TANGBOOK: HERRAMIENTA ANALÍTICA Y DE VISUALIZACIÓN PARA DATOS METABOLÓMICOS

Johana Constanza Ojeda Ávila

Trabajo Dirigido

Tutor
Ing. Wilmer Edicson Garzón Alfonso

UNIVERSIDAD DEL ROSARIO
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
PROGRAMA DE INGENIERÍA BIOMÉDICA
BOGOTÁ D.C
2022

AGRADECIMIENTOS

A mis padres y mi hermano principalmente, son ellos quienes creen en mí, me dan la motivación y fortaleza suficiente para superar cualquier reto que se me presente, me ayudan y aconsejan en todo momento. Son el motor que me impulsa a seguir y gracias a ellos pude culminar mi carrera. A mi apoyo emocional, testigos de situaciones clave en mi vida universitaria, que sin palabras siempre están ahí brindándome todo su cariño, mis mascotas.

A mi profesor Wilmer y a la Doctora Claudia Soler, por darme la oportunidad de desarrollar este proyecto. Por su asesoría y tiempo dedicado. Gracias a su profesionalismo y determinación, los consejos y comentarios siempre fueron de gran ayuda para manejar cualquier dificultad que se iba presentando. También por sus ideas y aportes que complementaron y mejoraron las mías e hicieron que se construyera una herramienta grandiosa.

A los profesores que me sirvieron de inspiración sin saberlo. A aquellos que se les nota la pasión al enseñar y hacen del aprendizaje, una experiencia increíble.

Por último, a mis amigos que me acompañaron a lo largo de mi carrera. Muchas veces fueron mi apoyo y la ayuda que necesitaba. Gracias a ellos encontraba seguridad en situaciones en las que sola, se me hubiese dificultado.

RESUMEN

Este proyecto tiene como objetivo desarrollar una herramienta de análisis de datos metabolómicos utilizando bases de datos generalizadas preexistentes. Recientemente, la metabolómica se ha vuelto relevante debido a su prometedora contribución al diagnóstico de enfermedades y la identificación de tratamientos. Sin embargo, los datos metabolómicos plantean un desafío debido a la gran cantidad de datos generados a partir de muestras biológicas de pacientes, lo que dificulta los análisis para médicos e investigadores. Por lo tanto, los investigadores de diferentes campos requieren herramientas bioinformáticas para procesar datos de manera rápida y confiable. Este proyecto utilizará datos metabolómicos sin procesar de pacientes con encefalopatía metabólica y arritmias relacionadas, afectados por *TANGO2*-RMEA. *TANGO2*-RMEA es una enfermedad genética rara descubierta recientemente causada por variantes patogénicas en el gen *TANGO2*. El mecanismo metabólico subyacente que conduce a los síntomas graves en *TANGO2*-RMEA no se comprende bien. El análisis tiene como objetivo ayudar a los médicos e investigadores a comprender mejor el defecto en los pacientes *TANGO2*-RMEA, lo que genera hipótesis que los médicos genetistas y otros especialistas explorarán en detalle, con el objetivo final de encontrar estrategias de tratamiento para esta afección.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

La capacidad de almacenar información ha cambiado conforme transcurre el tiempo. En la actualidad, gracias a los avances tecnológicos se tienen maneras más organizadas y eficientes para almacenar la información de un paciente, como su historia clínica, sus registros, fármacos que utiliza, entre otros datos. Una de las bondades de la capacidad de almacenamiento, es que no solo se limita a guardar información, pues con una correcta manipulación de los registros se pueden encontrar relaciones y patrones entre pacientes, tratamientos, síntomas, enfermedades, etc. Adicionalmente, la mayoría de información se encuentra hoy en día en la nube y de forma *online*, esto facilita su actualización y accesibilidad. Los registros de pacientes brindan posibilidades tanto prospectivas como retrospectivas para el seguimiento. Son la forma más adecuada para estudiar las características de la atención sanitaria, aportando información valiosa para la evaluación, auditoría y planificación de los servicios sanitarios y el seguimiento de su uso [1]. Sin embargo, no siempre es suficiente la información almacenada. En ciertos casos, se debe encontrar la manera de relacionarla, con el fin de resolver nuevos obstáculos que se presenten en el ámbito de la salud. Una enfermedad que se presente en una población muy reducida podría catalogarse como un obstáculo. Se puede enfrentar a 3 desafíos muy importantes como:

- la falta de información que permita entender la enfermedad
- la correcta clasificación de la información existente que ayude a determinar tratamientos potenciales, causas, síntomas, presentación clínica entre otras características
- el reconocimiento de herramientas y tecnologías que ayuden a recopilar, detectar y analizar la información.

Para estos desafíos existe la bioinformática, una disciplina resultante de la intersección de las ciencias biomédicas, la informática, la química y la biología. Con el fin de manejar adecuadamente los datos biológicos, y la información relacionada a las enfermedades, existe una gran variedad de herramientas que, en los últimos años, han sido clave para el análisis y comprensión de éstas. Se han determinado distintos tipos de enfermedades, y muchas de ellas se deben a desordenes en el metabolismo, éste básicamente se refiere a todos los procesos químicos y físicos del cuerpo que utilizan energía y son vitales para la supervivencia de muchas especies [2]. Los defectos metabólicos, pueden ser causados por defectos genéticos o el mal funcionamiento de alguna proteína [3].

1. Caso de estudio

Existen muchas enfermedades presentes en una cantidad muy reducida de pacientes, y se consideran como enfermedades raras. Entre estas se puede encontrar la enfermedad de *TANGO2* [4]. Investigadores del Baylor College of Medicine y del Texas Children's Hospital en Houston, están trabajando con pacientes que presentan alteraciones en el gen *TANGO2* a través del desarrollo de un estudio formal de historia natural con el objetivo de comprender el mecanismo de la enfermedad asociada a *TANGO2*.

2. Identificación del problema

Al ser una enfermedad rara, no hay mucha información disponible sobre *TANGO2*. Tampoco hay bases de datos de las cuales se pueda obtener referencias al caso de estudio. Sin embargo, si hay información acerca de los síntomas encontrados hasta ahora. También se encuentra información de los compuestos asociados al conjunto de datos que poseen los investigadores. Aunque existen herramientas para el análisis metabólico, ¿existe alguna que permita relacionar la información de los pacientes, metabolitos, vías e información complementaria como los genes asociados y análisis de enriquecimiento?

3. Planteamiento de la solución

Para dar solución a esta necesidad, después de hacer una extensa búsqueda de herramientas actuales, y de recolectar las necesidades del personal médico, se propuso crear una herramienta de análisis y visualización, intuitiva y de fácil manipulación, aplicable a casos relacionados con datos metabólicos. La herramienta tendrá distintas funciones que permitirán al usuario interpretar y analizar información metabólica de pacientes. La información de entrada a la herramienta se basa en los niveles de metabolitos en un grupo de pacientes.

Las funcionalidades y fases generales de la herramienta son:

1. Análisis entre pacientes.
    1.1. Determinación y visualización de los pacientes, vías y metabolitos más afectados teniendo como referencia el z – score.
    1.2. Análisis de enriquecimiento de compuestos metabólicos.
    1.3. Aplicación de machine learning para encontrar grupos con similitudes entre los datos.
2. Análisis cronológico.
    2.1. Análisis y visualización de un paciente con medición de datos metabólicos en distintos tiempos.
    2.2. Análisis de la evolución de los datos metabólicos con referencia al tiempo.

En el marco teórico, se muestra la búsqueda anteriormente mencionada, junto con la explicación de la terminología necesaria para la total comprensión del documento, se evidencian las capacidades y limitaciones que tienen las herramientas potenciales para ser utilizadas en enfermedades con datos metabólicos. También se explican detalles de personalización que el personal médico preferiría en la herramienta, y que estas no poseen. Posteriormente en la metodología, se muestra el paso a paso de la implementación de la herramienta. En los resultados, se da a conocer cómo se aplica la herramienta al caso de estudio para luego hacer un análisis en la discusión de los resultados. Finalmente, se comentan algunas funcionalidades a integrar y cambios próximos para la herramienta y se comparan los resultados con las expectativas generadas al plantear este proyecto y los propósitos propuestos.

Este proyecto establece una colaboración académica entre Baylor College of Medicine y dos universidades a las cuales pertenece la autora, la Escuela Colombiana de Ingeniería Julio Garavito (Bogotá) y la Universidad del Rosario (Bogotá).

OBJETIVOS

1. General

Construir una herramienta para el procesamiento de datos de metabolómica, incluyendo las fases de adquisición de datos, preprocesamiento, análisis y visualización de hallazgos para apoyar la interpretación y toma de decisiones del médico experto.

2. Específicos

a) Adquirir datos a partir de información estructurada de los pacientes representados en archivos tipo Excel. Integración de datos a partir de información de las fuentes metabolómicas más relevantes, como la base de datos Human Metabolome y KEGG *Pathway*.
b) Implementar y personalizar técnicas de preprocesamiento para mejorar la calidad de los datos como técnicas basadas en la normalización.
c) Analizar datos de pacientes contrastándolos con datos de fuentes metabolómicas para identificar metabolitos con comportamiento fuera del rango normal. El procesamiento debe conducir a resultados exploratorios, descriptivos o predictivos.
d) Comparar el progreso de cada paciente a partir de las muestras del estado basal junto con la evolución semana a semana para identificar hallazgos relevantes a partir de la información de los metabolitos.
e) Proporcionar gráficos para visualizar los resultados y hallazgos relevantes del análisis aplicado, incluido el análisis funcional y de ruta.
f) Debe estar desarrollada en un lenguaje de programación portable, que garantice la interacción a través de una interfaz gráfica y alta usabilidad por parte del médico experto.

MARCO TEÓRICO

1 Enfermedades Raras, TANGO2-RMEA y metabolómica

Una enfermedad rara es aquella que se puede definir como “una condición de baja prevalencia que afecta a menos de una de cada 2.000 personas” [5] ya que hay muy poca información y conocimiento acerca de ellas, es una tarea de gran dificultad establecer un tratamiento o llegar a sus causas [6]. Por esta razón, a menudo se les conoce como “enfermedades huérfanas” y la mayoría se clasifican como enfermedades genéticas [7]. Así mismo, no está muy lejos de la realidad pensar que las enfermedades raras son extremadamente difíciles de identificar entre un gran número de otros diagnósticos posibles. Con base en una encuesta de 2013, se necesitan, en promedio, más de 5 años, ocho médicos y dos o tres diagnósticos erróneos hasta que un paciente con una enfermedad rara recibe el diagnóstico correcto [8]. En dimensiones numéricas, antes de 1983, solo había 34 tratamientos para enfermedades raras y hasta el momento se han reconocido entre 6000 y 7000 distintas, que afectan aproximadamente 4-6 % de la población europea y a 300 millones de personas en todo el mundo [9]. En un escenario optimista, al diagnosticar correctamente alguna enfermedad rara, los desafíos para empezar un análisis que conlleve a un camino de entendimiento y tratamiento continúan, pues al ser una población tan pequeña, los incentivos comerciales para desarrollar medicamentos suelen ser bajos (aunque las políticas y legislaciones apuntan a aumentar los incentivos financieros para desarrollar tratamientos para enfermedades raras) [10]. Sumando lo complejo que es comprender el estudio de los procesos patológicos físicos y químicos subyacentes, muchas enfermedades raras carecen de opciones de tratamiento adecuadas dando como consecuencia un importante problema de salud pública al querer mejorar el diagnóstico y el tratamiento de las enfermedades de este tipo. Para esta problemática se han propuesto iniciativas que mejoren la atención médica de pacientes que las padezcan, con el fin de obtener y agrupar datos e información sobre la enfermedad en sí, para alimentar el conocimiento global de ésta. También se debe tener en consideración que algunas enfermedades son ultra raras afectando a menos de 100 personas en todo el mundo y, en algunos casos, a una sola persona [11,12], y que enfermedades como por ejemplo la malaria, o la enfermedad del sueño, son poco comunes en los países desarrollados, considerándose endémicas en otros países o utilizando el término “desatendidas” según la OMS ya que las grandes compañías farmacéuticas a menudo las pasan por alto [13]. Para el año 2021, gracias a los incentivos gubernamentales se aprobaron más de 600 terapias que fueron desarrolladas en los últimos 40 años, sin embargo, cabe resaltar que es un número muy reducido de pacientes los que se pueden tratar con un medicamento aprobado [14,15].

Como se ha mencionado, este proyecto tiene como caso de aplicación datos de una enfermedad rara llamada *TANGO2* (*TANGO2*-RMEA OMIM: 616878). Para una breve explicación, *TANGO2* es un gen codificador de proteínas de la familia de transporte y organización de Golgi, ubicado en la posición 11.21 del cromosoma 22 (22q11.21) [16], se prevé que sus miembros desempeñen funciones en la carga de proteínas secretoras en el retículo endoplásmico. Recientemente se ha identificado como un gen relacionado con enfermedades humanas, que conlleva a una enfermedad genética heredada de forma autosómica recesiva. Los efectos de la mutación de las variantes patogénicas bialélicas en el gen *TANGO2* conllevan a que los pacientes afectados experimenten una variedad de

síntomas como crisis metabólicas, encefalopatía, acidosis láctica, episodios agudos de rabdomiólisis marcados por arritmia cardíaca, retraso en el desarrollo, convulsiones, endocrinopatías y taquiarritmias potencialmente mortales [17,18]. Actualmente se sabe muy poco sobre la función de la proteína *TANGO2* y su papel en las descompensaciones metabólicas. En la literatura médica se han ilustrado 49 casos de mutación bialélica de *TANGO2*, identificados a través de un panel de genes o secuenciación del exoma [19,20]. A pesar de presentar síntomas sugestivos de un problema metabólico primario, las pruebas metabólicas convencionales no han revelado un patrón específico de anomalías en pacientes con *TANGO2*-RMEA. Por lo tanto, hasta la fecha no se ha identificado un tratamiento específico para esta afección. Estudios recientes indican que *TANGO2* también se localiza en las mitocondrias, pero la información disponible hasta la fecha no ha establecido una conexión con otros problemas mitocondriales similares a *TANGO2*. Asimismo, los intentos de aplicar los tratamientos disponibles para otras enfermedades diferentes similares a *TANGO2*-RMEA no han demostrado ser efectivos [21,22].

Para poder entender la información acerca de las enfermedades raras y especialmente de *TANGO2*-RMEA, es necesario comprender conceptos básicos de metabolismo y la importancia de este.

La metabolómica es el estudio de metabolitos, los cuales son sustancias presentes en células y tejidos. El metabolismo es una serie de procesos químicos que ocurren dentro de un organismo para mantener la vida. Por consecuencia, el estudio de la metabolómica se refiere a la identificación y cuantificación sistemática de todos los metabolitos en un determinado organismo o muestra biológica [23], es decir, reconoce actividades bioquímicas en un momento específico mediante el análisis de miles de moléculas pequeñas en células, tejidos, órganos o fluidos biológicos; después define biomarcadores de metabolitos haciendo uso de técnicas informáticas [24]. Esta colección de moléculas, conocida como metaboloma, identifica nuevos marcadores de diagnóstico o pronóstico, lo que puede conllevar a responder preguntas acerca de una enfermedad, y también mejorar la comprensión de los fenotipos de respuesta a medicamentos relacionados a esta. Existen dos enfoques diferentes para el estudio del metaboloma, que son la metabolómica dirigida y no dirigida (en inglés targeted y untargeted respectivamente). La metabolómica dirigida se enfoca en la cuantificación de compuestos conocidos, y la no dirigida tiene como objetivo detectar patrones de todo el conjunto de metabolitos desconocidos [25].

2. Técnicas Informáticas del estudio de los datos en la metabolómica

En la *Figura 1*, se encuentran los pasos a seguir para las técnicas informáticas del estudio de los datos en la metabolómica, empezando con el paso 1 que sería el planteamiento de un problema, o la identificación de una necesidad. Una vez planteado el problema biológico, seguimos con el paso 2 en donde se realiza la adquisición de los datos metabolómicos, para ello, las metodologías más utilizadas y frecuentes son el análisis no dirigido por espectroscopia de resonancia magnética nuclear (RMN) y la espectrometría de masas (MS). La recopilación y generación de conjuntos de datos espectrales a partir de las muestras experimentales diseñadas se hace por medio de principios cromatográficos que apuntan a la separación de los componentes por instrumentos cromatográficos, es decir, cromatografía líquida (LC), cromatografía (GC) o electroforesis capilar (CE) [26-28]. Generalmente se utiliza RMN para caracterizar estructuralmente compuestos desconocidos y analizar metabolitos en fluidos biológicos y extractos celulares. No obstante, tiene ciertas

desventajas como el costo del equipo que se utiliza ya que es mucho más alto en comparación con las técnicas basadas en MS, otras desventajas son el tiempo de ejecución por muestra que toma varias horas y la baja sensibilidad.

