Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional
Unidad Zacatenco

Departamento de Computación

Algoritmos de Teoría de la Información para analizar redes genéticas e identificar genes principales en cáncer de seno

Tesis que presenta

Moises Omar León Pineda

para obtener el Grado de

Maestría en Ciencias en Computación

Director de Tesis

Dr. Matías Alvarado Mentado

Ciudad de México **11 de Agosto de 2023**

Resumen

El cáncer es un desafío importante para la salud pública a nivel mundial debido a su rápido aumento en las últimas décadas. Los tumores primarios pueden dar lugar a réplicas cancerosas o metástasis heterogéneas, que son extremadamente difíciles de controlar e inhibir, y son la principal causa de muerte en pacientes con cáncer. En esta tesis, se utiliza información de conjuntos de biopsias de pacientes con cáncer de mama disponibles en las bases de datos GEO y GDCDP para desarrollar redes genéticas. Estas redes se crean para tres tipos de tejidos: 1) tejido mamario sin cáncer, 2) tejido mamario con tumor primario de cáncer y 3) tejido de la primera metástasis del cáncer de mama en los ganglios linfáticos. Se aplica el algoritmo ARACNe, basado en las ecuaciones de entropía e información mutua, a los perfiles de expresión genética de las biopsias utilizando su versión multinúcleo. Esto permite obtener matrices de co-expresíon genética y posteriormente generar las redes genéticas utilizando herramientas computacionales específicas. Mediante el análisis estructural de estas redes genéticas, se identifican los genes nodos con más información mutua en comparación con otros genes y con mayor grado dentro de la red, respectivamente. Se analiza la importancia de estos genes clave en los tres tipos de tejidos mediante el estudio de los principales procesos biológicos y funciones moleculares en los que participan. Además, se realiza una revisión de la escasa literatura existente sobre dichos genes. El aporte de esta tesis radica en la metodología utilizada para crear las redes, identificar los nodos con mayor grado e información mutua en los tres tipos de tejidos, que puede ser aplicable a cualquier tipo de cáncer, y validar su papel relevante en cada tipo de tejido utilizando herramientas de análisis genético, así como determinar su importancia en la formación del cáncer.

Palabras clave: Cáncer, metástasis del cáncer, teoría de la información, simulación computacional, redes genéticas.

Abstract

Cancer represents a significant challenge to global public health due to its accelerated growth in recent decades. Primary tumors can give rise to cancerous replications or heterogeneous metastases, which are extremely difficult to control and inhibit and are the leading cause of death in cancer patients. This thesis uses information from sets of breast cancer patient biopsies available in the GEO and GDCDP databases to develop genetic networks. These networks are created for three types of tissues: 1) non-breast cancer tissue, 2) breast cancer primary tumor tissue, and 3) breast cancer metastasis in lymph nodes. The ARACNe algorithm, based on entropy equations and mutual information, is applied in its multi-core version to the gene expression profiles of the biopsies. This allows for obtaining the genetic co-expression matrices and generating the genetic networks using specific computational tools. Through the structural analysis of these genetic networks, the gene nodes with the highest mutual information with other genes and the highest degree within the network are identified, respectively. The importance of these key genes in the three types of tissues is analyzed by studying the main biological processes and molecular functions in which they participate. The analysis is supplemented by reviewing the -limited- literature on these genes. The contribution of this thesis lies in the methodology used to create the networks, identify the nodes with the highest degree and mutual information in the three types of tissues, which can be applicable to any type of cancer, and validate their relevant role in each type of tissue using genetic analysis tools. Additionally, it aims to determine their relevance in cancer formation. Keywords: Cancer, cancer metastasis, information theory, computational simulation, genetic networks.

Keywords: Cancer, cancer metastasis, information theory, computational simulation, genetic networks.

Agradecimientos

Me gustaría agradecer a las siguientes personas e instituciones por su invaluable contribución y apoyo durante la realización de esta tesis:

- En primer lugar, deseo expresar mi profundo agradecimiento al CONACyT por otorgarme la beca que hizo posible llevar a cabo esta investigación. Su generoso apoyo financiero fue fundamental para el desarrollo de este proyecto.
- También quiero extender mi agradecimiento al CINVESTAV y al departamento de Computación por brindarme un entorno académico propicio para llevar a cabo mi investigación. Además agradezco a todos los doctores del Departamento por su guía y orientación durante todo el proceso y por ser parte en mi formación académica.
- A mi asesor de tesis, Dr. Alvarado, por su experiencia y orientación constante, estoy agradecido por su tiempo y esfuerzo invertido.
- También quiero reconocer y agradecer al Dr. Irving Martínez por su valiosa retroalimentación y consejos, los cuales contribuyeron significativamente a mejorar la calidad de esta tesis.
- A mi familia, quiero expresarles mi más sincero agradecimiento por su incondicional apoyo a lo largo de este camino. Su amor, comprensión y aliento constante fueron pilares fundamentales en mi búsqueda del conocimiento.
- Por último, pero no menos importante, quiero agradecer a mis amigos por su compañía, motivación y palabras de aliento en momentos clave de este proyecto. Su amistad ha sido un verdadero regalo que valoro enormemente.
- A todos y cada uno de ustedes, gracias de corazón. Su contribución y apoyo han dejado una huella imborrable en mi vida y en el desarrollo de esta tesis.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El cáncer es una enfermedad devastadora que representa un desafío importante en el ámbito de la salud. En México, en el año 2020, se registraron más de un millón de defunciones, y los tumores malignos fueron responsables de una proporción significativa de estas muertes. En particular, los tumores malignos de mama representaron el 8 % del total de fallecimientos, lo que resalta la gravedad de este tipo de cáncer. Además, la metástasis, que es la propagación del cáncer a otras partes del cuerpo, es la principal causa de muerte en pacientes con cáncer. Estas estadísticas reflejan la necesidad urgente de abordar el estudio del cáncer de mama y la identificación de los genes clave involucrados en su desarrollo y metástasis [1, 2, 3].

El cáncer es una enfermedad compleja y heterogénea, influenciada por factores genéticos y epigenéticos. Esta tesis se enfoca en la identificación de genes clave en el desarrollo del tumor de cáncer de seno y su primera metástasis en los ganglios linfáticos. Gen éticamente, el cáncer surge como resultado de la acumulación de mutaciones en los genes que regulan la división y la muerte celulares programada. La identificación de los genes especifico involucrados en la formación del cáncer y su metástasis es un desafío fundamental en la investigación del cáncer. Esta identificación proporciona información importante sobre los procesos biológicos subyacentes y puede facilitar la identificación de posibles terapias.

El cáncer de seno afecta a millones de mujeres en todo el mundo y es la principal causa de muerte por cáncer en mujeres, representando un desafío médico y científico de gran importancia. Comprender los mecanismos mo-

leculares y genéticos subyacentes en el cáncer de seno es fundamental para el desarrollo de estrategias de diagnóstico y tratamiento más efectivas. En los últimos años, se ha acumulado una gran cantidad de datos Genómicos y moleculares relacionados con el cáncer de seno, los cuales brindan información valiosa sobre la función de los genes y las interacciones entre ellos en las redes genéticas. Sin embargo, el análisis de estos datos masivos presenta desafíos significativos debido a su complejidad y dimensiones.

Los datos utilizados en este trabajo provienen del Genomic Data Commons Data Portal (GDCDP) del National Cancer Institute (NCI) [4]. Este portal proporciona un repositorio unificado y una base de conocimientos que facilita el intercambio de datos entre estudios genómicos del cáncer. El GDCDP respalda varios programas del genoma del cáncer en el Center for Cancer Genomics (CCG) del NCI, incluido el Cancer Genome Atlas (TCGA) y el Therapeutically Applicable Research to Generate Effective Treatments (TARGET). Las aplicaciones del GDC incluyen herramientas de análisis, visualización y exploración de datos (DAVE), así como la transferencia de datos.

El Gene Expression Omnibus (GEO) [5] es un repositorio público de datos de genómica funcional que admite el envío de datos compatibles con la Minimum Information About a Microarray Experiment (MIAME), que es un estándar creado por la Functional Genomics Data Society (FGED) para informar sobre experimentos de microarreglos. Se aceptan datos basados en matrices y secuencias, y se proporcionan herramientas para que los usuarios consulten y descarguen experimentos y perfiles de expresión génica seleccionados. Esta tesis se desarrolla utilizando datos de biopsias disponibles en la base de datos pública GEO.

La identificación de los genes principales en el tumor primario y la metástasis es crucial para comprender los mecanismos subyacentes del cáncer y desarrollar terapias más efectivas. El cáncer es una enfermedad compleja y multifactorial que involucra una interacción compleja entre factores genéticos, ambientales y epigenéticos. Identificar los genes específicos que contribuyen al desarrollo del tumor primario y su propagación a través de la metástasis es un desafío clave en la investigación del cáncer. En los

últimos años, los avances en el estudio de las enfermedades han tenido un mayor alcance gracias al desarrollo de modelos matemáticos y computacionales [6, 7, 8, 9, 10], los cuales han sido de gran ayuda para diseñar, analizar e implementar experimentos *in silico*, reduciendo así las costosas experimentaciones *in vitro* en términos de recursos humanos, económicos y de tiempo.

En este trabajo de tesis, se propone utilizar métodos y algoritmos avanzados basados en la Teoría de la Información y el análisis de redes genéticas para identificar los genes principales en el tumor primario y la metástasis del cáncer. Estos enfoques permiten evaluar la información mutua entre los genes, la distribución probabilística del grado de interconexión entre los genes y analizar las matrices de expresión genética. Al utilizar estas técnicas, podemos descubrir patrones funcionales y vías de señalización importantes en el desarrollo y la progresión del cáncer.

En el análisis de las redes genéticas, se identificarán los genes con más información mutua y grado de interconexión, lo que indica su importancia en el desarrollo del cáncer. Además, se explorarán las funciones moleculares, los procesos biológicos, las clases de proteínas y las vías de señalización asociadas a estos genes principales. Se utilizarán herramientas computacionales como NetworkX para realizar el análisis y visualización de las redes genéticas.

Con el objetivo de esta tesis, se aplicaron algoritmos de Teoría de la Información para generar y analizar redes genéticas, y posteriormente identificar los genes principales en el cáncer de seno y su importancia en el desarrollo del tumor primario y la primera metástasis. Se espera que los resultados de este trabajo de tesis contribuyan al conocimiento del cáncer de seno y proporcionen información valiosa para mejorar el diagnóstico y tratamiento de esta enfermedad.

Este trabajo de tesis busca contribuir al conocimiento del cáncer al identificar los genes principales en el tumor primario y la metástasis del cáncer. Los resultados obtenidos proporcionarán una comprensión más profunda de los mecanismos moleculares y genéticos subyacentes en el desarrollo y progresión de esta enfermedad.

En el próximo capítulo, se revisará el estado del arte y el marco teórico relacionado con la Teoría de la Información, los métodos computacionales para la identificación de cáncer y el análisis de redes genéticas. Se explorarán conceptos como los grafos, la distribución probabilística del grado y las matrices de expresión genética para sentar las bases teóricas de este estudio.

1.1. Problema: Identificación de los genes principales en los tejidos del cáncer de mama

El problema de investigación abordado en este trabajo de tesis consiste en responder a las siguientes preguntas:

- ¿Cómo se pueden utilizar de manera efectiva los algoritmos de Teoría de la Información para generar redes genéticas y posteriormente identificar los genes principales en el cáncer de seno?

Específicamente, se plantean las siguientes preguntas de investigación:

- ¿Qué algoritmos son útiles para la identificación del cáncer? ¿Cómo funcionan? ¿Cuáles son sus aportes y limitaciones?
- ¿Cómo se pueden identificar los principales genes implicados en la formación del cáncer? ¿Cómo varían según el tipo de tejido?
- ¿Existe suficiente información disponible para desarrollar un algoritmo capaz de identificar los genes principales en la formación del cáncer?

1.2. Propuesta de solución

En esta tesis, se propone resolver el problema planteado mediante el desarrollo y la aplicación de algoritmos basados en la Teoría de la Información. Estos algoritmos permitirán crear redes genéticas y posteriormente analizarlas para identificar los genes principales en el cáncer de seno. Los

algoritmos se fundamentan en principios y métricas de la Teoría de la Información, como la entropía, la información mutua y las redes de información. La propuesta de solución se divide en los siguientes pasos:

1. Recopilación de datos: Obtención de información de diversas bases de datos.
2. Preprocesamiento de datos: Generación de un diccionario y normalización de la información.
3. Construcción de redes genéticas: Desarrollo de una metodología para crear las redes genéticas.
4. Análisis de redes: Realización de un análisis estadístico para evaluar los resultados.
5. Validación y evaluación: Comparación de los resultados obtenidos con experimentos previamente validados en la literatura.
6. Interpretación de resultados: Discusión de los resultados obtenidos.
7. Informe y difusión: Elaboración de la tesis y redacción de artículos científicos para su divulgación.

1.3. Hipótesis y objetivos

Se plantea la hipótesis:

La aplicación de algoritmos basados en Teoría de la Información para la generación de redes genéticas, junto con el análisis estructural de estas redes, permitirá identificar genes relevantes para la formación del cáncer de mama y su metástasis.

Objetivo General: Con algoritmos basados en Teoría de la Información identificar los genes que tienen una mayor relevancia en la formación del cáncer de mama y su metástasis. Se hará análisis de información genética obtenida de las biopsias de tejido libre de cáncer, tumor primario y primera

metástasis del cáncer de mama.

Objetivos Particulares:

1. Describir el estado del arte sobre identificación de cáncer mediante la aplicación de herramientas computacionales.
2. Obtener información sobre cáncer y metástasis proveniente de diferentes bases de datos. Normalizar la información obtenida y generalizar las bases de datos para la creación de un diccionario.
3. Desarrollar e implementar una metodología para la generación de redes genéticas a partir de datos de biopsias.
4. Realizar un análisis estadístico mediante la simulación repetida en la generación de las redes genéticas con el propósito de recopilación, organización, presentación y análisis de datos obtenidos.
5. Evaluar la calidad y robustez de las redes genéticas generadas mediante el análisis estadístico.
6. Identificar los nodos clave y las interacciones más importantes en las redes genéticas generadas y validar su relevancia biológica mediante comparación con datos de la literatura.

1.4. Metodología

Aplicando conceptos de Teoría de la Información, como entropía, información mutua, distribución y redundancia, y codificados en algoritmos, analizamos la información de las bases de datos GEO y GDC sobre matrices de expresión genética obtenidas mediante microarreglos de biopsias de tejidos de pacientes con cáncer. A partir de estas matrices, generamos redes genéticas y, mediante su análisis, identificamos los genes principales involucrados en los procesos genéticos del cáncer. La identificación de genes relevantes en el cáncer de mama a través de la generación de redes genéticas proporciona información valiosa sobre el comportamiento de esta

enfermedad. El trabajo de tesis se desarrolló en las siguientes etapas, que se describen en detalle en el Capítulo 3:

1. Recopilación y preparación de datos.
2. Creación del diccionario de variables y genes.
3. Normalización y combinación de datos.
4. Construcción de redes genéticas.
5. Análisis topológico de las redes.
6. Identificación de genes principales y funciones biológicas.
7. Análisis estadístico de las asociaciones gen-función.
8. Validación funcional de los genes principales.
9. Interpretación de resultados y conclusiones.

En esta tesis se presentan los resultados obtenidos de acuerdo con la metodología descrita, las funciones biológicas asociadas a los genes principales y las conclusiones del estudio. Los hallazgos se han difundido y se continuará con esta labor en conferencias científicas especializadas, además de publicar los resultados en revistas relevantes en el campo de estudio. Las contribuciones de esta tesis se describen en la siguiente sección las siguientes.

1.5. Contribuciones

1. Desarrollo basado en el uso de algoritmos de Teoría de la Información para el análisis de redes genéticas para investigar el cáncer de seno. Esta integración permite revelar patrones y relaciones en los datos genómicos y moleculares asociados con esta enfermedad.

2. Creación de un diccionario para comparar y unificar diferentes bases de datos genéticas y moleculares. Esto garantizará la consistencia y confiabilidad de los resultados obtenidos al analizar la información proveniente de diversas fuentes.

3. Análisis topológico de las redes genéticas: para identificar las centralidades de los nodos y los grupos de genes altamente conectados, proporcionando una comprensión más profunda de la estructura y organización de las interacciones genéticas relacionadas con el cáncer de seno.

4. Identificación de los genes principales y sus funciones biológicas: con la aplicación de algoritmos de Teoría de la Información y análisis topológico en las redes genéticas del cáncer de seno, se comprenden mejor los mecanismos subyacentes de la enfermedad.

5. Análisis estadístico de las asociaciones gen-función: útil para evaluar la significación de las asociaciones identificadas entre los genes principales y las funciones biológicas. Es una evaluación cuantitativa de la relación entre los genes y su relevancia biológica en el contexto del cáncer de seno.

6. Validación funcional de los genes principales: Se compararon los resultados obtenidos con experimentos funcionales para validar la función biológica de los genes principales identificados para el cáncer de seno y de esa forma se respaldan aún más los hallazgos obtenidos a través del análisis de datos genómicos.

7. Comunicación de resultados: Además del trabajo de tesis, de esta investigación se han realizado las siguientes publicaciones:

   - ***Genetic Network of Breast Cancer Metastasis in Lymph Nodes via Information Theory Algorithms[11]***
     - M. Alvarado, I. Valdespin, M. León and S. A. Alcalá-Corona, "Genetic Network of Breast Cancer Metastasis in Lymph Nodes via Information Theory Algorithms,"2022 19th Internatio-

nal Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), México City, México, 2022, pp. 1-6.

- ***Patterns in Genesis of Breast Cancer Tumor[12]***
  - León, M., Alvarado, M. (2023). Patterns in Genesis of Breast Cancer Tumor. In: Rodríguez-González, A.Y., Pérez-Espinosa, H., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Olvera-López, J.A. (eds.) Pattern Recognition. MCPR 2023. Lecture Notes in Computer Science, vol. 13902. Springer, Cham.
- ***Main genes in breast cancer primary tumor and first metastasis in lymph nodes revealed by information-theory-based genetic networks analysis*** **[13]**
  - Irving Martínez-Vargas, Moises León-Pineda, Matías Alvarado-Mentado et al. Main genes in breast cancer primary tumor and first metastasis in lymph nodes revealed by information-theory-based genetic networks analysis, 03 jul. 2023, PREPRINT (Version 1) available at Research Square [`https://doi.org/10.21203/rs.3.rs-3126195/v1`]

8. Además con el artículo ***Patterns in Genesis of Breast Cancer Tumo***, se obtuvo el premio ***MCPR-IAPR Student Paper*** lo que nos dio la oportunidad de ser invitados a enviar un artículo extenso a la sección especial dedicada a MCPR en la revista *Pattern Recognition Letters*.

Capítulo 2

Estado del arte y marco teórico

En este estado del arte, se revisan artículos clásicos y nuevos sobre el uso de algoritmos de detección de cáncer, el modelado y la generación de redes genéticas. La detección temprana del cáncer es un aspecto crítico en la lucha contra esta enfermedad. Los métodos de diagnóstico actuales, como la biopsia y la tomografía computarizada tienen limitaciones, tal como la detección de síntomas tardíos. En consecuencia, los investigadores han estado buscando nuevas formas de detectar el cáncer en etapas iniciales, y los algoritmos de detección de cáncer se han convertido en una herramienta importante en este campo.

En los últimos años, se han publicado varios artículos que se centran en el uso de algoritmos de aprendizaje automático para identificar patrones en los datos de los pacientes que pueden indicar la presencia de cáncer. Estos algoritmos pueden analizar grandes cantidades de datos de pacientes, como resultados de pruebas de laboratorio e imágenes médicas, y detectar patrones que podrían indicar la presencia de cáncer en sus etapas tempranas. Los algoritmos de detección de cáncer han demostrado ser altamente efectivos en la identificación temprana del cáncer en diversos tipos de cáncer, como el de mama, pulmón y próstata. Además, estos algoritmos pueden actualizarse continuamente con nuevos datos para mejorar su precisión y eficacia. El objetivo principal es promover la comprensión científica de este proceso complejo mediante la generación de comportamientos emergentes in sillico, que son difíciles de observar directamente in vivo o in vitro.

Es oportuno decir que los datos son hechos o registros objetivos y sin

procesar que representan observaciones, medidas o símbolos. Los datos pueden ser numéricos, textuales, visuales u otro tipo de representación. Por otro lado, la información se obtiene al procesar y organizar los datos de manera significativa y relevante. La información implica dar estructura, contexto y sentido a los datos, lo que permite extraer conocimientos y comprender su significado. La información es el resultado de analizar y relacionar los datos para formar un mensaje coherente y comprensible [14].

La Teoría de la Información (TI) se ocupa de cuantificar, medir y transmitir la información de manera eficiente [15, 16, 17]. Claude Shannon la desarrolló en la década de 1940 [15] para estudiar la comunicación y la transmisión de señales. Un concepto fundamental en la TI es la entropía, que representa la incertidumbre asociada a una fuente de información. Cuanto mayor es la entropía, mayor es la cantidad de información contenida en el mensaje.

La aplicación de la TI permite la compresión de datos [15, 18], lo que reduce la cantidad de información necesaria para representar un conjunto de datos al encontrar patrones de redundancia y desarrollar algoritmos de compresión. Por otro lado, la TI permite el desarrollo de la codificación de la información [19, 20], especialmente en la implementación de códigos de corrección de errores que mejoran la fiabilidad de la comunicación y el almacenamiento de datos al detectar y corregir errores. Recientemente, la TI cuántica [21, 22], fusión de la TI con los principios de la mecánica cuántica, la información se basa en qubits o bits cuánticos, que pueden existir en múltiples estados simultáneos gracias a la superposición cuántica. Esto ha llevado al desarrollo de la criptografía cuántica y al diseño de algoritmos de computación cuántica más rápidos y eficientes.

Además, la TI ha encontrado aplicaciones en el campo del aprendizaje automático [23], donde se utiliza para analizar y abordar la selección de características. Asimismo, las redes de información [24, 25] son modelos que describen cómo la información se propaga y procesa en las redes sociales, las redes neuronales y las redes de comunicación. Estos modelos han sido utilizados para estudiar fenómenos como la difusión de información, la propagación de rumores y la detección de comunidades. La propuesta en

esta tesis aplica la TI en la biología de sistemas [26, 27, 28].

2.1. Teoría de la información y entropía para análisis de datos

Con algoritmos de TI se hace la generación de redes genéticas y la creación de un modelo donde el comportamiento emergente es impulsado por el crecimiento y las interacciones probabilísticas entre los genes. Las simplificaciones en el modelo para que sea manejable y comprensible no eliminan la información esencial sobre los tejidos cancerosos.

Para el desarrollo de este trabajo, el algoritmo que se utiliza aplica los conceptos de entropía e información mutua. Para especificar un sistema físico se requiere reducir la energía no útil o entropía.

La información es, modelada como entidad física, *una secuencia de símbolos que fluyen en un canal*. La TI se aboca a estudiar métodos de envío de cierta cantidad de información útil a través de un canal. Para extraer la información esencial del estado actual de un sistema requiere reducir su entropía de información.

La entropía $S(t)$ es el concepto fundamental para cuantificar la información de un sistema [29]. El antecedente de la entropía en la teoría de la información es la entropía de la termodinámica física, la energía no útil para el trabajo o energía desperdiciada [30]. Un sistema físico se caracteriza por tener la entropía mínima, lo que significa que se encuentra en estados donde el sistema tiene el mínimo de entropía. La ecuación de la entropía es:

$$S(t) = -\kappa \sum_i p_i(t) log p_i(t); \tag{2.1}$$

dondeκ es la constante de Boltzmann, propuesta por primera vez por el genio Ludwig Boltzmann en el siglo XIX y escrita en la superficie de su pulgar y $p_i(t)$ es la probabilidad de que S se encuentre en el estado i.

La entropía del sistema completo agrega las probabilidades de las variables que representan la ocurrencia de cada estado del sistema, ponderadas con el logaritmo de cada probabilidad. Una entropía menor en un sistema de información significa una menor cantidad de información no útil para caracterizar lo esencial de los estados del sistema.

La entropía permite cuantificar la información mutua (IM) entre las variables x, y en el sistema. La ecuación de IM es la siguiente (2.2):

$$I(g_i, g_j) = S(g_i) + S(g_j) - S(g_i, g_j). \tag{2.2}$$

donde $S(x, y)$ denota la entropía conjunta de x, y. Además de las correlaciones lineales, el aspecto destacado es que IM cuantifica las correlaciones no lineales entre pares de variables. Las variables son independientes si y solo si su IM es igual a cero, por lo que se cumple la siguiente ecuación:

$$S(x) + S(y) = S(x, y). \tag{2.3}$$

En este sistema de red genética, el análisis se basa en cuantificar la MI entre pares de genes, g_i, g_j; por lo tanto, en los valores de co-expresíon aplicando:

$$I(g_i, g_j) = S(g_i) + S(g_j) - S(g_i, g_j). \tag{2.4}$$

A continuación, se caracteriza el cáncer específico en las muestras de biopsias. La GEP (pérdida de expresión génica) entre pares con conexiones de un tercer gen se cuantifica en relación con la desigualdad de IM [31]:

$$I(g1, g2), I(g2, g3). \tag{2.5}$$

Esto significa que las interacciones de valores más bajos pueden no considerarse, sin perder conexiones esenciales en la red genética, hasta un umbral determinado. Así, podar los valores de IM más bajos no afecta la caracterización de la red genética. Después de esta poda, el archivo de salida asociado a la red genética es un archivo liviano que alivia la carga computacional, pero mantiene las centralidades de la red genética [32].

Para el problema a resolver utilizamos, en la primera etapa, ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks), algoritmo de aprendizaje automático, para procesar los datos e inferir las relaciones entre los genes a partir de datos de expresión génica. Este algoritmo se basa en la TI para identificar los pares de genes que tienen una relación significativa. ARACNe multi-core es una versión paralelizada del algoritmo ARACNe que utiliza múltiples núcleos de procesamiento para acelerar el cálculo de las redes de interacción génica. En lugar de analizar todos los pares de genes de manera secuencial, ARACNe multi-core divide los cálculos entre varios núcleos de procesamiento, lo que permite que el análisis se realice más rápidamente, muy útil en la comprensión de las complejas interacciones entre los genes en las células; y consta de tres etapas principales:

- Cálculo de correlaciones: se calculan las correlaciones entre todos los pares de genes en el conjunto de datos de expresión génica utilizando la teoría de la información.
- Eliminación de enlaces espurios: se utilizan técnicas de estadística y teoría de la información para eliminar los enlaces espurios, detectado en la etapa anterior, para garantizar que solo se conserven las interacciones relevantes y significativas.
- Construcción de la red de interacción: En esta etapa, se construye la red de interacción final a partir de las interacciones significativas que se han conservado en la etapa anterior.

La tabla 2.1 proporciona una visión general de tres algoritmos, su enfoque, ventajas y limitaciones en la identificación de cáncer. Estos algoritmos

se eligieron por estar basados en la Teoría de la información. Existen otros basados en otras teorías, asimismo, útiles en la investigación del cáncer; en la tabla 2.2 se muestra algunos de tales enfoques.

Algoritmo	**Enfoque**	**Ventajas**	**Limitaciones**
ARACNe [31]	Inferencia de redes de regulación génica utilizando datos de expresión génica.	- Capacidad para identificar relaciones de regulación génica precisas. - Robusto al ruido y a datos incompletos.	- Requiere gran cantidad de datos de expresión génica. - Sensible a falsos positivos por ausencia de la dirección de las relaciones.
MIFS [33] (Mutual Information Feature Selection)	Selección de características basada en la información mutua entre características y la variable objetivo.	- Maneja conjuntos de datos con gran número de características. - Identifica características relevantes y reduce la dimensionalidad.	- Sufre de sobreajustes en conjuntos de datos pequeños. - Puede ser computacionalmente costoso para conjuntos grandes de datos.
ReliefF [34]	Algoritmo basado en vecinos cercanos para estimar la importancia discriminativa de características.	- Robusto frente a ruido y datos incompletos. - Manejo de datos con redundancias.	- No maneja datos continuos directamente. - Sensible a la elección del número de vecinos.

