Modelo de redes complejas y aprendizaje automático para optimización in-silico de la resistencia de Zea mays (maíz) a Puccinia sorghi (roya común del maíz)

Oscar Mauricio Ramírez Rico

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Director

Jurado

Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ingeniería.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Cali, Colombia (07, 12, 2023)

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 07 diciembre 2023

Autor: Oscar Mauricio Ramirez Rico

Título del Trabajo de Grado: "Modelo de redes complejas y aprendizaje automático para optimización in-silico de la resistencia de Zea mays (maíz) a Puccinia sorghi (roya común del maíz)"

Director: Hernán Camilo Rocha Niño Ph. D.

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

Camilo Rocha

Firma del Director del Trabajo de Grado

Santiago de Cali, 11 de 12 de 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ingeniería, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado Modelo de redes complejas y aprendizaje automático para optimización in-silico de la resistencia de Zea mays (maíz) a Puccinia sorghi (roya común del maíz), el cual será realizado por el estudiante Oscar Mauricio Ramírez Rico con código 8962026 perteneciente al énfasis en Sistemas y Computación, bajo la dirección del profesor Camilo Rocha.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Oscar Ramirez

Firma
Oscar Mauricio Ramirez Rico

C.C. 1075681751 de Zipaquirá

Camilo Rocha

Firma
Camilo Rocha

C.C. 79948061 de Bogotá

Maestría en Ingeniería
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: "Modelo de redes complejas y aprendizaje automático para optimización in-silico de la resistencia de Zea Mays (maíz) a Puccinia Sorghi (roya común del maíz)"

1. ÉNFASIS: Sistemas y Computación
2. TIPO DE PROYECTO: Investigación
3. ÁREA DE TRABAJO: Agroindustrial
4. ESTUDIANTE: Oscar Mauricio Ramírez Rico
5. CORREO ELECTRÓNICO: oscarmramirez05@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Transversal 22 # 8-58 Torre 20 Apto 104. Zipaquirá, Cundinamarca. Cel: 3197390467
7. DIRECTOR: Camilo Rocha
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: camilo.rocha@javerianacali.edu.co
10. CO-DIRECTOR: Jorge Finke
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica):
12. OTROS GRUPOS O EMPRESAS: Fundación Ceiba
13. PALABRAS CLAVE (al menos 5): Zea Mays, Coexpresión, Genoma, Puccinia Sorghi, Machine learning.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Hambre Cero y Producción y Consumo Responsables
15. FECHA DE INICIO (Desarrollo del proyecto): 10/05/2022
16. RESUMEN (máximo 400 palabras).

La identificación del grupo de genes asociado a una función biológica específica desempeña un papel fundamental en la comprensión del funcionamiento del genoma de una especie. Este conocimiento acerca de la función de un genoma resulta relevante para intervenir en diversos procesos biológicos, como por ejemplo, los diferentes tipos de estrés que podrían afectar el desarrollo de un organismo. En este trabajo, se lleva a cabo un estudio sobre la expresión de genes y sus funciones biológicas en el maíz (Zea mays), enfocándose particularmente en su respuesta al estrés provocado por la roya común (Puccinia sorghi). El objetivo principal fue predecir las anotaciones funcionales del genoma, y a partir de estas, predecir la capacidad de un conjunto de genes para resistir el estrés. Esto permitirá disminuir la cantidad de genes que requieren validación experimental *in-vivo*, optimizando tanto el tiempo de investigación como la eficiente utilización de recursos disponibles. Para alcanzar tal propósito, se emplea una metodología *in-silico* fundamentada en la asociación de anotaciones funcionales del genoma y en la clasificación de genes que demuestran resistencia al estrés. Estas predicciones se realizan utilizando técnicas de aprendizaje automático supervisado, tomando en cuenta la información de anotaciones funcionales previamente identificadas en la literatura, así como bases de datos conocidas y las propiedades topológicas de la red de coexpresión del maíz.

Estas anotaciones funcionales resultan fundamentales para analizar los procesos biológicos relacionados con esta enfermedad y desempeñan un papel clave en la mejora de la resistencia al estrés ambiental en el maíz. Los genes identificados como resistentes pueden ser de gran ayuda al reducir el conjunto de genes candidatos que deben someterse a validación *in-vivo* en respuesta a un tratamiento específico.

Modelo de redes complejas y aprendizaje automático para optimización in-silico de la resistencia de *Zea mays* (maíz) a *Puccinia sorghi* (roya común del maíz)

Oscar Mauricio Ramírez Rico
Código estudiante: 8962026
Director: Camilo Rocha, Ph.D.
Co-director: Jorge Finke, Ph.D.

Facultad de Ingeniería y Ciencias
Maestría en Ingeniería
Énfasis en Ingeniería de Sistemas y Computación
Pontificia Universidad Javeriana Cali
Cali, Colombia, Diciembre 2023

Modelo de redes complejas y aprendizaje automático para optimización *in-silico* de la resistencia de *Zea mays* (maíz) a *Puccinia sorghi* (roya común del maíz)

Abstract

La identificación del grupo de genes asociado a una función biológica específica desempeña un papel fundamental en la comprensión del funcionamiento del genoma de una especie. Este conocimiento acerca de la función de un genoma resulta relevante para intervenir en diversos procesos biológicos, como por ejemplo, los diferentes tipos de estrés que podrían afectar el desarrollo de un organismo.

En este trabajo, se lleva a cabo un estudio sobre la expresión de genes y sus funciones biológicas en el maíz (*Zea mays*), enfocándose particularmente en su respuesta al estrés provocado por la roya común (Puccinia sorghi). El objetivo principal fue predecir las anotaciones funcionales del genoma, y a partir de estas, predecir la capacidad de un conjunto de genes para resistir el estrés. Esto permitirá disminuir la cantidad de genes que requieren validación experimental *in-vivo*, optimizando tanto el tiempo de investigación como la eficiente utilización de recursos disponibles.

Para alcanzar tal propósito, se emplea una metodología *in-silico* fundamentada en la asociación de anotaciones funcionales del genoma y en la clasificación de genes que demuestran resistencia al estrés. Estas predicciones se realizan utilizando técnicas de aprendizaje automático supervisado, tomando en cuenta la información de anotaciones funcionales previamente identificadas en la literatura, así como bases de datos conocidas y las propiedades topológicas de la red de coexpresión del maíz.

Estas anotaciones funcionales resultan fundamentales para analizar los procesos biológicos relacionados con esta enfermedad y desempeñan un papel clave en la mejora de la resistencia al estrés ambiental en el maíz. Los genes identificados como resistentes pueden ser de gran ayuda al reducir el conjunto de genes candidatos que deben someterse a validación *in-vivo* en respuesta a un tratamiento específico.

Palabras Clave: *Zea mays*, Genoma, *Puccinia sorghi*, Machine learning.

Dedicatoria

En memoria de Holly, quien me enseñó el verdadero significado de la valentía y me inspiró a perseverar sin cesar.

Agradecimientos

Agradezco a mi familia por su amor y apoyo incondicional a lo largo de mi carrera académica. Gracias a su constante ánimo y confianza en mí, pude completar este proyecto con éxito.

También quiero expresar mi gratitud a mi hermano, Sergio Ramirez, a mi amigo y maestro, Miguel Romero, a mi director de tesis Camilo Rocha y a mi codirector de tesis Jorge Finke por su orientación, enseñanzas y consejos expertos. Su dedicación y paciencia me ayudaron a avanzar en mi investigación y a mejorar mi trabajo.

Asimismo, quiero agradecer a Carlos Ramírez, Gloria Inés Alvarez, Mauricio Quimbaya, Frank Valencia, Hernan Benitez, Isabel Garcia, Camila Riccio y Mauricio Ramírez por sus aportes valiosos y comentarios constructivos.

Además a mis amigos Alejandro Santis, Mateo Pastran, Luis Felipe Guevara y Cesar Leal, mi primo Alejandro Murcia, quienes siempre encontraron palabras de apoyo para motivarme a continuar en el proceso, en especial a mi futura esposa Valentina, por su apoyo constante, y finalmente Holly, por acompañarme en cada momento y enseñarme a continuar sin importar lo que pase. Sin su ayuda no habría sido posible concluir con éxito esta tesis.

Por último, no puedo dejar de mencionar a la Pontificia Universidad Javeriana Cali y a la Fundación Ceiba por brindarme los recursos necesarios para realizar mi investigación.

Todos estos individuos y organizaciones han sido fundamentales en mi camino académico y en la culminación de mi tesis. Les estoy eternamente agradecido.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

El maíz (*Zea mays*) es uno de los cereales más consumidos en todo el mundo debido a su alto contenido nutricional, lo que lo convierte en un alimento esencial tanto para los seres humanos como para los animales. Adicionalmente, el maíz también se utiliza como materia prima en diversas industrias para producir endulzantes y almidón alimentario, fabricar aceites y otros productos derivados de su fermentación como etanol, alcohol industrial, antibióticos y plásticos; además, en el contexto de la transformación sostenible que se experimenta actualmente, el maíz también ha sido utilizado como sustituto del petróleo y sus derivados [17, 29], por ejemplo, el bioetanol de maíz es un compuesto químico obtenido a partir de la fermentación de los azúcares que luego puede utilizarse como combustible. En general, se mezclan cantidades variadas con gasolina y su uso se ha extendido principalmente para reemplazar el consumo de derivados del petróleo.

La producción de maíz en Colombia, desempeña un papel importante en la seguridad alimentaria, el maíz es uno de los cultivos fundamentales en la dieta de su población, proporcionando nutrientes esenciales para una alimentación equilibrada. Además de su importancia nutricional, la producción de maíz contribuye a la generación de empleo en las comunidades rurales del país, fortaleciendo la economía agrícola y promoviendo la inclusión social. Este cultivo también tiene un impacto significativo en la cadena agroindustrial, al ser utilizado como materia prima para la producción de alimentos procesados y otros productos derivados que impulsan la economía de diferentes regiones. Sin embargo, como fenalce y Bustos lo describen en [1, 6], la principal problemática observada en la baja producción del maíz en el país se debe a la importación del grano, que se estima aproximadamente en un 74 % del maíz que se consume en Colombia, además en el año 2022 se produjeron 1,9 millones de toneladas de maíz blanco y amarillo, cifra que no cumple con la demanda exigida por el país, convirtiéndolo en el principal importador de maíz de la región, y séptimo en el mundo.

El cultivo de maíz en la Provincia Sabana Centro se siembra transitoriamente, haciendo rotación con otros cultivos. El departamento de Cundinamarca presentó una producción de 75.000 toneladas en el año 2018 [16]. El trabajo presentado en [28] afirma que las diferencias más relevantes con los principales productores de maíz que son Estados Unidos y China [6], con el resto del mundo están marcadas por el uso de material genético modificado y la tecnificación de sus cultivos, todas encaminadas a la maximización de la productividad en cada uno de estos ambientes [11]. En ese

mismo contexto, la investigación relacionada con la producción de semillas de alta calidad, es un factor determinante en el aumento de la productividad y la calidad del cultivo de maíz; así mismo, debe ir acompañada de jornadas de capacitación, con el fin de socializar y transferir la tecnología al agricultor [46]. De otro modo, las practicas utilizadas para aumentar la producción agrícola del cultivo de maíz intensifican el deterioro del medio ambiente, aumentando el uso de insumos químicos y dejando de lado las practicas agroecológicas sostenibles de los cultivos. Sin embargo, en Cundinamarca la inversión en los predios por parte de los productores y del estado en la tecnología de estos cultivos se caracteriza por ser escasa y la baja productividad se ha agravado en los últimos años debido al cambio climático [34].

Los diversos usos del maíz han llevado a un aumento en su demanda a nivel mundial y a un cambio en el mercado de los granos. Estudios como [21, 18] sugieren que esto puede generar crisis en países que dependen de las importaciones del maíz debido al aumento de los precios y la reducción de la oferta. Esta dinámica del mercado sumado a las diversas enfermedades que afectan el rendimiento y la calidad del maíz, requieren acciones que permitan mitigar las afectaciones.

Entre las enfermedades que afectan el maíz, una de alto impacto es la *roya*, causada por el hongo *Puccinia sorghi*. La roya común del maíz es una enfermedad foliar que afecta tanto a las hojas como a las espigas y se caracteriza por la aparición de manchas de color marrón-rojizo en las hojas, esto reduce la capacidad fotosintética de la planta y afecta su desarrollo general. Además, la infección temprana de las espigas puede causar la reducción del rendimiento y la calidad del grano, lo que representa una amenaza significativa para la producción y la seguridad alimentaria [12].

Por esta razón, es importante implementar técnicas y herramientas computacionales que permitan analizar las características genéticas del maíz con el fin de acelerar el mejoramiento de parámetros de desempeño de los cultivos y reducir costos de experimentación *in-vivo*. La biotecnología presenta un gran potencial en este sentido, gracias a los notables avances en genética molecular, ingeniería genética, aprendizaje automático y bioinformática a nivel mundial, es posible desarrollar nuevas variedades e híbridos de maíz que tengan mejoras sustanciales, como la biofortificación del cultivo a situaciones de estrés biótico y abiótico, la obtención de mayor calidad en el grano y la producción de productos finales con valor agregado [38]. El aumento de la productividad es fundamental para garantizar la seguridad alimentaria y promover el crecimiento económico.