Figura 1 Técnicas informáticas en la metabolómica

Posterior a la adquisición de los datos, se realiza el paso 3 en donde se deben preprocesar, con el objetivo de convertir diferentes datos metabolómicos en una matriz adecuada y comparable para posteriores análisis estadísticos. Unos de los métodos más utilizados son la normalización y la identificación de metabolitos para la evaluación de la calidad de los datos de metabolómica. Con la normalización se desea hacer una comparación directa para el perfilado del metaboloma seleccionando un conjunto representativo de picos de cada conjunto de datos. Con la identificación de metabolitos, como resultado se desea obtener un perfil del metaboloma completo a partir de muestras biológicas. La identificación de metabolitos puede ser el paso más relevante en los estudios de metabolómica, porque de la información que se extraiga depende la identificación exacta de los metabolitos. Este proceso se lleva a cabo con espectrometría de masas en tándem (MS/MS). Por la complejidad de la matriz, se aplican técnicas avanzadas de RMN y estrategias analíticas para una mayor precisión en la identificación de metabolitos [29]. La aplicación de varias técnicas de identificación de metabolitos como experimentos de adición, espectros estándar, bancos de datos y bibliotecas de RMN ayudan a la solución de problemas para identificarlos como el apiñamiento espectral, la presencia de macromoléculas, la interacción molecular, entre otros [30-32].

Una vez se tiene hecho el procesamiento de los datos, se procede con el paso 4, en donde para el análisis, primero se establece cual es el resultado que se quiere obtener. Se determina si los resultados son de clasificación o predicción para implementar métodos de aprendizaje supervisado y no supervisado. Los métodos de aprendizaje supervisado poseen una clase, la cual permite diferenciar grupos en los datos. Mientras que los métodos de aprendizaje no supervisado no cuentan con esta, por lo que se basan en encontrar patrones para determinar la clasificación.
Después del preprocesamiento, si se desea distinguir por ejemplo en clases como “sanos” y “enfermos”, es recomendable descubrir marcadores biológicos o metabolitos relevantes que ayuden a identificar estas dos clases entre los datos. Estos biomarcadores pueden ayudar a una posterior predicción de éxito terapéutico o de enfermedades. Las

características seleccionadas se pueden considerar como biomarcadores en el ámbito biomédico con el fin de predecir el éxito terapéutico o diagnosticar y pronosticar enfermedades. La interpretación de estas características se puede realizar con análisis de enriquecimiento, mapeo y/o visualización de vías [33]. En la *Tabla 1*, se describen conceptos de los métodos más comunes de aprendizaje automático aplicados a estudios metabolómicos. Y en la *Figura* 2, se muestra la frecuencia de uso de métodos tanto supervisados como no supervisados destacados de la literatura relacionada [34-49].

Aprendizaje supervisado		Aprendizaje no supervisado	
Partial Least Squares (PLS)	Facilita la reducción dimensional y los datos también se pueden proyectar para su visualización en un espacio de baja dimensión. Se basa en la regresión lineal al proyectar las variables dependientes e independientes en un nuevo espacio.	Principal Component Analysis (PCA)	Reduce las dimensiones y proporciona una visualización de los datos proyectándolos en una dimensión más baja. Su objetivo es encontrar un subespacio unidimensional que capture la mayor variación de los datos, denominado componente principal (PC). Al usar el PCA, las muestras se pueden visualizar para evaluar similitudes y/o diferencias entre ellas.
Support Vector Machine (SVM)	Construye un hiperplano lineal óptimo que separa dos clases en un espacio de características. Busca hiperplano de margen máximo, con mayor separación entre las clases. El margen es la distancia máxima entre el hiperplano y los puntos de datos de cada lado.	Clustering	Su objetivo es dividir los datos en grupos de acuerdo con sus propiedades. Los agrupados en el mismo grupo son más similares entre sí que los objetos de otros grupos. Los algoritmos más conocidos son k-means donde el usuario define la cantidad de grupos para los n objetos, y agrupamiento jerárquico que como su nombre lo indica, jerarquiza los grupos.
Random Forest	Se basa en los resultados agregados de varios árboles de decisión individuales.	Self-Organizing Map (SOM)	Brinda una visualización de datos de alta dimensión en un espacio de baja dimensión poniendo las muestras que son similares entre sí en una región similar. Se le puede interpretar como una caja negra debido a que los datos clave para la separación están ocultos y sería una desventaja a la hora de dar una interpretación.

Tabla 1 Métodos más comunes en metabolómica para aprendizaje automático.

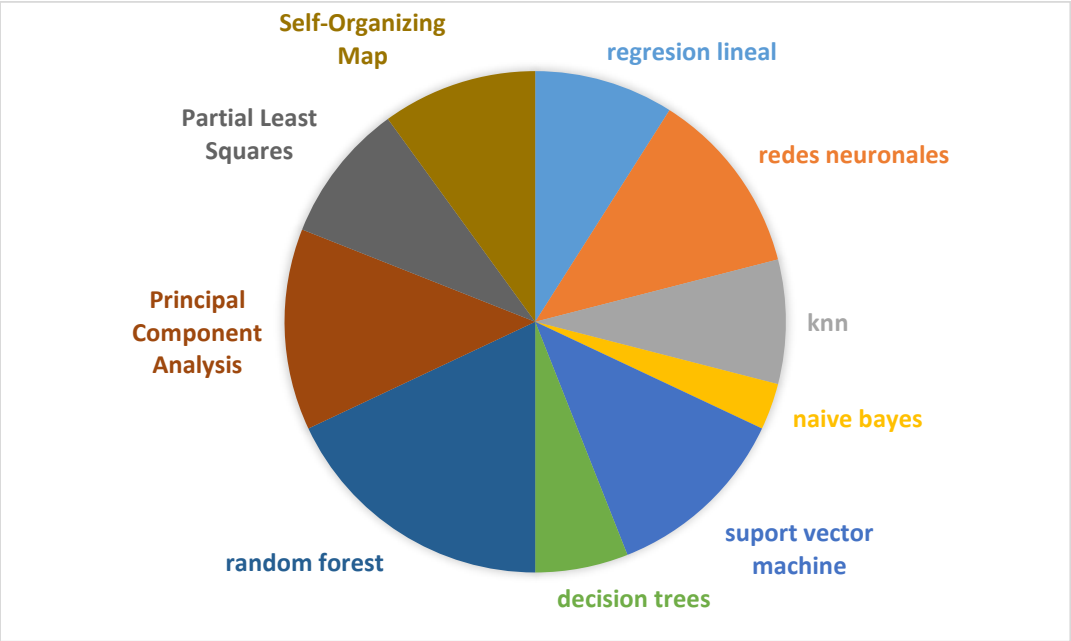


Figura 2 Métodos más frecuentes aprendizaje automático

Ahora bien, para el descubrimiento de biomarcadores se quiere seleccionar un marcador candidato que proporcione la mejor separación entre dos grupos. Esto se hace por medio de selección de características en donde se hace uso de pruebas de hipótesis estadísticas, como la prueba t que evalúa si las medias de dos grupos independientes son estadísticamente diferentes entre sí, la prueba Welch, la prueba de la suma de rangos de Wilcoxon en donde según la clasificación obtenida, los mejores metabolitos se denominan biomarcadores candidatos basándose en la puntuación significativa o *valor p* [50-52]. Existen biomarcadores de diagnóstico (distingue entre un paciente sano y uno enfermo), pronóstico (brinda información de cuál es el curso probable de la enfermedad) y predictivos (identifican subpoblaciones de pacientes que tienen más probabilidades de responder a una terapia determinada).

Por último, el paso 5, se requiere interpretar el análisis de los datos. Aunque se tenga una lista con los metabolitos ya identificados, no brinda un contexto de algún proceso biológico. Para esto se requiere información complementaria compuesta por vías metabólicas y anotaciones. En este punto es útil el mapeo y la visualización de las vías metabólicas que están relacionadas a los compuestos o metabolitos ya identificados, pues aporta una visión de como interactúan dichos metabolitos y la función o rol en determinado proceso biológico. Otro enfoque que brinda información complementaria de los posibles roles de los metabolitos es el análisis de enriquecimiento que permite determinar los procesos biológicos afectados por un experimento de expresión diferencial. Este enfoque requiere conjuntos de genes biológicamente relevantes, que actualmente se curan manualmente, lo que limita su disponibilidad y precisión en muchos organismos sin recursos cuidadosamente curados. Existen variantes como el análisis de sobrerrepresentación (ORA) el cual aplica una prueba estadística a metabolitos de entrada seleccionados por el usuario para determinar si un conjunto de metabolitos de entrada se enriquece en una anotación particular en comparación con un conjunto de fondo; y el análisis de enriquecimiento de

conjuntos (SEA) que utiliza una lista clasificada completa de genes expresados diferencialmente (GSEA) o metabolitos y los evalúa en un conjunto específico haciendo uso de un método estadístico que puntúa el enriquecimiento. Su comportamiento va a tender a organizarse estando en la parte superior o inferior de una lista clasificada (más regulados hacia arriba o hacia abajo) [53,54].

3. Estado del Arte

3.1 Herramientas

Con el objetivo de saber qué herramientas existen en la actualidad para la interpretación de datos, se hizo una búsqueda para determinar qué ventajas y desventajas se identificaban en cada una, y si daban solución a todos los requerimientos del personal médico del Baylor College (los cuales se detallarán en la metodología). La revisión de las herramientas actuales para analizar datos de metabolómica muestra una oportunidad para proporcionar una herramienta de libre acceso y fácil de usar para la comunidad médica. Se hizo una revisión de varias herramientas actuales comparando aproximadamente 86 herramientas de análisis metabólico. Del total, el 82% están enfocadas en el análisis de datos con cromatografía líquida-espectrómetro de masas (LC-MS).
Como el dominio de programación es en lenguaje Python por su flexibilidad, portabilidad y despliegue en aplicaciones web, el lenguaje de programación fue una característica a considerar. Con base a esto, fueron seleccionadas 11 herramientas que están basadas en Python, comparándolas para el propósito de análisis de datos. Como resultado de la comparación, las herramientas más completas son TidyMS, NeatMS, Biodendro, MRMkit y SmartPeak. Estas herramientas fueron elegidas por el tipo de análisis de datos ya que permiten un análisis estadístico, gráfico de fragmentación, metabolitos, picos y puntaje del análisis de componentes principales (PCA). Para una breve descripción de estas herramientas, presentamos algunas características relevantes para cada una.

TidyMS genera cromatogramas y espectros de iones totales, permite la selección de picos, la detección de características ofrece funcionalidades para la curación, normalización, imputación, escalado, métricas de calidad, correcciones por lotes basadas en control de calidad y visualización interactiva de resultados [55]. NeatMS permite el filtrado automático de falsos positivos en intensidad máxima. Esta herramienta se basa en el procesamiento de datos LC-MS de rutina [56]. Biodendro permite a los usuarios agrupar e interrogar de manera flexible miles de espectros e identificar rápidamente patrones de fragmentos centrales que causan grupos que conducen a la identificación de estructuras químicas centrales de una clase más grande, incluso cuando falta el metabolito individual de interés [57]. MRMkit realiza detección automática de picos, integración de picos, normalización, corrección de efectos por lotes, cálculos de métricas de calidad, visualizaciones de cromatogramas y eliminación de picos redundantes de clases multimodales [58]. Por último, Smartpeak ofrece algoritmos novedosos para la alineación, el ajuste de la curva de calibración y la interrogación de picos para facilitar la reproducibilidad al reducir el sesgo del operador. Smartpeak garantiza una alta calidad para el procesamiento automatizado de datos de EC, GC y LC-MS para experimentos dirigidos y semidirigidos de metabolómica, lipidómica y fluxómica [59]. Debido a la gran cantidad y variedad de datos generados por los análisis metabolómicos, las herramientas de software incluyen el procesamiento de

datos espectrales sin procesar, el análisis estadístico para encontrar metabolitos expresados significativamente y la conexión a bases de datos de metabolitos para su identificación. de metabolitos, integración y análisis de múltiples datos ómicos heterogéneos.
Finalmente, análisis bioinformático y visualización de redes de interacción molecular. El más completo de varias herramientas de software es MetaboAnalyst. Hay otras herramientas similares que son MetaCoreTM, Caleydo, IPATM, PathVisio, 3Omics e InCroMAP [60-67].

Se encontró que MetaboAnalyst [68] y Reactome [69] pueden hacer un análisis metabólico, análisis de enriquecimiento y también tienen integrado herramientas de visualización de vías metabólicas y mapeo. Aunque cuentan con un gran número de utilidades que se ajustan a ciertos requerimientos, no son del todo óptimas, pues la entrada de datos no es automática, es decir, hay que escribir uno a uno los compuestos a analizar. Esto no es lo ideal, pues se requiere que al subir un archivo Excel a la herramienta, ésta identifique los compuestos, *pathway*s y *super pathway*s sin necesidad de especificar códigos de identificación para cada base de datos. Algunas otras herramientas son KEGG, HumanCyc, HMDB y SMPDB que solo permiten la visualización, pero no tiene integrado un análisis de enriquecimiento, y la entrada de los datos de búsqueda tampoco es cómoda [70-73].

Para el análisis de enriquecimiento, se hizo una búsqueda de los métodos empleados con más frecuencia. Se encontraron alrededor de 43 métodos utilizados en diferentes estudios, tomando como base uno en específico titulado “Sobre la influencia de varios factores en el análisis del enriquecimiento de rutas” el cual ya había reunido información relevante [74].
Por ahora los métodos de interés son GSEA y ORA Fisher’s test. La elección de estos dos métodos se basó en la popularidad según la literatura, la capacidad de obtener la información necesaria para aplicarlos, las bases de datos con las que trabajan para complementar la información y el enfoque para la herramienta.
Para tener más claridad, ambos enfoques tienen como objetivo tomar niveles de expresión génica y utilizar el conocimiento del organismo dado para hacer una identificación de los procesos y mecanismos biológicos subyacentes en procesos como las vías metabólicas [75]. Una vía metabólica describe ciertos fenómenos e interacciones en genes, proteínas o metabolitos dentro de células, tejidos u organismos [76].
Se buscó información de cuales de estos fueron los métodos más repetidos, en donde encontramos que son:

- GSEA-S: determina si un conjunto de genes definido a priori muestra diferencias estadísticamente significativas y concordantes entre dos estados biológicos [77].
- ORA / Fisher's test: determina si las funciones o procesos biológicos conocidos están sobrerrepresentados [78].
- GSVA: análisis de variación del conjunto de genes realiza un análisis de datos moleculares centrados en la ruta al realizar un cambio en la unidad funcional de análisis, de genes a conjuntos de genes [79].
- PADOG: Análisis de rutas con reducción de peso de genes superpuestos, minimiza la importancia de los genes que aparecen a menudo en los conjuntos de genes que se analizarán [80].

3.2 Bases de Datos

KEGG es una colección de bases de datos en línea de genomas, rutas enzimáticas, y químicos biológicos. Puede ser utilizada para la modelización y la simulación, la navegación y extracción de datos. Reactome es una base de datos de vías gratuita, de código abierto, seleccionada y revisada por pares. Proporciona herramientas bioinformáticas intuitivas para la visualización, interpretación y análisis del conocimiento de la ruta para respaldar la investigación básica, el análisis del genoma, el modelado, la biología de sistemas y la educación [52]. Adicionalmente, en el data set en formato Excel generado por Metabolon Inc., se nombran otras bases de datos como PubChem la cual es la colección más grande del mundo de información química de libre acceso. Se encuentran productos químicos por nombre, fórmula molecular, estructura y otros identificadores. Y también información sobre propiedades químicas y físicas, actividades biológicas, información sobre seguridad y toxicidad, patentes y citas bibliográficas [81]; y HMDB la cual es actualmente la colección curada más completa y completa de metabolitos humanos y datos de metabolismo humano en el mundo. Contiene registros de más de 2180 metabolitos endógenos con información recopilada de miles de libros, artículos de revistas y bases de datos electrónicas [82].

METODOLOGÍA

Personal médico de Baylor College de Texas está en la búsqueda de información complementaria sobre la enfermedad rara llamada *TANGO2*-RMEA [4], pues como se sabe, no existe una cantidad suficiente de información para saber qué la produce, cual es el procedimiento que se debe seguir con los pacientes, y mucho menos cómo tratarla o curarla. Para ello, tomaron muestras de 10 pacientes con confirmación genética de la enfermedad, para su preprocesamiento y análisis las procesaron en una compañía llamada Metabolon Inc, la cual "descifra miles de señales químicas discretas de factores genéticos y no genéticos para descubrir biomarcadores y revelar vías biológicas. Hacen conexiones donde otras ómicas no pueden y proporcionan la representación definitiva del fenotipo" [83].

Como se había comentado en secciones anteriores, se encontró una oportunidad de creación de una herramienta principalmente para el análisis y la visualización de datos. No sólo de los datos metabólicos como los compuestos, vías y super vías; también los pacientes que se están analizando, junto con la información complementaria que proveen las bases de datos como KEGG o PubChem. Esta herramienta surge por la necesidad de comparar datos metabólicos a partir de un informe en Excel generado por Metabolon Inc, esta comparación y los procedimientos que la comprenden son en pocas palabras personalizados ya que deben ser lo más cómodos y sencillos de utilizar por parte del personal médico. Con ella se quiere dar la facilidad de visualizar y analizar los datos. Estos análisis se basan en técnicas de machine learning, o análisis de enriquecimiento de vías.

