**Análisis computacional de los interactores de efectores bacterianos en su planta hospedera: ¿Qué información se podrá dilucidar con estas herramientas de vanguardia?**

**1. Introducción**

La bioinformática es una disciplina científica que surgió por la necesidad de estudiar datos biológicos, que se pueden definir, de manera sofisticada, como información de secuencias moleculares, contenido genómico y análisis del producto génico, fenotípico (1). Así, esta disciplina integra interdisciplinariamente las ciencias exactas y naturales con la computación, en orden de solucionar cualquier problema relacionado a la biología o entender el funcionamiento de los fenómenos que componen a los seres vivos (1). Los métodos generados dentro de esta área han sido utilizados en una gran cantidad de áreas, incluyendo biología molecular, genómica, biología de sistemas, biotecnología, agricultura, biomédica e incluso medicina (1). Se puede entender, entonces, cómo la biología computacional ha revolucionado la forma en la cual se estudian los organismos vivos, de la mano de la biología molecular, resaltando así su gran importancia y contribución al estado del arte del conocimiento. Por ejemplo, la base de datos BLAST de ncbi es una herramienta computacional de análisis de secuencias, disponible para cualquiera que posea una conexión a internet (2). Sus algoritmos permiten buscar cualquier secuencia de ADN, ARN u proteínas en sus bases de datos, en un tiempo computacional muy rápido y, si se ha secuenciado previamente, los resultados proveerán el organismo al cual pertenece, en qué región de su genoma se encuentra, qué función posee e, incluso, el porcentaje de identidad con la secuencia buscada (2). Esto ha permitido soportar innumerables estudios en el campo de la biología.

Concomitante a la aparición de los medios moleculares para estudiar la composición más básica de los organismos, como la secuenciación de nucleótidos para el estudio de genes y genomas, por ejemplo, sucede el surgimiento de las “ómicas”, ramas de la biología molecular dedicadas al estudio del comportamiento de diversos sistemas moleculares que componen organismos y ecosistemas (3). En este caso de estudio es importante extender la discusión en el sentido de una de estas ramas, la interactómica (4). En específico, ésta integra la proteómica y a genómica, estudiando las redes de interacción proteica, las cuales, no sólo dependen de la expresión génica, sino, de muchos otros factores celulares. Estas redes de interacción buscan dilucidar, también, las interacciones proteína-proteína (PPI) (4). En el estudio de esta materia se han estandarizado tres formas de aproximación principales, métodos *in silico*, *in vitro* e *in vivo*.

En cuanto a los métodos *in silico* existen varios enfoques específicos (4). En general, se suelen hacer búsquedas en bases de datos por medio de minería de datos, esto, con la finalidad de hallar posibles interactores (término utilizado para referirse a proteínas que presentan PPI), como un indicio para el estudio posterior con otras herramientas (4). Usualmente, estos se pueden realizar como un complemento analítico a los métodos experimentales, puesto que aún se requiere validación en orden de disminuir la cantidad de falsos positivos (4).

Existen tres enfoques principales en la metodología *in silico*. Primero, la búsqueda de ortólogos, la cual se refiere a encontrar proteínas “homólogas”, es decir, que compartan origen evolutivo común con aquella de estudio (citar), en organismos ya caracterizados o “modelo”, con el fin de encontrar candidatos para PPI en el organismo de interés o, si ya se han validado estas interacciones obteniendo datos pertinentes, relacionar la funcionalidad de las proteínas halladas con ortólogos que ya han sido estudiados (4). Este enfoque halla su fundamento en el hecho que muchas proteínas ortólogas conservan su función en el proceso evolutivo, por lo que no deben usarse parálogos. Estos se refieren a homólogos que han surgido por medio de duplicaciones génicas (5). Esto implica que, en muchos casos, adquieren funcionalidades diferentes entre sí, por lo que no son ideales para este método. En segundo lugar, se encuentra el acercamiento basado en estructura. Este se refiere a la utilización de modelos tridimensionales computarizados de las proteínas, los cuales, poseen información muy precisa de los dominios proteicos su superficie y, en algunos casos, su afinidad de unión (4). Por este medio, se pueden predecir PPI, pues se ha observado que, si dos proteínas poseen dominios similares, es probable que éstas interactúen entre sí (4). Claramente, en muchos casos no existen modelos proteicos para aquella proteína que se esté estudiando, por lo que se pueden buscar indicios por ortólogos o, incluso, realizar un modelaje experimental *de novo*. Finalmente, el acercamiento basado en coevolución se encuentra basado en el hecho que, si dos proteínas han estado sometidas a un proceso de coevolución, dado por presión selectiva, en función de una relación funcional estas, se puede sospechar, van a encontrarse juntas en el proceso de heredabilidad (4). Así, al estudiar la presencia o ausencia de pares proteicos, de manera recurrente, se pueden predecir PPI´s entre ellas (4). Aun así, este método posee varias limitaciones relacionadas con eventos aleatorios que suceden en el proceso evolutivo (4).