Sin embargo, el costo y el tiempo necesarios para anotar grandes conjuntos de genes con sus funciones biológicas mediante experimentación *in-vivo* siguen siendo prohibitivamente altos debido a su demanda de tiempo y recursos económicos [7, 57], ya que un bajo porcentaje de los genomas quedan exitosamente anotados al realizar el proceso en el laboratorio, lo que conlleva a un enriquecimiento con grupos muy pequeños de genes. Para superar esta limitación, se han introducido enfoques híbridos que combinan el conocimiento existente sobre las asociaciones gen-función con métodos *in-silico* [8, 10, 26, 41]. Estos enfoques híbridos permiten enfrentar la naturaleza combinatoria de la anotación de genes, lo que posibilita que la experimentación computacional reduzca el esfuerzo, el tiempo y los costos asociados a la experimentación *in-vivo*.

Esta tesis se basa en el estudio de análisis genómicos y transcriptómicos para

desarrollar un clasificador que use la información de la coexpresión y sea capaz de predecir grupos de genes resistentes a la roya para optimizar la experimentación *in-vivo*.

El genoma es el conjunto de material genético (ADN) que almacena y codifica la información inherente a los seres vivos. Está compuesto por regiones codificantes, conocidas como genes, y regiones no codificantes. A través del genoma, se encuentran establecidos los mecanismos de regulación que permiten el adecuado funcionamiento de un organismo, así como el control en la transmisión de información, la coordinación de interacciones genéticas y proteicas, y las respuestas frente a cambios en el entorno, entre otros procesos fundamentales.

La coexpresión de genes se refiere a la activación simultánea de genes que están asociados en procesos biológicos. Por ejemplo, Glazebrook y colegas en [14] afirman que, en respuesta a un tipo de estrés, se activa una respuesta defensiva mediante la expresión de un grupo de genes. Representar la interacción de estos genes como una red de coexpresión (GCN, por sus siglas en inglés) resulta útil para identificar anotaciones funcionales. Las redes de coexpresión de genes se representan como grafos no dirigidos y ponderados. En estas redes, los nodos representan los genes y las aristas ponderadas reflejan el valor de la coexpresión. El análisis de las propiedades topológicas de estas redes permite realizar predicciones sobre las funciones de los genes [47]. Numerosos estudios han demostrado que las redes de coexpresión de genes (GCN) y los análisis basados en redes complejas son un marco valioso para guiar la anotación *in-silico* de genes [35, 40, 48]. Las anotaciones funcionales se definen en la ontología génica (GO, por sus siglas en inglés), la cual comprende tres tipos principales de anotaciones: procesos biológicos, funciones moleculares y componentes celulares [13].

La asociación de genes con funciones aún desconocidas es fundamental para comprender cómo el genoma sienta las bases de la vida. La investigación en el desarrollo de tratamientos que utilizan la información genómica de los organismos para abordar condiciones específicas, como mejorar la defensa a tensiones ambientales o enfermedades, ha generado un cuerpo significativo de estudios [44, 53, 49].

Un enfoque presentado en [56] propone la utilización de redes convolucionales de grafos para predecir funciones de proteínas de maíz. En particular, se emplea una secuencia de aminoácidos de las proteínas y la jerarquía de la ontología de genes para predecir las funciones mediante un modelo de red convolucional de grafos profundos denominado DeepGOA. Los resultados de este estudio demuestran que DeepGOA es una herramienta poderosa para integrar datos de aminoácidos y la estructura de GO en la anotación precisa de proteínas.

De manera similar, el trabajo presentado en [9] predijo los fenotipos y funciones asociados a los genes del maíz utilizando dos enfoques: (i) agrupación jerárquica basada en conjuntos de datos del transcriptoma y el metaboloma (conjunto de metabolitos presentes en un organismo); y (ii) análisis de enriquecimiento de la ontoloogía génica. Los resultados de este estudio sugieren que el perfilado de plantas individuales es un diseño experimental prometedor para reducir la brecha entre el laboratorio y el campo.

Gligorijevic et al. [15] propusieron un método de fusión de redes basado en autocodificadores profundos multimodales para extraer características de alto nivel de proteínas utilizando múltiples redes de interacción. Este enfoque, denominado

deepNF, se fundamentó en técnicas de aprendizaje profundo que capturan características relevantes de proteínas a partir de redes de interacción complejas y no lineales. Los resultados obtenidos demostraron la importancia de extraer nuevas características de las redes biológicas para la anotación de genes con funciones.

En resumen, el desarrollo de métodos computacionales y enfoques híbridos que combinan el conocimiento existente y la experimentación *in-silico* ha demostrado ser una estrategia efectiva para abordar la anotación de genes y predecir funciones en distintos organismos. Estos enfoques proporcionan alternativas más eficientes y económicas para comprender y aprovechar el potencial del genoma en la mejora de los cultivos.

Las anotaciones funcionales están estructuradas en una jerarquía y se definen como un grafo acíclico dirigido. Sin embargo, en los experimentos de anotación de genes, a menudo se pasan por alto las relaciones existentes entre los procesos biológicos, a pesar de que estas relaciones son fundamentales para mejorar la precisión y evitar inconsistencias en las predicciones. Se considera que una predicción es inconsistente con la jerarquía cuando se infiere que un gen tiene una función específica "a", pero no se infiere que también posee todos los ancestros de "a". El cumplimiento de las restricciones ancestrales se conoce comúnmente como la regla del camino verdadero en ontología génica [47, 2], y como restricción jerárquica en HMC [50].

Por otro lado, el trabajo presentado en [55] guarda una estrecha relación con este tema. En dicho estudio se propone un enfoque denominado Gene Ontology hierarchy preserving hashing (HPHash), que se utiliza para predecir funciones génicas y conservar el orden jerárquico entre los términos de Gene Ontology (GO). Este método se basa en la similitud taxonómica entre los términos para capturar la jerarquía existente en GO. Los resultados obtenidos evidenciaron que HPHash preserva la jerarquía GO y mejora el rendimiento de las predicciones.

Para desarrollar estrategias efectivas de manejo de la roya en el maíz, es fundamental comprender los mecanismos genéticos y moleculares que están involucrados en la resistencia del maíz a esta enfermedad. Aunque se han logrado avances en la identificación de genes asociados con la resistencia a la roya en el maíz, aún existen muchas incógnitas sobre los mecanismos subyacentes y cómo se pueden utilizar para mejorar la resistencia en las variedades de maíz. Por lo tanto, se requiere una mayor investigación en este campo para desarrollar estrategias de manejo más efectivas y contribuir al mejoramiento genético de variedades de maíz que sean más resistentes a la roya.

Uno de los posibles enfoques para aproximar a la respuesta de esta problemática es a través del análisis holístico de datos genómicos, transcriptómicos, y fenotípicos, que a través del uso de herramientas bioinformáticas permitan la predicción de funciones de genes poco conocidos o no caracterizados, como la posibilidad de candidatizar genes relevantes asociados a la resistencia a la roya. El estudio de la resistencia del maíz a la roya tiene implicaciones directas en la mejora de la productividad y calidad de los cultivos, y también contribuye al desarrollo de estrategias de manejo integrado de enfermedades en la agricultura. Además, al comprender mejor los mecanismos de resistencia, será posible diseñar estrategias de mejoramiento genético más eficientes y sostenibles, con el potencial de reducir la dependencia de pesticidas y mejorar la sostenibilidad ambiental de la producción de maíz.

En este trabajo se presenta un modelo de extracción de características para la

anotación *in-silico* de genes utilizando la clasificación jerárquica de multietiqueta (HMC). El enfoque propuesto consiste en diseñar un modelo de predicción que asignará funciones a los genes cumpliendo con la regla de la ruta verdadera. Para lograr esto, se realizará una búsqueda en la literatura para identificar cuerpos de anotación de genes y se recopilarán los genes que se hayan identificado como relevantes al estrés inducido por la infección de la roya.

Posteriormente, se construye una red de coexpresión de genes y se crea una matriz de afinidad que capturará las relaciones de coexpresión entre ellos. Estas relaciones son utilizadas para identificar grupos de genes que ayudarán a asociar funciones a los genes. Se aplica un algoritmo de agrupamiento con enriquecimiento de genes para extraer más características relevantes para la tarea de clasificación. Estudios anteriores, como el mencionado en [42], han demostrado que las nuevas características construidas a partir de redes de coexpresión de genes y las asociaciones entre genes y funciones utilizando algoritmos de agrupamiento espectral son efectivas para mejorar el rendimiento de la predicción en la anotación de genes. Con base en los resultados obtenidos, se realizará una predicción para identificar la asociación de funciones biológicas con los genes y finalmente se clasificarán grupos de genes como resistentes a la roya común del maíz.

En resumen, el objetivo principal de esta tesis de maestría es predecir las anotaciones funcionales del genoma de maíz, y a partir de estas, seleccionar genes relevantes que responden frente a un estrés biótico, tales como la infección por *Puccinia shorgi*, que llevaría a la enfermedad de la roya. Este nuevo conocimiento sobre la resistencia del maíz a la roya permitirá la evaluación de genes promisorios, a través de validación experimental *in-vivo*, que acelere el tiempo de investigación y optimice la eficiente utilización de recursos disponibles. Los resultados obtenidos pueden tener un impacto significativo en el mejoramiento genético de variedades de maíz más resistentes a la roya, lo que beneficiará tanto a los agricultores como a la seguridad alimentaria a nivel mundial.

Este documento se organiza de la siguiente manera:

El capítulo 1 es la indroducción.

El capítulo 2 presenta los métodos utilizados en el presente trabajo y la motivación para aplicarlos en las distintas etapas para la clasificación.

El capítulo 3 analiza el caso de estudio Zea Mays donde se observan los resultados de los modelos utilizados.

El capítulo 4 concluye el documento con los resultados obtenidos y propone el trabajo futuro

Capítulo 2

Preliminares

En esta sección se presentan los métodos utilizados en el trabajo de grado y la justificación para aplicarlos en las etapas para la clasificación. A continuación, se encuentran la agrupación espectral, las redes de coexpresión de genes, predicción de la función de los genes, la clasificación jerárquica multietiqueta, extracción de características basada en clusters, grafo de afinidad, agrupamiento de genes, enriquecimiento de genes, la aplicación de clasificación jerárquica multietiqueta para la predicción de funciones de genes, entrenamiento y predicción de resistencia, y la evaluación de los modelos de predicción.

2.1. Agrupación espectral

El propósito de realizar un análisis de agrupamiento a una red es identificar grupos de vértices que compartan una noción (paramétrica) de similitud [39, 5]. Normalmente, para el agrupamiento se utilizan medidas de centralidad o distancia. La agrupación espectral es un método discriminativo, es decir, no requiere conocimientos previos sobre las clases para clasificación, el agrupamiento se realiza utilizando únicamente la información contenida en los datos y algunos parámetros de partida como el número de agrupaciones. La inicialización es una etapa importante en los métodos no supervisados, ya que la mayoría de ellos son sensibles a los parámetros de partida [36]. La agrupación espectral es un método de agrupamiento con fundamentos en la teoría algebraica de grafos [20]. Se ha demostrado que la agrupación espectral ofrece un mejor rendimiento global en distintos ámbitos de aplicación que otros algoritmos de agrupamiento [30].

Las técnicas de agrupación espectral aprovechan la topología de los datos a partir de una representación basada en grafos no dirigidos y ponderados [36]. Dado un grafo G, la descomposición de agrupamiento espectral de G puede representarse mediante la ecuación $\mathbf{L} = \mathbf{D} - \mathbf{A}$, donde $\mathbf{L}$ es el Laplaciano, $\mathbf{D}$ es el grado (es decir, una matriz diagonal con el número de aristas incidentes en cada nodo), y $\mathbf{A}$ las matrices de adyacencia de G. La agrupación espectral utiliza, por ejemplo, los n vectores propios asociados a los n valores propios diferentes que cero más pequeños de $\mathbf{L}$. De este modo, cada nodo del grafo obtiene una coordenada en $\mathbb{R}^n$. La colección resultante de vectores propios sirve de entrada a un algoritmo de agrupación (por ejemplo, k-medias) que agrupa los nodos en n clusters. Esto significa que el análisis espectral puede apoyar a las técnicas convencionales generando una inicialización

que mejore la convergencia del algoritmo o enseñando caracteristicas de similitud o afinidad.

2.2. Red de coexpresión de genes

Los grafos son conjuntos, no vacíos, de objetos denominados nodos o vértices, y de líneas denominadas aristas que unen por pares los vértices. Las aristas indican la relación entre los vértices, de forma que si dos vértices están relacionados se traza una línea entre ellos, en caso contrario no debe existir ningún trazo. Los grafos pueden clasificarse en dirigidos y no dirigidos. Los grafos no dirigidos no tienen definido un sentido en la relación entre sus elementos. En los grafos dirigidos, las aristas tienen orientación definida [33]. Las redes de coexpresión de genes (GCN, por sus siglas en inglés) asocian los genes que tienen una expresión similar ante diferentes estímulos, lo que puede significar una relación funcional entre ellos. Las redes de coexpresión de genes son representadas como grafos no dirigidos donde cada vértice identifica un gen y las aristas la coexpresión entre dos genes.

