través de modelos biocomputacionales causa gran interés en la industria de los fármacos y la investigación farmacéutica sobre la esperanza de que estas herramientas de simulación emergentes apoyen a que el proceso de descubrimiento y prueba de fármacos sea mejor, más rápido y más barato [29].

Los analistas pronostican que el mercado global de biosimulación crezca a una tasa compuesta anual de 18.31 % durante el período 2013-2018. En 2022, se espera que el mercado para la biosimulación alcance 3,435 millones de dólares, 1,476 millones de dólares más de los alcanzados en 2017. Uno de los factores clave que contribuyen a este crecimiento del mer-

cado es la reducción en los costos de descubrimiento y desarrollo de fármacos, debido a la implementación de soluciones de biosimulación [33].

Es por esto, la relevancia de desarrollar modelos biocomputacionales predictivos que permitan entender la dinámica del cáncer de mama y la relación entre sus componentes y el entorno.

1.9. Organización del documento

El segundo capítulo lo integra el marco teórico donde se describen los sistemas biológicos y el conjunto de técnicas aplicadas, el tercer capítulo detalla la metodología aplicada a la regulación del mRNA-Smad7 dada por miRNAs, en el capítulo cuarto se muestran los resultados y la validación del modelo, y por último, las conclusiones y perspectivas se describen en el quinto capítulo.

Capítulo 2

Marco teórico

En este capítulo se abordan los fundamentos teóricos de los sistemas biológicos y del proceso de búsqueda en un espacio de soluciones, la teoría y técnicas de aprendizaje automático y computación natural, técnicas de preprocesamiento de datos y el algoritmo de importancia relativa desarrollado por Olden.

2.1. Los seres vivos y su interacción con el entorno

Hasta el momento, de la integración de las partículas fundamentales, los *quarks* y los *leptons*, se originan los electrones, protones y neutrones emergiendo el primer nivel de organización denominado subatómico. El nivel atómico es compuesto de átomos de carbono, fósforo, nitrógeno, hidrógeno, oxígeno, azufre, entre otros, denominados bioelementos que son integrados y organizados por elementos subatómicos. Como resultado de la unión de bioelementos emergen moléculas de carbohidratos, proteínas, lípidos, agua, sales, entre otras. Forman el nivel molecular y tienen características para producir y conserva energía, estructurales, térmicas, enzimáticas, de defensa y de comunicación [2].

Las células son unidades pequeñas rodeadas de una membrana que contiene una solución acuosa, concentrada de sustancias químicas compuesta por elementos de los niveles de organización anteriores. Denominadas unidades fundamentales de la vida con procesos y mecanismos para comunicarse, nutrirse y reproducirse. La teoría celular establece que todos los organismos vivos están formados por células. A partir de la organización de las células y de elementos de los niveles anteriores emerge el nivel tisular integrado por los distintos tipos de tejidos. Los tejidos en conjunto con elementos de los niveles anteriores dan forma, estructura y funcionalidad a los órganos. Éstos a su vez se relacionan con propósitos específicos emergiendo el nivel de aparatos o de sistemas funcionales. A partir de lo anterior se llega al nivel de organización superior en el cual las células, los tejidos, los órganos y los aparatos de funcionamiento forman a los seres vivos [2].

Al conjunto de organismos de la misma especie que forman un núcleo y coinciden en el tiempo y el espacio se le denomina población. Las comunidades se forman del conjunto de poblaciones de seres vivos diferentes de un lugar. La interacción con el medio físico por parte de las comunidades se le conoce como ecosistema. El conjunto ecosistemas diferentes dentro de una determinada superficie se le denomina paisaje, los cuales agrupados en una superficie geográfica forman una región. Los ecosistemas de gran tamaño asociados a unas determinadas características ambientales y el dominio de una especie aunque no sea homogéneo, son identificados como bioma. Hasta llegar a la biósfera que es todo el conjunto de seres vivos y componentes inertes que comprenden el planeta tierra, o de igual modo es la capa de la atmósfera en la que existe vida y que se sustenta sobre la litósfera [2].

Los seres vivos se integran en niveles que van de menor a mayor organización y complejidad. Los niveles superiores emergen como resultado de la interrelación, interconexión e interacción en forma no lineal en las dimensiones espacio y tiempo, entre un gran número de componentes de los distintos niveles de organización inferiores y el entorno, y llevan a cabo funciones de reproducción, crecimiento y desarrollo, homeóstasis y organización [2].

La reproducción es la producción de descendientes, el crecimiento es el aumento de la cantidad de materia viva en un organismo, el desarrollo es la serie de cambios que un organismo experimenta para alcanzar su forma final de adulto, la homeóstasis consiste en mantener las condiciones apropiadas para el funcionamiento interno del organismo compensando los cambios en su entorno mediante el intercambio regulado de materia y energía y la organización establece que todas las funciones o procesos vitales son fundamentales para la existencia e interactúan en conjunto para crear un sistema vivo único y ordenado [3].

2.2. La célula como elemento fundamental de vida

La célula es la unidad básica funcional y estructural más pequeña de los organismos vivos. Se compone de partes características, cuyo trabajo es coordinado de tal forma que cada tipo de célula lleva a cabo una función estructural o bioquímica única. Las células llevan a cabo numerosas reacciones químicas para dar origen al proceso vital mediante estructuras especializadas dentro de la misma, las cuales están coordinadas unas con otras para mantener con vida tanto a la célula como a los tejidos, los órganos, los sistemas y todo el organismo [4].

Las células efectúan diversas funciones básicas. Entre éstas, la regulación de el flujo de entrada y de salida de los materiales a fin de asegurar las condiciones óptimas para el proceso vital prevaleciente dentro de ellas. Asimismo, utilizan su información genética para guiar la síntesis de la mayoría de sus componentes y dirigir gran parte de sus actividades químicas. Entre esas actividades está la generación de adenosina trifosfato(ATP, por sus siglas en inglés), la síntesis molecular, la transportación de las moléculas dentro y entre las células, la remoción de los desechos y el movimiento parcial o incluso de toda la célula [4].

En la Figura 2.1 se muestra en forma abstracta las células con sus divisiones principales, características y funciones.

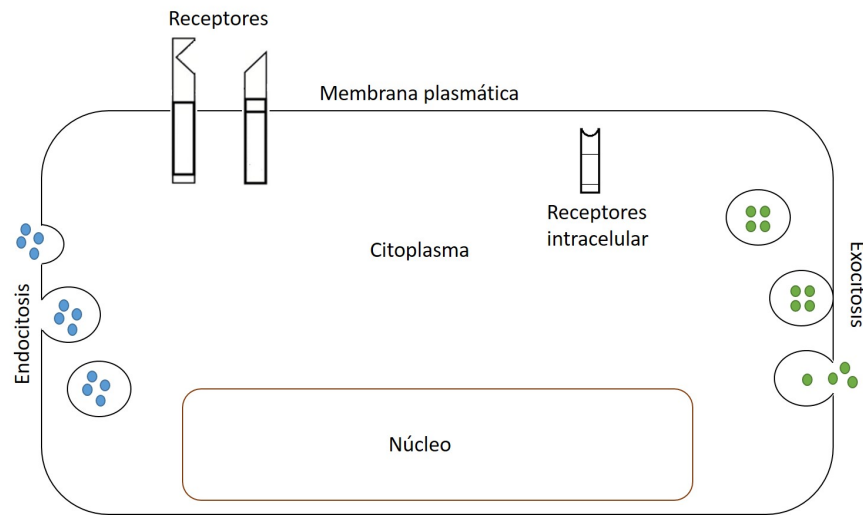


Figura 2.1: Descripción abstracta de una célula

La membrana plasmática conforma la superficie externa de la célula y separa su ambiente interno del externo. Constituye una barrera selectiva que organiza el flujo de materiales que entran y salen de la célula a fin de establecer y mantener un ambiente apropiado para las actividades que lleva a cabo. La membrana plasmática también desempeña una función clave en la comunicación entre células, y entre células y su ambiente externo.

El citoplasma comprende todo el contenido celular entre la membrana plasmática y el núcleo. Compuesto por el citosol y los organelos. El citosol consiste en la porción líquida del citoplasma que está formada principalmente por agua con partículas disueltas y suspendidas. Los organelos son estructuras celulares altamente organizadas, cada una de las cuales tiene forma y funciones específicas. Entre los ejemplos podemos citar citoesqueleto, ribosomas, retículo endoplásmico, complejo de Golgi, lisosomas, peroxisoma y mitocondria.

En el interior del núcleo se encuentran los genes que controlan la estructura de la célula y la mayoría de las funciones que lleva a cabo [4].

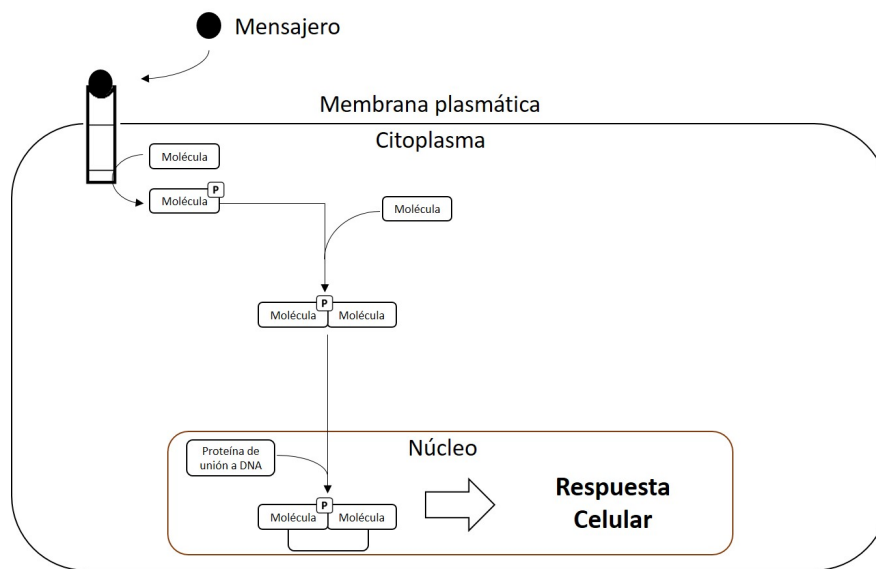


Figura 2.2: Vía de señalización celular

2.2.1. Las vías de señalización celular

En respuesta a las necesidades de auto regulación, el sistema endocrino envía señales químicas a las células para que lleven a cabo determinada función. Estos mensajeros pueden iniciar las vías de señalización al adherirse a los receptores en la membrana plasmática de la célula o internarse al citoplasma celular. Las vías de señalización son un grupo de moléculas que trabajan juntas para controlar las funciones celulares, donde la transducción de la señal se lleva a cabo mediante procesos de activación de moléculas por fosforilación, desactivación de moléculas, generación de complejos moleculares, entre otras, hasta llegar al núcleo y promover como respuesta celular la expresión génica, la reproducción celular, la apoptosis y demás funciones celulares.