En orden de permitir que estos métodos computacionales se nutran de información real y de validar las predicciones de posibles PPI´s, se deben llevar a cabo experimentos de interactómica. Uno de los diseños experimentales más utilizados es el ensayo de doble híbrido. Consiste en utilizar el sistema de regulación de la expresión génica de levaduras GAL4, específicamente, el dominio DBD de unión a DNA y el dominio AD de activación (7). El primero se encarga de reclutar ciertos factores de transcripción a la región génica específica y, el segundo, permite el reclutamiento de factores de activación (6). La base yace en el hecho que estos dominios, aunque existen dentro de una misma proteína en la naturaleza, pueden actuar en trans (estar presentes en diferentes proteínas (6). Así, se preparan dos plásmidos, como mínimo, pero dependiendo de la naturaleza del objetivo experimental y la cantidad de proteínas que se desean evaluar, esto puede cambiar. En uno se ubica el cDNA de la proteína “cebo” fusionada con el dominio BD (6, 7). En el otro, se ubica el cDNA de la proteína   
“presa” con el dominio AD. En caso de existir interacción entre ambas proteínas, el acercamiento físico de los dominios causará el evento de iniciación de la transcripción, en este caso, de un gen reportero previamente clonado en la célula de levadura corriente debajo de la secuencia a la que se une el BD (6, 7). Así, el gen reportero permitirá medir los niveles de interacción por medio del grado de expresión de este y su reporte (7). Una de las principales limitaciones de este método es la presencia de falsos positivos, debido a que suelen presentarse interacciones no específicas que pueden activar la cascada transcripcional generando señales que no necesariamente corresponden a interacciones entre las proteínas reportadas (7). No obstante, se pueden realizar correcciones estadísticas sobre los niveles de expresión en orden de corregir este tipo de error. Este método sigue siendo muy útil para el estudio de la interactómica.

Con estas técnicas, se pueden abordar diversos sistemas biológicos, estudiándolos desde la perspectiva de redes de interacción. Uno de los sistemas más estudiados por estos mecanismos es la interacción planta patógeno, dada su importancia académica y práctica, sobre todo en el campo de la agricultura, tema que se tratará más adelante. Dado esto, es imperativo entender el contexto de esta relación entre organismos. Esta relación es muy compleja debido a la gran cantidad de eventos físicos y químicos que la componen. Se puede abordar como una compleja red de interacciones entre moléculas producidas por la planta y aquellas secretadas por la bacteria. El primer contacto inicia con la llegada de la bacteria al tejido de la planta, donde debe superar las barreras físicas de ésta, compuesta por la cutícula, la pared celular, la membrana celular y los compuestos antimicrobianos de expresión constitutiva (8, 9). La principal forma en la cual superan las defensas físicas es ingresando por medio de las estomas, heridas, lenticelas u otros puntos débiles (10). La planta posee PRRs, receptores de membrana de reconocimiento de patrones moleculares que reconocen sustancias específicas, usualmente endógenas, de las bacterias patógenas (8). Estas moléculas se conocen como PAMPs, definidos como patrones moleculares asociados a patógenos (8). Así, después de este proceso de reconocimiento, la planta activa una respuesta conocida como PTI (Inmunidad desencadenada por PAMP) de carácter general y bastante efectiva en la mayoría de los casos, pues es suficiente para defender a la planta de la posible infección (8, 9). Adicionalmente, las células vegetales pueden generar más respuestas parecidas reconociendo DAMPs o marcadores de daño celular (10) que son producto del daño ejercido por el patógeno sobre la célula vegetal..

La respuesta PTI está asociada a la vía de señalización MAPK (10). Posterior a la transducción de señales intracelulares se activan respuestas celulares como el cierre estomático, limitando la carga bacteriana, producción de ROS, producción de compuestos antimicrobianos como las fitoalexinas y disminución del paso de nutrientes del citoplasma al apoplasto, disminuyendo así la tasa de crecimiento bacteriana, entre otros (10). En muchos casos, los patógenos producen proteínas efectoras (8). Su mecanismo de acción es la inhabilitación de la respuesta PTI, suprimiendo los componentes moleculares involucrados en la formación de la respuesta (10). Este mecanismo se conoce como susceptibilidad elicitada por Efectores (ETS por sus siglas en inglés)

El segundo nivel de respuesta vegetal es específico hacia una proteína efectora bacteriana. Éste se da por la expresión de un factor de resistencia (proteína R) activando así una segunda fase de inmunidad conocida como ETI (Inmunidad desencadenada por efector) (8, 9). En una gran cantidad de los casos, esta respuesta es acompañada de un efecto colateral conocido como respuesta hipersensible, un tipo de muerte celular programada , donde el tejido vegetal muere de forma localizada (8). Una ETI puede ser tan simple como la interacción de sólo un efector patógeno y una proteína R específica. Así, se hace énfasis en la importancia de considerar la respuesta ETI como una red de interacción de proteínas. Estas interacciones y respuestas de un lado al otro se conocen como el modelo de zig-zag (10).

