Pontificia Universidad JAVERIANA Cali

Facultad de Ingenieria y Ciencias

Acta de Correcciones al Proyecto de Grado

Fecha: Agosto 31 del 2022

Autores: Juan David Arce Rentería y Nicolas Ibagon Rivera

Nombre del Proyecto de Grado: Desarrollo e implementación de una estrategia integrativa para la detección de nuevos módulos genéticos y nuevos genes asociados al inicio y desarrollo del cáncer colorrectal

Director: Mauricio Alberto Quimbaya Gómez.
Codirector: Chrystian Camilo Sosa Arango.

Como indica el artículo 2.27 de las Directrices de Trabajo de Grado, he verificado que los estudiantes indicados arriba han implementado todas las correcciones que los Jurados del Proyecto de Grado definieron que se efectuaran, como consta en el Acta de Calificación correspondiente.

Firma del director del Proyecto de Grado

Nota de Aceptación

Aprobado por el Comité de Trabajo de Grado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana para optar el título de Ingeniero de Sistemas y Computación.

Dr. Camilo Rocha
Decano de la Facultad de Ingeniería

ING. GERARDO MAURICIO SARRIA
Director Carrera Ingeniería Sistemas y Computación.

Dr. Mauricio Alberto Quimbaya Gómez
Director(a) Trabajo

Dra. Gloria I Álvarez
Jurado 1

Dr. Henrry Fabián Tobar Tossé
Jurado 2

Santiago de Cali, septiembre 1 del 2022

Doctor Gerardo Sarria
Director del Programa de Ingeniería de Sistemas y Computación
Pontificia Universidad Javeriana Cali

Asunto: Entrega del trabajo de grado de los estudiantes Juan David Arce Rentería y Nicolas Ibagon Rivera.

Estimado Dr. Sarria,

Con la presente carta quiero manifestar que el proyecto de grado titulado "*Desarrollo e implementación de una estrategia integrativa para la detección de nuevos módulos genéticos y nuevos genes asociados al inicio y desarrollo del cáncer colorrectal*" presentado por los estudiantes Juan David Arce Rentería y Nicolas Ibagon Rivera se encuentra en su versión final, luego de incluir las correcciones sugeridas por los jurados de la tesis durante el proceso de disertación y evaluación del mismo.

Cualquier pregunta relacionada con el trabajo de grado de Nicolás y Juan David, estaré atento a solucionarla.

Sinceramente,

Mauricio A Quimbaya G-Ph.D.
Profesor asociado al Departamento de Ciencias Naturales y Matemáticas
maquimbaya@javerianacali.edu.co
(+57-2)321-82-00-Ext: 8636.

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Ingeniería de Sistemas y Computación
Trabajo de Grado

Desarrollo e implementación de una estrategia integrativa para la detección de nuevos módulos genéticos y nuevos genes asociados al inicio y desarrollo del cáncer colorectal

Juan David Arce Rentería
Nicolás Ibagón Rivera

Director: Dr. Mauricio Alberto Quimbaya Gómez
Co-director: Chrystian Camilo Sosa Arango

Julio 2022

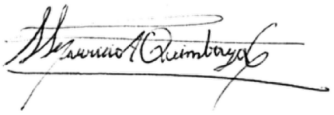
Santiago de Cali, Julio 2022.

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo Mauricio Sarria
Director Carrera de Ingeniería de Sistemas y Computación.
Cali.

Cordial Saludo.

Por medio de la presente me permito informarle que los estudiantes de Ingeniería de Sistemas Juan David Arce Rentería (cod: 8939473) y Nicolás Ibagón Rivera(cod: 8938633) trabajan bajo nuestra dirección, en el proyecto de grado titulado "Desarrollo e implementación de una estrategia integrativa para la detección de nuevos módulos genéticos y nuevos genes asociados al inicio y desarrollo del cáncer colorectal".

Atentamente,



Dr. Mauricio Alberto Quimbaya Gómez

Chrystian Camilo Sosa Arango

Santiago de Cali, Julio 2022.

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo Mauricio Sarria
Director Carrera de Ingeniería de Sistemas y Computación.
Cali.

Cordial Saludo.

Nos permitimos presentar a su consideración el trabajo de grado titulado "Desarrollo e implementación de una estrategia integrativa para la detección de nuevos módulos genéticos y nuevos genes asociados al inicio y desarrollo del cáncer colorectal" con el fin de cumplir con los requisitos exigidos por la Universidad para optar al título de Ingeniero de Sistemas.

Al firmar aquí, damos fe que entendemos y conocemos las directrices para la presentación de trabajos de grado de la Facultad de Ingeniería aprobadas el 26 de Noviembre de 2009, donde se establecen los plazos y normas para el desarrollo del anteproyecto y del trabajo de grado.

Atentamente,

Juan David Arce Rentería
Código: 8939473

Nicolás Ibagón Rivera
Código: 8938633

Agradecimientos

Agradecemos a nuestro director de tesis, Mauricio Alberto Quimbaya Gómez y nuestro Co-director Chrystian Camilo Sosa Arango, por su inmensurable guía, apoyo y conocimientos en la realización de este proyecto. Sin ellos nada de esto sería posible.

Damos gracias a nuestras familias, quienes nos apoyaron en todos los aspectos posibles durante todo el proceso de la carrera ayudándonos a ser mejores profesionales y personas cada día, también les agradecemos a nuestros amigos quienes nos han acompañado en todo el transcurso de la carrera, y nos han servido como apoyo emocional. Gracias a todos por sus palabras de animo, y esfuerzo.

Abstract

The advent of omics technologies, the development of computational techniques based on machine learning applied to biological systems, and the integration of both paradigms in mathematical models, has made it possible to advance in the causal understanding of complex diseases such as cancer. In this sense, from a systemic perspective, the use of biological networks and the representation of molecular systems such as genes, proteins and their interaction dynamics, has allowed an approach to biological systems from graph theory. From this perspective, in recent years a wide variety of strategies have been developed, which, from graph theory, have contributed to the understanding of the deleterious process that leads to disease and, equitably, to identify key nodes of the network. which could be related to different types of diseases, such as cancer. In the present work, we integrate different types of biological information associated with the genetic understanding of the origin and development of the disease, coupling it to the most detailed map of protein-protein interaction that exists. Subsequently, we performed fundamental analyzes on classical measures of the topology of the constructed network, which were useful to identify key elements of the network. We assigned weights to nodes and edges of the network based on biological information, which was a fundamental procedure for prioritizing network elements (proteins) associated with cancer and specifically with colorectal cancer. Based on this information and with the constructed network, we implemented modularity algorithms to identify specific communities that could be specifically associated with the development of colorectal cancer, and finally we implemented non-overlapping community characterization algorithms and specific machine learning strategies to find potential Colorectal cancer-associated proteins.

Resumen

El advenimiento de las tecnologías ómicas, el desarrollo de técnicas computacionales basadas en el aprendizaje de máquina aplicado a sistemas biológicos y la integración de ambos paradigmas en modelos matemáticos, ha permitido avanzar en el entendimiento causal de enfermedades complejas como el cáncer. En este sentido, desde de una perspectiva sistémica, el uso de redes biológicas y la representación de sistemas moleculares como genes, proteínas y sus dinámicas de interacción, ha permitido realizar una aproximación a los sistemas biológicos desde la teoría de grafos. Desde esta perspectiva, en los últimos años se han desarrollado una gran variedad de estrategias, las cuales, desde la teoría de grafos, han contribuido al entendimiento del proceso deletéreo que conduce a la enfermedad y, equitativamente, a identificar nodos clave de la red los cuales podrían estar relacionados con diferentes tipos de enfermedades, como lo sería el cáncer. En el presente trabajo, integramos distintos tipos de información biológica asociada a la comprensión genética del origen y desarrollo de la enfermedad, acoplándola al mapa más detallado de interacción proteína-proteína que existe. Posteriormente, realizamos análisis fundamentales sobre medidas clásicas de la topología de la red construida, que fueron útiles para identificar elementos claves de la red. Asignamos pesos a los nodos y a las aristas de la red según la información biológica, lo cual fue un procedimiento fundamental para priorizar elementos de la red (proteínas) asociadas al cáncer y específicamente al cáncer colorrectal. Con base en dicha información y con la red construida, implementamos algoritmos de modularidad para identificar comunidades específicas que pudieran estar específicamente asociadas al desarrollo de cáncer colorrectal, y finalmente implementamos algoritmos de caracterización de comunidades no sobrelapantes y estrategias específicas de aprendizaje de máquina para encontrar potenciales proteínas asociadas al cáncer colorrectal.

Palabras Clave: Cáncer, Redes Biológicas, Identificación de módulos de enfermedades, Redes Heterogéneas, Genética, Aprendizaje de máquina, Genes Conductores, Genética de Enfermedades, Biología de Redes y Sistemas.

Índice general

1. Descripción del Problema **17**
1.1. Planteamiento del Problema . . . 17
1.1.1. Formulación . . . 18
1.1.2. Sistematización . . . 18
1.2. Objetivos . . . 18
1.2.1. Objetivo General . . . 18
1.2.2. Objetivos Específicos . . . 19
1.3. Justificación . . . 19
1.4. Delimitaciones y Alcances . . . 20
1.4.1. Entregables . . . 20

2. Investigación de la Literatura **21**
2.1. Áreas Temáticas . . . 21
2.1.1. Trabajos Relacionados . . . 21

3. Marco Teórico **23**
3.1. Biología de Sistemas . . . 23
3.2. Cáncer como enfermedad multifactorial . . . 24
3.3. El cáncer entendido como una red de interacción . . . 25
3.4. Cáncer de colon . . . 26
3.5. Clasificación funcional de genes asociados a cáncer . . . 27
3.6. Algoritmos de detección de comunidades aplicados al entendimiento del cáncer . . . 28
3.6.1. Algoritmos de comunidades modulares . . . 29
3.6.2. Algoritmos de comunidades soprelapantes . . . 29
3.7. Aprendizaje de máquina y métodos de aprendizaje de máquina aplicados al entendimiento del cáncer . . . 30

4. Implementación **35**
4.1. Metodología . . . 35
4.1.1. Integración de distintos tipos de información biológica, asociada a la comprensión genética del origen y desarrollo del cáncer, a un mapa detallado de interacción proteína-proteína en humanos. . . . 35
4.1.2. Identificación de proteínas fundamentales, para el entendimiento del cáncer, a través del análisis topológico de la red de interacción construida. . . . 37
4.1.3. Identificación de módulos de interacción claves, para el entendimiento del cáncer, por medio de la identificación de comunidades, sobre la red de interacción construida . . . 38

4.1.4. Predicción de proteinas asociadas a cáncer colorectal usando métodos de aprendizaje de maquina . . . 39
4.2. Resultados . . . 41
4.2.1. Acoplamiento de distintos tipos de información biológica, asociada a la comprensión genética del origen y desarrollo del cáncer, a un mapa detallado de interacción proteína-proteína en humanos . . . 41
4.2.2. Identificación de proteínas fundamentales, para el entendimiento del cáncer, a través del análisis topológico de la red de interacción construida . . . 44
4.2.3. Identificación de módulos de interacción claves, para el entendimiento del cáncer, por medio de la identificación de comunidades, sobre la red de interacción construida . . . 49
4.2.4. Predicción de proteinas asociadas a cáncer colorectal usando métodos de aprendizaje de maquina . . . 52
4.3. Análisis de resultados . . . 63
4.3.1. Acoplamiento de distintos tipos de información biológica, asociada a la comprensión genética del origen y desarrollo del cáncer, a un mapa detallado de interacción proteína-proteína en humanos. . . . 63
4.3.2. Identificación de proteínas fundamentales, para el entendimiento del cáncer, a través del análisis topológico de la red de interacción construida . . . 63
4.3.3. Identificación de módulos de interacción claves, para el entendimiento del cáncer, por medio de la identificación de comunidades, sobre la red de interacción construida . . . 65
4.3.4. Predicción de proteínas asociadas a cáncer colorrectal usando métodos de aprendizaje de maquina . . . 67

5. Conclusiones **69**

Bibliografía **71**
.1. Anexo 1 . . . 75
.2. Anexo 2 . . . 75
.3. Anexo 3 . . . 75
.4. Anexo 4 . . . 75
.5. Anexo 5 . . . 76
.6. Anexo 6 . . . 76
.7. Anexo 7 . . . 76
.8. Anexo 8 . . . 76
.9. Anexo 9 . . . 76
.10. Anexo 10 . . . 76
.11. Anexo 11 . . . 76

Introducción

El proceso de carcinogénesis refleja una serie amplia y diversa de cambios moleculares, celulares y fisiológicos que se traducen en un complejo grupo de patologías, que si bien, desde perspectivas como su origen, desarrollo, tratamiento y evolución, podrían considerarse como distintas enfermedades, convergen en características definitorias que permite agruparlas como neoplasias. Sin embargo, pese a la gran variedad de factores que inciden directamente sobre el desarrollo de la enfermedad, es posible encontrar un punto de convergencia y, definir a la célula neoplásica como una célula anárquica que pierde control sobre sus funciones específicas dentro de un contexto tisular comportándose de una manera autónoma que desconoce las señales regulatorias que definen y limitan su destino celular. En el 2020 se reportaron cerca de 20 millones de casos de cáncer alrededor del mundo, de los cuales un 10 % corresponde al cáncer de colon [1].

En Colombia el Instituto Cancerológico Nacional, reporta que para el año 2020 se diagnosticaron 2.801 casos nuevos para mujeres y 1.970 para hombres. Para las mujeres, los carcinomas de mama (20 %), tiroides (12 %), piel (11.4 %) y cervicouterino (10.7 %), fueron las malignidades más frecuentemente diagnosticadas, mientras que para varones fueron los carcinomas de próstata (15.1 %), piel (14.5 %), estómago (11.2 %) y colorrectal (8.5 %), los que presentaron un mayor conteo de nuevos casos. Para este mismo año el Instituto Nacional de Cancerología reportó 1.314 defunciones asociadas al progreso de la enfermedad (Instituto Nacional de Cancerología-2020).

Según la plataforma para la Información del Cáncer en Colombia (INFOCANCER), para el año 2019, el Valle del Cauca fue el tercer departamento con mayor número de muertes por esta enfermedad, debido a que contó con 5045 defunciones asociadas a procesos carcinogénicos. En mujeres, los carcinomas más letales fueron el cáncer de mama (463 muertes), cuello y cuerpo del útero (266 defunciones), estómago (248 muertes) y carcinomas colorrectales (255 defunciones). Para los hombres las neoplasias con una mayor frecuencia de mortalidad fueron cáncer de estómago (337 muertes), cáncer de próstata (408 decesos), y adenocarcinomas colorrectales (230 defunciones).

En el contexto de la teoría de redes, las redes complejas pueden ser definidas como una colección de nodos o vértices conectados por vínculos que representan diversas interacciones complejas entre los nodos. Casi cualquier sistema a gran escala, ya sea natural o hecho por el hombre, puede ser visto como una red compleja de entidades que interactúan.

A nivel biológico, los distintos niveles organizacionales característicos de la vida, se estructuran en forma de redes. Esto sucede a nivel macro, al constituirse redes de depredación, redes poblacionales, comunitarias y ecosistémicas, y también ocurre a una escala microscópica, en donde la célula se organiza funcionalmente a partir de un intrincado patrón de interacciones moleculares que forman redes que ejercen su efecto en distintos niveles de regulación [2].

Molecularmente, uno de los tipos de redes biológicas que están directamente asociadas a la funcionalidad celular, son las redes de interacción proteína-proteína. Estas redes de interacción están definidas por el contacto físico existente entre un par o un grupo de proteínas específicas. En la representación gráfica de la red, los nodos representan las proteínas y los vínculos o aristas entre ellas, la interacción física entre las mismas. Dichas interacciones físicas son determinadas experimentalmente por métodos, o que bien determinan concretamente la interacción entre un par específico de proteínas, como lo son los ensayos de doble híbrido de levadura [3], o por estrategias que identifican grupos de proteínas que interactúan entre sí, pero que carecen de la habilidad de determinar interacciones específicas entre pares particulares de proteínas pertenecientes al complejo, como es el caso de la técnica de *Tag-Tagging* [4].

Dado que las proteínas son los elementos moleculares que finalmente son los catalizadores de los distintos procesos biológicos desarrollados por una célula, el estudio del proteoma y de sus dinámicas por medio de metodologías sistémicas como ensayos de doble híbrido de levadura a gran escala, ha tenido un gran auge en la última década como estrategia para tratar de dar una explicación a distintos fenómenos biológicos, incluido el proceso que conlleva al desarrollo de enfermedades particulares [5][6]. De la misma manera, en los últimos años el estudio del proceso cancerígeno se ha nutrido de estudios sistémicos e integrativos que han abordado la perspectiva de redes biológicas para tratar de elucidar las dinámicas características de la célula neoplásica. Así, la aplicación de las ciencias de la computación en el estudio del cáncer puede ayudar a comprender la enfermedad desde un punto de vista holístico. La aplicación directa de la perspectiva sistémica a la biología se conoce como biología de sistemas y define que la funcionalidad celular surge como consecuencia de la interacción precisa y coordinada de los distintos componentes celulares. Las funciones específicas que desempeña una célula particular son propiedades emergentes, derivadas de la dinámica de interacción de los componentes celulares. Por lo tanto, el entendimiento de las dinámicas que determinan la funcionalidad celular sólo cobra sentido en un contexto holístico y no pueden ser accedidas mediante un entendimiento aislado de los distintos componentes del sistema [7].