2.2.2. El ciclo celular

El ciclo celular es el mecanismo por el cual el organismo genera células nuevas que reemplazan a células muertas y permite que la cantidad de células en un organismo oscile entre

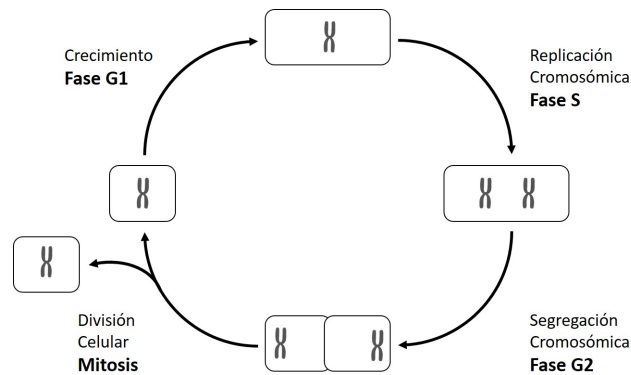


Figura 2.3: El ciclo celular

ciertos límites, cuenta con puntos de regulación que aseguran que las células no se dividan cuando el ácido desoxirribonucleico (DNA, por sus siglas en inglés) está dañado o cuando no hay espacio para más células en un tejido u órgano. En la Figura 2.3, se observan las cuatro etapas que integran el ciclo celular. Las células en fase G1 duplican su tamaño y aumenta la cantidad de organelos, enzimas y moléculas. En la fase S se duplica el DNA (replicación) y sus proteínas asociadas, de manera que, al final del proceso, la célula contará con el doble del material genético. Cada cromosoma posee cuatro cromátides en lugar de las dos originales. En la fase G2 las estructuras se preparan para la división, se empiezan a condensar los cromosomas y se corrigen los errores que pudieron ocurrir en la replicación. Por último, en la fase de mitosis inicia la división nuclear y la división citoplasmática dando como resultados dos células nuevas.

2.2.3. La apoptosis

Las células que han sufrido daños de forma natural por envejecimiento o debido a mutaciones por exposición a factores dañinos, y no pueden ser reparadas, entran en el proceso de muerte celular programada.

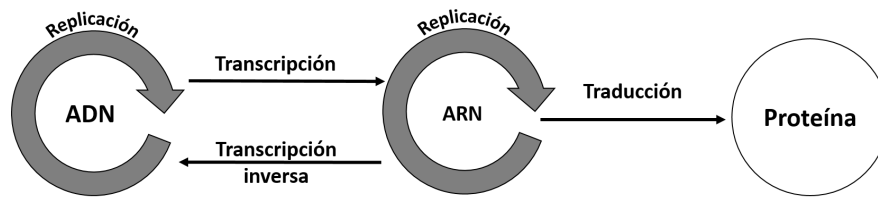


Figura 2.4: Dogma central de la biología molecular.

2.2.4. El dogma central de la biología molecular

La expresión génica está dada a partir del denominado dogma central de la biología molecular el cual decreta el flujo de información genética dentro de un sistema biológico y provee al organismo de las proteínas necesarias para su funcionamiento. Tal como se muestra en la Figura 2.4 puede ser visto en cuatro fases: la replicación del DNA, la codificación del DNA en la producción del ácido ribonucleico mensajero (mRNA, por sus siglas en inglés) en el proceso de transcripción, el procesamiento del mRNA y su migración del núcleo al citoplasma, y el proceso de síntesis de proteínas.

2.3. Aprendizaje automático

Es un área de la inteligencia artificial cuyo propósito es desarrollar técnicas y programas capaces de aprender y generalizar comportamientos a partir de la experiencia dada por un conjunto de datos ejemplo suministrados. El proceso de aprendizaje es el medio para alcanzar la capacidad de realizar la tarea [27].

En la Figura 2.5 se muestra el modelo genérico de aprendizaje automático, en el cual, a un algoritmo de entrenamiento se le suministra un conjunto de datos ejemplo, y éste es capaz de obtener una función que ajuste a los datos suministrados tal que sea capaz de predecir resultados a partir de datos no conocidos en el proceso de entrenamiento, denominado generalización.

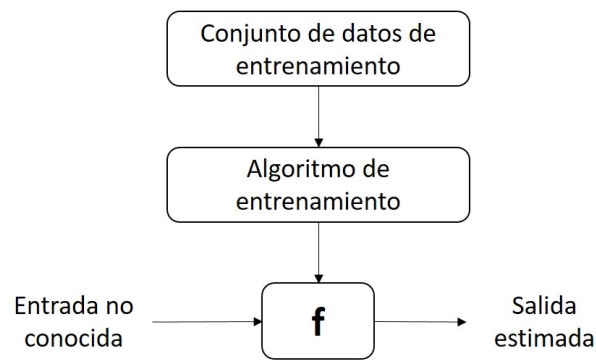


Figura 2.5: Modelo genérico de aprendizaje automático

Tal como se describe en el capítulo 1, las técnicas de aprendizaje automático requieren de identificar la clase de tarea a realizar, integrar y preprocesar un conjunto de datos de ejemplo y definir medidas de rendimiento que guíen el proceso de aprendizaje.

En las clases de tareas más comunes en el aprendizaje automático se le pide al programa de computadora que especifique la categoría K a la que pertenece alguna entrada (clasificación), prediga un valor numérico dado una entrada (regresión), observe una representación relativamente no estructurada de algún tipo de datos y transcriba la información en forma textual discreta (transcripción), convierta la entrada de símbolos en algún idioma en una secuencia de símbolos en otro idioma (traducción automática), identifique y filtre elementos atípicos e inusuales de un conjunto de eventos u objetos (detección de anomalías), entre otras [27].

En el conjunto de datos, cada ejemplo debe ser independiente de los demás y cada subconjunto obtenido a partir de éste deberá ser idénticamente distribuido. Cuando cada ejemplo contiene asociado un valor de salida a un valor de entrada se considera aprendizaje supervisado, de otra forma se considera aprendizaje no supervisado. Regularmente se divide en dos subconjuntos disjuntos, el primero para el proceso de entrenamiento y el segundo, denominado de prueba, aplicado para medir la capacidad de generalización [27].

Las medidas de rendimiento son cuantitativas y específicas de la clase de tarea y deben

ser apropiadas para medir el rendimiento en el proceso de entrenamiento y en el proceso de generalización [27].

La capacidad de un algoritmo de aprendizaje automático la integran el conjunto de todas las funciones que el modelo puede ajustar. La capacidad de representación está dada por el conjunto de funciones que el algoritmo puede aprender, y la capacidad efectiva es el conjunto de funciones que el algoritmo es capaz de encontrar; es probable que no encuentre la mejor función entre las posibles que pueda aprender, aunque puede aprender una que funciona excepcionalmente; el espacio de hipótesis está definido por el conjunto de funciones que el algoritmo está limitado a aprender [27].

En la Figura 2.6 se muestra un ejemplo de la capacidad de los algoritmos de aprendizaje automático, para una tarea de complejidad cuadrática se comparan los ajustes de tres predictores: lineal, cuadrático y un polinomio de grado nueve. En (a), la función lineal no puede capturar la curvatura que está presente en los datos, (b) la función cuadrática ajusta adecuadamente y en (c) el predictor de grado nueve es capaz de representar la función correcta, pero también es capaz de representar infinitamente muchas otras funciones que pasan exactamente a través de los puntos de entrenamiento sin que la generalización sea dada [27].

El desafío central en el aprendizaje automático es la capacidad de generalización, para lo cual el algoritmo debe tener un buen rendimiento en entradas previamente no observadas, y no solo en aquellas en las que el modelo fue entrenado. Un algoritmo adquiere la capacidad de resolver un tipo de tarea específico, mediante un proceso de entrenamiento iterativo por épocas [27].

Una época ocurre cuando el conjunto de datos completo es expuesto al algoritmo de aprendizaje. En cada época, el algoritmo es expuesto a datos ejemplo con el fin de optimizar la medida de rendimiento y al final aprender una función del espacio de hipótesis cuyo error de entrenamiento sea suficientemente pequeño. Una vez entrenado, se mide la capacidad de

generalización y si el error de prueba es suficientemente pequeño se considera que la capacidad del algoritmo es apropiada para la verdadera complejidad de la tarea y la cantidad de datos de entrenamiento proporcionados, Por lo tanto, el algoritmo ajusta en forma adecuada al conjunto de datos [27].

De lo anterior, se desprenden otras dos posibles condiciones: el bajoajuste (*underfitting*), ocurre cuando la capacidad del algoritmo no es suficiente para ajustar una función al conjunto de datos y se podría lograr una mejor generalización con más entrenamiento o incremento de la capacidad. Un indicador de esta condición es la incapacidad para obtener un error suficientemente bajo en el conjunto de datos de entrenamiento. El sobreajuste (*overfitting*) se produce cuando la capacidad del algoritmo es mayor que la complejidad de la tarea y éste puede memorizar las propiedades del conjunto de entrenamiento y no ajusta al conjunto de prueba, ésta situación es manifestada cuando la brecha entre el error de entrenamiento y el error de prueba es demasiado grande, ver la Figura 2.7 [27].

Existen parámetros que no son aprendidos, se denominan hiperparámetros y son sintonizados por el investigador. Los métodos de validación son mecanismos aplicados en el proceso de entrenamiento, tienen el propósito de evitar el sobreajuste y la sintonización de hiperparámetros. En la validación *hold-out* el conjunto de datos de entrenamiento se divide en dos conjuntos disjuntos, uno para entrenamiento y otro para validación, sobre el cual se estima el error de generalización durante el entrenamiento [27].

La validación cruzada se basa en la idea de repetir el entrenamiento y la prueba sobre subconjuntos aleatorios distintos o por ranuras del conjunto original. En el método (*K-fold*), el conjunto de datos es particionado en K conjuntos no superpuestos, el predictor es entrenado K veces, donde el K-ésimo subconjunto de datos se usa como conjunto de prueba, y el resto de los datos se usa como conjunto de entrenamiento. El error de prueba se estima por el promedio de los K-ésimos errores de prueba [27].