Esta interacción también se debe abordar desde la naturaleza del patógeno. Existen tres categorías principales. Los biótrofos se alimentan de tejido vegetal vivo (11). Los necrótrofos, por otra parte, requieren que las células mueran induciendo este evento por medio de toxinas producidas por la bacteria invasora (11). Finalmente, los hemibiótrofos combinan ambas estrategias, donde, en la primera fase biótrofa, secretan sustancias que inhiben la muerte celular programada de la planta y, posterior a esto, una segunda fase necrótrofa, donde secretan toxinas para generar la muerte de la célula y alimentarse de los compuestos degradados (11). En el caso de las bacterias es difícil clasificar de forma rígida a determinada especie debido a que muchas presentan resultados experimentales variables (12). Es importante mencionar que *Xanthomonas.spp*, bacteria objeto en este estudio, suele ser considerada como un patógeno hemibiótrofo, a tal punto, que produce efectores, como AvrBs3, para producir hipertrofia celular en su hospedero (12, 13). La importancia de esta clasificación para patógenos yace en la respuesta que la planta dispone para combatir a cada tipo de agente infeccioso (12). La respuesta sistémica adquirida (SAR) está determinada por la señalización por parte del ácido salicílico y es activada por un agente biótrofo o hemibiótrofo, relacionado, por ende, con la respuesta hipersensible localizada (16). Por otro lado, la respuesta de resistencia sistémica inducida (ISR) es activada por patógenos biótrofos por medio de sustancias como el metil-jasmonato y el etileno, por lo que suele tener como objetivo la manutención de sus células vivas y la producción de agentes antimicrobianos os sustancias inhibitorias (14, 16) (15). Usualmente, algunos organismos de la rizosfera generan un tipo de predisposición activando la vía del metil-jasmonato/etileno cuyo efecto es una posterior respuesta ISR mucho más eficiente y rápida al entrar la planta en contacto con un patógeno (14, 15). Las interacciones planta patógeno son muy complejas dada su larga historia de coevolución, así que es en este punto donde se recalca la importancia de herramientas operacionales eficientes de bioinformática que permiten estudiar estos fenómenos de forma mucho más fácil y completa.

Los efectores bacterianos tienen blancos proteicos en la planta muy específicos, así, como la funcionalidad que modifican en base a esta interacción. Aún así, se ha descubierto que efectores con historias evolutivas independientes muchas veces convergen en cuanto a sus blancos, de manera directa e indirecta, afectando sistemas celulares inmunes de la planta en conjunto, concepto reconocido como “hubs celulares” (17). En cuanto a efectores, se debe considerar el hecho que el proceso de coevolución planta patógeno ha sucedido por millones de años, por lo que, por medio de adaptaciones y exaptaciones, la variedad de estas proteínas que ha surgido es muy alta. Por lo tanto, su mecanismo de acción varía mucho en función de la fisiología del hospedero. Muchos patógenos poseen efectores cuya función es superar la barrera física descrita previamente, estos, entonces, son proteínas de degradación, por ejemplo, de la pared celular vegetal, que permitirán la entrada del parásito (17). Como respuesta a mecanismos PTI descritos, algunos efectores evitan el cierre de estomas alterando la vía de señalización del metil-jasmonato (16). Relacionado con esta primera línea de defensa, algunos otros bloquean la vía de señalización MAPK, disminuyendo la efectividad del reconocimiento de PAMP´s actuando sobre los PRR, enmascaran PAMPs específicos o inhiben enzimas proteolíticas del hospedero disminuyendo así la efectividad de la respuesta PTI (16). Dependiendo del tipo de patógeno y su clasificación según el estado celular de la planta, los efectores pueden ser producidos con el objetivo de generar la muerte celular o, incluso, mantener a la célula viva inhibiendo la respuesta de hipersensibilidad (16). Estos efectos se consiguen por medio de la disrupción de las vías de metil-jasmonato/etileno y ácido salicílico ya descritas, dependiendo de tipo de patógeno y sus requerimientos nutricionales (16). Por último, con una gran cantidad de fines, muchos efectores han evolucionado con la capacidad de alterar la expresión génica en su hospedero de diversas maneras, por ejemplo, alterando el silenciamiento por medio de RNA, inhibiendo de manera directa algunos factores de transcripción de la planta o, incluso, por medio de TALE´s, efectores que actúan directamente como factores de transcripción específicos para diversos promotores en el genoma vegetal. Dada la complejidad de estos mecanismos, se requiere un gran poder de análisis para poder estudiar estos fenómenos.

La relevancia del estudio de interacciones planta-patógeno radica en muchos aspectos. En cuanto al impacto que las enfermedades en plantas producen en nuestra sociedad, las implicaciones son inmensas. Este hecho se evidencia con la estimación que, debido sólo a las enfermedades en cultivos de consumo humano, existen alrededor de 800 millones de personas a nivel mundial que no poseen una alimentación adecuada (18). Sólo en relación con la enfermedad causada por el patógeno en estudio, la “Mancha foliar”, se puede aseverar que comprende una alta proporción de los casos de enfermedad en cultivos y afecta tanto a monocotiledóneas como dicotiledóneas (19). En orden de solucionar este problema, los esfuerzos se han decantado a la producción de organismos genéticamente modificados por medio del uso de DNA recombinante, produciendo cepas con mejor resistencia a enfermedades (20). Teniendo esto en cuenta, es claro que, si se logran identificar las interacciones específicas entre los efectores de patógenos y sus blancos en las plantas, estas modificaciones genéticas para conferir resistencia a los cultivos se harían inmensamente más eficientes. Es aquí donde un estudio como este adquiere una gran relevancia a nivel práctico.