Tabla 2.1: Comparativo entre ARACNe, MIFS y ReliefF en la identificación de cáncer

2.2. Métodos computacionales para la identificación de cáncer

La minería de datos es una disciplina que combina la estadística, la informática y la inteligencia artificial para analizar grandes cantidades de datos y extraer información valiosa. Permite encontrar patrones y relaciones que no son evidentes a simple vista y que pueden ser utilizados para tomar decisiones informadas. Los datos pueden provenir de diversas fuentes, como archivos de texto, redes sociales y transacciones en línea. La minería de datos utiliza una variedad de técnicas y algoritmos para analizar los datos, como el análisis de clustering, el análisis de regresión, la clasificación, la detección de anomalías y la visualización de datos. Estas técnicas permiten

identificar patrones y relaciones en los datos, como tendencias de ventas, preferencias de los clientes y perfiles de comportamiento. A continuación, se describen los métodos más comunes utilizados en la minería de datos:

- Árboles de decisión: se construye mediante la selección de variables que mejor separan los datos en grupos. Estos grupos se utilizan para crear una estructura de árbol, que se puede utilizar para clasificar nuevos datos.
- Clustering: utilizado para agrupar datos basados en la similitud; dividen los datos en grupos que son similares entre sí, y diferentes a los demás grupos.
- Reglas de asociación: para descubrir patrones de asociación entre diferentes variables en un conjunto de datos. Pueden ser utilizadas para predecir la probabilidad de que una variable ocurra dado el conocimiento de otra variable.
- Redes neuronales: es un algoritmo que imita la forma en que funciona el cerebro humano. Estas redes se utilizan para clasificar, predecir y modelar datos complejos. Las redes neuronales aprenden de los datos y ajustan sus pesos para mejorar la precisión de las predicciones.
- Análisis de regresión: para analizar la relación entre dos o más variables. Se utiliza para predecir los valores de una variable en función de los valores de otras variables.
- Análisis de texto: para extraer información de grandes cantidades de texto. Utiliza técnicas de procesamiento de lenguaje natural para identificar patrones y relaciones entre las palabras en el texto.

En esta tesis, se lleva a cabo la minería de datos en las plataformas GDCDP (Genomic Data Commons Data Portal) [4] y GEO (Gene Expression Omnibus) [5]. GDC alberga una gran cantidad de datos genómicos y clínicos de diferentes tipos de cáncer, recopilados de diversos proyectos de investigación previos. Los datos incluyen información sobre perfiles de

expresión génica, datos de secuenciación del ADN y ARN, datos de metilación, datos de anotaciones clínicas y mucho más. Con diversas técnicas de minería de datos, como análisis de expresión génica diferencial, identificación de biomarcadores, análisis de correlación de datos clínicos y genómicos, y clasificación de subtipos de cáncer. GEO es una base de datos pública que almacena una amplia gama de perfiles de expresión génica de diferentes tipos de muestras biológicas, incluidos datos relacionados con el cáncer. Proporciona acceso a datos de microarreglos y secuenciación de ARN de alta calidad, así como información asociada, como anotaciones y metadatos. Estos datos son útiles para identificar patrones de expresión génica, descubrir genes diferencialmente expresados en diferentes tipos de cáncer, realizar análisis de vías biológicas y explorar correlaciones entre genes y condiciones clínicas.

2.3. Redes genéticas(RG)

El análisis de redes genéticas se centra en el desarrollo de técnicas y herramientas para comprender y analizar la complejidad de las interacciones genéticas y su impacto en la función biológica. A continuación, se presentan algunos aspectos clave del análisis de redes genéticas:

1. **Construcción de redes genéticas:** con diversos enfoques, a partir de datos: experimentales de expresión génica, de interacción proteína-proteína y de interacción gen-regulador. Estos enfoques incluyen métodos basados en correlación, inferencia de redes bayesianas, algoritmos de aprendizaje automático y técnicas de análisis de co-expresíon.

2. **Análisis topológico de redes genéticas:** son métricas y algoritmos para revelar características estructurales y funcionales importantes. Esto incluye la identificación de nodos y enlaces clave, la detección de comunidades y la caracterización de los patrones de conectividad.

3. **Predicción de funciones génicas y descubrimiento de módulos funcionales:** para predecir las funciones de genes desconocidos

y descubrir módulos funcionales o vías biológicas relevantes. Esto se logra mediante técnicas de enriquecimiento funcional, análisis de enriquecimiento de genes, análisis de perfiles de expresión y métodos de agrupamiento basados en la topología de la red.

4. **Modelado dinámico y simulación de redes genéticas:** modelos matemáticos y algoritmos de simulación para estudiar la dinámica de las redes genéticas y predecir el comportamiento de los sistemas biológicos. Estos modelos incluyen redes booleanas, redes de regulación génica y modelos de redes de interacción proteína-proteína.

5. **Visualización y herramientas de análisis:** El análisis de redes genéticas también implica el desarrollo de herramientas y software especializados para visualizar y analizar las redes genéticas. Estas herramientas proporcionan interfaces interactivas y visualizaciones intuitivas para explorar y comprender la estructura y función de las redes genéticas.

En resumen, el análisis de redes genéticas involucra el desarrollo de enfoques computacionales avanzados, modelos matemáticos y herramientas de visualización para comprender la complejidad de las interacciones genéticas y su impacto en la función biológica. Estos avances permiten descubrir nuevos conocimientos sobre los mecanismos moleculares y las bases genéticas de las enfermedades. En la tabla 2.3 se analizan aspectos de las redes genéticas se hace la descripción y se resalta la importancia de cada uno.

2.3.1. Grafos

Para el estudio sistemático de las redes complejas se aplica la teoría de grafos [35, 24, 36]. Los grafos son objetos matemáticos constituidos por un conjunto de nodos o vértices conectados mediante aristas. La teoría de grafos es una rama de las matemáticas que comenzó a desarrollarse en el siglo XVIII. Fue el gran genio Leonhard Euler (1707-1783) quien primero resolvió un problema combinatorio y definió algunas propiedades estructurales de los grafos.

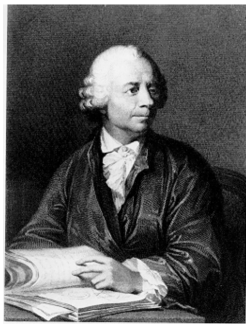


Figura 2.1: Retrato de L. Euler (1707-1783), fundador de la teoría de grafos.

A continuación, se dan definiciones formales de un grafo y otros conceptos relacionados:

- **Grafo:** Un grafo se define formalmente como un par ordenado $G = (V, E)$, donde V es un conjunto finito de elementos llamados vértices (o nodos) y E es un conjunto de pares no ordenados de vértices, llamados aristas. Matemáticamente, se puede expresar como $G = (V, E)$, donde $V = v_1, v_2, \ldots, v_n$ y $E = e_1, e_2, \ldots, e_m$. Cada arista e_i es un par no ordenado de vértices, es decir, $e_i = (v_j, v_k)$, donde v_j y v_k son vértices del grafo.

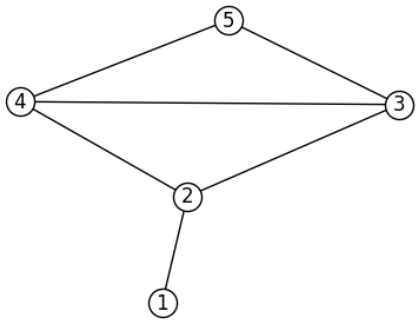


Figura 2.2: Ejemplo de un grafo

- **Vértice:** Un vértice (también conocido como nodo) es un elemento individual que representa una entidad o un punto de conexión en un

grafo. En términos formales, un vértice es un elemento del conjunto V en el par ordenado que define el grafo $G = (V, E)$. Los vértices se representan típicamente mediante símbolos, como letras, números o cualquier otro identificador único.

- **Arista:** Una arista es un par no ordenado de vértices que representa una conexión o relación entre dos vértices en un grafo. Formalmente, una arista es un elemento del conjunto E en el par ordenado que define el grafo $G = (V, E)$. Una arista se puede representar como $e_i = (v_j, v_k)$, donde v_j y v_k son vértices del grafo. En un grafo no dirigido, la arista no tiene dirección, mientras que, en un grafo dirigido, la arista tiene una dirección específica.

- **Matriz de adyacencia:** La matriz de adyacencia es una representación de un grafo mediante una matriz bidimensional. Si un grafo tiene n vértices, la matriz de adyacencia es una matriz cuadrada de tamaño $n \times n$. La entrada en la posición (i, j) de la matriz indica si hay una arista entre los vértices v_i y v_j. Si hay una arista, el valor en la posición (i, j) puede ser 1 o algún otro valor no nulo que indique la presencia de la arista. Si no hay una arista, el valor suele ser 0.

$$a_{ij} = \begin{cases} 1 & \text{, si el nodo } j \text{ está unido al } i \text{ por una arista } (i,j), \\ 0 & \text{, cualquier otra situación} \end{cases}$$

  En el grafo simple de la figura 2.2 la matriz de adyacencia es:

$$\mathsf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

- **Lista de adyacencia:** La lista de adyacencia es una representación de un grafo donde se utiliza una lista o arreglo para cada vértice del

grafo. Cada elemento de la lista o arreglo corresponde a un vértice adyacente al vértice actual. Esta representación es útil para grafos con un número reducido de aristas, ya que permite un acceso eficiente a los vecinos de cada vértice.

- **Matriz de incidencia:** La matriz de incidencia es otra forma de representar un grafo mediante una matriz bidimensional. Si un grafo tiene n vértices y m aristas, la matriz de incidencia es una matriz de tamaño $n \times m$. Cada fila de la matriz representa un vértice y cada columna representa una arista. Los elementos de la matriz indican la incidencia de un vértice en una arista específica. Para el caso de una red no-dirigida se pueden definir dos tipos de matrices de incidencia: las no orientadas y las orientadas. Las matrices de incidencia no orientadas es una matriz **B** N $\times$ L donde N es el número de vértices y L el número de uniones o aristas l = (i, j) y tal que:

$$b_{il} = \begin{cases} 1 & \text{, si la unión l contiene al vértice i,} \\ 0 & \text{, cualquier otra situación} \end{cases}$$

  La matriz de incidencia orientada de una red no dirigida es la matriz de incidencia de cualquier orientación del grafo.

Algunos de los principales conceptos y propiedades en la teoría de grafos, son:

- **Conectividad o grado**: Un grafo es conexo si existe un camino entre cualquier par de nodos. En caso contrario, se dice que es un grafo des conexo.
- **Ciclos**: Un ciclo es una secuencia de aristas que forma un circuito cerrado en el grafo. Un grafo sin ciclos se denomina grafo acíclico.
- **Caminos más cortos**: Se refiere a encontrar la ruta más corta entre dos nodos en un grafo, considerando la longitud de las aristas.

- **Árboles**: Un árbol es un grafo acíclico y conexo. En un árbol, cualquier par de nodos está conectado por un único camino.
- **Grafos bipartitos**: Un grafo es bipartito si sus nodos se pueden dividir en dos conjuntos disjuntos, de manera que todas las aristas conectan nodos de conjuntos diferentes.

Algoritmos y Técnicas clásicos para resolver problemas de grafos, son:

- **Algoritmo de Dijkstra [37]**: Se utiliza para encontrar el camino más corto entre dos nodos en un grafo ponderado.
- **Algoritmo de Kruskal [38]**: Permite encontrar un árbol de expansión mínimo en un grafo ponderado.
- **Algoritmo de Ford-Fulkerson [39]**: Se emplea para encontrar el flujo máximo en una red.

2.3.2. Distribución probabilística del grado

- **Número total de aristas**: en una red compleja se puede calcular directamente a partir de la matriz de adyacencia:

$$\sum_{i=1}^{n} i^2$$

- **Grado de un nodo**: En una red no dirigida se define el grado de un nodo i como el número total de aristas incidentes en dicho nodo y se denota por k_i.
- **Distribución de probabilidad de grados de un grafo**: El grado de un nodo es una propiedad local de la red, pero considerando la secuencia de grados podemos determinar algunas propiedades globales de la red. La organización y estructura global de una red, inducida por su secuencia de grados, está caracterizada por la distribución de probabilidad de grados de la red $P(k)$. Para un grafo no dirigido, $P(k)$ se define como la fracción de nodos de grado k y representa la probabilidad de que un nodo elegido al azar en la red tenga un grado k.

- **Camino en un grafo**: De manera formal, se define un camino en un grafo como una secuencia de nodos tal que dos nodos consecutivos en la secuencia están conectados por una arista. Un camino dirigido en una red dirigida es un camino con las aristas dirigidas desde cada nodo al siguiente en la secuencia. Cada camino tiene su longitud definida como el número de aristas que se pasan al seguirlo, incluyendo posibles repeticiones en caminos que se interceptan.

- **Distancia**: es la longitud más corta de los caminos que conectan dos nodos i y j en una red. Por ejemplo, en redes de aeropuertos, las ciudades con vuelos directos están a una distancia más corta que las ciudades que se alcanzan a través de otras intermedias, aunque puedan estar físicamente más cercanas entre sí. Muchas redes complejas se caracterizan por tener distancias cortas entre nodos, por ejemplo, cada dos personas en el mundo están separadas por unos pocos saludos de manos a través de otras personas.

- **Longitud promedio de caminos:** es la distancia más pequeña entre un nodo i y j, y la denotamos como $d(n_i, n_j) = d(i, j)$, como la longitud del camino más corto entre dichos nodos. La longitud promedio de caminos o distancia promedio es el promedio de las distancias más pequeñas entre dos nodos de la red (si existe un camino conectando un par de nodos de la red).

- **Coeficiente de agrupamiento (clustering coefficient)**: mide el nivel en que los nodos de una red se agrupan entre sí. Dado un grafo $G(N, L)$ no dirigido, con L_i es el número de aristas que existen entre los vecinos del nodo i, y k_i es el grado del nodo i, el coeficiente de agrupamiento de i, se denota C_i, y es el cociente:

$$C_i = \frac{2L_i}{k_i(k_i-1)}$$

.

2.3.3. Matrices de expresión genética

Las micro matrices de ADN nos permiten visualizar la expresión potencial de todos los genes dentro de una población celular o muestra de tejido que revela el *transcriptoma*. Este tipo de datos se denomina GEP (perfil de expresión génica) y proporciona una imagen completa del patrón de expresión génica en una muestra biológica. Desde 2004, GEP se está adoptando cada vez más como una herramienta de toma de decisiones en el entorno clínico [40], particularmente en la investigación relacionada con el cáncer [41, 42, 43, 44]. Generalmente, el GEP se mide en transcripciones por millón (TPM), $\frac{1}{1{,}000{,}000}$. La cantidad de TPM cuantifica el nivel de expresión de cada gen y es la proporción de moléculas de ARN en la muestra del gen RNA-seq. Cuanto mayor sea el valor de TPM, más expresado es el gen. Para manejar mejor la información TPM de los genes en las biopsias, se ingresa a la función logaritmo en base 2, log_2. Los valores GPE son la entrada específica de ARACNe.

Las matrices de expresión génica son una representación tabular de datos que capturan los niveles de expresión de genes en muestras biológicas. Estas matrices son ampliamente utilizadas en la investigación genómica para el análisis y la interpretación de los perfiles de expresión génica. En una matriz de expresión génica cada fila representa un gen y cada columna representa una muestra biológica. Los valores en la matriz representan los niveles de expresión del gen en cada muestra, que generalmente se obtienen mediante técnicas de secuenciación de ARN o Microarray de ADN. Estas matrices contienen información cuantitativa que refleja la abundancia relativa de ARN mensajero (ARNm) transcritos a partir de los genes en una muestra específica. Los niveles de expresión pueden variar ampliamente entre los genes y entre las muestras, según la actividad génica en diferentes condiciones biológicas o estados patológicos.

El análisis de matrices de expresión génica permite identificar patrones de expresión, diferencias entre grupos de muestras, genes co-expresados y vías biológicas relevantes. Algunas de las técnicas comunes utilizadas para analizar matrices de expresión génica incluyen:

- **Análisis de expresión diferencial:** para identificar genes cuya expresión difiere significativamente entre dos o más condiciones biológicas. Esto puede revelar genes específicos asociados con una enfermedad o fenotipo particular.

- **Clustering y análisis de agrupamiento:** para identificar grupos de genes con perfiles de expresión similares o muestras biológicas que comparten patrones de expresión. Esto puede ayudar a identificar subtipos de enfermedades o grupos funcionales de genes.

- **Análisis de enriquecimiento de genes:** para determinar si un conjunto de genes se encuentra enriquecido en una vía biológica específica o en funciones génicas particulares. Esto ayuda a comprender los procesos biológicos subyacentes asociados con los genes diferencialmente expresados.

- **Construcción de redes de co-expresíon:** para identificar relaciones funcionales entre genes en función de sus patrones de co-expresíon dentro de una red de genes.

Las matrices de expresión génica también se pueden combinar con otros tipos de datos, como datos clínicos, datos de variantes genéticas o datos de interacción proteína-proteína, para obtener una visión más completa de los mecanismos moleculares subyacentes a una condición biológica o enfermedad; son una representación tabular de los niveles de expresión de genes en muestras biológicas. Su análisis permite obtener información valiosa sobre la actividad génica, identificar genes relevantes y comprender los procesos biológicos subyacentes en diferentes condiciones biológicas o enfermedades.

2.4. Análisis estadístico

El análisis estadístico es una herramienta fundamental en el campo de la investigación y toma de decisiones. Se utiliza para extraer información y co-

nocimiento a partir de datos, permitiendo comprender patrones, relaciones y tendencias que subyacen en los fenómenos estudiados [45]

El análisis estadístico involucra diversas técnicas y métodos que ayudan a resumir y visualizar los datos, identificar la variabilidad presente en ellos, realizar inferencias y tomar decisiones basadas en evidencia. Algunas de las técnicas más comunes incluyen la descripción de datos mediante medidas de tendencia central y dispersión, la construcción de gráficos y visualizaciones para representar la distribución de los datos, y el cálculo de correlaciones y regresiones para explorar relaciones entre variables [46].

Además, el análisis estadístico también abarca el diseño de experimentos, donde se establecen condiciones controladas para recopilar datos y evaluar el impacto de diferentes variables en el resultado de interés. Esto implica la planificación de muestras representativas, la asignación aleatoria de tratamientos y la aplicación de pruebas de hipótesis para determinar si las diferencias observadas son estadísticamente significativas [47].

En el análisis estadístico, la interpretación adecuada de los resultados es esencial. Esto implica evaluar la significancia estadística, es decir, determinar si las diferencias observadas son el resultado del azar o si realmente representan diferencias o relaciones reales en la población objetivo. Además, el análisis estadístico también puede proporcionar intervalos de confianza para estimar la precisión de los resultados y la incertidumbre asociada.

El análisis estadístico es aplicable en una amplia gama de disciplinas, como la medicina, la economía, la psicología, la biología y la ingeniería, entre otras. Permite tomar decisiones informadas, respaldadas por evidencia empírica, y es una herramienta fundamental para el avance de la ciencia y la toma de decisiones basada en datos.

En resumen, el análisis estadístico es una herramienta poderosa que permite comprender y extraer información valiosa de los datos. Ayuda a revelar patrones, relaciones y tendencias, y proporciona una base sólida para la toma de decisiones informadas. Su aplicación abarca diversas áreas y disciplinas, y es esencial para el avance del conocimiento y la toma de decisiones basadas en evidencia empírica.

Método	**Descripción**	**Ventajas**	**Desventajas**	**Aplicaciones**
Machine Learning	Utiliza algoritmos para analizar datos e identificar patrones en los datos.	- Puede detectar patrones complejos en grandes conjuntos de datos.	- Requiere una gran cantidad de datos para un entrenamiento preciso. - Requiere experiencia en algoritmos y programación.	- Clasificación de tumores. - Predicción de respuesta a tratamientos.
Deep Learning	Utiliza redes neuronales profundas para aprender características y realizar predicciones.	- Puede aprender representaciones jerárquicas de los datos.	- Requiere una gran cantidad de datos etiquetados para un entrenamiento preciso. - Requiere capacidad computacional y recursos de memoria significativos.	- Segmentación de imágenes médicas. - Detección de metástasis.
Análisis de imágenes	Procesamiento de imágenes médicas para identificar patrones y anomalías.	- Permite la detección temprana y precisa de anomalías.	- Requiere una gran cantidad de imágenes etiquetadas para entrenar algoritmos de forma efectiva. - Sensible a la calidad de las imágenes y las condiciones de iluminación.	- Diagnóstico de cáncer de piel. - Detección de tumores en imágenes de resonancia magnética.
Bioinformática	Utiliza técnicas computacionales para analizar datos biológicos, como secuencias de ADN.	- Permite el análisis de grandes conjuntos de datos biológicos.	- Requiere una comprensión sólida de la biología molecular y la genómica. - Requiere herramientas y recursos computacionales especializados.	- Análisis de expresión génica. - Identificación de mutaciones somáticas.

Método	**Descripción**	**Ventajas**	**Desventajas**	**Aplicaciones**
Métodos de minería de datos	Utiliza algoritmos para descubrir patrones y conocimiento útil en grandes conjuntos de datos.	- Puede descubrir relaciones complejas entre diferentes variables.	- La calidad de los datos y la elección de los algoritmos pueden afectar la precisión de los resultados. - Requiere una comprensión sólida de la estructura y los patrones de los datos.	Descubrimiento de patrones en datos clínicos. Predicción de supervivencia de pacientes.
Generación de redes genéticas	Construcción de redes basadas en la interacción de genes para comprender la regulación y las vías biológicas.	- Permite el estudio de interacciones complejas entre genes y proteínas.	- Requiere una integración y validación cuidadosa de los datos experimentales. - Requiere conocimientos de biología de sistemas y análisis de redes.	- Identificación de biomarcadores para el diagnóstico y pronóstico del cáncer. - Descubrimiento de nuevas vías terapéuticas.

Tabla 2.2: Métodos computacionales para la identificación de cáncer, incluye la generación de redes genéticas

Aspecto	**Descripción**	**Importancia**
Estructura	Compuesta por genes, proteínas y elementos de control como regiones promotoras y represoras del ADN.	Determina la interacción y regulación de los genes en la red.
Dinámica	Sistema dinámico en el que la expresión génica puede cambiar en respuesta a estímulos internos y externos.	Permite la adaptación y respuesta de los organismos a su entorno.
Modelización	Utiliza modelos matemáticos y computacionales basados en teoría de grafos y teoría de control para comprender y predecir el comportamiento de las redes genéticas.	Ayuda a estudiar y simular el funcionamiento de los genes y proteínas en la red.
Funciones	Regulan procesos biológicos como el desarrollo embrionario, la diferenciación celular, la respuesta inmune, la cicatrización de heridas y la regulación del ciclo celular.	Cruciales para el desarrollo, funcionamiento y mantenimiento de los organismos vivos.
Enfermedades	Alteraciones en las redes genéticas se han asociado con diversas enfermedades como el cáncer, enfermedades cardiovasculares y enfermedades genéticas hereditarias.	Proporciona información sobre las causas y mecanismos subyacentes de las enfermedades.

Tabla 2.3: Información sobre las redes genéticas.

Capítulo 3

Métodos y algoritmos para el análisis genético

Se combina el uso de algoritmos basados en Teoría de la Información con técnicas estadísticas para obtener una comprensión más profunda de los genes relevantes en la formación del cáncer de mama y su metástasis.

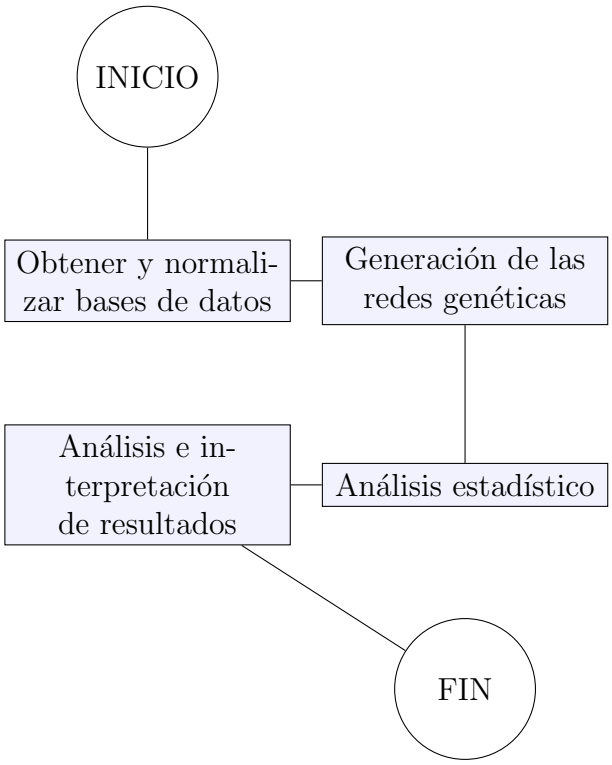


Figura 3.1: Diagrama de flujo general

El proceso general se puede ver en la Figura 3.1 y se describe en el pseudocódigo del Algoritmo 1.

Algorithm 1 Pseudocódigo de la Metodología.

```
INICIO
   Proceso: Obtener_y_normalizar_las_bases_de_datos
        leer base_de_datos_1 = archivo1
            obtener lista de genes=genes
        leer base_de_datos_2 = archivo2
            obtener lista de genes=genes2
            Si genes1=genes2
                generar diccionario
                agregar a Entrada_Red
            si no
                igualar genes1 y genes2
        extraer diccionario, Entrada_Red

    Proceso: Generar_las_Redes_Geneticas
        De Obtener_y_normalizar_las_bases_de_datos obtener Entrada_Red
        Obtener ARACNe
        Entrada_Red= entrada
        introducir entrada a ARACNe
            si salida ARACNe >>> 100 MB
                hacer una poda
            si no
                hacer analisis_de_salida_ARACNe

            extraer analisis_de_salida_ARACNe

    Proceso: Generar_analisis_estadistico
        De Generar_las_Redes_Geneticas obtener analisis_de_salida_ARACNe
        Simulacion = Repetir n veces Generar_las_Redes_Geneticas obtener
        Promediar Simulacion
        Generar Analisis de Simulacion
        extraer Analisis_estadistico

    Proceso: Valdidar_resultados
        De Generar_analisis_estadistico obtener Analisis_estadistico
        Comparar= Comparar Analisis_estadistico con Estado del arte
        Reporte= Reportar informacion relevande de Comparar
        extraer Reporte

FIN
```

Los pasos que conlleva el proceso se describen detalladamente a continuación:

1. Recopilación y preparación de datos:

    - Recolectar datos genómicos y moleculares relacionados con el cáncer de seno de diferentes bases de datos públicas y fuentes relevantes.
    - Realizar un proceso de limpieza de datos para eliminar valores faltantes, *outliers* y datos irrelevantes en cada conjunto de datos.
    - Identificar las variables y características comunes en las diferentes bases de datos para establecer la correspondencia entre ellas.

2. Creación del diccionario de variables y genes:

    - De las variables que representan los genes, analizar las vías de señalización, funciones biológicas, entre otros aspectos relevantes para el estudio.
    - A las variables de las bases de datos, asignar un nombre común a las que representen la misma información biológica.

3. Normalización y combinación de datos:

    - Realizar una normalización de cada conjunto de datos, asegurando que las variables tengan la misma escala y unidad de medida.
    - Utilizando el diccionario, combinar los conjuntos de datos en una matriz unificada con la información consistente de todas las bases de datos.