Los requerimientos básicos de la herramienta son los siguientes:

- Encontrar similitudes y diferencias entre metabolitos y sujetos de forma interactiva y rápida y amigable para el médico.
- Comparación de rutas metabólicas.
- Detectar de forma visual y más sencilla cuales metabolitos están afectados (elevados o disminuidos) y en cuales pacientes.
- Conexión entre metabolitos y presentación clínica del paciente.
- Tratamientos potenciales.

Para empezar, se hizo una búsqueda de información de los procesos que utiliza Metabolon, actualmente generan informes tipo Excel con información compuesta por las bases de datos que utilizan, plataformas, metabolitos y *pathway*s encontrados, también generan una interpretación junto con los gráficos correspondientes, sin embargo, en este informe de *TANGO2*, no se obtuvo una interpretación de los datos metabólicos y no se fue especificada la razón. Para la obtención de los datos, procesaron las muestras y aplicaron una técnica analítica que implica la separación física de compuestos objetivo seguida de su detección basada en masas. LC hace referencia a cromatografía líquida y MS a espectrometría de masas. Por la sensibilidad, selectividad y precisión detecta cantidades de microgramos de una variedad de metabolitos [84]. Posterior a ello, el preprocesamiento consistió en la eliminación del ruido, reducción de datos, comparación y búsqueda de información complementaria, normalización, entre otros [85]. Para el estudio, no se proporcionó información más detallada del preprocesamiento realizado por Metabolon Inc. Para finalizar,

hicieron un análisis de los datos para determinar un rango o z-score. Un z-score es el número de desviaciones estándar de la media de un punto de información [86]. Se determinó que si el valor del metabolito en un paciente es mayor a 2 y menor a -2, quiere decir que es un valor fuera del rango de normalidad. En la *Figura 3*, se muestra el encabezado y la información que aparece en las primeras celdas del reporte hecho por Metabolon Inc.

									TANG_0051	TANG_0052
Pathway Sort Order	Super Pathway	Sub Pathway	Biochemical Name	Platform	Comp ID	KEGG	HMDB	PUBCHEM		
1			glycine	LC/MS pos early	58	C00037	HMDB0	750	0,24	0,22
2			N-acetylglycine	LC/MS polar	27710		HMDB0	10972	-0,26	0,99
5			sarcosine	LC/MS pos early	1516	C00213	HMDB0	1088	-0,11	-0,17
6			dimethylglycine	LC/MS pos early	5086	C03626	HMDB0	673	0,07	-0,6
7			betaine	LC/MS pos early	3141	C00719	HMDB0	247	1,58	0,38
10		Glycine, Serine an	serine	LC/MS pos early	1648	C00065	HMDB0	5951	0,22	0,31
11			N-acetylserine	LC/MS polar	37076		HMDB0	65249	-0,14	0,06
17			threonine	LC/MS pos early	1284	C00188	HMDB0	6288	0,68	0,77
18			N-acetylthreonine	LC/MS neg	33939		HMDB0	152204	0,34	0,45
23			O-acetylhomoserin	LC/MS polar	31539	C01077	HMDB0	439389		
29			alanine	LC/MS pos early	1126	C00041	HMDB0	5950	0,86	0,89

Figura 3 Reporte por Metabolon Inc

Después de conocer el procedimiento llevado a cabo para tener un conjunto de datos en una tabla o también llamado *dataset,* se plantea un plan de acción para abordar el problema inicial. Se tienen los siguientes aspectos para tener en cuenta:

- Búsqueda y obtención de información en Python.
- Adecuar el *dataset*
- Integrar la tabla de Excel y las bases de datos
- Seleccionar la información relevante
- Automatizar la información
- Implementar gráficos interactivos
- Encontrar similitudes y diferencias entre los sujetos y metabolitos
- Comparar vías, super vías y compuestos
- Conexión de la presentación clínica y los metabolitos
- Análisis de enriquecimiento
- Aplicación de técnicas de machine learning
- Integración de la página web

De los anteriores aspectos, surgen 4 grandes pasos para la creación de la herramienta.

1. Comprensión del dataset

En esta sección, se describirá la información contenida en las columnas del *dataset*, para determinar la importancia de cada una dependiendo la información que aporte. De la *Figura 3*, los datos más relevantes con los que se trabaja son:

- *Super Pathway* que indica la super vía metabólica o macromolécula, puede estar compuesta por aminoácidos, péptidos, carbohidratos, cofactores y vitaminas, energía, xenobióticos, nucleótidos, lípidos y por último moléculas parcialmente caracterizadas.
- *Sub Pathway* que son las vías metabólicas que componen las macromoléculas.
- *Biochemical Name*, que hace referencia a los compuestos o metabolitos.
- Las bases de datos como KEGG, PubChem o HMDB que permiten la identificación de los compuestos en las diferentes bases de datos.
- TANG_00XX, hace referencia a los pacientes, en este caso son 10 pacientes que irían de TANG_0051 a TANG_0060.

Para utilizar al máximo la información entregada en la tabla de pacientes, se procedió a buscar la información con la que la alimentaron, es decir, buscar aquella que específica los compuestos, *pathways* y *super pathway*s en las bases de datos KEGG, PubChem y HMDB, pero en programación. La manipulación se hizo uso de cuadernillos en Google Colab, el cual tiene la misma funcionalidad y sintaxis que un cuadernillo en Jupyter Notebook.

- Para KEGG se encontraron diversas funcionalidades programáticas que se desarrollan a través de las bibliotecas de programación de Biopython. Es sencilla la configuración de búsqueda para una o más colecciones de secuencias [87]. Algunas funciones son:

    - Convertir los identificadores de KEGG a los de otras bases de datos.
    - Encontrar entradas KEGG con datos de consulta coincidentes.
    - Recuperar datos para una entrada específica de KEGG.
    - Obtener información sobre una base de datos KEGG.
    - Encontrar entradas en KEGG usando una referencia cruzada de base de datos.
    - Obtener lista de entradas en una base de datos.

- Para PubChem también se encontraron funcionalidades para Python haciendo uso de la librería PubChemPy [88]. Algunas de estas son:

    - Obtener la información de un compuesto por su nombre o Id.
    - Encontrar características asociadas al compuesto como la fórmula, peso molecular, isómeros, nombre iupac, sinónimos, entre otros.
    - Obtener información de sustancias y compuestos relacionados.

- Para HMDB se encontraron varias librerías, una llamada Bio2bel [89] y otra denominada Bioconductor [90], ambas dan información química y general de compuestos.

2. Preprocesamiento

En esta sección se utilizan técnicas como análisis de datos exploratorios (EDA por su nombre en inglés *Exploratory Data Analysis*), lo cual nos brinda un enfoque para el análisis

de conjuntos de datos con el fin de resumir sus características principales haciendo uso de métodos visuales complementarios [91], en donde los pasos más comunes son:

- Recopilación de datos: este proceso consiste recopilar información de una manera sistemática establecida que permite probar hipótesis y evaluar resultados fácilmente. Posterior a esto, se identifican el tipo de características de los datos, es decir, si se trabaja con variables de tipo numéricas, categóricas, de tiempo, ordinales, entre otras.
- Limpieza de datos: Este paso consiste en la eliminación de información que no es necesaria, para garantizar que los datos sean correctos y utilizables al identificar cualquier error en los datos o datos faltantes.
- Preprocesamiento de datos: es una técnica del campo de minería de datos la cual consiste en transformar datos sin procesar en un formato comprensible. Algunas técnicas usadas más comunes son normalización y estandarización, transformación, extracción y selección de características, entre otras.
- Visualización de datos: representa gráficamente la información y los datos anteriormente manipulados. Utiliza gráficos estadísticos, diagramas, gráficos de información y otras herramientas para comunicar información de manera clara y eficiente.

Para empezar, aplicando los pasos y metodologías de EDA, se ordenaron las filas y columnas del *dataset* haciendo uso de Pandas, una librería para manipular tablas de datos en Python. De esta forma, los índices o nombres de las columnas quedan en la primera fila, así es más fácil de identificar un subgrupo del *dataset*. Después de esto, se procedió a eliminar la información irrelevante. Fueron eliminadas las columnas *pathway sort order*, *platform, comp ID, TANG_0061* y *TANG_0062*. Estas dos ultimas no correspondían a pacientes. Las variables de los datos discretos son categóricas, por ejemplo, en *super pathway* son 9 los posibles a los que puede pertenecer un compuesto. Y en los datos continuos son variables de tipo numéricas.

Ya identificados los tipos de datos con los que se está trabajando, se procede a limpiar el *dataset*. Para empezar, se desea conocer la cantidad aproximada de datos faltantes para saber cómo manejar la falta de información. Pues, teniendo datos faltantes, se puede llegar a realizar análisis muy alejados de la realidad y tener una perspectiva errónea del *dataset*. En la figura 4, se muestra en color claro los datos faltantes en las columnas.
Existen varias formas para lidiar con los valores faltantes, se pueden reemplazar con cero, la mediana, moda, valor mínimo, valor máximo, entre otros; de la fila o columna en la que se encuentra el valor faltante [92]. Sin embargo, para aplicar algún método se debe saber con certeza cual afectaría menos el contexto y análisis de los datos. Por ejemplo, en este caso de estudio es relevante saber qué metabolitos son anormales, y con el z-score se sabe que se determina anormal si el valor del metabolito en el paciente es mayor a dos y menor a menos dos. Por lo tanto, no es útil reemplazar los valores faltantes con el valor mínimo o máximo, ya que también se tendría en cuenta el metabolito con el valor reemplazado en el conjunto de afectados y puede conducir a una conclusión errónea. Con la anterior información, se decide que para las columnas *super pathway* y *sub pathway* se reemplazarán los valores faltantes con el dato anterior. La elección se debe a que se sabe que estas columnas representadas en Excel tienen algunas filas combinadas, pero la visualización del *dataset* con Pandas, las filas combinadas en una columna muestran el

contenido en la primera celda, y en las siguientes celdas muestra el valor faltante. Para los datos faltantes en las columnas de los pacientes, se decide que se reemplazarán con cero, de esta forma no afecta al análisis en términos de que es un valor “normal”.

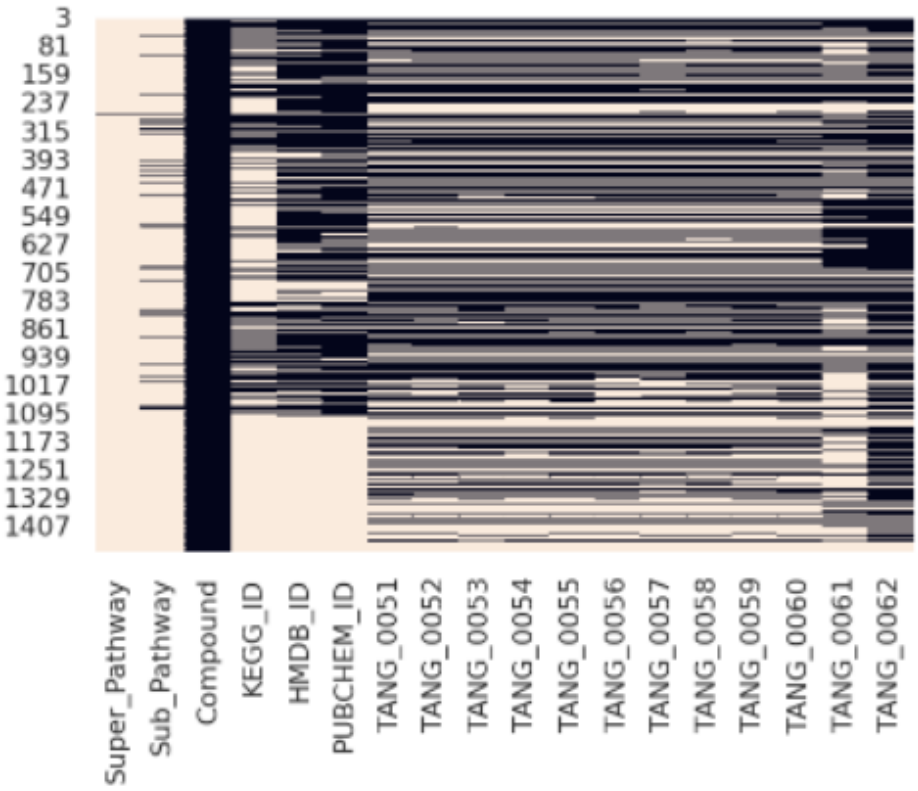


Figura 4 Heatmap de valores nulos

Una vez se tiene el *dataset* ordenado y sin datos faltantes, se extraen las filas pertenecientes a los metabolitos que no se presentan en ningún paciente, y se ordenaron en una lista, pues al tratarse de datos de una enfermedad rara, por el hecho de que no haya información en los pacientes, no significa que estos compuestos no den indicios o patrones para encontrar pistas que conlleven a información sobre los pacientes con *TANGO2*-RMEA. Ya que surgen preguntas que quizá sirvan al personal médico. Algunas de ellas pueden ser: ¿estos compuestos siempre están ausentes en los pacientes que la padecen? ¿existe relación con estos compuestos, vías o super vías con los pacientes de *TANGO2*?

Para dar una breve explicación de cómo cambia la tabla de información al haber hecho los cambios en los valores faltantes, el tamaño inicial era de 1448 filas por 21 columnas, y la tabla de datos a utilizar, quedó de 841 filas por 13 columnas

Para integrar la tabla o *dataset* con la información en Python de la base de datos KEGG se hizo uso de la casilla KEGG_ID. Ya que, en la integración se conectan las bases de datos disponibles de forma *online* con los códigos de cada compuesto que en este caso están en la columna de KEGG_ID.

No se va a trabajar con la base de datos de HMDB ni PubChem por ahora, ya que se consideró que la primera versión de la herramienta será con KEGG. Cabe resaltar, que en una actualización futura si se puede hacer uso de estas bases de datos.

3. Diseño de la herramienta

Como la herramienta está pensada para personal médico, y los primeros usuarios están en Houston, se desarrolló en inglés. A continuación, se menciona la forma en la que se diseñó la herramienta y el orden de aparición de las funciones.

3.1 Menú

Al utilizar la herramienta, el primer vistazo se compone de dos columnas. La columna izquierda posee el menú, el cual está conformado por el logo de la herramienta, y una lista de páginas. Cada página tiene una funcionalidad diferente. Estas páginas son:

- Home
- Data Analysis
- Pathway Analysis
- Enrichment Analysis
- Chronological Analysis
- Tutorials
- About
- Contact Us

Y la columna derecha en donde se mostrará la información perteneciente a la pestaña elegida en el menú para la interacción del usuario.

3.2 Home

En este espacio, se muestra información descriptiva de la herramienta. Invita al usuario a probar diferentes funciones, de forma gráfica da ejemplos de lo que se puede realizar, da consejos y breves descripciones para aclarar dudas que puedan surgir, también da a conocer funciones futuras pensadas para ser integradas a la herramienta.

3.3 Data Analysis

En esta sección, se encuentra la función del análisis de datos del archivo importado por el usuario. Acá se desplegarán las tablas y gráficos correspondientes, para la comparación entre metabolitos, pacientes, vías y super vías. Después del usuario subir el archivo ya sea tipo Excel o csv, se le da la elección de elegir la información que quiere representar en un gráfico, y también el tipo de grafico a utilizar. Esta información corresponde a graficar los pacientes seleccionados en función de compuestos, vías o super vías. El tipo de gráfica está a disposición de la preferencia del usuario ya que puede elegir entre gráficos de barras, puntos, caja y mapas de calor.
Como se quiere ayudar a encontrar la mayor cantidad de información posible y la relación entre los datos introducidos, se buscó la manera de implementar machine learning para encontrar patrones o alguna clasificación entre ellos. El uso de machine learning es muy

popular por su efectividad, pues ayuda a encontrar y relacionar información de grandes conjuntos de datos. Sin embargo, para que sea óptimo el utilizar estos métodos se debe contar con una cantidad adecuada de información. Como se ha descrito, *TANGO2*-RMEA no cuenta con mucha información, y al momento de querer aplicar estos métodos a los data sets, no es posible porque hay información faltante como clases o etiquetas en caso de aprendizaje supervisado, o información complementaria.
Se hizo una revisión literaria de cómo proceder en casos donde se estudian enfermedades raras, y entre muchos artículos científicos se destaca uno por los consejos y estrategias que propone, este artículo se denomina “Oportunidades y desafíos para el aprendizaje automático en enfermedades raras” [93] y la información más conveniente en este caso se basa en extender la información, y lo propone de varias maneras:

- Usando datos disponibles fuera de la enfermedad de interés con el fin de recolectar ejemplos no etiquetados.
- Heredar y ajustar modelos de aprendizaje.
- Aumentar los datos probando el conjunto con muestras construidas artificialmente.