El objetivo de este estudio radica en dilucidar redes de interacción planta-patógeno dados los datos de un ensayo doble híbrido entre *Manihot esculenta* y *Xanthomonas*, utilizando una gran cantidad de herramientas bioinformáticas y computacionales, encontrar los blancos específicos en términos de rutas metabólicas y enzimas específicas de los 16 efectores bacterianos estudiados y determinar qué factores a nivel de biología de sistemas e interacción de dominios proteicos poseen una alta relevancia en estos fenómenos de patogenicidad. Con base en los resultados de estos análisis, se estudiarán en detalle algunos puntos específicos de esta interactómica estudiada, los cuales, posean importancia biológica y funcional relevante y hubs de interacción que resulten de interés académico o práctico. Se espera generar una metodología y flujo de trabajo simplificado, por medio del cual se puedan estudiar relaciones de este tipo en otros casos de interacciones planta-patógeno con un valor predictivo y eficiencia altos, facilitando así el estudio en esta área. En este caso de estudio específico, se postula que existen relaciones claras en cuanto a los blancos metabólicos y redes enzimáticas en función de los efectores estudiados y que estos interactúan indirectamente entre sí, por lo cual, se prevé que, por medio del análisis computacional, se encontrarán diferentes blancos en común para grupos de efectores en particular. Paralelo a esto, se postula que existen dominios proteicos en los efectores que determinan sus blancos en el hospedero y, en gran parte, las relaciones propuestas anteriormente. Con base en este hecho, se encontrarán redes de interacción, en el sentido conceptual de la biología de sistemas, con un alto grado de importancia a nivel del desarrollo de la enfermedad infecciosa por *Xanthomonas* en plantas de este tipo. Por último, se postula que existen interacciones específicas de dominio entre diferentes efectores e interactores, que se repiten con mayor frecuencia. Esto significa que la identificación de estos dominios “usuales” y su modelamiento fisicoquímico permitirá dilucidar nuevos aspectos de este tipo de interacciones que no se han estudiado a profundidad.

**2. Resumen**

Gracias a la revolución de la tecnología en las últimas décadas, la ciencia también ha sufrido un cambio impresionante. La biología no es la excepción y es aquí donde nace la bioinformática. No obstante, así como las herramientas disponibles ahora son inconmensurables, así mismo lo son las posibilidades que todavía existen. Esto es aun más notorio en campos como la interactómica, donde hacen falta metodologías y algoritmos de análisis, diseñados específicamente, para suplir las necesidades de este campo. El estudio de la interacción planta patógeno se presenta como un área perfecta para este propósito. No solo es extremadamente compleja, sino que, además, sus posibles aplicaciones acogen desde la mejoría significativa de la agroindustria y todas las repercusiones en la sociedad que esto conlleve, hasta áreas como ecología y conservación. Es por esto que este estudio propone la integración de muchas herramientas computacionales utilizando como caso de estudio la interacción planta patógeno entre *Xanthomonas spp.* y *M. esculenta* con el din de dilucidar parte de los fenómenos que componen la interacción planta patógeno. Se postula que los efectores bacterianos interactúan con redes metabólicas muy específicas y que esta especificidad se explica, en gran parte, por la naturaleza de los dominios proteicos involucrados.

**3.Objetivo general**

-Caracterizar y dilucidar redes de interacción metabólica en *Manihot esculenta* por la bacteria fitopatógena, *Xanthomonas phaseoli pv. manihotis (Xpm)*. con el fin de aportar evidencia y conocimiento al fenómeno de interacción planta patógeno.

**4. Objetivos específicos**

-Estudiar rutas y enzimas específicas, cuyos análisis demuestren una alta relevancia dentro del mecanismo de patogenicidad de *Xpm* en yuca.

-Crear un flujo de trabajo, metodología, métodos y funciones que faciliten el análisis bioinformático de datos de estudios de interacción proteína-proteína, enumerando las herramientas utilizadas de forma clara.

-Generar un modelo de análisis fisicoquímico para la interacción entre dominios proteicos que se encuentren en una alta frecuencia en el análisis de interacción P-P.

**5. Métodos**

***Análisis inicial: DESeq2***

La información obtenida a partir del ensayo doble híbrido provee resultados parecidos a los que se obtendrían de un experimento de expresión de RNA. Gracias a este hecho, se utiliza el algoritmo DESeq2, en forma de librería en el lenguaje R, para analizar estos datos y transformarlos a valores cuantificables de expresión génica, evaluados entre la interacción de cada efector bacteriano y cada uno de los genes de la planta. Esto, gracias al análisis de expresión diferencial que provee este método, donde se comparan los resultados experimentales con valores de referencia y se provee un estimado de expresión en forma de valor p y expresión logarítmica en base dos, denominada como LogFoldChange2. Además, se ajusta el valor P normalizando los datos en base a la distribución binomial negativa, usual en este tipo de estudios.