4. Construcción de redes genéticas:

    - Utilizar los datos unificados y normalizados para construir redes genéticas que representen las interacciones entre los genes asociados al cáncer de seno.
    - Aplicar el algoritmo de ARACNe para estimar la IM entre pares de genes y construir la estructura de la red.

5. Análisis topológico de las redes:

- Calcular medidas de centralidad, el grado, la distribución del grado y el coeficiente de agrupamiento, para identificar los nodos - genes más importantes en la red.
- Realizar un análisis de modularidad para identificar grupos de genes altamente conectados en la red, lo que puede indicar vías biológicas o funciones específicas.

6. Identificación de genes principales y funciones biológicas:

   - Utilizar las medidas de centralidad para identificar los genes principales en la red, con mayor grado e IM.
   - Identificar los genes principales a sus funciones biológicas.

7. Análisis estadístico de las asociaciones gen-función:

   - Realizar decenas de simulación, y de los resultados, identificar su dispersión y el promedio.
   - Aplicar las medidas estadísticas usuales para determinar la significación de las asociaciones entre los genes principales y las funciones biológicas.

8. Validación funcional de los genes principales:

   - Comparar resultados con experimentos y reportes científicos previos para validar la función biológica de los genes principales identificados.
   - Analizar si los resultados respaldan la función propuesta de los genes principales conforme el análisis genético.

9. Interpretación de resultados y conclusiones:

   - Interpretar los resultados considerando el análisis topológico de las redes y las asociaciones gen-función identificadas.

3.1. Datos de biopsias de tejido libre de cáncer, tumor primario y primera metástasis

Se preparan diferentes números de muestras para aplicar ARACNe, utilizando el conjunto de datos *GEO GSE32490* que contiene perfiles de expresión de metástasis de cáncer de mama en los ganglios linfáticos y tejidos de autopsia fijados en formalina e incrustados en parafina (FFPE), y el proyecto *TCGA-BRCA* del GDCDP que contiene datos de expresión génica, perfiles de metilación del ADN, datos de mutaciones, notas clínicas y datos de imágenes radiológicas, entre otros. En total se obtuvieron un total de:

- 97 muestras de tejido libre de cáncer (libre de cáncer).
- 130 muestras de tumor primario de cáncer de mama y,
- 90 muestras de las primeras metástasis de cáncer de mama en los tejidos de los ganglios linfáticos.

Los archivos GEO tienen formato diferentes formatos con información sobre el título, el número de serie, el tipo de muestra de tejido, el método normalizado y las transcripciones de cada gen por millón (TPM) de las biopsias. Los archivos GDC comprenden la cuantificación de ARNm y la secuencia de análisis para medir la expresión del nivel de genes a partir de datos sin procesar. Posteriormente, los recuentos se aumentan con varias transformaciones que incluyen arreglos por kilo base de transcripción por millón de lecturas mapeadas (FPKM), FPKM normalizado de cuartil superior (FPKM-UQ). Estos valores se anotan con el biotipo del gen en un archivo. tsv. El símbolo del gen abrevia el nombre científico del gen, y ambos se utilizan en el diccionario estándar de genes, creado para no depender de la nomenclatura GEO o GDC.

Los resultados son archivos *.sif* y *. sort* con el MI de pares de genes. Constituyen la matriz de expresión del conjunto de genes. En el archivo *. sort* los datos están en orden descendente MI; en el archivo *.sif* los primeros elementos son los más conectados con los otros genes y así sucesivamente en orden descendente del grado de los nodos. A partir de esta información

ordenada se generan las redes genéticas (RG). El tamaño del archivo de salida depende de los datos de entrada y es diferente para cada tipo de tejido. Ver Tabla 3.1.

Información de archivos de entrada y de salida					
	Entrada			Salida	
Tipo de tejido	Número de muestras	Número de genes	Tamaño de archivo. tsv	Tamaño de archivo .sif y. sort	Número de pares de genes
Tejido sano	97	14,996	21.6 MB	2.4 GB	110,319,334
Tumor primario	130	14,996	28.5 MB	2.5 GB	109,517,928
Primera metástasis	90	14,996	16.2 MB	2.5 GB	112,417,355

Tabla 3.1: Información de datos de entrada y salida de ARACNe.

Luego de identificar, descargar y almacenar localmente la base de datos, el pseudocódigo del Algoritmo 2 describe las etapas del algoritmo para generar el diccionario y normalizar las bases de datos, así mismo se puede ver el diagrama de flujo de este proceso en la figura 3.2

El logaritmo no se utiliza específicamente para normalizar datos, sino más bien para realizar una transformación en los datos con el fin de obtener ciertos beneficios o propiedades deseadas. En muchos casos, los datos pueden tener una distribución sesgada o muy dispersa, lo que dificulta su análisis o interpretación. La aplicación de una función logarítmica a los datos puede ayudar a mitigar estos problemas.

Aquí hay algunas razones comunes por las cuales se utiliza el logaritmo para transformar los datos:

- Reducción del sesgo: Si los datos tienen una distribución sesgada, es decir, están desplazados hacia un extremo, aplicar el logaritmo puede reducir ese sesgo. Al tomar el logaritmo de los valores, se comprimen los valores más altos y se expanden los valores más bajos, lo que puede ayudar a obtener una distribución más simétrica.

- Estabilización de la varianza: En algunos casos, los datos pueden tener una varianza desigual a medida que aumenta su magnitud. Al aplicar el logaritmo, se puede reducir la escala de los valores más altos, lo que

Algorithm 2 Generar diccionario y normalizar bases de datos.

Descargar y procesar los datos de la base de datos

1: Exploración inicial de los datos para comprender su estructura y características.
2: Colocar los datos en una estructura de programación adecuada (matrices).
3: Identificar y eliminar los atributos o registros necesarios con información irrelevante.
4: Verificación final tal que la base de datos resultante contenga la información relevante.
5: Guardar la base de datos depurada en un archivo local o en una base de datos adecuada.
6: Fin del proceso de descarga y depuración de la base de datos.

Generar diccionario

7: Identificar y comparar los genes de cada base de datos.
8: Extraer solamente los genes que se repitan en todas las bases de datos.
9: Verificar que no se repitan los datos.
10: Guardar el diccionario en un archivo local o en una base de datos para que después sea comparada con las demás bases de datos.

Normalizar base de datos

11: Comparar cada base de datos con el diccionario creado.
12: Identificar los parámetros en los que están expresados los Transcritos por millón (TMP) en los perfiles de expresión genética.
13: Aplicar el $\log_2$ para normalizar los datos.

a su vez puede estabilizar la varianza y mejorar la interpretación y el análisis estadístico.

- Linealización de relaciones no lineales: En ciertas situaciones, los datos pueden mostrar una relación no lineal entre las variables. Al aplicar el logaritmo a una o ambas variables, es posible transformar la relación no lineal en una forma más lineal. Esto puede facilitar el modelado y el análisis posterior.

- Reducción de la amplitud de los valores: Si los datos tienen una amplitud muy grande, con valores extremadamente altos o bajos, aplicar el logaritmo puede reducir esa amplitud. Esto puede ser útil para visualizar los datos en una escala más manejable o para mejorar la interpretación de los resultados.

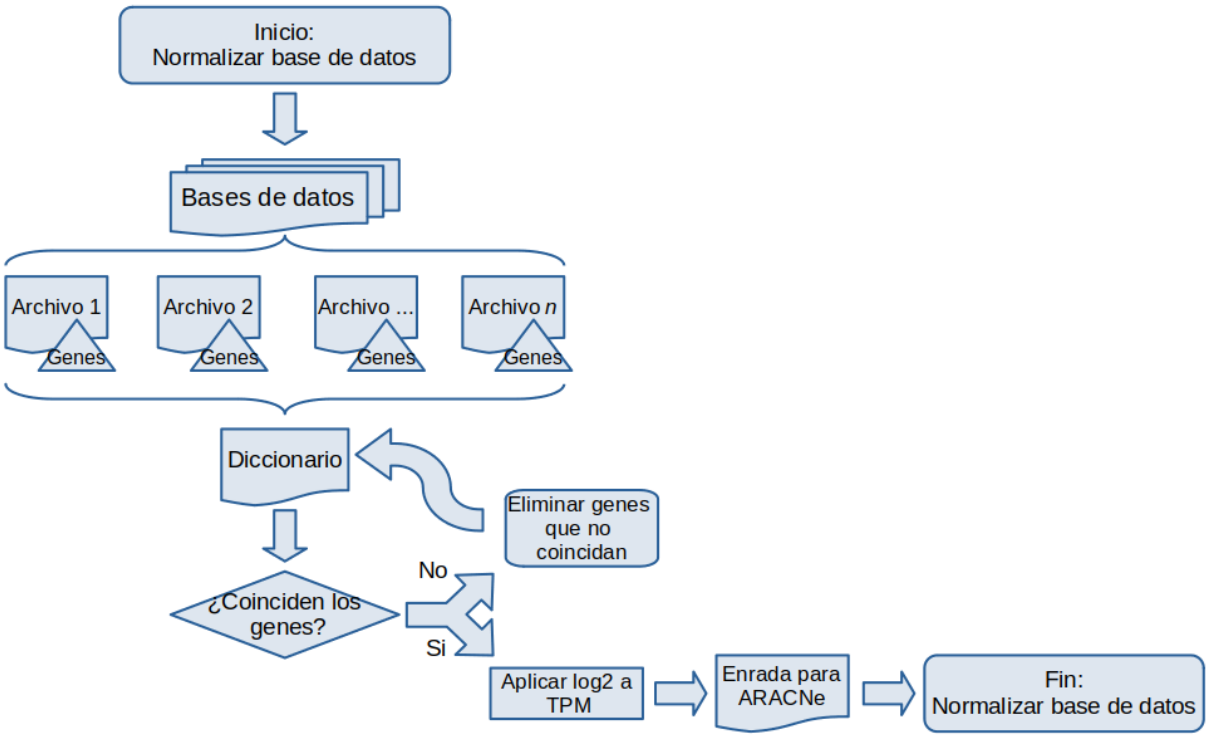


Figura 3.2: Diagrama de flujo de la obtención y normalización de la base de datos.

3.2. Generación y análisis de redes genéticas

Los pasos que se siguieron para generar las redes genéticas son los siguientes:

- Introduce datos a ARACNe para obtener toda la matriz de correlación.
- Podar la matriz anterior con respecto al umbral deseado.
- Construcción y análisis de las redes genéticas
- Repetición de las simulaciones para generar el análisis estadístico.

El código está en el apéndice y el pseudocódigo en el Algoritmo 3

Para el análisis estructural de las redes genéticas se utilizó la herramienta **NetworkX** Se toma cada matriz de correlación recortada, se crea la red y se sacan las siguientes centralidades:

- Grado de cada gen.

Algorithm 3 Flujo de trabajo para obtener matrices de correlación con ARACNe

for i desde 1 hasta n simulaciones **do**
    $ftsv \leftarrow$ nombre del archivo tsv
    Iniciando Simulacion
    $nom \leftarrow$ nombre base del archivo sin extensión
    Extraer la primera columna de $ftsv$ y guardarla en *node.list*
    cname $\leftarrow$ nombre de índice de columna
    Ejecutar ARACNe paralelo con $ftsv$, *node.list*, *cname*, y número de procesadores disponibles
    Obtener parámetros para unir las matrices de la carpeta *adj*
    Unir matrices ADJ usando *nom*, *n*, *node.list*, *cname*
    Mover la matriz de adyacencia a la carpeta actual
    Convertir la matriz de adyacencia en formato TSV
    Renombrar el archivo resultante a $nom - complete\_i.tsv$
    Convertir la matriz de adyacencia en formato SIF
    Guardar el archivo como $nom\_i.sif$
    Ordenar el archivo $nom\_i.sif$ en orden descendente según la tercera columna
    Guardar el archivo ordenado como $nom\_i.sort$
    Realizar la poda en el archivo $nom\_i.sort$ seleccionando los pares de genes con valor mayor o igual a 0.35
    Guardar los pares de genes podados en el archivo Recorte_de_$nom_i.txt$
    Eliminar archivos temporales y de limpieza
    Fin de prueba
end for

- Grado promedio de la red.
- Distribución de grado de la red.
- Información mutua entre genes.
- Coeficiente de agrupamiento.

El código está en el apéndice y el pseudocódigo se puede ver en el Algoritmo 4

Se realiza un flujo de trabajo para obtener 100 simulaciones con el propósito de hacer un análisis exploratorio para identificar la distribución, la media, la desviación estándar y otros aspectos relevantes de los datos y las redes genéticas que se obtuvieron, en la figura 3.3 se describe el proceso para generar las redes genéticas y desarrollar el análisis estadístico.

Algorithm 4 Análisis de redes con NetworkX

```
for cada archivo en el directorio actual do
    counter ← 1
    limit ← 10
    while counter < limit do
        if el archivo termina con '.txt' then
            Imprimir el nombre del archivo
            Abrir el archivo y leer su contenido
            Cerrar el archivo
            Crear un diccionario vacío llamado dgenes para almacenar los grados de los nodos
            for cada línea en el contenido del archivo do
                Dividir la línea en dos nodos y su peso
                if los nodos no están en dgenes then
                    Agregar los nodos a dgenes con un grado inicial de 1
                else
                    Incrementar en 1 el grado de los nodos en dgenes
                end if
            end for
            Calcular el promedio de los grados de los nodos (k) como la suma de los grados dividido por la cantidad de nodos
            Calcular la centralidad promedio (C) como el promedio de la agrupación local de los nodos
            Imprimir "Grado promedio de los nodos: "seguido de k
            Imprimir Çentralidad promedio de los nodos: "seguido de C
            Crear una lista vacía llamada seqgenes
            for cada grado en los valores de dgenes do
                Agregar el grado a seqgenes
            end for
            Ordenar seqgenes en orden ascendente
            Imprimir el histograma con los valores de seqgenes y su frecuencia
            Imprimir los límites y etiquetas del eje x e y del histograma
            Incrementar counter en 1
        end if
        Incrementar counter en 1
    end while
end for
```

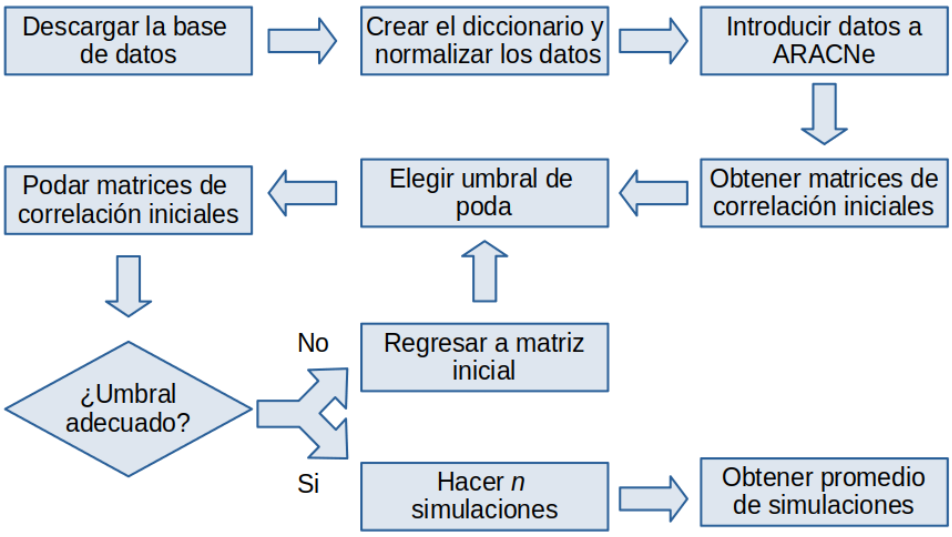


Figura 3.3: Diagrama de flujo para general las Redes Genéticas.

3.3. Análisis molecular y funcionalidad de los genes

Gene Ontology (GO) es una herramienta valiosa para la anotación y el análisis de genes y proteínas, con una ontología estandarizada para describir las funciones biológicas, procesos y ubicaciones celulares de los productos génicos. GO es una iniciativa internacional colaborativa cuyo objetivo es proporcionar una ontología estandarizada y bien definida para la anotación de genes y proteínas en todos los organismos. Con GO se analiza la función molecular del grupo de genes y su relación en la formación del cáncer.

GO y su base de datos de anotaciones se utilizan ampliamente en la investigación en biología molecular y celular, y han sido utilizadas en estudios genómicos, transcriptómicas y proteómicos de una amplia variedad de organismos. Además, la ontología GO es una herramienta importante para la interpretación de los resultados de experimentos de alta capacidad, como la secuenciación masiva de ARN y ADN. La base de datos de anotaciones de genes y proteínas de GO se construye recopilando información relevante de la literatura científica y fusionando datos experimentales obtenidos a partir de técnicas avanzadas de análisis de alto rendimiento. Este proceso riguroso y sistemático de recopilación y fusión de datos se conoce como cu-

ración, y es esencial para asegurar la calidad y precisión de las anotaciones de GO. La curación de la literatura científica y la integración de datos de experimentos de alto rendimiento son procesos continuos y en constante evolución para garantizar que la base de datos de GO refleje los avances más recientes en la investigación científica. Los términos de GO se utilizan para anotar genes y proteínas en bases de datos públicas, lo que permite la integración de datos y el análisis de conjuntos de genes.

La descripción de la funcionalidad genética utilizando Gene Ontology (GO) se basa en un sistema de vocabulario controlado que proporciona términos estandarizados y jerarquizados para describir las funciones de los genes y sus productos en tres ramas principales: función molecular, proceso biológico y componente celular. Cada rama se subdivide en términos específicos que describen funciones, procesos y ubicaciones celulares:

- **Procesos biológicos:** Esta categoría describe las funciones que los genes desempeñan en los procesos biológicos a nivel molecular y celular. Incluye términos como metabolismo, señalización celular y desarrollo embrionario. Los términos están jerarquizados, lo que significa que términos más específicos están subordinados a términos más generales.

- **Componentes celulares:** Esta categoría describe las partes subcelulares o estructuras celulares en las que los genes o proteínas están localizados y ejercen su función. Incluye términos como núcleo, mitocondria y membrana plasmática. Los términos también están organizados jerárquicamente.

- **Función molecular:** Esta categoría describe las actividades moleculares de los genes o proteínas, como actividad enzimática, unión a proteínas o transporte de iones. Los términos también están organizados jerárquicamente, con una descripción de las funciones moleculares.

Así, al utilizar GO se asignan anotaciones sobre la función biológica potencial de los genes y proteínas en el contexto de un estudio genómico,

lo cual facilita la comparación y el análisis de conjuntos de genes o proteínas. Mediante términos estándares y un marco de referencia común pueden identificarse patrones funcionales, asociaciones entre genes y procesos biológicos, y realizar análisis de enriquecimiento funcional para determinar si hay sobre o subrepresentación en un conjunto de genes de interés. El diagrama de flujo de la Figura 3.4 ilustra el proceso en el que se aplica GO. Este diagrama de flujo proporciona una estructura visual para comprender el proceso general de aplicación de Gene Ontology en el análisis de genes y la interpretación de los resultados obtenidos.

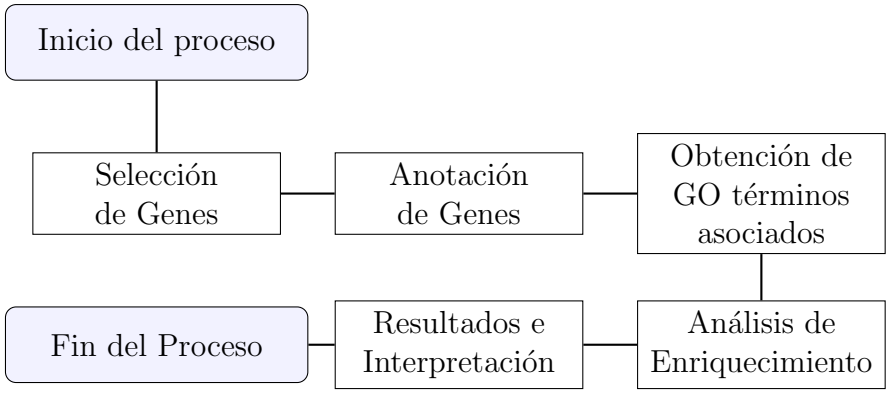


Figura 3.4: Diagrama de flujo del proceso de aplicación de Gene Ontology

A continuación, se describen brevemente las etapas del diagrama de flujo:

1. **Selección de Genes:** En esta etapa, se seleccionan los genes de interés basándose en algún criterio, como los resultados de un estudio de expresión génica diferencial o análisis genómicos.

2. **Anotación de Genes:** Se realiza la anotación de los genes seleccionados utilizando bases de datos y herramientas bioinformáticas. La anotación implica asignar información relevante a cada gen, como su identificador, función conocida, ubicación, entre otros.

3. **Obtención de GO términos asociados:** En esta etapa, se busca obtener los términos de Gene Ontology (ontología génica) asociados a los genes seleccionados. Esto se puede hacer utilizando bases de datos o herramientas específicas que proporcionan información de GO.

4. **Análisis de Enriquecimiento:** Se realiza un análisis estadístico para determinar si ciertos términos de GO están enriquecidos en los genes seleccionados en comparación con un conjunto de genes de fondo. Esto permite identificar los términos de GO que están significativamente asociados con los genes de interés.

5. **Resultados e interpretación:** El análisis de enriquecimiento se interpretan para comprensión de la función biológica y los procesos asociados a los genes estudiados.

Capítulo 4

Análisis de las redes genéticas

La salida de ARACNe, un algoritmo utilizado para inferir redes de regulación génica a partir de datos de expresión génica incluye dos tipos de archivos: *.sif* y *. sort.*

El archivo *.sif* (Simple Interaction File) es un formato de archivo utilizado para representar las interacciones entre nodos en una red. En el contexto de ARACNe, el archivo .sif contiene las relaciones de regulación génica predichas por el algoritmo. Cada línea en el archivo .sif representa una interacción entre dos nodos, donde cada nodo representa un gen y la interacción indica una relación de regulación (por ejemplo, activación o represión) entre los genes. El archivo .sif proporciona una representación simplificada de la estructura de la red, lo que facilita su visualización y análisis.

Por otro lado, el archivo *. sort* contiene información adicional sobre los genes en la red. Este archivo proporciona una lista de genes ordenados según su importancia en la red inferida por ARACNe. El orden se basa en medidas de relevancia o centralidad de los genes en la red, como su grado, entre otros. El archivo. sort puede ayudar a identificar los genes más influyentes o centrales en la red, lo que es útil para priorizar genes candidatos y comprender mejor la estructura y función de la red de regulación génica.

La desigualdad del procesamiento de datos (DPI) establece que si un par de genes (individuales) g_1 y g_3 interactúan solo a través de otro gen g_2, ocurre que

$$I(g1, g3) \leq min\{I(g1, g2), I(g2, g3)\}[31]. \quad (4.1)$$

Por lo tanto, los valores MI más pequeños y menos importantes provienen de interacciones genéticas indirectas. Con esta premisa y para evitar operaciones innecesarias, se aplicaron varios umbrales de poda sobre el archivo de salida con el fin de obtener un archivo delgado con la información relevante para generar la RG. Para el análisis estadístico en la siguiente sección, para generar el RG se aplica el mismo umbral cada vez en el archivo de salida ARACNe. De esta forma, el conjunto de RG de cada simulación, obtenido después de podar cada matriz de expresión completa diferente, merece un conjunto de muestras representativo. Los tamaños de los archivos después de podar una matriz de expresión completa con diferentes umbrales se muestran en la Tabla 4.1.

Con base en estos resultados, se decide utilizar un umbral de 0.35 ya que como se ve en la Tabla 4.2, si se tratan de emparejar el tamaño de los archivos y, por lo tanto, en número de pares de genes, los umbrales varían mucho, entonces, una vez que se elige el umbral de poda, se lleva a cabo un análisis estructural de RG para identificar los genes con el grado más alto junto con los pares de genes con el MI más alto.

4.1. Genes con mayor información mutua (IM)

La Tabla 4.3 que presenta los 30 pares de genes con más información mutua en las redes de tejido libre de cáncer proporciona información valiosa sobre las interacciones genéticas significativas en el contexto de un tejido sin cáncer. La información mutua es una medida que captura la dependencia y la relación entre dos genes en una red biológica.

La Tabla 4.4 de los 30 pares de genes con más información mutua en las redes del tumor primario de cáncer de mama muestra las interacciones genéticas más relevantes en el sitio del tumor primario. Estas interacciones pueden implicar vías biológicas específicas que están asociadas con la

Umbral	Tejido	Tamaño del archivo	Pares de genes
0	Tejido Sano	2.4 GB	110,319,334
	Tumor primario	2.5 GB	109,517,928
	Primera metástasis	2.5 GB	112,417,355
0.1	Tejido Sano	821.6 MB	38,156,368
	Tumor primario	325.9 MB	15,158,367
	Primera metástasis	726.2 MB	33,744,090
0.2	Tejido Sano	348.1 MB	16,149,425
	Tumor primario	58.7 MB	2,739,837
	Primera metástasis	158.5 MB	7,366,824
0.3	Tejido Sano	140.7 MB	6,523,887
	Tumor primario	9.9 MB	461,157
	Primera metástasis	23 MB	1,070,951
0.4	Tejido Sano	74.0 MB	2,055,279
	Tumor primario	44.3 MB	112,717
	Primera metástasis	937.0 kB	43,630
0.5	Tejido Sano	7.3 MB	338,606
	Tumor primario	698.5 kB	32,492
	Primera metástasis	6.4 kB	301

Tabla 4.1: Tamaño de archivos después de la poda de RG con diferentes umbrales.

progresión y el desarrollo del cáncer de mama.

Por otro lado, la Tabla 4.5 de los 30 pares de genes con más información mutua en las redes de la primera metástasis del cáncer de mama revela las interacciones genéticas clave en los tejidos donde se ha producido la diseminación inicial del cáncer. Estos genes pueden ser cruciales en la invasión y migración celular y en la formación de metástasis en otros tejidos distantes del sitio primario.

Al comparar estas tres tablas, se pueden identificar los pares de genes que muestran una alta información mutua tanto en el tejido libre de cáncer como en el tumor primario y la primera metástasis. Estos pares de genes pueden tener un papel importante en la regulación de las vías biológicas relevantes para el cáncer de mama y podrían servir como objetivos te-

Umbral	Tejido	Tamaño del archivo	Pares de genes
0.44	Tejido Sano	24.1 MB	1,117,645
0.25	Tumor primario	23.8 MB	1,111,467
0.3	Primera metástasis	23 MB	1,070,951
	Tejido Sano	137.37 MB	2,000,645
0.35	Tumor primario	4.45 MB	213,103
	Primera metástasis	5.76 MB	274,067

Tabla 4.2: Tamaño de archivos después de la poda de RG con diferentes umbrales.

rapéuticos potenciales o biomarcadores para un diagnóstico más preciso. En resumen, estas tablas proporcionan una visión detallada de los pares de genes con mayor información mutua en diferentes etapas del cáncer de mama, desde el tejido libre de cáncer hasta el tumor primario y la primera metástasis. Estas interacciones genéticas destacadas pueden ayudar a comprender mejor los mecanismos subyacentes del cáncer de mama y brindar información valiosa para la identificación de objetivos a analizar más profundamente.

4.2. Genes con mayor grado

La Tabla 4.6, que muestra los 30 genes con mayor grado en las redes de tejido libre de cáncer proporciona información sobre los genes que tienen una mayor conectividad en el tejido no afectado por el cáncer. Estos genes con alto grado pueden desempeñar un papel importante en la función normal del tejido y pueden servir como marcadores de referencia para comparar con los tejidos cancerosos.

La Tabla 4.7, de los 30 genes con mayor grado en las redes del tumor primario de cáncer de mama destaca los genes que tienen una mayor conectividad específicamente en el sitio del tumor primario. Estos genes pueden estar involucrados en la progresión del cáncer de mama, el crecimiento del tumor y la regulación de las vías biológicas asociadas con el cáncer de mama.