Definición 2.2.1. Sea V un conjunto de genes, E un conjunto de aristas que conectan pares de genes, y $w : E \to \mathbb{R}_{\geq 0}$ una función de peso. Se dice que una red de coexpresión de genes pesada es un grafo pesado $G = (V, E, w)$.

El conjunto de genes V de una red de coexpresión es particular del genoma estudiado. La correlación de los perfiles de expresión entre cada par de genes se mide, por lo general, mediante el coeficiente de correlación de Pearson. Cada par de genes se asigna y clasifica según una medida de relación, y se utiliza un umbral como valor de corte para determinar E. La función de peso w denota la fuerza de la coexpresión entre cada par de genes en V. Por ejemplo, en la base de datos ATTED-II, la relación de coexpresión entre cualquier par de genes se mide como un z-score expresada como una función del índice de coexpresión LS (Logit Score) [32, 31].

2.3. Predicción de Funciones de Genes

En una red de coexpresión de genes anotada, cada gen se asocia al conjunto de funciones biológicas con las que está relacionado (por ejemplo, mediante experimentos *in-vitro*).

Definición 2.3.1. Sea A un conjunto de funciones biológicas. Una *red de coexpresión de genes anotada* es una red de coexpresión de genes $G = (V, E, w)$ complementada con una función de anotación $\phi : V \to 2^A$.

Dada una red de coexpresión anotada $G = (V, E, w)$ con función de anotación ϕ, el objetivo es utilizar la información representada por ϕ, junto con información adicional (por ejemplo, características de G), para obtener una función $\psi : V \to 2^A$ que extiende ϕ. Las asociaciones entre genes y funciones que no se encuentran anotadas en ϕ o bien no se han encontrado mediante experimentación en laboratorio, o bien no existen en un sentido biológico. Las nuevas asociaciones identificadas por ψ son una sugerencia de funciones que deben verificarse mediante experimentos *in-vitro* o *in-vivo*. La función ψ se obtiene a partir de un predictor de funciones asociadas a los genes, por ejemplo, basado en un modelo de aprendizaje automático supervisado.

2.4. Clasificación jerárquica multietiqueta

La *Clasificación de nodos* se encarga de predecir una clase de nodo para unos datos de entrada basándose en la información de otros nodos de la red [3]. En general, los problemas de clasificación de nodos se pueden clasificar en tres tipos diferentes: *clasificación binaria* se refiere a predecir un atributo (objetivo) con dos clases (por ejemplo, positivo y negativo) [22]; *clasificación multiclase* se refiere al caso en que el atributo a predecir tiene más de dos clases y son mutuamente excluyentes (por ejemplo, la marca de un automovil) [27]; y *clasificación multietiqueta* se refiere a la predicción de un atributo con al menos dos clases, pero en el que una instancia puede asociarse a más de una clase (por ejemplo, el problema de predicción de la función génica) [52].

Aunque los métodos de predicción mencionados se utilizan con frecuencia, generalmente no tienen en cuenta las relaciones jerárquicas entre las clases. Para tales escenarios, la clasificación jerárquica multietiqueta (HMC) aborda la tarea de predicción de resultados estructurados donde las clases se organizan en una jerarquía y una instancia puede pertenecer a múltiples clases. Los autores en [45] exponen que existen dos tipos de métodos para explorar la estructura jerárquica. En primer lugar, *top down o clasificadores locales* se refieren a predecir parcialmente las clases en la jerarquía de arriba a abajo. En segundo lugar, *big bang o clasificadores globales* se refiere al uso de un único clasificador que considera toda la jerarquía a la vez.

Los clasificadores que ignoran las relaciones de clase, prediciendo sólo las clases del último nivel en la jerarquía o prediciendo cada clase de forma independiente, a menudo conducen a *predicciones inconsistentes*. Esto se refiere al hecho de que se infiere que un nodo tiene una clase particular a, pero el resultado del clasificador no infiere la asociación del nodo con todas las clases antepasadas de a en la jerarquía. En otras palabras, una predicción incoherente indica que la predicción no satisface la jerarquía para alguna clase a. La satisfacción de las restricciones ancestrales a menudo se denomina *regla del camino verdadero* en GO [47, 2] y *restricción jerárquica* en HMC [50].

En trabajos se utilizaron cuatro métodos de clasificación jerárquica multietiqueta que son: *Clasificador local por nodo* (*lcn*) el cual consiste en entrenar un clasificador binario para cada clase de la jerarquía excepto la raíz. *Clasificador local por nodo padre* (*lcpn*) consiste en entrenar un clasificador multietiqueta para cada nodo padre de la jerarquía para distinguir entre sus clases hijas. *Clasificador local por nivel* (*lcl*) consiste en entrenar un clasificador multietiqueta para cada nivel de la jerarquía de clases excepto para la raíz. *Clasificador global* consiste en construir un único clasificador multietiqueta teniendo en cuenta la jerarquía en su conjunto durante una única ejecución [41, 43]. Observe en la figura 2.1 una ilustración de como funcionan los métodos explicados.

Sin embargo, los resultados presentados en [43], demuestran que el mejor desempeño de clasificación jerárquica multietiqueta se obtuvo reiterativamente del clasificador global. El clasificador global puede asignar a una instancia clases potencialmente de todos los niveles de la jerarquía, por esta motivación en el trabajo presente se utiliza únicamente este método.

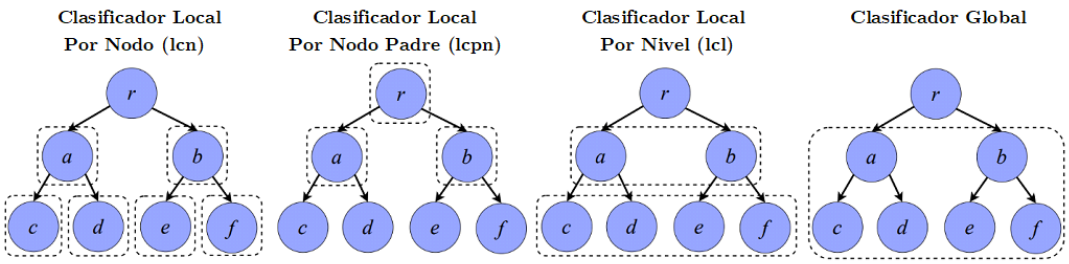

Figura 2.1: Ejemplo de métodos globales y locales para la clasificación jerárquica multietiqueta. Dada una jerarquía de clases (r, a, b, c, d, e y f), los cuadros discontinuos muestran el número de clasificadores necesarios para cada método [43].

2.5. Extracción de características basada en clusters

Se presenta la motivación para extraer características de la GCN utilizando un algoritmo de agrupamiento y el enriquecimiento de términos de la ontología de genes. En este proceso se combina la información de la GCN y las asociaciones entre genes y funciones, para crear nuevas características que capturen las propiedades topológicas de la GCN.

Las entradas del modelo son una GCN, denotada por $G = (V, E, w)$, un conjunto de funciones biológicas A, una función de anotación $\phi : V \to 2^A$, un grupo de genes relevantes al estrés por infección de la roya común del maíz B en la literatura [23] y un conjunto $K = k_0, \ldots, k_{m-1}$ para muestrear el número de clusters. La función de anotación ϕ debe satisfacer la regla del camino verdadero para la jerarquía GO [2, 47]. Es decir, si un gen está asociado a una función, entonces también debe estar asociado a cada ancestro de la función en la jerarquía, y si un gen no está asociado a una función, entonces no debe estar asociado a ninguno de sus descendientes.

La salida es una matriz de características J_F, de dimensión $|V| \times |A||K| \to [0, 1]$, que especifica la probabilidad de que los genes V se asocien a las funciones en A cuando el grafo se descompone en m clusters. La matriz J_F corresponde a la GCN (es decir, el grafo G) y a un grafo de afinidad definido en la siguiente subsección.

El proceso de extracción de características consta de tres etapas, que se representan en la Figura 2.2. En primer lugar, se crea un grafo de afinidad F con información en ϕ y B a partir de G. En segundo lugar, se aplica el algoritmo de agrupación espectral a F para el número m diferente de agrupaciones especificado en K. En tercer lugar, se utiliza la técnica de enriquecimiento de términos de la Ontología Génica para crear m características para cada función $a \in A$, correspondientes al número de clusters en K.

2.6. Grafo de afinidad

Se construye un grafo de afinidad $F = (V, E, w_F)$ entre G y ϕ. Su función de peso se define como la media entre el valor de la coexpresión especificado por w y

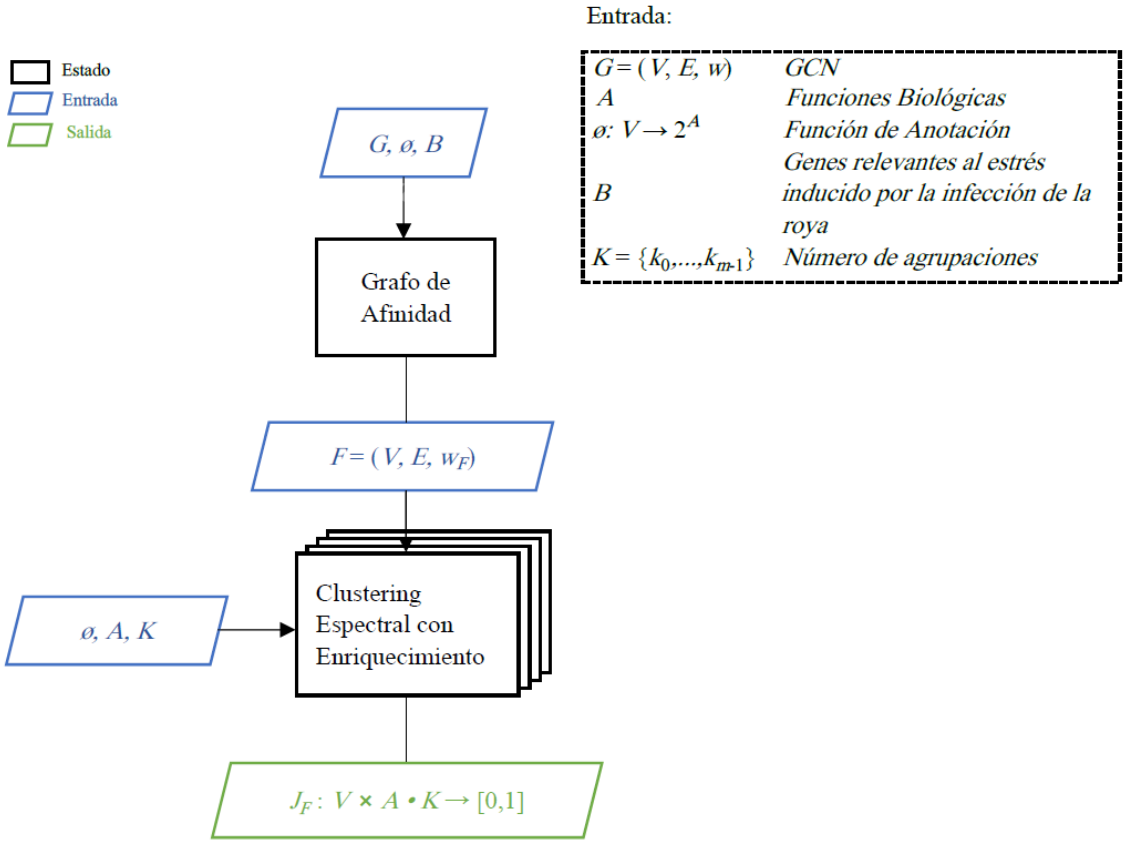

Figura 2.2: El método de extracción de características basado en la agrupación consta de tres etapas. La creación del grafo de afinidad, el proceso de agrupamiento espectral y el enriquecimiento de términos de la ontología de genes. Sus entradas son una GCN, denotada por $G = (V, E, w)$, un conjunto de funciones A, una función de anotación $\phi : V \to 2^A$, un grupo de genes identificados como relevantes al estrés por infección de la roya común del maíz B y un conjunto $K = k_0, \ldots, k_{m-1}$. Su salida es una matriz de características de dimensión $V \times A \cdot K \to [0, 1]$ que especifica la probabilidad de que los genes estén asociados a las funciones en A cuando el grafo se descompone en m clusters, cada uno de tamaño k_i, para $0 \leq i \leq m$.

la proporción de funciones compartidas entre genes especificada por ϕ.

Definición 2.6.1. La función de peso $w_F : V \times V \to [0, 1]$ se define para cualquier $u, v \in V$ como

$$w_F(u, v) = \frac{1}{2}\left(\frac{w(u, v) - 1}{\text{máx}(w) - 1} + \frac{|\phi(u) \cup \phi(v)|}{\phi(u) \cap \phi(v)|}\right),$$

donde máx(w) indica el valor máximo en el rango de w (que existe porque w es finito).