***Filtrado de interactores y extracción de datos relevantes***

El lenguaje R provee muchas herramientas nativas para la manipulación de las estructuras de datos de tipo DataFrame, es decir, matrices donde, las columnas, representan las variables de cada punto de datos, es decir, cada fila. Así, cada columna posee un valor que identifica a la variable y puede ser referenciado como una “llave”. El flujo de trabajo para completar este objetivo se establece de forma que se obtienen nuevos DataFrames para cada efector, conteniendo tres variables, identificador de cada gen de yuca, su valor p de interacción con el efector y su valor de LogFoldChange2. Esto se lleva a cabo con la función subset(), que permite discriminar sólo los datos que cumplan cierta condición de acuerdo al valor de las variables especificadas. Para el filtrado en este caso, se utiliza valor p < 0.01 y LogFoldChange2 > a 2. Se utiliza la función quantile(), para obtener los cuantiles 0.05, 0.25, 0.5, 0.75, 0.95, de cada uno de los subsets obtenidos. Se realiza un histograma para cada set de datos que muestra la distribución de los valores p y del LogFoldChange2, por medio de la librería ggplot2. Finalmente, se utiliza la función write.table(), para obtener un documento en formato csv de cada uno de los DataSets.

**Extracción de identificadores en base de datos genómica**

Se obtiene la base de datos del genoma de *Manihot esculenta* de phytozome en su versión 7.1. Para estos análisis, se utiliza el clúster de alto rendimiento de la Universidad de los Andes. Este posee una versión de servidor del sistema operativo GNU/Linux, por lo cual provee una gran cantidad de software para diferentes tipos de análisis, entre ellos, programas como Awk, Sed, grep, entre otros. Para utilizar este tipo de servidores se debe establecer una conexión remota ssh, esto permite la entrada y salida de paquetes de información desde la maquina local y el servidor de forma remota y segura. Se crea un directorio de trabajo donde se ubica la base de datos y cada uno de los data sets. A partir de esto, se crea un script shell que permite realizar acciones consecutivas en la terminal del sistema operativo. Esto facilita el trabajo automatizado y la eficacia de los análisis. En este script se establecen varios comandos, donde su pseudocódigo puede ser explicado de la siguiente manera:

-Crear las variables que representen a cada data set, facilitando su diferenciación posterior.

-Por medio de la herramienta grep, buscar cada elemento de la columna que representa el identificador de *M. esculenta y* compararlos con esta misma variable, posición por posición, en la base de datos. Si y solo si las dos cadenas de caracteres comparadas son exactamente iguales, extraer, a partir de la misma fila en la base de datos, el identificador de *Arabidopsis thaliana* (organismo modelo) y el número “ec” (identificador enzimático). Esta es una implementación del algoritmo de Boyer-Moore, ampliamente utilizado para la identificación de expresiones regulares.

-Por medio de la herramienta Awk, concatenar estas columnas extraídas como variables a los Data Sets de cada efector. Este comando es muy simple, ya que Awk tiene la capacidad de manipular columnas como si fueran variables individuales.

-Por medio de un ciclo que recorre la lista de identificadores de cada data set, se busca cada uno de estos elementos en una base de datos en formato FASTA para proteínas. El algoritmo programado busca en cada línea que comience con el símbolo ‘>’ la expresión regular que representa a cada identificador. Si y sólo si son exactamente iguales, esta se copiará en un nuevo archivo, incluyendo el símbolo y las siguientes líneas, que representan la secuencia FASTA, hasta que se identifique un nuevo símbolo ‘>’. Al suceder esto, se procede al siguiente identificador y así sucesivamente. Si un identificador no se encuentra en la base de datos, éste se escribe en el nuevo archivo en formato FASTA, sin su respectiva secuencia.

**Anotación enzimática en KAAS**

Con el fin de identificar las rutas metabólicas y las enzimas dentro de ellas que se obtuvieron en el análisis, se utiliza la herramienta de anotación en línea automática de KAAS. El formato de ingreso de los datos en un archivo FASTA que incluya las proteínas a identificar. Como argumento, se escoge el identificador de esta especie, en este caso, ‘mesc’. Como resultado, se obtiene cada una de las rutas identificadas que contengan una o más enzimas provenientes del archivo de entrada. Se generan imágenes que representan cada una de las rutas y resaltan las enzimas identificadas, así como un archivo con cada identificador y su número ‘ko’. Este número representa, con base en su ortólogo, en este caso, *A. thaliana*, las rutas metabólicas en donde se encuentra involucrado cada interactor.

**Elaboración de matriz de rutas metabólicas**

Se realiza una matriz donde se ubican como puntos de datos las filas, representando cada una de las rutas identificadas en la base de KAAS, cuyas columnas son cada uno de los efectores bacterianos. En cada casilla se realiza un conteo manual de la cantidad de interactores por cada ruta por efector. Esta matriz se archiva como un archivo csv. Así mismo, identificando la clasificación de KAAS para estas rutas, se realizan dos nuevas matrices, con la misma estructura, sumando las filas cuyas rutas pertenezcan a estos grupos de metabolismo general. Dada la falta de información para algunas de las enzimas, la base de KAAS identifica rutas que no pertenecen a plantas. Estas son removidas, a excepción de aquellas que puedan pertenecer a organelos endosimbiontes.