En los últimos 15 años, la producción masiva de datos derivados de la aplicación de tecnologías ómicas las cuales son experimentalmente holísticas, han facilitado el desarrollo de estrategias y aproximaciones basadas en los principios integradores de la biología de sistemas. La genómica, la transcriptómica, la metabolómica y la proteómica han permitido acceder al conocimiento de la dinámica celular desde una perspectiva de totalidad en donde, por primera vez en la historia experimental de la biología, se ha pasado del análisis puntual de los elementos celulares a la comprensión genómica, transcriptómica, metabolómica y proteómica de la célula. Con el desarrollo de las herramientas ómicas se ha dado el primer paso para realizar una reinterpretación sistémica de las dinámicas celulares. Sin embargo, aunque el conocimiento de la mayoría de los componentes celulares y de sus dinámicas ha sido fundamental para avanzar hacia el conocimiento holístico de la célula, también es primordial mencionar que a la par con el desarrollo de las tecnologías ómicas ha habido un gran avance en el desarrollo de herramientas computacionales y en el poder de computo que han permitido organizar sistemáticamente y analizar eficientemente los distintos datos gene-

rados [8][9]. De la misma manera, la combinación de los resultados experimentales con desarrollos matemáticos específicos, han permitido el desarrollo de modelos que mediante la integración de datos y el análisis de las dinámicas de los mismos, han permitido no sólo describir el comportamiento global de las dinámicas celulares que convergen en una función biológica particular, sino que, también son herramientas predictivas que ayudan a pronosticar estados y comportamientos celulares particulares, inaccesibles antes del desarrollo e implementación de esta estrategia holística [10]. Así, las estrategias de aprendizaje de máquina como método de análisis de datos permiten crear modelos de aprendizaje automático que pueden encontrar correlaciones no tan evidentes entre los nodos de las redes analizadas, información que ayudaría a comprender mejor las propiedades del cáncer y así, potencialmente, encontrar o caracterizar nuevos genes, circuitos génicos o redes de interacción gen-proteína o proteína-proteina, que permitan un entendimiento sistémico de la enfermedad que pueda derivar en nuevos y mejores tratamientos.

El advenimiento de las tecnologías ómicas y la avalancha de datos genómicos, transcriptómicos y proteómicos generaron la necesidad de buscar nuevas estrategias para dar una adecuada interpretación y sentido biológico a los datos producidos. Un claro ejemplo de esta necesidad fue la generación de redes de coexpresión basadas en el análisis de diferentes experimentos de microarreglos [11] y las redes de interacción proteína-proteína fundamentadas en las interacciones existentes en el proteoma celular [12]. En este sentido el producto directo de las tecnologías ómicas puede representarse en redes de interacción gen-gen, gen-proteína o proteína-proteína, cuya caracterización topológica y posterior análisis biológico permiten el entendimiento de un fenómeno particular desde una perspectiva holística y sistémica por medio de la caracterización de patrones funcionales derivados de la organización de la red [13]. Teniendo estas consideraciones en mente, el objetivo de esta investigación es diseñar una estrategia integrativa que utilice distintos tipos de información y que acople la teoría de grafos con métodos de aprendizaje de máquina, para identificar elementos y módulos genéticos asociados al origen o progresión del cáncer de colon.

Capítulo 1

Descripción del Problema

1.1. Planteamiento del Problema

A pesar de los grandes avances que han surgido en campos como la biología celular y molecular, la oncología, las ciencias médicas y las ciencias de la computación en los últimos años, sigue siendo de gran esfuerzo el intentar entender el cáncer como una enfermedad multifactorial sobre la cual convergen distintas capas de complejidad desde una perspectiva molecular y celular y con ello encontrar una posible cura. En el año 2020 alrededor del mundo se reportaron 19'292.789 casos de cáncer de los cuales casi 2 millones de casos están relacionados con el cáncer de colon, siendo estos cerca del 10 % del total de los casos [1]. En Colombia el Instituto Cancerológico Nacional, reporta que para el año 2020 se diagnosticaron 2.801 casos nuevos para mujeres y 1.970 para hombres. Para las mujeres, los carcinomas de mama (20 %), tiroides (12 %), piel (11.4 %) y cervicouterino (10.7 %), fueron las malignidades más frecuentemente diagnosticadas, mientras que para varones fueron los carcinomas de próstata (15.1 %), piel (14.5 %), estómago (11.2 %) y colorrectal (8.5 %), los que presentaron un mayor conteo de nuevos casos. Para este mismo año el Instituto Nacional de Cancerología reportó 1.314 defunciones asociadas al progreso de la enfermedad (Instituto Nacional de Cancerología-2020).

El siglo 21 ha sido dominado por la investigación en el campo de la biología, y se debe en parte al planteamiento de nuevos paradigmas biológicos y metodológicos como las ciencias ómicas que han tenido una influencia directa sobre las ciencias médicas y sobre el entendimiento de enfermedades particulares. Desde una perspectiva sistémica, la aplicación de la teoría de redes sobres sistemas biológicos ha permitido aproximar sistemas complejos de interacción molecular desde la teoría de grafos, permitiendo un nuevo enfoque más integrativo y holístico sobre problemas complejos, antes enfocados desde una perspectiva mecanisista y por lo tanto, reduccionista. [6].

El interactoma humano es el conjunto de interacciones proteína-proteína que ocurre en las células humanas en un momento particular y bajo unas condiciones específicas. Las redes de interacción proteína-proteína representan el producto final del flujo de la información genética, en donde las distintas proteínas, como los principales elementos que ejecutan las funciones biológicas, y su interacción física, modulan los distintos procesos biológicos que ocurren en una célula bajo un contexto tisular definido. Gracias a los mapas globales de interacción proteína-proteína, las enfermedades se han empezado a entender como perturbaciones puntuales y definidas sobre dichos mapas de interacción. El entendimiento del cáncer como una enfermedad basada en la interacción de proteínas, ha permitido la identificación de módulos funcionales que son alterados durante el inicio o progreso

de la enfermedad, lo que ha sido fundamental para la identificación de nuevos elementos genéticos o circuitos genéticos que podrían ser fundamentales para el entendimiento causal del cáncer.

Para el estudio de estas enfermedades se han elaborado distintas metodologías, las cuales han dado grandes resultados a la hora de identificar y extraer los llamados "módulos de enfermedades", los cuales están conformados por grupos de proteínas y sus respectivas interacciones que al modificar sus propiedades topológicas, inciden directamente sobre el origen o progreso de la enfermedad. Entre los distintos métodos que se han formulado, uno de los más recientes, diseñado por Beethika Tripathi, ha sido proficiente en comparación a otros [14].

El cáncer es un fenómeno progresivo que evoluciona por medio de la acumulación de mutaciones en distintas partes del genoma, que repercute en la producción de proteínas funcionales, y por ende, en las dinámicas de interacción de la red [15]. Por ello, en el presente trabajo, se pretende diseñar y estructurar una estrategia integrativa que utilice la teoría de grafos y modelos de aprendizaje de máquinas para identificar nuevas proteínas y módulos de interacción asociados al orígen o progreso del cáncer de colon.

1.1.1. Formulación

¿De qué forma se puede acoplar la teoría de grafos, implementada sobre redes de interacción proteína-proteína, con técnicas de aprendizaje de máquina para identificar proteínas y módulos de interacción asociados al origen o progreso del cáncer de colon?

1.1.2. Sistematización

- ¿De qué forma se pueden acoplar distintos tipos de información biológica, asociada a la comprensión genética del origen y desarrollo del cáncer, a un mapa detallado de interacción proteína-proteína en humanos?
- ¿De qué forma se pueden identificar elementos claves para el entendimiento del cáncer sobre la red de interacción construida?
- ¿De qué forma se pueden identificar módulos de interacción claves para el entendimiento del cáncer sobre la red de interacción construida?
- ¿De qué forma se pueden implementar estrategias de aprendizaje de máquina sobre la red de interacción construida para encontrar potenciales proteínas asociadas al cáncer de colon?

1.2. Objetivos

1.2.1. Objetivo General

Diseñar y estructurar una estrategia integrativa, que utilice la teoría de grafos y estrategias de aprendizaje de máquinas, para identificar nuevas proteínas y módulos de interacción asociados al

origen o progreso del cáncer de colon.

1.2.2. Objetivos Específicos

- Acoplar distintos tipos de información biológica, asociada a la comprensión genética del origen y desarrollo del cáncer, a un mapa detallado de interacción proteína-proteína en humanos.
- Identificar proteínas fundamentales, para el entendimiento del cáncer, a través del análisis topológico de la red de interacción construida.
- Identificar módulos de interacción claves, para el entendimiento del cáncer, por medio de la identificación de comunidades, sobre la red de interacción construida.
- Implementar estrategias de aprendizaje de máquina, sobre la red de interacción construida, para la identificación de nuevas proteínas asociadas al cáncer de colon.

1.3. Justificación

Como se planteó anteriormente, el cáncer sigue siendo una enfermedad mortal y, a su vez, el cáncer de colon representa uno de los tipos de cáncer con más diagnósticos y defunciones, tanto a nivel global como nacional y regional, y es a partir de las cifras de mortalidad registradas por el cáncer que surge la pertinencia de esta investigación. Esto, aunado al hecho de que el cáncer como enfermedad multifactorial que se compone de distintas capas de complejidad molecular y que con el advenimiento de las ciencias ómicas, como la proteómica y la interactómica, se puede aproximar desde una visión holística (top-down), utilizando aplicaciones que vienen desde la matemática o las ciencias computacionales. Entender la enfermedad desde la perspectiva de la ingeniería, abre puertas a la aplicación de herramientas computacionales sobre datos biológicos para entender desde una perspectiva sistémica el origen o progreso de la enfermedad. Dado que el cáncer posee características que lo convierten en una enfermedad muy variable en cuanto a sus causas y su forma de desarrollarse y manifestarse, es pertinente identificar elementos que influyan de una manera causal sobre el desarrollo de la enfermedad.

Específicamente para el cáncer de colon, no existe una gran variedad de este tipo de aproximaciones, por lo que es importante plantear estrategias integrativas para entender particularmente el cáncer de colon. En este caso particular, pretendemos identificar proteínas que intervengan en el inicio y desarrollo de cáncer colorrectal. Parte de los beneficios que supone realizar esta investigación desde nuestra disciplina es precisamente la integración de la biología y la ingeniería de sistemas en una relación interdisciplinar, ya que la colaboración entre dos áreas del conocimiento permite encontrar respuestas a problemas que normalmente serían más difíciles de encontrar desde una sola perspectiva.

El centro de esta investigación es diseñar y estructurar una estrategia integrativa para la detección de nuevos genes, representados por las proteínas que éstos codifican, que se puedan asociar,

por lo menos desde una perspectiva de experimentación in silico, al origen o progresión del cáncer de colon. Esto puede ser logrado utilizando distintas estrategias de aprendizaje de máquina, como lo es la metodología propuesta por Tripathi, en la que estrategias de aprendizaje de máquina se aplican sobre características específicas de distintas redes de interacción, con el fin de proponer módulos de interacción que puedan estar asociados con el inicio o avance del cáncer colorrectal.

La aplicación de estrategias complementarias para la detección de nuevos módulos y nuevos genes para cáncer colorectal puede ser un avance muy importante en el campo de la biología y la oncología, puesto que una mejora en la identificación de estos módulos, acoplada también la identificación de proteínas fundamentales para la fisiología de la enfermedad, permitiría un mejor entendimiento de patrones de interacción molecular que potencialmente podrían ser asociados a terapias para la enfermedad.[15].

1.4. Delimitaciones y Alcances

Considerando que el desarrollo de este proyecto está basado en las teorías de redes, algoritmia y aprendizaje de máquina, el alcance del mismo está delimitado por los límites mismos de la biología de sistemas, la biología integrativa, y los de las aplicaciones directas de la matemática y de la ingeniería de sistemas a la resolución de problemas biológicos.

De esta forma se quiere demostrar el éxito de integrar distintos tipos de información biológica, tales como redes de interacción proteína-proteína, y distintos tipos de funcionalidades biológicas, asociadas a los genes y proteínas que participan del proceso de carcinogénesis, en una estrategia metodológica basada en aprendizaje de máquina para postular nuevas proteínas y módulos de interacción asociados al inicio o desarrollo del cáncer de colon.

En el futuro, se espera la validación experimental de estos nuevos elementos postulados.

1.4.1. Entregables

- Implementación de los algoritmos de detección de comunidades aplicados a la red con su código fuente. (*Deteccion_de_comunidades.py*, *Louvain.gephi*)
- Implementación de los algoritmos de aprendizaje de máquina aplicados a la red con su código fuente. (*SMOTE_GSMOTE.ipynb*, *PU_Learning.ipynb*)
- Lista de proteínas y módulos posiblemente asociados a cáncer colorrectal. sus implicaciones funcionales y verificación sobre si los genes encontrados ya han sido asociados a un proceso de cáncer. (*GENE_LISTS.xlsx*)

Capítulo 2

Investigación de la Literatura

2.1. Áreas Temáticas

En esta sección se presentan las categorías relacionadas al proyecto.

- *Computación aplicada → Vida y ciencias médicas → Biología Computacional → Genómica Computacional, Reconocimiento de genes y elementos reguladores, Redes Biológicas, Análisis de secuencias moleculares.*
- *Computación aplicada → Vida y ciencias médicas → Bioinformática.*
- *Metodologías de Computación → Aprendizaje de máquinas → Algoritmos de Aprendizaje de máquinas.*
- *Matemáticas de la Computación → Matemáticas discretas → Teoría de grafos → Hipergrafos, Algoritmos de grafos.*

2.1.1. Trabajos Relacionados

- Paul *et al* [16]. En este artículo se propone un algoritmo llamado μSim el cual permite la identificación de miRNAs con funcionalidades similares asociados al CCR (*Colorectal cancer*). Se hicieron pruebas en cuatro grupos de datos distintos referentes al cáncer colorrectal las cuales dieron como resultado que los miRNAs seleccionados por el método propuesto son más significativos. A su vez, se dieron a conocer más miRNAs los cuales eran funcionalmente similares a los que están asociados con el cáncer colorrectal.
- Tripathi *et al* [14]. En esta investigación se propone un algoritmo para identificar y extraer módulos de enfermedades en respuesta al desafío DREAM, el cual planteaba el problema de identificar dichos módulos de seis redes heterogéneas de proteínas/genes. El algoritmo resulta de solapar las diferentes técnicas que habían utilizado otros competidores en el desafío con una serie de modificaciones adicionales. Este algoritmo obtiene los resultados más acertados en comparación a los demás.
- Deng *et al* [17]. En este artículo se describe un método para identificar genes clave involucrados en la expresión de cáncer de hígado. Para identificar dichos genes con precisión se integraron múltiples conjuntos de datos de genes en micro-arreglos, que resulta ser más preciso que al

hacerlo con conjuntos de datos individuales. Se construyó una red de interacción proteína-proteína basada en esos genes para identificar los genes clave combinando la influencia que tienen de forma local y global en la red. Los resultados de este estudio sugieren que el método descrito puede usarse en otros tipos de conjuntos de datos para encontrar genes clave en otras enfermedades.

- Peng Ni *et al* [18]. En este estudio se propone un método llamado *ModuleSim*, el cual mide las asociaciones entre enfermedades utilizando datos de asociación enfermedad-gen e interacciones en redes proteína-proteína (PPIN) basado en la teoría de módulos de enfermedades. La cuantificación de estas asociaciones ha sido utilizada en varios estudios, siendo uno de ellos la predicción de genes que estén relacionados con enfermedades. Como resultados se demostró que el método propuesto supera a otros cuatro métodos populares, que a su vez es capaz de encontrar asociaciones potenciales entre enfermedades.

- Cheng Peng *et al* [19]. En este articulo se propone un método de aprendizaje profundo para el descubrimiento de genes relacionados con el cáncer de mama mediante el uso de una red de cápsulas basada en datos multiómicos (CapsNetMMD). Como resultado, el método propuesto identificó genes relacionados con el cáncer de mama con un mejor rendimiento que otros métodos propuestos.

- Han *et al* [15]. En esta investigación se propone un modelo de aprendizaje de máquina para identificar genes conductores asociados al cáncer llamado *DriverML*. Esta aproximación se desarrolló integrando la prueba de puntaje de Rao y aprendizaje automático supervisado. Basado en una prueba de rendimiento rigurosa entre *DriverML* y otras 20 herramientas existentes en más de 30 conjuntos de datos independientes, se demostró que este modelo obtiene resultados más acertados en comparación a los demás, siendo una herramienta más robusta y poderosa, con un mejor balance entre precisión y sensibilidad al momento de detectar genes conductores.

- Lin *et al* [20]. En este estudio se desarrolla una matriz de factorización corregularizada para identificar asociaciones entre enfermedades y lo que se conoce como lincRNAs, integrando la expresión genética de los lincRNA, la red de interacción genética para genes mRNA y las asociaciones entre enfermedades y genes conocidas. La regularización que se hace no solo preserva la estructura topológica de la red coexpresión de los lincRNA, sino que también mantiene el orden entre los lincRNA, el gen respectivo y su enfermedad.