Bajo el supuesto de que al menos un elemento en el rango de w es mayor que 1, se garantiza que el rango de w_F es $[0, 1]$ (porque $w : V \times V \to [1, \infty)$). Este es el caso en la práctica, porque la coexpresión entre dos genes en la GCN se cuantifica en términos de z-score, que es muy poco probable que sea 1 para todos los pares de genes.

2.7. Agrupamiento de genes

El algoritmo de agrupación espectral se aplica al grafo F para descomponerlo (es decir, agrupar los genes V) utilizando el número de clusters especificado por $K = k_0, \ldots, k_{m-1}$. La descomposición de F se realiza m veces, una vez por cada k en K. La matriz de adyacencia del grafo pesado y no dirigido F se utiliza como una matriz de afinidad precalculada necesaria para el algoritmo de agrupación espectral. El resultado del algoritmo de clustering es una asignación de nodos a clusters de tamaño k, para cada k en K. Más concretamente, la salida de esta etapa es la matriz $I : V \times K \to [0, 1]$, donde cada columna $0 \leq i < m$ representa la descomposición de F en k_i clusters.

2.8. Enriquecimiento de genes

El objetivo de esta etapa es producir una matriz $J_F : V \times A \cdot K \to [0, 1]$, especificando la probabilidad de que los genes se asocien a cada función $a \in A$ cuando F se descompone en el número determinado de clusters.

Para cada descomposición de la etapa anterior (es decir, cada columna de la matriz I) y función $a \in A$, las agrupaciones resultantes se utilizan para calcular si un número significativo de miembros asociados a la función a está (localmente) presente. Intuitivamente, si los genes agrupados tienen una fuerte relación de coexpresión y la mayoría del grupo está asociada a la función biológica a, entonces es probable que los genes restantes también estén asociados a a (es decir, por asociación, véase [37]). De este modo, para cada $v \in V$, $a \in A$, y $k \in K$, la entrada $J_F(v, a \cdot k)$ es un p-valor que indica si la función a está sobrerrepresentada en la descomposición de k clusters de F. Este proceso se conoce comúnmente como enriquecimiento de términos de Ontología de Genes y puede utilizar diferentes pruebas estadísticas, como la prueba exacta de Fisher [54].

2.9. Clasificación jerárquica multietiqueta para la predicción de funciones de genes

Esta sección presenta el proceso de predicción de funciones de genes utilizando la clasificación jerárquica multietiqueta para crear un predictor, enriquecido con la información de las características creadas en la Sección 2.5.

La jerarquía de la Ontología de genes se define como un grafo acíclico dirigido que contiene tres tipos principales de anotaciones: procesos biológicos, funciones moleculares y componentes celulares [13]. Este trabajo se centra en los procesos biológicos, es decir, un subgrafo de la jerarquía GO. Este subgrafo se denota como $H = (A, R)$, donde A es el conjunto de procesos biológicos y R la relación binaria que representa las relaciones ancestrales entre pares de procesos biológicos (es decir, $(a, b) \in R$ significa que la función b es ancestro de la función a en la jerarquía GO). Para transformar la jerarquía GO de procesos biológicos en un árbol, se utiliza el algoritmo de clasificación topológica presentado en [41]. Como resultado, la jerarquía se divide en varios componentes, es decir, subárboles de H llamados subjerarquías. Cada subjerarquía, $H' = (A', R')$ con $A' \subseteq A$, $R' \subseteq R$, y $r \in A'$ la raíz, se asocia a

un subgrafo $G' = (V', E', w)$ que contiene todos los genes $v \in V$ asociados a r, es decir, $V' = \phi^{-1}(r)$.

Las entradas del enfoque son una subjerarquía $H' = (A', R')$, un subgrafo de la GCN, denotado por $G' = (V', E', w)$, donde $V' \subseteq V$ y $E' \subseteq E$, una función de anotación $\phi : V \to 2^{A'}$ y la matriz J_F resultante de la sección 2.5. La salida es una función $\psi : V' \times A' \to [0, 1]$, especificando, para cada gen $v \in V'$, la probabilidad $\psi(v, a)$ de que v esté asociado a función $a \in A'$.

En primer lugar, se crea la submatriz J'_F a partir de J_F, considerando únicamente los genes $V' \subseteq V$ y las funciones $A' \subseteq A$. Esta submatriz representa propiedades estructurales del subgrafo de la GCN G', y asociaciones entre genes y funciones basadas en múltiples particiones del grafo. La figura 2.3 ilustra el proceso de predicción de funciones biológicas.

Según los resultados obtenidos en el trabajo presentado en [43], se seleccionan las características resultantes de la matriz de afinidad para la predicción, esto debido a que mediante el método SHAP las diferentes combinaciones de selección de características reflejaron que los mejores resultados se obtienen con la matriz de afinidad. (SHapley Additive exPlanation, es un marco que calcula los valores de importancia de cada atributo o característica en un conjunto de datos para el desempeño de la predicción, utilizando conceptos de la teoría de juegos [24, 25]. SHAP asigna valores para explicar qué características del modelo son las más importantes para la predicción calculando los cambios en la predicción cuando se condicionan las características). No obstante, para el presente trabajo no se ejecutó el algoritmo SHAP debido a su alto costo computacional, ya que las combinaciones de los atributos tienen una complejidad alta. Asimismo, en la etapa de clasificación jerárquica el clasificador global fue seleccionado para predecir asociaciones entre genes y funciones sin inconsistencias (es decir, cumpliendo la regla del camino verdadero), y se utiliza únicamente este método ya que los resultados obtenidos son consistentemente mejores que el clasificador local por nodo (lcn), el clasificador local por nodo padre (lcpn) y el clasificador local por nivel (lcl).

2.10. Entrenamiento y predicción

2.10.1. Funciones Biológicas

Esta etapa comprende un proceso que combina dos técnicas de aprendizaje automático supervisado para construir el predictor ψ. En concreto, el método random forest con validación cruzada estratificada k-fold y la clasificación jerárquica multietiqueta se utilizan secuencialmente en un proceso.

El proceso toma como entrada la matriz J, que especifica las características significativas de J'_F, la subjerarquía H' y la función de anotación ϕ. En primer lugar, se aplica k-fold para dividir el conjunto de datos en k diferentes divisiones para la validación cruzada (tenga en cuenta que k no está relacionado con la entrada K). Es decir, cada división se utiliza como conjunto de prueba, mientras que las restantes $k-1$ divisiones se utilizan para el entranamiento. Recuerde que la validación cruzada tiene como ventaja evitar el overfitting en el entrenamiento. Para la tarea de predicción se seleccionó el método random forest ya que es un algoritmo de clasificación basado en árboles y multietiqueta [4], [51].

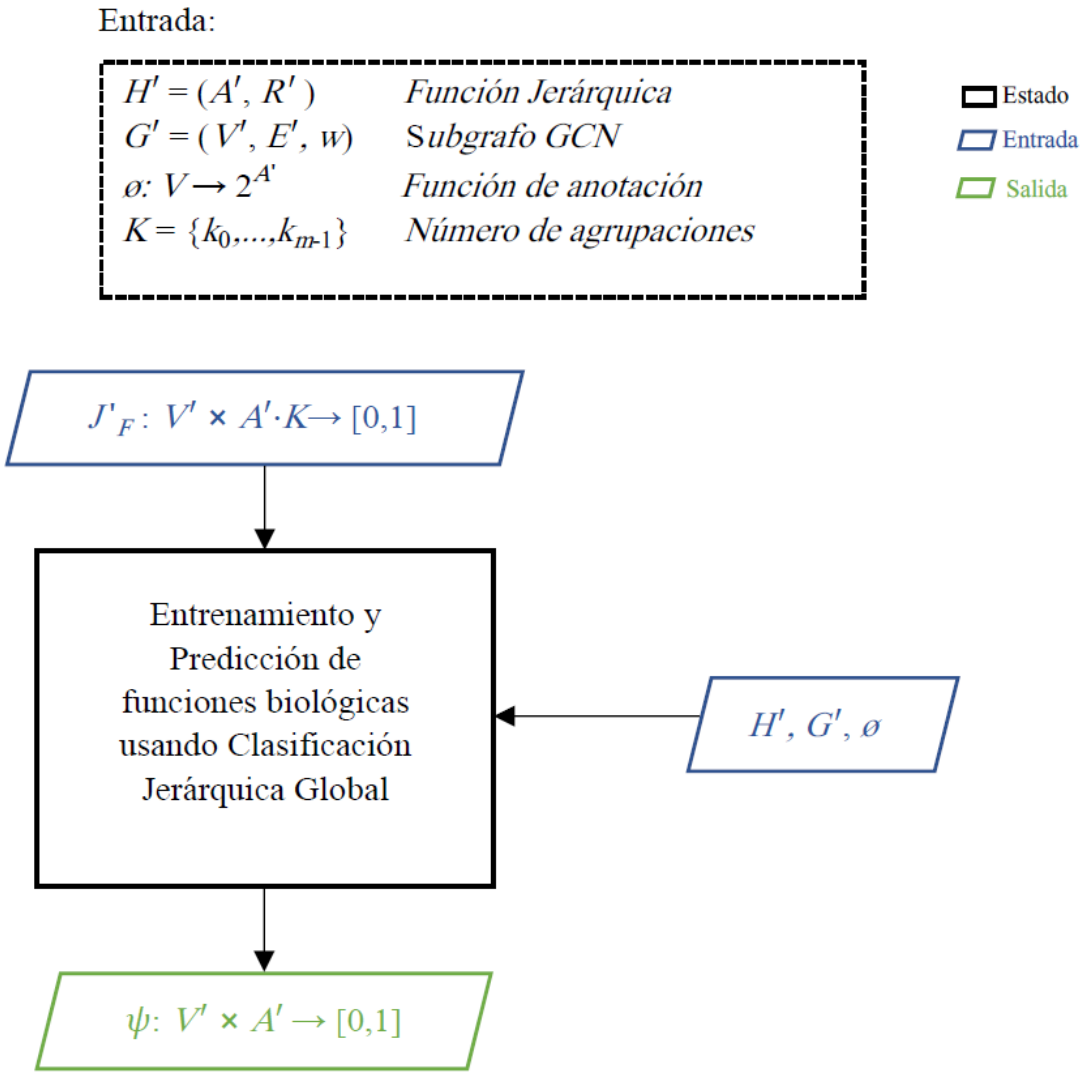

Figura 2.3: El proceso de clasificación jerárquica global tiene como entrada una subjerarquía $H' = (A', R')$, un subgrafo de la GCN $G' = (V', E', w)$, una función de anotación $\phi : V \to 2^{A'}$ que satisface la subjerarquía H', la submatriz J_F que contiene únicamente las funciones A' y los genes V'. Su salida es una función $\psi : V' \times A' \to [0, 1]$, que indica para cada gen $v \in V'$, la probabilidad $\psi(v, a)$ de que v esté asociado a cierta función $a \in A'$.

Los valores de los parámetros utilizados para random forest, a diferencia de los valores predeterminados de scikit-learn, son: 200 estimadores ($n\_estimators$) y un número mínimo de muestras de 5 ($min\_samples\_split$).

La clasificación jerárquica multietiqueta global cumple con la regla del camino verdadero. Es decir, para cada gen $v \in V$ y funciones $(a, b) \in R$, la asociación entre v y a se cumple también la asociación entre v y b (su antepasado). El resultado de esta etapa es el predictor ψ, es decir, las probabilidades de asociación entre los genes de V' y las funciones A'.

2.10.2. Resistencia a la Roya

En la etapa final, se lleva a cabo la predicción de genes relevantes que responden al ataque de la roya común en el maíz. Este proceso se divide en seis experimentos diferentes con el objetivo de evaluar y comparar su rendimiento.

En el primer experimento, se utiliza como entrada la función $\psi : V' \times A' \rightarrow [0, 1]$, que se obtiene en la subsección previa de predicción de asociaciones entre funciones biológicas y genes específicos del maíz. Junto con los datos de relevancia al estrés inducido por la infección de la roya disponible en el artículo de Saet y colaboradores [23], y finalmente se emplea el método random forest para realizar la clasificación y obtener la probabilidad de resistencia.

En el segundo experimento, se repite el mismo proceso utilizando el método random forest con k-folding estratificado.

En el tercer experimento, se repite nuevamente el procedimiento utilizando el método XGBoost para la tarea de predicción de resistencia.

Asímismo, para el cuarto enfoque propuesto, se utilizan las asociaciones entre genes y funciones, y el método XGBoost con k-folding estratificado para clasificar muestras como resistentes a la roya.

Para el quinto enfoque propuesto, se utilizan los datos de la matriz $I : V \times K \rightarrow [0, 1]$, la cual se obtiene en la etapa de clustering, aprovechando las características extraídas del GCN G, el grafo de afinidad F, y los datos de genes relevantes al estrés inducido por la infección de la roya B. Se realizan clasificaciones utilizando los métodos XGBoost y random forest con k-folding estratificado para predecir la resistencia a la roya común del maíz.

Finalmente, en el sexto experimento, se realiza nuevamente el proceso de predicción, tomando como entrada tanto las asociaciones de genes con funciones biológicas ψ como la matriz I del clustering y se usa el método XGBoost con k-folding estratificado para clasificar muestras como resistentes a la roya.