Siguiendo lo anterior, se aplicaron dos modelos con el fin de encontrar similitudes entre los datos. Esto se hizo para las super vías, ya que todos los metabolitos pertenecen a alguna de las 9 encontradas en el archivo importado. Los dos métodos son:

- Clustering aglomerativo, el cual va uniendo grupos con base a características para encontrar algún tipo de homogeneidad [94].
- K-means, asigna cada observación a grupos en función de la proximidad de la observación a la media del grupo [95].

El número de clusters o grupos (n), es el número de grupos diferentes que se pueden encontrar. Luego de determinar el número de clusters, se aplican ambos modelos, dando como resultado dos tablas con *n* grupos. En la tabla aparece el nombre de las super vías por cluster, el número de compuestos perteneciente, y el porcentaje de cuantos compuestos de cada super vía del total aparecen en el cluster. Para una mejor comprensión también aparece un gráfico de barras. En cada cluster se eligen las super vías con mayor porcentaje, pues quiere decir que encontró similitudes entre estas.

3.4 Pathway Analysis

Esta pestaña se dedica exclusivamente a la visualización de vías metabólicas. Estas vías fueron hechas por KEGG. Adicionalmente, hay tablas y buscadores complementarios para entender la información de cada vía, pues están compuestas por códigos de enzimas, genes, compuestos, reacciones, entre otra información metabólica. También hay buscadores que generan listas de la información contenida, es decir, si el usuario quiere tener todos los compuestos asociados a la vía metabólica, basta con oprimir un botón.

3.5 Enrichment Analysis

Al encontrar herramientas como MetaboAnalyst, la cual obtiene la información de los compuestos de entrada, y de ellos hace un análisis de enriquecimiento, se quiso recrear el procedimiento, pues esta herramienta está hecha en otro lenguaje de programación. Se

implementó la búsqueda de librerías que al complementarlas se lograra el mismo resultado de MetaboAnalyst. Se partió desde la información ya obtenida de la base de datos KEGG, se hace una lista de los compuestos a los que se quieren hacer un análisis de enriquecimiento. Para cada compuesto se buscan los *pathway*s o vías metabólicas asociadas y se obtiene su código de identificación de la base de datos. En este momento, se necesitaba un "puente" que determinara los genes asociados a los *pathway*s. Para ello se hizo una extensa búsqueda de bases de datos, librerías y documentos que conllevaran a obtenerlos. Se encontró una librería llamada Keggx la cual no se actualiza desde el año 2019, aun así, se decidió analizar su proceder pues inicialmente su información sólo decía "Paquete Python para manipulación y visualización de rutas KEGG." [96] Observando el código y sus funciones, había una parte en la que lograba obtener los genes en una lista.
Ya obtenido este puente, se necesitaba una forma de obtener todos los *pathway*s relacionados con los genes encontrados, para encontrar la importancia de cada *pathway* y la relación de estos independientemente de los compuestos de entrada. MetaboAnalyst utiliza Enrichr [97], así que se retomó la tarea de búsqueda en bases de datos, librerías y documentos, encontrando GSEAPY, un método computacional que determina si un conjunto de genes definido a priori muestra diferencias estadísticamente significativas y concordantes entre dos estados biológicos (por ejemplo, fenotipos). [98]
Finalmente, toda la información se agrupó en una tabla con el fin de graficar el análisis de enriquecimiento.

3.6 Chronological Analysis

Adicional a la tabla en Excel en donde se presentan los niveles metabólicos de varios pacientes, se tiene la opción de importar otros archivos también en formato Excel que representan los niveles de un conjunto de compuestos en un paciente tomados en varios tiempos. La herramienta permite ver los compuestos seleccionados libremente por el usuario, en una gráfica interactiva, y también una tabla con la información, incluyendo una columna que determina si el paciente ha mejorado o no con base a la media de los distintos tiempos.

3.7 Tutorials

Este espacio está designado a mostrar los tutoriales de forma sencilla al usuario. Puede ser en distintos formatos como imágenes, videos, diapositivas, entre otros. De igual forma se encontrará el *link* del repositorio de GitHub para más información.

3.8 About

Toda la información relevante sobre la herramienta, las publicaciones relacionadas, información profesional de la autora, entre otras cosas se encuentran en este apartado.

3.9 Contact Us

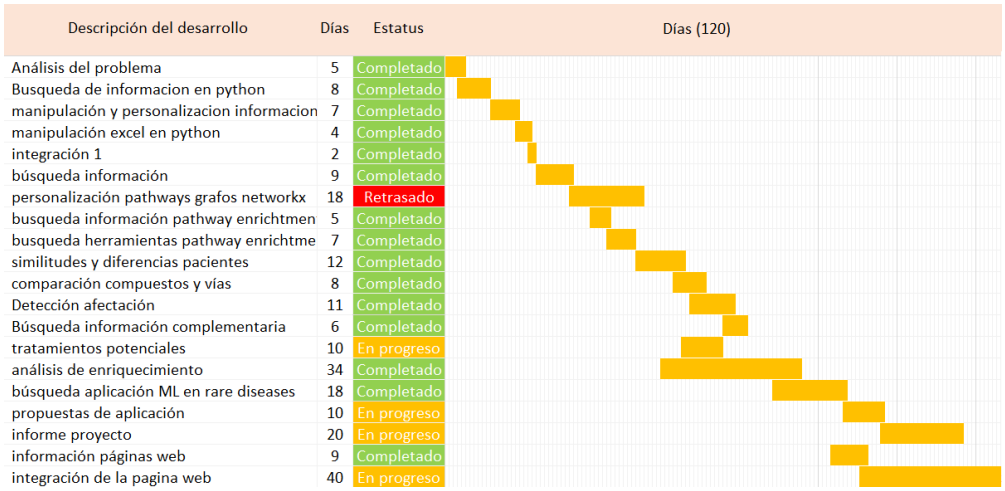
Por último, esta sección permite al usuario contactarse con la persona a cargo de la herramienta. Puede ser para formular preguntas, dar sugerencias, hacer comentarios, y todo lo relacionado.

3.10 Integración página web.

Para la creación de la página web se utilizó un framework llamado Dash. Esta se utiliza para crear e implementar aplicaciones de datos con interfaces de usuario personalizadas utilizando Python, R, Julia o F# (experimental). Está escrito sobre Plotly.js y React.js, Dash es ideal para. Es particularmente adecuado para cualquiera que trabaje con datos [99].
Se utilizó Spyder y Jupyter notebook. Spyder es una extensión de Anaconda, este entorno científico gratuito y de código abierto está escrito en Python, para Python y diseñado por y para científicos, ingenieros y analistas de datos. Cuenta con una combinación única de la funcionalidad avanzada de edición, análisis, depuración y creación de perfiles de una herramienta de desarrollo integral con la exploración de datos, la ejecución interactiva, la inspección profunda y las hermosas capacidades de visualización de un paquete científico [100]. Para la migración y portabilidad, se escogió Heroku, es una plataforma de servicios en la nube que permite manejar los servidores y sus configuraciones, escalamiento y la administración. Su popularidad ha crecido en los últimos años debido a su facilidad de uso y versatilidad para distintos proyectos [101]. De esta forma se puede utilizar la herramienta con un link, de la misma manera que una red social.

DIAGRAMA DE GANTT

El cronograma inicial se muestra en la *Figura 5*, en él también aparecen funciones y tareas pensadas para la herramienta que no se han llevado a cabo, debido a que para su realización se necesitan otro tipo de análisis y un tiempo considerable. Por eso no se tuvieron en cuenta en la proyección de la herramienta. No obstante, son funciones que si se tienen pensadas implementar (*véase la sección de Recomendaciones y Trabajos Futuros*).
En el cronograma, aparece que por semana se tenían por lo general dos reuniones para la discusión del avance de la herramienta, y cada mes aproximadamente se tenía una reunión con la Doctora para tener en cuenta sus consejos, comentarios y aportes, debido a que como ya se ha mencionado, pero no sobra recordar, el procedimiento con una enfermedad rara puede llegar a ser muy diferente y más complejo a un procedimiento con una enfermedad tratable.



Descripción del desarrollo	Días	Estatus
Análisis del problema	5	Completado
Busqueda de informacion en python	8	Completado
manipulación y personalizacion informacion	7	Completado
manipulación excel en python	4	Completado
integración 1	2	Completado
búsqueda información	9	Completado
personalización pathways grafos networkx	18	Retrasado
busqueda información pathway enrichtmen	5	Completado
busqueda herramientas pathway enrichtme	7	Completado
similitudes y diferencias pacientes	12	Completado
comparación compuestos y vías	8	Completado
Detección afectación	11	Completado
Búsqueda información complementaria	6	Completado
tratamientos potenciales	10	En progreso
análisis de enriquecimiento	34	Completado
búsqueda aplicación ML en rare diseases	18	Completado
propuestas de aplicación	10	En progreso
informe proyecto	20	En progreso
información páginas web	9	Completado
integración de la pagina web	40	En progreso

Figura 5 Diagrama de Gantt

RESULTADOS

1. Integración de las bases de datos, la información y las gráficas interactivas

La herramienta se denominó “TANGBOOK”, es la unión de “TANG” haciendo referencia al caso de estudio de *TANGO2* y “BOOK” porque fue desarrollada en Jupyter notebook.
La primera impresión de TANGBOOK.

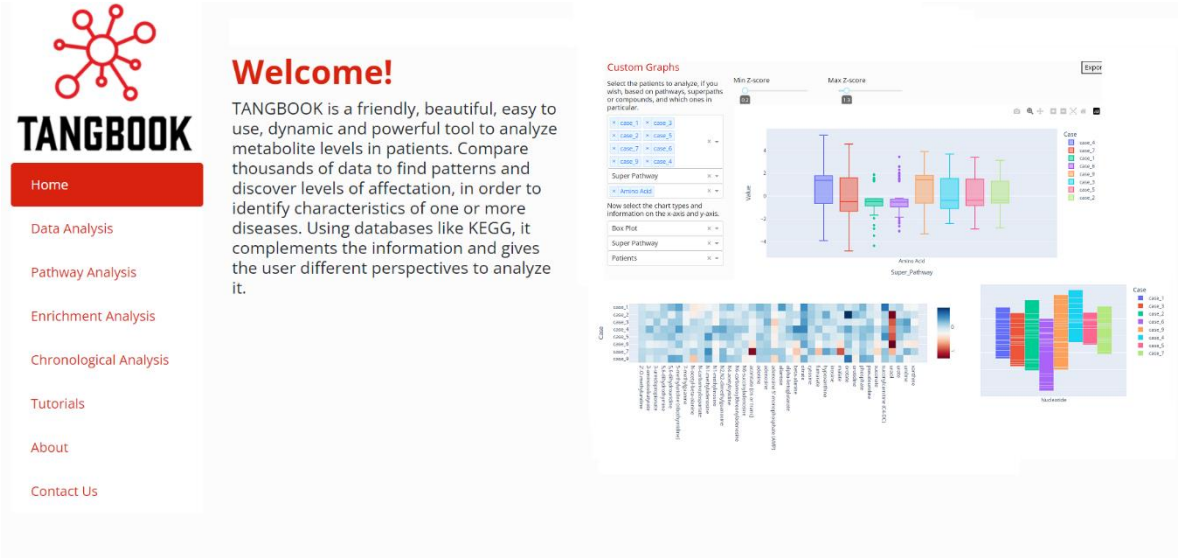


Figura 6 Interfaz inicial

En la pestaña Data Analysis, aparece información para subir el archivo y un botón de la parte superior derecha para seleccionar los documentos. Si el usuario tiene alguna duda sobre la imagen ejemplo de la estructura que debe cumplir el archivo a importar, puede hacer click en el botón ”info” para una explicación más detallada. Una vez importado el documento, aparece una gráfica de burbújas general, que permite ver todos los compuestos afectados de cada paciente, entre más afectado esté el compuesto, el tamaño de la burbuja es mayor. Los colores hacen referencia a las super vías.

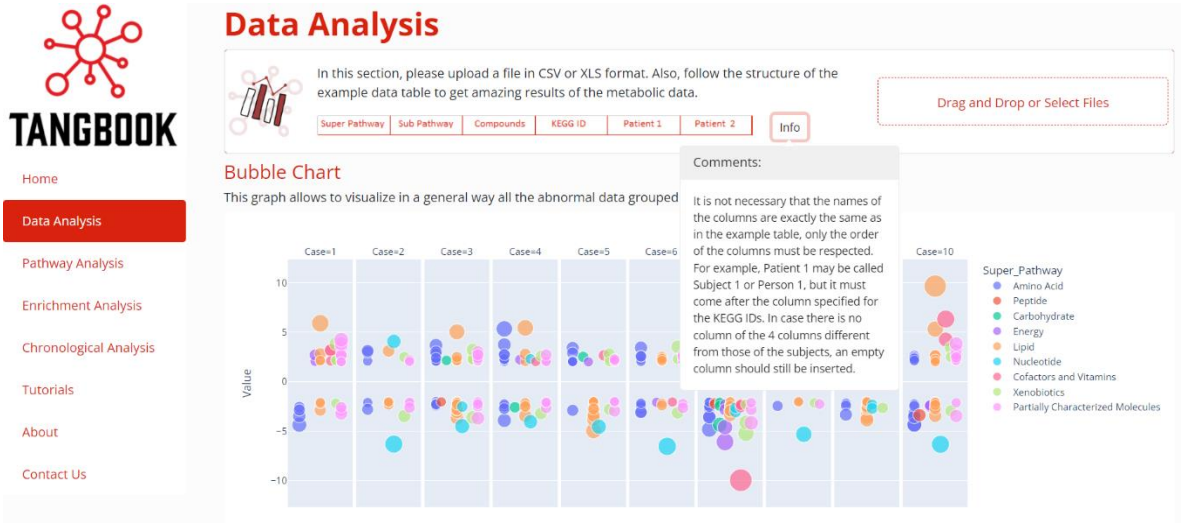


Figura 7 Visualización del archivo importado en la sección de Data Analysis

Luego de la gráfica de burbujas, aparece la tabla principal. Los compuestos superior o inferior al z-score, se marcan en azul y rojo respectivamente. La tabla resultante es dinámica, por lo que el usuario la puede organizar y filtrar. Por ejemplo, en la *Figura 8.a*, se puede observar que la flecha hacia arriba de "case_4" está presionada. El primer compuesto es el más alto, y la tabla se ordena de los compuestos más altos en "case_4" a los más bajos. En el caso de *8.b* está presionada la flecha hacia debajo de "case_2", por lo que se arregla la tabla de menor a mayor dependiendo de los valores en "case_2".
Por último, en *8c* se hizo una búsqueda en la columna "compound" teniendo aún el arreglo del caso *8.b*. De esta forma, la organización será de forma alfabética en los compuestos que empiecen por "pal" y dependiendo de cuál es menor en la columna de "case_2".

a

Super_Pathway	Sub_Pathway	Compound	KEGG_ID	case_1	case_2	case_3	case_4	case_5	case_6	case_7	case_8	case_9
filter data...												
Lipid	Phospholipid Met	trimethylamine N	C01104	5.91	3.11	5.04	5.44	1.25	-0.61	0.99	-0.28	2.01
Amino Acid	Glutamate Metabc	2-pyrrolidinone	none	0.01	0.73	2.52	5.34	3.4	-0.21	-2.14	-0.72	0.01

b

Super_Pathway	Sub_Pathway	Compound	KEGG_ID	case_1	case_2	case_3	case_4	case_5	case_6	case_7	case_8	case_9
filter data...												
Nucleotide	Pyrimidine Metab	uracil	C00106	0	-6.3	-4.49	-4.04	-4.55	-6.53	-3.01	-5.31	0
Xenobiotics	Food Component/F	tartronate (hydr	C02287	-2.19	-3.47	-3.61	-3.2	-2.8	-3.14	-5.22	-1.61	-2.65
Amino Acid	Methionine, Cyst	hypotaurine	C00519	0	-2.79	0	-0.68	0	-1.11	0	-1.66	0

c

Super_Pathway	Sub_Pathway	Compound	KEGG_ID	case_1	case_2	case_3	case_4	case_5	case_6	case_7	case_8	case_9	case_10
filter data...		pal											
Lipid	Lysophospholipic	1-palmitoyl-GPI	none	-1.18	-2.06	-1.63	-0.97	-1.85	-1.1	-1.52	0.34	-3.06	-2.44
Lipid	Phosphatidylchol	1-palmitoyl-2-ar	C05208	-2.88	-1.58	-0.44	-0.48	-0.56	-0.22	0.47	0.55	-2.85	-0.43
Lipid	Phosphatidyletha	1-palmitoyl-2-ar	C05210	-2.21	-1.37	-1.66	0.04	-0.18	0.06	0.96	-0.99	-2.78	-0.98

Figura 8 Tabla dinámica del archivo importado

Después de la tabla dinámica, se tiene la sección de "*Custom Graphs*". De esta forma, el usuario elige los pacientes que quiere analizar, la columna de información (Compound, Pathway y Super Pathway) y la información específica dependiendo de la anterior selección. Por ejemplo, en la *Figura 9* se seleccionaron los pacientes 1,5,8 y 10; se desea ver la

información de las super vías y específicamente “Energy”, “Cofactors and Vitamins”, “Péptide” y “Amino Acid”.
Aparte de escoger la información a graficar, el usuario puede elegir cómo quiere graficarla, ya que puede decidir qué información va en cada eje, el tipo de gráfica y también el z-score mínimo y máximo con ayuda de deslizadores.