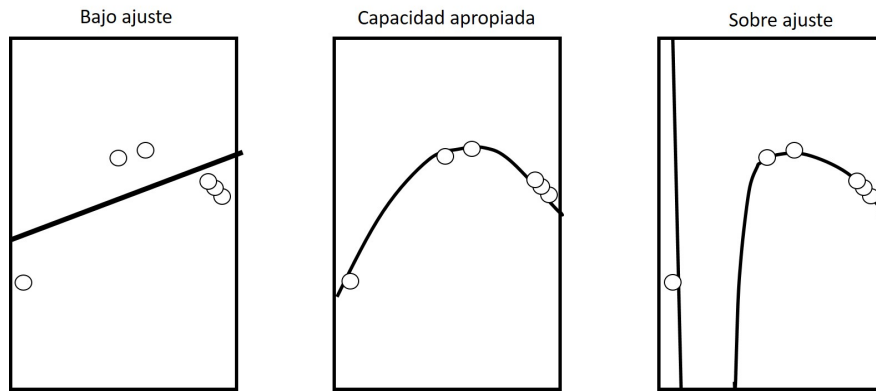


Figura 2.6: Capacidad de algoritmo de aprendizaje automático [27]

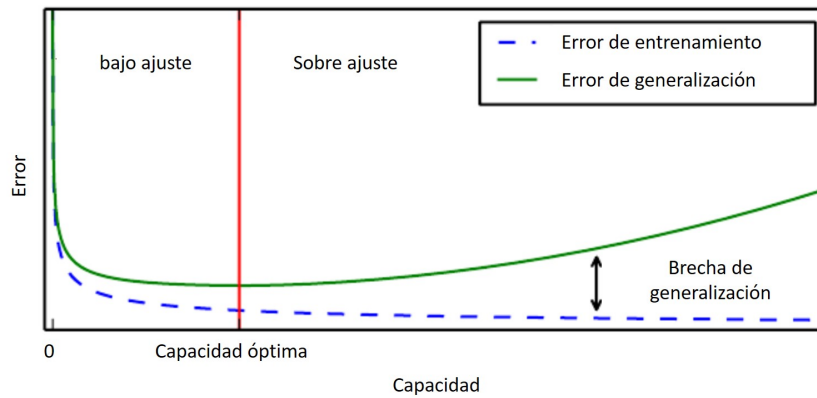


Figura 2.7: Relación típica entre la capacidad y el error [27]

2.3.1. Regresión lineal

La regresión lineal resuelve un problema de clase regresión, mediante un sistema que toma un vector de entrada $x \in \mathbb{R}^n$ y prediga el valor escalar de salida $y \in \mathbb{R}$, donde la salida es una función lineal de la entrada, caracterizada en la ecuación 2.1.

$$\hat{y} = w^T x, \quad (2.1)$$

Donde $w \in \mathbb{R}^n$ es un vector de parámetros y x el conjunto de características.

Los valores de los parámetros controlan el comportamiento del sistema, cada w_i es el coeficiente que multiplica la característica x_i , y sumados resumen las contribuciones de todas las características. Se puede ver a w como un conjunto de ponderaciones o pesos, los cuales determinan cómo cada característica afecta la predicción. Si una característica x_i recibe un peso positivo w_i , entonces al aumentar el valor de esa característica aumenta el valor de nuestra predicción. Si una característica recibe un peso negativo, entonces al aumentar el valor de esa característica disminuye el valor de nuestra predicción. Si el peso de una característica es grande en magnitud, entonces tiene un gran efecto en la predicción. Si el peso de una característica es cero, no tiene efecto en la predicción [27].

El rendimiento es medido por el error cuadrático medio (MSE, por sus siglas en inglés), el cual mide la distancia euclidiana entre el valor predicho y el valor real. Tal como se muestra en la ecuación 2.2 cuando las predicciones son muy cercanas a los valores reales el error se reduce [27].

$$MSE = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2 \quad (2.2)$$

Otra medida de rendimiento es la bondad de ajuste, describe el grado de acoplamiento entre los valores predichos y valores reales. El coeficiente de determinación se muestra en la ecuación 2.3, es un estimador de la bondad de ajuste, mide la proporción de varianza explicada por la regresión. Es un valor real en el rango de 0 a 1, entre más cerca esté de uno mayor acoplamiento entre los valores reales y los predichos, de forma inversa, entre más cerca de cero 0 menos ajustado será el algoritmo [27].

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (2.3)$$

2.4. Preprocesamiento de datos

2.4.1. Transformada wavelets

De manera muy general, la transformada wavelet de una función $f(t)$ es la descomposición de $f(t)$ en un conjunto de funciones $\Psi_{s,t}(t)$, que forman una base y son llamadas las wavelet [34]. Definida en la ecuación 2.4.

$$W_f(s, \tau) = \int f(t) \Psi_{s,\tau}^*(t) dt \quad (2.4)$$

Las wavelets son generadas a partir de la traslación y cambio de escala de una misma función wavelet $\Psi(t)$, llamada la wavelet madre [34], y se define en la ecuación 2.5

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \Psi\left(\frac{t - \tau}{s}\right) \quad (2.5)$$

donde s es el factor de escala, y τ es el factor de traslación.

Las wavelets $\Psi_{s,\tau}(t)$ generadas de la misma función wavelet madre $\Psi(t)$ tienen diferente

escala s y ubicación τ , pero todas tienen la misma forma. Se utilizan siempre factores de escala $s > 0$. Las wavelets son dilatadas cuando la escala $s > 1$, y son contraídas cuando $s < 1$. Así, cambiando el valor de s se cubren rangos diferentes de frecuencias. Valores grandes del parámetro s corresponden a frecuencias de menor rango, o una escala grande de $\Psi_{s,\tau}(t)$. Valores pequeños de s corresponden a frecuencias de mayor rango o una escala muy pequeña de $\Psi_{s,\tau}(t)$ [34].

Las transformadas wavelet cuenta con características de conservación de la información y retienen las características de la señal después de la transformación, y permiten reconstruir la señal original a través de la transformada wavelet inversa [34].

Éstas presentan mejor desempeño en relación a métodos de filtrado convencionales aplicados sobre datos genómicos [35] y han sido utilizadas en el filtrado de señales de datos biológicos [36], en la clasificación de tumores usando microarreglos de datos de expresión de genes [37], entre otros.

En particular, la wavelet discreta de Meyer [38] está definida en la ecuación (2.6).

$$\Psi(\omega) = \begin{cases} \frac{1}{\sqrt{2\pi}} \sin\left(\frac{\pi}{2}\nu\left(\frac{3|\omega|}{2\pi} - 1\right)\right) \varrho^{\frac{j\omega}{2}} & \text{if } \frac{2\pi}{3} < |\omega| < \frac{4\pi}{3}, \\ \frac{1}{\sqrt{2\pi}} \cos\left(\frac{\pi}{2}\nu\left(\frac{3|\omega|}{4\pi} - 1\right)\right) \varrho^{\frac{j\omega}{2}} & \text{if } \frac{4\pi}{3} < |\omega| < \frac{8\pi}{3}, \\ 0 & \text{De otra forma} \end{cases} \quad (2.6)$$

$$\text{donde :} \quad \nu(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x > 1 \end{cases}$$

2.4.2. Multicolinealidad

La multicolinealidad está definida por la correlación entre variables predictoras. La relación entre variables predictoras hace difícil cuantificar con precisión el efecto que cada variables predictoras ejerce sobre la variable dependiente. La multicolinealidad severa es problemática, porque puede incrementar la varianza de los coeficientes de regresión, haciéndolos inestables y como consecuencia: los coeficientes pueden parecer insignificantes incluso cuando exista una relación significativa entre el predictor y la respuesta, los coeficientes de los predictores altamente correlacionados variarán ampliamente de una muestra a otra, la eliminación de cualquier término altamente correlacionado del modelo afectará considerablemente los coeficientes estimados de los demás términos altamente correlacionados y los coeficientes de los términos altamente correlacionados incluso pueden tener el signo equivocado.

2.4.3. Escalamiento de datos

La ecuación 2.7 describe la función de escalamiento con base en la media μ y rango en el intervalo cerrado de -1 a 1 . De igual forma, la función de normalización basada en la unidad con rango en el intervalo cerrado de 0 a 1 se muestra en la ecuación 2.8.

$$f(x) = \frac{x - \mu(x)}{\max(x) - \mu(x)} \quad (2.7)$$

$$f(x) = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2.8)$$

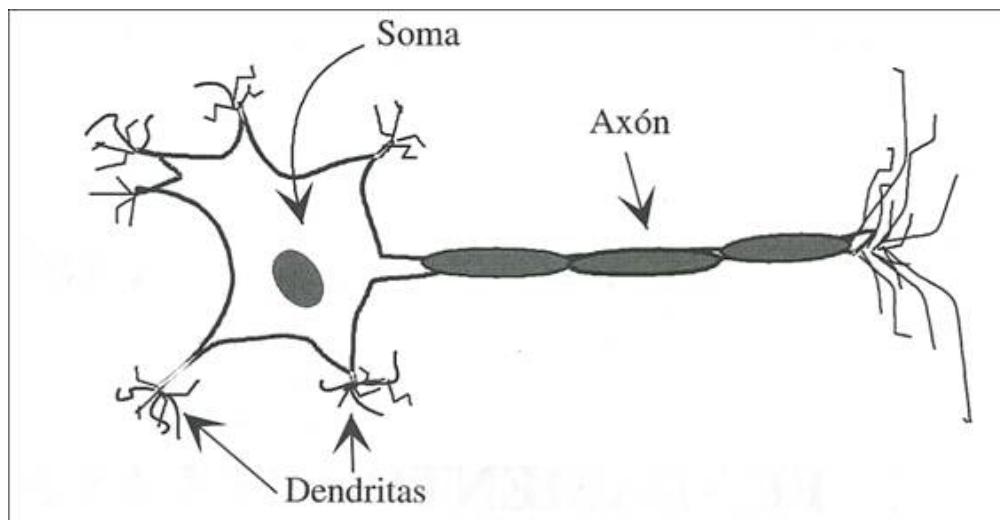


Figura 2.8: Descripción de las neuronas biológicas.

Fuente: <http://grupo.us.es/gtocom/pid/pid10/RedesNeuronales.htm>

2.5. Técnicas de computación natural

2.5.1. Redes neuronales artificiales

Las redes neuronales artificiales (ANN, por sus siglas en inglés) están inspiradas en el sistema nervioso del humano, responsable de controlar y regular al organismo mediante el sensado del entorno en que vive y se mueve, procesar la información de entrada, relacionarla a la experiencia previa y transformarla en memoria o acciones de respuesta apropiadas.

Las unidades de procesamiento elementales del sistema nervioso son las neuronas, como se ilustra en la Figura 2.8, que son células con tres elementos principales: las dendritas especies de fibras que alimentan de señales eléctricas al cuerpo de la célula, el cuerpo celular o soma que se encarga de procesar las señales de entrada y el axón es una fibra larga que propaga la señal desde el cuerpo de la célula hacia otras células.

La sinapsis permite la comunicación de las células nerviosas a través de transmisión de señales y puede ser modulada en distintas formas lo que permite adaptarse a distintas

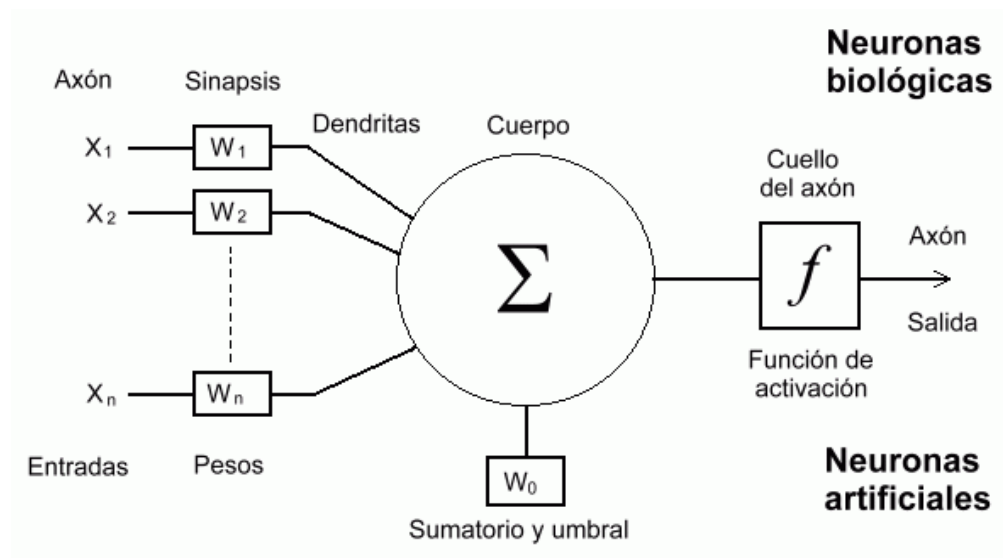


Figura 2.9: Descripción de las neuronas artificiales y su relación con las neuronas biológicas.