La descomposición en seis experimentos permite evaluar y comparar de manera exhaustiva el rendimiento del proceso de predicción de los genes relevantes frente al estrés por infección a la roya común del maíz.

La figura 2.4 ilustra el proceso de predicción de dichos genes.

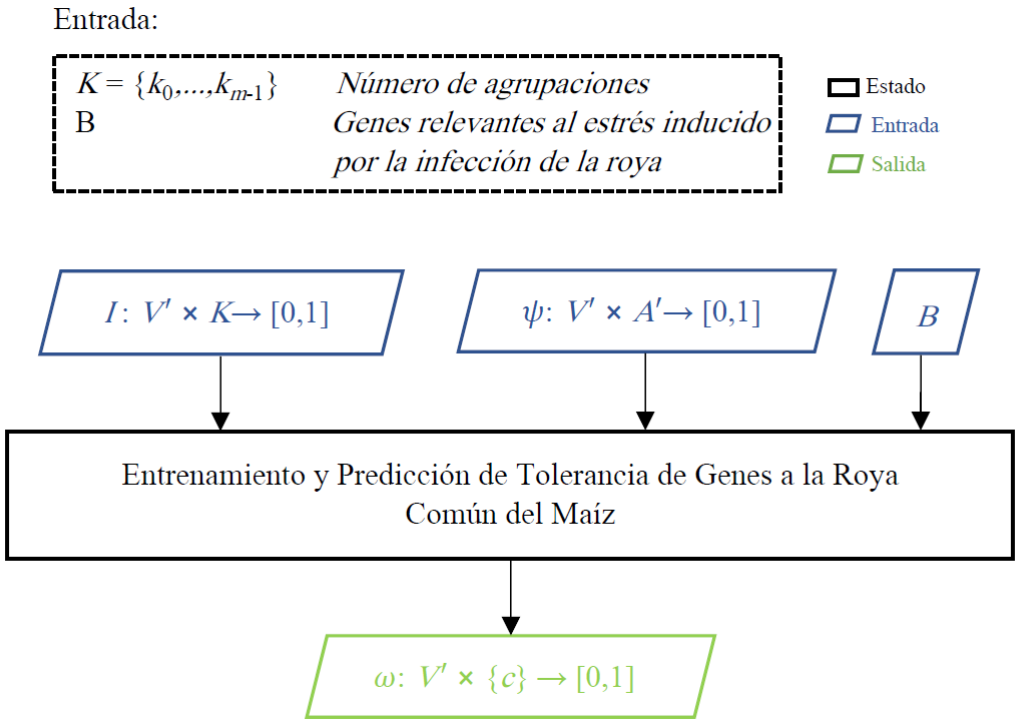

Figura 2.4: El proceso de predicción de la resistencia de los genes al ataque de la roya común del maíz se basa en una serie de elementos de entrada. Estos incluyen una matriz $I : V' \times K \rightarrow [0, 1]$ con los k agrupamientos, una función de anotación $\phi : V' \rightarrow 2^{A'}$ que representa las relaciones entre los genes y los grupos de funciones biológicas y cumple con la regla del camino verdadero, los datos de genes relevantes al estrés inducido por la infección de la roya B, las asociaciones A' y los genes V'. El resultado de este proceso es una función $\omega : V' \times c \rightarrow [0, 1]$, la cual se representa como un vector columna. Para cada gen $v \in V'$, esta función indica la probabilidad $\omega(v)$ de que el gen sea resistente o susceptible al ataque de la roya común en el maíz.

Capítulo 3

Caso de estudio: *Zea mays*

En la siguiente sección se presenta la aplicación del modelo descrito en las secciones 2.5 y 2.9 para realizar la predicción de características de asociación de funciones biológicas y selección de genes relevantes frente a estrés por infección con la roya común en el cultivo de maíz (*Zea mays*). En primer lugar, se describen en detalle los datos relacionados con el maíz utilizados en este estudio. A continuación, se implementa el método propuesto utilizando los datos mencionados. Por último, se compara el rendimiento de los modelos entrenados de forma independiente utilizando los conjuntos de características I y $\psi(v, a)$.

3.1. Descripción de datos y extracción de características

La información de coexpresión utilizada en el estudio se obtiene de la base de datos de ATTED-II [32]. La red de coexpresión de genes $G = (V, E, w)$ comprende 26131 nodos (genes) y 44621533 aristas. En este caso, se utiliza un umbral de puntuación z-score de 1 como medida de corte para G, es decir, E contiene aristas e que satisfacen $w(e) \geq 1$ (la mayoría de ellas satisfacen $w(e) > 1$). Obsérvese que el valor más alto se asigna a las conexiones más fuertes. La información funcional de esta red se ha tomado de DAVID Bioinformatics Resources [19] (actualización de 2021); contiene anotaciones de procesos biológicos, es decir, vías a las que contribuye un gen. Es importante señalar que los genes pueden estar asociados a varios procesos biológicos, y los procesos biológicos pueden estar asociados a múltiples genes. La base de datos comprende 3924 procesos biológicos A y 7021 relaciones ancestrales R entre estas funciones, que representan la jerarquía $H = (A, R)$ de la GO [13]. Se consideran un total de 255865 asociaciones entre genes y funciones, estas asociaciones representan la función de anotación $\phi : V \to 2^A$.

Posteriormente, se identifica el grupo de genes relevantes al estrés por infección por la roya común del maíz en el trabajo de Kim et al. [23], de todo el conjunto que ellos enseñan como relevantes, solamente 161 se encuentran en la red de coexpresión de genes G que es el conjunto de datos con el que se realiza la predicción. De los 161 genes, 61 tienen al menos una función biológica en A. Luego, es necesario filtrar los otros genes de la red de coexpresión G que comparten exactamente las mismas funciones a los 61 identificados, este proceso se realiza para obtener un mejor

desempeño en la predicción ya que entre estos grupos de genes se comparten más características. Como resultado de este proceso de filtrado se reducen 26131 genes a solamente 8265 con 5654974 arcos, y 3703 funciones. Esto permite que el proceso de clasificación pueda desarrollarse de una manera más eficiente con dicha base de datos. Sin embargo, es necesario revisar que la nueva red de coexpresión de genes sea conexa para que tenga sentido asociar funciones biológicas, lo cual se cumple al verificarlo con las propiedades topológicas de la red resultante.

El método de extracción de características se aplica con las entradas G, A, ϕ y $K = \{10, 20, \ldots, 100\}$ (los valores se incrementan en pasos de 10 hasta 100). El resultado es la matriz de características J_F, que especifica la probabilidad que los genes de maíz V se asocien a los procesos biológicos A cuando el gráfico se descompone en el número de agrupaciones en K.

Es importante mencionar que la jerarquía de ontología génica se divide en 28 subjerarquías cuando se consideran sólo los procesos biológicos. Además, se descartan todas las subjerarquías con menos de 10 funciones y se utiliza el algoritmo de ordenación topológica introducido en [41] para transformar las subjerarquías, representadas como DAGs (Grafos Acíclicos dirigidos), en árboles. Para cada relación ancestral (a, b) enR (b es ancestro de a), el algoritmo asigna un peso como el cociente entre el número de genes asociados a a y el número de genes asociados a b. A continuación, para cada función $a \in A'$ con más de un ancestro, sólo se conserva el de mayor peso (los empates se deshacen arbitrariamente).

Raíz	Descripción	Funciones	Funciones por nivel
GO:0050896	respuesta a estimulos	13	5/5/2
GO:0051179	localización	25	3/5/9/6/1
GO:0065007	regulación biologica	37	2/5/11/10/4/2/2
GO:0008152	procesos metabólicos	92	8/18/38/12/7/6/2
GO:0009987	procesos celulares	92	13/19/19/17/13/8/2

Cuadro 3.1: Subjerarquías resultantes H' de procesos biológicos para el maíz. El identificador y la descripción de cada función raíz r se presentan en la primera y segunda columnas, respectivamente. La tercera columna muestra el número de funciones A' dentro de cada subjerarquía. La última columna muestra el número de funciones por nivel, por ejemplo, la primera subjerarquía tiene 3 niveles y hay 5, 5 y 2 funciones en cada nivel.

Como resultado, hay 5 subjerarquías de procesos biológicos. La tabla 3.1 describe cada subjerarquía H', empezando por el término raíz r y su descripción, siguiendo por el número de funciones A' en el subgrafo de la red de coexpresión de genes (GCN) asociado G'. El enfoque de predicción se aplica a cada subjerarquía H' de forma independiente. Es importante mencionar que el método global requiere un clasificador por jerarquía.

3.2. Resumen de resultados

La figura 3.1 muestra los resultados de predicción obtenidos mediante el primer enfoque propuesto (ver Sección 2.10.2). En este proceso, se utilizan las asociaciones entre genes y funciones $\psi(v, a)$, aprovechando las características extraídas del GCN G y el grafo de afinidad F. Además, se consideran las relaciones ancestrales de los procesos biológicos y los genes relevantes al estrés B.

El método seleccionado para la clasificación de la resistencia a la roya común del maíz es el random forest ya que es un método eficiente cuando se extraen características de los nodos o aristas de un grafo y luego se predicen variables basadas en esas características. Se realiza un aumento gradual de la muestra, incrementando en 10 genes en cada etapa.

En general, se puede observar que el rendimiento obtenido no supera el 50 %. Esto significa que, de los 10 genes identificados, solamente 5 demuestran ser efectivamente relevantes. Además, a medida que aumenta la cantidad de genes considerados, la precisión disminuye, llegando a alcanzar valores inferiores al 10 %.

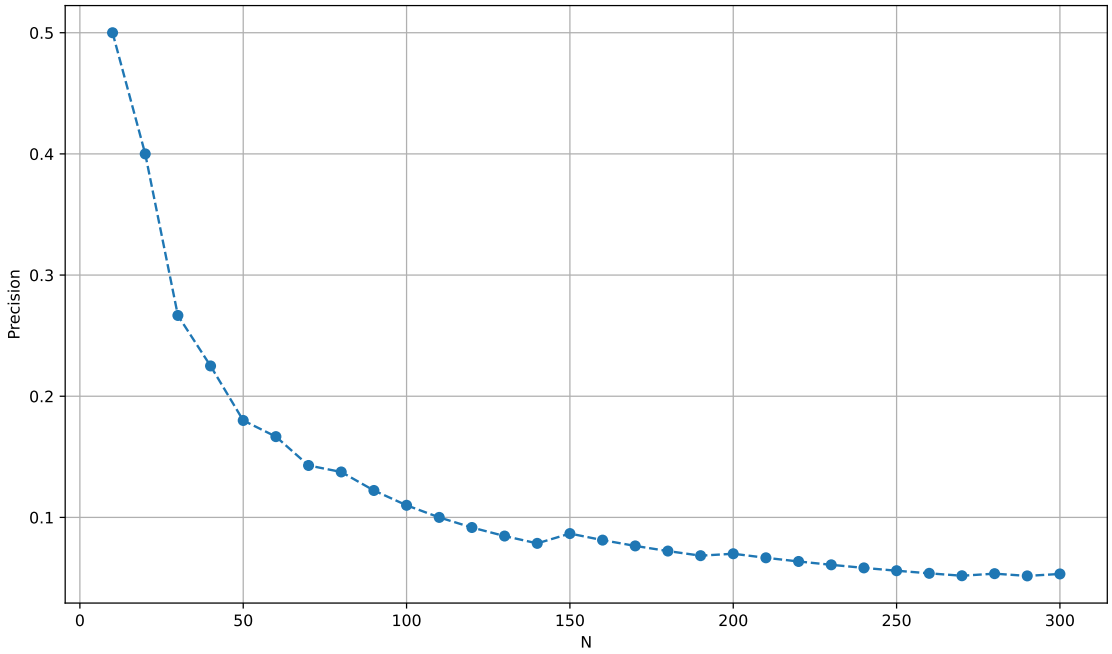

Figura 3.1: Predicción del primer enfoque propuesto, el cual utiliza la función de asociacion entre genes y funciones $\psi(v, a)$, y el método random forest para clasificar muestras como resistentes a la roya, cuyo tamaño incrementa 10 genes en cada etapa.

La figura 3.2 muestra los resultados de predicción obtenidos mediante el segundo enfoque propuesto (ver Sección 2.10.2). En este proceso, se utilizan las asociaciones entre genes y funciones $\psi(v, a)$, aprovechando las características extraídas del GCN G y el grafo de afinidad F. Además, se consideran las relaciones ancestrales de los procesos biológicos y los datos de relevancia B.

El método seleccionado para la clasificación de la relevancia al estrés por infección a la roya común del maíz es random forest con k-folding estratificado. Se realiza un aumento gradual de la muestra, incrementando en 10 genes en cada etapa.

En este caso, se puede observar que el rendimiento obtenido mejora hasta alcanzar aproximadamente un 70 %. Esto significa que, de los 10 genes identificados, alrededor de 7 demuestran estar efectivamente asociados a la resistencia. Además, a medida que aumenta la cantidad de genes considerados, la precisión muestra un mejor comportamiento en comparación con el primer modelo. Sin embargo, para muestras muy grandes, el rendimiento aún se mantiene cerca del 10 %.