Tag	Gen 1	Gen 2	IM
1	RPL31	RPL37	0.691509
2	RPL5	RPS6	0.69091
3	RPL13A	RPL7A	0.68961
4	IFNA13	LECT2	0.68835
5	CIDEC	GPD1	0.687827
6	AQP7	GPD1	0.685024
7	GPD1	LIPE	0.684901
8	RPL32	RPL37	0.683978
9	RPL13A	RPL18	0.683371
10	GYG2	KCNIP2	0.683165
11	RPL31	RPL32	0.682688
12	AQP7	KCNIP2	0.682154
13	CIDEC	KCNIP2	0.680789
14	RPL32	RPL35A	0.680065
15	C1QB	C1QC	0.679639
16	RPL5	RPL6	0.679447
17	RPL34	RPS27A	0.679395
18	RPL3	RPS3	0.679336
19	IFNA13	ZNF593	0.678812
20	IFNA13	VN1R4	0.678812
21	IFNA13	VCY1B	0.678812
22	IFNA13	VCY	0.678812
23	IFNA13	UTP14C	0.678812
24	IFNA13	UGT1A5	0.678812
25	IFNA13	UGT1A3	0.678812
26	IFNA13	TSSK2	0.678812
27	IFNA13	TSPY2	0.678812
28	IFNA13	TSPY1	0.678812
29	IFNA13	TRIM6TRIM34	0.678812
30	IFNA13	TREX1	0.678812

Tabla 4.3: 30 genes con mayor IM en redes de tejido libre de cáncer

La Tabla 4.8, que muestra los 30 genes con mayor grado en las redes de la primera metástasis de cancer de mama en nodos linfáticos revela los genes que tienen una mayor conectividad en los tejidos donde se ha producido la diseminación inicial del cáncer. Estos genes pueden desempeñar un papel crucial en la invasión y migración celular, así como en la formación de metástasis en otros tejidos distantes del sitio primario.

Al comparar estas tres tablas, se pueden identificar genes que muestran un alto grado de conectividad tanto en el tejido libre de cáncer como en el tumor primario y la primera metástasis. Estos genes podrían considerarse candidatos clave para el estudio de la progresión del cáncer, la identificación de biomarcadores y el desarrollo de terapias dirigidas. En resumen, estas tablas proporcionan una visión detallada de los genes con mayor grado en diferentes etapas del cáncer de mama, desde el tejido libre de cáncer hasta el tumor primario y la primera metástasis. Estos genes con alto grado son objetivos importantes para la investigación y el análisis que se realiza en

Tag	Gen 1	Gen 2	IM	Tag	Gen 1	Gen 2	IM
1	OR10P1	PCDHA1	0.771405	16	OR10P1	TSPY2	0.741574
2	KRTAP63	OR10P1	0.771405	17	OR10P1	TSPY1	0.741574
3	KRTAP202	OR10P1	0.771405	18	OR10P1	TRIM6TRIM34	0.741574
4	KLK9	OR10P1	0.771405	19	OR10P1	TREX1	0.741574
5	GLT6D1	OR10P1	0.771405	20	OR10P1	TNFRSF6B	0.741574
6	OR10P1	ZNRF4	0.741574	21	OR10P1	TMSB4Y	0.741574
7	OR10P1	ZNF593	0.741574	22	OR10P1	TMEM207	0.741574
8	OR10P1	ZCCHC13	0.741574	23	OR10P1	TGM3	0.741574
9	OR10P1	WFDC9	0.741574	24	OR10P1	TEX13A	0.741574
10	OR10P1	VN1R4	0.741574	25	OR10P1	TBL1Y	0.741574
11	OR10P1	VCY1B	0.741574	26	OR10P1	TAS2R16	0.741574
12	OR10P1	VCY	0.741574	27	OR10P1	TAAR9	0.741574
13	OR10P1	UTP14C	0.741574	28	OR10P1	TAAR8	0.741574
14	OR10P1	UNCX	0.741574	29	OR10P1	TAAR2	0.741574
15	OR10P1	TSSK2	0.741574	30	OR10P1	SSX4B	0.741574

Tabla 4.4: Primeros 30 pares de genes con el IM más alto de la red de tumor primario.

el siguiente capítulo.

4.3. Distribución de grado y centralidades

El grado de nodo promedio y el coeficiente de agrupamiento son medidas importantes para analizar las redes genéticas en el contexto de tejidos relacionados con el cáncer. Estas medidas proporcionan información clave sobre la estructura y las propiedades de la red, lo que nos ayuda a comprender mejor la organización y la dinámica de los genes involucrados. Estos valores están expresados en la Tabla 4.9.

El grado de nodo promedio se refiere al promedio de conexiones que tiene cada gen en la red. Representa el nivel de conectividad de los genes y nos indica cuántos otros genes están directamente vinculados a cada gen en la red. Un grado de nodo alto puede sugerir que un gen está involucrado en múltiples interacciones y puede desempeñar un papel central en la red. Por otro lado, un grado de nodo bajo puede indicar que un gen tiene menos conexiones y puede estar más aislado en la red. El grado de nodo pro-

Tag	Gen 1	Gen 2	IM
1	FGA	SALL3	0.687776
2	FGA	P2RX7	0.669893
3	SALL3	P2RX7	0.667756
4	FGA	LHFPL3	0.6657
5	LHFPL3	SALL3	0.660608
6	EFNB2	PLK1	0.646231
7	LHFPL3	P2RX7	0.641282
8	LHFPL3	EFNB2	0.641274
9	FGA	EFNB2	0.628311
10	LHFPL3	PLK1	0.626748
11	EFNB2	SALL3	0.626642
12	EFNB2	P2RX7	0.615593
13	FGA	CDC20B	0.61367
14	LHFPL3	CDC20B	0.603493
15	SALL3	CDC20B	0.601245

Tag	Gen 1	Gen 2	IM
16	FGA	PLK1	0.598333
17	P2RX7	CDC20B	0.597947
18	CARM1	SALL3	0.594405
19	FGA	CARM1	0.591954
20	SALL3	PLK1	0.59087
21	PLK1	P2RX7	0.584459
22	LHFPL3	CARM1	0.581333
23	LHFPL3	PRSS35	0.580244
24	SALL3	PRSS35	0.579253
25	FGA	PRSS35	0.57771
26	ROCK2	DAPP1	0.575295
27	RHOC	RXRB	0.573585
28	EFNB2	CDC20B	0.571676
29	CARM1	P2RX7	0.570414
30	LASP1	NEDD8	0.570045

Tabla 4.5: Primeros 30 pares de genes con el IM más alto de la red de primera metástasis.

medio nos ayuda a identificar los genes más conectados y potencialmente importantes en la red genética del tejido estudiado.

El coeficiente de agrupamiento, por otro lado, es una medida que indica cuánto se agrupan los genes en la red. Representa la tendencia de los genes a formar grupos o subredes altamente interconectadas. Un coeficiente de agrupamiento alto sugiere que los genes tienden a interactuar entre sí y formar módulos o clústeres en la red, lo que indica una organización más densa y funcionalmente coherente. Por el contrario, un coeficiente de agrupamiento bajo indica una red más dispersa, con menos agrupamiento de genes. El coeficiente de agrupamiento nos proporciona información sobre la modularidad de la red genética y nos ayuda a identificar subredes funcionales y posibles vías de señalización específicas.

El análisis del grado de nodo promedio y el coeficiente de agrupamiento en las redes genéticas de tejidos relacionados con el cáncer nos permite comprender la organización y las propiedades emergentes de estas redes. Estas medidas pueden revelar características distintivas de las redes en diferentes tejidos, como la presencia de genes altamente conectados, módulos o

Pos	Symbol	Grado	Promedio
1	CTCF	371400	3714
2	GOSR1	365200	3652
3	RTF1	364300	3643
4	GPBP1L1	361800	3618
5	API5	360700	3607
6	PLRG1	360400	3604
7	EIF4ENIF1	360200	3602
8	PRPF38A	360100	3601
9	GPBP1	359600	3596
10	CTR9	359200	3592
11	ADNP	357800	3578
12	RBBP4	357200	3572
12	PHF14	357200	3572
14	SS18	356200	3562
15	TFCP2	355800	3558
16	SIN3A	355600	3556
17	YY1AP1	354200	3542
18	TOP2B	354100	3541
19	NBR1	352900	3529
20	WIPF2	352800	3528
21	HNRNPL	352700	3527
22	PAPOLA	351900	3519
23	DHX9	351500	3515
24	USP10	351400	3514
25	PARG	351000	3510
26	PHF12	350900	3509
27	FAM120B	350100	3501
28	USP7	349800	3498
29	CNOT10	349700	3497
30	VPS4B	349500	3495

Tabla 4.6: 30 genes con mayor grado en redes de tejido libre de cáncer

clústeres funcionales, o cambios en la conectividad y agrupamiento asociados con la progresión del cáncer. Estas medidas son fundamentales para el estudio de las redes genéticas y nos proporcionan una visión más completa de la estructura y la función de los genes en el contexto del cáncer.

La ilustración de la red de genes y la figura que muestra la distribución de grados para las redes genéticas de tejido libre de cáncer (ver Figura 4.2), tumor primario de cáncer de mama (ver Figura 4.3) y primera metástasis de cáncer de mama (ver Figura 4.4) tiene varios propósitos importantes en la investigación del cáncer.

En primer lugar, la representación visual de la red de genes permite visualizar las interacciones y conexiones entre diferentes genes en un tejido específico. Estas redes genéticas capturan las relaciones funcionales y regulatorias entre los genes, lo que ayuda a comprender cómo trabajan juntos en el contexto de la biología celular y el desarrollo del tejido. Al ilustrar estas redes, se pueden identificar patrones, subredes o módulos de genes que desempeñan un papel clave en procesos específicos, como la proliferación celular, la apoptosis o la respuesta inmune.

Pos	Symbol	Grado	Promedio
1	ADIG	137900	1379
2	AVP	131500	1315
3	ACTL7B	128300	1283
4	NRN1L	128100	1281
5	FABP1	124900	1249
6	KRTAP102	123000	1230
7	NPPB	121600	1216
8	IFNA1	120600	1206
9	BARHL1	119600	1196
10	CSN2	119400	1194
11	DPRX	119200	1192
12	FGF3	119000	1190
13	FABP2	118000	1180
14	GSX2	117200	1172
15	OR10P1	116000	1160
15	WNT8A	116000	1160
17	CARD18	115700	1157
18	CST8	114900	1149
19	KLK15	114500	1145
20	SLCO1B3	114200	1142
21	SSTR4	113100	1131
22	GAL3ST3	111700	1117
23	TBC1D3	110400	1104
24	GALR3	110100	1101
24	ALPP	110100	1101
26	MRGPRX4	108500	1085
27	TRIM43	107000	1070
28	NPY	106100	1061
29	FAM181A	105000	1050
30	ALLC	104400	1044

Tabla 4.7: 30 genes con mayor grado en redes de tumor primario

En segundo lugar, la figura que muestra la distribución de grados es una herramienta importante para analizar la topología de la red genética. El grado de un nodo en la red se refiere al número de conexiones o enlaces que tiene con otros nodos. La distribución de grados revela la frecuencia con la que se encuentran diferentes niveles de conectividad en la red. Esto puede proporcionar información sobre la centralidad de ciertos genes en la red, es decir, su importancia en la comunicación y coordinación de la función génica en el tejido.

Al comparar las redes y las distribuciones de grados entre diferentes tejidos, como el tejido libre de cáncer, el tumor primario de cáncer de mama y la primera metástasis de cáncer de mama, se pueden obtener conocimientos valiosos sobre los cambios genéticos y las alteraciones de la red asociadas con la progresión del cáncer. Estas comparaciones pueden revelar diferencias en la conectividad de genes específicos, la presencia de subredes aberrantes o la reorganización de rutas de señalización clave.

En resumen, la ilustración de la red de genes y la figura de distribu-

Pos	Symbol	Grado	Promedio
1	LHFPL3	205100	2051
2	FGA	194900	1949
3	SALL3	194800	1948
4	EFNB2	178700	1787
5	P2RX7	177000	1770
6	PLK1	171900	1719
7	PRSS35	161100	1611
8	CDC20B	151700	1517
9	CDH26	144600	1446
10	GFI1B	141100	1411
11	CARM1	139200	1392
12	SLCO4A1	135500	1355
13	OR4C15	128500	1285
14	NPBWR2	126900	1269
15	OR8D2	123900	1239

Pos	Symbol	Grado	Promedio
16	LETM2	122700	1227
17	FBF1	122000	1220
18	ITLN1	120500	1205
19	OR9I1	119600	1196
20	DRGX	117800	1178
21	TTC29	114400	1144
21	PTGES2	114400	1144
23	RBM15	112900	1129
24	DCK	112400	1124
25	LYL1	111100	1111
26	SPRYD3	110500	1105
27	SFTPC	109700	1097
28	POLR2H	109300	1093
29	OR52B6	108900	1089
30	ITGA4	107700	1077

Tabla 4.8: 30 genes con mayor grado en redes de primera metástasis

Tejido	**Grado promedio**	**Coeficiente de agrupamiento**
Sin cancer	164.047857	0.474653
Tumor primario	56.971528	0.396157
Primera metástasis	154.273572	0.634926

Tabla 4.9: Centralidades de las 3 RG

ción de grados proporcionan una representación visual y cuantitativa de las redes genéticas en diferentes tejidos cancerosos y no cancerosos. Estas herramientas permiten identificar genes clave, patrones de conectividad y alteraciones genéticas asociadas con el cáncer de mama, lo que contribuye a una mejor comprensión de la biología del cáncer y a la identificación de posibles objetivos que serán abordados en el siguiente capítulo.

4.4. Patrones en las redes.

A continuación, se muestra una comparación entre los 10 primeros genes con mayor grado de cada tipo de tejido frente a los otros dos

(a) IM genes libres de cáncer.

(b) IM genes tumor primario.

(c) IM genes primera metástasis.

(d) Grado genes tejido libre de cáncer.

(e) Grado genes tumor primario.

(f) Grado genes primera metástasis.

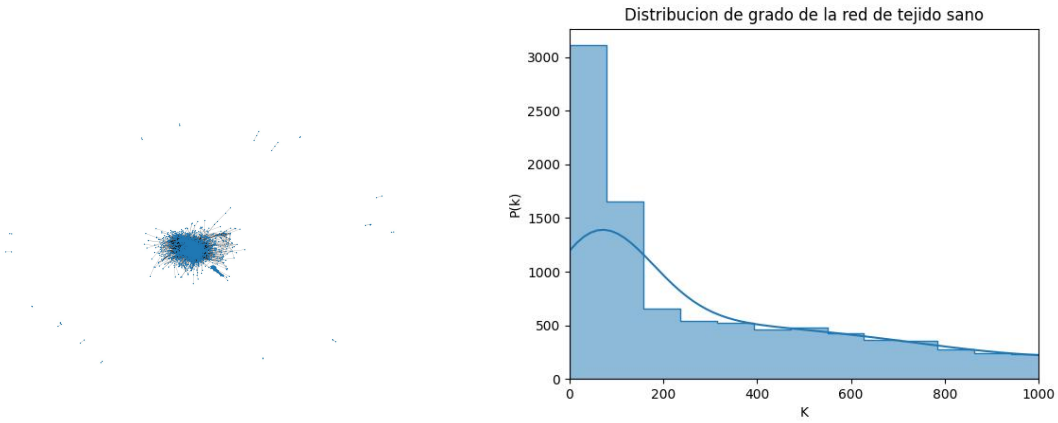
Figura 4.1: Códigos QR para la lista completa de pares de genes con su respectiva IM y lista completa de grado por tipo de tejido

El análisis de la Tabla 4.10 muestra una comparación entre los 10 primeros genes con mayor grado de expresión en diferentes tipos de tejidos: tejido libre de cáncer, tumor primario y primera metástasis. Cada columna representa un tipo de tejido, y los valores dentro de cada columna indican el lugar que ocupa cada gen en términos de grado de expresión, donde el primer gen tiene el mayor grado de expresión y así sucesivamente.

Algunas observaciones importantes que destacar del análisis son:

- **CTCF:** El gen CTCF ocupa el primer lugar en el tejido libre de cáncer, pero su grado de expresión disminuye significativamente en el tumor primario y la primera metástasis, donde ocupa los lugares 815 y 2760, respectivamente. Esta disminución podría estar asociada con su papel como regulador epigenético y supresor tumoral, ya que se ha demostrado que CTCF juega un papel crucial en la regulación de la expresión génica.

- **Genes N/D (No Disponible):** En la tabla, algunos genes se in-



(a) Red genética.

(b) Distribución de grado.

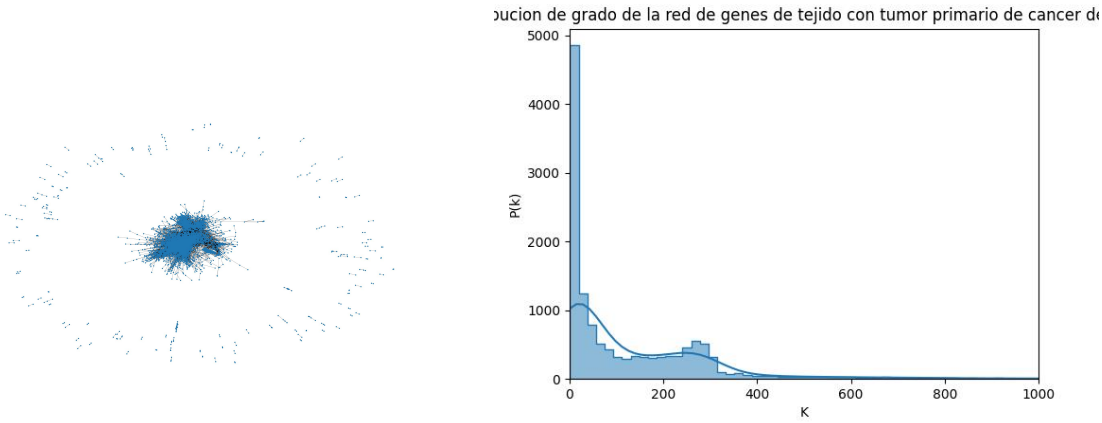
Figura 4.2: Ilustraciones para red libre de cáncer

dican como N/D, lo que significa que no están disponibles o no se detectaron en el tejido correspondiente. Estos genes podrían estar implicados en funciones específicas del tejido o podrían no tener una función relevante en ese contexto particular.

- **Diferencias entre tejido libre de cáncer y tumoral:** Se puede observar que los lugares de los genes cambian significativamente entre el tejido libre de cáncer y el tumor primario. Esto sugiere que la expresión génica puede estar regulada de manera diferente en el tejido sin cáncer y en el tumor primario, lo que podría ser relevante para entender la transformación tumoral y la progresión del cáncer.

- **Genes con alto grado en primera metástasis:** Algunos genes, como EIF4ENIF1, PRPF38A y IFNA1, muestran un alto grado de expresión en la primera metástasis en comparación con el tejido libre de cáncer y el tumor primario. Esto sugiere que estos genes pueden estar involucrados en el proceso de metástasis y podrían ser objetivos terapéuticos potenciales para abordar la propagación del cáncer.

- **Genes específicos para tejidos:** Algunos genes, como ADIG, AVP, ACTL7B, y CSN2, tienen un alto grado de expresión en tejidos es-

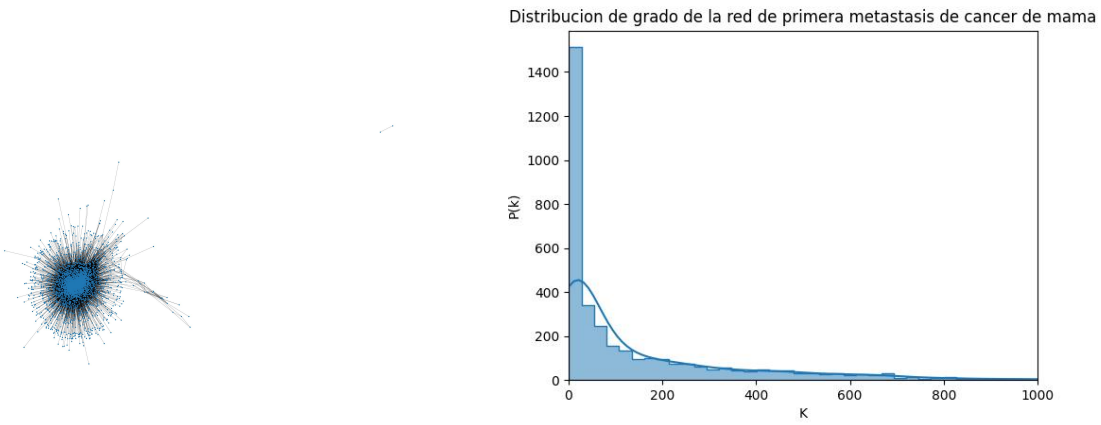


(a) Red genética. (b) Distribución de grado.

Figura 4.3: Ilustraciones para red de tumor primario.

pecíficos (ya sea tejido libre de cáncer, tumor primario o primera metástasis), lo que indica que estos genes podrían desempeñar un papel importante en la función y el mantenimiento de la homeostasis de esos tejidos específicos.

En general, esta comparación de los 10 primeros genes con mayor grado de expresión en diferentes tejidos proporciona una visión general de los cambios en la expresión génica asociados con la transformación tumoral y la metástasis. Es importante tener en cuenta que este análisis es una instantánea y que la regulación génica es un proceso complejo y dinámico que puede variar en diferentes etapas del cáncer y en diferentes tejidos. Por lo tanto, se requeriría un estudio más detallado para comprender completamente el papel de estos genes en el contexto del cáncer y su posible relevancia clínica.



(a) Red genética. (b) Distribución de grado.

Figura 4.4: Ilustraciones para red de primera metástasis.

Gene symbol	Lugar en tejido libre de cáncer	Lugar en tumor primario	Lugar en primera metástasis
CTCF	1	815	2760
GOSR1	2	1782	N/D
RTF1	3	3174	N/D
GPBP1L1	4	2602	N/D
API5	5	718	N/D
PLRG1	6	3683	N/D
EIF4ENIF1	7	4253	1892
PRPF38A	8	5384	3142
GPBP1	9	616	N/D
CTR9	10	2148	N/D
ADIG	4712	1	N/D
AVP	11808	2	N/D
ACTL7B	9953	3	N/D
NRN1L	10587	4	596
FABP1	N/D	5	2969
KRTAP102	10556	6	N/D
NPPB	10827	7	N/D
IFNA1	11204	8	2990
BARHL1	9840	9	1503
CSN2	12871	10	N/D
LHFPL3	N/D	52	1
FGA	N/D	N/D	2
SALL3	5836	N/D	3
EFNB2	12316	N/D	4
P2RX7	5573	6945	5
PLK1	10589	1014	6
PRSS35	N/D	N/D	7
CDC20B	N/D	N/D	8
CDH26	12526	N/D	9
GFI1B	N/D	N/D	10

Tabla 4.10: Comparación de los 10 genes con mayor grado en cada red genética de cada tejido frente a los otros 2

Capítulo 5

Análisis de genes principales

Gene [48] es una base de datos del Centro Nacional de Información sobre Biotecnología (NCBI). Proporciona información detallada sobre los genes identificados en diversos organismos; la utilizan ampliamente los investigadores para estudiar la función de los genes y su papel en la salud y la enfermedad. Cada registro en la base de datos está asociado con un identificador único de gene (GeneID). La información se organiza en varias secciones, e incluye del gen: la secuencia, la ubicación cromosómica, la función, la co-expresíon con otros genes, la variación genética, las relaciones de homología y las anotaciones de literatura científica. Los términos de búsqueda son el nombre, el identificador GeneID o términos asociados con la función o enfermedad relacionada con el gen. Los resultados reportan nombre, símbolo, longitud, función y ubicación cromosómica del gen, entre otros.

5.1. Funciones moleculares

El mapa de calor de la Figura 5.1 revela que los tejidos libres de cáncer presentan una mayor actividad en procesos dependientes del ATP, como el metabolismo celular, así como la actividad de moléculas adaptadoras que regulan la transducción de señales dentro de la célula y la transcripción del ADN. Además, se observa un aumento en la actividad de regulación transcripcional, que está asociada a las moléculas adaptadoras. En cuanto

al tumor primario, se observa un enriquecimiento en las funciones de regulación molecular y la actividad de moléculas transductores de señales, lo cual podría indicar una señalización celular mejorada o sobre activada, característica de la proliferación sostenida en el cáncer y, por lo tanto, una mayor actividad celular. En el tejido de la primera metástasis, se encuentra un aumento en la actividad de regulación de la traducción de proteínas y en la actividad de transporte. La regulación de la traducción refleja una regulación de las funciones celulares independiente de las modificaciones genéticas, mientras que la actividad de transporte indica un incremento en el transporte de lípidos y otros metabolitos intermedios dentro y fuera de la célula, lo que podría reflejar la característica de desregulación del metabolismo celular.

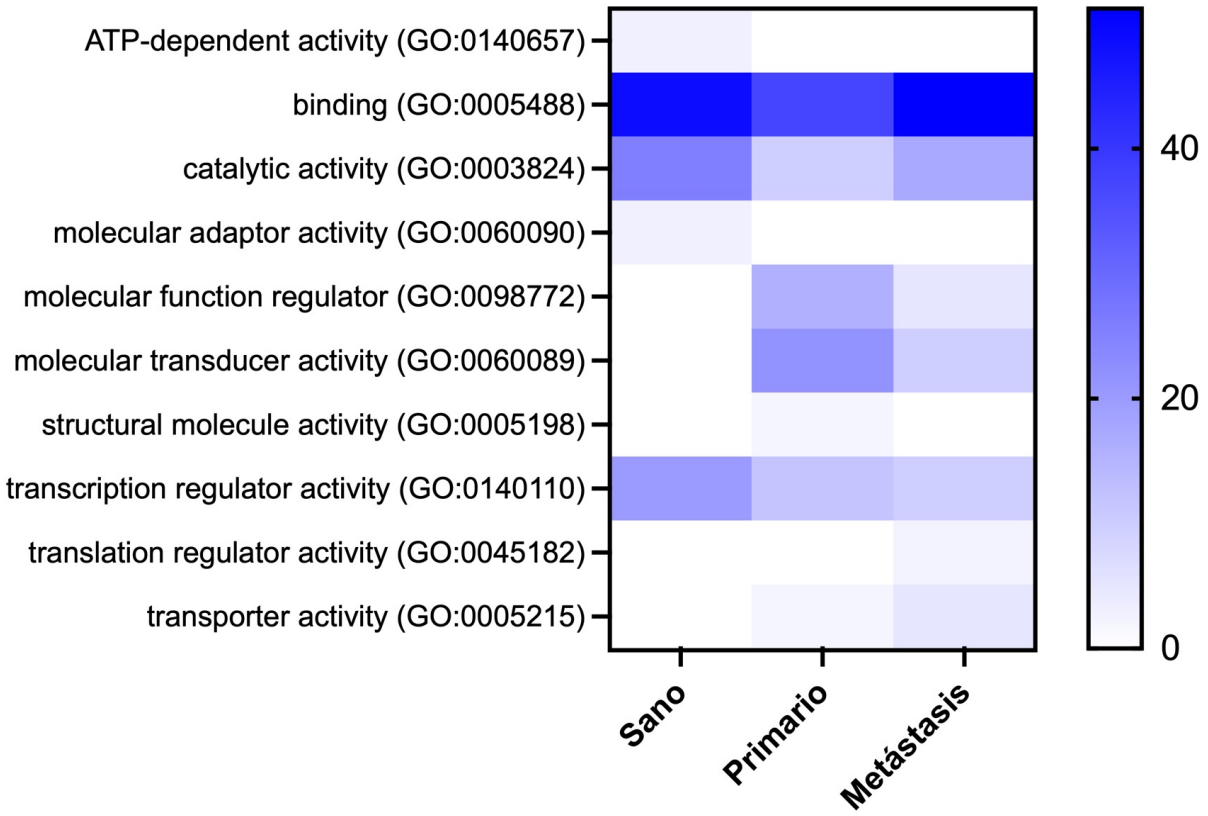


Figura 5.1: Mapa de calor de GO asociado a las funciones moleculares.