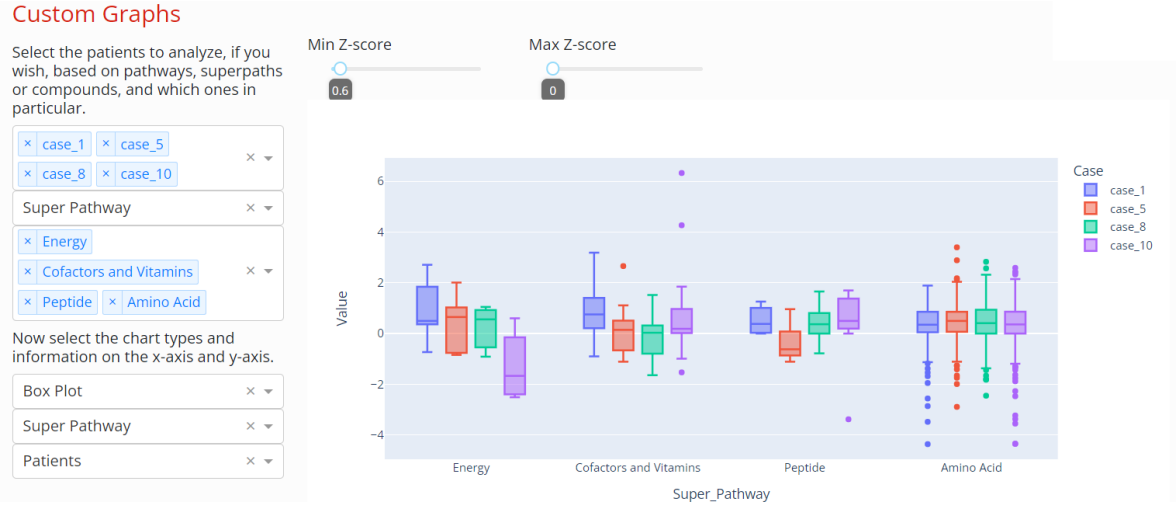


Figura 9 Dataframe preprocesado.

En la Figura 10, se muestran las 4 formas de representar la información. En el gráfico de barras, burbujas y caja, los colores hacen referencia a los pacientes. Mientras que en el mapa de calor, el rojo y azul más intenso representa un valor muy pequeño o muy grande respectivamente.

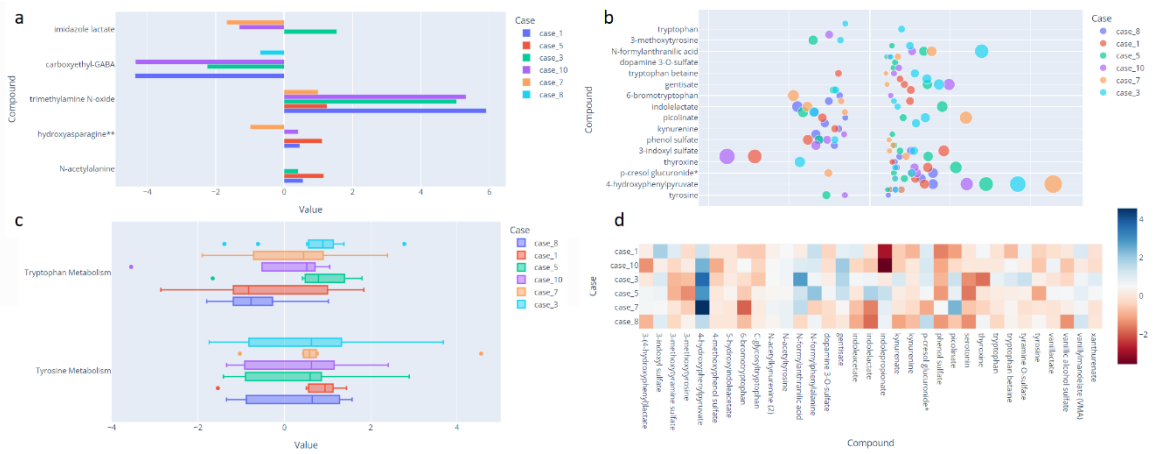


Figura 10 Diferentes formas de graficar la información

Los compuestos que no se presentaron en ningún paciente se apartaron en una tabla, junto con las columnas de Super_Pathway, Pathway y KEGG_ID (*Figura 11*). Para esta tabla, se hizo un grafico de pastel que indica cual vía o sub vía tiene mayor ausencia en los datos. También se tiene la elección de cuanta información se desea ver en el gráfico de pastel para las vías. Por defecto grafica las 20 vías con más compuestos.

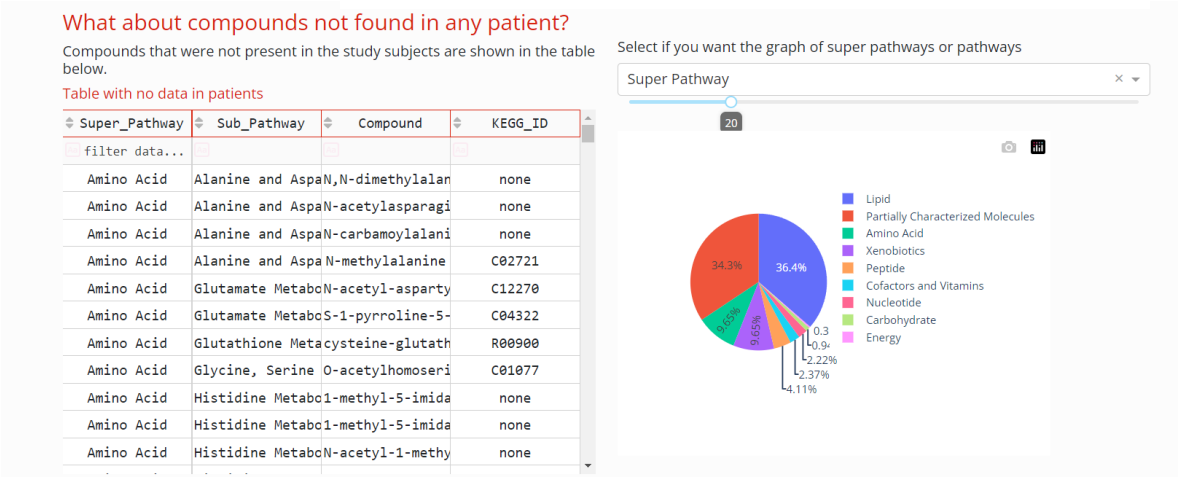


Figura 11 Visualización de la información seleccionada

Por último, esta sección nos brinda un pequeño análisis utilizando machine learning. Con los dos métodos mencionados anteriormente de aprendizaje no supervisado, se obtienen dos gráficas que agrupan la información en clusters (*Figura 12).*

Figura 12 Uso de Machine Learning para encontrar homogeneidad en los datos

En la siguiente pestaña denominada Pathway Analysis, el usuario se encuentra con un buscador y unas tablas complementarias que ayudan a entender la información de las vías metabólicas *Figura 13*. Como se observa, de la tabla de Pathways se eligió la vía metabólica del metabolismo del glutatión (00440), en la tabla de compuestos se buscó el primer compuesto encerrado en el ovalo azul utilizando la palabra “Oxopro” en el buscador. Y en la tabla de enzimas se buscó el código encerrado en el ovalo rojo.
En la *Figura 14* se muestra una tabla, con los compuestos presentes en la vía metabólica escogida, y la imagen explicativa de la simbología.

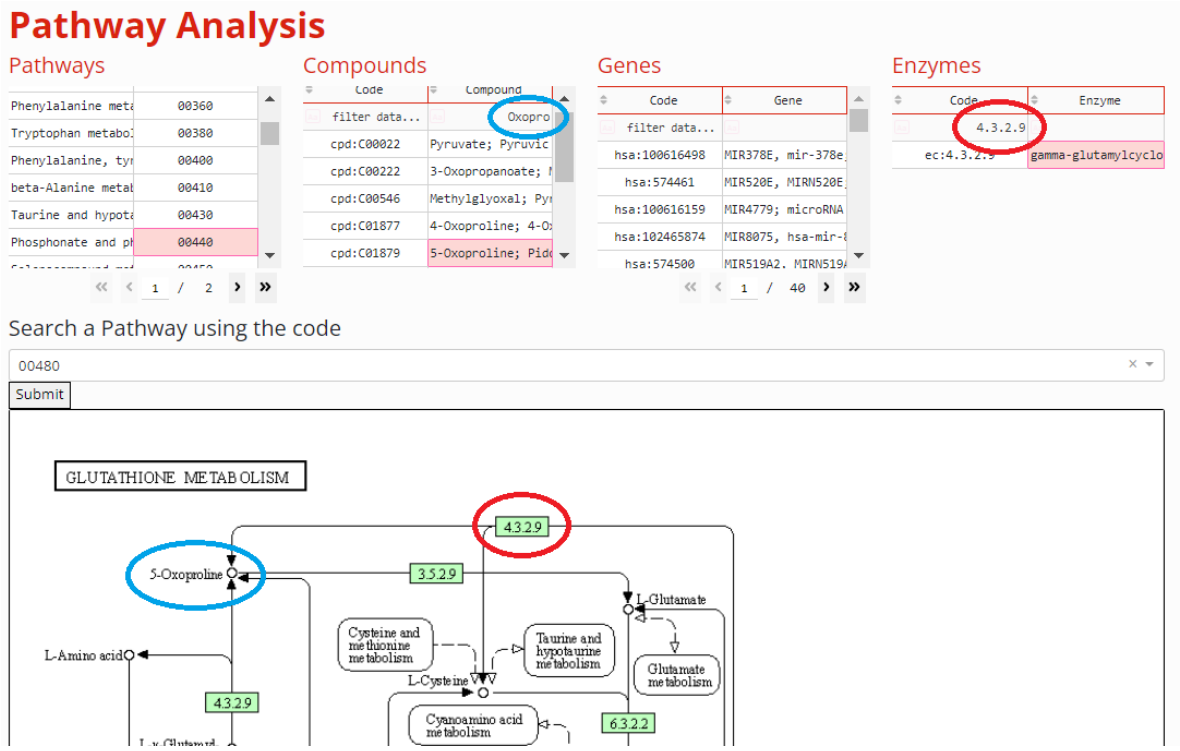


Figura 13 Vía metabólica del metabolismo del glutatión.

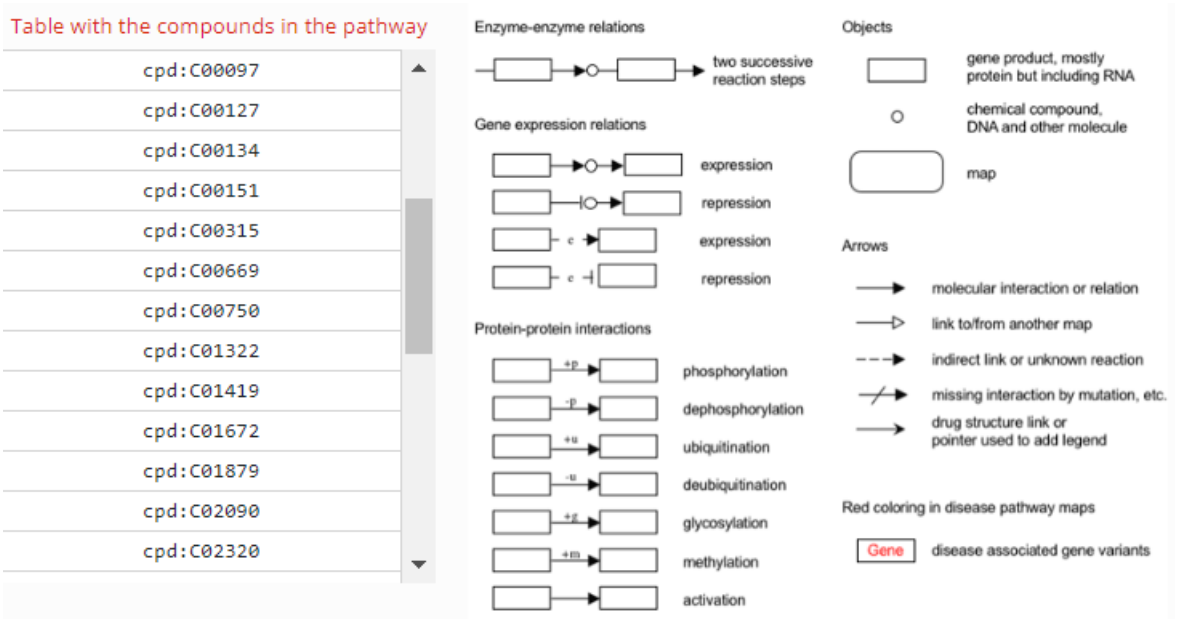


Figura 14 Simbología de las vías metabólicas de KEGG

La siguiente pestaña corresponde a “Enrichment Analysis”. En esta se hace el análisis de enriquecimiento de los compuestos seleccionados en el buscador. Inicialmente, se quería obtener una gráfica similar a la de MetaboAnalyst. Como ejemplo, se usarán 3 compuestos aleatorios los cuales son Glycine, Sarcosine y L-Phenylalanine.

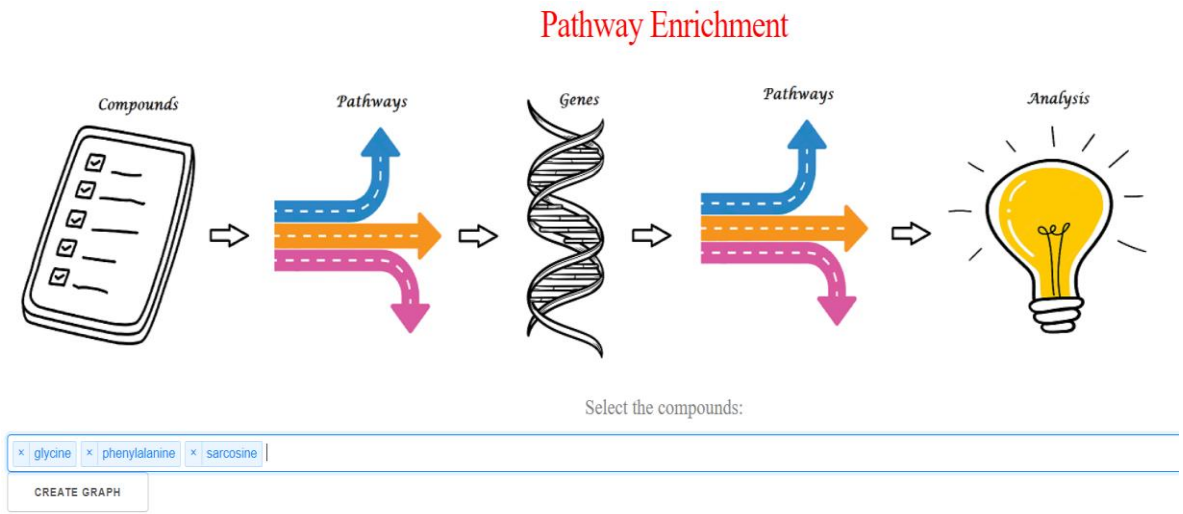


Figura 15 Análisis de enriquecimiento

Al ingresar esta lista de compuestos, en MetaboAnalyst resultan 8 *pathway*s del análisis de enriquecimiento, y escritos de mayor importancia a menor.