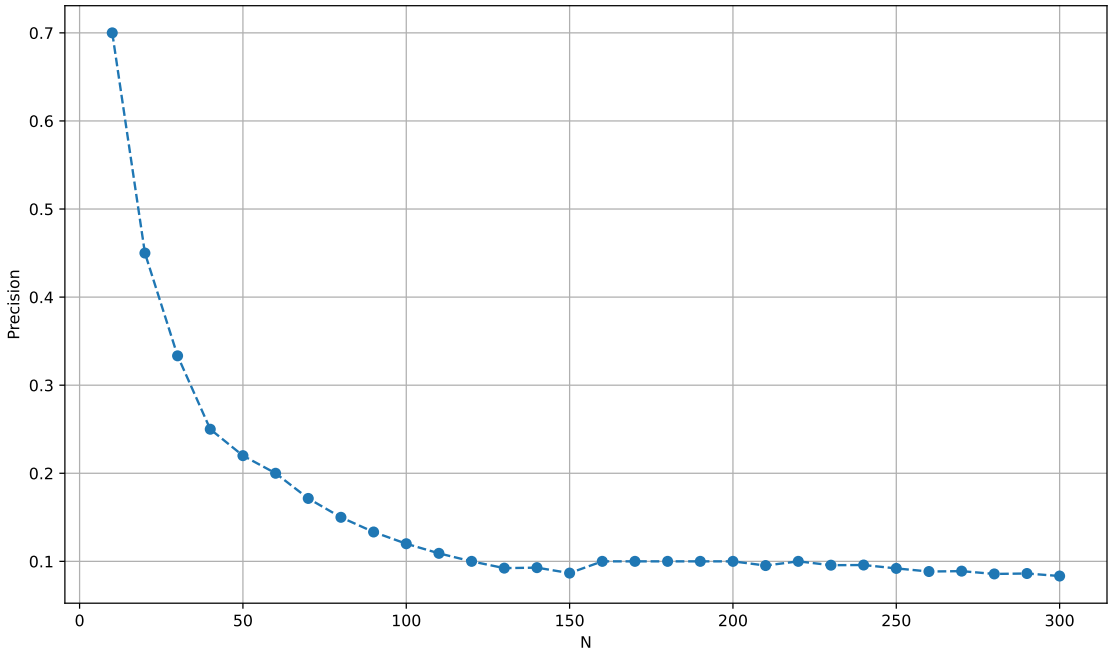

Figura 3.2: Predicción del segundo enfoque propuesto, el cual utiliza las asociaciones entre genes y funciones $\psi(v, a)$, y el método random forest con k-folding estratificado para clasificar muestras como resistentes a la roya, cuyo tamaño incrementa 10 genes en cada etapa.

La figura 3.3 muestra los resultados de predicción obtenidos mediante el tercer enfoque propuesto (ver Sección 2.10.2). En este proceso, se utilizan las asociaciones entre genes y funciones $\psi(v, a)$, aprovechando las características extraídas del GCN G y el grafo de afinidad F. Además, se consideran las relaciones ancestrales de los procesos biológicos y los datos de relevancia B.

El método seleccionado para la clasificación de la resistencia a la roya común del maíz es el XGBoost, este método resulta útil cuando se trabaja con características derivadas de un grafo para resolver problemas de clasificación o regresión. Se realiza un aumento gradual de la muestra, incrementando en 10 genes en cada etapa.

En este caso, se puede observar que el rendimiento es muy similar al segundo enfoque, pero se observa una mejora en las muestras de 30 a 100 genes. Sin embargo, a medida que aumenta la cantidad de genes considerados, la precisión muestra un comportamiento muy similar al modelo anterior. Al obtener resultados muy similares con distintos métodos de clasificación se puede considerar que el desempeño de clasificación se debe principalmente al experimento realizado, es decir, la tarea de predicción de genes con una buena respuesta al estrés producido por la roya común del maíz es más eficiente con muestras pequeñas.

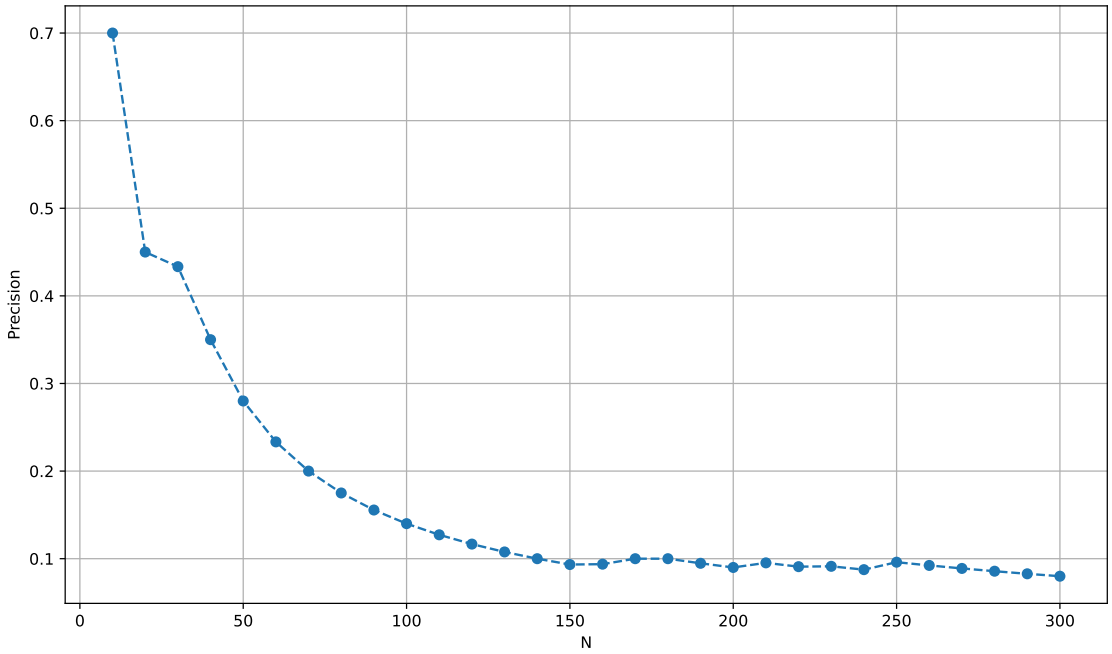

Figura 3.3: Predicción del tercer enfoque propuesto, el cual utiliza las asociaciones entre genes y funciones $\psi(v, a)$, y el método XGBoost para clasificar muestras como resistentes a la roya, cuyo tamaño incrementa 10 genes en cada etapa.

La figura 3.4 muestra los resultados de predicción obtenidos mediante el cuarto enfoque propuesto. En este proceso, se utilizan las asociaciones entre genes y funciones $\psi(v, a)$, aprovechando las características extraídas del GCN G y el grafo de afinidad F. Además, se consideran las relaciones ancestrales de los procesos biológicos y los datos de relevancia B.

El método seleccionado para la clasificación de la resistencia a la roya común del maíz es el XGBoost con k-folding estratificado. Se realiza un aumento gradual de la muestra, incrementando en 10 genes en cada etapa.

En este caso, se puede observar que el rendimiento es el mejor hasta ahora. Aunque el mayor rendimiento obtenido siempre ha sido de alrededor del 70 %, se logra un porcentaje considerablemente alto para muestras pequeñas (menores que 50). Sin embargo, a medida que aumenta la muestra, la precisión muestra un desempeño más bajo.

La figura 3.5 muestra los resultados de predicción obtenidos mediante el quinto enfoque propuesto. En este proceso, se utilizan los datos de la matriz $I : V \times K \to [0, 1]$, la cual se obtiene en la etapa de clustering, aprovechando las características extraídas del GCN G, el grafo de afinidad F, y los datos de resistencia B. Se realizan clasificaciones utilizando los métodos XGBoost y random forest con k-folding estratificado para predecir la resistencia a la roya común del maíz. La muestra se incrementa en 10 genes en cada etapa.

En este caso, la predicción no muestra un buen desempeño con ninguno de los clasificadores. Aunque random forest muestra el resultado más favorable, la precisión se reduce significativamente de alrededor del 70 % a solo 20 % para muestras pequeñas. Además, se puede observar que el rendimiento general para muestras grandes es

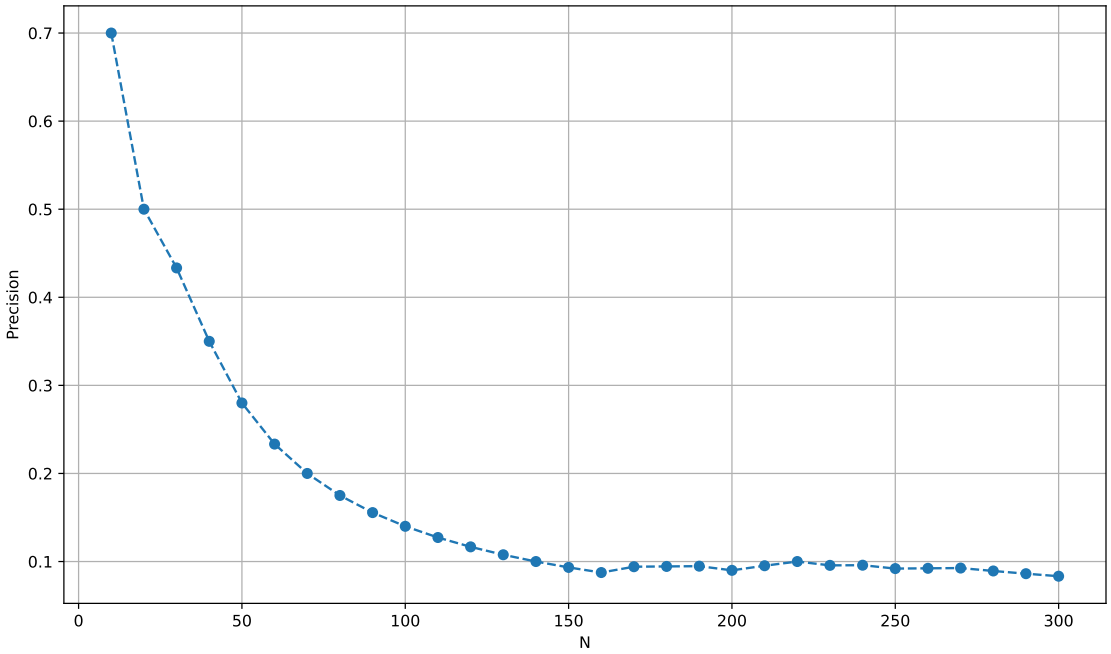

Figura 3.4: Predicción del cuarto enfoque propuesto, el cual utiliza las asociaciones entre genes y funciones $\psi(v, a)$, y el método XGBoost con k-folding estratificado para clasificar muestras como resistentes a la roya, cuyo tamaño incrementa 10 genes en cada etapa.

inferior al 10 %, llegando incluso a valores cercanos al 5 %, en la figura 3.5 solamente se enseñan los resultados del método que obtuvo un mejor desempeño.

Finalmente, la figura 3.6 muestra los resultados de predicción obtenidos mediante el sexto y último enfoque propuesto. En este proceso, se utilizan los datos tanto de las asociaciones entre genes y funciones $\psi(v, a)$, como los de clustering, aprovechando las características extraídas del GCN G, el grafo de afinidad F y los datos de genes relevantes al estrés B. Se realiza la clasificación utilizando el método XGBoost con k-folding estratificado, el cual ha demostrado consistentemente los mejores resultados en la tarea de predicción de resistencia a la roya común del maíz. La muestra se incrementa en 10 genes en cada etapa.

En este caso, la predicción muestra el mejor desempeño de todos los enfoques. Aunque para muestras pequeñas la precisión disminuye ligeramente en comparación con el cuarto enfoque, esta reducción es mínima. Además, se puede observar que el rendimiento para muestras grandes mejora significativamente para números mayores a 100, alcanzando como mínimo un 10 % de precisión.

De acuerdo con Romero et al. [42], se ha sugerido que las características generadas mediante la utilización de la GCN, así como las conexiones entre genes y funciones y el algoritmo de agrupación espectral, desempeñan un papel fundamental en la mejora del rendimiento de la predicción en el problema de anotación de genes. Estas características y conexiones resultan más efectivas en comparación con otras características de la GCN y la información funcional de los genes.

Sin embargo, dependiendo del enfoque de extracción de características se produce dos conjuntos distintos de rasgos que se combinan y utilizan para la predicción [43].