Fuente: <http://www.um.es/LEQ/Atmosferas/Ch-VI-3/F63s4p3.htm>

situaciones, es la base de las capacidades en el procesamiento de información del cerebro y se considera como los fundamentos de asociación, memoria y muchas otras habilidades mentales.

Las neuronas pueden tener conexiones hacia adelante y de retroalimentación con otras neuronas, dando origen a las redes neuronales, las cuales pueden exhibir comportamientos complejos y capacidades de procesamiento de información que no puede ser observada en una sola neurona, la representación de la información está distribuida en toda la red y es procesada en forma paralela.

Las neuronas artificiales, se ilustran en la Figura 2.9, son unidades sencillas de procesamiento con capacidad para recibir datos de entrada, calcular los pesos en la conexión de cada entrada hacia ésta, integrar todos los pesos mediante la función de agregación y generar datos de salida aplicando la función de activación.

Al igual que las neuronas, las neuronas artificiales tienen capacidades de comunicación y por lo tanto integrar redes. Las ANN presentan características y funcionalidades en común

con el sistema nervioso, neuronas artificiales como unidades de procesamiento de información, con capacidad de enviar y recibir estímulos a otras neuronas artificiales, conectarse con otras neuronas artificiales formando arquitecturas de ANN. La información es transmitida entre neuronas artificiales vía enlaces de conexión sinápticos; la eficiencia de una sinapsis se representa por pesos asociados o fuerza y corresponde a la información almacenada en la neurona artificial; el conocimiento es adquirido mediante la adaptación de las fuerzas o pesos asociados a los estímulos del entorno, denominado proceso de aprendizaje.

Las ANN tienen tres características principales: un conjunto de neuronas artificiales organizadas en capa de entrada, capas ocultas y capa de salida denominado arquitectura; el patrón de conectividad entre neuronas artificiales; y un método de entrenamiento que determina los pesos asociados a la conexiones sinápticas.

En las ANN las neuronas artificiales están integradas en capas funcionales, las capas de entrada reciben los estímulos directamente del entorno, las capas ocultas no tienen contacto directo con el entorno y las capas de salida emiten la respuesta al entorno.

Tal como se muestra en la Figura 2.10 la arquitectura de ANN de alimentación directa con múltiples capas (MLFFN, por sus siglas en inglés) esta formada por múltiples capas ocultas, las señales son propagadas desde la capa de entrada hasta las capa de salida, la salida de cada capa oculta es la entrada de la siguiente capa.

2.5.2. Perceptrón multicapas

Un perceptrón multicapas es una red neuronal con varias ocultas, una capa de entrada y una capa de salida y utiliza funciones de activación no lineales. Las más comunes son las funciones: sigmoidea con rango de -1 a 1 (Ecuación 2.9), tangente hiperbólica con rango de -1 a 1 y logística con rango de 0 a 1 (Ecuación 2.10).

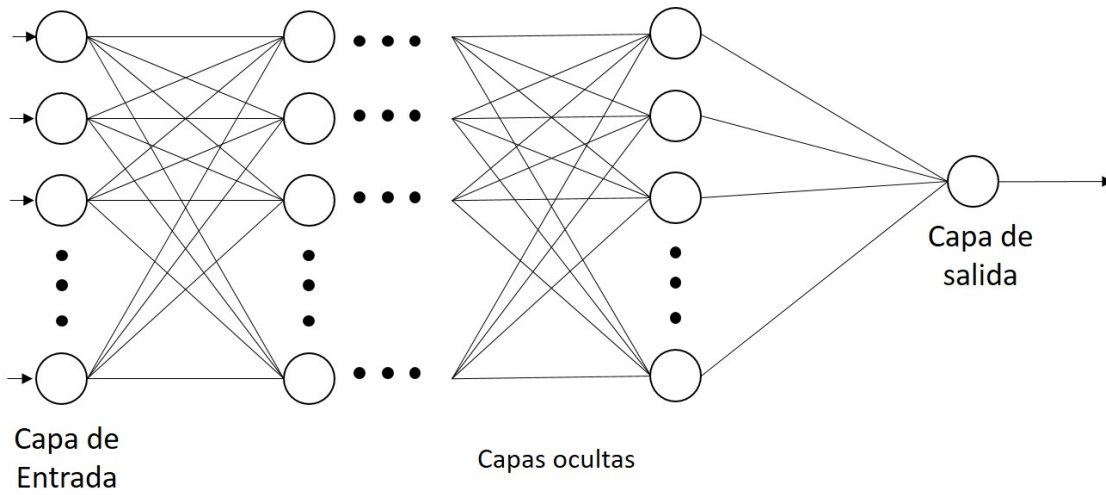


Figura 2.10: Arquitectura de ANN con múltiples capas y alimentación directa.

Este utiliza el paradigma de aprendizaje supervisado, son entrenados en modo lote (*batch*) en el cual para calcular la modificación de los pesos se deben presentar a la red cada uno de los ejemplos en el conjunto de datos, y se actualizan los pesos con las modificaciones y aplica método de propagación del error hacia atrás (*Backpropagation*)

$$S(t) = \frac{L}{1 + e^t} \quad (2.9)$$

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \quad (2.10)$$

2.5.3. Redes neuronales profundas (DNN)

Aún y cuando no existe un total acuerdo sobre lo que son las redes neuronales profundas (DNN, por sus siglas en inglés), en esta investigación se entienden como perceptrón multicapa, entrenado con algoritmos de tipo mini-lote en el cual el conjunto de datos se divide en lotes de tamaño fijo preestablecido y el cálculo de la modificación de los pesos se realiza después de procesar cada uno de los lotes.

2.5.4. Algoritmos genéticos

La computación evolutiva tiene sus fundamentos en los postulados de la biología evolutiva con el propósito de desarrollar técnicas de búsqueda y optimización para resolver problemas complejos.

La biología evolutiva es una ciencia que estudia la diversidad de la vida, las diferencias y similitudes entre los organismos y las características adaptativas y no adaptativas de los organismos.

Los sistemas evolutivos son aquellos donde existe descendencia generación tras generación y las características de la descendencia difieren entre generaciones, la población está integrada por un conjunto de individuos con capacidad de reproducirse, con similitud hereditaria, con capacidad de variar y con la capacidad de preservar las características originadas por las variaciones (selección natural). El proceso de evolución es el resultado de la herencia y las variaciones del material genético entre generaciones [28].

Los algoritmos genéticos (GA por sus siglas en inglés) son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización. Están basados en el proceso genético de los organismos vivos, con el paso de las generaciones las poblaciones evolucionan en la naturaleza de acuerdo con los principios de la selección natural y la supervivencia de los más fuertes, postulados por Darwin en 1859 [39] .

En los GA, los individuos de la población son genotipos y tienen cromosomas haploides donde cada unidad del cromosoma es un gen localizado en determinada posición denominado locus y puede asumir formas distintas. Se representan generalmente por cadenas de bits y puede asumir los valores de 0 o 1 [28, 39].

A partir de la información codificada en el genotipo se manifiesta el fenotipo y se evalúa la capacidad de adaptación al entorno. Con el objetivo de preservar las características de los in-

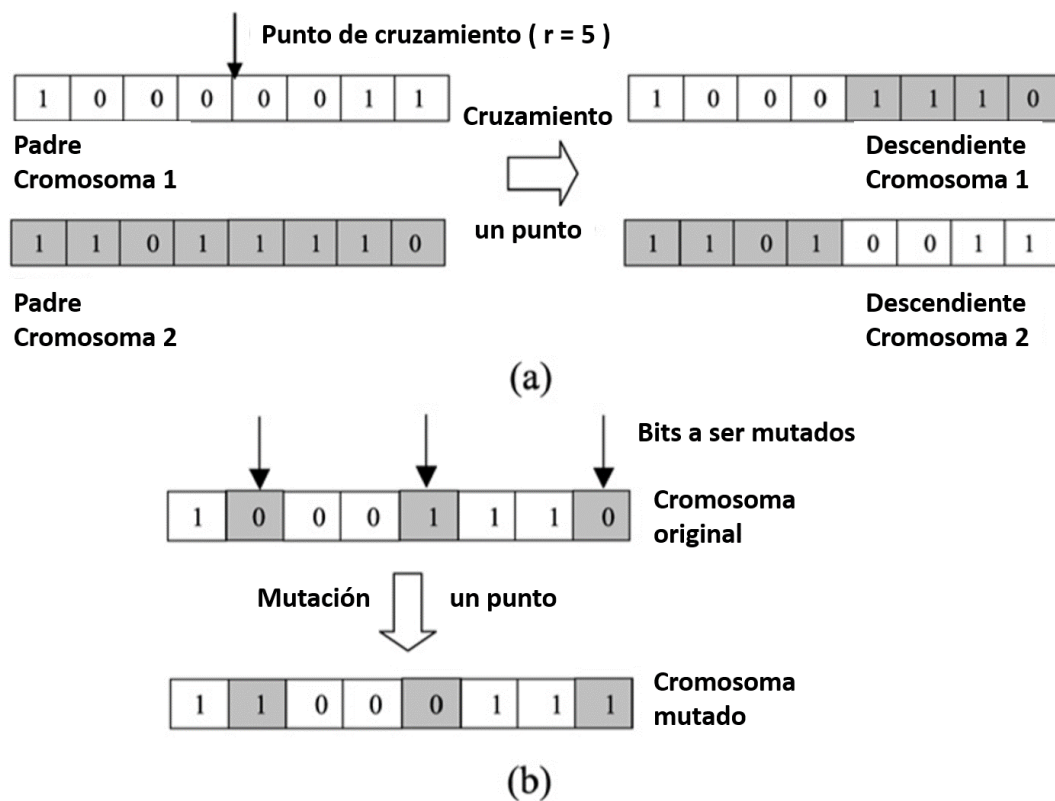


Figura 2.11: Operadores genéticos. Proceso de cruzamiento en un punto (a) y proceso de mutación en un punto (b).

dividuos mejores adaptados, el proceso de selección le asigna a cada individuo la probabilidad de ser seleccionado con relación a la capacidad de adaptación del mismo [28, 39].