Capítulo 3

Marco Teórico

3.1. Biología de Sistemas

La biología de sistemas tiene sustento en un movimiento científico surgido en la década de los 50s, el cual, propuso un nuevo paradigma basado en la integración de las ciencias. El proyecto sistémico, intentaba una reorientación científica tendiente a lograr la integración del conocimiento en una unidad armoniosa. El gestor de este nuevo paradigma fue el biólogo Austriaco Ludwig von Bertalanffy quien en su libro, "la teoría general de los sistemas" postuló los principios básicos que argumentaban esta percepción holística de pensamiento [21].

La perspectiva sistémica surgió en contraposición a la perspectiva mecanicista que dominó el mundo científico desde el renacimiento. Bajo el paradigma mecanicista, la realidad se concibe como un compuesto de piezas básicas o elementales que sostienen una interacción mecánica para realizar algún proceso. De esta manera, conociendo las propiedades de las piezas individualmente y su interacción inmediata con otras piezas, es posible acceder al entendimiento del sistema. En un contexto biológico, al conocimiento del funcionamiento celular se puede acceder mediante el entendimiento de los componentes celulares. Bajo esta premisa, conociendo la función particular de un gen podríamos acceder a las dinámicas del proceso en el cuál dicho gen está implicado. Por ejemplo, en el caso del estudio del cáncer, se hace referencia a genes particulares denominados oncogenes, los cuales son moléculas puntuales, que al alterarse funcionalmente, inducen el inicio o progreso del proceso carcinogénico. En este contexto, se trata de explicar al cáncer como un estado alterado que subyace en el mal funcionamiento de un componente celular particular. Sin embargo, los ejemplos biológicos en los cuales es posible atribuir un fenotipo específico a la acción directa de una molécula particular son escasos.

En contraposición a esta perspectiva reduccionista, la teoría de sistemas propone que la funcionalidad de un sistema, sólo puede ser explicada por la acción coordinada en el tiempo y el espacio de todos sus componentes. Estos procesos de interacción, generan dinámicas específicas que generan propiedades emergentes que no poseen sus partes separadamente. La acción de una parte dentro del sistema sólo puede ser entendida cuando dicha parte encaja en un todo perfectamente engranado [22].

La aplicación directa de la perspectiva sistémica a la biología se conoce como biología de sistemas y define que la funcionalidad celular surge como consecuencia de la interacción precisa y coordinada de los distintos componentes celulares. Las funciones específicas que desempeña una

célula particular son propiedades emergentes, derivadas de la dinámica de interacción de los componentes celulares. Por lo tanto el entendimiento de las dinámicas que determinan la funcionalidad celular sólo cobra sentido en un contexto holístico y no pueden ser accedidas mediante un entendimiento aislado de los distintos componentes del sistema [23].

3.2. Cáncer como enfermedad multifactorial

El estudio de las características y de las dinámicas moleculares que son inherentes a la transformación neoplásica ha sido exhaustivo. En este sentido, distintas técnicas moleculares que parten de las aproximaciones tradicionales basadas en las denominadas *forward* y la *reverse genetics* y que en los últimos años han llegado al análisis genómico, transcriptómico, metabolómico y proteómico de la célula cancerígena, han contribuido ampliamente al entendimiento de las dinámicas celulares que gobiernan la enfermedad. Sin embargo, pese a la variedad de enfoques, aproximaciones experimentales y técnicas utilizadas, son limitados los casos en los que un determinado elemento molecular clave y el entendimiento de sus dinámicas en células sanas y enfermas, se han traducido en una herramienta útil para combatir la enfermedad. De la misma manera, cuando dichos logros son alcanzados, la funcionalidad de la herramienta caracterizada es aplicable a un único o a un grupo limitado de cánceres. Gran parte de la dificultad de encontrar elementos moleculares que puedan asociarse directamente no a uno, sino a una gran mayoría de cánceres se basa en la ausencia de enfoques sistémicos que permitan el entendimiento del proceso carcinogénico como una dinámica holística de interacciones complejas. En este sentido, durante los últimos años, el desarrollo de técnicas moleculares holísticas, asociadas a procesos de simulación y a procesos de construcción y análisis de redes han permitido acceder al proceso de transformación neoplásica desde una perspectiva integrativa [24].

Si bien, el proceso de carcinogénesis es altamente heterogéneo, regido por eventos moleculares propios de cada tipo de cáncer, la célula cancerígena es distintivamente caracterizada por su pérdida de identidad celular, en el marco de un contexto tisular específico. Esta característica propia, presupone la existencia de mecanismos concretos que regulan la transformación neoplásica. Douglas Hanahan y Robert Weinberg, propusieron diez marcas distintivas del proceso carcinogénico (*cancer hallmarks*), las cuales son procesos convergentes característicos de la transformación cancerígena. Dichos procesos moleculares, son comunes sino a todos, a la mayoría de cánceres [25][26]. Los procesos propuestos por Hanahan y Weinberg, representan una serie de eventos que son típicos de la célula malignizada y por lo tanto, el entendimiento de las dinámicas moleculares asociadas a dichos eventos convergentes, permitiría acceder al esqueleto molecular que subyace en el centro del proceso de carcinogénesis, facilitando el entendiendo los procesos de transformación neoplásica desde la unidad y no subdividiéndolo en eventos moleculares independientes y aislados, asociados a canceres particulares.

3.3. El cáncer entendido como una red de interacción

El cáncer puede ser entendido como un conjunto específico de funciones celulares que al alterarse desembocan en la formación de una célula independiente, asilada de las señales de control que determinan su destino celular. En el campo de la bilogía, y más específicamente en la biología molecular, se han llevado diversos estudios los cuales proponen distintos puntos de vista para poder estudiar esta enfermedad, entre ellos se encuentra la célula vista como un conjunto de interacciones moleculares, entre las cuales se encuentran las interacciones proteína-proteína, las cuales determinan el interactoma funcional de la célula en la que se determinan dichas interacciones.

Se define como interactoma, al conjunto total de relaciones físicas directas entre el compendio de proteínas que ejecutan su acción celular, en un momento dado y bajo unas circunstancias celulares, fisiológicas y ambientales determinadas. Sobre el mapa de interacciones físicas de proteínas en humanos (interactoma humano) es posible identificar y mapear, las distintas proteínas que son afectadas, alteradas, hiperactivadas o suprimidas en el contexto de alguna enfermedad. De tal manera que, al ejercer un cambio sobre el interactoma, debido a la ausencia o cambio de una proteína específica, se puede apreciar cualitativamente y calcular cuantitativamente el efecto de dicho cambio sobre una parte o sobre el patrón global del interactoma analizado. El interactoma humano puede representarse en forma de una red de interacción la cual puede representarse en forma de grafo, siendo una pareja de conjuntos $G = (V, E)$, siendo V el conjunto de vértices y E el conjunto de aristas formando parejas de vértices de la forma (u, v), donde u, v $\in$V.

En el contexto de la teoría de grafos, las redes poseen principios y propiedades particulares. Para esta investigación, las redes analizadas están definidas por el interactoma proteína-proteína construido experimentalmente para la célula humana, en donde los nodos representan proteínas y los arcos que conectan a dichas proteínas, representan interacciones físicas directas. Bajo este escenario, a continuación se definen algunas propiedades que se deben tener en cuenta a la hora de trabajar en las redes biológicas, dada su significancia.[27]:

- **Módulos:** Un conjunto de nodos que son representativos de una característica funcional, agrupándose en una zona local.
- **Grado de distribución y *hubs*:** Las distribuciones en las redes siguen un principio llamado la distribución de Poisson, el cual está dado por $P(k)$, donde el grado k sigue la forma $P(k) \sim k^{-y}$, donde y es llamado un grado de exponente. Como consecuencia de esta propiedad, se puede notar en diferentes estudios como existe una gran cantidad de *hubs* los cuales interconectan casi toda la red.
- **Fenómeno mundo-pequeño:** La propiedad mundo-pequeño, o *small-world* en inglés, define que existen caminos relativamente cortos entre cualquier par de nodos. Lo que implica que cualquier perturbación en la red puede afectar la manera como se comunican los genes.
- **Comunidad:** Se dice que una red tiene una estructura de comunidades si los nodos se pueden

agrupar fácilmente, incluso de forma superpuesta. Es más probable que un par de nodos estén conectados si pertenecen a la misma comunidad.

- **Centralidad de intermediación:** Se define como una medida del número de caminos más cortos que van por cada nodo. Los nodos con una alta centralidad de intermediación son llamados cuellos de botella.

Gracias a la capacidad de representar interacciones moleculares en forma de grafos y luego de analizar metódicamente algunas de sus características topológicas, se ha podido entender como distintas enfermedades pueden ser entendidas como procesos y dinámicas de interacción entre distintos componentes celulares, lo que nos lleva a utilizar las redes de interacción (interactomas) como una herramienta para encontrar nuevas proteínas asociadas a distintos tipos de enfermedades, particularmente, el cáncer.

3.4. Cáncer de colon

Más que una única enfermedad, el proceso de carcinogénesis refleja una serie amplia y diversa de cambios moleculares, celulares y fisiológicos que se traducen en un complejo grupo de patologías. En el 2020 se reportaron cerca de 20 millones de casos de cáncer alrededor del mundo, de los cuales un 10 % corresponde al cáncer de colon [1]. El cáncer de colon es el que se deriva de la transformación neoplásica en el intestino grueso o el recto. También se le conoce como cáncer colorectal agrupando los dos tejidos en donde es más frecuentemente originado, además de compartir múltiples características fisiológicas y moleculares que permiten tenerlos como equivalentes [28].

La transformación neoplásica del intestino, se origina a partir de un grupo de células que empiezan procesos de división descontrolados, que están en la capacidad de evadir señales antiproliferativas o evitar la muerte celular programada. Estos grupos de células forman estructuras conocidas clinicamente como pólipos que, al detectarse a tiempo, son frecuentemente benignos permitiendo su extirpación sin implicaciones negativas para el paciente. Sin embargo, si los pólipos no son removidos y persisten en el tiempo, algunos pueden originar tejido tumoral, característico de la transformación neoplásica y evolucionar a adenocarcinomas de colon con potencial invasivo que frecuentemente comprometen la vida de la persona afectada.

De acuerdo con el modelo clásico de carcinogénesis del cáncer de colon, se requieren varios cambios genéticos para la iniciación del crecimiento tumoral y progresión de la enfermedad. El desencadenante genético más frecuentemente asociado al inicio y desarrollo del cáncer de colon es la inactivación del gen APC. A la par con esta alteración, existen otras mutaciones que conducen a la inactivación de supresores tumorales como SMAD2, SMAD4, DCC y TP53, y la hiperactivación de oncogenes como KRAS y BRAF [29].

Desde el punto de vista funcional, una de las principales características asociadas al desarrollo del cáncer de colon es la inestabilidad genómica, un proceso complejo en el que convergen diversas

rutas asociadas a la repartición equitativa del material genético y que al verse alteradas, se traducen en fragmentaciones cromosómicas, retrasos anafásicos y disyunciones anómalas de las cromátides hermanas que originan células aneuploides.

3.5. Clasificación funcional de genes asociados a cáncer

Dado que el cáncer es una enfermedad multigénica, es decir, su origen y progresión depende de miles de genes cuya alteración funcional converge en la transformación neoplásica de la célula, es posible clasificar dichos genes y a las proteínas que son codificadas por éstos, en distintas categorías funcionales, según la perspectiva y el efecto de su acción, tanto en una célula fisiológicamente normal, como en una célula transformada.

A grandes rasgos, existe una primera división que agrupa en dos categorías los genes y las proteínas codificadas por éstos, según su efecto directo sobre el proceso de carcinogénesis. Así, tenemos los supresores de tumores, que son genes que al expresarse codifican para proteínas que en su acción evitan el proceso de transformación neoplásica o protegen de la iniciación del estado tumoral. Éstos son por ejemplo proteínas que reparan el ADN cuando existen daños sobre las hebras, proteínas que detienen el avance del ciclo celular o que inducen la muerte celular programada. En contraposición, están los oncogenes que son genes que al alterarse mutacionalmente, codifican para proteínas anómalas que ayudan o contribuyen al proceso de transformación neoplásica. Son proteínas que, al alterarse funcionalmente, contribuyen a la inducción del ciclo celular, incluso cuando no hay condiciones fisiológicas para que la célula avance a través de éste o que evitan la inducción del proceso apoptótico cuando la célula, debido a sus daños genómicos o estructurales, debería entrar a su proceso de muerte.

Estas dos categorías son las más amplias y la mayoría de los genes y proteínas asociadas al proceso de carcinogénesis, pueden dividirse como oncogenes o supresores de tumores. Pero existen otras categorías funcionales, más específicas, que también permiten reconocer ciertas propiedades o funcionalidades de los genes y proteínas asociados a la transformación neoplásica. Para la realización del presente trabajo, consideramos las siguientes categorías:

- **Genes canónicos implicados en cáncer:** contiene a los genes y a las proteínas codificadas por estos que por evidencia experimental múltiple y sistematizada son reconocidos por la comunidad científica como genes y proteínas directamente implicados en el proceso de carcinogénesis. Esta lista se actualiza constantemente y está disponible en `http://ncg.kcl.ac.uk/citation.php` y en las publicaciones asociadas, de la cual la más reciente es la propuesta por Dressler y colaboradores en el 2022 [30].
- **Genes candidatos asociados a cáncer:** la lista de genes está formada por genes con investigaciones iniciales que los vinculan al proceso de carcinogénesis, sin embargo, al ser

propuestos como tales por pocas publicaciones, requieren de un mayor esfuerzo investigativo para asociarlos de una forma causal a los distintos eventos de transformación neoplásica. Al igual que el grupo precedente, Dressler y colaboradores los curan y agrupan en su recurso web [30].

- **Genes asociados a las marcas distintivas del cáncer (*cancer hallmarks*):** A través de los últimos 25 años, Robert Weinberg y Douglas Hanahan, han propuesto, que, pese a la diversidad de carcinomas reportados, existen ciertas propiedades genómicas, fisiológicas y celulares que son comunes a los distintos tipos de cáncer [25][26]. Estas propiedades son denominadas las marcas distintivas del cáncer y cada una de ellas (son diez marcas distintivas), agrupa una serie de genes implicados en la emergencia de dicha propiedad. Este grupo de genes agrupa a todas aquellas proteínas asociadas con cada una de las diez marcas distintivas del cáncer.

- **Genes conductores en el proceso de transformación neoplásica:** Los genes, y proteínas codificados por estos, que al alterarse funcionalmente, directamente contribuyen al proceso de carcinogénesis, se les denomina genes conductores (*driver genes*) del proceso de transformación, se distinguen porque son los más frecuentemente mutados en cáncer y se distinguen de los genes pasajeros (*passenger*) que aunque son alterados en el proceso de carcinogénesis, su alteración aparece en estados muy avanzados de la enfermedad, como una consecuencia secundaria del proceso, guiada inicialmente por los genes conductores. Dietlein y colaboradores propusieron en el 2020 una lista que proponía el grupo de genes conductores en cáncer, la cual fue tenida en cuenta para la presente investigación [31].

- **Genes implicados en la propiedad de inestabilidad genómica:** La inestabilidad genómica se describe como la propensión del genoma a sufrir mutaciones y daños estructurales que comprometen la división equitativa del material genético, durante el proceso de división celular. Dentro de las diez marcas distintivas del cáncer, es la marca de inestabilidad genómica, la que es más frecuentemente observada en el cáncer de colon. Esta lista, contiene las proteínas propuestas por Knijnenburg y colaboradores que están implicadas en la emergencia de la inestabilidad genómica en la célula cancerosa [32].

- **Genes asociados al cáncer de colon:** Esta lista contiene los genes propuestos por Zhunussova y colaboradores en el 2019, que agrupa a los genes, y proteínas codificadas por éstos, que, por evidencia experimental, se asocian causalmente con la aparición de pólipos y posteriormente del surgimiento y progreso del cáncer de colon [29].

3.6. Algoritmos de detección de comunidades aplicados al entendimiento del cáncer

En los últimos años el problema computacional en detección de comunidades ha recibido una atención considerable. Detectar comunidades nos brinda la oportunidad de analizar el comportamiento que poseen los nodos por medio de sus interacciones, con esto se espera poder encontrar

comunidades las cuales compartan características comunes, atributos o inclusive relaciones funcionales que puedan brindar un mejor entendimiento de su funcionamiento. En el contexto de las redes biológicas se espera poder encontrar comunidades las cuales puedan brindar información asociadas a las interacciones que poseen las proteínas en un interactoma humano, o al mismo tiempo en el estudio de enfermedades como lo sería el cáncer se pueden detectar genes los cuales puedan estar relacionados a la carcinogénesis. A continuación se presentan los algoritmos de comunidades los cuales se utilizaron en esta tesis.

3.6.1. Algoritmos de comunidades modulares

3.6.1.1. Algoritmo de propagación de etiquetas

El algoritmo algoritmo de propagación de etiquetas asigna inicialmente una etiqueta única a cada nodo, para luego, repetidamente, asignar la etiqueta que aparece más repetidamente entre los nodos vecinos a cada nodo. Las comunidades se asumen formadas y el algoritmo termina cuando todos los nodos cumplen esta condición. Este algoritmo es probabilístico, por lo que sus resultados pueden variar en cada ejecución. [33].