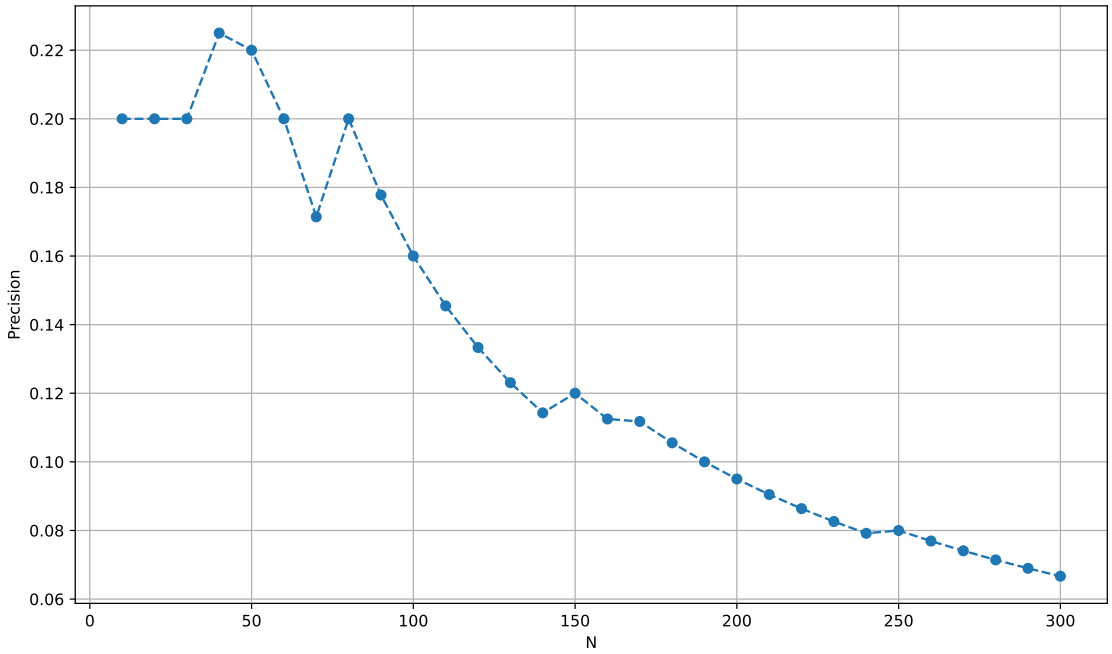

Figura 3.5: Predicción del quinto enfoque propuesto, el cual utiliza la matriz $I : V \times K \to [0, 1]$, la cual se obtiene en la etapa de clustering, y el método random forest con k-folding estratificado para clasificar muestras como resistentes a la roya, cuyo tamaño incrementa 10 genes en cada etapa.

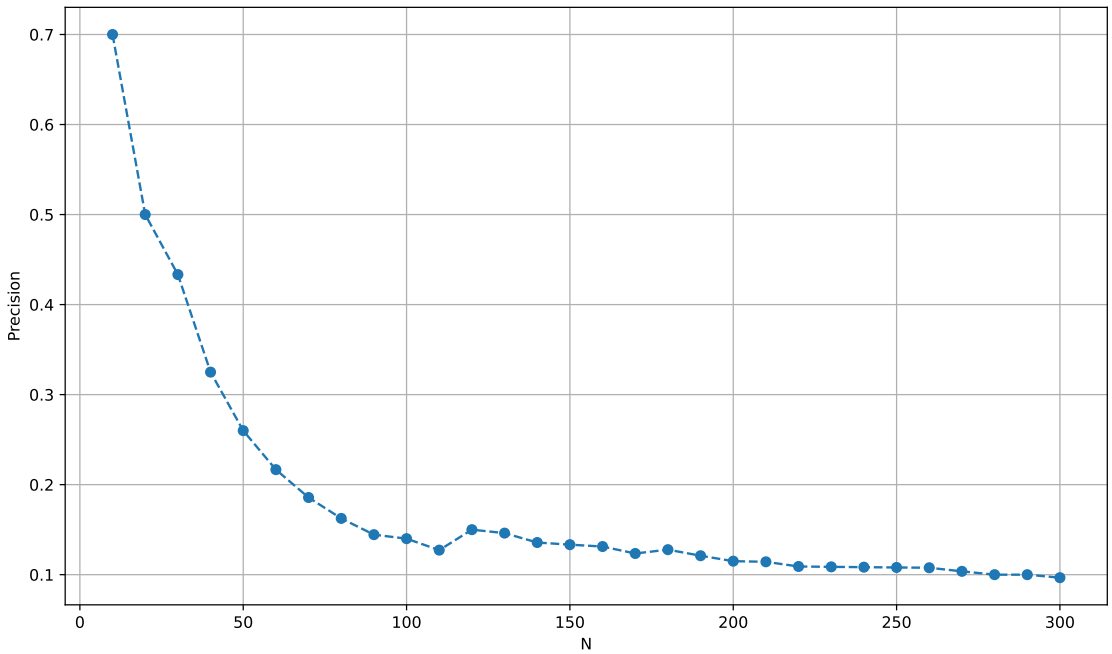

Figura 3.6: Predicción del sexto enfoque propuesto, el cual utiliza las asociaciones entre genes y funciones $\psi(v, a)$, los datos de clustering y el método XGBoost con k-folding estratificado para clasificar muestras como resistentes a la roya, cuyo tamaño incrementa 10 genes en cada etapa.

Estas características son filtradas mediante los valores SHAP medios para seleccionar las más relevantes. Para analizar la relevancia individual de cada conjunto de características en el problema de anotación de genes, se llevan a cabo dos enfoques: (i) se observa la distribución de las características filtradas para el método global, y (ii) se compara el rendimiento de la tarea de predicción utilizando cada conjunto de características de forma independiente.

En dicho estudio se concluye que las características del grafo de afinidad F desempeñan un papel crucial en la mejora del rendimiento del enfoque propuesto en todas las subjerarquías. Además, se encontró que a medida que una función se encuentra más profunda en una subjerarquía, las probabilidades predichas tienden a ser más bajas. Asimismo, se determinó que el método global supera consistentemente a los métodos locales. Estas conclusiones se respaldan mediante tres métricas y sugieren que el uso de técnicas de agrupación para extraer características de la GCN y considerar la estructura jerárquica de los procesos biológicos resulta clave en la tarea de predicción de funciones génicas.

Basándose en los resultados obtenidos en los estudios previos mencionados, se ha implementado un procedimiento de extracción y selección de características para abordar los problemas de la asociación de los genes responsivos a la roya común del maíz y anotación de genes, teniendo en cuenta tanto la jerarquía de las funciones biológicas. En este caso, se ha optado por omitir el proceso de filtrado de características utilizando el método de SHAP debido a su alta complejidad computacional. No obstante, se utilizan las mismas características obtenidas por Romero et al. [43] para la tarea de predicción, ya que se comparten los atributos relevantes.

Además, se ha decidido utilizar exclusivamente el método global en lugar de los métodos locales de jerarquía, ya que el rendimiento del método global fue consistentemente superior en los experimentos previos reportados por Romero et al. [43]. Este enfoque combinado permite lograr una clasificación eficiente de pequeños grupos de genes al utilizar tanto los datos del agrupamiento espectral como la asociación de genes con funciones biológicas y los datos de resistencia al ataque de la roya. Para ello, se emplean métodos de XGBoost y se aplica la técnica de k-folding estratificado para validar el modelo de manera robusta.

Capítulo 4

Conclusiones y trabajo futuro

4.1. Conclusiones

Mediante la combinación de modelado basado en redes, análisis de agrupamiento, Machine Learning y clasificación jerárquica multietiqueta, este documento presenta un enfoque novedoso para abordar el desafío de predecir las funciones biológicas de los genes del maíz y su relevancia a un tipo específico de estrés. Para lograr esto, se utiliza la descomposición espectral del grafo GCN, así como información relevante sobre expresión génica diferencial frente al estrés, asociaciones entre genes y funciones, y las relaciones ancestrales entre las funciones, es decir, la jerarquía GO. En conjunto, esta metodología ofrece un enfoque efectivo para abordar la complejidad de la predicción de funciones biológicas en el contexto del maíz y su respuesta a diferentes tipos de estrés.

En el presente documento se realiza un estudio exhaustivo del genoma del maíz (*Zea mayz*). Para mejorar el rendimiento del enfoque utilizado, se emplea información estructural de la red de coexpresión de genes, obtenida mediante un algoritmo de agrupación espectral. Además, se considera la estructura jerárquica de los procesos biológicos utilizando HMC (Clasificación Jerárquica Multietiqueta).

El método de clasificación jerárquica global aprovecha todas las características disponibles de una subjerarquía para construir un único clasificador, lo cual resulta en un rendimiento significativamente alto en la asociación de genes con funciones. Finalmente, el algoritmo de clasificación XGBoost se utiliza de manera exitosa para predecir genes relevantes frente al ataque de la roya.

Esta combinación de estrategias, basada en el uso de información estructural de la red de coexpresión de genes, la consideración de la estructura jerárquica de los procesos biológicos y el algoritmo de clasificación XGBoost, se revela como una aproximación altamente efectiva para abordar la predicción de resistencia al ataque de la roya en el maíz.

Los resultados presentados en [42] demuestran que al extraer características del GCN utilizando agrupamiento espectral se logra un rendimiento mejorado en la predicción de funciones de genes abordada como un problema de clasificación binaria independiente por función. En otro estudio [41, 43], se ha comprobado que al considerar las relaciones ancestrales entre funciones y asegurar la coherencia con la estructura jerárquica (es decir, cumplir la regla del camino verdadero), utilizando las características extraídas del GCN, se obtiene una mejora en el rendimiento de la

predicción en un enfoque de clasificación jerárquica multietiqueta.

En este trabajo, se utilizan tanto el agrupamiento espectral como la predicción de asociación de funciones de genes mediante clasificación jerárquica multietiqueta para identificar genes relevantes a un estrés específico, en particular, la roya común del maíz. El clasificador XGBoost se emplea con éxito para realizar esta tarea utilizando los datos de entrada mencionados. Los genes identificados como resistentes pueden ser de gran ayuda al reducir el conjunto de genes candidatos que deben someterse a validación *in-vivo* en respuesta a un tratamiento específico.

4.2. Trabajo Futuro

El enfoque propuesto en este trabajo tiene un potencial amplio, ya que puede aplicarse a diferentes bases de datos, no solo limitado a distintos cultivos como el maíz o el arroz, sino también a diversos tipos de estrés, como bajas temperaturas, sequías, salinidad, entre otros. Al utilizar este proceso para identificar genes asociados a funciones biológicas y predecir su relevancia frente a infecciones (en este caso particular), se pueden obtener mejoras significativas en el sector agroindustrial y contribuir al desarrollo de la seguridad alimentaria. En este sentido, los genes seleccionados a partir de los resultados de este trabajo fueron predichos completamente con herramientas computacionales, y se requiere de experimentos en laboratorio para confirmar la relevancia, o no, mostrada en este trabajo.

La aplicabilidad de este enfoque a diversas bases de datos y escenarios de estrés permite una adaptabilidad y generalización a diferentes contextos agrícolas. Al identificar genes relacionados con funciones biológicas y su respuesta a diferentes tipos de estrés, se puede estrechar el rango de genes responsivos, evaluar de forma direccionada en laboratorio, o en campo, y así tomar medidas más precisas y específicas, todo con el fin de acelerar la respuesta y, quizás, optimizar la producción agrícola.

En resumen, la aplicación de este enfoque no se limita a cultivos específicos ni a un tipo particular de estrés, sino que puede ser una herramienta valiosa en el sector agroindustrial para mejorar la seguridad alimentaria a través de la identificación de genes asociados a funciones biológicas y su respuesta a diversos desafíos ambientales.

Bibliografía

[1] Maíz - Fenalce. URL: https://fenalce.co/product-category/semillas/maiz/.

[2] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. doi:10.1038/75556.

[3] Smriti Bhagat, Graham Cormode, and S. Muthukrishnan. Node Classification in Social Networks. In Charu C. Aggarwal, editor, *Social Network Data Analytics*, pages 115–148. Springer US, Boston, MA, 2011. doi:10.1007/978-1-4419-8462-3_5.

[4] Rohit Uttam Bhagwat and B. Uma Shankar. A novel multilabel classification of remote sensing images using XGBoost. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pages 1–5, Bombay, India, March 2019. IEEE. URL: https://ieeexplore.ieee.org/document/9033768/, doi:10.1109/I2CT45611.2019.9033768.

[5] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 874–883. PMLR, 13–18 Jul 2020. URL: https://proceedings.mlr.press/v119/bianchi20a.html.

[6] Carlos Andres Bustos Gonzalez. *Estudio De Competitividad Agrícola De La Provincia De Sabana Centro, Cundinamarca, Colombia. Caso Tipo De Lechuga (Lactuca Sativa), Maíz (Zea Mayz L.) Y Papa (Solanum Tuberosum).* PhD thesis, 2021.

[7] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Diffusion Component Analysis: Unraveling Functional Topology in Biological Networks. In *RECOMB 2015*, pages 62–64. Springer, Cham, April 2015.

[8] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Systems*, 3(6):540–548.e5, December 2016. doi:10.1016/j.cels.2016.10.017.

[9] Daniel Felipe Cruz, Sam De Meyer, Joke Ampe, Heike Sprenger, Dorota Herman, Tom Van Hautegem, Jolien De Block, Dirk Inzé, Hilde Nelissen, and Steven Maere. Using single-plant-omics in the field to link maize genes to functions and phenotypes. *Molecular Systems Biology*, 16(12), December 2020. doi:10.15252/msb.20209667.