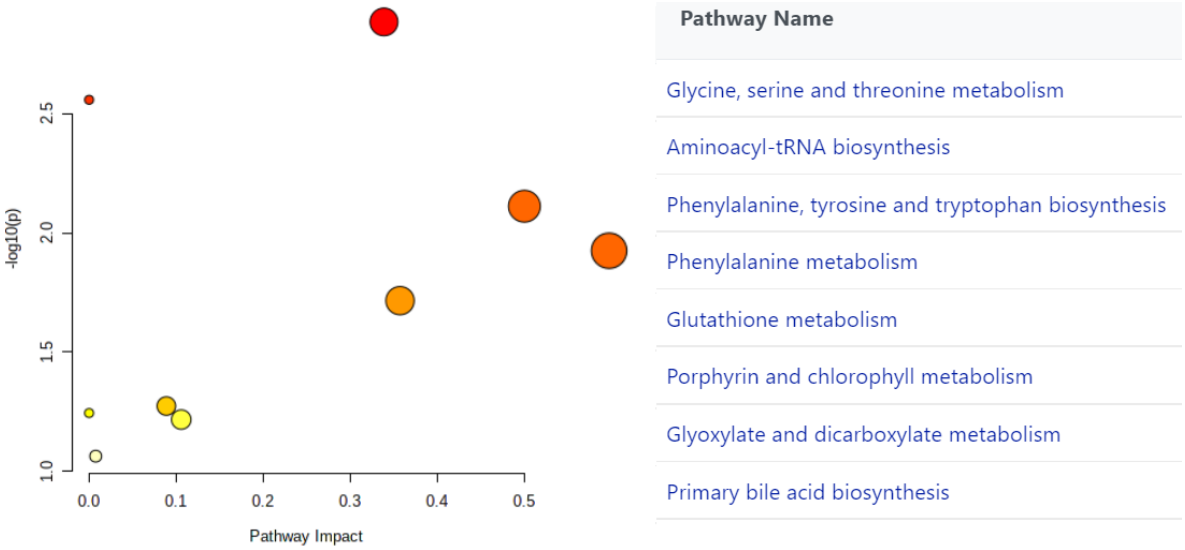


Figura 16 Análisis de enriquecimiento hecho por MetaboAnalyst.

Después de conectar toda la información de distintas fuentes, se obtuvo un dataframe de aproximadamente 200 filas, quiere decir que encontró esa cantidad de *pathway*s asociados, el impacto se obtuvo calculando el *overlap* y se graficó modificando los valores según el impacto pues, entre menor sea el impacto mínimo y mayor el impacto máximo, más *pathway*s se verán en la gráfica. Esto se hace con el fin de ignorar los más obvios y los menos significativos, aunque, por ser un análisis de una enfermedad rara se puede considerar valiosa toda la información al respecto.

Min 0.39 Max 0.76

impact	Term	Overlap	P-value	Adjusted P-value	Odds Ratio	Combined Score	Genes	log_value
0.539823	Carbon metabolism Homo sapiens hsa01200	61/113	1.262408e-55	9.678458e-54	31.093840	3930.546150	GPI;GLDC.	122.069692
0.554054	Biosynthesis of amino acids Homo sapiens hsa01230	41/74	2.537532e-38	1.459081e-36	32.077287	2776.835844	TAT;SHM.	82.515257
0.725000	Glycine, serine and threonine metabolism Homo ...	29/40	1.483939e-32	6.826119e-31	67.061409	4914.798072	DMGDH;ALA.	69.459382
0.515152	Aminoacyl-tRNA biosynthesis Homo sapiens hsa00970	34/66	1.571944e-30	6.025784e-29	27.177333	1865.051019	MTFMT;TR.	64.978920
0.612245	N-Glycan biosynthesis Homo sapiens hsa00510	30/49	3.397866e-30	1.116442e-28	40.200335	2727.769827	DPAGT1;B.	64.362236

Figura 17 Tabla de análisis de enriquecimiento obtenida.

Ahora se procede a graficar tomando el impacto y el cálculo de $-log10(adjusted\ pvalue)$ que son los datos de la columna log_value.

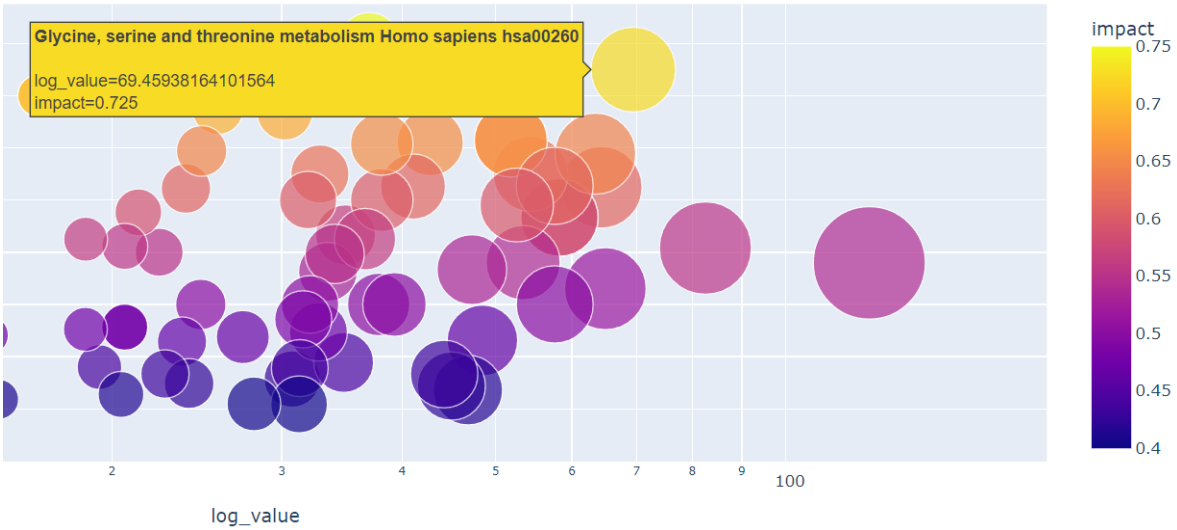


Figura 18 Visualización análisis de enriquecimiento

En la pestaña de “Chronological Analysis” se muestra la conexión de la presentación clínica y los metabolitos de forma cronológica, para esta sección, el usuario puede insertar varios archivos en Excel, en donde se organizan los datos y se crea una tabla con los más relevantes, y al final de esta hay una columna llamada “hs” que determina si el paciente ha mejorado o no, donde 1 es que si, y 0 es que no. (Figura 20)
La información de esta gráfica también se puede determinar por medio de una tabla que indica exactamente la misma información de la gráfica (*Figura 19*).

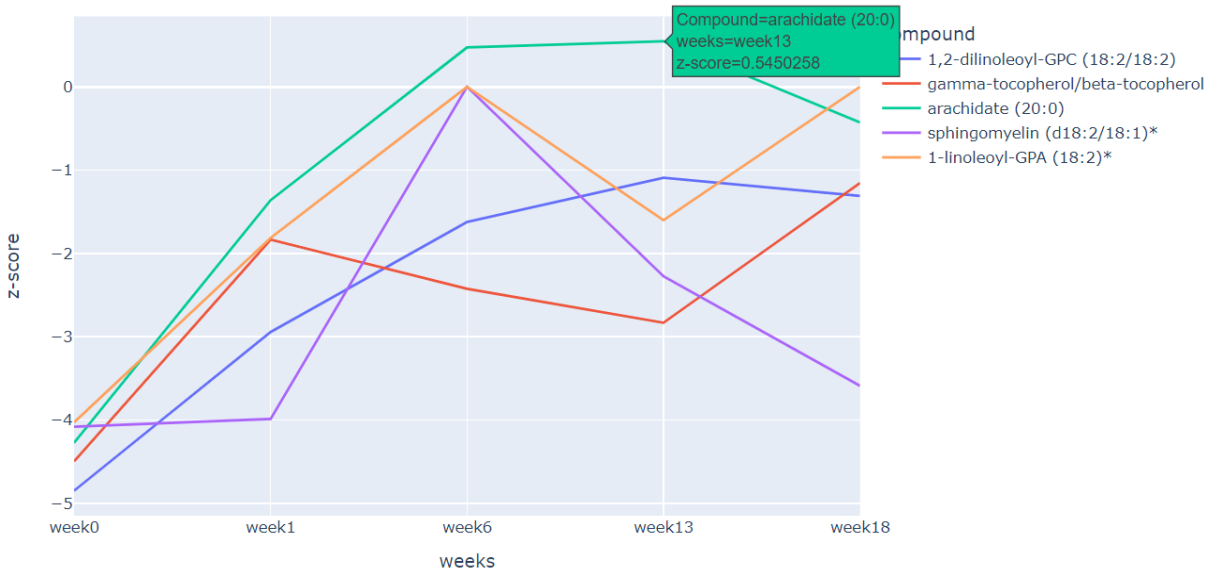


Figura 19 Visualización datos cronológicos

	BIOCHEMICAL	SUPER_PATHWAY	SUB_PATHWAY	week0	week1	week6	week13	week18	hs
0	1,2-dilinoleoyl-GPC (18:2/18:2)	Lipid	Phosphatidylcholine (PC)	-4.849025	-2.944256	-1.622565	-1.092238	-1.308942	1
1	p-cresol sulfate	Xenobiotics	Benzoate Metabolism	-4.699576	-4.699576	-4.699576	-4.699576	-1.511223	1
2	gamma-tocopherol/beta-tocopherol	Cofactors and Vitamins	Tocopherol Metabolism	-4.496610	-1.835366	-2.425313	-2.832784	-1.153454	1
3	arachidate (20:0)	Lipid	Long Chain Saturated Fatty Acid	-4.273723	-1.361284	0.471657	0.545026	-0.428861	1
4	sphingomyelin (d18:2/18:1)*	Lipid	Sphingomyelins	-4.081903	-3.987498	-3.987498	-2.276046	-3.589586	0
5	1-linoleoyl-GPA (18:2)*	Lipid	Lysophospholipid	-4.029421	-1.815643	-1.815643	-1.601743	-1.601743	1
6	kynurenate	Amino Acid	Tryptophan Metabolism	-3.781672	0.906115	0.384713	0.580021	0.768195	1

Figura 20 Tabla de datos cronológicos

N llas pestañas de “Tutorials” y “About Us” se encuentra información básica ya mencionada y por último en la pestaña de “Contact Us” se le brinda al usuario un formulario para contactarnos. (*Fígura 21*)

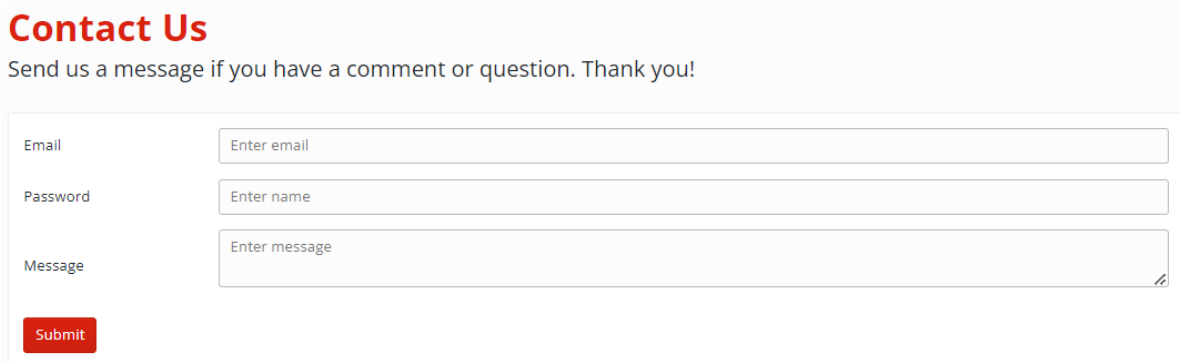


Figura 21 Pestaña "Contact Us"

DISCUSIÓN

1. Diferencias entre TANGBOOK y otras herramientas

- Esta herramienta de forma visual muestra los datos de los pacientes del *dataset* de diversas formas, el usuario puede cambiar la representación dependiendo de su preferencia. Después de comparar TANGBOOK con las herramientas en el estado del arte, se aprecian las diferencias en cuanto a la funcionalidad. Aunque tanto TANGBOOK como las herramientas vistas se centran en análisis de datos metabólicos, todas muestran cierta complejidad en el sentido de ingresar la información menos TANGBOOK. Por ejemplo, para el análisis de enriquecimiento en MetaboAnalyst se debe ingresar uno a uno el código del compuesto, mientras que en TANGBOOK hay una tabla que muestra los compuestos con su nombre y su código, por lo que solo es cuestión de oprimir en el botón desplegable.

- Las otras herramientas no dan la libertad de graficar la información de forma personalizada, tampoco cuentan con tablas dinámicas.

- El tipo de análisis que utiliza TANGBOOK por ahora es mucho más sencillo, pero esto es una ventaja pensada para el caso de estudio el cual está basado en una enfermedad rara. Básicamente partiendo desde cero, se buscan patrones y similitudes en los valores metabólicos de los pacientes, o en conjuntos de vías y super vías. Al ser una funcionalidad sencilla que posiblemente hasta se pueda hacer en Excel, otras herramientas no la tienen, y en Excel llega a ser muy tedioso teniendo una tabla tan amplia. Eso sin contar que el usuario debe tener un nivel de manejo de Excel superior al promedio.

- MetaboAnalyst muestra las mismas vías metabólicas hechas por KEGG, pero no da a conocer cierta información complementaria como la simbología o los buscadores de términos que el usuario no conoce, como lo hace TANGBOOK. Sin embargo, MetaboAnalyst en ciertos casos muestra en pequeñas ventanas información del grafo al parar el cursor en el término.

- En el análisis de enriquecimiento, aunque no dio exactamente la misma gráfica que realiza MetaboAnalyst, se considera que aborda más información porque tiene en consideración valores muy pequeños, y no obvia información que puede llegar a ser casi imperceptible. Las razones por las que no se obtuvo el mismo resultado es porque MetaboAnalyst en el análisis de enriquecimiento utiliza análisis de sobrerrepresentación junto con perfilado de muestra única y un análisis de enriquecimiento cuantitativo, mientras que TANGBOOK solo utiliza un análisis de enriquecimiento de genes.

- Se considera que es crucial para el análisis general de una enfermedad rara tener en cuenta toda la información obtenida en un análisis y no descartarla sin tener una razón, ya que brinda más caminos que paso a paso pueden culminar en brindar la información sobre la enfermedad que hoy en día es faltante y necesaria. Por último, TANGBOOK permite la comparación de varias muestras de pacientes al mismo tiempo, lo que corresponde una ventaja con respecto a MetaboAnalyst.

2. ¿Es TANGBOOK una herramienta reproducible?

Quien determina si los resultados fueron buenos, es decir, si se cumplen los objetivos de la mejor manera es el usuario en este caso el usuario es el personal médico.
Para determinar qué tan sencilla y amigable es TANGBOOK, dos personas utilizaron la herramienta sin supervisión del experto, y gracias a la información que ésta misma brinda se tuvieron comentarios positivos, calificándola de intuitiva, sencilla de utilizar, expresiva, entre otras características. Por tal motivo, y teniendo en cuenta la funcionalidad no encontrada en otras herramientas, TANGBOOK puede ser relevante para otros tipos de casos con objetivos comparativos sin importar la complejidad de los datos.

Para saber si TANGBOOK es reproducible, se debe determinar si funciona. Es decir, si los análisis y resultados no están muy lejos de la realidad. Para esto, la Doctora Claudia propuso utilizar la herramienta con un dataset ya analizado. Esto con el fin de comparar los resultados obtenidos en TANGBOOK y el estudio previo.

Resultados de TANGBOOK:

Al cargar el dataset y observar el primer grafico correspondiente a Bubble Chart, se nota que hay un compuesto que en todos los pacientes está muy por encima del z-score. Este compuesto corresponde a “citrulline” y en los 10 casos se presenta por encima de 10.

Figura 22 Bubble Chart del ciclo de urea

Gracias a esta gráfica se determinaron algunos otros compuestos afectados notoriamente. Para complementar, se procedió a revisar y filtrar la tabla dinámica.

