Irene Castellanos González

ANÁLISIS DE DATOS ÓMICOS PEC 1

Tabla de contenidos

1. Resumen ejecutivo

Este informe presenta un análisis exploratorio del dataset '2023-CIMCBTutorial', centrado en las concentraciones de metabolitos en diferentes tipos de muestras relacionadas con cáncer gástrico. Los objetivos principales incluyen la creación de un objeto llamado SummarizedExperiment para organizar los datos, la exploración de patrones en la distribución de los metabolitos y la identificación de datos faltantes. Utilizando herramientas de R, como readxl y SummarizedExperiment, se han realizado visualizaciones como boxplots para explorar las relaciones entre metabolitos y las diferentes clases de muestras, detectando variaciones en la distribución de algunos metabolitos y el porcentaje de valores faltantes. El análisis muestra cómo ciertos metabolitos presentan diferencias significativas entre clases, lo que subraya la necesidad de estudios adicionales para determinar su relevancia biológica.

2. Objetivo del estudio

El principal objetivo del estudio es llevar a cabo un análisis exploratorio del dataset '2023-CIMCBTutorial', que contiene información sobre concentraciones de metabolitos en muestras de pacientes con cáncer gástrico y controles. El análisis incluye:

1. La creación de un objeto SummarizedExperiment para organizar y manejar los datos de metabolitos.
2. La exploración de patrones en la distribución de los metabolitos según las diferentes clases de muestras.
3. La identificación y análisis de datos faltantes en las variables de interés.
4. La comparación de las concentraciones de metabolitos entre las diferentes clases de muestra, utilizando gráficos y tablas de frecuencia para visualizar tendencias clave.

Este estudio busca proporcionar una visión general de los datos y servir como base para futuros análisis más detallados en el contexto del cáncer gástrico.

3. Materiales y métodos

- **Elección del dataset y descarga**

  El dataset utilizado en este estudio, '2023-CIMCBTutorial', fue descargado del repositorio de GitHub nutrimetabolomics/metaboData. Este dataset contiene datos obtenidos mediante resonancia magnética nuclear (NMR) sobre las concentraciones de metabolitos en diferentes muestras de pacientes con cáncer gástrico y controles sanos. La hoja de datos correspondiente para el estudio es 'GastricCancer_NMR.xlsx'.

- **Preparación de los datos**

  Se utilizaron dos hojas del archivo Excel; 'Data', que contiene tanto las identificaciones de las muestras (SampleID), el tipo de muestra (SampleType) y la clase (Class), como las mediciones de los metabolitos (columnas M1 a M149); y 'Peak', que proporciona los nombres de los metabolitos. Sin embargo, el análisis se centró en la hoja 'Data', ya que incluye la información relevante para los análisis posteriores.

  Al cargar los datos en R, se identificó que algunos valores numéricos en las columnas de metabolitos estaban formateados con comas (","), lo que impedía su correcta lectura. Para solucionar esto, se reemplazaron las comas por puntos (".") usando la función gsub(), y se utilizó as.numeric() para asegurar que los valores fueran interpretados como números decimales.

A continuación, se extrajeron las columnas de metabolitos y se organizó una matriz numérica con las muestras como filas y los metabolitos como columnas. Los nombres de las filas de esta matriz se asignaron a partir de la columna SampleID.

- **Ajustes en la orientación de la matriz**
  Durante el análisis, se encontró un error relacionado con la orientación de la matriz de datos, específicamente con los nombres de las filas. Para evitar este problema, se transpuso la matriz, de modo que las filas correspondieran a los metabolitos y las columnas a las muestras, lo que permitió su correcto manejo en análisis posteriores.

- **Creación del contenedor 'SummarizedExperiment'**
  Se generó un objeto colData que contiene información relevante sobre las muestras (SampleID, SampleType y Class). Con estos datos organizados, se construyó un objeto SummarizedExperiment, un contenedor especializado para almacenar datos biológicos en R, que incluyó:

  - Una matriz de metabolitos (metabolite_data_t).
  - Un conjunto de metadatos de las muestras (colData).