5.2. Procesos biológicos

En la Figura 5.2 se puede observar que, dentro de los tejidos libres de cáncer, se observa un enriquecimiento en diversos procesos biológicos, tales como la regulación biológica, los procesos celulares, la localización celular, los procesos metabólicos, la respuesta a estímulos y la señalización. Tanto el tumor primario como la primera metástasis comparten procesos enriquecidos, como la adhesión biológica, lo que indica que la adhesión celular es una de las principales características de los tumores. Además, ambos presentan un aumento en los procesos de desarrollo y procesos de organismos multicelulares, lo que se relaciona con la expresión de genes ligados al desarrollo embrionario y por ende refleja un estado de desdiferenciación celular. Sin embargo, se observan diferencias en el grado de enriquecimiento de estos procesos entre el tumor primario y la metástasis. Por ejemplo, los procesos de desarrollo y procesos de organismos multicelulares están más enriquecidas en el tumor primario que en la metástasis, mientras que el proceso de adhesión tiene un mayor enriquecimiento en la metástasis en comparación con el tumor primario. Por otro lado, se encuentra un enriquecimiento exclusivo de la locomoción en el tumor primario. Estos hallazgos sugieren que el proceso de formación del cáncer comparte diversos procesos biológicos, pero algunos de ellos pueden ser más relevantes para el tumor primario o la metástasis

5.3. Clases de proteínas

Las diferentes clases de proteínas revelan la función específica de las proteínas codificadas por los genes enriquecidos en nuestro análisis. Al igual que las funciones moleculares y los procesos biológicos, las clases de proteínas muestran enriquecimientos específicos en cada tipo de muestra, así como enriquecimientos compartidos tal y como se puede ver en la Figura 5.3. En el tejido libre de cáncer, destacan las proteínas relacionadas con la regulación de la transcripción específica de genes, las enzimas modificadoras de proteínas necesarias para activar vías de señalización y procesos

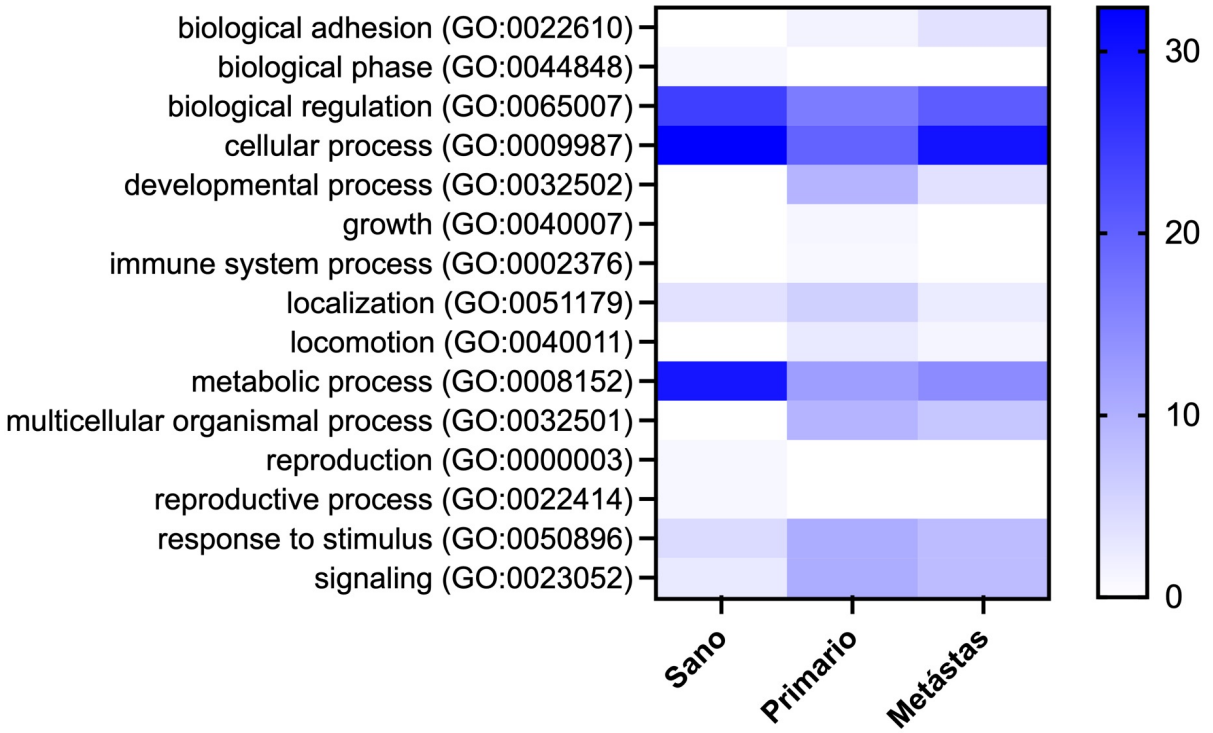


Figura 5.2: Mapa de calor de GO asociado a los procesos biológicos.

metabólicos, las proteínas involucradas en el metabolismo de ARN y ADN, y las proteínas relacionadas con el tráfico membranal. En el tumor primario, sobresalen las proteínas de unión al calcio, las proteínas chaperonas (que ayudan en el plegamiento de proteínas), las proteínas relacionadas con la inmunidad, las moléculas implicadas en la comunicación célula-célula, las proteínas de almacenamiento y las proteínas transportadoras de iones. Estas proteínas son más destacadas en comparación con la metástasis. Por otro lado, en la metástasis, se observa un enriquecimiento de las proteínas de adhesión celular y de unión célula-célula, lo cual sugiere el proceso de transición mesénquima-epitelio necesario para el establecimiento de un nuevo tumor. El enriquecimiento de proteínas reguladoras de la cromatina sugiere la participación de mecanismos de transformación dependientes de la regulación epigenética. Además, el enriquecimiento de proteínas adaptadoras, estructurales y aquellas que participan en la regulación de la expresión de proteínas sugiere una regulación postranscripcional compleja. Por últi-

mo, el enriquecimiento en actividades de transportadores indica un mayor intercambio de componentes tanto extracelulares como intracelulares en el microambiente tumoral.

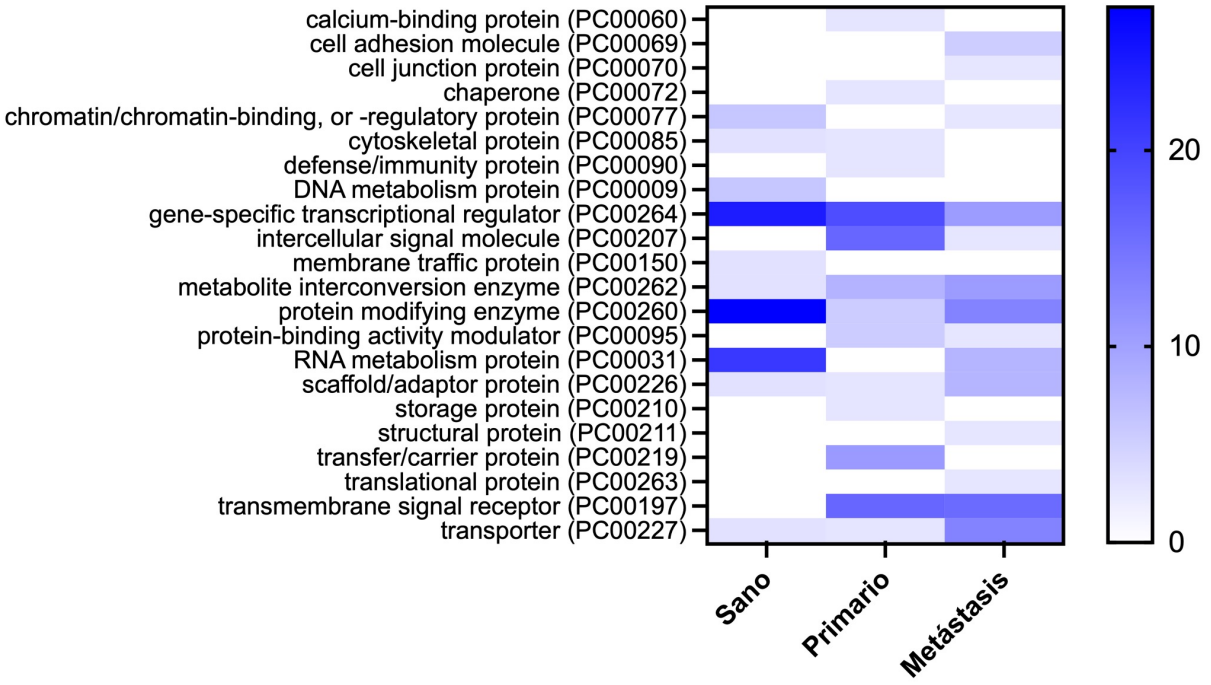


Figura 5.3: Mapa de calor de GO asociado a las clases de proteínas.

5.4. Vías de señalización

Las vías de señalización celular son procesos mediante los cuales las señales extracelulares se transmiten al interior de la célula, lo que resulta en modificaciones de las proteínas intracelulares y cambios en la expresión génica, que a su vez culminan en alteraciones en el comportamiento o fenotipo celular. A diferencia de los mapas de calor presentados anteriormente, nuestro análisis de las vías de señalización de los genes, ver Figura 5.4, revela una firma prácticamente exclusiva de vías presentes en cada tipo de tejido. De las 29 vías de señalización analizadas, solo 3 se comparten entre dos tipos de muestras: la vía de la enfermedad de Huntington, relacionada

con proteínas que regulan la forma celular; la vía de citocinas y quimiocinas mediadas por la inflamación; y la vía Wnt, involucrada en el control de la diferenciación, proliferación y migración celular. En el tejido libre de cáncer, se observa la activación de las vías de interleucinas y receptores tipo Toll, lo cual indica la presencia de procesos de activación y regulación de la inflamación. Además, la activación de la vía de la MAPK (cinasa activada por mitógenos) y la activación de TGF-b (factor transformante b) sugieren un proceso de reparación celular en curso. Por otro lado, la activación de la vía p53, también conocida como el guardián del genoma, revela un proceso de control del ciclo celular. En el tumor primario, destacan la vía de señalización mediada por el factor de crecimiento de fibroblastos (FGF), el cual está relacionado con la proliferación sostenida en procesos tumorales. La señalización de cadherinas también se destaca, ya que está relacionada con el proceso de transición epitelio-mesénquima, requerida para el movimiento celular. La señalización mediada por proteínas G indica una transducción de señales elevada a través de receptores acoplados a proteínas G (GPCR), que son los receptores de transducción de señales más abundantes en las células eucariotas. La vía de señalización mediada por el factor de crecimiento derivado de plaquetas (PDGF) sugiere que las células están en constante proliferación. Por otro lado, en la metástasis, destacan la vía de la angiogénesis, la coagulación sanguínea y la activación del plasminógeno, lo que indica un proceso activo de vascularización. La regulación del citoesqueleto sugiere procesos de movimiento celular. A diferencia del tumor primario, en la metástasis, la proliferación sostenida puede estar impulsada por el factor de crecimiento epidemial (EGF), como se muestra en el enriquecimiento de esta vía. El enriquecimiento de la vía de señalización de las integrinas también podría indicar un proceso de proliferación inducido por la adhesión de las células en el sitio de la metástasis. El enriquecimiento de la vía de señalización de Ras está bien definido como un proceso asociado a la oncogénesis, que induce la proliferación, diferenciación y supervivencia celular.

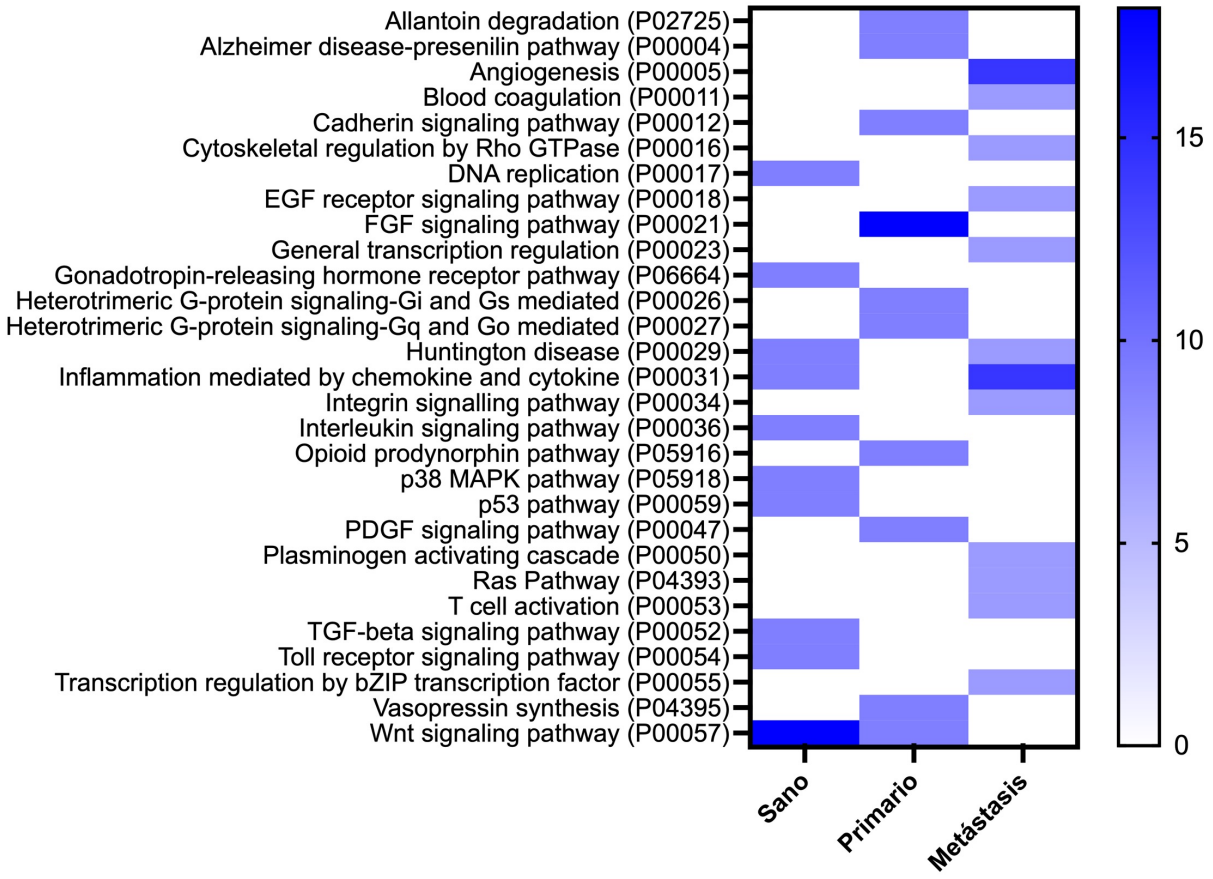


Figura 5.4: Mapa de calor de GO asociado a las vías de señalización.

5.5. Genes Principales por Tipo de Tejido

Nuestro análisis de Gene Ontology revela que el análisis de las redes genéticas identifica genes con un mayor grado de asociación a procesos oncogénicos, como los descritos anteriormente, y nos permite identificar genes exclusivos en cada tipo de muestra, principalmente en relación con las vías de señalización y otros procesos biológicos.

La Tabla 5.1 muestra las principales vías de señalización y los genes asociados en el tejido libre de cáncer. La sección se titula "Libre de cáncerz está enfocada en las vías de señalización y los genes principales relacionados con la ausencia de cáncer en el tejido. Los genes principales asociados con cada vía de señalización mencionada. El gen MAP3K7 aparece en varias

vías de señalización, lo que sugiere su papel relevante en el tejido libre de cáncer. El gen SIN3A también se encuentra en dos vías de señalización. Además, se mencionan otros genes específicos, como SMARCA5 en la vía de señalización de Wnt y TOP2B en la replicación del ADN. En resumen, proporciona una visión de las principales vías de señalización y los genes asociados que se encuentran en el tejido libre de cáncer. Estos genes y vías de señalización desempeñan un papel crucial en el mantenimiento de la salud y la prevención del desarrollo de células cancerosas en el tejido en cuestión.

Libre de cáncer.	
Vías de Señalización	**Genes Principales**
Vía de señalización del receptor de la hormona liberadora de gonadotropinas (P06664)	MAP3K7
Vía de señalización de la interleucina (P00036)	MAP3K7
Vía del p53 (P00059)	SIN3A
Enfermedad de Huntington (P00029)	SIN3A
Inflamación mediada por la señalización de quimiocinas y citocinas (P00031)	MAP3K7
Vía de señalización de p38 MAPK (P05918)	MAP3K7
Vía de señalización de Wnt (P00057)	MAP3K7 SMARCA5
Vía de señalización del receptor Toll (P00054)	MAP3K7
Vía de señalización del TGF-beta (P00052)	MAP3K7
Replicación del ADN (P00017)	TOP2B

Tabla 5.1: Principales genes y vías de señalización en la RG de tejido libre de cáncer

La Tabla 5.2 proporciona información sobre los principales genes y vías de señalización en el tumor primario. Al igual que en las descripciones

anteriores, esta tabla se divide en dos columnas: "Vías de Señalizaciónz "Genes Principales". En la columna "Vías de Señalización", se enumeran diversas vías de señalización que están asociadas con funciones específicas en el tumor primario. En la columna "Genes Principales", se mencionan los genes principales asociados con cada vía de señalización. Estos genes son los siguientes:

- ALLC: Relacionado con la degradación de la alantoína [49, 50].
- AVP: Asociado tanto con la síntesis de la vasopresina como con la vía de prodiomorfina opioides [51].
- WNT8A: Presente en la vía de la enfermedad de Alzheimer-presenilina y la vía de señalización de la cadherina [52, 53].
- SSTR4: Vinculado con la vía de señalización de proteínas G heterotriméricas mediada por la proteína alfa Gi y la proteína alfa Gs, así como la vía mediada por la proteína alfa GQ y la proteína alfa Go [54].
- FGF8 y FGF3: Ambos genes están presentes en la vía de señalización del factor de crecimiento de fibroblastos [55, 56].
- FEV: Relacionado con la vía de señalización del factor de crecimiento derivado de las plaquetas (PDGF) [57]. .

En resumen, la tabla muestra las vías de señalización y los genes principales asociados con el tumor primario. Se destacan varios genes, como AVP, WNT8A y SSTR4, que están presentes en múltiples vías de señalización. Esto sugiere su importancia en la regulación de funciones específicas en el contexto del tumor primario. Además, los genes FGF8, FGF3 y FEV también se identifican como genes clave en vías de señalización específicas relacionadas con el crecimiento y la proliferación celular.

La Tabla 5.3 muestra las vías de señalización y los genes principales asociados a la primera metástasis. Aquí hay algunos puntos destacados basados en la información de la tabla:

Tumor primario	
Vías de Señalización	**Genes Principales**
Degradación de la alantoína (P02725)	ALLC
Síntesis de la vasopresina (P04395)	AVP
Vía de la enfermedad de Alzheimer-presenilina (P00004)	WNT8A
Vía de la prodiomorfina opioides (P05916)	AVP
Vía de señalización de la cadherina (P00012)	WNT8A
Vía de señalización de proteínas G heterotriméricas - vía mediada por la proteína alfa Gi y la proteína alfa Gs (P00026)	SSTR4
Vía de señalización de proteínas G heterotriméricas - vía mediada por la proteína alfa GQ y la proteína alfa Go (P00027)	SSTR4
Vía de señalización de Wnt (P00057)	WNT8A
Vía de señalización del factor de crecimiento de fibroblastos (FGF) (P00021)	FGF8 FGF3
Vía de señalización del factor de crecimiento derivado de las plaquetas (PDGF) (P00047)	FEV

Tabla 5.2: Principales genes y vías de señalización en la RG de tumor primario

- **Vías de Señalización:** La tabla enumera varias vías de señalización que están implicadas en la primera metástasis. Algunas de estas vías incluyen la coagulación sanguínea, la angiogénesis, la vía de señalización de integrinas, la enfermedad de Huntington, la inflamación mediada por la señalización de quimiocinas y citocinas, la regulación de la transcripción por factores de transcripción bZIP, la activación de células T, la regulación general de la transcripción, la cascada de activación del plasminógeno, la vía de señalización del receptor del factor de crecimiento epidérmico y la vía de señalización de Ras y la regulación del citoesqueleto por Rho GTPasa.

- **Genes Principales:** La tabla también muestra los genes principales asociados a cada vía de señalización mencionada. Estos genes representan las moléculas clave que desempeñan un papel importante en la primera metástasis. Algunos de los genes mencionados incluyen FGA, PAK3, EFNB2, ITGA4, RHOQ y POLR2H.

Es importante tener en cuenta que esta es solo una vista general y simplificada de los genes y vías de señalización asociados a la primera metástasis. Para obtener una comprensión más completa y precisa, se requeriría un análisis más detallado e integrado de múltiples fuentes de datos y literatura científica.

La Tabla 5.4 muestra los genes principales por tipo de tejido. Está dividida en tres secciones: "Libre de Cáncer", "Tumor Primarioƶ "Primera Metástasis". Cada sección contiene una lista de genes y sus respectivas anotaciones.

En la sección "Libre de Cáncer", se mencionan cinco genes principales: MAP3K7, SIN3A, SMARCA5, TOP2B y CTCF. Estos genes están presentes en las principales vías de señalización, lo que sugiere su importancia en el funcionamiento normal del tejido. Además, se mencionan tres genes adicionales: RPL31, RPL37 y CTCF. El gen CTCF se destaca por tener el mayor grado, lo que puede indicar su influencia significativa en el tejido, mientras que RPL31 y RPL37 son genes con el mayor Índice de Mutación (IM).

En la sección "Tumor Primario", se mencionan varios genes, incluyendo ALLC, AVP, FEV, FGF3, FGF8 y SSTR4, que también están presentes en las principales vías de señalización. Además, se mencionan los genes ADIG, OR10P1 y PCDHA1, con ADIG como el gen con el mayor grado y OR10P1 y PCDHA1 como los genes con el mayor IM.

En la sección "Primera Metástasis", se mencionan los genes EFNB2, FGA, ITGA4, PAK3, POLR2H y RHOQ, que están presentes en las principales vías de señalización. Además, se mencionan los genes LHFPL3 y

Primera metástasis	
Vías de Señalización	**Genes Principales**
Coagulación sanguínea (P00011)	FGA
Angiogénesis (P00005)	PAK3 EFNB2
Vía de señalización de integrinas (P00034)	ITGA4
Enfermedad de Huntington (P00029)	RHOQ
Inflamación mediada por la señalización de quimiocinas y citocinas (P00031)	PAK3 ITGA4
Regulación de la transcripción por factores de transcripción bZIP (P00055)	POLR2H
Activación de células T (P00053)	PAK3
Regulación general de la transcripción (P00023)	POLR2H
Cascada de activación del plasminógeno (P00050)	FGA
Vía de señalización del receptor del factor de crecimiento epidérmico (P00018)	RHOQ
Vía de señalización de Ras (P04393) Regulación del citoesqueleto por Rho GTPasa (P00016)	PAK3

Tabla 5.3: Principales genes y vías de señalización en la RG de primera metástasis

SALL3, con LHFPL3 como el gen con el mayor grado y SALL3 como el gen con el mayor IM. También se menciona un segundo gen con el mayor grado, pero no se especifica su nombre.

A continuación, se hace una revisión de la literatura sobre los genes principales, para corroborar el resultado de salida de GO.

Genes Principales por Tipo de Tejido		
Tipo de Tejido	**Genes Principales**	**Anotaciones**
Libre de Cáncer	MAP3K7	Presente en principales vías de señalización
	SIN3A	Presente en principales vías de señalización
	SMARCA5	Presente en principales vías de señalización
	TOP2B	Presente en principales vías de señalización
	CTCF	Gen con mayor grado
	RPL31	Gen con mayor IM
	RPL37	Gen con mayor IM
Tumor Primario	ALLC	Presente en principales vías de señalización
	AVP	Presente en principales vías de señalización
		Segundo gen con mayor grado
	FEV	Presente en principales vías de señalización
	FGF3	Presente en principales vías de señalización
	FGF8	Presente en principales vías de señalización
	SSTR4	Presente en principales vías de señalización
	ADIG	Gen con mayor grado
	OR10P1	Gen con mayor IM
	PCDHA1	Gen con mayor IM
Primera Metástasis	EFNB2	Presente en principales vías de señalización
	FGA	Presente en principales vías de señalización
		Segundo gen con mayor grado
		Gen con mayor IM
	ITGA4	Presente en principales vías de señalización
	PAK3	Presente en principales vías de señalización
	POLR2H	Presente en principales vías de señalización
	RHOQ	Presente en principales vías de señalización
	LHFPL3	Gen con mayor grado
	SALL3	Gen con mayor IM

Tabla 5.4: Genes principales por tipo de tejido

- **CTCF**: Ocupa el primer lugar en la red de tejido libre de cáncer de seno y está presente en la red de tumor primario y primera metástasis, pero en las posiciones 815 y 2760 respectivamente. El gen CTCF, que significa ÇCCTC-binding factor"(Factor de unión a CCCTC), es un gen clave en la regulación de la estructura del ADN y la expresión génica [58]. El factor de transcripción CTCF codificado por este gen desempeña un papel crucial en la organización tridimensional del ge-

noma, facilitando la interacción entre las regiones del ADN distantes y regulando la accesibilidad de los genes a los factores de transcripción. Se ha descubierto que las mutaciones o alteraciones en el gen CTCF están asociadas con varios tipos de cáncer [59]. Estas alteraciones pueden afectar la función normal de CTCF y contribuir al desarrollo y progresión de los tumores. Algunas de las implicaciones más comunes del gen CTCF en el cáncer incluyen:

- Regulación génica: CTCF juega un papel crítico en la regulación de la expresión génica. Las mutaciones en el gen CTCF pueden alterar la unión de CTCF al ADN, afectando la expresión de genes involucrados en la proliferación celular, la apoptosis y otros procesos celulares relacionados con el cáncer.
- Alteraciones de la estructura del ADN: CTCF participa en la organización tridimensional del genoma y en la formación de estructuras como los bucles de ADN. Las mutaciones en el gen CTCF pueden interferir con la formación adecuada de estas estructuras, lo que puede tener un impacto en la regulación de los genes relacionados con el cáncer.
- Inestabilidad genómica: La pérdida o alteración de la función de CTCF puede contribuir a la inestabilidad genómica, que es una característica común de muchos tipos de cáncer. Esto puede llevar a cambios en el número de copias de los genes, re-arreglos cromosómicos y otras aberraciones genómicas asociadas con la progresión del cáncer.

- **GOSR1**: Ocupa el segundo lugar en la red de tejido libre de cáncer de seno, además de ser uno de los genes principales en las funciones molecular, es un gen que participa en la actividad del adaptador molecular y también está presente en la red de tejido de tumor primario en la posición 1782. El gen GOSR1, también conocido como golgi SNAP receptor complex member 1, es un gen que codifica una proteína que participa en el proceso de transporte vesicular en el complejo de Golgi. La proteína GOSR1 es un componente esencial del complejo de

fusión de membrana que media la transferencia de proteínas y lípidos desde el Golgi hasta otras estructuras celulares, como las vesículas secretoras. Se ha demostrado que las alteraciones en el gen GOSR1 pueden estar asociadas con diferentes enfermedades y trastornos. Si bien la mayoría de las investigaciones se han centrado en su relación con enfermedades neurológicas, como la enfermedad de Alzheimer y la enfermedad de Parkinson, también se ha estudiado su participación en el cáncer. En el contexto del cáncer, se ha observado que el gen GOSR1 puede estar implicado en la proliferación celular, la invasión y la metástasis [60, 61]. Algunos estudios han mostrado una sobreexpresión de GOSR1 en diversos tipos de cáncer, como cáncer de mama, cáncer de colon y cáncer de pulmón, y se ha sugerido que puede estar asociado con un peor pronóstico en estos pacientes.