En caso de reproducción sexual, tiene lugar el cruzamiento (*cross-over*), para lo cual se seleccionan dos individuos de la población actual, se obtiene un número aleatorio entre uno y la longitud del genotipo menos 1, se intercambian los segmentos a partir de esta posición entre ambos individuos y generan dos nuevos descendientes. En el proceso de mutación, por cada individuo se seleccionan en forma aleatoria las posiciones a mutar, con base en la probabilidad de mutación el alelo cambia de forma [28, 39]. Tal como se muestra en la Figura 2.11.

2.5.5. Importancia relativa Olden

Una vez que la DNN ha sido entrenada y ha cumplido con el proceso de generalización se entiende entonces que la red ha ajustado al conjunto de datos y los pesos relacionados con cada una de las variables de entrada han sido establecidos.

Con base en lo anterior y con el propósito de determinar la importancia relativa de cada una de las variables independientes en la variable dependiente, se han desarrollado algunos métodos: pesos de la conexión, algoritmo de garson, derivadas parciales, perturbación de entrada, análisis de sensibilidad, entre otros.

Olden et al en [40] realizan un comparativo de métodos de importancia relativa y nos muestran que la mejor metodología para cuantificar con precisión la importancia de las variables es el método basado en los pesos de conexión. Éste se basa en el principio de los pesos de las conexiones de cada nodo en la DNN y toma en cuenta la magnitud y la dirección, donde el peso representa la intensidad, y la dirección la excitación de la señal, por lo tanto, a mayor peso con dirección positiva la importancia relativa es mayor e incrementa la capacidad predictiva del modelo [40].

Capítulo 3

Metodología

En el desarrollo de esta investigación, se utilizó la lógica de la simulación como método de investigación [41]. Tal como se muestra en la Figura 3.1 consta de cuatro etapas. En la primera etapa se recolectan datos de experimentos *in vitro* e *in vivo* del sistema biológico a estudiar. En la segunda etapa se analiza el sistema en conjunto con los datos colectados y se obtiene una expresión abstracta, a partir de la cual se desarrolla el modelo biocomputacional en forma de programa de computadora. En la tercera etapa el modelo es ejecutado *in virtuo* para generar el conjunto de datos simulados. En la cuarta etapa se analizan los conjuntos de datos en busca de similitudes.

3.1. Recopilación y tratamiento de datos

3.1.1. Conjunto de datos

Con base en el objetivo de este estudio y con el propósito de ser exhaustivo se buscó en distintas bases de datos de interacción entre mRNAs y miRNAs. Fueron identificaron 179 miRNAs que interactúan con el mRNA-Smad7 en las bases de datos mirDB [42, 43],

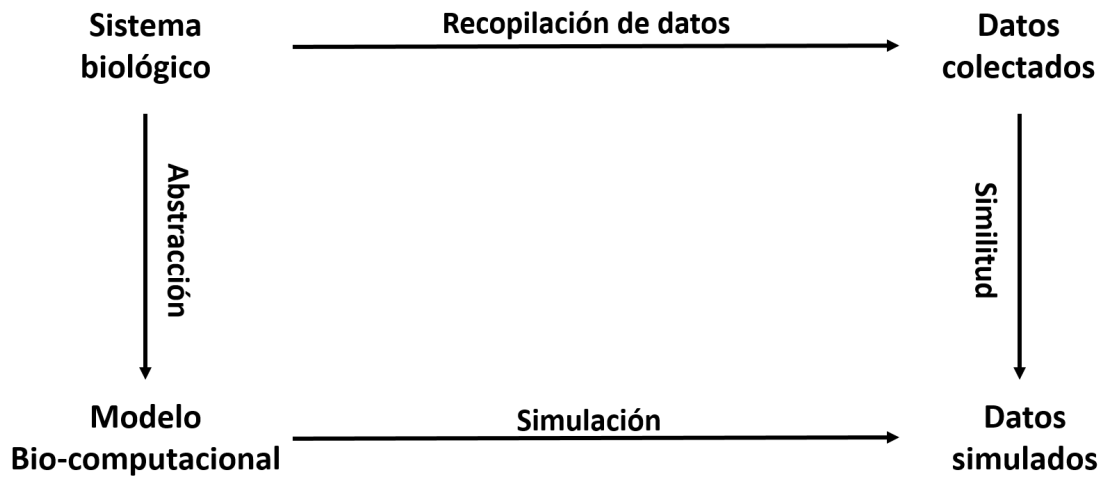


Figura 3.1: La lógica de la simulación como método de investigación.

microRNA [44], y mirtarBase [45]. Así como también se descargó el conjunto de archivos de expresión de 1074 muestras de pacientes con cáncer de mama del sitio web del proyecto “El Atlas del Genoma del Cáncer” [46], cada conjunto de archivos de una muestra contiene la expresión base y normalizada de los miRNAs, los mRNAs, entre otros datos que no fueron considerados para este trabajo. El conjunto de datos inicial se integró por los valores base de mRNA-Smad7 y los valores base de los miRNAs de todos los conjunto de archivos [47] con tamaño de 1074 registros (renglones) x 180 campos (columnas) (un campo con la expresión del mRNA-Smad7 y 179 expresiones de miRNAs) de valores enteros.

3.1.2. Preprocesamiento de datos

Dado el conjunto de datos inicial en cada muestra se validó la existencia de expresión de mRNA-Smad7 y la existencia de expresión de los miRNAs, de lo cual se obtuvo que las 1074 muestras cuentan con expresión del mRNA-Smad7 y que un total de 39 miRNAs no cuentan con expresión en las muestras, reduciendo de 179 a 140 miRNAs, por lo tanto el tamaño del conjunto de datos se redujo a 1074 registros $\times$ 141 campos de valores enteros (mRNA-Smad7 y 140 miRNAs) . Partiendo de este nuevo conjunto de datos se validó la expresión de cada

miRNA en todas las muestras y se descartaron 41 miRNAs que no aportan estadísticamente en la explicación de la variabilidad de la expresión de mRNA-Smad7 dado que su expresión tiene moda igual a cero y la frecuencia de la moda es de 800 en promedio. El tamaño del conjunto de datos cambió a 1074 registros $\times$ 100 campos con valores enteros (mRNA-Smad7 y 99 miRNAs).

Los vectores de expresión de cada miRNA se consideraron como series de datos en el tiempo y fueron transformados mediante la wavelet discreta de Meyer [38] para eliminar posible ruido en los datos, definida en la ecuación (2.6).

La técnica de regresión lineal simple se aplicó al conjunto de datos para identificar y eliminar la relación lineal exacta y/o la alta correlación entre los miRNAs y de esta forma validar el supuesto de multicolinealidad, obteniendo como resultado la inexistencia de colinealidad entre los miRNAs (resultados omitidos en este documento).

Por último, la expresión de los miRNAs fueron escaladas en el rango de -1 a 1 ver ecuación (2.7) y así empatar con la función de activación sigmoidea de las DNN. La expresión del mRNA-Smad7 se escaló en el rango de 0 a 1 , dado que la función de activación es lineal y los valores esperados están en ese rango.

El conjunto de datos después del preprocesamiento se integró por 1074 muestras que cuentan con expresión de mRNA-Smad7 como variable dependiente y la expresión de 99 miRNAs como variables independientes predictoras.

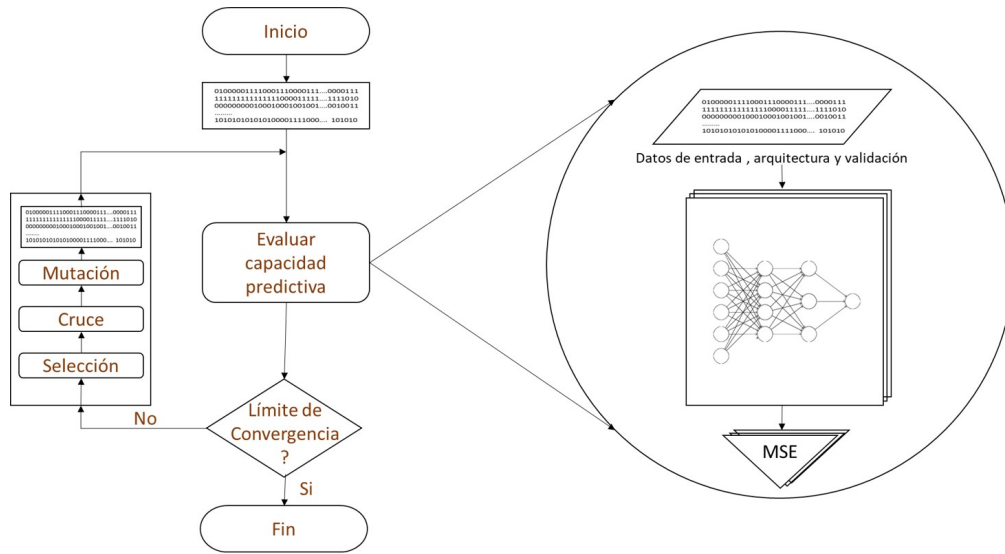


Figura 3.2: Método evolutivo del conjunto de datos,arquitectura y validación (EMDAV).

3.2. Diseño del modelo de predicción

3.2.1. Método de búsqueda evolutivo

El diseño del modelo predictivo fue tratado como un problema de optimización de búsqueda automática en un espacio de soluciones, donde cada punto (solución) representa un modelo de predicción con características propias y capacidad predictiva. Para lo cual se desarrolló un método evolutivo basado en GA como técnica de optimización de búsqueda automática para diseñar automáticamente DNN evolutivas como modelos predictivos, tal como se muestra en la Figura 3.2.

Este método establece los mecanismos para evolucionar el conjunto de datos de entrada, la arquitectura de la DNN y el método de validación en el entrenamiento de la DNN y es denominado método evolutivo del conjunto de datos, arquitectura y validación (EMDAV). El EMDAV se ejecutó 20 veces y en cada una, la evolución máxima fue hasta por 100 generaciones, con 5 generaciones como umbral de paro sin cambio significativo en la capacidad predictiva de los modelos evaluados.

3.2.2. Algoritmo genético

La búsqueda del modelo con mejores capacidades predictivas se trató como un problema de minimización sobre el error cuadrático medio (MSE) como valor de adaptación, se utilizó un algoritmo genético de tipo binario con población inicial de 50 individuos con operador de cruzamiento uniforme con 80 % de probabilidad, operador de selección de rango lineal, operador de mutación aleatorio uniforme con 10 % de probabilidad y factor de elitismo de 10 %. En el proceso de sintonizar los parámetros con mejor ajuste a la evolución de las características deseadas, se realizaron experimentos con los factores selección con cuatro niveles (*linear rank*, *non linear rank*, *roulette wheel* y *tournament*) y cruzamiento con dos niveles (*Single point* y *uniform*), mientras que el tratamiento fue la evolución del GA.