3.6.1.2. Comunidades modulares codiciosas

El algoritmo de comunidades modulares codiciosas encuentra comunidades en el grafo usando la maximización de modularidad de Clauset-Newman-Moore. Empieza con cada nodo en su propia comunidad, y luego, repetidamente, junta el par de comunidades que llevan a la modularidad más grande hasta que no se pueda conseguir más. Este algoritmo puede parametrizarse para detenerse antes de cumplir esta condición y ahorrar tiempo de cómputo [34].

3.6.1.3. Comunidades de Louvain

El algoritmo de detección de comunidades de Louvain es un algoritmo de detección de comunidades usado en redes grandes. El algoritmo busca maximizar un puntaje de modularidad para cada comunidad, donde la modularidad determina la calidad de la asignación de un nodo a una comunidad en cada paso. Dicha optimización se realiza en una fase local y una fase de agregación. En la fase local, cada nodo se agrega a la comunidad que convenga más de acuerdo a la calidad de dicha agregación en términos de la modularidad resultante. En la fase de agregación se crea una red agregada a partir de la partición obtenida en la fase de movimiento local. Cada comunidad en esta partición se convierte en un nodo en la red agregada. Las dos fases se repiten hasta que la función de calidad no se puede aumentar más [35].

3.6.2. Algoritmos de comunidades soprelapantes

3.6.2.1. Algoritmo IPCA

IPCA es una versión modificada de DPClus. A diferencia de DPClus, este método se centra en mantener el diámetro de un cluster, definido como la distancia máxima más corta entre todos

los pares de vértices, en lugar de su densidad. Al hacerlo, el aspecto de crecimiento de semillas de IPCA enfatiza la cercanía estructural de un complejo proteico predicho, así como la conectividad estructural. Al igual que DPClus, IPCA calcula los pesos locales de vértices y bordes contando el número de vecinos comunes compartidos entre dos vértices. Sin embargo, IPCA calcula estos valores solo una vez al comienzo del algoritmo, en lugar de actualizarlos cada vez que se elimina del gráfico un clúster descubierto. Esto permite que se produzca una superposición natural entre los clústeres, ya que los nodos del clúster no se eliminan permanentemente del gráfico; sin embargo, también puede conducir a una gran cantidad de superposición de clústeres.[36].

3.6.2.2. Algoritmo de Angel

El algoritmo de Angel es un algoritmo de descubrimiento comunitario de abajo hacia arriba centrado en nodos. Aprovecha las estructuras de red de ego y la propagación de etiquetas superpuestas para identificar comunidades de microescala que posteriormente se fusionan en las de mesoescala, como hiperparametros se le pasaron el minimo número de comunidades a crear siendo 3, un puntaje limite(*threshold*) el cual a mayor cercanía a un valor de 1 prioriza encontrar un gran número de comunidades pero con menor tamaño, siendo en este caso con un valore de 0.7, y el grafo. Esto con el fin de no generar un gran número de comunidades las cuales sean muy pequeñas [37].

3.7. Aprendizaje de máquina y métodos de aprendizaje de máquina aplicados al entendimiento del cáncer

El cáncer ha sido objeto de estudio desde hace ya muchos años, diversas técnicas se han desarrollado con el fin de poder diagnosticar, pronosticar o incluso encontrar tratamientos para esta enfermedad. Las ultimas investigaciones han decidido tomar el camino de la computación y guiarse por el estudio de la aplicación del aprendizaje de máquina. El aprendizaje de máquina es una rama de la inteligencia artificial la cual se centra en el uso de datos y algoritmos con el fin de aprender comportamientos y replicarlos. El sistema de aprendizaje de un algoritmo de aprendizaje de máquina puede dividirse en tres partes [38]:

- **Proceso de decisión:** Generalmente los algoritmos de aprendizaje automático se usan para clasificar o predecir algo. Basado en una entrada de datos, el algoritmo produce una estimación acerca de los datos.
- **Función de error:** Una función de error sirve para evaluar la calidad de la predicción o clasificación del modelo.
- **Proceso de optimización del modelo:** Si se decide que el modelo puede ajustarse mejor a los datos en el conjunto de entrenamiento, entonces se realizan ajustes para reducir las diferencias entre el ejemplo conocido y el estimado resultado del modelo.

Gracias al aprendizaje de máquina se han podido llevar a cabo varios estudios y aplicar una gran variedad de técnicas con los cuales poder diagnosticar el cáncer en las personas, como por ejemplo

la detección de biomarcadores de ADN, siendo considerado un factor importante para la detección de cáncer, por medio de varios métodos aprendizaje de máquina se pueden detectar biomarcadores a partir de secuencias de ADN [39]. Al mismo tiempo también se puede llegar a predecir factores como variables clínicas, así como parámetros histológicos que pueden hacer parte de los conjuntos de datos de entrada, utilizados para implementar las estrategias de aprendizaje de máquina. [40]. Para el desarrollo de esta tesis se hará uso de lo siguientes algoritmos de aprendizaje de maquina.

3.7.0.1. Regresión Logística

El algoritmo de Regresión Logística ha sido conocido por su simplicidad en el funcionamiento y por la variedad de predicciones a las cuales sometida, entre ellos la clasificación de spam, compras online o incluso detección de cancer. Este algoritmo utiliza la función Logística o también conocida por el nombre de función de Sigmoide la cual es una curva en forma de **S** que puede tomar cualquier valor en un rango entre 0 y 1 pero jamás por fuera de este rango. Cuando un valor se encuentra por debajo de 0.5 se dice que corresponde a la clase cero mientras que si son superiores a 0.5 a la clase uno. La función se describe de la siguiente manera. [41]

$$f(x) = \frac{1}{1 + e^{-x}}, x \in \mathbb{R}$$

Si x tiende a menos infinito el resultado de la ecuación tiende a cero, pero si x tiende a infinito el resultado tiende a uno. Ahora bien, como se quiere identificar la probabilidad de que cierto evento ocurra con un resultado binario en base a ciertas variables independientes se parte de la siguiente formula.

$$log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Donde p es la probabilidad del evento, β es un parámetro, y x es una variable independiente

Al hacer despejes algebraicos se llega a la siguiente expresión.

$$p = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}$$

3.7.0.2. Bosques aleatorios

Debido a que se posee un conjunto de datos de gran tamaño y es bien conocido este algoritmo por dar buenos resultados independiente del tamaño del conjunto de datos es utilizado este algoritmo para la predicción de esta tesis. El algoritmo Bosques aleatorios clasifica en base a un número de arboles de decisión en varios subconjuntos del conjunto de datos, tomando la predicción del mayor número de votos de los arboles de decisión. [42]

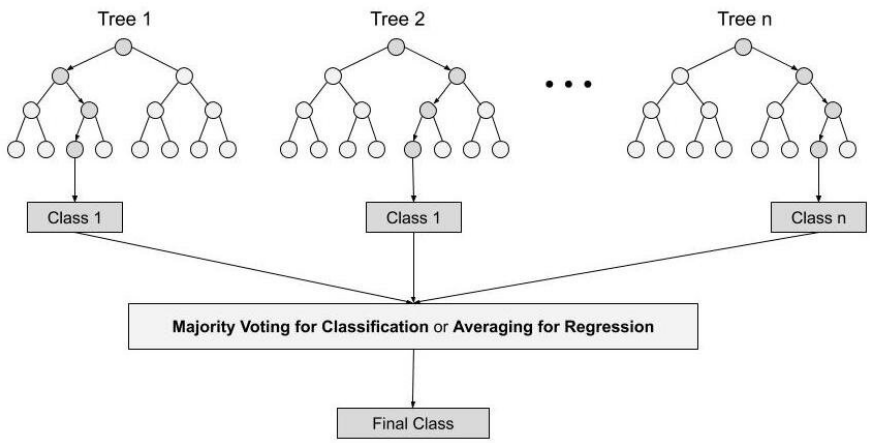


Figura 3.1: **Procedimiento Bosques aleatorios**: La figura muestra una representación de los árboles de decisión clasificando un conjunto de datos. La clase final de un dato se decide mediante votación o promediado de los resultados de cada árbol individual. [42]

3.7.0.3. K-vecinos más cercanos

El algoritmo de K-vecinos hace la suposición que puntos similares tienden a estar muy cerca de otros, esto lo hace en base a la distancia euclidiana entre dos puntos, para ello busca k puntos cercanos al punto que se quiere categorizar. Recordemos que la distancia euclidiana funciona por medio de la siguiente formula. [43]

$$d(P,Q) = \sqrt{(x_Q - x_P)^2 + (y_Q - y_P)^2}$$

Calculando la distancia euclidiana tenemos los vecinos más cercanos a nuestro objetivo, dependiendo a que clase tenga más vecinos cercanos es hacía donde se va a clasificar. Se escogió hacer una predicción con este modelo con un k entre 3 y 19 y tomar la mejor predicción y escoger el mejor k como nuestro punto objetivo para hacer el análisis de las métricas.

3.7.0.4. Análisis Discriminante Lineal

El análisis discriminante lineal o también conocido en inglés por Linear Discrimiant Analysis (LDA) hace uso del teorema de Bayes, dado la observación de un valor x cual es la probabilidad de que pertenezca a una clase k, por medio de la siguiente formula. [44]

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{P(X = x)}$$

3.7.0.5. PU Learning

Al momento de implementar un modelo de aprendizaje de máquina para clasificación, es ideal contar con un set de datos de entrenamiento en el que todos sus elementos estén etiquetados de alguna forma, de modo que las clases a tratar sean totalmente distinguibles y la naturaleza de cada elemento del set esté bien definida. Sin embargo, existen muchos escenarios en los que esto no sucede así. En lugar de ello, por ejemplo, puede existir un escenario en el que se quiera clasificar unos elementos de acuerdo a la información dada por un conjunto X en el que hay un subconjunto de datos etiquetados como positivos (que llamaremos P), mientras que los demás elementos no están etiquetados de ninguna forma (que llamaremos U). Estos datos U, al no tener una etiqueta, podrían ser positivos como negativos, y esto afecta el rendimiento del modelo de clasificación. Este problema es conocido como *PU Learning* y se diferencia del problema de clasificación supervisada estándar por la falta de ejemplares negativos en el set de datos.

Nuestro objetivo es aprender, a partir de P y U, una manera de identificar nuevos datos positivos. Así, distinguimos dos formas de *PU Learning*:

- ***PU Learning* inductivo**: En este caso, el objetivo es aprender una función que nos permita asociar un puntaje o probabilidad de que un dato nuevo $x \in X$, que puede no estar el set de datos sin etiquetar U de entrenamiento del modelo, sea positivo. Este escenario suele presentarse en sistemas diseñados para la clasificación de imágenes o documentos, donde un subconjunto de los set de datos principal es usado para entrenar el sistema, el cual debe poder clasificar nuevas imágenes fuera de dicho set de entrenamiento.
- ***PU Learning* transductivo**: En este caso, el objetivo es estimar una función de puntaje en la que solo nos interesa encontrar datos positivos en el set dado U. Este caso suele darse en problemas de clasificación de genes, donde el set completo de genes se conoce durante el entrenamiento y se clasifica entre genes de enfermedades P y el resto del genoma U, y nos interesa entonces encontrar nuevos genes de enfermedades en U.

Dada la naturaleza del problema que queremos tratar, notamos que una estrategia aplicada al *PU Learning* transductivo se adapta mejor a nuestras necesidades, pues queremos descubrir genes posiblemente implicados en cáncer colorrectal en un set de datos sin etiquetar. Basados en esta idea, se propone un método llamado *Bagging* (*Bootstrap aggregation*) que consiste en la agregación de clasificadores entrenados para distinguir elementos positivos P de una muestra del subconjunto de elementos sin etiquetar U y promediar sus resultados individuales.

En el caso particular del *PU Learning* transductivo, cada vez que se genera una muestra U_t de U, un clasificador se entrena para discriminar los datos positivos P de U_t, y se usa para asignar una probabilidad a cada elemento de U de que sea positivo. Al final, el puntaje de cada elemento x de U se obtiene promediando las predicciones de los clasificadores que no contenían a x.

Capítulo 4

Implementación

4.1. Metodología

4.1.1. Integración de distintos tipos de información biológica, asociada a la comprensión genética del origen y desarrollo del cáncer, a un mapa detallado de interacción proteína-proteína en humanos.

El interactoma humano utilizado se configuró a partir de 21 bases de datos públicas que compilan datos de interacciones proteína-proteína (PPI) obtenidos experimentalmente:

- PPI binarias, derivadas de experimentos de doble híbrido de levadura de alto rendimiento.
- PPI identificadas por purificación por afinidad seguida de espectrometría de masas.
- Interacciones sustrato quinasa.
- Interacciones de señalización.
- Interacciones regulatorias.

Para usar ciertos algoritmos de detección de comunidades en la red proteína-proteína, era necesario que las conexiones en el grafo tuvieran un peso, el cual no tenían en la red original. Para establecer los pesos de las conexiones en la red de interacción proteína-proteína, utilizamos información preexistente sobre las proteínas incluidas en el grafo: Hanahan & Weinber reportan los genes asociados a las marcas distintivas del cáncer; Dressler y colaboradores definen los genes y proteína codificadas por éstos, implicadas en el proceso de transformación neoplásica; Knijnenburg establece las proteínas implicadas en el proceso de inestabilidad genómica; Dietley y colegas establecen la lista de genes y proteínas codificantes, de los genes conductores del proceso de carcinogénesis; y Zhunussova establece las proteína hasta ahora implicadas directamente en el desarrollo del cáncer de colon. Utilizando dichas implicaciones funcionales, asignamos un valor o peso a cada nodo, considerando si dicha proteína esta relacionada con la carcinogénesis a través de una o varias categorías funcionales, así, el valor del nodo aumenta entre más implicaciones funcionales aporte al proceso de transformación neoplásica. A cada categoría se le asignó un mismo valor, a excepción de las categorías asociadas como proteínas implicadas en el proceso de desarrollo de cáncer de colon y la categoría de genes conductores, a las cuales se les asignó un mayor peso, teniendo en cuenta que la primera identifica las proteínas implicadas directamente en la transformación maligna del epitelio del colon, y la segunda define las proteínas que conducen y direccionan el proceso de transformación

neoplásica (ver tabla 4.1). A través de dichas implicaciones asignamos un valor o peso a cada nodo, de modo que este aumente dependiendo de su pertenencia a grupos determinados de genes, como se muestra en la siguiente tabla:

Tabla 4.1: Pesos por set de datos usados

Característica funcional en el proceso de carcinogénesis	Número de proteínas pertenecientes a la categoría	Referencia bibliográfica asociada a la lista de la categoría	Peso dado al nodo dentro de la categoría
Proteínas asociadas a Cáncer	591	[Dressler et al; 2022][30]	0.125
Proteínas candidatas potencialmente asociadas a cáncer	2613	[Dressler et al; 2022][30]	0.125
Proteínas asociadas al proceso de inestabilidad genómica	439	[Knijnenburg et al; 2015][32]	0.125
Proteínas implicadas en el desarrollo de cáncer de colon	85	[Zhunussova et al; 2019][29]	0.25
Proteínas codificadas por genes conductores asociados al desarrollo de cáncer	446	[Dietlein et al; 2020][31]	0.25
Proteínas asociadas a las marcas distintivas del cáncer	5215	[Hanahan and Weinberg; 2000; 2011][25][26]	0.125

Como preprocesamiento para el presente estudio se obtuvo el componente gigante más conectado de la red de interacción proteína-proteína (interactoma) propuesto por Gysi y Colaboradores [45] con el fin de obtener resultados solo basados en conexiones existentes entre proteínas y no realizar las predicciones sobre subgrafos disponibles dentro del interactoma original. Este análisis fue realizado en el modulo de Python networkx 2.6.3.

La asignación de pesos de aristas usados en la detección de comunidades y predicción de proteínas asociadas a cáncer colorrectal mediante algoritmos de aprendizaje de maquinas se realizó en Python 3.7. Como primer paso se calcularon los pesos de nodos para usarlos en el calculo de pesos de aristas mediante el siguiente procedimiento:

$$W(u) = x; x = sumw_i$$

donde u es un nodo (una proteína del interactoma)

$W(u)$ representa el peso del nodo u

x es la suma de los valores w_i

w_i representa un valor dado a u si pertenece a uno de los set de datos i (ver tabla 4.1)

Como pesos de aristas se usaron los pesos de los nodos usados previamente mediante la siguiente ecuación:

$$W(e) = W(u) + W(v)$$

donde $W(e)$ es el peso de la arista e compuesta de los nodos u y v respectivamente.

Finalmente, se obtuvieron el número de conexiones *degree*, Centralidad de intermediación *betweness Centrality* y centralidad de cercanía *closeness Centrality* para ser usados en el análisis de características topológicas del interactoma analizado y como variables predictoras usando estrategias de aprendizaje de maquina mediante Cytoscape 3.9.2.

4.1.2. Identificación de proteínas fundamentales, para el entendimiento del cáncer, a través del análisis topológico de la red de interacción construida.

Con el objetivo de tener un mejor entendimiento del cáncer en el interactoma humano se procedió a identificar proteínas las cuales tuviesen una implicación importante en el grafo, por ello se realizaron los siguientes análisis: Luego de determinar el peso total de cada nodo de acuerdo a sus características funcionales asociadas, se realizó una clasificación por pesos, determinando aquellos nodos de mayor peso y la composición de las categorías funcionales asociadas al nodo.