- **RTF1**: Ocupa el tercer lugar en la red de tejido libre de cáncer de seno, también está presente en la red de tejido de tumor, pero en la posición 3174. El gen RTF1, también conocido como RNA polymerase-associated protein RTF1, es un gen que codifica una proteína implicada en la regulación de la transcripción génica [62]. RTF1 es un componente esencial del complejo de elongación de ARN polimerasa II (ARNpol II) y juega un papel crucial en la transcripción de genes codificantes de ARN mensajero (ARNm).En relación con el cáncer, se ha investigado el papel del gen RTF1 en la tumor-génesis y la progresión del cáncer. Se han identificado alteraciones en la expresión de RTF1 en varios tipos de cáncer, incluyendo cáncer de mama, cáncer de colon y cáncer de pulmón. Estudios han demostrado que RTF1 puede regular la expresión de genes involucrados en la proliferación celular, la invasión y la metástasis, lo que sugiere su posible contribución a la progresión tumoral. [63].

- **RPL31**: Junto con RPL37 forman el par de genes con mayor MI en la red libre de cáncer de mama. Los ribosomas, los orgánulos que catalizan la síntesis de proteínas, constan de una subunidad 40S pequeña y una subunidad 60S grande. Juntas, estas subunidades están

compuestas por 4 especies de ARN y aproximadamente 80 proteínas estructuralmente distintas. Este gen codifica una proteína ribosómica que es un componente de la subunidad 60S. La proteína pertenece a la familia L31E de proteínas ribosómicas. Se encuentra en el citoplasma. Se han observado niveles más altos de expresión de este gen en pólipos adenomatosos familiares en comparación con tejidos normales emparejados. Como es típico de los genes que codifican proteínas ribosómicas, existen múltiples pseudogenes procesados de este gen dispersos por el genoma.

- **RPL37**: Junto con RPL31 forman el par de genes con mayor MI en la red libre de cáncer de mama. La proteína pertenece a la familia L37E de proteínas ribosómicas. Se encuentra en el citoplasma. La proteína contiene un motivo similar a un dedo de zinc de tipo C2C2. Como es típico de los genes que codifican proteínas ribosómicas, existen múltiples pseudogenes procesados de este gen dispersos por el genoma.

- **MAP3K7**:Lugar 39 en la red de tejido libre de cáncer con mayor grado, se presenta como uno de los genes con función biológica de señalización en este mismo tejido

- **MAPKAPK5**:Lugar 42 en la red de tejido libre de cáncer con mayor grado, se presenta como uno de los genes con función biológica de señalización en este mismo tejido

- **AVP**: Este gen es el 2do con mayor grado en redes de tumor primario y aparece en la red libre de cáncer, es el encargado un miembro de la familia de la vasopresina/oxitocina y una preproproteína que se procesa proteolíticamente para generar múltiples productos proteicos. Las células de cáncer de mama expresan anormalmente vasopresina (AVP) y sus receptores. El efecto de AVP se orquesta en gran medida a través de su señalización aguas abajo y por endocitosis mediada por receptor (RME), en la que Dynamin 2 (Dyn2) desempeña un papel integral en el cierre de vesículas [64].

- **ACTL7B**: Ocupa el tercer lugar en la red de tumor primario. El ACTL7B (Actin Like 7B) es un gen que codifica una proteína perteneciente a la familia de las actinas. Las actinas son proteínas fundamentales en la organización y función del citoesqueleto, que es una red de filamentos proteicos que proporciona soporte estructural a la célula y participa en procesos celulares como la división, migración y contracción. Se ha observado que el ACTL7B desempeña varios roles en diferentes contextos biológicos. Por ejemplo, se ha encontrado que está involucrado en la organización del núcleo celular y en la regulación de la transcripción génica. Además, se ha identificado que el ACTL7B interactúa con otras proteínas y participa en la formación de complejos multiproteicos relacionados con la remodelación de la cromatina. En cuanto a su relación con el cáncer, se han realizado investigaciones que sugieren que el ACTL7B podría desempeñar un papel en la progresión tumoral y la metástasis en ciertos tipos de cáncer. Por ejemplo, se ha observado que la expresión alterada del ACTL7B está asociada con el crecimiento y la invasión celular en el cáncer de mama y el cáncer de colon [65]

- **OR10P1**:Ocupa el primer lugar, junto con OR10P1, de los pares de genes con mayor MI junto en la red de tumor primario, además es el numero quince con mayor grado en esta misma red.

- **PCDHA1**:Ocupa el primer lugar, junto con PCDHA1, de los pares de genes con mayor MI junto en la red de tumor primario y está en la posición 782 de los genes con mayor grado.

- **KRT74**: Ocupa el lugar 45 en los genes de mayor grado en la red de tumor primario. Es uno de los genes que tiene presencia en los procesos biológicos diferenciales a los otros dos tipos de tejido, participa en la actividad de la molécula estructural.

- **SLCO1B3**: Es uno de los genes que tiene presencia en los procesos biológicos de la red de tumor primario participando en la actividad de transporte. Además, ocupa la posición veinte de los genes con mayor

grado en esta misma red

- **LHFPL3**: Ocupa el primer lugar en la red de primera metástasis y está presente en varias ocasiones dentro de los genes con mayor MI en la red de este mismo tejido. En 2019, LHFPL3 aparece en un trabajo de investigación como objetivo para regular la proliferación, la migración y las transiciones epitelio-mesenquimatosas de células de glioma humano[66]. Lo llamativo de este gen es que está presente en las muestras de tumor primario y en las muestras de metástasis en un lugar relativamente bajo en los genes con mayor grado en la RG.

- **FGA**: Ocupa el segundo lugar en la red de primera metástasis, además de ser el que mayor MI presenta junto con SALL3. El gen FGA codifica la subunidad alfa del fibrinógeno, una proteína involucrada en la coagulación sanguínea. El fibrinógeno es convertido en fibrina durante la cascada de coagulación, lo que conduce a la formación de coágulos sanguíneos. Además de su función en la coagulación, se ha encontrado que el gen FGA tiene implicaciones en diversas enfermedades, incluido el cáncer. En relación con el cáncer, se han realizado investigaciones que han demostrado una asociación entre variaciones en el gen FGA y ciertos tipos de cáncer. Por ejemplo, estudios han identificado polimorfismos en el gen FGA que pueden estar relacionados con un mayor riesgo de desarrollar cáncer de mama, cáncer de colon y otros tipos de cáncer [67, 68]

- **SALL3**: Ocupa el tercer lugar en la red de primera metástasis, además de ser el que mayor MI presenta junto con FGA

- **P2RX7**:Ocupa el segundo lugar en los pares de genes con mayor MI junto con FGA en la red de primera metástasis, además de ser uno de los genes principales en las funciones molecular y procesos biológicos, es un gen que participa en la actividad de transporte.

- **SEC61A1**: Además de ser uno de los genes principales en las funciones molecular y procesos biológicos, es un gen que participa en

la actividad de transporte, se encuentra también entre los genes con mayor grado, la posición 48 para ser más específicos.

- **MAP3K7**: Es una serina/treonina quinasa, también conocida como TAK1, que, como su nombre indica, agrega grupos fosfato a los aminoácidos serina y treonina de las proteínas [69] (Entrada: O43318). Esta modificación es esencial para la activación o desactivación de las funciones de las proteínas dentro de las vías de señalización de citocinas, factores de crecimiento y receptores inmunológicos innatos [70]. Según el Atlas de Proteínas Humanas [71], esta proteína se expresa principalmente en tejido respiratorio sano, estómago, senos y trompas de Falopio. Dentro de la mama, las principales células que expresan este gen son los fibroblastos y las células glandulares. Se expresa moderadamente en el cáncer de mama, pero no se sabe cómo contribuye al desarrollo del cáncer. Sin embargo, su expresión es un factor pronóstico desfavorable para pacientes con cáncer de hígado [72]. Probablemente porque esta proteína promueve la proliferación, migración e invasión de las células hepatocelulares. El análisis genético identificó mutaciones en esta proteína en pacientes con mesotelioma [73]. Además, los pacientes con mieloma múltiple y una expresión elevada de esta proteína tienen una menor supervivencia libre de progresión y una menor supervivencia general [74]. También se asocia con la progresión del melanoma [75]. Por el contrario, se ha demostrado que esta proteína tiene una función supresora de tumores en el cáncer de próstata [76, 77, 78]. A pesar de lo anterior, hasta la fecha, no hay asociación de esta proteína con el desarrollo de cáncer de mama, lo cual es consistente con nuestros resultados.

Capítulo 6

Discusión

La aplicación de algoritmos de Teoría de la Información en el análisis de redes genéticas ha demostrado ser una herramienta eficaz para identificar genes principales en el cáncer de seno. Estos algoritmos permiten cuantificar la correlación y la información mutua entre los genes, lo que proporciona una medida de su interacción y grado de relevancia en el contexto del cáncer de seno.

Al utilizar estos algoritmos, se procesan los datos de expresión génica de pacientes con cáncer de seno y se construye una red génica que representa las relaciones entre los genes. La Teoría de la Información proporciona una base matemática sólida para medir la información compartida entre los genes y su dependencia mutua. Esto permite identificar los genes con el mayor grado de conexión con otros genes y los pares de genes más fuertemente relacionados en la red génica.

La identificación de los genes principales en el cáncer de seno es de vital importancia, ya que estos genes desempeñan un papel crucial en los procesos biológicos asociados con la enfermedad. Estos genes pueden estar involucrados en la regulación del crecimiento celular, la proliferación, la apoptosis y otros procesos clave en el desarrollo y progresión del cáncer de seno.

Al utilizar algoritmos de Teoría de la Información, se puede obtener una visión más completa de la interacción entre los genes y su contribución a la enfermedad. Esto permite identificar los genes con mayor influencia y potencialmente dirigirse a ellos para el desarrollo de terapias más efectivas

y personalizadas.

Además, el análisis de redes genéticas basado en Teoría de la Información supera las limitaciones de los enfoques convencionales de análisis de expresión génica que se centran en la sobreexpresión o regulación descendente de genes individuales. Estos enfoques no tienen en cuenta las interacciones entre genes y pueden pasar por alto genes importantes que no muestran cambios significativos en su expresión. En cambio, la forma en que se abordó este problema en la tesis toma en cuenta la complejidad de las interacciones genéticas y pueden revelar conexiones sutiles y relevantes entre genes.

Adicionalmente, mostramos un análisis de la función conocida de las proteínas codificadas por los genes encontrados. Sin embargo, no hay información extensa sobre su función específica en el cáncer de mama. Por lo tanto, como próximo paso, sugerimos validar la expresión de estos genes a nivel de proteínas en cada uno de los tipos de tejido analizados. Además, se requieren cohortes de pacientes con diferentes etapas de cáncer para determinar si alguno de los genes encontrados está asociado positiva o negativamente con la supervivencia sin progresión y la supervivencia global en el cáncer de mama y validar su uso clínico

En resumen, el algoritmo desarrollado es una poderosa herramienta para analizar redes genéticas e identificar genes principales en el cáncer, en este caso, se hizo el análisis con cáncer de seno, sin embargo, se podría aplicar para cualquier tipo de cáncer. Estos algoritmos permiten una comprensión más completa de las interacciones genéticas y pueden revelar nuevos objetivos terapéuticos y biomarcadores para mejorar el diagnóstico, pronóstico y tratamiento de esta enfermedad. Su aplicación promete avances significativos en la investigación del cáncer de seno y en la búsqueda de terapias más efectivas y personalizadas.

Capítulo 7

Conclusiones y trabajos futuros

A manera de conclusión se exponen los siguientes puntos:

- Con algoritmos basados en Teoría de la Información se identificaron los genes que tienen una mayor relevancia en la formación del cáncer de mama y su metástasis.

- La obtención y normalización adecuada de datos de GDCDP y GEO para muestras de cáncer de mama es crucial en la formación de las redes genéticas para evitar falsos positivos que afectan a los resultados de la investigación.

- Con el diccionario creado se vinculan los nombres de los genes coincidentes, lo cual facilita la integración y comparación de los datos entre diferentes bases de datos.

- El grado de conexión, la distribución del grado y el coeficiente de agrupamiento, permitieron identificar los nodos más importantes en las RG generadas. Estas medidas cuantifican la influencia y el papel que desempeñan los nodos individuales en la propagación de información, la estructura global de la red asociadas a cada tejido.

- A través del análisis estructural de estas redes genéticas, se han identificado los nodos-gen con más información mutua y mayor grado dentro de la red.

- Se analizó la importancia de estos genes clave en los tres tipos de tejidos mediante el estudio de los principales procesos biológicos y funciones moleculares en los que participan.
- El aporte de esta tesis es la metodología: crear las redes genéticas, identificar los nodos con mayor grado e información mutua en los tres tipos de tejidos, y validar su relevancia utilizando herramientas de análisis genético. Esta metodología puede ser aplicable no solo al cáncer de seno, sino también a otros tipos de cáncer.

Posibles trabajos futuros factibles de abordar con el material de esta tesis:

- Expansión a otros tipos de cáncer: Se podría considera aplicar la metodología utilizada en esta tesis a otros tipos de cáncer. Explora cómo las redes genéticas difieren entre diferentes tipos de tejidos cancerosos y cómo los genes clave pueden variar en su importancia y función molecular en cada caso.
- Análisis de interacciones gen-gen en las redes: Profundizar en el análisis de las interacciones gen-gen en las redes genéticas que se han construido. Utilizando técnicas avanzadas de análisis de redes para identificar subredes específicas de interés y comprender cómo las interacciones entre genes pueden influir en la formación y progresión del cáncer.
- Incorporación de datos clínicos: Dada la robustez de la metodología, se podrían integrar datos clínicos de los pacientes, como información sobre el estadio del cáncer, la respuesta al tratamiento o los resultados de supervivencia y relacionar estos datos con la información genética para obtener una visión más completa de los factores que contribuyen a la progresión del cáncer y la respuesta al tratamiento.
- Validación experimental de genes clave: Realizar experimentos de validación para confirmar la importancia funcional de los genes identificados como clave en este análisis. Esto puede implicar estudios in vitro o in vivo para evaluar el efecto de la modulación de la expresión

de estos genes en la proliferación celular, la invasión o la resistencia al tratamiento.

En resumen, esta tesis proporciona una base sólida para continuar investigando y comprendiendo mejor los mecanismos genéticos subyacentes al cáncer de seno, y abre nuevas oportunidades para el desarrollo de enfoques terapéuticos más precisos y efectivos.

Bibliografía

[1] Estadísticas a propósito del día mundial de la lucha contra el cáncer de mama (19 de octubre), 18 de octubre de 2021. accessed: 01.09.2022.

[2] Olga Blomberg, Lorenzo Spagnuolo, and Karin de Visser. Immune regulation of metastasis: mechanistic insights and therapeutic opportunities. *Disease Models y Mechanisms*, 11:dmm036236, 10 2018.

[3] Alex Jinnah, Benjamin Zacks, Chukwuweike Gwam, and Bethany Kerr. Emerging and established models of bone metastasis. *Cancers*, 10:176, 06 2018.

[4] National Cancer Institute. Genomic data commons, sep 2022.

[5] Gene expression omnibus. accessed: 09.2022.

[6] El comerco Agencia EFE. Científicos mexicanos crean modelos computacionales para mejorar medicamentos, 07 de enero de 2020. accessed: 12.2022.

[7] European Comission. Grandes logros de proyectos - la medicina in silico llega a la clínica, 31 Marzo 2017. accessed: 12.2022.

[8] Paul K. Newton, Jeremy Mason, Kelly Bethel, Lyudmila A. Bazhenova, Jorge Nieva, and Peter Kuhn. A stochastic markov chain model to describe lung cancer growth and metastasis. *PLOS ONE*, 7(4), 04 2012.

[9] Jeffrey West and et al. Cellular interactions constrain tumor growth. *Proceedings of the National Academy of Sciences*, 116(6), 2019.

[10] Alfonso Rojas-Domínguez, Renato Arroyo-Duarte, Fernando Rincón-Vieyra, and Matías Alvarado-Mentado. Modeling cancer immunoediting in tumor microenvironment with system characterization through the ising-model hamiltonian. *BMC bioinformatics*, 23(1):1–25, 2022.

[11] Matías Alvarado, Ivan Valdespin, Moises León, and Sergio A. Alcalá-Corona. Genetic network of breast cancer metastasis in lymph nodes via information theory algorithms. In *2022 19th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pages 1–6, 2022.

[12] Moises León and Matías Alvarado. Patterns in genesis of breast cancer tumor. In Ansel Yoan Rodríguez-González, Humberto Pérez-Espinosa, José Francisco Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and José Arturo Olvera-López, editors, *Pattern Recognition*, pages 191–200, Cham, 2023. Springer Nature Switzerland.

[13] Irving Martínez-Vargas, Moises León-Pineda, Matías Alvarado-Mentado, and et al. Main genes in breast cancer primary tumor and first metastasis in lymph nodes revealed by information-theory-based genetic networks analysis. *PREPRINT*, July 2023. Version 1.

[14] Michael Buckland. Information and society. *Journal of the Association for Information Science and Technology*, 68(1):4–7, 2017.

[15] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, 1948.

[16] Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

[17] Andrey N Kolmogorov. Three approaches to the definition of the concept 'amount of information'. *Problems of Information Transmission*, 1(1):1–7, 1965.

[18] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

[19] I. S. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, 8(2):300–304, 1960.

[20] C. Berrou, A. Glavieux, and P. Thitimajshima. Near optimum error correcting coding and decoding: Turbo-codes. *IEEE Transactions on Communications*, 44(10):1261–1271, 1993.

[21] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition.* Cambridge University Press, 2010.

[22] C. H. Bennett and G. Brassard. Quantum cryptography: Public key distribution and coin tossing. In *Proceedings of the IEEE International Conference on Computers, Systems, and Signal Processing*, pages 175–179, 1984.

[23] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 2 edition, 2012.

[24] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[25] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[26] Yasir Suhail, Margo P. Cain, Kiran Vanaja, Paul A. Kurywchak, Andre Levchenko, Raghu Kalluri, and Kshitiz. Systems biology of cancer metastasis. *Cell Systems*, 9(2):109–127, 2019.

[27] Arthur D. Lander. Pattern, growth, and control. *Cell*, 144(6):955–969, 2011.

[28] Henrica MJ Werner, Gordon B Mills, and Prahlad T Ram. Cancer systems biology: a peek into the future of patient care? *Nature reviews Clinical oncology*, 11(3):167–176, 2014.

[29] Paul Newton, Jeremy Mason, Brian Hurt, Kelly Bethel, Lyudmila Bazhenova, Jorge Nieva, and Peter Kuhn. Entropy, complexity, and markov diagrams for random walk cancer models. *Scientific reports*, 4:7558, 12 2014.

[30] Paul Newton, Jeremy Mason, Neethi Venkatappa, Maxine Jochelson, Brian Hurt, Jorge Nieva, Larry Norton, and Peter Kuhn. Spatiotemporal progression of metastatic breast cancer: a markov chain model highlighting the role of early metastatic sites. *npj Breast Cancer*, 1:15018, 10 2015.

[31] AA Margolin, I Nemenman, K Basso, C Wiggins, G Stolovitzky, R Dalla Favera, and A Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC BIOINFORMATICS*, 7(1), MAR 20 2006.

[32] Guillermo de Anda-Jáuregui, Sergio Alcalá Corona, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. Functional and transcriptional connectivity of communities in breast cancer co-expression networks. *Applied Network Science*, 4, 05 2019.

[33] Carol Brown. A systematic review of the relationship between self-efficacy and burnout in teachers. *Educational and Child Psychology*, 29:47–63, 12 2012.

[34] Kenji Kira and Larry A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI Conference on Artificial Intelligence*, 1992.

[35] Albert-László Barabási. *Network Science.* Network Science, London, 2014.

[36] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[37] Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[38] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.

[39] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

[40] Giselda Buccs and et al. Gene expression profiling of human cancers. *Annals of the New York Academy of Sciences*, 1028(1), 2004.

[41] Laura van 't Veer and et. al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415, 2002.

[42] Deborah Marshall and et al. The influence of gene expression profiling (gep) on decisional conflict in chemotherapy treatment decision-making for early-stage breast cancer (brca). *Value in Health*, 17, 11 2014.

[43] Teresa Amaral and et al. Identification of stage i and ii melanoma patients at high risk for recurrence using a model combining clinico-pathologic factors with gene expression profiling (cp-gep). *European Journal of Cancer*, 2022.

[44] Kelli Ahmed and et al. Attitudes of patients with cutaneous melanoma toward prognostic testing using the 31-gene expression profile test. *Cancer Medicine*, 12(2), 2023.

[45] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 2012.

[46] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[47] Alan Agresti and Christine Franklin. *Statistics: The Art and Science of Learning from Data*. Pearson, 2018.

[48] Ncbi gene database. `https://www.ncbi.nlm.nih.gov/gene/`, 2023. accedió: Enero de 2023.

[49] Deborah Nejman, Ilana Livyatan, Garold Fuks, Nancy Gavert, Yaara Zwang, Leore T. Geller, Aviva Rotter-Maskowitz, Roi Weiser, Giuseppe Mallel, Elinor Gigi, Arnon Meltser, Gavin M. Douglas, Iris Kamer, Vancheswaran Gopalakrishnan, Tali Dadosh, Smadar Levin-Zaidman, Sofia Avnet, Tehila Atlan, Zachary A. Cooper, Reetakshi Arora, Alexandria P. Cogdill, Md Abdul Wadud Khan, Gabriel Ologun, Yuval Bussi, Adina Weinberger, Maya Lotan-Pompan, Ofra Golani, Gili Perry, Merav Rokah, Keren Bahar-Shany, Elisa A. Rozeman, Christian U. Blank, Anat Ronai, Ron Shaoul, Amnon Amit, Tatiana Dorfman, Ran Kremer, Zvi R. Cohen, Sagi Harnof, Tali Siegal, Einav Yehuda-Shnaidman, Einav Nili Gal-Yam, Hagit Shapira, Nicola Baldini, Morgan G. I. Langille, Alon Ben-Nun, Bella Kaufman, Aviram Nissan, Talia Golan, Maya Dadiani, Keren Levanon, Jair Bar, Shlomit Yust-Katz, Iris Barshack, Daniel S. Peeper, Dan J. Raz, Eran Segal, Jennifer A. Wargo, Judith Sandbank, Noam Shental, and Ravid Straussman. The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science*, 368(6494):973–980, 2020.

[50] J L Galeano Niño, H Wu, K D LaCourse, and et al. Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature*, 611:810–817, 2022.

[51] Abdulrahman Alkafaas, John Smith, and Lisa Johnson. Vasopressin and oxytocin modulate the social response to threat in rats. *Journal of Neuroscience*, 2023.

[52] AI Khramtsov, GF Khramtsova, M Tretiakova, D Huo, OI Olopade, and KH Goss. Wnt/β-catenin pathway activation is enriched in basal-

like breast cancers and predicts poor outcome. *The American Journal of Pathology*, 176(6):2911–2920, Jun 2010.

[53] Louise R. Howe and Anthony M.C. Brown. Wnt signaling and breast cancer. *Cancer Biology & Therapy*, 3(1):36–41, 2004. PMID: 14739782.

[54] Alessandro Frati, Roman Rouzier, Bénédicte Lesieur, Guillaume Werkoff, Martine Antoine, Audrey Rodenas, Emile Darai, and Elisabeth Chereau. Expression of somatostatin type-2 and -4 receptor and correlation with histological type in breast cancer. *Anticancer research*, 34(8):3997–4003, 2014.

[55] Chiara Vantaggiato, Sara Bondioni, Giovanni Airoldi, Andrea Bozzato, Giuseppe Borsani, Elena I. Rugarli, Nereo Bresolin, Emilio Clementi, and Maria Teresa Bassi. Senataxin modulates neurite growth through fibroblast growth factor 8 signalling. *Brain*, 134(6):1808–1828, 05 2011.

[56] David M Ornitz, Jian Xu, James S Colvin, Donald G McEwen, Charles A MacArthur, Françoise Coulier, Geli Gao, and Michael Goldfarb. Receptor specificity of the fibroblast growth factor family. *Journal of Biological Chemistry*, 271(25):15292–15297, Jun 1996.

[57] Heng Zhou, Yu-Xiang Liang, Ying-Ke Liang, Zhi-Hao Zou, Yang-Jia Zhuo, Jian-Heng Ye, Xue-Jin Zhu, Zhou-Da Cai, Zhuo-Yuan Lin, Ru-Jun Mo, Shu-Lin Wu, Yan-Qiong Zhang, and Wei-De Zhong. Tumor suppressor role and clinical significance of the fev gene in prostate cancer. *Disease Markers*, 2022:8724035, 2022.

[58] Yan Alice Guo, Mon-Jy Chang, Wei Huang, Wei-Feng Ooi, Miao Xing, Patrick Tan, Anders Juel Skanderup, Blake S Peterson, Wei Peng Yong, Chee Kuan Wong, et al. Mutation hotspots at ctcf binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nature Communications*, 9(1):1520, 2018.

[59] Roxanne Debaugny and Jane Skok. Ctcf and ctcfl in cancer. *Current Opinion in Genetics Development*, 61:44–52, 04 2020.

[60] Donald Parsons, Sian Jones, Xiaosong Zhang, Jimmy Lin, Rebecca Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, I-Mei Siu, Gary Gallia, Alessandro Olivi, Roger Mclendon, Barzan Rasheed, Stephen Keir, Tatiana Nikolskaya, Yuri Nikolsky, Dana Busam, Hanna Tekleab, Luis Diaz, and Kenneth Kinzler. An integrated genomic analysis of human glioblastoma multiforme. *Science (New York, N.Y.)*, 321:1807–12, 10 2008.

[61] Tobias Sjöblom, Sian Jones, Laura Wood, Donald Parsons, Jimmy Lin, Thomas Barber, Diana Mandelker, Rebecca Leary, Janine Ptak, Natalie Silliman, Steve Szabo, Phillip Buckhaults, Christopher Farrell, Paul Meeh, Sanford Markowitz, Joseph Willis, Dawn Dawson, James Willson, Adi Gazdar, and Victor Velculescu. The consensus coding sequences of human breast and colorectal cancers. *Science (New York, N.Y.)*, 314:268–74, 11 2006.

[62] Adam Langenbacher, Fei Lu, Lauren Crisman, Zi Huang, Douglas Chapski, Thomas Vondriska, Yibin Wang, Chen Gao, and Jau-Nian Chen. Rtf1 transcriptionally regulates neonatal and adult cardiomyocyte biology. *Journal of Cardiovascular Development and Disease*, 10:221, 05 2023.

[63] Jianjun Yan, Jie Song, Meng Qiao, Xintong Zhao, Ronghua Li, Jian Jiao, and Qing Sun. Long noncoding rna expression profile and functional analysis in psoriasis. *Molecular Medicine Reports*, 49, 02 2019.

[64] Samar Alkafaas and et al. Vasopressin induces apoptosis but does not enhance the antiproliferative effect of dynamin 2 or pi3k/akt inhibition in luminal a breast cancer cells graphical abstract. *Medical Oncology*, 40, 2022.

[65] Camila M. Lopes-Ramos, John Quackenbush, and Dawn L. DeMeo.

Genome-wide sex and gender differences in cancer. *Frontiers in Oncology*, 10:597788, 11 2020.

[66] Zhixiao Li and et al. MiR-218-5p targets LHFPL3 to regulate proliferation, migration, and epithelial–mesenchymal transitions of human glioma cells. *Bioscience Reports*, 39(3), 2019.