Se estableció el operador de selección en *linear rank* y el operador de cruzamiento *uniform* dado que son los niveles de los factores del tratamiento con mayor R^2 y menor MSE. De igual forma, el tamaño de la población de 50 individuos fue establecido a través de la evolución del GA con tamaño de población de 20, 50 y 100 individuos. La codificación base para los individuos de la población es de longitud variable integrado por 99 bits que codifican el conjunto de datos de entrada donde un uno (1) significa que el miRNA es parte de este conjunto y cero (0) que no es; un bit para los métodos de validación (0 para Hold-out y 1 para K-folds), la cantidad máxima de capas ocultas está codificado en seis bits y la cantidad máxima de nodos por cada capa oculta está codificado en diez bits, tal como se describe en la ecuación (3.1).

$$individual_length = 99bits + 1bit + 6bits * 10bits \quad (3.1)$$

La cantidad de capas ocultas resulta de la conversión a decimal de 6 bits + 1 decimal evitando la existencia de 0 capas, de igual forma el número de nodos por capa oculta se

obtiene de la conversión a decimal de $10 \text{ bits} + 1 \text{ decimal}$. La variabilidad de la longitud del individuo está dada por los últimos dos términos de la ecuación (3.1) que establecen la longitud necesaria para los valores máximos en ambos términos, cuando no es de esa forma a partir del bit 107 solo se procesan los primeros $(n + 1) * 10 \text{ bits}$, donde $n + 1$ es el número de capas ocultas y por lo tanto $10 * (64 - (n + 1)) \text{ bits}$ no son parte del individuo en cuestión.

3.2.3. Redes neuronales profundas

Cada individuo del GA se manifestó en una DNN con arquitectura dada por el conjunto de datos de entrada de tamaño variable de miRNAs, función de transferencia sigmoidea (Ecuación 2.9) entre la capa de entrada y capas ocultas, función de transferencia lineal entre la última capa oculta y la capa de salida y fueron entrenados con el algoritmo de optimización de descenso de gradiente de tipo mini-lote denominado RMSprop [48] con tamaño de lote fijo en 50 y con tasa de aprendizaje fija en 0.01; se ejecutaron 20 repeticiones cuando se usó validación hold-out y 10 pliegues con validación K-folds hasta por un máximo de 5000 épocas en ambos tipos de validación, en cada una de las corridas el conjunto de datos fue dividido de forma aleatoria en tres subconjuntos: el 50 % para entrenamiento, un 25 % para validación y el 25 % restante para probar la capacidad predictiva determinada por el MSE. Con el fin de evitar el sobreentrenamiento se estableció un umbral de 10 actualizaciones de gradiente como parada por divergencia en resultados del conjunto de validación o falta de convergencia significativa del conjunto de entrenamiento y el valor de 50 % para abandono de pesos (*Dropout*) entre la última capa oculta y la capa de salida.

Capítulo 4

Resultados y validación del modelo

En esta sección se discuten los resultados obtenidos mediante el método EMDAV y el modelo de predicción basado en DNN.

4.1. Evolución del método EMDAV

El método EMDAV se ejecuto 20 veces y se alcanzó como máximo la generación número 24 antes del umbral de paro; en la ejecución con mayor número de generaciones se realizaron 17,760 evaluaciones de fenotipos hasta la generación $24 + 5$, de los cuales 1,560 usaron validación k-folds con 10 pliegues y los otros 16,200 restantes con validación hold-out y 20 repeticiones. La variabilidad del rasgo de método de validación fue segregado a partir de la generación 14, la arquitectura quedó establecida en la generación 21, las generaciones restantes se usaron en forma exclusiva para la evolución del conjunto de datos de entrada en cada una de las 20 ejecuciones.

De lo anterior, en la Tabla 4.1 se muestran los 10 modelos con mayor capacidad predictiva y se puede apreciar que en la evolución de las características hay una tendencia a converger

hacia los mismos valores.

Los resultados mostrados en la Tabla 4.1 solamente representan una parte minúscula de los experimentos realizados en el proceso de puesta a punto del algoritmo. Dicho proceso se realizó con cuidado debido al costo computacional, llevando por ejemplo, a la realización de ejecuciones exploratorias sobre espacios de búsqueda incrementales en el rango de 4 capas ocultas y 64 nodos por capa hasta 64 capas ocultas y 1024 nodos por capa con ambos tipos de validación y un solo pliegue o repetición. En las ejecuciones exploratorias el GA convergía a la primer arquitectura mostrada en la Tabla 4.1. Dado esto, se realizó la exploración final con 20 repeticiones para hold-out y 10 pliegues para K-fold y los resultados fueron consistentes con las evoluciones exploratorias. Lo cual permite establecer el modelo predictivo basado en DNN descrito en la siguiente sección.

Tabla 4.1: Los mejores modelos predictivos basados en DNN

Generación	Individuo	miRNAs	Validación	Capas Ocultas	MSE	R ²
23	17	45	Hold-out	42:63	0.003877	0.9275
19	16	53	Hold-out	42:63	0.004425	0.9224
21	28	53	Hold-out	42:63	0.003982	0.9174
21	6	52	Hold-out	42:63	0.003612	0.9167
17	36	52	Hold-out	42:63	0.003668	0.9139
18	20	54	Hold-out	42:63	0.004021	0.9138
22	7	50	Hold-out	42:63	0.004431	0.9137
20	8	55	Hold-out	46:61	0.004826	0.9128
17	5	53	Hold-out	42:63	0.004176	0.9119
19	25	55	Hold-out	42:63	0.004045	0.9112

4.2. Modelo predictivo basado en DNN

En la figura 4.1 se muestra el modelo predictivo basado en DNN con arquitectura de dos capas ocultas con 42 y 63 nodos respectivamente, tipo de validación hold-out, un conjunto de datos de entrada integrado por 44 miRNAs como se muestra en la Tabla 4.2 y predice la

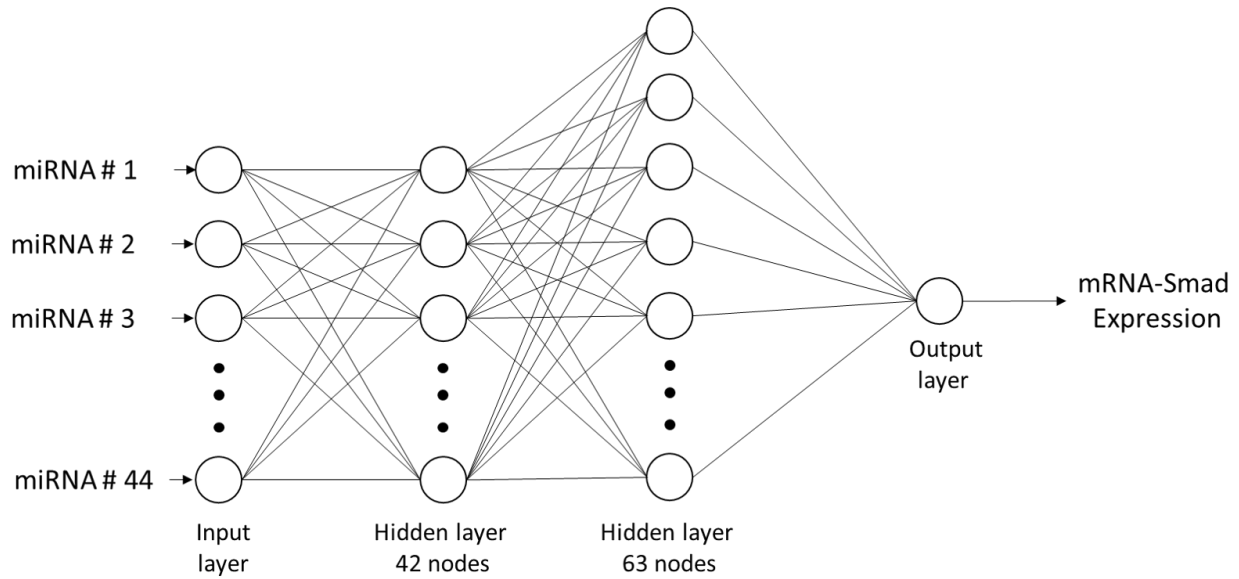


Figura 4.1: Modelo predictivo basado en DNN

regulación dada por los miRNA sobre los transcritos de mRNA que codifican para la proteína Smad7 en pacientes con cáncer de mama con $R^2 = 0,99$ y MSE de 0.00001.

Tabla 4.2: Conjunto de datos evolucionado del modelo predictivo

hsa-let-7c	hsa-let-7d	hsa-let-7g	hsa-mir-100	hsa-mir-107	hsa-mir-10b
hsa-mir-125a	hsa-mir-126	hsa-mir-140	hsa-mir-145	hsa-mir-146a	hsa-mir-148a
hsa-mir-149	hsa-mir-15a	hsa-mir-15b	hsa-mir-183	hsa-mir-185	hsa-mir-191
hsa-mir-200c	hsa-mir-205	hsa-mir-20a	hsa-mir-21	hsa-mir-212	hsa-mir-22
hsa-mir-224	hsa-mir-23a	hsa-mir-26b	hsa-mir-27a	hsa-mir-29a	hsa-mir-29c
hsa-mir-30a	hsa-mir-32	hsa-mir-33a	hsa-mir-34a	hsa-mir-374a	hsa-mir-375
hsa-mir-378	hsa-mir-425	hsa-mir-429	hsa-mir-497	hsa-mir-503	hsa-mir-625

4.3. Importancia relativa de miRNAs con mRNAs-Smad7

La importancia relativa de cada miRNAs sobre la expresión de mRNA-Smad7 fue evaluada mediante el algoritmo de Olden [49, 40].

Los miRNAs que incrementan la capacidad predictiva del modelo se muestran en la Ta-

bla 4.3 y corresponden a los valores promedio de 20 ejecuciones del modelo predictivo como perceptrón multicapas (MLP), con función de activación logística (Ecuación 2.10) y entrenado con el algoritmo *resilient backpropagation with weight backtracking* denominado Rprop+ [50].

Considerando la información contenida en las muestras de pacientes con cáncer de mama tomados del proyecto TCGA, los resultados muestran que los miRNAs hsa-mir-146a, hsa-mir-93, hsa-mir-375, hsa-mir-205, hsa-mir-15a, hsa-mir-21, hsa-mir-20a, hsa-mir-503, hsa-mir-29c, hsa-mir-497, hsa-mir-107, hsa-mir-125a, hsa-mir-200c, hsa-mir-212, hsa-mir-429, hsa-mir-34a, hsa-let-7c, hsa-mir-92b, hsa-mir-33a, hsa-mir-15b, hsa-mir-224, hsa-mir-185 y hsa-mir-10b integran un perfil que regula de forma crítica la expresión del mRNA-Smad7 y la expresión de la proteína Smad7 en cáncer de mama y pueden ser usados como biomarcadores o como blancos en terapias dirigidas.