Utilizando el patrón de interacción determinado por el interactoma de relaciones proteína-proteína, para cada uno de los nodos, se determinó el número de conexiones realizadas por cada nodo, para identificar los nodos más conectados (hubs de red). De la misma manera se calculó el peso total del nodo, teniendo en cuenta la sumatoria total del peso de las aristas de conexión, dadas por las características funcionales evaluadas. Se estableció una correlación entre el grado de los nodos del interactoma y sus pesos asociados, según su funcionalidad dentro del proceso de carcinogénesis y, finalmente, se estableció la misma correlación pero teniendo en cuenta exclusivamente las proteínas asociadas al proceso de inestabilidad genómica en conjunto con las proteínas implicadas en la aparición del cáncer de colon.

Finalmente, utlizando los resultados del cálculo de peso funcional, determinamos el coeficiente de relación que tiene el número de conexiones de todos los nodos en la red con su peso funcional, y

luego de la misma manera para los genes de inestabilidad genómica o cáncer colorrectal en específico. Adicionalmente, Sobre la red de interacción construida se determinaron algunos descriptores básicos de la topología de la red como, grados de los nodos, grados de la red, frecuencias de distribución de grados y centralidad.

Para calcular los coeficientes de determinación y correlación de Pearson se usaron (i) El número de conexiones para el componente gigante obtenido en el paso anterior (ii) los pesos de nodos calculados $w(u)$ con base en la suma de las interacciones entre proteínas donde estuviera presente un nodo u para las proteinas identificadas como presentes en cáncer colorrectal e inestabilidad genómica. El calculo de los coeficientes de determinación y correlación fueron realizados en Microsoft Excel y R 4.1.3.

4.1.3. Identificación de módulos de interacción claves, para el entendimiento del cáncer, por medio de la identificación de comunidades, sobre la red de interacción construida

Con el objetivo de determinar los patrones de distribución de las distintas proteínas asociadas a las categorías funcionales consideradas, y obtener comunidades de proteínas modulares, es decir comunidades de proteínas las cuales tuvieran una modularidad alta (separación bien definida dentro del componente gigante del interactoma como criterio de inclusión) se usaron cuatro algoritmos: (i) Algoritmo de propagación de etiquetas (*asynchronous Label Propagation Communities*, por sus siglas en inglés). (ii) Algoritmo de comunidades modulares codiciosas con los siguientes parámetros: Sin límite en el número de comunidades a ser obtenidas (iii) Algoritmo de comunidades de Louvain. El parámetro de resolución fue conservado en 1 (resolución igual 1) para (ii) y (iii) con el fin de no sesgar la búsqueda a comunidades pequeñas (resolución menor que 1) o comunidades grandes (resolución mayor que 1) dentro de la red de interacción de proteínas. Estos análisis se realizaron usando la librería *networkx* 2.6.3 en *Python* 3.7 y el programa *Gephi*.

Con el fin de obtener comunidades que se sobrelapen unas a otras (es decir donde no se usa la modularidad como criterio de aglomeración) se usaron dos algoritmos de comunidades: (i) IPCA con los siguientes parámetros: un puntaje limite(T_in)=0.7 para no generar un gran número de comunidades las cuales sean muy pequeñas [36] y (ii) el algoritmo de Angel [37] para observar el comportamiento de las proteínas dentro del interactoma sin el efecto de la presencia de los pesos de aristas e identificar las posibles agrupaciones de los genes. Estos análisis fueron realizados en el modulo de Python cdlib.

4.1.4. Predicción de proteinas asociadas a cáncer colorectal usando métodos de aprendizaje de maquina

4.1.4.1. Preprocesamiento para clasificación

Una vez hecho el estudio de los resultados de los algoritmos en la detección de comunidades sobrelapantes y no sobrelapantes se procedió a hacer los preparamientos para hacer aprendizaje de máquina en donde se hará uso de las metodologías de clasificación por clases. Para ello se usó el flujo de trabajo *framework node2vec*[46] el cual es un algoritmo semi-supervisado para el aprendizaje de características escalables en redes. Un embebido de grafos usando *Node2vec+* [47] fue ejecutado para conseguir una matriz de 128 columnas la cual representa las características de vecindades del interactoma analizado tomando en cuenta los pesos de aristas calculados previamente en toda el interactoma. Para este fin se usaron los siguientes hiperparámetros: *PreCompFirstOrder*, para usar el precomputo de orden para acelerar el proceso de embebido del grafo, *extend weighted*, con el fin de usar la aproximación *Node2vec+*, la cual toma los pesos de las aristas del grafo calculadas anteriormente en consideración, p (parámetro de retorno) y q (parámetro de entrada y salida) con valor 1 respectivamente. El resultado del embebido del grafo fue obtenido usando el modulo pecanpy [48]. Este fue tomado como el set de datos (variable X) usado en los métodos de aprendizaje de maquinas a través de la siguiente linea de comando:

```
pecanpy --input entrada.txt --output salida.emb --mode PreCompFirstOrder
--extend --weighted --p 1 --q 1 --verbose
```

A este conjunto de predictores se le agregaron tres características topológicas para obtener una matriz de 128 columnas y 17,784 filas, la cual posteriormente fue escalada usando la desviación estandar para obtener la variables predictoras para la clasificación mediante algoritmos de aprendizaje automático.

Para la implementación de los modelos de aprendizaje de máquina en la red, se realizó primero un rebalanceo de las clases con el método SMOTE y en segundo lugar con el método G-SMOTE. Finalmente, se implementaron los mismos modelos a través de la estrategia *PU Learn* con el objetivo de mejorar los resultados del aprendizaje.

A continuación, presentamos los métodos de rebalanceo de clases y los modelos de aprendizaje de máquina usados en el proyecto.

4.1.4.2. SMOTE y G-SMOTE

Dado que el proceso de la clasificación presentó un evidente desbalance en cuanto a datos de entrenamiento (85 proteínas asociadas a cáncer colorrectal en todo el interactoma) se usaron dos estrategias para solventar el problema; (i) sobremuestreo de minorías sintéticas *(SMOTE)* [49] y (ii) sobremuestreo de minorías sintéticas geometrico *(GSMOTE)* [50]. Las cuales crean datos sintéticos usando como base la distancia euclidiana entre datos de entrenamiento p y una región geométrica determinada respectivamente con el fin de disminuir el desequilibrio de etiquetas para predicciones

de aprendizaje de maquina (es decir una categoría o etiqueta posee menos datos que otra). Los procedimientos de SMOTE y GSMOTE fueron realizados en los módulos de Python imblearn y gsmote usando los parámetros por defecto de las funciones respectivas (número de vecinos=5).

Para evaluar si la estrategia de sobremuestreo (SMOTE y GSMOTE) podrían funcionar para la clasificación de proteínas de cáncer colorrectal, cuatro algoritmos fueron ejecutados usando un 80 % de los datos para entrenamiento y 20 % para evaluación. Los algoritmos usados fueron los siguientes: (i) Regresión logística usando maximo 500 iteraciones para lograr una convergencia de resultados. (ii) Bosques aleatorios *(random forest)*, con los siguientes hiperparámetros: 1,000 estimadores y 1,000 pasos de *bootstrap*. (iii) K Vecinos más cercanos con los siguientes hiperparámetros:una búsqueda en grilla de valores entre 3 a 18 para encontrar número k de vecinos más óptimos. (iv) Análisis discriminante lineal con una tolerancia de 0.001. Los algoritmos de aprendizaje de maquinas fueron ejecutados en el modulo de Python scikitlearn 1.0.2.

4.1.4.3. PU Learning

Como ultima estrategia para la predicción de proteínas asociadas a cáncer colorrectal se utilizó una estrategia basada en *PU Learning*. Para dicho fin se usó como algoritmo de clasificación un árbol de decisiones con 10,000 estimadores, *boostraping* y el doble de datos usados para entrenamiento con etiqueta 1 (Proteínas asociadas a cáncer colorrectal con evidencia experimental) usando como datos de entrada la matriz de 131 variables predictoras escaladas obtenida en el preprocesamiento. Se implemento una estrategia en dos pasos para aumentar las probabilidades de obtener una clasificación óptima actualizando las etiquetas positivas (1) mediante las probabilidades obtenidas de correr iterativamente el arbol de clasificación durante 20 veces. La clasificación mediante *PU Learning* fue realizada en el modulo de Python pulearn 0.0.7.

4.1.4.4. Validación y predicción de los modelos de clasificación de proteínas asociadas a cáncer colorrectal

Para evaluar la idoneidad de los modelos de clasificación para predecir posibles proteínas asociados a cáncer colorrectal, se evaluó cada modelo a través de una matriz de confusión, así como el calculo de precisión, exhaustividad (*recall*) y puntaje F1 para las dos etiquetas (Proteínas que se desconoce si están asociados a cáncer colorrectal=0 y Proteínas asociadas a cáncer colorrectal con evidencia experimental=1). Las validaciones fueron llevadas a cabo en el modulo de Python scikitlearn 1.0.2.

En caso de que un modelo de clasificación tuviese un valor de *recall* igual o mayor a 0.8 para la etiqueta positiva (Proteínas asociadas a cáncer colorrectal con evidencia experimental) se procedió a considerar solo aquellas proteínas con probabilidades iguales o superiores a 0.9 como candidatas para un modelo de clasificación.

Con el fin de obtener una lista final de proteínas asociadas a cáncer colorrectal se utilizó una aproximación de ensamblaje de modelos, solo se tomaron en cuenta como proteínas candidatas

finales aquellas proteínas que tuvieran probabilidad igual o mayor a 0.9 para al menos dos algoritmos de aprendizaje automático de maquina con el fin de disminuir el número de candidatos disponibles a aquéllas proteínas que tuviesen mejor certeza estadística de la predicción.

4.2. Resultados

4.2.1. Acoplamiento de distintos tipos de información biológica, asociada a la comprensión genética del origen y desarrollo del cáncer, a un mapa detallado de interacción proteína-proteína en humanos

4.2.1.1. Genes con mayores pesos de nodo junto con las características funcionales en el proceso de carcinogénesis.

Luego de mapear los distintos tipos de información funcional asociada al desarrollo del proceso carcinogénico sobre el interactoma analizado, se obtuvieron los pesos asociados a dichos procesos de transformación, para cada nodo de la red. La tabla 4.2 muestra el top 20 de las proteínas con mayor peso. A continuación se indican las abreviaciones contenidas en la tabla:

- **Proteínas candidatas potencialmente asociadas a cáncer**: Candidatas.
- **Proteínas asociadas al proceso de inestabilidad genómica**: Inestabilidad genómica
- **Proteínas implicadas en el desarrollo de cáncer de colon**: Cáncer de colon
- **Proteínas asociadas a las marcas distintivas del cáncer**: Marcas distintivas
- **Proteínas codificadas por genes conductores asociados al desarrollo de cáncer**: Genes conductores

Tabla 4.2: Top 20 de genes con mayores pesos de nodo, se muestran Entrez ID(NCBI), nombre oficial (Alias), valor de peso $w(u)$, y set de datos en el que la proteina estaba presente. Parte 1

Rango	Entrez ID	Alias	Peso	Set de datos de orígen
1	675	BRCA2	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
2	1956	EGFR	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
3	672	BRCA1	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
4	2068	ERCC2	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
5	7157	TP53	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
6	2956	MSH6	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
7	472	ATM	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
8	9401	RECQL4	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
9	11200	CHEK2	0.875	Candidatas + Inestabilidad genómica + Cáncer de colon + Marcas distintivas + Genes conductores
10	4233	MET	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores

Tabla 4.3: Continuación Tabla 4.2: Top 20 de genes con mayores pesos de nodo, se muestran Entrez ID(NCBI), nombre oficial (Alias), valor de peso $w(u)$, y set de datos en el que la proteina estaba presente.

Rango	Entrez ID	Alias	Peso	Set de datos de orígen
11	7428	VHL	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
12	5728	PTEN	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
13	1029	CDKN2A	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
14	5925	RB1	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
15	1050	CEBPA	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
16	4771	NF2	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
17	7248	TSC1	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
18	324	APC	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
19	3815	KIT	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores
20	6598	SMARCB1	0.75	Candidatas + Cáncer de colon + Marcas distintivas + Genes conductores

4.2.2. Identificación de proteínas fundamentales, para el entendimiento del cáncer, a través del análisis topológico de la red de interacción construida

4.2.2.1. Nodos por orden de grado, peso ponderado, y conexiones ordenadas por peso

Al calcular la estadística descriptiva sobre la red, se obtuvieron los siguientes datos:

Tabla 4.4: Estadística descriptiva sobre las conexiones en la red

Métrica	Valor
Promedio de pesos y desviación estándar	0.2410 ±0,2209
Mediana y desviación alrededor de la mediana de pesos (MAD)	0.25 ±0,125
Aristas bajo el promedio de pesos	156406
Aristas sobre el promedio de pesos	165753
Aristas bajo la desviación de la mediana de pesos	66466

A través del uso de Networkx, se obtuvo la siguiente información topológica de interacción proteína-proteína. Las tablas a continuación muestran los nodos en orden de grado, peso ponderado, y por último las conexiones ordenadas por peso:

La figura 4.1 muestra una parte de la distribución del grado de conexión de los nodos en la red. Aquí se puede apreciar que los nodos tienden a tener pocas conexiones (entre 0 y 20, aproximadamente, y en contraste, los nodos que son hubs de red, dado su alto número de conexiones, representa menos del 2 % de los nodos de la red.) La gráfica con la representación completa se encuentra en el *anexo .1*

La figura 4.2 muestra que el comportamiento de la distribución de los pesos de los nodos es similar al de sus grados de conexión. El 61 % de los nodos no están involucrados en ningún set de datos y por ende no tienen peso (peso 0). Los nodos de gran peso (más de 0.625) representan solo el 0.4 % de la red. La información completa sobre este apartado se encuentra en el *anexo .2*

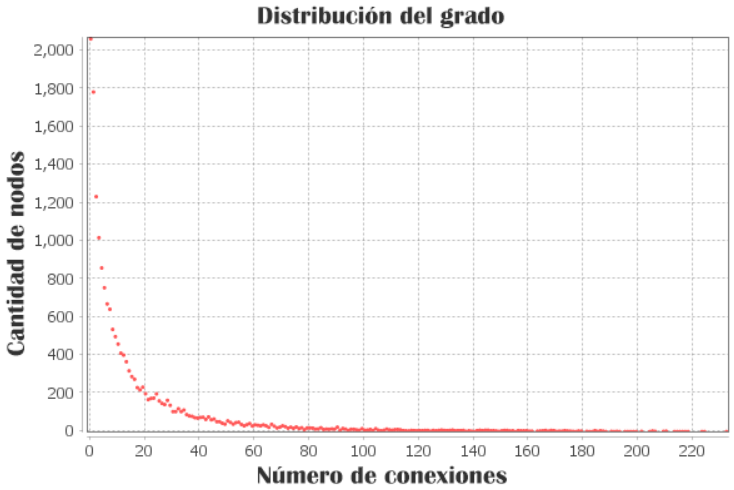


Figura 4.1: **Distribución de frecuencias en relación a los grados de los nodos representados en la red de interacción.** La figura muestra en el eje X el grado alcanzado por un nodo particular y el eje Y, hace relación a la cantidad de nodos que tienen un determinado grado dentro de la red. Se observa como a medida que el grado de los nodos aumenta, su presencia en la red disminuye, es decir, la red de interacción analizada, presenta muchos nodos con bajo número de conexiones, mientras que los hubs de red son escasos.

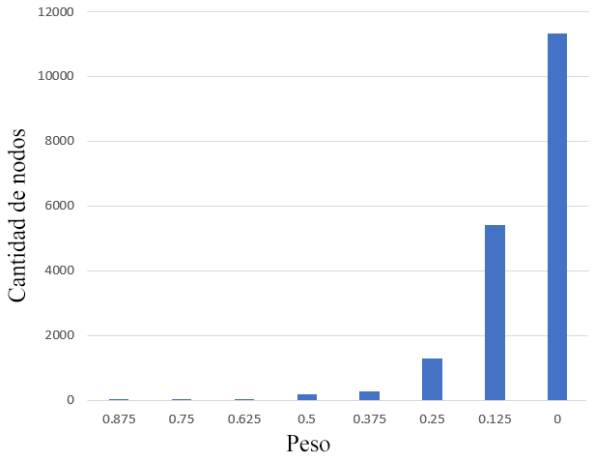


Figura 4.2: **Distribución del peso de los nodos en la red**: La figura muestra en el eje X el peso alcanzado por un nodo particular y el eje Y, hace relación a la cantidad de nodos que tienen un determinado peso en la red. La figura evidencia la alta presencia de nodos que están involucrados en en pocos o ningún set de datos.

4.2.2.2. Peso funcional de todos los nodos de acuerdo a la suma de los pesos de sus conexiones

La tabla 4.5 muestra, entre todos los genes, cuáles son los que tienen un mayor peso funcional de acuerdo a la relevancia de sus conexiones. Dicho peso funcional se halló sumando los pesos de cada conexión de un nodo.