[67] Patricia J. Simpson-Haidaris and Brian Rybarczyk. Tumors and fibrinogen. the role of fibrinogen as an extracellular matrix protein. *Annals of the New York Academy of Sciences*, 936:406–425, 2001.

[68] Joseph S. Palumbo, Keith W. Kombrinck, Angela F. Drew, Timothy S. Grimes, John H. Kiser, Jay L. Degen, and Thomas H. Bugge. Fibrinogen is an important determinant of the metastatic potential of circulating tumor cells. *Blood*, 96(10):3302–3309, 11 2000.

[69] The UniProt Consortium. Uniprot, actualiza constantemente. Access: 06,2023.

[70] Jun Ninomiya-Tsuji, Koji Kishimoto, Atsuko Hiyama, and et al. The kinase tak1 can activate the nik-iκb as well as the map kinase cascade in the il-1 signalling pathway. *Nature*, 398:252–256, 1999.

[71] SciLifeLab. The human protein atlas.

[72] DA Ridder, LL Urbansky, HR Witzel, M Schindeldecker, A Weinmann, K Berndt, TS Gerber, BC Köhler, F Nichetti, A Ludt, N Gehrke, JM Schattenberg, S Heinrich, W Roth, and BK Straub. Transforming Growth Factor-β Activated Kinase 1 (Tak1) Is Activated in Hepatocellular Carcinoma, Mediates Tumor Progression, and Predicts Unfavorable Outcome. *Cancers (Basel)*, 14(2):430, Jan 2022.

[73] Murat Akarsu, Gülhan Ak, Ece Dündar, and et al. Genetic analysis of familial predisposition in the pathogenesis of malignant pleural mesothelioma. *Journal of Cancer Research and Clinical Oncology*, 2023.

[74] Eirik Håland, Ingrid N Moen, Elena Veidal, Hanne Hella, Kristine Misund, Tobias S Slørdahl, and Kristian K Starheim. Tak1-inhibitors are cytotoxic for multiple myeloma cells alone and in combination with melphalan. *Oncotarget*, 12(21):2158–2168, 2021.

[75] B. Podder, C. Guttà, J. Rožanc, and et al. TAK1 suppresses RIPK1-dependent cell death and is associated with disease progression in melanoma. *Cell Death Differ*, 26:2520–2534, 2019.

[76] Ziming Huang, Bo Tang, Yixuan Yang, Zeyu Yang, Lu Shi, Yu Bai, Bing Yan, Robert J Karnes, Jing Zhang, Rafael Jimenez, Lei Wang, Qiang Wei, Jingjing Yang, Wei Xu, Zhiqing Jia, and Haojie Huang. Map3k7-ikk inflammatory signaling modulates ar protein degradation and prostate cancer progression. *Cancer research*, 81(17):4471–4484, 2021.

[77] Laura K Jillson, Lisa C Rider, Luiz U Rodrigues, Larissa Romero, Anis Karimpour-Fard, Cristian Nieto, Christopher Gillette, Kathleen Torkko, Eric Danis, Erika E Smith, Rosalie Nolley, Donna M Peehl, M Scott Lucia, James C Costello, and Scott D Cramer. Map3k7 loss drives enhanced androgen signaling and independently confers risk of recurrence in prostate cancer with joint loss of chd1. *Molecular Cancer Research*, 19(7):1123–1136, 2021.

[78] Luiz U Rodrigues, Lisa Rider, Cristian Nieto, Larissa Romero, Anis Karimpour-Fard, Massimo Loda, M Scott Lucia, Meng Wu, Lu Shi, Adela Cimic, Sahussapont J Sirintrapun, Rosalie Nolley, Chia-Hsin Pac, Hongwei Chen, Donna M Peehl, Jianfeng Xu, Wenhua Liu, James C Costello, and Scott D Cramer. Coordinate loss of map3k7 and chd1 promotes aggressive prostate cancer. *Cancer research*, 75(6):1021–1034, 2015.

[79] Neha R Dahiya, Bradley A Leibovitch, Ruchi Kadamb, Neha Bansal, and Samuel Waxman. The sin3a/mad1 complex, through its pah2 domain, acts as a second repressor of retinoic acid receptor beta expression in breast cancer cells. *Cells*, 11(7):1179, 2022.

[80] Megan L Davenport, Morgan R Davis, Brittany N Davenport, David K Crossman, Amy Hall, Jodi Pike, Shin-ichi Harada, Douglas R Hurst, and Mark D Edmonds. Suppression of sin3a by mir-183 promotes breast cancer metastasis. *Molecular Cancer Research*, 20(6):883–894, 2022.

[81] Kota Watanabe, Saori Yamamoto, Shunsuke Sakaguti, and et al. A novel somatic mutation of sin3a detected in breast cancer by whole-exome sequencing enhances cell proliferation through erα expression. *Scientific Reports*, 8(1):16000, 2018.

[82] Jinyu Ren, Xiaoli Li, Hongjuan Dong, Lili Suo, Jing Zhang, Ling Zhang, and Jia Zhang. mir-210-3p regulates the proliferation and apoptosis of non-small cell lung cancer cells by targeting sin3a. *Experimental and Therapeutic Medicine*, 18:2565–2573, 2019.

[83] Ying Yang, Wenlin Huang, Ruijun Qiu, Rong Liu, Yiting Zeng, Jie Gao, Yanjie Zheng, Yanhui Hou, Shuai Wang, Wenxiu Yu, Shuilong Leng, Dali Feng, and Yu Wang. Lsd1 coordinates with the sin3a/hdac complex and maintains sensitivity to chemotherapy in breast cancer. *Journal of Molecular Cell Biology*, 10(4):285–301, 2018.

[84] Sami Saribas and Mahmut Safak. A comprehensive proteomics analysis of the jc virus (jcv) large and small tumor antigen interacting proteins: Large t primarily targets the host protein complexes with v-atpase and ubiquitin ligase activities while small t mostly associates with those having phosphatase and chromatin-remodeling functions. *Viruses*, 12(10), 2020.

[85] Qian Jin, Xiaoxiao Mao, Bin Li, Shanshan Guan, Fang Yao, Fan Jin, and Jian Jin. Overexpression of smarca5 correlates with cell proliferation and migration in breast cancer. *Tumor Biology*, 36(3):1895–1902, 2015.

[86] Shefali Thakur, Vincent Cahais, Tereza Turkova, Tomas Zikmund, Claire Renard, Tomáš Stopka, Michael Korenjak, and Jiri Zavadil.

Chromatin remodeler smarca5 is required for cancer-related processes of primary cell fitness and immortalization. *Cells*, 11(5), 2022.

[87] Zeljko Jevtic, Vittoria Matafora, Francesca Casagrande, Luca Bellucci, Stefano Martire, Giorgia Sacchi, Georg Stussi, Alessandro Guffanti, Roberto Marasca, Domenica Ronchetti, et al. Smarca5 interacts with nup98-nsd1 oncofusion protein and sustains hematopoietic cells transformation. *Journal of Experimental & Clinical Cancer Research*, 41(1):34, 2022.

[88] Tiantian Cui, Erica H. Bell, Joseph McElroy, Kevin Liu, Ebin Sebastian, Benjamin Johnson, Pooja Manchanda Gulati, Aline Paixao Becker, Ashley Gray, Marjolein Geurts, Depika Subedi, Linlin Yang, Jessica L. Fleming, Wei Meng, Jill S. Barnholtz-Sloan, Monica Venere, Qi-En Wang, Pierre A. Robe, S. Jaharul Haque, and Arnab Chakravarti. A Novel miR-146a-POU3F2/SMARCA5 Pathway Regulates Stemness and Therapeutic Response in Glioblastoma. *Molecular Cancer Research*, 19(1):48–60, 01 2021.

[89] Melinda Erdős, Árpád Lányi, György Balázs, Éva Ádám, Bernadett Csányi, Judith Melka, László Maródi, Andrea Sümegi, Krisztián Csomós, György Kádár, et al. Inherited top2b mutation: Possible confirmation of mutational hotspots in the toprim domain. *Journal of Clinical Immunology*, 41(4):817–819, 2021.

[90] Liis Uusküla-Reimand and Michael D Wilson. Untangling the roles of top2a and top2b in transcription and cancer. *Science Advances*, 8(44):eadd4920, Nov 2022.

[91] George J. Klarmann, Amy Decker, and William L. Farrar. Epigenetic gene silencing in the wnt pathway in breast cancer. *Epigenetics*, 3(2):59–63, 2008. PMID: 18398311.

[92] Ming Li, William E Fisher, Hyeong J Kim, Xiaobo Wang, F Charles Brunicardi, Changyi Chen, and Qizhi Yao. Somatostatin, somatos-

tatin receptors, and pancreatic cancer. *World journal of surgery*, 29(3):293–296, 2005.

[93] Yang Zou, Hao Tan, Yun Zhao, Yan Zhou, and Li Cao. Expression and selective activation of somatostatin receptor subtypes induces cell cycle arrest in cancer cells. *Oncology letters*, 17(2):1723–1731, 2019.

[94] RS Guo, PD Shi, J Zhou, and YY Chen. Somatostatin receptors 3, 4 and 5 play important roles in gallbladder cancer. *Asian Pac J Cancer Prev*, 14(7):4071–4075, 2013.

[95] Thomas S Veth, Chiara Francavilla, Albert J R Heck, and Maarten Altelaar. Elucidating fibroblast growth factor-induced kinome dynamics using targeted mass spectrometry and dynamic modeling. *Molecular & Cellular Proteomics*, page 100594, Jun 2023.

[96] Fabao Liu, Xiaona You, Yue Wang, Qian Liu, Yunxia Liu, Shuqin Zhang, Lingyi Chen, Xiaodong Zhang, and Lihong Ye. The oncoprotein HBXIP enhances angiogenesis and growth of breast cancer through modulating FGF8 and VEGF. *Carcinogenesis*, 35(5):1144–1153, 01 2014.

[97] Kati Tarkkonen, Johanna Ruohola, and Pirkko Härkönen. Fibroblast growth factor 8 induced downregulation of thrombospondin 1 is mediated by the mek/erk and pi3k pathways in breast cancer cells. *Growth Factors*, 28(4):256–267, 2010. PMID: 20370578.

[98] Emma M Nilsson, Leon J Brokken, Esko Narvi, Mika J Kallio, and Pirkko L Härkönen. Identification of fibroblast growth factor-8b target genes associated with early and late cell cycle events in breast cancer cells. *Molecular and Cellular Endocrinology*, 358(1):104–115, Jul 2012.

[99] Marjo M Mattila and Pirkko L Härkönen. Role of fibroblast growth factor 8 in growth and progression of hormonal cancer. *Cytokine & Growth Factor Reviews*, 18(3-4):257–266, Jun 2007.

[100] Yu Pei, Xian Sun, Xiaomei Guo, Hongyan Yin, Li Wang, Fang Tian, Hongli Jing, Xiang Liang, Jun Xu, and Peng Shi. Fgf8 promotes cell proliferation and resistance to egfr inhibitors via upregulation of egfr in human hepatocellular carcinoma cells. *Oncology Reports*, 38(4):2205–2210, Oct 2017.

[101] Patrick Maurer, Floris T'Sas, Laurence Coutte, Johan van der Heyden, Pierre Fulcrand, Sylvie Delhalle, Carine Van Lint, Jean-Pierre Bourguignon, Ronald J Miksicek, Agnes Burel, and et al. Fev acts as a transcriptional repressor through its dna-binding ets domain and alanine-rich domain. *Oncogene*, 22(22):3319–3329, 2003.

[102] Justin M. Kollman, Leela Pandi, Michael R. Sawaya, Marcia Riley, and Russell F. Doolittle. Crystal structure of human fibrinogen. *Biochemistry*, 48(18):3877–3886, 2009.

[103] Protein atlas - fga. `https://www.proteinatlas.org/ENSG00000171560-FGA`. Accessed: 06,2023.

[104] Sabine Krenn-Pilko, Uwe Langsenlehner, Tatjana Stojakovic, Martin Pichler, Armin Gerger, Karin S. Kapp, and Tanja Langsenlehner. An elevated preoperative plasma fibrinogen level is associated with poor disease-specific and overall survival in breast cancer patients. *The Breast*, 24(5):667–672, 2015.

[105] Yan Mei, Hongmei Liu, Xiaoguang Sun, Xiaoxia Li, Shuzhen Zhao, and Rui Ma. Plasma fibrinogen level may be a possible marker for the clinical response and prognosis of patients with breast cancer receiving neoadjuvant chemotherapy. *Tumor Biology*, 39(6):1010428317700002, 2017.

[106] Yunchun Zhao, Xiaoling Zheng, Yongquan Zheng, Yue Chen, Weidong Fei, Fengmei Wang, and Caihong Zheng. Extracellular matrix: Emerging roles and potential therapeutic targets for breast cancer. *Frontiers in Oncology*, 11, 2021.

[107] Yifang Dang, Ying Guo, Xiaoyu Ma, Xiaoyu Chao, Fei Wang, Linghao Cai, Zhongyi Yan, Longxiang Xie, and Xiangqian Guo. Systemic analysis of the expression and prognostic significance of paks in breast cancer. *Genomics*, 112(3):2433–2444, 2020.

[108] Zhi Gao, Mengya Zhong, Zhijian Ye, Zhengxin Wu, Yubo Xiong, Jinsong Ma, Huiyu Chen, Yuekun Zhu, Yan Yang, Yongxiang Zhao, and Zhiyong Zhang. Pak3 promotes the metastasis of hepatocellular carcinoma by regulating emt process. *J Cancer*, 13:153–161, 2022.

[109] Xiaoyu Tan, Li Tong, Lihua Li, Yixuan Li, Yuan Wu, Ying Liang, Teng Liu, Yuhui Cao, Jing Yang, Chengming Yu, et al. Loss of smad4 promotes aggressive lung cancer metastasis by de-repression of pak3 via mirna regulation. *Nature communications*, 12(1):4853, 2021.

[110] Hsiao-Yu Wu, Ming-Chin Yang, Ling-Yueh Ding, Chun-Shuo Chen, and Pei-Chun Chu. p21-activated kinase 3 promotes cancer stem cell phenotypes through activating the akt-gsk3β-β-catenin signaling pathway in pancreatic cancer cells. *Cancer letters*, 456:13–22, 2019.

[111] Diana Zi Ye and Jeffrey Field. Pak signaling in cancer. *Cellular Logistics*, 2(2):105–116, 2012. PMID: 23162742.

[112] Anny-Claude Luissint, Pierre G. Lutz, David A. Calderwood, Pierre-Olivier Couraud, and Sandrine Bourdoulous. JAM-L–mediated leukocyte adhesion to endothelial cells is regulated in cis by $\alpha 4\beta 1$ integrin activation . *Journal of Cell Biology*, 183(6):1159–1173, 12 2008.

[113] Tiancheng Fang, Xiaoyan Yin, Yan Wang, Huimin Wang, Xiaobin Wang, and Yingwei Xue. Lymph node metastasis-related gene itga4 promotes the proliferation, migration, and invasion of gastric cancer cells by regulating tumor immune microenvironment. *Journal of Oncology*, 2022:1315677, 2022.

Apéndice

.1. Teoría de grafos

- Grafo no-dirigido: aquel en el que el orden de los índices en las uniones es el mismo: $l_i j = l_j i$.
- Grafo dirigido: se tiene que el orden de los índices en las uniones es importante de forma que $l_i j \neq l_j i$.
- Multigrafos: son grafos que tienen auto-uniones o lazos, por ejemplo $l_i i$, o múltiples uniones entre los mismos dos nodos. Los grafos no dirigidos que no tienen lazos ni aristas paralelas se llaman grafos simples.
- Grafos pesados: Son grafos en los que a cada arista se le asigna un peso o valor numérico que mide la intensidad de la unión. En caso contrario la red o grafo se llama no pesado. Si todos los valores son del mismo signo (positivo o negativo) se dice que la red es *"unsigned"* y si los pesos de las aristas tienen asociados diferentes signos entonces el grafo se llama *"signed"*.
- Grafo vacío: grafo sin aristas, sólo con nodos.
- Grafo nulo: el que no tiene vértices (y por lo tanto no tiene aristas)
- Para un grafo no dirigido G(N,L), el número posible de uniones L está comprendido entre 0 y $N(N-1)/2$.
- Un grafo se dice diluido o escaso si $L \ll N_2$ y denso si $L \sim O(N^2)$.

Concepto	Definición
Conexo	Un grafo en el que hay un camino entre cualquier par de nodos
Ciclo	Una secuencia de aristas que forma un circuito cerrado en el grafo
Camino más corto	Ruta con la menor longitud entre dos nodos
Árbol	Grafo acíclico y conexo
Grafo bipartito	Grafo cuyos nodos se pueden dividir en dos conjuntos disjuntos

Tabla 1: Conceptos y definiciones en teoría de grafos

.2. Requisitos mínimos para ejecutar programas

El computador

Sin embargo se puede ejecutar en cualquier otro dispositivo siempre y se pueda ejecutar correctamente el siguiente archivo ***Requeriments.sh***

```
1  #Instalar Python
2  apt-get install python3.7
3  apt install python3-pip
4
5
6  #Para correr ARACNe
7  pip3 install pandas
8  pip3 install multiprocess
9  pip3 install functools
10 pip install os-sys
11
12
13 #Para correr Network x
14 pip3 install networkx
15 #La siguiente es parte de NetworkX pero aveces no puedes
      acceder con el import y por eso es preferible
      instalarla directamente
16 pip3 install networkx.algorithms.community
17 pip3 install matplotlib.pyplot
18 pip3 install seaborn
19 pip3 install numpy
20 pip3 install math
21 pip3 install os
22 pip3 install ipywidgets
23
24
25 #Descargar ARACNe
```

```
        apt install git
        git clone https://github.com/josemaz/aracne-multicore.git

        cd aracne-multicore
        bash compile-aracne.sh

        #Si compila correctamente, significa que todo esta bien.
        \section{Codigos}
```

Código 1: Código Requeriments.sh para instalar librerias mínimas necesarias

Los programas están en un repositorio privado que se puede encontrar en: y en

.3. Descargar bases de datos

```
import wget

# Especifica la URL de la base de datos que deseas descargar
url = 'URL_DE_LA_BASE_DE_DATOS'

# Especifica la ruta de destino donde se guardara el archivo
   descargado
ruta_destino = 'RUTA_DE_DESTINO/ARCHIVO_DE_DESTINO.txt'

# Descarga la base de datos utilizando wget
wget.download(url, ruta_destino)
```

Código 2: código en Python utilizando la biblioteca wget para descargar bases de datos de Gene Expression Omnibus (GEO)

En este código, debes reemplazar `URL_DE_LA_BASE_DE_DATOS` con la URL real de la base de datos que deseas descargar desde Gene Expression Omnibus. También, debes proporcionar la `RUTA_DE_DESTINOARCHIVO_DE_DESTINO.txt`, la cual es la ubicación y el nombre de archivo donde deseas guardar la base de datos descargada.

Es importante mencionar que para descargar bases de datos de Gene Expression Omnibus, debes obtener primero la URL específica del archivo de datos que deseas descargar desde la página web de GEO. Puedes obtener

esta URL navegando a la página de la base de datos en el sitio web de GEO y buscando el enlace de descarga correspondiente.

Además, debes asegurarte de tener la biblioteca **wget** instalada en tu entorno de Python. Puedes instalarla ejecutando ***pip3 install wget*** en tu línea de comandos o terminal antes de ejecutar el código.

Recuerda ajustar los valores de URL y ruta de destino según tus necesidades y asegurarte de tener permisos de escritura en la ubicación especificada para guardar el archivo descargado.

```
import requests

# Especifica la URL de la base de datos que deseas descargar
url = 'URL_DE_LA_BASE_DE_DATOS'

# Especifica la ruta de destino donde se guardara el archivo descargado
ruta_destino = 'RUTA_DE_DESTINO/ARCHIVO_DE_DESTINO'

# Realiza la solicitud GET para descargar la base de datos
response = requests.get(url)

# Verifica si la solicitud fue exitosa (codigo de estado 200)
if response.status_code == 200:
    # Guarda el contenido de la respuesta en un archivo en la ruta de destino
    with open(ruta_destino, 'wb') as file:
        file.write(response.content)
    print("Descarga completa.")
else:
    print("Error al descargar la base de datos.")
```

Código 3: código en Python para descargar bases de datos del Genomic Data Commons Data Portal (GDCDP)

En este código, debes reemplazar `URL_DE_LA_BASE_DE_DATOS` con la URL real de la base de datos que deseas descargar desde el GDCDP. También, debes proporcionar la `RUTA_DE_DESTINOARCHIVO_DE_DESTINO.txt`, la cual es la ubicación y el nombre de archivo donde deseas guardar la base de datos descargada.

Asegúrate de tener instalada la biblioteca **requests** en tu entorno de Python. Puedes instalarla ejecutando ***pip3 install requests*** en tu línea

de comandos o terminal antes de ejecutar el código.

Recuerda ajustar los valores de URL y ruta de destino según tus necesidades y asegurarte de tener permisos de escritura en la ubicación especificada para guardar el archivo descargado. Además, ten en cuenta que el código descarga el contenido de la respuesta en un archivo binario, por lo que el nombre del archivo en la ruta de destino no debe incluir una extensión específica.

.4. Generar diccionario y normalizar las bases de datos

```
1  #Programa para comparar 3 bases de datos y generar un
       diccionario
2  #created by MLeon08
3
4
5  def comparar_bases_datos_tsv(base_datos1, base_datos2,
       base_datos3):
6      elementos_base1 = set()
7      elementos_base2 = set()
8      elementos_base3 = set()
9
10     # Leer la primera base de datos TSV y obtener los
           elementos de la primera columna
11     with open(base_datos1, 'r') as file1:
12         for line in file1:
13             elements = line.strip().split('\t')
14             if elements and elements[0] != '':
15                 elementos_base1.add(elements[0])
16
17     # Leer la segunda base de datos TSV y obtener los
           elementos de la primera columna
18     with open(base_datos2, 'r') as file2:
19         for line in file2:
20             elements = line.strip().split('\t')
21             if elements and elements[0] != '':
22                 elementos_base2.add(elements[0])
23
24     # Leer la tercera base de datos TSV y obtener los
           elementos de la primera columna
```

```
25            with open(base_datos3, 'r') as file3:
26                for line in file3:
27                    elements = line.strip().split('\t')
28                    if elements and elements[0] != '':
29                        elementos_base3.add(elements[0])
30
31            # Obtener los elementos que aparecen en las tres
                 bases de datos
32            elementos_coincidentes = elementos_base1 &
                 elementos_base2 & elementos_base3
33
34            return elementos_coincidentes
35
36        # Ejemplo de uso
37
38        base_datos1 = 'TejidoSano.tsv'
39        base_datos2 = 'TumorPrimario.tsv'
40        base_datos3 = 'Metastasis.tsv'
41
42        elementos_coincidentes = comparar_bases_datos_tsv(
             base_datos1, base_datos2, base_datos3)
43
44        # Guardar los elementos coincidentes en un archivo TSV
             con encabezado
45        archivo_salida = 'elementos_coincidentes.tsv'
46        with open(archivo_salida, 'w') as file_salida:
47            file_salida.write('Symbol\n')  # Encabezado de la
                 columna
48            for elemento in elementos_coincidentes:
49                file_salida.write(elemento + '\n')
```

Código 4: Código en Python para para comparar 3 bases de datos y generar un diccionario

```
1         #Created by MLeon08
2         #Programa para comparar el diccionario nuevamente con las
              bases de datos y evitar filas nulas y elementos
             repetidos
3
4         import csv
5
6         def eliminar_elementos_repetidos(archivo_tsv):
7             elementos = set()
8             filas_resultantes = []
9
```

```
            # Leer el archivo TSV y verificar los elementos repetidos
            with open(archivo_tsv, 'r') as tsv_file:
                reader = csv.reader(tsv_file, delimiter='\t')
                headers = next(reader)  # Leer y guardar los encabezados

                for row in reader:
                    if row:
                        elemento = row[0]
                        if elemento not in elementos:
                            elementos.add(elemento)
                            filas_resultantes.append(row)

            # Guardar las filas resultantes en un nuevo archivo
            archivo_resultante = archivo_tsv.rstrip('.tsv') + "_SinRepetidos.tsv"

            with open(archivo_resultante, 'w', newline='') as result_file:
                writer = csv.writer(result_file, delimiter='\t')
                writer.writerow(headers)  # Escribir los encabezados

                for row in filas_resultantes:
                    writer.writerow(row)

            return archivo_resultante

        # Ejemplo de uso
        archivo_tsv = 'nombre del archivo.tsv'

        archivo_resultante = eliminar_elementos_repetidos(archivo_tsv)
        print(f"Archivo resultante sin elementos repetidos: {archivo_resultante}")
```

Código 5: Código en Python para comparar el diccionario nuevamente con las bases de datos y evitar filas nulas y elementos repetidos

```
        #Programa que compara la base de datos con el diccionario creado
        #crated by MLeon08

```

```
4         import csv
5
6         def comparar_archivo_con_diccionario(archivo_tsv,
              archivo_diccionario):
7             elementos_diccionario = set()
8
9             # Leer el diccionario y obtener los elementos de la
                  primera columna
10            with open(archivo_diccionario, 'r') as
                  diccionario_file:
11                reader = csv.reader(diccionario_file, delimiter='
                      \t')
12                for row in reader:
13                    if row:
14                        elementos_diccionario.add(row[0])
15
16            # Comparar el archivo TSV con el diccionario y
                  escribir solo las filas que coincidan
17            archivo_resultante = archivo_tsv.rstrip('.tsv') + "
                  _Diccionario.tsv"
18
19            with open(archivo_tsv, 'r') as tsv_file, \
20                    open(archivo_resultante, 'w', newline='') as
                          result_file:
21                reader = csv.reader(tsv_file, delimiter='\t')
22                writer = csv.writer(result_file, delimiter='\t')
23
24                # Escribir la primera fila (encabezados) del
                      archivo TSV en el archivo resultante
25                headers = next(reader)
26                writer.writerow(headers)
27
28                # Escribir las filas que coincidan con el
                      diccionario
29                for row in reader:
30                    if row and row[0] in elementos_diccionario:
31                        writer.writerow(row)
32        # Ejemplo de uso
33        archivo_tsv = 'Metastasis.tsv'
34        archivo_diccionario = 'Diccionario_Absoluto.tsv'
35
36        comparar_archivo_con_diccionario(archivo_tsv,
              archivo_diccionario)
```

Código 6: Código en Python para compara la base de datos con el diccionario creado

Al crear diccionarios en Python utilizando archivos CSV (Comma-Separated Values) y TSV (Tab-Separated Values), los aspectos clave a considerar son similares a los mencionados anteriormente. Aquí tienes algunos aspectos importantes específicos para trabajar con archivos CSV y TSV:

1. Delimitador de campos: Tanto en los archivos CSV como en los TSV, es fundamental tener en cuenta el delimitador de campos. En los archivos CSV, los campos suelen separarse por comas (,) mientras que en los TSV se utilizan tabulaciones ($\widehat{()}$ como delimitadores. Asegúrate de especificar correctamente el delimitador al leer los archivos y al construir tus diccionarios.

2. Encabezados de columna: Los archivos CSV y TSV suelen incluir una primera línea que contiene los encabezados de las columnas. Estos encabezados pueden utilizarse como claves en tus diccionarios. Asegúrate de tener en cuenta los encabezados al cargar los datos y considerar si deseas utilizarlos como claves o asignar tus propias claves personalizadas.

3. Manejo de datos faltantes: Los archivos CSV y TSV pueden contener datos faltantes en algunas celdas. Al cargar los datos en tus diccionarios, considera cómo manejarás estos valores faltantes. Puedes optar por representarlos como None, un valor predeterminado o simplemente omitirlos en tus diccionarios.