**Identificación general de categorías enzimáticas**

En orden de identificar las categorías enzimáticas generales se utiliza el lenguaje de programación Python. Se cargan los datasets obtenidos, previamente, por cada efector, como objetos de la clase DataFrame de la librería Pandas. Así, esta información se encuentra disponible para posteriores análisis. Se crea una función que recibe los DataFrames como parámetro, toma la columna cuya llave representa los números “ec” y, por medio de un ciclo de recorrido total, de acuerdo al primer carácter de cada secuencia, asigna el tipo de enzima basado en la clasificación de la Unión Internacional de Bioquímica y Biología molecular. Estos tipos se guardan en una lista que es retornada por esta función. Se utiliza la librería Matplotlib y su capa de “scripting” para realizar diagramas de barras que representan la cantidad de enzimas identificadas por cada uno de sus tipos por efector.

**Complementación de bases de datos de efectores con números ko**

Por medio del método merge(), de la clase DataFrame de Pandas, se realiza un “inner join”, es decir, una concatenación de bases de datos basada en los puntos de datos cuyos valores son iguales, lo cual representa sus puntos de intersección. Este método se aplica sobre la base de cada efector y se utiliza como parámetro el archivo obtenido de KAAS que contiene los identificadores de genes con su respectivo número ko. Esta unión se realiza con base en los valores de los identificadores y se concatena la columna de ko con la base original, en cada valor respectivo.

**Análisis de clusters de rutas enzimáticas: K-means**

En este caso, se utiliza la librería Sci-kit learn de Python. Se escoge el método constructor Kmeans(), para crear un modelo y ajustarlo a los datos. Los métodos pertenecientes a esta clase permiten determinar la agrupación de clusters basados en cada una de las rutas metabólicas evaluadas y su respectiva posición en un espacio de 16 dimensiones, cada una, representa el número de enzimas identificadas en ésta para cada efector. En base a la cantidad de clusters especificados como parámetros en el constructor, se crea esta misma cantidad de centroides en el espacio. Así, a partir de la medición de las distancias euclidianas a partir de cada data point, se calcula una agrupación específica para todos los datos, de tal manera que se optimicen estas distancias. Para determinar el número óptimo de centroides se utiliza el método de la máxima silueta. A continuación se describe el pseudocódigo de este flujo de trabajo:

-Se normaliza la matriz de rutas enzimáticas por medio de su conversión a valores porcentuales por cada columna (efector).

-Se genera un ciclo cuyo elemento de control será una lista de valores posibles para el centroide a utilizar. Dentro del ciclo, se asigna un objeto KMeans(), cuyo argumento para el número de clusters será el mismo contador del ciclo. Se utiliza el método fit\_predict(), para ajustar el modelo a los datos de la matriz enzimática. Posterior a esto, se utiliza la función silhouette\_score(), cuyos argumentos son los datos de la matriz y el modelo ajustado, permitiendo obtener el score de silueta de este modelo. Este valor es el promedio de la distancia euclidiana de los datos con respecto a su centroide en el espacio de 16 dimensiones. Este procedimiento se itera según la cantidad de valores de centroide que se requiera probar. Estos valores se guardan en listas para facilitar su análisis. Según una gráfica donde el puntaje de silueta dependa del número de clusters, se escoge este último valor basado en el mayor valor de la silueta.