Tabla 4.5: Top 20 de genes con mayor peso funcional. Se muestran rango, nombre oficial (Alias), peso funcional, grado del nodo

Rango	Alias	Peso funcional	Grado
1	UBC	1851.75	5198
2	TP53	798.5	762
3	EP300	788	1124
4	CTCF	769.25	1361
5	GATA2	758.5	1369
6	RAD21	686.875	969
7	EGFR	610.25	623
8	BRCA1	594.75	705
9	ETS1	552.75	1496
10	MYC	410.5	1107
11	AR	390.875	1099
12	CTNNB1	387.5	712
13	PIK3R1	370.375	572
14	ESR1	367.625	841
15	STAT3	356.75	604
16	SUMO2	321.375	850
17	EGR1	321	910
18	FOXA1	292.25	498
19	GATA1	268.5	740
20	MAPK1	259.125	462

4.2.2.3. Peso funcional de los nodos implicados en cáncer colorrectal o inestabilidad genómica de acuerdo a la suma de los pesos de sus conexiones

La tabla 4.6 muestra, entre los genes implicados en inestabilidad genómica o cáncer colorrectal, cuáles son los que tienen un mayor peso funcional de acuerdo a la relevancia de sus conexiones.

Tabla 4.6: Top 20 de genes implicados en inestabilidad genómica o cáncer colorrectal con mayor peso funcional. Se muestran rango, nombre oficial (Alias), valor de peso funcional, grado del nodo

Rango	Alias	Peso funcional	Grado
1	UBC	1851.75	5198
2	TP53	798.5	762
3	EP300	788	1124
4	GATA2	758.5	1369
5	RAD21	686.875	969
6	EGFR	610.25	623
7	BRCA1	594.75	705
8	SUMO2	321.375	850
9	E2F1	250.125	673
10	CDK2	232.75	569
11	E2F4	231.5	628
12	SUMO1	212.375	543
13	NPM1	206.5	419
14	MET	201.875	233
15	RB1	197.75	294
16	UBE2I	186.875	512
17	YY1	182.625	643
18	UBB	172.625	414
19	VHL	168.25	195
20	ABL1	165.75	413

4.2.2.4. Coeficiente de determinación entre el peso funcional del nodo y el grado del nodo de toda la red

La figura 4.3 muestra el coeficiente de determinación de toda la red entre el peso del nodo y el grado del nodo, dando un R^2 de 0.5913.

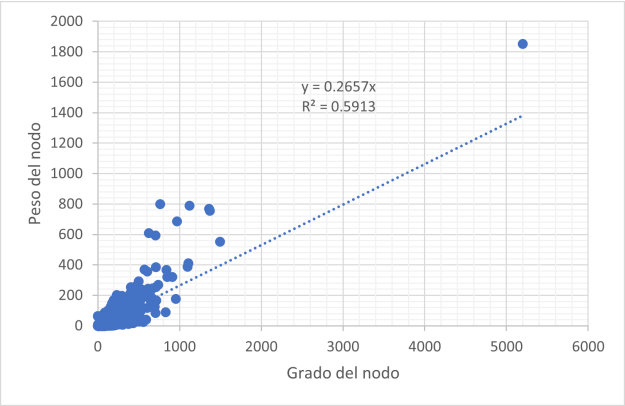


Figura 4.3: **Coeficiente de determinación de toda la red**: La figura muestra en el eje X el grado alcanzado por un nodo particular y el eje Y hace referencia al peso de dicho nodo en la red. Se observa cómo existe una parte de los nodos cuya correlación entre el grado y peso es alta, mientras que otra parte se encuentra más dispersa.

4.2.2.5. Coeficiente de determinación entre el peso funcional del nodo y el grado del nodo implicados en cáncer colorrectal o inestabilidad genómica

La figura 4.4 muestra el coeficiente de determinación de genes implicados en cáncer colorrectal o inestabilidad genomica entre el peso del nodo y el grado del nodo, dando un R^2 de 0.8666.

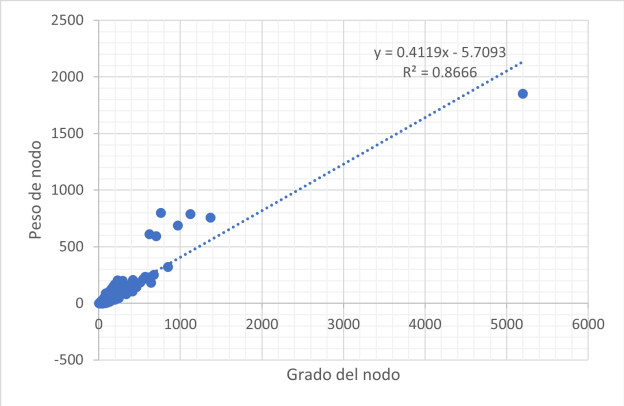


Figura 4.4: **Coeficiente de determinación de genes implicados en cáncer colorrectal o inestabilidad genomica**: La figura muestra en el eje X el grado alcanzado por un nodo particular y el eje Y hace referencia al peso de dicho nodo en la red, específicamente para los nodos implicados en cáncer colorrectal o inestabilidad genómica. La dispersión en este caso es menor en relación a la de la figura 4.3, en la que se presenta este mismo análisis pero considerando todos los nodos de la red. Esta propiedad se evidencia en un R^2 de 0.8666.

4.2.3. Identificación de módulos de interacción claves, para el entendimiento del cáncer, por medio de la identificación de comunidades, sobre la red de interacción construida

Luego de implementar los algoritmos de búsqueda de comunidades sobre el interactoma analizado, se obtuvieron 5 conjuntos de datos. La tabla 4.7 muestra el las diferentes características encontradas en las comunidades por cada algoritmo. Si se quiere ver más a detalle el resultado de las comunidades se encuentran en los *anexos .3, .4, .5, .6, .7*. A continuación se indican las abreviaciones contenidas en la tabla:

- **NCD**: Número de comunidades detectadas.
- **NC 1000**: Número de comunidades con más de 1000 nodos.
- **NC 500-999**:Número de comunidades que contienen entre 500 y 999 nodos.
- **NCD 10-99**:Número de comunidades que contienen entre 10 y 99 nodos.
- **NC-CC**:Número de comunidades con presencias de proteínas asociadas a cáncer de colon.
- **NC-CM**:Número de proteínas asociadas al cáncer de colon en la comunidad más representada por este tipo de proteínas (número total de nodos en la comunidad).

Tabla 4.7: Resultados algoritmos comunidades modulares y sobrelapantes

Algoritmo	NCD	NC 1000	NC 500-999	NC 100-499	NC 10-99	NC 1-9	NC-CC	NC-CM
ALPC	122	1	0	0	1	120	3	80(13477)
GMC	44	10	0	2	6	26	11	30(2069)
Louvain	9	7	0	1	0	0	9	30(2389)
IPCA	1709	0	0	0	0	1709	742	4(6)
Angel	128	1	0	1	7	119	3	83(12416)

A continuación se muestra una tabla haciendo referencia al grado y peso del nodo para los 30 genes asociados al cáncer de colon los cuales se obtuvieron con el algoritmo de comunidades modulares codiciosas(GMC).

Tabla 4.8: Conformación del principal modulo asociado al cáncer de colon

Alias	Peso	Grado
ATM	0.875	249
BRCA1	0.875	705
BRCA2	0.875	85
RECQL4	0.875	42
MSH6	0.875	59
CHEK2	0.875	135
TP53	0.875	762
CDH1	0.75	155
CDKN2A	0.75	176
STK11	0.75	196
TSC1	0.75	158
VHL	0.75	195
BLM	0.625	72
BRIP1	0.625	56
ERCC4	0.625	34
FANCA	0.625	86
FANCD2	0.625	75
MEN1	0.625	71
NBN	0.625	79
PMS2	0.625	39
PALB2	0.625	43
NSD1	0.625	67
WRN	0.625	42
XPC	0.625	87
BUB1B	0.5	95
EXT2	0.5	22
SLX4	0.5	31
RAD51C	0.5	22
FANCI	0.5	24
DIS3L2	0.5	6

4.2.4. Predicción de proteinas asociadas a cáncer colorectal usando métodos de aprendizaje de maquina

Presentaremos el desempeño de los modelos de aprendizaje individualmente a través de sus matrices de confusión. Después, presentaremos algunas métricas de aprendizaje automático sobre los modelos en conjunto a modo de comparación.

4.2.4.1. SMOTE

La tabla 4.9 muestra la redistribución de las clases en la red al aplicar la estrategia SMOTE sobre ella. Nótese que tras aplicar la estrategia, ambas clases tienen el mismo tamaño. Posterior a este balanceo de las clases, se aplicaron los algortimos de aprendizaje de máquina.

Tabla 4.9: Distribución de clases en la red

Clase	Tamaño
Mayoritaria (Previo a SMOTE)	3472
Minoritaria (Previo a SMOTE)	84
Mayoritaria (Posterior a SMOTE)	3472
Minoritaria (Posterior a SMOTE)	3472

4.2.4.2. Regresión logística

Tabla 4.10: Matriz de confusión de la Regresión Logística

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	148	264	412
	Negativos	2307	11509	13816

En la matriz de confusión de la regresión logística aplicada a la red, podemos ver que el modelo tiene más falsos positivos que verdaderos positivos. Exactamente un 78 % extra. La cantidad de falsos negativos es baja, en comparación, con solo un 17 % de falsos negativos respecto al total.

4.2.4.3. Bosques aleatorios

En la matriz de confusión del Bosques aleatorios aplicado a la red, podemos ver que el modelo tiene, una vez más, más falsos positivos que verdaderos positivos. Sin embargo, en este caso representan un 99 % de los datos positivos totales. La cantidad de falsos negativos es muy baja, en comparación, con solo un 0.13 % de falsos negativos respecto al total.

Tabla 4.11: Matriz de confusión del algoritmo Bosques aleatorios

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	2	410	412
	Negativos	19	13797	13816

Tabla 4.12: Matriz de confusión de K-vecinos más cercanos

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	53	359	412
	Negativos	1223	12593	13816

4.2.4.4. K-vecinos más cercanos

En la matriz de confusión de los K-vecinos más cercanos aplicado a la red, podemos ver que los falsos positivos en el modelo representan un 87 % de los datos positivos totales. La cantidad de falsos negativos es bastante baja, con solo un 8.8 % de falsos negativos respecto al total.

4.2.4.5. Análisis Discriminante Lineal

Tabla 4.13: Matriz de confusión del Análisis Discriminante Lineal

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	162	250	412
	Negativos	2802	11014	13816

En la matriz de confusión del análisis discriminante lineal aplicado a la red, podemos ver que los falsos positivos en el modelo representan un 60 % de los datos positivos totales, siendo el más bajo en los modelos aplicados. Sin embargo, la cantidad de falsos negativos es la más alta con un 20 % de falsos negativos respecto al total.

4.2.4.6. Métricas de aprendizaje automático

En la siguiente tabla se muestra los resultados de las métricas dados por los algoritmos implementados de aprendizaje de máquina. Las métricas utilizadas son Precisión, Recall, Puntaje F1 para las clases a clasificar siendo para la predicción de los 1 y para la predicción de los 0. Los resultados de los algoritmos descritos anteriormente son respectivamente Regresión Logística, Bosques aleatorios, K-vecinos más cercanos, y Análisis discriminante lineal.

Tabla 4.14: Cuadro comparativo de métricas de aprendizaje automático sobre los modelos implementados con SMOTE

Algoritmo	Clase	Precisión	Recall	Puntaje F1
Regresión Logística	Genes no asociados a cáncer (0)	0.98	0.83	0.90
Regresión Logística	Genes asociados a cáncer colorrectal (1)	0.06	0.36	0.10
Bosques Aleatorios	Genes no asociados a cáncer (0)	0.97	1.00	0.98
Bosques Aleatorios	Genes asociados a cáncer colorrectal (1)	0.10	0	0.01
K-vecinos más cercanos	Genes no asociados a cáncer (0)	0.97	0.91	0.94
K-vecinos más cercanos	Genes asociados a cáncer colorrectal (1)	0.04	0.13	0.06
Análisis discriminante lineal	Genes no asociados a cáncer (0)	0.98	0.80	0.88
Análisis discriminante lineal	Genes asociados a cáncer colorrectal (1)	0.05	0.39	0.10

4.2.4.7. G-SMOTE

La tabla 4.15 muestra la redistribución de las clases en la red al aplicar la estrategia G-SMOTE sobre ella. Nótese que tras aplicar la estrategia, ambas clases tienen el mismo tamaño. Posterior a este balanceo de las clases, se aplicaron los algortimos de aprendizaje de máquina.

4.2.4.8. Regresión logística

En la matriz de confusión de la regresión logística aplicada a la red, podemos ver que el modelo tiene la misma cantidad de falsos positivos y verdaderos positivos. La cantidad de falsos negativos es baja, en comparación, con solo un 4 % de falsos negativos respecto al total.

Tabla 4.15: Distribución de clases en la red

Clase	Tamaño
Mayoritaria (Previo a G-SMOTE)	3537
Minoritaria (Previo a G-SMOTE)	19
Mayoritaria (Posterior a G-SMOTE)	3537
Minoritaria (Posterior a G-SMOTE)	3537

Tabla 4.16: Matriz de confusión de la Regresión Logística

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	33	33	66
	Negativos	593	13569	14162

Tabla 4.17: Matriz de confusión del algoritmo Bosques aleatorios

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	27	39	66
	Negativos	403	13759	14162

4.2.4.9. Bosques aleatorios

En la matriz de confusión del modelo de Bosques aleatorios aplicado a la red, podemos ver que el modelo tiene un poco más de falsos positivos que verdaderos positivos, representando el 59 % sobre los datos verdaderos totales. La cantidad de falsos negativos es baja, en comparación, con solo un 2 % de falsos negativos respecto al total.

4.2.4.10. K-vecinos más cercanos

Tabla 4.18: Matriz de confusión de K-vecinos más cercanos

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	0	66	66
	Negativos	0	14162	14162

En la matriz de confusión del modelo K-vecinos más cercanos aplicado a la red, podemos ver que el modelo falla completamente en la clasificación de los valores positivos, mientras que acierta en todos los valores negativos.

4.2.4.11. Análisis discriminante lineal

Tabla 4.19: Matriz de confusión del análisis discrimnante lineal

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	38	28	66
	Negativos	665	13497	14162

En la matriz de confusión del modelo de análisis discriminante lineal aplicado a la red, podemos ver que tiene un mejor rendimiento en la clasificación de los valores positivos, con una porcentaje de verdaderos positivos del 57 % con respecto a los valores positivos totales. Al igual que en los demás modelos, tiene un bajo porcentaje de falsos negativos, con solo el 4 %.

4.2.4.12. Métricas de aprendizaje automático

En la siguiente tabla se muestra los resultados de las métricas dados por los algoritmos implementados de aprendizaje de máquina. Las métricas utilizadas son Precisión, Recall, Puntaje F1 para las clases a clasificar siendo para la predicción de los 1 y para la predicción de los 0. Los resultados de los algoritmos descritos anteriormente son respectivamente Regresión Logística, Bosques aleatorios, K-vecinos más cercanos, y Análisis discriminante lineal.

Tabla 4.20: Cuadro comparativo de métricas de aprendizaje automático sobre los modelos implementados con G-SMOTE

Algoritmo	Clase	Precisión	Recall	Puntaje F1
Regresión Logística	Genes no asociados a cáncer (0)	1.00	0.96	0.98
Regresión Logística	Genes asociados a cáncer colorrectal (1)	0.05	0.46	0.10
Bosques Aleatorios	Genes no asociados a cáncer (0)	1.00	0.98	0.99
Bosques Aleatorios	Genes asociados a cáncer colorrectal (1)	0.10	0.41	0.16
K-vecinos más cercanos	Genes no asociados a cáncer (0)	1.00	1.00	1.00
K-vecinos más cercanos	Genes asociados a cáncer colorrectal (1)	0.00	0.00	0.00
Análisis discriminante lineal	Genes no asociados a cáncer (0)	1	0.95	0.98
Análisis discriminante lineal	Genes asociados a cáncer colorrectal (1)	0.05	0.52	0.10

4.2.4.13. PU Learning

4.2.4.14. Regresión logística

Tabla 4.21: Matriz de confusión de la Regresión Logística

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	85	0	85
	Negativos	2728	14971	17699

En la matriz de confusión de la regresión logística aplicada a la red mediante PU Learning,

podemos ver que el modelo acierta completamente en la predicción de datos positivos con el 100 % de aciertos. La cantidad de falsos negativos es baja, con solo un 15 % de falsos negativos respecto al total.

4.2.4.15. Bosques aleatorios

Tabla 4.22: Matriz de confusión de la Bosques aleatorios

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	85	0	85
	Negativos	3736	13963	17699

En la matriz de confusión de la Bosques aleatorios aplicado a la red mediante PU Learning, podemos ver que el modelo acierta completamente en la predicción de datos positivos con el 100 % de aciertos. La cantidad de falsos negativos es moderada, con un 21 % de falsos negativos respecto al total.

4.2.4.16. K-vecinos más cercanos

Tabla 4.23: Matriz de confusión de K-vecinos más cercanos

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	36	49	85
	Negativos	168	17531	17699

En la matriz de confusión de K-vecinos más cercanos aplicado a la red mediante PU Learning, podemos ver que el modelo acierta en el 42 % de los datos positivos en la predicción. Los falsos negativos representan solo el 0.9 % respecto al total.

4.2.4.17. Análisis discriminante Lineal

Tabla 4.24: Matriz de confusión de la Análisis discriminante Lineal

		Datos reales		
		Positivos	Negativos	Total
Predicción	Positivos	85	0	85
	Negativos	2728	14971	17699

En la matriz de confusión de la Análisis discriminante Lineal aplicado a la red mediante PU Learning, podemos ver que el modelo también acierta completamente en la predicción de datos

positivos con el 100 % de aciertos. Los falsos negativos representan un 15 % respecto al total.