[10] Minghua Deng, Kui Zhang, Shipra Mehta, Ting Chen, and Fengzhu Sun. Prediction of Protein Function Using Protein–Protein Interaction Data. *Journal of Computational Biology*, 10(6):947–960, December 2003. doi:10.1089/106652703322756168.

[11] Olaf Erenstein, Moti Jaleta, Kai Sonder, Khondoker Mottaleb, and B.M. Prasanna. Global maize production, consumption and trade: trends and R&D implications. *Food Security*, 14(5):1295–1319, October 2022. URL: https://link.springer.com/10.1007/s12571-022-01288-7, doi:10.1007/s12571-022-01288-7.

[12] A Norma Formento. Enfermedades foliares reemergentes del cultivo de maíz: royas (puccinia sorghi y puccinia polysora), tizón foliar (exserohilum turcicum) y mancha ocular (kabatiella zeae). *INTA, Argentina*, 2010.

[13] Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, January 2019. doi:10.1093/nar/gky1055.

[14] Jane Glazebrook. Genes controlling expression of defense responses in arabidopsis — 2001 status. *Current Opinion in Plant Biology*, 4(4):301–308, 2001. URL: https://www.sciencedirect.com/science/article/pii/S1369526600001771, doi:https://doi.org/10.1016/S1369-5266(00)00177-1.

[15] Vladimir Gligorijević, Meet Barot, and Richard Bonneau. deepNF: Deep network fusion for protein function prediction. *Bioinformatics*, 34(22):3873–3881, 2018. doi:10.1093/bioinformatics/bty440.

[16] B. Govaerts, D. Vega, X. Chávez, L. Narro, F. M. San Vicente, N. Palacios, M. Pérez, G. González, P. Ortega, A. Carvajal, A. L. Arcos, J. Bolaños, N. Romero, J. Bolaños, Y. F. Vanegas, R. G. Echeverria, A. Jarvis, D. Jiménez, J. Ramírez-Villegas, W. Kropff, C. E. González, C. Navarro-Racines, L. Ordóñez, S. D. Prager, and J. Tapasco. *Maíz para Colombia Visión 2030*. CIMMYT, 2019. URL: https://repository.cimmyt.org/handle/10883/20218.

[17] Carlos David Grande Tovar and Brigitte Sthepani Orozco Colonia. Producción y procesamiento del maíz en Colombia. *Revista Guillermo de Ockham*, 11(1):97, June 2013. URL: http://revistas.usb.edu.co/index.php/GuillermoOckham/article/view/604, doi:10.21500/22563202.604.

[18] Gloria Marcela Hoyos Gómez and Juan Esteban Ocampo. Producción y consumo del maíz en Colombia, descripción de la cadena y propuesta de estrategias para un mejor desempeño de la misma. *Fondo Editorial Biogénesis*, pages

95–112, October 2018. URL: https://revistas.udea.edu.co/index.php/biogenesis/article/view/336225.

[19] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, January 2009. doi:10.1038/nprot.2008.211.

[20] Hongjie Jia, Shifei Ding, Xinzheng Xu, and Ru Nie. The latest research progress on spectral clustering. *Neural Computing and Applications*, 24(7-8):1477–1486, June 2014. doi:10.1007/s00521-013-1439-2.

[21] González Rojas K, García Salazar JA, Matus Gardea JA, and Martínez Saldaña T. Vulnerabilidad del mercado nacional de maíz (zea mays l.) ante cambios exógenos internacionales. *Agrociencia [online]*, 45(6):733–44, September 2011. URL: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-31952011000600008&lng=es&nrm=iso.

[22] Shehroz S. Khan and Michael G. Madden. A Survey of Recent Trends in One Class Classification. In Lorcan Coyle and Jill Freyne, editors, *Artificial Intelligence and Cognitive Science*, volume 6206, pages 188–197. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi:10.1007/978-3-642-17080-5_21.

[23] Saet-Byul Kim, Lisa Van den Broeck, Shailesh Karre, Hoseong Choi, Shawn A. Christensen, Guan-Feng Wang, Yeonhwa Jo, Won Kyong Cho, and Peter Balint-Kurti. Analysis of the transcriptomic, metabolomic, and gene regulatory responses to *Puccinia sorghi* in maize. *Molecular Plant Pathology*, 22(4):465–479, April 2021. URL: https://onlinelibrary.wiley.com/doi/10.1111/mpp.13040, doi:10.1111/mpp.13040.

[24] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*, November 2017. arXiv:1705.07874.

[25] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

[26] Feng Luo, Yunfeng Yang, Jianxin Zhong, Haichun Gao, Latifur Khan, Dorothea K Thompson, and Jizhong Zhou. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics*, 8(1):299, December 2007. doi:10.1186/1471-2105-8-299.

[27] Peter Mills. Solving for multi-class: A survey and synthesis. *arXiv:1809.05929 [cs, stat]*, January 2021. arXiv:1809.05929.

[28] Cristian Millán Hernández. Análisis de la estructura y competitividad de la cadena productiva de maíz Zea mays. *Administración de Agronegocios*, January 2015. URL: https://ciencia.lasalle.edu.co/administracion_agronegocios/95.

[29] Luz Eliana Hernández Montoya, Eduardo Javid Corpas Iguarán, and Katherin Castro Ríos. Vida útil en masas y productos derivados del maíz: estudio bibliométrico. *Brazilian Journal of Food Technology*, 23:e2019023, 2020. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1981-67232020000100473&tlng=es, doi:10.1590/1981-6723.02319.

[30] Nivedha Murugesan, Irene Cho, and Cristina Tortora. Benchmarking in Cluster Analysis: A Study on Spectral Clustering, DBSCAN, and K-Means. In *Data Analysis and Rationality in a Complex World*, pages 175–185. Springer, Cham, 2021. URL: http://link.springer.com/10.1007/978-3-030-60104-1_20.

[31] T. Obayashi and K. Kinoshita. COXPRESdb: A database to compare gene coexpression in seven model animals. *Nucleic Acids Research*, 39(Database):D1016–D1022, January 2011. doi:10.1093/nar/gkq1147.

[32] Takeshi Obayashi, Yuichi Aoki, Shu Tadaka, Yuki Kagaya, and Kengo Kinoshita. ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index. *Plant and Cell Physiology*, 59(1):e3–e3, January 2018. doi:10.1093/pcp/pcx191.

[33] Diego Hernán Peluffo Ordóñez. Agrupamiento Espectral Multiclase Basado en Particiones Normalizadas. *Cuaderno activa*, 5:39–49, 2013. URL: https://ojs.tdea.edu.co/index.php/cuadernoactiva/article/view/112.

[34] Jesús Efrén Ospina Noreña, Gustavo Adolfo Ligarreto Moreno, Nixon Flórez Velasco, Andrés Leonardo Leguizamón García, Christian Camilo Pimentel Ladino, Saúl Roberto Murcia López, and Alefsi David Sánchez Reinoso. *Diagnóstico socio - técnico y de clima en los cultivos de frijol y maíz en las regiones de Ubaté y Guavio en Cundinamarca*. Centro Editorial Facultad de Ciencias Agrarias, December 2017. URL: https://repositorio.unal.edu.co/handle/unal/77630.

[35] Martin Oti, Jeroen van Reeuwijk, Martijn A Huynen, and Han G Brunner. Conserved co-expression for candidate disease gene prioritization. *BMC Bioinformatics*, 9(1):208, 2008. doi:10.1186/1471-2105-9-208.

[36] Diego Hernán Peluffo Ordoñez. *Agrupamiento espectral de datos dinamicos*. PhD thesis, Universidad Nacional de Colombia Sede Manizales Facultad de Ingeniería y Arquitectura Departamento de Ingeniería Eléctrica, Electrónica y Computación, 2013.

[37] Gregory A Petsko. Guilt by association. *Genome Biology*, 10(4):104, 2009. doi:10.1186/gb-2009-10-4-104.

[38] Iris Pérez Almeida and Pedro García Mendoza. Aportes de la biotecnología al mejoramiento del maíz. *Revista Peruana de Innovación Agraria - ISSN: 2810-8876 (En línea)*, 1(1):130 – 150, nov. 2020. URL: http://200.123.25.14/index.php/REVINIA/article/view/10.

[39] Mayra Z. Rodriguez, Cesar H. Comin, Dalcimar Casanova, Odemir M. Bruno, Diego R. Amancio, Luciano da F. Costa, and Francisco A. Rodrigues. Clustering algorithms: A comparative approach. *PLOS ONE*, 14(1):e0210236, January 2019. doi:10.1371/journal.pone.0210236.

[40] Miguel Romero, Jorge Finke, Mauricio Quimbaya, and Camilo Rocha. In-silico Gene Annotation Prediction Using the Co-expression Network Structure. In *Complex Networks and Their Applications VIII*, pages 802–812. Springer, 2020.

[41] Miguel Romero, Jorge Finke, and Camilo Rocha. A top-down supervised learning approach to hierarchical multi-label classification in networks. *Applied Network Science*, 7(1):8, December 2022. doi:10.1007/s41109-022-00445-3.

[42] Miguel Romero, Óscar Ramírez, Jorge Finke, and Camilo Rocha. Supervised Gene Function Prediction Using Spectral Clustering on Gene Co-expression Networks. In Rosa Maria Benito, Chantal Cherifi, Hocine Cherifi, Esteban Moro, Luis M. Rocha, and Marta Sales-Pardo, editors, *Complex Networks & Their Applications X*, volume 1016, pages 652–663. Springer International Publishing, Cham, 2022. doi:10.1007/978-3-030-93413-2_54.

[43] Miguel Romero, Oscar Ramírez, Jorge Finke, and Camilo Rocha. Feature extraction with spectral clustering for gene function prediction using hierarchical multi-label classification. *Applied Network Science*, 7(1):28, December 2022. URL: https://appliednetsci.springeropen.com/articles/10.1007/s41109-022-00468-w, doi:10.1007/s41109-022-00468-w.

[44] Alistair G. Rust, Emmanuel Mongin, and Ewan Birney. Genome Annotation Techniques: New Approaches and Challenges. *Drug Discovery Today*, 7(11):S70–S76, May 2002. doi:10.1016/S1359-6446(02)02289-4.

[45] Carlos N. Silla and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, January 2011. doi:10.1007/s10618-010-0175-9.

[46] Sri Bananiek Sugiman, Zainal Abidin, and Muh. Asaad. Implementation and farmer perception of corn seed production technology in Southeast Sulawesi. *IOP Conference Series: Earth and Environmental Science*, 484(1):012128, April 2020. URL: https://iopscience.iop.org/article/10.1088/1755-1315/484/1/012128, doi:10.1088/1755-1315/484/1/012128.

[47] Giorgio Valentini. True Path Rule Hierarchical Ensembles. In *Multiple Classifier Systems*, pages 232–241. Springer, 2009.

[48] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, page bbw139, January 2017. doi:10.1093/bib/bbw139.

[49] D. Van Den Poel, C. Chesterman, M. Koppen, and M. Ballings. Equity price direction prediction for day trading: Ensemble classification using technical analysis indicators with interaction effects. pages 3455–3462.

Institute of Electrical and Electronics Engineers Inc., 2016. cited By 0; Conference of 2016 IEEE Congress on Evolutionary Computation, CEC 2016 ; Conference Date: 24 July 2016 Through 29 July 2016; Conference Code:124911. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85008245281&doi=10.1109%2fCEC.2016.7744227&partnerID=40&md5=19d391540e137abbd214f4dd4d4d3873, doi:10.1109/CEC.2016.7744227.

[50] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, November 2008. doi:10.1007/s10994-008-5077-3.

[51] Xin Wu, Yuchen Gao, and Dian Jiao. Multi-Label Classification Based on Random Forest Algorithm for Non-Intrusive Load Monitoring System. *Processes*, 7(6):337, June 2019. URL: https://www.mdpi.com/2227-9717/7/6/337, doi:10.3390/pr7060337.

[52] Donna Xu, Yaxin Shi, Ivor W. Tsang, Yew-Soon Ong, Chen Gong, and Xiaobo Shen. Survey on multi-output learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2409–2429, 2020. doi:10.1109/TNNLS.2019.2945133.

[53] Mark Yandell and Daniel Ence. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5):329–342, May 2012. doi:10.1038/nrg3174.

[54] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–515, July 2008. doi:10.1038/nrg2363.

[55] Yingwen Zhao, Guangyuan Fu, Jun Wang, Maozu Guo, and Guoxian Yu. Gene function prediction based on Gene Ontology Hierarchy Preserving Hashing. *Genomics*, 111(3):334–342, 2019. doi:10.1016/j.ygeno.2018.02.008.

[56] Guangjie Zhou, Jun Wang, Xiangliang Zhang, Maozu Guo, and Guoxian Yu. Predicting functions of maize proteins using graph convolutional network. *BMC Bioinformatics*, 21(S16):420, December 2020. doi:10.1186/s12859-020-03745-6.

[57] Y. Zhou, J. A. Young, A. Santrosyan, K. Chen, S. F. Yan, and E. A. Winzeler. In-silico gene function prediction using ontology-based pattern identification. *Bioinformatics*, 21(7):1237–1245, April 2005. doi:10.1093/bioinformatics/bti111.