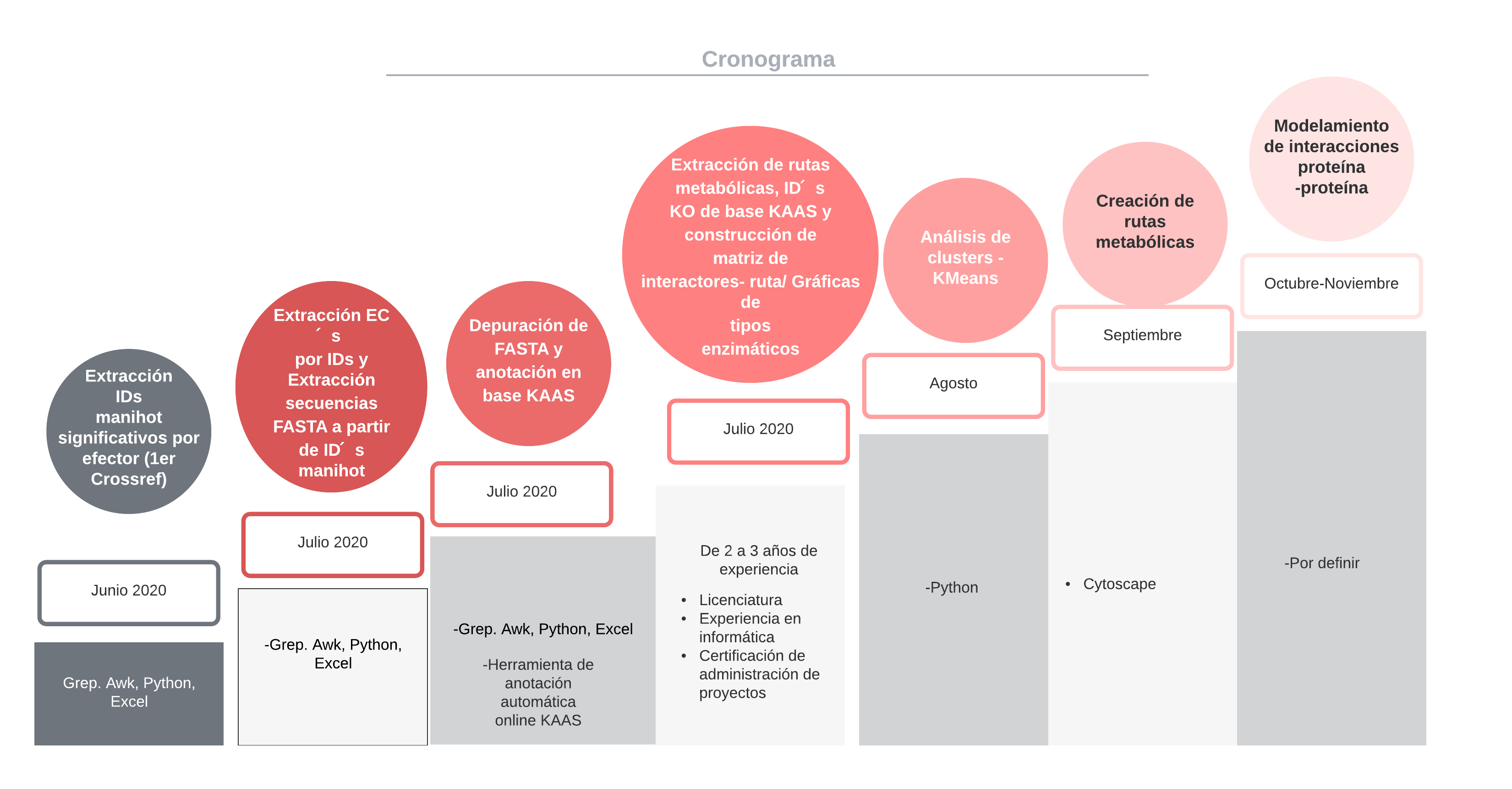
-Dado el valor de clusters óptimo, se crea un nuevo modelo KMeans con este número de centroides como argumento. Se ajusta por medio del método fit(), cuyo input será la matriz de enzimas. Finalmente, se extrae la lista de clusters generados por medio de su atributo labels\_ y se genera un ciclo que, para cada cluster, busque los valores de las rutas enzimáticas que pertenezcan al mismo. Finalmente, estos datos se escriben en un archivo de texto.

**6. Resultados esperados**

Dado que el algoritmo DESeq2 usualmente se utiliza para el análisis de datos provenientes de experimentos de RNA-seq, no existe mucha información con respecto a su utilización en ensayos de interacción proteína-proteína. Aun así, los datos de conteo provenientes de dicho estudio son parecidos a los que se obtendrían en un análisis de expresión de RNA. Usualmente, se observan distribuciones de valores p y LogFoldChange2 cuya forma describe una distribución normal (21). En este caso, se esperan valores de interacción similares, debido a que la mayoría de los interactores no interactúan con los efectores bacterianos. Aún así, posterior al filtrado, se esperan distribuciones torcidas a la derecha puesto que habrán varios interactores cuyos valores serán muy altos.

En cuanto a las enzimas interactoras y sus rutas respectivas, se espera una gran variedad de agentes involucrados. Esto se debe a que el experimento trata con una alta cantidad de efectores bacterianos que, debido a su rango de acción y efectos patogénicos en la planta hospedera, afectan una gran cantidad de procesos metabólicos intrínsecos del organismo. Como se mencionó previamente, la naturaleza de la bacteria patógena específica del estudio indica que sus efectores alterarán factores que permitan la hipertrofia de las células e inhiban la respuesta inmune del hospedero. Claramente, se espera entonces que se encuentren implicadas rutas metabólicas relacionadas con el catabolismo y anabolismo central, de carbohidratos, proteínas y lípidos, además de rutas de señalización celular y activación de las vías de jasmonato y ácido salicílico. Predecir un nivel de detalle mayor a este es, ciertamente, imposible, debido a que los alcances de este estudio pretenden la formación de redes de interacción a través de grafos, que identificarán hasta el nivel de detalle dado por cada proteína individual y su red metabólica. En base a esto, la información de cada par de proteínas interactuantes es fundamental para construir una red fidedigna (22), que permita cumplir el objetivo del estudio.

**7. Cronograma de actividades**

****

**8. Presupuesto**

Para la elaboración de este análisis computacional, se obtuvo financiación por parte de la Universidad de Los Andes, gracias a su programa de prestación de servicios informáticos para proyectos de grado por parte de la computación de alto desempeño de la vicerrectoría de Investigaciones. Así, a nombre del estudiante se solicitaron dos millones de pesos, los cuales se invirtieron en tiempo de cómputo en servidores de alto rendimiento por un año.