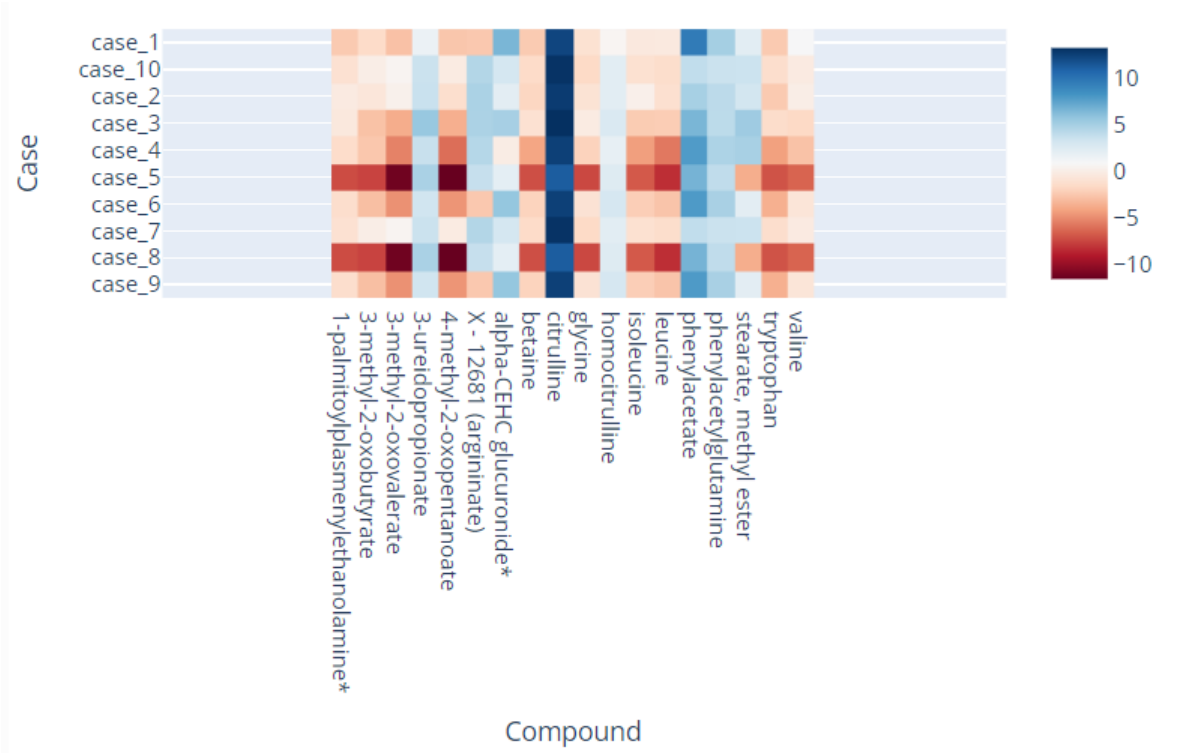
Compound	KEGG_ID	case_1	case_2	case_3	case_4	case_5	case_6	case_7	case_8	case_
citrulline	C00327	12.27621122	12.73555034	13.25363624	12.53104613	11.20138731	12.5222674	13.10739276	11.20138731	12.5222
phenylacetate	C07086	9.618151357	4.99219915	6.573528284	7.871869969	6.680241582	7.957546137	3.942985524	6.680241582	7.957546
X - 12681 (argir	none	-2.507611569	4.821118857	4.801810879	4.419156279	3.76300075	-2.507611569	4.509489037	3.76300075	-2.50761
phenylacetylglut	C04148	5.05318716	4.163084023	4.141468849	4.758029589	4.015767872	4.92148228	3.573197194	4.015767872	4.921482
3-ureidopropiona	C02642	1.60674839	3.635550848	5.56811488	3.692385875	4.87220485	3.321818388	3.547411935	4.87220485	3.321818
N-acetylarginine	C02562	-0.217192668	3.441295816	0.791512092	1.760914929	-0.219352676	0.557523372	2.822826644	-0.219352676	0.557523
stearate, methyl	none	2.245836354	3.217216299	5.31718812	4.936824492	-3.714026464	2.184786319	3.511071544	-3.714026464	2.184786
palmitoyl sphing	none	0.480018282	2.923422129	1.295280919	-0.034030761	-3.110172989	0.570736065	2.475035991	-3.110172989	0.570736

Figura 23 Tabla dinámica filtrada por compuestos más altos

Compound	KEGG_ID	case_1	case_2	case_3	case_4	case_5	case_6	case_7	case_8	case
4-methyl-2-oxope	C00233	-2.604659482	-1.253601223	-3.644489089	-6.088730268	-11.5585304	-4.696387324	-0.240558966	-11.5585304	-4.69638
3-methyl-2-oxova	C00671	-2.836306066	0.235645128	-3.766505206	-5.365478107	-11.27334371	-4.820251429	0.527391603	-11.27334371	-4.82025
leucine	C00123	-0.382537191	-1.189418102	-2.28388415	-5.715124961	-8.330634537	-2.685878138	-1.353278469	-8.330634537	-2.68587
3-methyl-2-oxobu	C00141	-1.553447344	-0.647279725	-2.802483336	-2.570565307	-7.578625873	-2.937826729	-0.03414509	-7.578625873	-2.93782
glycine	C00037	-1.057176581	-0.954676692	-0.250261541	-1.998552697	-7.435897269	-0.984976682	-1.640970756	-7.435897269	-0.98497
1-palmitoylplasm	none	-2.433248956	-0.314272897	-0.483820576	-1.429260162	-7.29218064	-1.327937277	-1.07849637	-7.29218064	-1.32793
betaine	C00719	-2.371300799	-1.81650308	-1.072252545	-4.048687169	-7.182659352	-1.966393765	-1.550470986	-7.182659352	-1.96639
tryptophan	C00078	-2.413173998	-2.415753768	-1.424000208	-4.249779682	-7.053336298	-3.622129902	-1.340918562	-7.053336298	-3.62212

Figura 24 Tabla dinámica filtrada por compuestos más bajos

Como resultado los siguientes compuestos fueron los más afectados en la mayoría de pacientes.



Ahora, se compara con el estudio realizado previamente denominado “Untargeted metabolomic profiling reveals multiple pathway perturbations and new clinical biomarkers in urea cycle disorders” [102].
Haciendo una comparación, los compuestos identificados fueron los mismos, es decir, que el investigador puede hacer análisis complejos, partiendo de uno relativamente sencillo hecho con TANGBOOK.

RECOMENDACIONES Y TRABAJOS FUTUROS

Para mejorar la herramienta se tienen varios cambios:

- Implementar más técnicas de machine learning con el fin de complementar la información ya obtenida y dar una visión más amplia de los datos.

- Agregar otros análisis de enriquecimiento haciendo uso de otros métodos.

- Complementar la información y la automatización. Esto se quiere lograr, empezando por adicionar más información relevante y complementaria a la base de datos y/o a los gráficos interactivos, por ejemplo, de forma interactiva en los compuestos agregar su fórmula química, su estructura 3D, y en lo posible también para los *pathway*s y *super pathway*s.

- También se espera ampliar y mejorar la adquisición de los tratamientos potenciales, posiblemente aplicando técnicas de selección o de aprendizaje automático.

CONCLUSIONES

- Los gráficos fueron los esperados, cumplen la función de determinar de forma visual qué pacientes están más afectados que otros, qué metabolitos están disminuidos o elevados y cuanto están afectados los *pathway*s y/o *super pathway*s. Además, no solo se puede aplicar al análisis de pacientes de *TANGO2*-RMEA, también se puede aplicar a otras enfermedades y problemas similares.
- Se proyectó e implementó una herramienta capaz de procesar, analizar y visualizar datos metabólicos.
- Haciendo uso de varios métodos y funciones, se logró analizar y visualizar la información requerida para el caso de estudio de *TANGO2*-RMEA.
- Se analizaron y complementaron los datos metabólicos, contrastándolos entre sí, encontrando similitudes y diferencias.
- Se logró hacer una comparación del avance de un paciente dependiendo del tiempo transcurrido, para determinar si mejoró o si sus niveles metabólicos estaban más afectados.
- Se hizo un exitoso análisis de enriquecimiento con una basta cantidad de información.
- Los gráficos obtenidos son claros, y fácilmente se puede contrastar la información de un paciente con otro.
- Los gráficos generados son interactivos, y son más funcionales con ayuda de los widgets.
- La herramienta es sencilla de utilizar, intuitiva y cómoda.
- La herramienta está programada en un lenguaje muy utilizado a nivel mundial el cual es Python.

ANEXOS

1. Tratamientos potenciales.

Al ser una enfermedad rara, no se puede decir con exactitud si los tratamientos o procedimientos propuestos son correctos, pues se necesitaría una investigación por parte de personal médico, químico y biológico; después pruebas que den indicios de su funcionamiento y por último una aceptación por parte de entes reguladores de fármacos. Sin embargo, al existir muy poca información sobre la enfermedad y el tratamiento, este paso se hace con la intención de generar ideas, o abrir caminos que puedan conducir a un exitoso resultado. Para esta sección se utilizó una herramienta llamada Ontobee, la cual tiene como objetivo facilitar el intercambio, la visualización, la consulta, la integración y el análisis de datos ontológicos [103]. Con ella se buscan los códigos de identificación de las enfermedades asociadas a *TANGO2*-RMEA. Después, como referencia a procesos con fármacos por parte de ROBOKOP (Razonamiento sobre objetos biomédicos vinculado en Vías orientadas al conocimiento) es un sistema abierto de preguntas y respuestas, construido sobre múltiples bases de datos biomédicas abiertas y diseñado para explorar las relaciones entre una variedad de tipos de datos biomédicos. [104], se identificaron posibles drogas o tratamientos químicos para las enfermedades que componen a *TANGO2*-RMEA, esto se hizo utilizando técnicas de Deep Learning con nodos o neuronas.

2. Información complementaria

En la información complementaria, para los códigos identificadores de las bases de datos se obtiene la información química correspondiente. En ella aparece la información relacionada al compuesto elegido anteriormente, por ejemplo, se escogieron los compuestos con código “C00037” y “C00719”, y al obtener la información de la base de datos KEGG, sabemos que se trata de los compuestos Glycine y Betaine respectivamente. También aparece información como la fórmula química, el peso, masa, reacciones, *pathway*s en los que está involucrado el compuesto, la clasificación de drogas o fármacos denominado “brite”, los códigos de identificación en otras bases de datos como PubChem, enzimas asociadas, entre otra información (*Figura 17*).

```
ENTRY       C00037                      Compound
NAME        Glycine;
            Aminoacetic acid;
            Gly
FORMULA     C2H5NO2
EXACT_MASS  75.032
MOL_WEIGHT  75.0666
REMARK      Same as: D00011
REACTION    R00364 R00365 R00366 R00367 R00368 R00369
            R00373 R00374 R00395 R00478 R00497 R00565
            R00611 R00652 R00751 R00775 R00830 R00899
            R01424 R01723 R01766 R01957 R02452 R02551
            R03284 R03425 R03579 R03654 R03718 R03956
            R04486 R04777 R04951 R05055 R05704 R05835
            R07226 R07463 R08195 R08196 R08252 R08701
            R09717 R09718 R10060 R10062 R10179 R10685
            R10994 R11664 R12026 R12514 R12692 R12693
            R12723 R12941 R12967
```

```
ENTRY       C00719                      Compound
NAME        Betaine;
            Trimethylaminoacetate;
            Glycine betaine;
            N,N,N-Trimethylglycine;
            Trimethylammonioacetate
FORMULA     C5H11NO2
EXACT_MASS  117.079
MOL_WEIGHT  117.1463
REMARK      Same as: D07523
REACTION    R02565 R02566 R02821 R07228 R07244 R08211 R08212 R10061
            R10062 R12614 R12787
PATHWAY     map00260  Glycine, serine and threonine metabolism
            map01100  Metabolic pathways
            map02010  ABC transporters
MODULE      M00555  Betaine biosynthesis, choline => betaine
ENZYME      1.1.3.17        1.2.1.8         1.13.11.90      1.21.4.4
            2.1.1.5         2.1.1.157       2.1.1.161       2.1.1.162
```

Figura 25 Información resumida obtenida de la base de datos KEGG

Toda la información acerca de este proyecto se encuentra en:
https://github.com/johaojeda/TANGBOOK

REFERENCIAS

[1] Ministerio de la Protección Social. (2017). Pautas de Auditoría para el Mejoramiento de la Calidad de la Atención en Salud. Imprenta Nacional de Colombia, Bogotá, D. C., 2007

[2] Chakraborty, C. (2014). Bioinformatics: Approaches and Applications. Biotech.

[3] Boron, W. F., & Boulpaep, E. L. (2017). Medical physiology (Tercera edición). Editorial Elsevier.

[4] Ramirez, W. (2017). Rare Diseases: Prevalence, Treatment Options and Research Insights (1.a ed.). Nova Science Pub Inc.

[5] 1. Sernadela P, González-Castro L, Carta C, van der Horst E, Lopes P, Kaliyaperumal R, et al. Linked registries: connecting rare diseases patient registries through a semantic web layer. Biomed Res Int. (2017) 2017:8327980. doi: 10.1155/2017/8327980

[6] Decherchi, S., Pedrini, E., Mordenti, M., Cavalli, A., & Sangiorgi, L. (2021). Opportunities and Challenges for Machine Learning in Rare Diseases. Frontiers in Medicine, 8. https://doi.org/10.3389/fmed.2021.747612

[7] Ekins S. Industrializing rare disease therapy discovery and development. Nat Biotechnol. (2017) 35:117–8. doi: 10.1038/nbt.3787

[8] Colbaugh, R., Glass, K., Rudolf, C., & Tremblay Volv Global Lausanne Switzerland, M. (2018). Learning to Identify Rare Disease Patients from Electronic Health Records. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2018, 340–347.

[9] Shire, Rare Disease Impact Report. https://globalgenes.org/wp-content/ uploads/2013/04/ShireReport-1.pdf. Accessed 16 Apr 2020.

[10] Wakap SN, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28:165–173.

[11] S.T. Crooke, A call to arms against ultra-rare diseases, Nat Biotechnol 39 (6) (2021 396 2021,) 671–677.

[12] WHO. Neglected tropical diseases. www.who.int/news-room/q-a-detail/neglectedtropical-diseases [accessed October 21, 2021].

[13] E.D. Kakkis, M. O'Donovan, G. Cox, M. Hayes, F. Goodsaid, P. Tandon, et al., Recommendations for the development of rare disease drugs using the accelerated approval pathway and for qualifying biomarkers as primary endpoints, Orphanet J Rare Dis10 (1) (2015) 16.

[14] FDA. Developing products for rare diseases and conditions. www.fda.gov/industry/ developing-products-rare-diseases-conditions [accessed October 21, 2021]

[15] NIH. FAQs About Rare Diseases | Genetic and Rare Diseases Information Center (GARD) – an NCATS Program. https://rarediseases.info.nih.gov/diseases/pages/ 31/faqs-about-rare-diseases [accessed October 21, 2021].

[16] Lalani SR, Liu P, Rosenfeld JA, et al. Recurrent muscle weakness with rhabdomyolysis, metabolic crises, and cardiac arrhythmia due to Bi-allelic TANGO2 mutations. Am J Hum Genet 2016;98:347–57.

[17] Kremer, L.S., F. Distelmaier, B. Alhaddad, M. Hempel, et al., Bi-allelic Truncating Mutations in TANGO2 Cause Infancy-Onset Recurrent Metabolic Crises with Encephalocardiomyopathy. Am J Hum Genet, 2016. 98(2): p. 358-62. 10.1016/j.ajhg.2015.12.009

[18] Lalani, Seema R and Graham, Brett and Burrage, Lindsay and Lai, Yi-Chen and Scaglia, Fernando and Soler-Alfonso, Claudia and Miyake, Christina Y and Yang, Yaping. TANGO2-related metabolic encephalopathy and arrhythmias. 2018.
[19] Dines JN, Golden-Grant K, Lacroix A, et al. TANGO2: expanding the clinical phenotype and spectrum of pathogenic variants. Gen Med 2019;21:601–7.
[20] Jennions E, Hedberg-Oldfors C, Berglund AK, et al. TANGO2 deficiency as a cause of neurodevelopmental delay with indirect effects on mitochondrial energy metabolism. J Inherit Metab Dis 2019;42:898–908
[21] Dr. Felix Distelmaier - Putting TANGO2-deficiency into the right metabolic context (2021), TANGO2 research conference.
[22] Dines, Jennifer N and Golden-Grant, Katie and LaCroix, Amy and Muir, Alison M and Cintr´on, Dianne Laboy and McWalter, Kirsty and Cho, Megan T and Sun, Angela and Merritt, J Lawrence and Thies, Jenny and others. TANGO2: expanding the clinical phenotype and spectrum of pathogenic variants Genetics in Medicine. 2019. Nature Publishing Group.
[23] F. Giacomoni, G. Le Corguillé, M. Monsoor. M. Landi, P. Pericard, M. Pétéra, C. Dupcrier, M. Tremblay-Franco, J.F. Martin, D. Jacob, S. Goulitquer, E.A. Thévenot, C. Caron, Workflow+Metabolomics: a collaborative research infrastructure for computational metabolomics, Bioinformatics 31 (2015) 1493–1495.
[24] Chumnanpuen P, Hansen MAE, Smedsgaard J, Nielsen J. Dynamic metabolic footprinting reveals the key components of metabolic network in yeast Saccharomyces cerevisiae. Int J Genomics. 2014;2014:14.
[25] Fuhrer, T. and N. Zamboni, High-throughput discovery metabolomics. Curr Opin Biotechnol, 2015. 31: p. 73-8. 10.1016/j.copbio.2014.08.006
[26] Miller, M.J., A.D. Kennedy, A.D. Eckhart, L.C. Burrage, et al., Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. J Inherit Metab Dis, 2015. 38(6): p. 1029-39. 10.1007/s10545-015-9843-7
[27] Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, Bjorndahl TC, Krishnamurthy R, Saleem F, Liu P, et al. The human urine metabolome. PLoS One. 2013;8 (9):e73076.
[28] Xu, J., Li, J., Zhang, R., He, J., Chen, Y., Bi, N., Song, Y., Wang, L., Zhan, Q., & Abliz, Z. (2019). Development of a metabolic pathway-based pseudo-targeted metabolomics method using liquid chromatography coupled with mass spectrometry. Talanta, 192, 160–168. https://doi.org/10.1016/j.talanta.2018.09.021.
[29] Smedsgaard J, Nielsen J. Metabolite profiling of fungi and yeast: from phenotype to metabolome by MS and informatics. J Exp Bot. 2005;56(410):273–86
[30] Shen, B., Tang, H., & Jiang, X. (2016). Translational Biomedical Informatics. Springer Publishing.
[31] Blekherman G, Laubenbacher R, Cortes DF, Mendes P, Torti FM, Akman S, Torti SV, Shulaev V. Bioinformatics tools for cancer metabolomics. Metabolomics. 2011;7(3):329–43.
[32] Jiang, L., Sullivan, H., & Wang, B. (2022). Principal Component Analysis (PCA) Loading and Statistical Tests for Nuclear Magnetic Resonance (NMR) Metabolomics Involving Multiple Study Groups. Analytical Letters, 55(10), 1648–1662. https://doi.org/10.1080/00032719.2021.2019758
[33] Wu, H.-Y., Nöllenburg, M., & Viola, I. (2021). Graph Models for Biological Pathway Visualization: State of the Art and Future Challenges.
[34] Kohonen T, Schroeder MR, Huang TS. Self-organizing maps. New York: Springer; 2001.