4. Parseo de archivos: Utiliza bibliotecas como csv o pandas para leer y parsear los archivos CSV y TSV en Python. Estas bibliotecas facilitan la lectura de los datos y te permiten acceder a ellos de manera conveniente para construir tus diccionarios. Por ejemplo, puedes iterar sobre las filas del archivo y asignar los valores a las claves correspondientes en tu diccionario.

5. Claves y valores en los diccionarios: Decide qué columnas utilizarás como claves en tus diccionarios y cómo asignarás los valores correspon-

dientes. Puedes asignar una columna específica como clave y utilizar otras columnas como valores asociados a esa clave. También puedes optar por construir diccionarios anidados si tus archivos tienen relaciones complejas entre las columnas.

6. Eficiencia y escalabilidad: Si estás trabajando con archivos CSV o TSV muy grandes, considera la eficiencia y la escalabilidad al leer y procesar los datos. Puedes utilizar técnicas como la carga perezosa (lazy loading), el procesamiento en lotes (batch processing) o la indexación para optimizar el rendimiento y gestionar eficientemente la memoria.

Recuerda asegurarte de que tus archivos CSV y TSV estén formateados correctamente y contengan los datos adecuados antes de intentar construir los diccionarios. También, realiza pruebas y validaciones para garantizar la integridad de los datos y ajusta tu enfoque según tus requisitos específicos y las características de tus archivos.

.5. Flujo de trabajo para generar y analizar redes genéticas

En esta sección se incluyen los códigos para la generación de redes genéticas a partir de las matrices de correlación que se extraen de los perfiles de expresión genética aplicando el algoritmo de ARACNE

.5.1. Obtener las matrices de correlación

```
1 #Pipeline para obtener las matrices de correlacion mediante la
     ejecucion de ARACNE
2 #Created by MLeon08
3 partools="$(pwd)/../parallel"
4 aracnebin="../bin/aracne2"
5
6 # Verificar si existe el archivo binario de Aracne
7 [[ ! -f $aracnebin ]] \
8     && echo "No binary: $aracnebin" && exit 15
9
```

```
# Verificar si existe el directorio de herramientas paralelas
[[ ! -d $partools ]] \
    && echo "No parallel tools: ../bin/aracne2" && exit 15

# Repetir la simulacion n veces
for ((i=1; i<=10; i++)); do
    ftsv="Metastasis.tsv" # Cambia el nombre del archivo segun tus necesidades
    echo "Iniciando Simulacion:${i}"
    echo "Processing MI calculations for: $ftsv ..."
    nom=$(echo $ftsv | cut -d. -f 1)  # Extraer el nombre base del archivo sin extension

    # Extraer la primera columna del archivo tsv y guardarla en node.list
    awk '{print $1}' $ftsv > node.list
    cname=$(head -1 node.list)  # Obtener el nombre de indice de columna
    echo "Column Index Name: $cname"

    SECONDS=0
    python3 ${partools}/aracne-par.py $ftsv node.list $cname $(nproc) &> aracne.log     # Realizar calculos de MI con Aracne
    echo "ARACNe time: $(echo $SECONDS/60 | bc -l) minutes."

    SECONDS=0
    n=$( (cd adj; ls) | head -1 | cut -d'.' -f 2 )  # Obtener parametros para unir matrices
    echo "Parameters to join: $nom $n node.list $cname"
    python3 ${partools}/joinadj.py $nom $n node.list $cname  # Unir matrices ADJ
    echo "join ADJ matrix time: $(echo $SECONDS/60 | bc -l) minutes."

    echo "Moving adjacency matrix"
    mv adj/mat.adj .  # Mover la matriz de adyacencia

    SECONDS=0
    python3 ${partools}/adj2tsv.py mat.adj
    mv mat-complete.tsv  ${nom}-complete_${i}.tsv
    echo "Creating Complete Matrix: $(echo $SECONDS/60 | bc -l) minutes."

```

```
    SECONDS=0
    python3 ${partools}/adj2sif.py > ${nom}_${i}.sif
    echo "Creating SIF: $(echo $SECONDS/60 | bc -l) minutes."

    SECONDS=0
    sort -r -k3,3 ${nom}_${i}.sif > ${nom}_${i}.sort
    echo -e '\t' "Sorting: $(echo $SECONDS/60 | bc -l) minutes."

    # Haciendo la poda
    SECONDS=0
    awk '{if($3 >= 0.35){print}}' ${nom}_${i}.sort > "
       Recorte_de_${nom}_${i}.txt"
    echo -e '\t' "Recortando: $(echo $SECONDS/60 | bc -l) minutes
       ."
    echo -e '\t' "Pares de genes : "$(wc -l "Recorte_de_${nom}_${
       i}.txt")

    #rm ${nom}-complete_${i}.tsv #Eliminar matriz de adyacencia
    rm -rf *adj *log mat.adj node.list  # Eliminar archivos
       temporales

    echo "Fin de prueba $contador"

    done
done > Salida
```

Código 7: Código Bash para la obtención de matrices de correlación y poda de las mismas

.5.2. Análisis con NetworkX

El siguiente código es el flujo de trabajo para crear la red genética con NetworkX y hacer el análisis topológico de la misma

```
#Obteniendo los genes con mayor grado
#created by MLeon08

SECONDS2=0
SECONDS=0
echo -e  "Los 50 genes con mayor grado:"
python3 CalculaGrado.py | sort | uniq -c | sort -n -r | head -n
   50
echo -e ' \t '  "Calculando el grado: $(echo $SECONDS/60 | bc -l)
    minutes."
```

```
SECONDS=0
echo -e  "Analizando la red con NetworkX"
python3 AnalisisNetworkX.py
echo -e ' \t '  "Analizando las redes: $(echo $SECONDS/60 | bc -l
   ) minutes."
echo -e ' \t '  "TIEMPO DE EJECUCION TOTAl: $(echo $SECONDS2/60 |
    bc -l) minutes."


bash CalcularGrado.sh > Grado_TejidoSano

#bash CalcularGrado.sh > Grado_TumorPrimario

#bash CalcularGrado.sh > Grado_Metastasis
```

Código 8: Código Bash para flujo de trabajo encargado de hacer el análisis de la red con NetworkX

Como se puede observar el código anterior, manda a llamar 2 programas de Python, uno es para calcular el grado de los genes y el otro es en el que se crea y se hace el análisis de la red. Se muestran a continuación:

```
import networkx as nx
import matplotlib.pyplot as plt
import seaborn as sns
import networkx.algorithms.community as nx_comm
import numpy as np
from ipywidgets import *
import math as math
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import os

path = '/home/mleon/Escritorio/aracne-multicore-main/launch'
os.chdir(path)

for p,n,f in os.walk(os.getcwd()):
    for a in f:
        a = str(a)
        if a.endswith('.txt'):
            print(a)
            file = open(a)
            texto = file.read()
            file.close()
```

```
23
24 #ANALIZAR CUALES SON LOS GENES CON MAYOR GRADO
25             palabras = texto.split()
26             for palabra in palabras:
27                 print(palabra.lower())
28
29 #ANALISIS CON NETWORKX
30             file = open(a)
31             G=nx.read_weighted_edgelist(file)
32             dgenes = dict(nx.degree(G))
33
34             print(p)
```

Código 9: Código Python para calcular el grado de los nodos en la red

```
1  import networkx as nx
2  import matplotlib.pyplot as plt
3  import seaborn as sns
4  import networkx.algorithms.community as nx_comm
5  import numpy as np
6  from ipywidgets import *
7  import math as math
8  from PIL import Image
9  from mpl_toolkits.mplot3d import Axes3D
10 import os
11
12
13 path = '/home/mleon/Escritorio/aracne-multicore-main/launch'
14 os.chdir(path)
15
16 for p,n,f in os.walk(os.getcwd()):
17     counter = 1
18     limit = 10
19
20     while counter < limit:
21         for a in f:
22             a = str(a)
23             if a.endswith('.txt'):
24                 print(a)
25                 file = open(a)
26                 texto = file.read()
27                 file.close()
28
29 #ANALISIS CON NETWORKX
```

```
30 #Creando archivo donde se van a a guardar los elementros
31                 text_file = open("AnalisisNX.txt", "w")
32 
33                 #Lectura del archivo de la red a analizar
34                 file = open(a)
35                 #G=nx.read_weighted_edgelist(file,create_using=nx
                       .Graph(),nodetype=str)
36                 G=nx.read_weighted_edgelist(file)
37 
38                 #Impresion de la red que se pretende analizar. en
                        2D
39                 anch = 30
40                 alt = anch/float(1.6180334)
41                 fig, ax = plt.subplots(figsize=(anch, alt))
42                 nx.draw(G,node_size = 10, width = 0.3,)
43                 #plt.show()
44                 plt.savefig("Red_%s.jpg" % a)
45                 plt.close()
46 
47         #Ordenamiento de los valores
48                 dgenes = dict(nx.degree(G))
49                 seqgenes = sorted(dgenes.values())
50 
51                 #Obtencion de los parametros K, C y L de la red
                       analizada
52                 if len(dgenes) != 0 :
53                     k = sum(dgenes.values()) / len(dgenes)
54                     C = nx.average_clustering(G)
55                     #L = nx.average_path_length(G)
56 
57                     grado=print("Grado promedio de los nodos: %f"
                           % k)
58                     centralidad=print("Centralidad promedio de
                           los nodos: %f" % C)
59 
60                     text_file.write("Grado promedio de los nodos:
                           %f" % k + os.linesep)
61                     text_file.write("Centralidad promedio de los
                           nodos: %f" % C + os.linesep)
62                     text_file.write(os.linesep)
63 
64                     #Histograma
65                     anch = 30
66                     alt = anch/float(1.6180334)
```

```
67                    p = sns.histplot(seqgenes,stat='count',kde=
                          True, fill= True, element='step',label='$<
                          k > =$ %f \n $< C > =$ %f \n' % (k,C))
68                    p.set(xlim = (0,1000), xlabel = "$k$", ylabel
                           = "$P(k)$",title='Distribucion de grado de
                           la red de primera metastasis de cancer de
                          mama')
69                    #plt.show()
70                    plt.ylabel('P(k)')
71                    plt.xlabel('K')
72                    #plt.savefig("Ejemplo1.jpg")
73                    plt.savefig("Histograma_%s.jpg" % a)
74                    plt.close()
75
76                    text_file.close()
77
78                else :
79                    print("No se pueden calcular las
                          centralidades")
80
81                    text_file.write("No se pueden calcular las
                          centralidades")
82                    text_file.write(os.linesep)
83
84                print(counter)
85                counter += 1
```

Código 10: Código Python para generar las redes genéticas y hacer el análisis topológico de las mismas

.5.3. Calcular grado

Este código realiza algunas operaciones relacionadas con el cálculo del grado de los genes a partir de un archivo de datos y luego realiza una filtración de los resultados.

```
1 #Obteniendo los genes con mayor grado
2
3 SECONDS2=0
4 SECONDS=0
5 echo -e  "El grado de los genes es:"
6 python3 CalculaGrado_Promedio.py | sort | uniq -c | sort -n -r |
     head -n 100000000
```

```
echo -e ' \t '  "Calculando el grado: $(echo $SECONDS/60 | bc -l)
    minutes."
SECONDS=0

awk '{if($2 >= 1){print}}'  Grado_Metastasis  > Grado__Metastasis
```

Código 11: Código Bash para calcular grado

```
import networkx as nx
import matplotlib.pyplot as plt
import seaborn as sns
import networkx.algorithms.community as nx_comm
import numpy as np
from ipywidgets import *
import math as math
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import os

path = 'Direccion/archivos/.sort&.sif'
os.chdir(path)

for p,n,f in os.walk(os.getcwd()):
    for a in f:
        a = str(a)
        if a.endswith('.txt'):
            print(a)
            file = open(a)
            texto = file.read()
            file.close()

#ANALIZAR CUALES SON LOS GENES CON MAYOR GRADO
            palabras = texto.split()
            for palabra in palabras:
                print(palabra.lower())

#ANALISIS CON NETWORKX
            file = open(a)
            G=nx.read_weighted_edgelist(file)
            dgenes = dict(nx.degree(G))

            print(p)
```

Código 12: Código Python para calcular grado de los genes

.5.4. Información adicional sobre los genes

Genes libres de tejido canceroso

- **MAP3K7**: Es una serina/treonina quinasa, también conocida como TAK1, que, como su nombre indica, añade grupos fosfato a los aminoácidos serina y treonina de las proteínas. Esta modificación es esencial para la activación o desactivación de las funciones de la proteína dentro de las vías de señalización de las citocinas, los factores de crecimiento y los receptores inmunitarios innatos, que desempeñan diversas funciones en el tejido sano. Según el Human Protein Atlas [71], esta proteína se expresa principalmente en tejido respiratorio sano, estómago, mama y trompas de Falopio. Dentro de la mama, las principales células que expresan este gen son los fibroblastos y las células glandulares. Se expresa moderadamente en el cáncer de mama, pero no se sabe cómo contribuye al desarrollo del cáncer. Sin embargo, su expresión es un factor pronóstico desfavorable para pacientes con cáncer de hígado. Probablemente porque esta proteína promueve la proliferación, migración e invasión de células hepatocelulares. El análisis genético identificó mutaciones en esta proteína en pacientes con mesotelioma. Además, los pacientes con mieloma múltiple y expresión elevada de esta proteína tienen una supervivencia libre de progresión y una supervivencia general más bajas . También está asociado con la progresión del melanoma. Por el contrario, se ha encontrado que esta proteína tiene una función supresora de tumores en el cáncer de próstata. A pesar de lo anterior, a la fecha no existe asociación de esta proteína con el desarrollo de cáncer de mama, lo cual es consistente con nuestros resultados.[69, 70, 71, 72, 73, 75, 76, 77, 78]
- **SIN3A**: Es un represor transcripcional, lo que indica que su función principal está en el núcleo de las células. Interactúa con la proteína MXI1 para reprimir la expresión génica en respuesta a MYC, uno de los principales oncogenes descritos en la literatura. De esta forma esta proteína participa en la prevención de la generación de células tumo-

rales. Además, esta proteína es necesaria para la expresión de genes asociados al ciclo circadiano y coopera con FOXK1 para la regulación del ciclo celular. Esta proteína se expresa en varios órganos, pero se expresa más en la nasofaringe, los testículos y la placenta. Dentro del tejido mamario, se expresa principalmente por células glandulares. Sin embargo, esta proteína no funciona como factor pronóstico para ningún tipo de cáncer hasta la fecha [71]. En el cáncer de mama, esta proteína forma un complejo con MAD1 para reprimir la expresión del receptor de ácido retinoico. La expresión reducida de esta proteína se ha asociado con una mayor progresión del cáncer, mientras que los pacientes con una expresión más baja de esta proteína tienen una tasa de supervivencia general más baja. Una mutación en esta proteína que impide su localización en el núcleo está asociada con una mayor proliferación de células de cáncer de mama. El silenciamiento de SIN3A en NSCLC conduce a una disminución de la proliferación celular y a un aumento de la apoptosis. Mientras que el complejo SIN3A/HDAC mantiene su sensibilidad a la quimioterapia. Al mismo tiempo, algunos estudios asocian esta proteína con el crecimiento, la migración y la invasión de células de cáncer de mama. Sin embargo, las funciones opuestas de SIN3A y SIN3B se han indicado en el cáncer de mama, lo que confirma que SIN3A funciona como un represor de la migración y SIN3B como un amplificador de la migración celular. Por el contrario, en el cáncer colorrectal se ha asociado con la progresión del cáncer. Además, en el melanoma, el complejo SIN3A-HDAC1/2 está asociado con un mayor crecimiento tumoral y diseminación metastásica. Así, SIN3A, además de encontrarse en tejido mamario sano, funciona como un potencial represor tumoral. [79, 80, 81, 82, 83]

- **SMARCA5**: Pertenece a una familia de proteínas con propiedades helicasa y ATPasa, diseñadas para regular la transcripción génica y regular la estructura de la cromatina. Esta proteína se expresa en diversos tejidos sanos, con expresión principal en la glándula paratiroides, nasofaringe, esófago, aparato digestivo, vejiga, testículos, cuello uterino, trompa de Falopio y en células del sistema inmunitario.

Curiosamente, en la mama, la principal expresión de este gen se encuentra en los linfocitos T, un componente importante del sistema inmunitario adaptativo. Aunque se puede encontrar en tejido tumoral mamario, ocurre en pocos casos. Por otro lado, funciona como un factor pronóstico desfavorable para pacientes con cáncer de hígado [71]. Estos resultados sugieren que la expresión de este gen en tejido sano puede ser principalmente por células del sistema inmunológico, los linfocitos T. Se ha demostrado que esta proteína se expresa en células de cáncer de mama MCF-7 sin que se haya descrito su función. Al contrario de nuestros resultados, se ha demostrado que SMARC5 se sobreexpresa en los tumores de mama y contribuye a la proliferación e invasión de estas células, lo que afecta negativamente el pronóstico de los pacientes. Además, se ha demostrado que las células deficientes en SMARCA5 reducen su capacidad para evitar la senescencia y convertirse en inmortales. Mientras que en otros tipos de cáncer esta proteína sustenta la proliferación de células hematopoyéticas. Por otro lado, SMARCA5 suprime la generación de células madre en el glioblastoma. La posible explicación de estas diferencias en la función puede deberse al origen de la línea celular utilizada y al tipo de cáncer que representa. Así, esta proteína puede tener diferentes funciones según el tipo de cáncer de mama. Sin embargo, se requieren más estudios para confirmar esta hipótesis [84, 85, 86, 87, 88]

- **TOP2B**: esta proteína es una topoisomerasa, que controla y altera los estados topológicos del ADN durante la transcripción [69] (Q02880). Curiosamente, esta función se ha descrito en los linfocitos B. Algunas variantes de esta proteína están relacionadas con el desarrollo de inmunodeficiencias de linfocitos B [89]. Según el Human Protein Atlas tiene expresión ubicua, es decir, se expresa en la mayoría de los tejidos, con menor expresión en el tejido adiposo. Dentro del tejido mamario, se expresa principalmente por células glandulares y linfocitos T. Su expresión se encuentra en la mayoría de los cánceres, incluido el cáncer de mama. Sin embargo, funciona como un factor de pronóstico desfavorable para el cáncer de hígado y favorable para el cáncer renal

[71]. Esta proteína no está directamente asociada con el cáncer de mama. Sin embargo, hay mutaciones en algunas de las proteínas que interactúan con él, como CTCF que está mutado en el cáncer gastrointestinal [58]. Además, hay una gran cantidad de copias de este gen en varios tipos de cáncer, incluido el cáncer de mama; sin embargo, su contribución al desarrollo del cáncer no está claramente definida [90]. En base a lo anterior, destacamos la identificación de MAP3K7 y SIN3A como genes claramente asociados a tejido mamario sano y ubicamos a SMARCA5 y TOP2B como potenciales funciones distintas dependiendo del subtipo de cáncer de mama analizado. Sin embargo, son necesarios más estudios sobre el papel de TOP2B en el cáncer de mama.

Genes tumorales primarios

- **ALLC**:Hasta la fecha, solo se conoce la expresión del ARNm, lo que impide conocer su función precisa. Según el Human Protein Atlas, este gen no se expresa en ningún tejido humano excepto en el testículo y su expresión no se ha detectado en ningún tipo de cáncer [71, 69](Q8N6M5). Esta enzima se encuentra en bacterias, por lo que los hallazgos recientes del genoma bacteriano en tumores [49, 50] podrían explicar este hallazgo en nuestros datos.

- **AVP**: La vasopresina actúa como agente antidiurético [69](P01185). Esta proteína se expresa en el hipotálamo y los testículos. En tejido mamario su expresión es casi nula, pero las pocas células que la expresan son los fibroblastos. Su expresión está ausente en todos los tipos de cáncer por lo que no es factor pronóstico en ninguno de ellos [71]. Las células de cáncer de mama expresan anormalmente AVP y sus receptores. Se han encontrado efectos antiproliferativos y antimetastásicos en células luminales. Además, se ha demostrado que los análogos sintéticos de AVP tienen un efecto negativo sobre la proliferación y la metástasis en modelos murinos de cáncer de mama. Aunque los análogos de esta proteína tienen un efecto negativo sobre

la tumorogénesis, la vasopresina expresada por las células tumorales parece tener un efecto positivo sobre la proliferación y el desarrollo tumoral.

- **WNT8A**: Esta proteína está asociada con el desarrollo embrionario [69]. Según el Human Protein Atlas, no se expresa en ningún tejido sano en adultos y se expresa en tumores colorrectales y de ovario, pero no es un factor pronóstico para ningún tipo de cáncer [71]. El mensajero Wnt8A se sobreexpresa en respuesta al estradiol, una hormona sexual femenina involucrada en el metabolismo de los lípidos. Pero su papel en este y otros tipos de cáncer no ha sido dilucidado. Sin embargo, otros miembros de esta misma familia de proteínas participan en el desarrollo del cáncer de mama [52, 53]. Contribuye principalmente a la activación de la transición epitelio-mesenquimatosa, un fenómeno necesario para iniciar el desprendimiento de las células tumorales primarias y la búsqueda de un nuevo sitio para la metástasis [91].

- **SSTR4**: Es el receptor de somastatina-14. Su actividad está mediada por proteínas G. Participa en la liberación de ácido araquidónico y en la activación de MAPK (Proteína Quinasa Asociada a Mitógenos) [69]. No se expresa en ningún tejido sano, incluida la mama. La mayoría de los tumores son negativos para su expresión, excepto el cáncer de páncreas y pulmón. No está asociado con ningún factor pronóstico [71]. Aunque se ha confirmado su expresión en tejido tumoral de mama en algunas muestras [54], no se ha determinado su función específica en tumores de mama. En otros tipos de cáncer, los miembros de esta familia de proteínas diferentes a SSTR4 participan en el control del crecimiento en el cáncer de páncreas [92]. Además, esta proteína participa con SSTR1 para inducir la detención celular [93] y en t. El desarrollo de cáncer de vesícula biliar [94].

- **FGF3**: El factor de crecimiento de fibroblastos 3, es una proteína implicada en el desarrollo embrionario, la proliferación celular y la diferenciación celular [56]. Según el atlas de proteínas humanas, en tejido normal su expresión es exclusivamente en el cerebelo. La mayoría

de los cánceres son negativos para esta proteína, incluido el cáncer de mama, y no está asociada con ningún factor pronóstico [71].

- **FGF8**: Similar a FGF3, esta proteína tiene funciones en el desarrollo embrionario, proliferación celular, diferenciación celular y migración celular [55]. Según el Human Protein Atlas, en tejido sano su expresión está restringida al músculo esquelético. Hay poca expresión de estos en tumores y no se asocia con ningún pronóstico [71]. Varios miembros de esta familia de proteínas están involucrados en la activación de la proliferación y la transición epitelial-mesenquimatosa [95]. En el cáncer de mama, se ha descubierto que FGF8 aumenta la angiogénesis dependiente de HBXIP [96] e independiente [97]. También promueve el ciclo celular [98] y, por lo tanto, la proliferación celular debido a la regulación negativa de la muerte celular. Esta proteína está asociada con la tumorogénesis en diferentes tipos de cáncer [99]. Específicamente, FGF8 promueve la proliferación y la resistencia a la terapia con inhibidores de EGFR en células de hepatocarcinoma [100]. A pesar de la evidencia de la función de FGF3 en el cáncer de mama u otros tipos de cáncer, los hallazgos con otros miembros de la familia sugieren un papel similar en el cáncer.

- **FEV**: esta proteína funciona como un regulador transcripcional [69]. Tiene un papel en el crecimiento celular [101]. Basado en el Human Protein Atlas, la expresión de esta proteína se enriquece en el cerebro sano. En el cáncer, se encuentra que su expresión está enriquecida en el cáncer de páncreas, próstata y estómago y su expresión se considera un factor de pronóstico favorable para el cáncer de páncreas [71]. No existe evidencia experimental que evalúe la función de esta proteína en el cáncer de mama. Por otro lado, en el cáncer de páncreas se ha encontrado que la expresión de este gen inhibe el crecimiento, migración e invasión de células tumorales, pero su expresión está disminuida en pacientes con metástasis [57]. Sin embargo, el conocimiento sobre las funciones de este gen en la mayoría de los cánceres es limitado.

Genes de metástasis

- **FGA**: La cadena alfa de la proteína fibrinógeno es una proteína que junto con la cadena beta se polimeriza para formar una red de fibras insolubles. Su función principal es la coagulación de la sangre. Sin embargo, durante las primeras etapas de la reparación celular, guía la migración de las células epiteliales [69](P02671). Es una proteína secretada [102] y según el Human Protein Atlas, esta proteína se expresa principalmente en el hígado. La mayoría de los cánceres son negativos para esta proteína, excepto el cáncer de pulmón, riñón e hígado. Funciona como un factor pronóstico desfavorable para el cáncer renal y favorable para el cáncer de mama [103] Se ha confirmado la producción de esta proteína por parte de las células tumorales y su papel potencial en el crecimiento tumoral y la metástasis [67]. En modelos murinos de melanoma y carcinoma de Lewis se ha demostrado que esta proteína es necesaria para el crecimiento tumoral [68]. Debido a que esta proteína es soluble y se encuentra en la circulación, los niveles plasmáticos de esta proteína se han asociado con un mal pronóstico en el cáncer de mama y se ha propuesto como un marcador de pronóstico para pacientes con quimioterapia [104, 105]. Como componente de la matriz extracelular, esta proteína se ha asociado con la promoción de células madre cancerosas y metástasis [106].

- **PAK3**: una serina/treonina quinasa involucrada en varias vías de señalización, incluidas las que regulan el citoesqueleto de actina, la migración celular y el ciclo celular [69](O75914). Esta proteína normalmente se expresa en el cerebro, el páncreas , placenta y ganglios linfáticos. En mama, la expresión del ARN mensajero se encuentra principalmente en fibroblastos y células glandulares. Su expresión es negativa en casi todos los cánceres y por lo tanto no se asocia con ningún factor pronóstico [71]. La expresión de esta familia de proteínas se ha asociado con mal pronóstico en algunos pacientes [107]. Aunque no se ha identificado la función de esta proteína en el cáncer de mama, se ha demostrado que promueve la metástasis de las células del carci-

noma hepatocefálico al inducir la transición epitelial-mesenquimatosa [108]. De manera similar, en el cáncer de pulmón, PAK3 promueve la migración celular [109]. Por otro lado, promueve la generación de células madre cancerosas en el cáncer de páncreas [110]. En general, esta familia de proteínas promueve la migración, supervivencia y proliferación en células tumorales [111].

- **ITGA4**: Esta proteína forma parte de los receptores a7b1 (también conocido como antígeno muy tardío 4, VLA-4) y a4b7, que reconocen la fibronectina, un componente de la matriz extracelular. También reconocen la molécula de adhesión de células vasculares 1 (VCAM-1), que se expresa en la superficie de varios tipos de células, incluidas las células endoteliales [69] (P13612). Su interacción con estos ligandos induce la extravasación de células inmunitarias [112]. La expresión del mensajero se encuentra principalmente en el tejido linfoide. Dentro del tejido mamario, su expresión se encuentra en linfocitos T y macrófagos. Su expresión se encuentra en la mayoría de los cánceres, pero se destaca como factor pronóstico desfavorable para el cáncer renal y favorable para el cáncer de cabeza y cuello [71]. La expresión de esta proteína se ha demostrado en algunos tumores de mama, sin embargo, no se expresa en tejido sano. Por otro lado, se ha demostrado que esta proteína promueve la proliferación, migración y metástasis del cáncer gástrico al regular el inmuno-microambiente [113] del tumor. Debido a que esta proteína es expresada esencialmente por las células inmunes, y el tejido metastásico analizado en nuestro estudio es la primera metástasis de ganglios linfáticos, dicha expresión podría reflejar el sitio de la metástasis. Se requieren más estudios para comprender el papel de esta proteína en la metástasis del cáncer de mama.