4.2.4.18. Métricas de aprendizaje automático

En la siguiente tabla se muestra los resultados de las métricas dados por los algoritmos implementados de aprendizaje de máquina. Las métricas utilizadas son Precisión, Recall, Puntaje F1 para las clases a clasificar siendo para la predicción de los 1 y para la predicción de los 0. Los resultados de los algoritmos descritos anteriormente son respectivamente Regresión Logística, Bosques aleatorios, K-vecinos más cercanos, y Análisis discriminante lineal.

Tabla 4.25: Cuadro comparativo de métricas de aprendizaje automático sobre los modelos implementados con PU Learning

Algoritmo	**Clase**	**Precisión**	**Recall**	**Puntaje F1**
Regresión Logística	Genes no asociados a cáncer (0)	1.00	0.85	0.92
Regresión Logística	Genes asociados a cáncer colorrectal (1)	0.03	1.00	0.06
Bosques Aleatorios	Genes no asociados a cáncer (0)	1.00	0.79	0.88
Bosques Aleatorios	Genes asociados a cáncer colorrectal (1)	0.02	1.00	0.04
K-vecinos más cercanos	Genes no asociados a cáncer (0)	1.00	0.99	0.99
K-vecinos más cercanos	Genes asociados a cáncer colorrectal (1)	0.18	0.42	0.25
Análisis discriminante lineal	Genes no asociados a cáncer (0)	1.00	0.83	0.90
Análisis discriminante lineal	Genes asociados a cáncer colorrectal (1)	0.03	1.00	0.05

4.2.4.19. Predicción de proteínas asociadas a cáncer colorrectal

A continuación, presentamos las nuevas proteínas potencialmente asociadas a cáncer colorrectal, resultado de los modelos entrenados a través de la estrategia *PU Learning*.

Tabla 4.26: Nuevas proteínas detectadas asociadas al cáncer de colon utilizando estrategias de aprendizaje de máquina

Id NCBI	Nombre oficial	Número de conexiones	Número de categorías funcionales asociadas
2113	ETS1	1496	2
1499	CTNNB1	712	3
367	AR	1099	2
1387	CREBBP	616	3
207	AKT1	537	3
5594	MAPK1	462	3
4869	NPM1	419	4
6597	SMARCA4	448	3
1874	E2F4	628	2
6938	TCF12	399	3
6714	SRC	585	2
3326	HSP90AB1	603	2
1017	CDK2	569	2
3725	JUN	588	2
7314	UBB	414	3
50943	FOXP3	663	1
2801	GOLGA2	512	2
5430	POLR2A	245	3
841	CASP8	231	3
3091	HIF1A	345	2
3320	HSP90AA1	422	1
4904	YBX1	830	0
142	PARP1	214	2
3932	LCK	221	2
5590	PRKCZ	195	2
83737	ITCH	199	2
2264	FGFR4	115	3
5777	PTPN6	175	2
6117	RPA1	154	2
6187	RPS2	202	2
8878	SQSTM1	245	1
4176	MCM7	170	2

Id NCBI	Nombre oficial	Número de conexiones	Número de categorías funcionales asociadas
9212	AURKB	235	1
84445	LZTS2	299	1
6189	RPS3A	170	2
4691	NCL	212	1
8850	KAT2B	203	1
8358	H3C2	187	1
8451	CUL4A	134	2
3727	JUND	534	0
2648	KAT2A	210	1
8968	H3C7	327	0
3297	HSF1	182	1
388324	INCA1	247	1
7507	XPA	66	3
6654	SOS1	90	2
2074	ERCC6	55	3
81620	CDT1	67	2
8453	CUL2	100	1
8065	CUL5	108	1
4173	MCM4	87	1
6921	ELOC	105	1
23368	PPP1R13B	93	1
27030	MLH3	46	3
3566	IL4R	53	2
56949	XAB2	59	2
55824	PAG1	52	2
388677	NOTCH2NLA	368	0
2355	FOSL2	160	0
50855	PARD6A	53	1
23236	PLCB1	32	1
2961	GTF2E2	109	0
2966	GTF2H2	20	2
107080638	TBC1D7	24	2

Id NCBI	Nombre oficial	Número de conexiones	Número de categorías funcionales asociadas
100499483	CCDC180	22	0
378708	CENPS	30	0
24137	KIF4A	37	1
220082	CBY2	131	0
4656	MYOG	66	0
4250	SCGB2A2	77	0
23595	ORC3	32	1
10714	POLD3	16	2
84875	PARP10	22	1
267004	PGBD3	21	0
9837	GINS1	10	1
51659	GINS2	13	0
390916	NUDT19	11	1
79786	KLHL36	19	0
55654	TMEM127	3	3

4.3. Análisis de resultados

4.3.1. Acoplamiento de distintos tipos de información biológica, asociada a la comprensión genética del origen y desarrollo del cáncer, a un mapa detallado de interacción proteína-proteína en humanos.

Aunque distintos tipos de análisis se han implementado sobre redes de interacción proteína-proteína, este es el primer estudio en el cual se integra sobre el interactoma distintos tipos de información funcional asociada al efecto y características de los distintos genes, y proteínas por éstos codificadas, en el proceso de carcinogénesis. De esta manera, aunque se conocía que tanto supresores tumorales como oncogenes específicos, resultaban ser nodos centrales y altamente interconectados en distintas redes de interacción, se desconocía, desde la perspectiva funcional, que conjunto de procesos o características biológicas decisivas podían ser influenciados por alteraciones en dichos oncogenes o supresores de tumores [51]. Por ejemplo, aunque se conocía que las proteínas codificadas por genes como BRCA2, BRCA1, TP53, RB1, PTEN o EGFR, eran nodos de red se desconocía, el conjunto de información funcional, que más allá de la misma conectividad, podría estar contribuyendo al proceso de transformación neoplásica [52]. Con el presente trabajo en el cual se complementa el grafo de interacción generado con información asociada a cada proteína, se postula que sobre estos nodos confluyen procesos y características que los asocian como factores decisivos para el progreso del cáncer, ya que son proteínas ya establecidas como participantes del proceso carcinogénico, pero que a su vez, participan activamente de las marcas distintivas del cáncer, son genes que codifican para proteínas conductoras del proceso de transformación, y además, guían el proceso de inestabilidad genómica y contribuyen a la malignización de las células del colon (tabla 4.2).

El top 20 (tabla 4.2) de los genes con mayor peso funcional (inclusive el top 100) está formado por las principales proteínas asociadas con el proceso de carcinogénesis. El hecho que la metodología planteada, identifique este grupo fundamental de moléculas gracias a los pesos asignados indica que la metodología planteada es congruente con la investigación científica en cáncer. En concordancia, las proteínas de la lista están asociadas a procesos biológicos como la regulación del ciclo celular (EGFR, TP53, PTEN, CDKN2A y RB1) y la detección y reparación de daños en el ADN (BRCA1, BRCA2, ERCC2, MSH6, ATM, CHEK2), dos de los eventos más representativos en los procesos tempranos de transformación neoplásica.

4.3.2. Identificación de proteínas fundamentales, para el entendimiento del cáncer, a través del análisis topológico de la red de interacción construida

La caracterización de los nodos en la red, teniendo en cuenta el número de conexiones generadas por éstos, muestra que los nodos de alto grado (hubs de red) son escasos; que entre más interconectado esté un nodo, es menos frecuente en la red y que aquellos nodos con un grado bajo, son los más abundantes en la red construida (figura 4.1). Este fenómeno es típico de redes biológicas

y fue descrito por Barabasi y colaboradores, en donde describe a la mayoría de las redes biológicas, incluidas las redes de interacción proteína-proteína, como redes que siguen o tienden a seguir una ley de potencias, denominándolas redes libres de escala (scale-free networks) [53][54][55]. Se ha demostrado como este tipo de redes, son más resistentes y conservan mejor las características generales de la topología de la red cuando los nodos son alterados o eliminados. En un sentido biológico, dado que los hubs de red son muy infrecuentes y el proceso de mutación es aleatoria, es más probable que alteraciones funcionales afecten nodos pocos conectados, y, por lo tanto, no se afecte la estructura de la red y la funcionalidad celular que se deriva de dicha arquitectura, se mantenga. Este tipo de estructura de red genera redes más robustas [6]. En concordancia, los efectos deletéreos sobre la estructura de la red ocurren cuando los nodos altamente interconectados son atacados, lo cual ocurre en el contexto de la enfermedad y de enfermedades específicas como el cáncer, cuando los nodos de alto grado son alterados funcionalmente. Las tablas 4.5 y 4.6 muestra las proteínas de mayor grado en la red, las cuales coinciden con moléculas implicadas en la transformación neoplásica, así, proteínas como TP53, GATA2, RAD21, EGFR, BRCA1, MYC, STAT3 o MAPK1, reconocidas como primordiales en el proceso de conversión celular, son claros hubs de red, demostrando que, aunque infrecuentes, si son alterados, pueden determinar el destino celular.

Dado que los análisis realizados nos permitieron calcular para cada nodo, tanto su grado como el peso funcional del nodo según sus interacciones, fue posible establecer qué tipo de relación podía existir entre ambas variables, así, graficamos el grado de los nodos vs el peso funcional total y establecimos el coeficiente de determinación, para saber si existe una relación de proporcionalidad entre las variables. A partir de las figura 4.6 se puede evidenciar el comportamiento presentado por el coeficiente de determinación al calcularse sobre las proteínas de toda la red, con un valor de R^2 de 0.59, implicando que en aproximadamente un 41 % de los casos el peso del nodo no necesariamente indica que sea de alta conectividad y por ende también tenga un mayor peso funcional.

Por otro lado, como se puede apreciar en la figura 4.7, cuando se toma el coeficiente de determinación exclusivamente de proteínas las cuales están implicadas en cáncer colorrectal o inestabilidad genómica, esa relación sufre un incremento notable, alcanzando un R^2 de 0.8666. Por lo tanto, cuando se limita a estas dos características, sí existe una relación considerable, implicando que específicamente en las proteínas de cáncer colorrectal o características de inestabilidad genómica sí es importante el número de conexiones.

En ese sentido, consideramos que si bien la relación entre el peso y el grado de conexiones no es tan fuerte en las proteínas en el cáncer como fenómeno general, esta relación se fortalece para el caso de las proteínas involucradas en cáncer colorrectal o en inestabilidad genómica. Esta observación es interesante, ya que los resultados encontrados señalan que en la red general de interacción existen proteínas altamente interconectadas que funcionalmente no participan del proceso de carcinogénesis. Biológicamente, esta característica se puede explicar desde la funcionalidad celular, en donde proteínas que se comportan como hubs de red participan de procesos celulares que no están implicados en la transformación maligna de la célula, pero que son fundamentales para mantener

la homeostasis celular. Por ejemplo, el Factor Nuclear Respiratorio 1 (NRF1) tiene un grado de 521, pero un peso funcional de 0, indicando que es una proteína densamente conectada, pero que su funcionalidad como factor de transcripción involucrado en la activación de proteínas respiratorias, es vital para la supervivencia celular, pero no para la carcinogénesis. Algo similar ocurre con el Factor de Transcripción Sp1 (SP1), el cual controla procesos de diferenciación y senescencia celular, fundamentales para regular distintos destinos de la célula, pero cuya alteración no conduce a la transformación neoplásica. SP1 tiene un grado de 705, pero un peso funcional de 0. En contraposición, al analizar los nodos asociados directamente con el desarrollo de cáncer de colon aunados a las proteínas implicadas en el proceso de inestabilidad genómica, existe una buena correlación positiva entre los grados del nodo y su peso funcional, indicando que para éstas proteínas, al aumentar sus conexiones también aumenta su relación funcional con la transformación neoplásica, esto quiere decir, que se conectan con nodos que a su vez participan funcionalmente del proceso de carcinogénesis, sugiriendo que el proceso de transformación en el epitelio del colon es ejecutado por una alta proporción de proteínas conductoras de la transformación celular.

4.3.3. Identificación de módulos de interacción claves, para el entendimiento del cáncer, por medio de la identificación de comunidades, sobre la red de interacción construida

Al momento de compararar los resultados en la maximización de modularidad por parte de todos los algoritmos implementados, podemos notar varios puntos. Primero, el algoritmo algoritmo de propagación de etiquetas (ALPC) presenta una comunidad que ocupa el 97 % de los nodos, mientras que las otras comunidades detectadas son definidas como comunidades muy pequeñas. Este comportamiento se puede ver de igual manera con el algoritmo sobrelapante de Angel, esto debido a que utiliza Label Propagation como principio para la detección de las comunidades, explicando así que presente un comportamiento similar Y una complejidad algoritmica similar al de ALPC (aproximadamente $O(n)$) [37]. Cabe resaltar que ALPC considera el peso de los nodos del grafo en el cálculo y definición de sus comunidades, mientras Angel no. Es importante resaltar que en esta gran comunidad la cual ocupa la mayoría de nodos en el interactoma se encuentran los nodos asociados a cáncer colorrectal, dando a entender que estas proteínas llegan a estar presentes e interactúan en comunidades de gran tamaño.

Por otro lado el algoritmo de comunidades modulares codiciosas tiene una complejidad de $O(nlog^2n)$ [34], y si bien es computacionalmente más pesado, este presenta resultados muchos más favorables en cuanto a la creación de comunidades modulares se refiere. Las proteínas relacionadas a cáncer colorrectal se encuentran uniformemente distribuidos en las primeras comunidades las cuales se componen de una gran cantidad de nodos. Esto se puede interpretar como que estos genes relacionados a cáncer llegan a participar en varias comunidades con gran magnitud de interacciones y no necesariamente solo en una comunidad como se muestra en los anteriores dos algoritmos.

Por otra parte, en el algoritmo IPCA se observa una gran cantidad de comunidades detectadas,

con la gran diferencia que estas son muy pequeñas en comparación a los demás algoritmos analizados. Estas comunidades se componen de 2 a 6 nodos, probablemente debido a el valor utilizado en el hiperparámetro T_in, el cual es de 0.7, priorizando así un gran número de comunidades pero con menor tamaño [36].

Se ha demostrado que las funciones celulares también están organizadas en un sistema altamente modular. donde cada módulo es un objeto discreto compuesto por un grupo de componentes estrechamente vinculados que realiza una tarea relativamente independiente. Se ha sugerido como esta modularidad en las funciones celular, surge a su vez, de la modularidad en la interacción molecular de redes biológicas como las redes reguladora de la transcripción y las redes de interacción proteína-proteína [56].

En una red de interacción proteína-proteína los módulos corresponden a complejos proteicos o grupos funcionales de proteínas. Los módulos funcionales participan en un proceso celular particular o interactúan para ejecutar una misma función biológica [57]

Algunos estudios de interacción proteína-proteína, demuestran que dependiendo del tipo celular los módulos funcionales que controlan la fisiología de una célula, pueden variar desde algunas decenas hasta cientos de módulos [58][59]. En este sentido, de los algoritmos evaluados, los que mejor se acercan a esta distribución, fueron los algoritmos de ALPC, GMC y Angel (tabla 4.7). Sin embargo, tanto el algoritmo de ALPC como el de Angel se caracterizan por haber detectado múltiples comunidades con un número muy bajo de nodos, que no podrían sustentar una función biológica. De la misma manera, ambos algoritmos detectaron una única macro-comunidad, que agrupa la mayoría de nodos de la red, presentando una súper comunidad multifuncional, que no describe el comportamiento modular de los sistemas biológicos. En contraste, el algoritmo de GMC, presenta una distribución más balanceada de comunidades, en donde existen diversas comunidades conformadas por más de mil nodos, que podrían sustentar una funcionalidad biológica, pero también detectó distintas comunidades con un número variado de nodos (tabla 4.7), reflejando una distribución más modular en relación con los otros algoritmos analizados. Por lo tanto, los resultados obtenidos luego de la implementación de este algoritmo son los que mejor reflejan la modularidad funcional de las redes biológicas de interacción.

Uno de los objetivos principales del presente trabajo de investigación consistía en proponer circuitos génicos o módulos asociados al desarrollo del proceso de carcinogénesis. Teniendo en cuenta los resultados obtenidos luego de la aplicación del algoritmo de GMC, proponemos el módulo conformado por 30 genes asociados al proceso de cáncer de colon, embebidos en la comunidad conformada por 2069 proteínas (tabla 4.8), como el principal módulo que podría asociarse causalmente al origen o desarrollo del cáncer de colon.

Es evidente, luego de la implementación de los distintos algoritmos para la detección de comunidades que las proteínas directamente asociadas al origen o progreso del cáncer de colon, no se

ubican en una única comunidad, en contraste, se distribuyen en distintas comunidades. Por ejemplo, luego de aplicar el algoritmo de GMC, las proteínas asociadas al cáncer de colon se distribuyen en 11 de las 44 comunidades detectadas. De la misma manera, existen 742 comunidades, de las 1709 detectadas por el algoritmo de IPCA en donde hay presencia de proteínas asociados al cáncer de colon. Esta distribución de proteínas en diversas comunidades es característico de los sistemas biológicos para mejorar su robustes, en donde una determinada función es distribuida en módulos funcionalmente equivalentes, pero estructurados por grupos de proteínas diferentes, de tal manera, que al alterarse funcionalmente una proteína clave del módulo, que anule la función celular soportada por éste, dicha función puede ser soportada y sostenida por los otros módulos asociados a dicha función evitando el colapso de la célula [54].