[35] Rawlinson, C., Jones, D., Rakshit, S., Meka, S., Moffat, C. S., & Moolhuijzen, P. (2020). Hierarchical clustering of MS/MS spectra from the firefly metabolome identifies new lucibufagin compounds. Scientific Reports, 10(1), 1–9. https://doi.org/10.1038/s41598-020-63036-1

[36] . Wold H. Path models with latent variables: the NIPALS approach. New York: Acad Press; 1975.

[37] Nguyen, H. T., Lee, D.-K., Lee, W. J., Lee, G., Yoon, S. J., Shin, B., Nguyen, M. D., Park, J. H., Lee, J., & Kwon, S. W. (2016). UPLC-QTOFMS based metabolomics followed by stepwise partial least square-discriminant analysis (PLS-DA) explore the possible relation between the variations in secondary metabolites and the phylogenetic divergences of the genus Panax. Journal of Chromatography B, 1012–1013, 61–68. https://doi.org/10.1016/j.jchromb.2016.01.002

[38] Spiga, O. (1), Cicaloni, V. (1,2), Trezza, A. (1), Visibelli, A. (1,4), Millucci, L. (1), Bernardini, G. (1), Bernini, A. (1), Marzocchi, B. (1,5), Braconi, D. (1), Santucci, A. (1), Fiorini, C. (3), & Prischi, F. (6). (n.d.). Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. Orphanet Journal of Rare Diseases, 15(1). https://doi.org/10.1186/s13023-020-1305-0

[39] Chardin, D., Humbert, O., Bailleux, C., Burel-Vandenbos, F., Rigau, V., Pourcher, T., & Barlaud, M. (2021). Primal-dual for classification with rejection (PD-CR): a novel method for classification and feature selection—an application in metabolomics studies. BMC Bioinformatics, 22(1), 1–17. https://doi.org/10.1186/s12859-021-04478-w

[40] Han, S., Huang, J., Foppiano, F., Prehn, C., Adamski, J., Suhre, K., Li, Y., Matullo, G., Schliess, F., Gieger, C., Peters, A., & Wang-Sattler, R. (2022). TIGER: technical variation elimination for metabolomics data using ensemble learning architecture. Briefings in Bioinformatics, 23(2), 1–16. https://doi.org/10.1093/bib/bbab535

[41] Garg, R., Dong, S., Shah, S., & Jonnalagadda, S. R. (2016). A Bootstrap Machine Learning Approach to Identify Rare Disease Patients from Electronic Health Records.

[42] Sergio Decherchi, Elena Pedrini, Marina Mordenti, Andrea Cavalli, & Luca Sangiorgi. (2021). Opportunities and Challenges for Machine Learning in Rare Diseases. Frontiers in Medicine, 8. https://doi.org/10.3389/fmed.2021.747612

[43] Julia Schaefer, Moritz Lehne, Josef Schepers, Fabian Prasser, & Sylvia Thun. (2020). The use of machine learning in rare diseases: a scoping review. Orphanet Journal of Rare Diseases, 15(1), 1–10. https://doi.org/10.1186/s13023-020-01424-6

[44] Mukherjee, S., Cogan, J. D., Newman, J. H., Phillips, I. J. A., Hamid, R., Meiler, J., & Capra, J. A. (2021). Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network. The American Journal of Human Genetics, 108(10), 1946–1963. https://doi.org/10.1016/j.ajhg.2021.08.010

[45] Kitty Yu-Yeung Lau, Kei-Shing Ng, Ka-Wai Kwok, Kevin Kin-Man Tsia, Chun-Fung Sin, Ching-Wan Lam, & Varut Vardhanabhuti. (2022). An Unsupervised Machine Learning Clustering and Prediction of Differential Clinical Phenotypes of COVID-19 Patients Based on Blood Tests—A Hong Kong Population Study. Frontiers in Medicine, 8. https://doi.org/10.3389/fmed.2021.764934

[46] Gal, J., Bailleux, C., Chardin, D., Pourcher, T., Gilhodes, J., Jing, L., Guigonis, J.-M., Ferrero, J.-M., Milano, G., Mograbi, B., Brest, P., Chateau, Y., Humbert, O., & Chamorey, E. (2020). Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer.

Computational and Structural Biotechnology Journal, 18, 1509–1524. https://doi.org/10.1016/j.csbj.2020.05.021

[47] Alves, V. M., Korn, D., Pervitsky, V., Thieme, A., Capuzzi, S. J., Baker, N., Chirkova, R., Ekins, S., Muratov, E. N., Hickey, A., & Tropsha, A. (2022). Knowledge-based approaches to drug discovery for rare diseases. Drug Discovery Today, 27(2), 490–502. https://doi.org/10.1016/j.drudis.2021.10.014

[48] John L. Jefferies, Alison K. Spencer, Heather A. Lau, Matthew W. Nelson, Joseph D. Giuliano, Joseph W. Zabinski, Costas Boussios, Gary Curhan, Richard E. Gliklich, & David G. Warnock. (2021). A new approach to identifying patients with elevated risk for Fabry disease using a machine learning algorithm. Orphanet Journal of Rare Diseases, 16(1), 1–8. https://doi.org/10.1186/s13023-021-02150-3

[49] Sandra Brasil, Cátia José Neves, Tatiana Rijoff, Marta Falcão, Gonçalo Valadão, Paula A. Videira, & Vanessa dos Reis Ferreira. (2021). Artificial Intelligence in Epigenetic Studies: Shedding Light on Rare Diseases. Frontiers in Molecular Biosciences, 8. https://doi.org/10.3389/fmolb.2021.648012

[50] Meinshausen N, Meier L, Bu¨hlmann P. P-values for high-dimensional regression. Journal of the American Statistical Association. 2009; 104(488):1671–1681. https://doi.org/10.1198/jasa.2009. tm08647

[51] Bin Masud, S., Jenkins, C., Hussey, E., Elkin-Frankston, S., Mach, P., Dhummakupt, E., & Aeron, S. (2021). Utilizing machine learning with knockoff filtering to extract significant metabolites in Crohn's disease with a publicly available untargeted metabolomics dataset. PLoS ONE, 16(7), 1–13. https://doi.org/10.1371/journal.pone.0255240

[52] Vettore, M. V., Borges-Oliveira, A. C., Prado, H. V., Lamarca, G. de A., & Owens, J. (2020). Rare genetic diseases affecting skeletal development and oral health disparities among children and adolescents: a pathway analysis. International Dental Journal, 70(6), 469–476. https://doi.org/10.1111/idj.12583

[53] Maertens, A., Bouhifd, M., Zhao, L., Odwin-DaCosta, S., Kleensang, A., Yager, J. D., & Hartung, T. (2017). Metabolomic network analysis of estrogen-stimulated MCF-7 cells: a comparison of overrepresentation analysis, quantitative enrichment analysis and pathway analysis versus metabolite network analysis. Archives of Toxicology, 91(1), 217–230. https://doi-org.ez.urosario.edu.co/10.1007/s00204-016-1695-x

[54] Sebastian Canzler, & Jörg Hackermüller. (2020). multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. BMC Bioinformatics, 21(1), 1–13. https://doi-org.ez.urosario.edu.co/10.1186/s12859-020-03910-x

[55] Riquelme, G.; Zabalegui, N.; Marchi, P.; Jones, C.M.; Monge, M.E. A Python-Based Pipeline for Preprocessing LC–MS Data for Untargeted Metabolomics Workflows. Metabolites 2020, 10, 416, doi:10.3390/metabo10100416.

[56] https://github.com/bihealth/NeatMS.

[57] https://github.com/ccdmb/BioDendro

[58] Teo, G., Chew, W. S., Burla, B. J., Herr, D., Tai, E. S., Wenk, M. R., et al. (2020). MRMkit: Automated data processing for large-scale targeted metabolomics analysis. Analytical Chemistry. https:// doi. org/ 10. 1021/ acs. analc hem. 0c030 60

[59] Kutuzova, S., Colaianni, P., R¨ost, H., Sachsenberg, T., Alka, O., Kohlbacher, O., et al. (2020).SmartPeak automates targeted and quantitative metabolomics data processing. Analytical Chemistry, 92(24), 15968–15974. https:// doi. org/ 10. 1021/ acs. analc hem.https://github.com/AutoFlowResearch/SmartPeak

[60] Cirillo E, Parnell LD, Evelo CT. A Review of Pathway-Based Analysis Tools That Visualize Genetic Variants. Front Genet. 2017 Nov 7;8:174. doi: 10.3389/fgene.2017.00174. PMID: 29163640; PMCID: PMC5681904.
[61] Kotelnikova, E., Frahm, K.M., Lages, J. et al. Statistical properties of the MetaCore network of protein–protein interactions. Appl Netw Sci 7, 7 (2022). https://doi.org/10.1007/s41109-022-00444-4
[62] JKU Visual Data Science Lab. (s. f.). https://jku-vds-lab.at/
[63] IPA TM of QIAGEN's Ingenuity Pathway Analysis (2016). Calculating and Interpreting the p-values for Functions, Pathways and Lists in IPA. Available at: https://www.ingenuity.com/wp- content/themes/ingenuityqiagen/pdf/ipa/functions-pathways-pval-whitepaper.pdf.
[64] PathVisio biological pathway editor. (s. f.). pathvisio.github.io. https://pathvisio.org/
[65] 3Omics: A web based systems biology visualization tool for integrating human transcriptomic, proteomic and metabolomic data. (s. f.).https://3omics.cmdm.tw/
[66] Clemens Wrzodek, clemens.wrzodek@uni-tuebingen.de. (s. f.). InCroMAP. http://www.ra.cs.uni-tuebingen.de/software/InCroMAP/index.htm
[67] Cambiaghi A, Ferrario M, Masseroli M. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Briefings in Bioinformatics. 2017;18(3):498-510. doi:10.1093/bib/bbw031
[68] Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. Nucleic Acids Res. 2015;43(W1):W251–7.
[69] Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39(Database issue):D691–7.
[70] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000). [pubmed] [doi] Kanehisa, M; Toward understanding the origin and evolution of cellular organisms. Protein Sci. 28, 1947-1951 (2019) [pubmed] [doi] Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M.; KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 49, D545-D551 (2021). [pubmed] [doi]
[71] Caspi, R et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 36:D623-31.
[72] Wishart DS, Tzur D, Knox C, et al., HMDB: the Human Metabolome Database. Nucleic Acids Res.
[73] Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., Liu, P., Gautam, B., Ly, S., Guo, A. C., Xia, J., Liang, Y., Shrivastava, S., & Wishart, D. S. (2010). SMPDB: The Small Molecule Pathway Database. Nucleic acids research, 38(Database issue), D480–D487. https://doi.org/10.1093/nar/gkp1002
[74] Mubeen, S., Tom Kodamullil, A., Hofmann-Apitius, M., & Domingo-Fernández, D. (2022). On the influence of several factors on pathway enrichment analysis. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbac143
[75] Enrico Glaab, Anaıs Baudot, Natalio Krasnogor, and Alfonso Valencia. TopoGSA: network topological gene set analysis. Bioinformatics, 26(9):1271–1272, 2010.
[76] Nelson DL, Cox MM (2008). Lehninger principles of biochemistry (5th ed.). New York: W.H. Freeman. ISBN 978-0-7167-7108-1.
[77] Jelle J. Goeman, Sara A. van deGeer,Floor deKort, and Hans C. vanHouwelingen. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, 20(1):93–99, 2004.

[78] Khatri P., M. Sirota, and A. J. Butte, 2012 Ten years of pathway analysis: Current approaches and outstanding challenges. PLOS Computational Biology 8: e1002375. https://doi.org/10.1371/journal.pcbi.1002375
[79] Hänzelmann S., Castelo R. and Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics, 14:7, 2013.
[80] Armstrong, Scott A, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. 2002. “MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia.” Nature Gen. 30 (1): 41–47. https://doi.org/10.1038/ng765.

[81] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 Jan 8; 49(D1):D1388–D1395. doi:10.1093/nar/gkaa971.[PubMed PMID: 33151290] [PubMed Central PMCID: PMC7778930] [Free Full Text]
[82] Wishart DS, Tzur D, Knox C, et al., HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007 Jan;35(Database issue):D521-6. 17202168

[83] Metabolon, Inc. Metabolon – Small Molecules, Big Insights. Metabolon. https://www.metabolon.com/
[84] Dincel D, Olgan H, Canbaloğlu Z et al. Determinación de la adulteración de dihidrocapsaicina en suplementos dietéticos mediante LC-MS / MS . J. Chem. Metrol. 2020; 14 1: 77–82. doi : 10.25135 / jcm.36.20.01.1532
[85] Dried Blood Spot (DBS) Card Analysis on Metabolon’s Precision MetabolomicsTM Platform. (2021, 5 febrero). [Vídeo]. YouTube. https://www.youtube.com/watch?v=qL6ZPIKJutI
[86] Team, D. S. (2020, 23 noviembre). ¿Qué es un Z-Score? DATA SCIENCE. https://datascience.eu/es/matematica-y-estadistica/que-es-un-z-score/

[87] 09-KEGG_programming. (2018, 6 marzo). James Hutton Institute. https://widdowquinn.github.io/2018-03-06-ibioic/02-sequence_databases/09-KEGG_programming.html
[88] https://pubchempy.readthedocs.io/en/latest/index.html
[89] Wishart DS, et al., HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007 Jan;35(Database issue):D521-6
[90] Bioconductor - FAQ. (s. f.). Bioconductor. https://bioconductor.org/help/faq/
[91] Sharma, N. (2018, 25 noviembre). Exploratory Data Analysis (EDA) techniques for kaggle competition beginners – ConfusedCoders. https://confusedcoders.com/data-science/exploratory-data-analysis-eda-techniques-for-kaggle-competition-beginners
[92] Kumar, S. (2021, 15 diciembre). 7 Ways to Handle Missing Values in Machine Learning. Medium. https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e
[93] Sergio Decherchi, Elena Pedrini, Marina Mordenti, Andrea Cavalli, & Luca Sangiorgi. (2021). Opportunities and Challenges for Machine Learning in Rare Diseases. Frontiers in Medicine, 8. https://doi.org/10.3389/fmed.2021.747612

[94] sklearn.cluster.AgglomerativeClustering. (s. f.). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
[95] K-means Clustering. (s. f.). scikit-learn. https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html?highlight=k+mean
[96]Li, J. (2019). keggx: Manipulate KEGG networks as NetworkX objects in in Python. keggx. https://github.com/iamjli/keggx
[97] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013; 128(14).
[98] Zhuoqing Fang, Alex Wolf, Yuxing Liao, Austin McKay, Fabian Fröhlich, Jacob Kimmel, Liu Xiaohui, & sorrge. (2020). zqfang/GSEApy: gseapy-v0.10.1. https://doi.org/10.5281/zenodo.3983639
[99] Dash Documentation & User Guide | Plotly. (s. f.). https://dash.plotly.com/
[100] Spyder IDE. © 2022 Spyder Website Contributors. https://www.spyder-ide.org/
[101] Platform as a Service | Heroku. (s. f.). https://www.heroku.com/platform
[102] Lindsay C. Burrage, Lillian Thistlethwaite, Bridget M. Stroup, Qin Sun, Marcus J. Miller, Sandesh C.S. Nagamani, William Craigen, Fernando Scaglia, V. Reid Sutton, Brett Graham, Adam D. Kennedy, Aleksandar Milosavljevic, Brendan H. Lee, Sarah H. Elsea, Untargeted metabolomic profiling reveals multiple pathway perturbations and new clinical biomarkers in urea cycle disorders, (2019), https://doi.org/10.1038/s41436-019-0442-0.
[103] Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, Mungall C, Courtot M, Ruttenberg A, He Y. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query, and integration. Nucleic Acid Research. 2017 Jan 4;45(D1):D347-D352. PMID: 27733503. PMCID: PMC5210626.
[104] Morton, K. (1), Wang, P. (1), Bizon, C. (2), Cox, S. (2), Balhoff, J. (2), Kebede, Y. (2), Fecho, K. (2), & Tropsha, A. (3). (n.d.). ROBOKOP: An abstraction layer and user interface for knowledge graphs to support question answering. Bioinformatics, 35(24), 5382–5384. https://doi.org/10.1093/bioinformatics/btz604