Estos resultados son acordes a algunos trabajos reportados previamente que asocian a miRNAs con el mRNA-Smad7 y la proteína Smad7 en cáncer de mama, tales como: El hsa-mir-146a es un *oncomir* que regula la expresión del mRNA y la proteína Smad7 en esteatohepatitis fibrosante no alcohólica [51] y está sobreexpresado en el plasma de pacientes con cáncer de mama [52]. El hsa-mir-93 es un *oncomir* parte del *cluster* 106b-25 implicado en la transición epitelial mesenquimal suprimiendo la expresión de Smad7 y activando la vía de señalización TGF-Beta en cáncer de mama [53]. El hsa-mir-21 es un *oncomir* que promueve la proliferación y migración en cáncer de mama mediante la supresión de Smad7, lo cual mejora las vías de señalización del factor de crecimiento epidérmico (EGF, por sus siglas en inglés) y TGF-Beta [54]. El *cluster* miR424-503 contiene al *oncomir* hsa-mir-503, el cual está sobreexpresado en cáncer de mama metastásico y suprime los niveles de expresión de Smad7 y Smurf2 [55]. El hsa-mir-497 es un *oncomir* bajoexpresado en cáncer de mama [18], el mRNA-Smad7 es un blanco de éste y están correlacionados en forma negativa en cáncer de mama [56].

De igual forma, se reportan algunos trabajos que asocian a miRNAS con el con el mRNA-Smad7 y la proteína Smad7 en distintos tipos de cáncer u otras malignidades, tales como: el hsa-mir-375 que es un *oncomir* asociado con polimorfismos de un solo nucleótido de Smad7 en cáncer colorectal (CRC, por sus siglas en inglés) y en conjunto pueden considerarse como biomarcadores no invasivos en la detección y diagnóstico de CRC [57]. El hsa-mir-15a es un *oncomir* sin reporte asociado a cáncer de mama y que en infecciones del virus de la hepatitis B regula la apoptosis y tumorigenesis con base en la regulación de Smad7 [58]. El hsa-mir-212 activa células stellate hepatic y promueve fibrosis en el hígado mediante la supresión de Smad7 [59]. El hsa-mir-92b promueve la progresión del carcinoma hepatocelular vía la represión de Smad7 [60]. El hsa-mir-15b hace blanco en mRNA-Smad7 en la angiogénesis en el infarto de miocardio [61]. El hsa-mir-185-3p predice la radiosensibilidad del carcinoma nasofaríngeo y modulan el crecimiento y la apoptosis de células cancerosas al regular a Smad7 [62].

Por otra parte, algunos miRNAS son reportados asociados solamente a cáncer de mama, tal es el caso de: el hsa-mir-205 es supresor de tumor que no ha sido reportado en relación con Smad7, supresor de crecimiento celular e invasión en cáncer de mama, regulador de ErbB3 y el factor de crecimiento endotelial vascular [63], y se observó bajoexpresado en cáncer de mama [18]. El hsa-mir-20a se encuentra sobreexpresado de forma significativa en células de cáncer de mama [64]. El hsa-mir-29c es un *oncomir* sobreexpresado en cáncer de mama [17]. El hsa-mir-107 es un *oncomir* que se encuentra bajoexpresado en células madre en cáncer de mama [65], asociado con altas probabilidades de recurrencia en cáncer de mama triple negativo [66] y sobreexpresado en el carcinoma hepatocelular humano acelera la progresión tumoral mediante la represión del gene CPEB3 [67]. El hsa-mir-125a es un supresor de tumores cuya sobreexpresión decrementa la movilidad e invasión celular mediante la supresión de los niveles de transcrito y proteínas her2 y her3 y se encuentra significativamente bajoexpresado en cáncer de mama her2-positivo [68]. El hsa-mir-200c se encuentra bajoexpresado

en cáncer de mama [69].

Contrario a lo anterior, los miRNAs has-mir-429, has-mir-34a, has-let-7c y has-mir-33a no están reportados en la literatura directamente asociados a Smad7.

Tabla 4.3: Importancia relativa de miRNAs que hacen blanco en mRNA-Smad7

miRNAs	% Relativo	miRNAs	% Relativo	miRNAs	% Relativo
hsa.mir.146a	12.00 %	hsa.mir.29c	4.13 %	hsa.let.7c	1.52 %
hsa.mir.93	10.71 %	hsa.mir.497	3.95 %	hsa.mir.92b	1.32 %
hsa.mir.375	10.18 %	hsa.mir.107	3.66 %	hsa.mir.33a	1.18 %
hsa.mir.205	7.93 %	hsa.mir.125a	3.40 %	hsa.mir.15b	1.10 %
hsa.mir.15a	7.75 %	hsa.mir.200c	3.16 %	hsa.mir.224	1.09 %
hsa.mir.21	7.72 %	hsa.mir.212	1.97 %	hsa.mir.185	0.74 %
hsa.mir.20a	7.70 %	hsa.mir.429	1.68 %	hsa.mir.10b	0.22 %
hsa.mir.503	5.31 %	hsa.mir.34a	1.59 %		

Capítulo 5

Conclusiones y perspectivas

Se desarrolló un modelo biocomputacional de regresión no-lineal basado en DNN para predecir la expresión del mRNA-Smad7 dada por los miRNAs, validado a través de los resultados de experimentos *in vivo* e *in vitro* reportados en la literatura.

Se presentó un método evolutivo de búsqueda basado en GA binario denominado EMDAV, donde cada individuo se manifiesta en DNN y se evolucionan el conjunto de datos de entrada, la arquitectura y el método de validación de entrenamiento como características que definen la capacidad predictiva del modelo. Este método es capaz de encontrar un modelo de predicción basado en DNN que se ajusta a los datos biológicos con $R^2 = 0,99$ y MSE de 0.00001.

Se establece un perfil de regulación crítica para la expresión de mRNA-Smad7 y la expresión de la proteína Smad7 en cáncer de mama integrado por hsa-mir-146a, hsa-mir-93, hsa-mir-375, hsa-mir-205, hsa-mir-15a, hsa-mir-21, hsa-mir-20a, hsa-mir-503, hsa-mir-29c, hsa-mir-497, hsa-mir-107, hsa-mir-125a, hsa-mir-200c, hsa-mir-212, hsa-mir-429, hsa-mir-34a, hsa-let-7c, hsa-mir-92b, hsa-mir-33a, hsa-mir-15b, hsa-mir-224, hsa-mir-185 y hsa-mir-10b.

A partir de este perfil, con base en estudios previamente reportados en la literatura, se observa que cinco miRNAs están asociados con el mRNA-Smad7 y la proteína Smad7 en

cáncer de mama, seis miRNAs están asociados con el mRNA-Smad7 y la proteína Smad7 en diferentes tipos de cáncer u otras malignidades, seis miRNAs se relacionan solo con cáncer de mama, y por último, que cuatro miRNAs no están directamente asociados con Smad7.

Como trabajo futuro se propone la aplicación de esta metodología a otros conjuntos de datos relacionados con el cáncer de mama u otras enfermedades, así como el desarrollo de experimentos de laboratorio que permitan confirmar la relevancia de cada miRNA en la regulación del mRNA-Smad7.

Bibliografía

- [1] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, 136(5):E359–86, mar 2015.
- [2] Teresa Audesirk, Gerald Audesirk, and Bruce E. Byers. *Biología. La vida en la tierra*. México, 2008.
- [3] Raymond F Oram. *Biología de sistemas vivos*. 2007.
- [4] Bruce Alberts, Dennis Bray, Karel Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Introducción a la biología celular*. Médica Panamericana, 2011.
- [5] Lodish Harvey, Berk Arnold, Matsudaira Paul, Kaiser Chris A., Krieger Monty, Scott Matthew P., Zipursky S. Lawrence, and Darnell James. *Biología Celular y Molecular*. Editorial Médica Panamericana, 2004.
- [6] National Cancer Institute. Que es el cancer. [ONLINE] Disponible en: <https://www.cancer.gov/espanol/cancer/naturaleza/que-es> [Consultado: 2017-04-27].
- [7] Winship Cancer Institute. Apoptosis. [ONLINE] Disponible en:

- <https://www.cancerquest.org/es/biologia-del-cancer/apoptosis> [Consultado: 2017-04-27].
- [8] ¿Qué es el cáncer de mama? [ONLINE] Disponible en: http://www.breastcancer.org/es/sintomas/cancer_de_mama/que_es_cancer_mama [Consultado: 2017-04-27].
- [9] J. Massague. Tgf- β signal transduction. *Annual Review of Biochemistry*, 67(1):753–791, 1998. PMID: 9759503.
- [10] E Christeli, V Zoumpourlis, H Kiaris, M Ergazaki, S Vassilaros, and D Spandidos. TGF-beta 1 overexpression in breast cancer. *Oncology Reports*, 3(6):1115–1118, nov 1996.
- [11] C. Stolfi, I. Marafini, V. De Simone, F. Pallone, and G. Monteleone. The dual role of smad7 in the control of cancer growth and metastasis, int. *J. Mol. Sci*, 14:12, 2013.
- [12] L. Lin, H. Gan, H. Zhang, W. Tang, Y. Sun, X. Tang, D. Kong, J. Zhou, Y. Wang, and Y. Zhu. Microrna-21 inhibits smad7 expression through a target sequence in the 3' untranslated region and inhibits proliferation of renal tubular epithelial cells, mol. In *Med. Rep*, vol. 10, no. 2, pp. 707-712, 2014.
- [13] G. Desmeulles, G. Querrec, P. Redou, S. Kerdélo, L. Misery, V. Rodin, and J. Tisseau. The virtual reality applied to biology understanding: The in virtuo experimentation. *Expert Systems with Applications*, 30(1):82–92, 2006.
- [14] American Society of Clinical Oncology. Qué es la terapia dirigida — Cancer.Net. [ONLINE] Disponible en: <http://www.cancer.net/es/desplazarse-por-atención-del-cáncer/cómo-se-trata-el-cáncer/qué-es-la-terapia-dirigida> [Consultado: 2017-04-27].
- [15] Q. Li, D. Zhang, Y. Wang, P. Sun, X. Hou, J. Larner, W. Xiong, and J. Mi. MiR-21/Smad. 7 signaling determines tgf- β 1-induced caf formation., sci. *Rep.*, 3, 2013.