**9. Referencias:**

1. Abdurakhmonov, I. (2016). Bioinformatics: Basics, Development, and Future, Bioinformatics - Updated Features and Applications, IntechOpen. 10.5772/63817. <https://www.intechopen.com/books/bioinformatics-updated-features-and-applications/bioinformatics-basics-development-and-future>
2. Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: improvements for better sequence analysis. Nucleic acids research, 34(Web Server issue), W6–W9. https://doi.org/10.1093/nar/gkl164
3. Vailati-Riboni M., Palombo V., Loor J.J. (2017) What Are Omics Sciences? In: Ametaj B. (eds) Periparturient Diseases of Dairy Cows. Springer, Cham. <https://doi.org/10.1007/978-3-319-43033-1_1>
4. Shachuan Feng, Li Zhou, Canhua Huang, Ke Xie & Edouard C Nice (2015) Interactomics: toward protein function and regulation, Expert Review of Proteomics, 12:1, 37-60, DOI: 10.1586/14789450.2015.1000870
5. Jensen R. A. (2001). Orthologs and paralogs - we need to get it right. Genome biology, 2(8), INTERACTIONS1002. <https://doi.org/10.1186/gb-2001-2-8-interactions1002>
6. Paiano, A., Margiotta, A., De Luca, M., & Bucci, C. ( 2019). Yeast two‐hybrid assay to identify interacting proteins. Current Protocols in Protein Science, 95, e70. doi: 10.1002/cpps.70
7. Brückner, A., Polge, C., Lentze, N., Auerbach, D., & Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. International journal of molecular sciences, 10(6), 2763–2788. <https://doi.org/10.3390/ijms10062763>
8. Gupta, R., Lee, S. E., Agrawal, G. K., Rakwal, R., Park, S., Wang, Y., & Kim, S. T. (2015). Understanding the plant-pathogen interactions in the context of proteomics-generated apoplastic proteins inventory. Frontiers in plant science, 6, 352. <https://doi.org/10.3389/fpls.2015.00352>
9. Peyraud, R., Dubiella, U., Barbacci, A., Genin, S., Raffaele, S., & Roby, D. (2017). Advances on plant-pathogen interactions from molecular toward systems biology perspectives. The Plant journal : for cell and molecular biology, 90(4), 720–737. <https://doi.org/10.1111/tpj.13429>
10. Bigeard, J., Colcombet, J., Hirt, H. (2015) Signaling Mechanisms in Pattern-Triggered Immunity (PTI), Molecular Plant,Volume 8, Issue 4, 2015, Pages 521-539. ISSN 1674-2052. <https://doi.org/10.1016/j.molp.2014.12.022>.
11. Koeck, M., Hardham, A. R., & Dodds, P. N. (2011). The role of effectors of biotrophic and hemibiotrophic fungi in infection. Cellular microbiology, 13(12), 1849–1857. <https://doi.org/10.1111/j.1462-5822.2011.01665.x>
12. Kraepiel, Y. and Barny, M.‐A. (2016), Gram‐negative phytopathogenic bacteria, all hemibiotrophs after all?. Molecular Plant Pathology, 17: 313-316. doi:10.1111/mpp.12345
13. Marois, E, (2002). The Xanthomonas Type III Effector Protein AvrBs3 Modulates Plant Gene Expression and Induces Cell Hypertrophy in the Susceptible Host. Molecular Plant-Microbe Interactions® 2002 15:7, 637-646
14. Van Wees, S. C., de Swart, E. A., van Pelt, J. A., van Loon, L. C., & Pieterse, C. M. (2000). Enhancement of induced disease resistance by simultaneous activation of salicylate- and jasmonate-dependent defense pathways in Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America, 97(15), 8711–8716. <https://doi.org/10.1073/pnas.130425197>
15. Choudhary, D. K., Prakash, A., & Johri, B. N. (2007). Induced systemic resistance (ISR) in plants: mechanism of action. Indian journal of microbiology, 47(4), 289–297. <https://doi.org/10.1007/s12088-007-0054-2>
16. Toruño, T. Y., Stergiopoulos, I., & Coaker, G. (2016). Plant-Pathogen Effectors: Cellular Probes Interfering with Plant Defenses in Spatial and Temporal Manners. Annual review of phytopathology, 54, 419–441. <https://doi.org/10.1146/annurev-phyto-080615-100204>
17. Mukhtar, M. S., Carvunis, A. R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., … Payne, T. (2011). Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science*, *333*(6042), 596–601. <https://doi.org/10.1126/science.1203659>
18. Taylor, W. I., & Achanzar, D. (1972). Catalase test as an aid to the identification of Enterobacteriaceae. *Applied Microbiology*, *24*(1), 58–61. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=380547&tool=pmcentrez&rendertype=abstract>
19. Hayes, R. J., Trent, M. A., Truco, M. J., Antonise, R., Michelmore, R. W., & Bull, C. T. (2014). The inheritance of resistance to bacterial leaf spot of lettuce caused by Xanthomonas campestris pv. vitians in three lettuce cultivars. Horticulture research, 1, 14066. <https://doi.org/10.1038/hortres.2014.66>
20. Buiatti, M., Christou, P., & Pastore, G. (2013). The application of GMOs in agriculture and in food production for a better nutrition: two different scientific points of view. *Genes & nutrition*, *8*(3), 255–270. <https://doi.org/10.1007/s12263-012-0316-4>.
21. Love MI, Huber W and Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, pp. 550. <http://doi.org/10.1186/s13059-014-0550-8>.
22. De Las Rivas, J., & Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, *6*(6), e1000807. https://doi.org/10.1371/journal.pcbi.1000807