4.3.4. Predicción de proteínas asociadas a cáncer colorrectal usando métodos de aprendizaje de maquina

Considerando que existía un desbalance considerable en los datos con respecto a las clases que queríamos trabajar en este problema de clasificación y que esto podría generar un sobreajuste en las predicciones de los diferentes modelos, utilizamos las estrategias de sobremuestreo SMOTE y GSMOTE para intentar balancear los datos. Sin embargo, como lo evidencian los resultados arrojados por los diferentes modelos que usaron los set de datos sobremuestreados, no se logró conseguir valores satisfactorios de exhaustividad para la etiqueta positiva (proteínas asociadas a cáncer colorrectal) en ninguno de los cuatro modelos, por lo que no fueron considerados para la proposición de proteínas nuevas.

Por otro lado, al utilizar como alternativa la estrategia de clasificación de *PU Learning*, el valor de la exhaustividad incrementó considerablemnte para la etiqueta deseada. Esto debido a que esta estrategia se beneficia de sets de datos "poco contaminados", refiriéndose a que la proporción de valores positivos respecto al total de datos es baja, y ese es el caso de el set de datos utilizado en esta investigación. [60]. Si bien decidimos utilizar estos modelos para predecir nuevas proteínas potencialmente asociadas a cáncer colorrectal, es importante resaltar que para ganar exhaustividad, los modelos parecen sacrificar algo de precisión, lo cual podría afectar la calidad de las predicciones obtenidas en el estudio.

Considerando lo anterior, proponemos 79 proteínas candidatas que podrían estar potencialmente asociados al desarrollo de cáncer de colon (tabla 4.26). De estas proteínas, el 50 % están asociadas tan solo a un par de características funcionales utilizadas en la estructuración del proyecto, pero más interesante aún, es que 14 de las nuevas proteínas detectadas no tienen ninguna implicación funcional asociada, las que las transforman en candidatas ideales para corroborar su función y su participación del proceso de transformación del epitelio del colon experimentalmente. En este sentido, literatura reciente ha asociado a NOTCH2NLA con el desarrollo de adenocarcinoma de ovario [61] y a GINS2 directamente con el cáncer de colon [62].

Capítulo 5

Conclusiones

Basados en el análisis topológico realizado sobre la red de interacciones entre proteínas, demostramos que la red posee una cantidad limitada de proteínas con un alto número de conexiones que aparecen tener una correlación moderada con los valores de pesos de nodos (posible participación en procesos cancerígenos). Es decir, para todo el interactoma se sugiere que hay proteínas que no necesariamente que poseen un alto número de conexiones pero que participen en diversos tipos de cáncer. En relación a esto y a pesar de que la evidencia muestra una correlación moderada entre el número de conexiones y los pesos de nodos asociados a la participación en procesos cancerígenos, para el caso de cáncer colorrectal e inestabilidad genómica se sugiere que estas proteínas tienen un rol como reguladores de diversas rutas metabólicas e interacciones basado en el alto valor de coeficiente de determinación observado.

En cuanto a la modularidad en la red, encontramos que el uso de algoritmos de comunidades no demostró ser del todo adecuada para la predicción de proteínas, ya que fue indistinguible la consecución de comunidades con una predominancia de proteínas catalogadas experimentalmente para cáncer colorrectal en la mayoría de algoritmos implementados. Sin embargo, luego de aplicar el algoritmo de GMC, logramos proponer el módulo conformado por 30 genes asociados al proceso de cáncer de colon, embebidos en una comunidad conformada por 2069 proteínas, como el principal módulo que podría asociarse causalmente al origen o desarrollo del cáncer de colon.

Considerando que el uso de las estrategias de sobre-muestreo *SMOTE* y *GSMOTE* no demostró ser adecuado para la tarea de predicción de proteínas ya que ninguna de los tres algoritmos de clasificación implementados demuestran un valor de exhaustividad (*recall*) alto para la etiqueta 1 (Proteínas asociadas a cáncer colorrectal), preferenciamos el uso de las estrategias basadas en *PU Learning*, las cuales demostraron su éxito para la predicción de proteínas asociadas a cáncer colorrectal. Esto se evidencia en valores altos de exhaustividad (*recall*). Sin embargo los modelos sugieren una perdida de precisión para ganar en exhaustividad (número de proteínas identificadas correctamente del total de positivos verdaderos) lo cual podría privilegiar el sobre-entrenamiento afectando así la calidad de las predicciones obtenidas en este estudio.

Finalmente, se propone una lista de 79 proteínas candidatas nuevas a través de un ensamblaje de métodos de clasificación de aprendizaje de maquinas que podrían participar en el desarrollo de cáncer colorrectal. Dichos genes tienen el potencial de ser validados experimentalmente como blancos terapéuticos y contribuir al conocimiento del desarrollo de la enfermedad en el futuro.

Se sugiere el uso de un nuevo set de datos con un mayor número de proteínas asociadas a cáncer colorrectal o la inclusión de datos de entrenamiento nuevos con el fin de mejorar la predicción de nuevas proteínas candidatas mediante el uso de *PULearn*.

Bibliografía

[1] World Health Organization, "international agency for research on cancer"." https://gco.iarc.fr/. Accessed: 2021-05-08.

[2] C. G, *Scale-Free Networks: Complex Webs in Nature and Technology.* Oxford University Press, 2007.

[3] D. G and T. Bouwmeester, *Global approaches to protein-protein interactions.* Curr Opin Cell Biol, 2003.

[4] C. V. and G. Boissy, *From protein-protein complexes to interactomics. Subcell Biochem.* Subcell Biochem, 2007.

[5] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[6] M. C. Vidal M. and A. Barabasi, *Interactome Networks and Human Disease.* Cell, 2011.

[7] D. Gatherer, "So what do we really mean when we say that systems biology is holistic?," 2010.

[8] e. a. Ng, A., "Resources for integrative systems biology: from data through databases to networks and dynamic system models. Brief Bioinform," 2006.

[9] S. Ghosh, Y. Matsuoka, Y. Asai, K.-Y. Hsin, and H. Kitano, "Software for systems biology: from tools to integrated platforms," *Nature Reviews Genetics*, vol. 12, pp. 821–832, Dec 2011.

[10] J. Dada and P. Mendes, "Multi-scale modelling and simulation in systems biology," 2011.

[11] e. a. Stuart, J.M., "A gene-coexpression network for global discovery of conserved genetic modules," 2003.

[12] M. B. Zhu, H. and M. Snyder, "Proteomics," 2003.

[13] A. Ma'ayan, "Insights into the organization of biochemical regulatory networks using graph theory analyses," 2009.

[14] B. Tripathi, *Adapting Community Detection Algorithms for Disease Module Identification in Heterogeneous Biological Networks.* Frontier in Genetics, 2019.

[15] Y. Han, *DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies.* Oxford University, 2019.

[16] S. Paul and D. Brahma, "An Integrated Approach for Identification of Functionally Similar MicroRNAs in Colorectal Cancer," 2019.

[17] S. P. Deng and W. L. Guo, "Identifying Key Genes of Liver Cancer by Networking of Multiple Data Sets," 2019.

[18] P. Ni, J. Wang, P. Zhong, Y. Li, F.-x. Wu, and Y. Pan, "on Disease Module Theory," vol. XX, no. X, pp. 1–11, 2018.

[19] C. Peng, Y. Zheng, and D. S. Huang, "Capsule Network Based Modeling of Multi-omics Data for Discovery of Breast Cancer-Related Genes," 2020.

[20] Y. Lin and X. Ma, "Predicting lincRNA-Disease Association in Heterogeneous Networks Using Co-regularized Non-negative Matrix Factorization," *Frontiers in Genetics*, no. January, 2021.

[21] L. v. Bertalanffy, *General system theory: Foundations, development, applications.* G. Braziller, 1968.

[22] M. D. Mesarovic, S. N. Sreenath, and J. D. Keene, "Search for organising principles: understanding in systems biology.," *Systems biology*, vol. 1 1, pp. 19–27, 2004.

[23] F. Bruggeman and H. Westerhoff, "The nature of systems biology," *Trends in microbiology*, vol. 15, pp. 45–50, 02 2007.

[24] R. D. Prasasya, D. Tian, and P. K. Kreeger, "Analysis of cancer signaling networks by systems biology to develop therapies," *Semin Cancer Biol*, vol. 21, pp. 200–206, Apr. 2011.

[25] D. Hanahan and R. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57–70, 02 2000.

[26] D. Hanahan and R. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[27] A. L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," jan 2011.

[28] M. Li, J. Y. Li, A. L. Zhao, and J. Gu, "Colorectal cancer or colon and rectal cancer?," *Oncology*, vol. 73, no. 1-2, pp. 52–57, 2007.

[29] G. Zhunussova, G. Afonin, S. Abdikerim, A. Jumanov, A. Perfilyeva, D. Kaidarova, and L. Djansugurova, "Mutation spectrum of cancer-associated genes in patients with early onset of colorectal cancer," *Frontiers in Oncology*, vol. 9, 2019.

[30] L. Dressler, M. Bortolomeazzi, M. R. Keddar, H. Misetic, G. Sartini, A. Acha-Sagredo, L. Montorsi, N. Wijewardhane, D. Repana, J. Nulsen, J. Goldman, M. Pollitt, P. Davis, A. Strange, K. Ambrose, and F. D. Ciccarelli, "Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the network of cancer genes (ncg) resource," *Genome Biology*, vol. 23, p. 35, Jan 2022.

[31] F. Dietlein, D. Weghorn, A. Taylor-Weiner, A. Richters, B. Reardon, D. Liu, E. S. Lander, E. M. Van Allen, and S. R. Sunyaev, "Identification of cancer driver genes based on nucleotide context," *Nature Genetics*, vol. 52, pp. 208–218, Feb 2020.

[32] T. Knijnenburg, T. Bismeijer, L. Wessels, and I. Shmulevich, "A multilevel pan-cancer map links gene mutations to cancer hallmarks," *Chinese journal of cancer*, vol. 34, p. 48, 09 2015.

[33] y. S. K. Usha Nandini Raghavan, Réka Alber, "Near linear time algorithm to detect community structures in large-scale networks," 2007.

[34] M. E. J. N. Aaron Clauset and C. Moore, "Finding community structure in very large networks," 2004.

[35] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, oct 2008.

[36] J.-x. W. B. H. y. G. C. Min Li, Jian-er Chen, "Modifying the DPClus algorithm for identifying protein complexes based on new topological structures," 2008.

[37] G. Rossetti, "Exorcising the Demon: Angel, Efficient Node-centric Community Discovery," 2020.

[38] IBM, "What is machine learning?." https://www.ibm.com/cloud/learn/machine-learning. Accessed: 2021-05-22.

[39] D. Bither and T. Suat, "Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences," 2022.

[40] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[41] Javatpoint, "Logistic regression in machine learning." https://www.javatpoint.com/logistic-regression-in-machine-learning. Accessed : 2022-06-07.

[42] Sruthi E.R., "Understanding random forest." https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/, 2021. Accessed : 2022-06-07.

[43] Javatpoint, "K-nearest neighbor(knn) algorithm for machine learning." https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning. Accessed : 2022-06-07.

[44] Joaquín Amat Rodrigo, "Análisis discriminante lineal (lda) y análisis discriminante cuadrático (qda)." https://www.cienciadedatos.net/documentos/28_linear_discriminant_analysis_lda_y_quadratic_discriminant_analysis_qda. Accessed : 2022-06-07.

[45] D. M. Gysi, Ítalo do Valle, M. Zitnik, A. Ameli, X. Gan, O. Varol, S. D. Ghiassian, J. J. Patten, R. A. Davey, J. Loscalzo, and A.-L. Barabási, "Network medicine framework for identifying drug-repurposing opportunities for covid-19," *Proceedings of the National Academy of Sciences*, vol. 118, no. 19, p. e2025581118, 2021.

[46] A. Cobián and L. E. Eguiarte, "Estructura y complejidad del genoma humano," 2002.

[47] R. Liu, M. Hirn, and A. Krishnan, "Accurately modeling biased random walks on weighted graphs using *Node2vec+*," 2021.

[48] R. Liu and A. Krishnan, "PecanPy: a fast, efficient and parallelized Python implementation of node2vec," *Bioinformatics*, vol. 37, pp. 3377–3379, 03 2021.

[49] Vivek Vinushanth Christopher, "Handling imbalanced data using geometric smote." https://towardsdatascience.com/handling-imbalanced-data-using-geometric-smote-770b49d5c7b5, 2020. Accessed : 2022-06-05.

[50] F. B. Georgios Douzas, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," 2019.

[51] G. Gulfidan, B. Turanli, H. Beklen, R. Sinha, and K. Y. Arga, "Pan-cancer mapping of differential protein-protein interactions," *Scientific Reports*, vol. 10, p. 3272, Feb 2020.

[52] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, B. Raphael, D. Marks, B. Ouellette, A. Valencia, G. Bader, P. Boutros, J. Stuart, R. Linding, N. Lopez-Bigas, and L. Stein, "Pathway and network analysis of cancer genomes," *Nature Methods*, vol. 12, pp. 615–621, June 2015. Funding Information: We gratefully acknowledge the assistance of J. Jennings during preparation of this manuscript. J.M.S. acknowledges support from the US National Cancer Institute (R01-CA180778 and U24-CA143858), Stand Up To Cancer, the Prostate Cancer Foundation and the Movember Foundation. P.C. is currently funded by a Ludwig Fund Postdoctoral Fellowship. P.C.B. and L.D.S. were supported by the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. P.C.B. was also supported by a Terry Fox Research Institute New Investigator Award and a Canadian Institutes of Health Research New Investigator Award. L.D.S. and G.W. acknowledge support from the US National Institutes of Health (NIH) and National Human Genome Research Institute (P41 HG003751). G.D.B. is supported by NRNB (NIH, National Institute of General Medical Sciences grant number P41 GM103504). Publisher Copyright: © 2015 Nature America, Inc. All rights reserved.

[53] E. Almaas, A. Vazquez, and A.-L. Barabasi, "Scale-free networks in biology," *Biological networks*, vol. 3, 01 2013.

[54] S. Wuchty, E. Ravasz, and A. L. Barabasi, "The architecture of biological networks," 2006.

[55] A.-L. Barabási, "Scale-free networks: A decade and beyond," *Science*, vol. 325, no. 5939, pp. 412–413, 2009.

[56] A.-L. Barabasi and Z. Oltvai, "Network biology: Understanding the cell's functional organization," *Nature reviews. Genetics*, vol. 5, pp. 101–13, 03 2004.

[57] X. Meng, W. Li, X. Peng, Y. Li, and M. Li, "Protein interaction networks: centrality, modularity, dynamics, and applications," *Frontiers of Computer Science (print)*, vol. 15, 01 2021.

[58] Y. Qi and H. Ge, "Modularity and dynamics of cellular networks," *PLoS computational biology*, vol. 2, p. e174, 01 2007.

[59] Z. Wang and J. Zhang, "In search of the biological significance of modular structures in protein networks," *PLoS computational biology*, vol. 3, p. e107, 07 2007.

[60] F. Mordelet and J.-P. Vert, "A bagging svm to learn from positive and unlabeled examples," *Pattern Recognition Letters*, vol. 37, pp. 201–209, 2014. Partially Supervised Learning for Pattern Recognition.

[61] W. Guo, Z. Sun, N. Zhao, Y. Zhou, J. Ren, L. Huang, and Y. Ping, "NOTCH2NLA silencing inhibits ovarian carcinoma progression and oncogenic activity in vivo and in vitro," *Ann Transl Med*, vol. 9, p. 1669, Nov. 2021.

[62] H. Hu, L. Ye, and Z. Liu, "GINS2 regulates the proliferation and apoptosis of colon cancer cells through PTP4A1," *Mol Med Rep*, vol. 25, Feb. 2022.

.1. Anexo 1

La información requerida en este anexo se encuentra en el archivo de excel *Criterios.xslx* en la hoja *node_frecuency*.

.2. Anexo 2

La información requerida en este anexo se encuentra en el archivo de excel *Criterios.xslx* en la hoja *nodes_by_weight*.

.3. Anexo 3

La información requerida en este anexo se encuentra en el archivo de excel *asyn_lpa_communities_excel.xslx*.

.4. Anexo 4

La información requerida en este anexo se encuentra en el archivo de excel *greedy_modularity_excel.xslx*

.5. Anexo 5

La información requerida en este anexo se encuentra en el archivo de excel *Louvain_algorithm.xslx*

.6. Anexo 6

La información requerida en este anexo se encuentra en el archivo de excel*v2_ipca_algorithm*

.7. Anexo 7

La información requerida en este anexo se encuentra en el archivo de excel*new_angel_algorithm.*

.8. Anexo 8

La información requerida en este anexo se encuentra en el archivo Python *Deteccion_de_comunidades.py*

.9. Anexo 9

La información requerida en este anexo se encuentra en el archivo Gephi *Louvain.gephi*

.10. Anexo 10

La información requerida en este anexo se encuentra en el archivo ipynb *Pu_Learning.ipynb*

.11. Anexo 11

La información requerida en este anexo se encuentra en el archivo ipynb *SMOTE_GSMOTE.ipynb*