- [16] M. Han, F. Wang, Y. Gu, X. Pei, G. Guo, C. Yu, L. Li, M. Zhu, Y. Xiong, and Y. Wang. "MicroRNA-21 induces breast cancer cell invasion and migration by suppressing smad7 via egf and $\text{tgf-}\beta$ pathways". *Oncology Reports*, 35(1):73–80, 2016.
- [17] L. Yan, X. Huang, Q. Shao, M. Huang, and L. Deng. "mir-21 overexpression in human breast cancer is associated with advanced clinical stage , lymph node metastasis and patient poor prognosis". *pp*, 2348, 2008.
- [18] Apostolos Zaravinos Radojicic, Thomas Vrekoussis, Maria Kafousi, and A. Demetrios. Spandidos y efstathios n. *Stathopoulos MicroRNA expression analysis in triple-negative (ER, PR and Her*, 10(3):507–517, 2011.
- [19] Computational Biology . [ONLINE] Disponible en: <https://irp.nih.gov/our-research/scientific-focus-areas/computational-biology> [Consultado: 2018-05-07].
- [20] Warren Weaver. Science and Complexity. pages 1–11, 1948.
- [21] John H. Holland. Complex adaptive systems. *Daedalus*, 121(1):17–30, 1992.
- [22] Melanie Mitchell. *Complexity: A Guided Tour*. Oxford University Press, 2009.
- [23] Serena Chan. Complex Adaptive Systems. *Research Seminar in Engineering Systems*, pages 1–9, 2001.
- [24] S Boccaletti, G Bianconi, R Criado, and C I Genio. The structure and dynamics of multilayer networks. *Physics Reports*, 2014.
- [25] Lorena Caballero. La búsqueda del comienzo. El pensamiento complejo en biología. page 73, 2008.
- [26] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

- [27] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [28] Leandro Decastro. Fundamentals of natural computing: an overview. *Physics of Life Reviews*, 4(1):1–36, mar 2007.
- [29] Wayne Materi and David S Wishart. Computational systems biology in drug discovery and development: methods and applications. *Drug discovery today*, 12(7-8):295–303, April 2007.
- [30] Lian En Chai, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, 48(1):55–65, 2014.
- [31] OMS — Cáncer de mama: prevención y control. [ONLINE] Disponible en: <http://www.who.int/topics/cancer/breastcancer/es/index1.html> [Consultado: 2018-04-17].
- [32] INEGI. Estadísticas a propósito del Día Mundial contra el Cancer: Datos Nacionales. Technical report, 2018.
- [33] Research and Markets. Global Biosimulation Market 2014-2018 - Research and Markets.
- [34] C Valens. A Really Friendly Guide to Wavelets. 1999(C):23, 1999.
- [35] Mohamed Nounou, Hazem Nounou, Nader Meskin, and Aniruddha Datta. Wavelet-based multiscale filtering of genomic data. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, pages 804–809, 2012.
- [36] M N Nounou, H N Nounou, N Meskin, a Datta, and E R Dougherty. Multiscale denoising of biological data: a comparative analysis. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 9(5):1539–44, 2012.

- [37] Jayakishan Meher. Mixed PCA and Wavelet Transform based Effective Feature Extraction for Efficient Tumor Classification using DNA Microarray Gene Expression Data. 2:110–116, 2013.
- [38] Yves Meyer. Wavelets and operators. *Analysis at Urbana*, 1(1):256–365, 1989.
- [39] John H. Holland. Genetic Algorithms, 1992.
- [40] Julian D. Olden, Michael K. Joy, and Russell G. Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3-4):389–397, 2004.
- [41] Nigel Gilbert and Klaus G. Troitzsch. *Simulation for the Social Scientist*. 1999.
- [42] Aigli Korfiati, Konstantinos Theofilatos, Dimitrios Klefogiannis, Christos Alexakos, Spiros Likothanassis, and Seferina Mavroudi. Predicting human miRNA target genes using a novel computational intelligent framework. *Information Sciences*, 294:576–585, 2015.
- [43] Xiaowei Wang. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, 32(9):1316–1322, may 2016.
- [44] D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander. The microrna.org resource: Targets and expression, nucleic acids res., vol. 36, no. *SUPPL.*, 1, 2008.
- [45] Chih Hung Chou, Nai Wen Chang, Sirjana Shrestha, Sheng Da Hsu, Yu Ling Lin, Wei Hsiang Lee, Chi Dung Yang, Hsiao Chin Hong, Ting Yen Wei, Siang Jyun Tu, Tzi Ren Tsai, Shu Yi Ho, Ting Yan Jian, Hsin Yi Wu, Pin Rong Chen, Nai Chieh Lin, Hsin Tzu Huang, Tzu Ling Yang, Chung Yuan Pai, Chun San Tai, Wen Liang Chen, Chia Yen Huang, Chun Chi Liu, Shun Long Weng, Kuang Wen Liao, Wen Lian Hsu, and Hsien Da Huang. miRTarBase 2016: Updates to the experimentally validated

- miRNA-target interactions database. *Nucleic Acids Research*, 44(D1):D239–D247, jan 2016.
- [46] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016. PMID: 27653561.
- [47] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.
- [48] Tijmen Tielemand and Geoffrey Hinton. Lecture 6.5 - RMSProp, COURSERA: Neural Networks for Machine Learning. Technical report. 2012.
- [49] Julian D Olden and Donald A Jackson. Illuminating the black box : a randomization approach for understanding variable contributions in artificial neural networks. 154:135–150, 2002.
- [50] Martin Riedmiller. A Direct Adaptive Method for Faster Backpropagation Learning : The RPROP Algorithm.
- [51] Jinghua Du, Xuemin Niu, Yang Wang, Lingbo Kong, Rongqi Wang, Yuguo Zhang, Su-xian Zhao, and Yuemin Nan. MiR-146a-5p suppresses activation and proliferation of hepatic stellate cells in nonalcoholic fibrosing steatohepatitis through directly targeting Wnt1 and Wnt5a. *Scientific Reports*, 5(October):1–14, 2015.
- [52] Subramanian Kumar, Ranganathan Keerthana, Annamalai Pazhanimuthu, and Pachiap-pan Perumal. Overexpression of circulating miRNA-21 and miRNA-146a in plasma samples of breast cancer patients. *Indian Journal of Biochemistry and Biophysics*, 50(3):210–214, 2013.

- [53] Anna L Smith, Ritsuko Iwanaga, David J Drasin, Douglas S Micalizzi, Aik-choon Tan, Heide L Ford, Scientist Training Program, and Molecular Genetics. characteristics downstream of Six1 in human breast cancer. 31(50):5162–5171, 2013.
- [54] Mingli Han, Fang Wang, Yuanting Gu, Xinhong Pei, Guangcheng Guo, Chao Yu, Lin Li, Mingzhi Zhu, Youyi Xiong, and Yimeng Wang. MicroRNA-21 induces breast cancer cell invasion and migration by suppressing smad7 via EGF and TGF- β pathways. *Oncology Reports*, 35(1):73–80, 2016.
- [55] Yun Li, Wei Li, Zhe Ying, Han Tian, Xun Zhu, Jun Li, and Mengfeng Li. Metastatic heterogeneity of breast cancer cells is associated with expression of a heterogeneous TGF β -activating miR424-503 gene cluster. *Cancer Research*, 74(21):6107–6118, 2014.
- [56] Jingjing Liu, Yang Zhou, Zhendong Shi, Yunhui Hu, Tingting Meng, Xiaobei Zhang, Sheng Zhang, and Jin Zhang. microRNA-497 Modulates Breast Cancer Cell Proliferation, Invasion, and Survival by Targeting SMAD7. *DNA and Cell Biology*, 35(10):dna.2016.3282, 2016.
- [57] Olfat Gamil Shaker, Shereen Rashad Mohammed, Asmaa Mohammed Mohammed, and Zeinab Mahmoud. Impact of microRNA-375 and its target gene SMAD-7 polymorphism on susceptibility of colorectal cancer. *Journal of Clinical Laboratory Analysis*, 7(February):1–5, 2017.
- [58] Yang Yang, Shiao Ding, Gaojun Xu, Fei Chen, and Fangbao Ding. MicroRNA-15a inhibition protects against hypoxia/reoxygenation-induced apoptosis of cardiomyocytes by targeting mothers against decapentaplegic homolog 7. *Molecular Medicine Reports*, 15(6):3699–3705, 2017.
- [59] Jie Zhu, Ziqiang Zhang, Yitong Zhang, Wenshuai Li, Wanwei Zheng, Jianghong Yu, Bangting Wang, Lirong Chen, Qin Zhuo, Lin Chen, Jun Zhang, and Jie Liu. MicroRNA-

- 212 activates hepatic stellate cells and promotes liver fibrosis via targeting SMAD7. *Biochemical and Biophysical Research Communications*, (2017):1–8, 2018.
- [60] L. K. Zhuang, Y. T. Yang, X. Ma, B. Han, Z. S. Wang, Q. Y. Zhao, L. Q. Wu, and Z. Q. Qu. MicroRNA-92b promotes hepatocellular carcinoma progression by targeting Smad7 and is mediated by long non-coding RNA XIST. *Cell Death and Disease*, 7(4):1–12, 2016.
- [61] Zhihua Liu, Dan Yang, Ping Xie, Guomin Ren, Guibo Sun, Xu Zeng, and Xiaobo Sun. MiR-106b and miR-15b modulate apoptosis and angiogenesis in myocardial infarction. *Cellular Physiology and Biochemistry*, 29(5-6):851–862, 2012.
- [62] Jianhua Xu, Qin Ai, Hanhai Cao, and Quan Liu. MiR-185-3p and miR-324-3p Predict Radiosensitivity of Nasopharyngeal Carcinoma and Modulate Cancer Cell Growth and Apoptosis by Targeting SMAD7. *Medical science monitor : international medical journal of experimental and clinical research*, 21:2828–36, 2015.
- [63] Scott B Reeder, Houchun H Hu, Claude B Sirlin, Liver Imaging Group, and San Diego. Suppression of cell growth and invasion by miR-205 in breast cancer. 36(5):1011–1014, 2016.
- [64] Chung-Man Leung, Ting-Wen Chen, Sung-Chou Li, Meng-Ru Ho, Ling-Yueh Hu, Wen-Shan Liu, Tony T Wu, Ping-Chi Hsu, Hong-Tai Chang, and Kuo-Wang Tsai. MicroRNA expression profiles in human breast cancer cells after multifraction and single-dose radiation treatment. *Oncology reports*, 31(5):2147–56, 2014.
- [65] Qingbao Cheng, Bin Yi, Aihua Wang, and Xiaoqing Jiang. Exploring and exploiting the fundamental role of microRNAs in tumor pathogenesis. *OncoTargets and Therapy*, 6:1675–1684, 2013.

- [66] Andrea Ravelli, James M. Reuben, Francesco Lanza, Simone Anfossi, Maria Rosa Cappelletti, Laura Zanotti, Angela Gobbi, Chiara Senti, Paola Brambilla, Manuela Milani, Daniele Spada, Paolo Pedrazzoli, Massimo Martino, Alberto Bottini, and Daniele Generali. Breast cancer circulating biomarkers: advantages, drawbacks, and new insights. *Tumor Biology*, 36(9):6653–6665, 2015.
- [67] Chen-Dan Zou, Wei-Ming Zhao, Xiao-Na Wang, Qiang Li, Hui Huang, Wan-Peng Cheng, Jian-Feng Jin, He Zhang, Ming-Juan Wu, Sheng Tai, Chao-Xia Zou, and Xu Gao. MicroRNA-107: a novel promoter of tumor progression that targets the CPEB3/EGFR axis in human hepatocellular carcinoma. *Oncotarget*, 7(1):266–278, 2016.
- [68] S Vimalraj, P J Miranda, B Ramyakrishna, and N Selvamurugan. Regulation of Breast Cancer and Bone Metastasis by MicroRNAs. 35(5):369–387, 2013.
- [69] María Ángeles Castilla, Juan Díaz-Martín, David Sarrió, Laura Romero-Pérez, María Ángeles López-García, Begoña Vieites, Michele Biscuola, Susana Ramiro-Fuentes, Clare M. Isacke, and José Palacios. MicroRNA-200 Family Modulation in Distinct Breast Cancer Phenotypes. *PLoS ONE*, 7(10), 2012.