  Finalmente, se imprimió el objeto SummarizedExperiment para verificar su correcta creación, confirmando la estructura y el contenido del mismo, incluyendo el número de metabolitos, las muestras y los metadatos asociados.

4. Resultados

- **Exploración de los datos**

Para una exploración del dataset, he utilizado varias herramientas de R para tener una visión general de los datos:

- **Resumen de los datos de metabolitos**: La primera fase consistió en obtener un resumen estadístico de los valores de concentración de los metabolitos en el dataset, para ello se usó la función summary(). Este resumen incluyó medidas como la media, el rango y los valores mínimos y máximos para cada metabolito. Los resultados muestran una amplia variabilidad en las concentraciones de los metabolitos, como puede verse en el siguiente resumen:

En esta primera parte del análisis, se realizó un resumen descriptivo de los metabolitos para entender mejor la distribución y variabilidad de las concentraciones de metabolitos en las muestras. A continuación, se muestra un resumen estadístico de los 9 primeros metabolitos con los valores mínimos, máximos, medianas, medias y cuartiles:

  - M1: Las concentraciones de este metabolito varían entre un mínimo de 0.40 y un máximo de 909.90. La media es de 101.07 y la mediana de 60.35, lo que sugiere una distribución sesgada hacia valores más bajos, ya que la media es considerablemente mayor que la mediana.
  - M2: Este metabolito muestra una mayor variabilidad, con un mínimo de 3.1 y un máximo de 26195.8. La media de 642.0 también es significativamente mayor que la mediana de 270.2, lo que indica la presencia de valores atípicos altos que están influyendo en la media.
  - M3: Las concentraciones de M3 oscilan entre 0.1 y 862.5, con una media de 146.4 y una mediana de 105.1. Este metabolito presenta una distribución más simétrica en comparación con M1 y M2, aunque también se observan valores altos que elevan la media.

- M4: El rango de concentración va de 0.10 a 242.50, con una media de 43.83 y una mediana de 35.70, sugiriendo que la mayoría de las muestras se concentran alrededor de la mediana.

- **Estructura y contenido de los metadatos**: a través del comando head() exploré los metadatos asociados con las muestras para ver las primeras filas, y str() para observar la estructura del objeto col_data. Esto permite confirmar que las columnas clave (SampleID, SampleType, y Class) están correctamente formateadas y contienen los tipos de datos esperados.

  Las primeras filas del dataframe muestran que hay 140 muestras en total, de las cuales algunas pertenecen al tipo QC (control de calidad) y otras son Sample (muestras de pacientes o grupos de estudio) y que las clases están distribuidas en cuatro categorías: BN, GC, HE, y QC.

  En cuanto a la estructura del dataframe muestra que cada columna contiene datos del tipo character (cadena de caracteres), lo que es adecuado para las categorías de muestras y que no hay metadatos adicionales asociados a este objeto, lo que sugiere que los datos principales están bien estructurados.

- **Análisis de la unicidad en las columnas de metadatos**. Mediante la función sapply(), se cuenta el número de valores únicos presentes en cada columna de los metadatos. Este análisis revela que había 140 muestras únicas en la columna SampleID, 2 tipos distintos de muestras en SampleType (QC y Sample) y 4 clases diferentes en Class (BN, GC, HE, y QC).

- **Frecuencia de los tipos y clases de muestras**. Se realizaron tablas de frecuencia para las variables categóricas presentes en los metadatos del estudio, **SampleType** y **Class**, con el objetivo de entender mejor la distribución de las muestras.

La columna **SampleType** clasifica las muestras como "Sample" (muestras biológicas) o "QC" (controles de calidad). La tabla de frecuencias muestra que se incluyen un total de **123 muestras** y **17 controles de calidad.** Esta distribución refleja una proporción adecuada de muestras biológicas con respecto a los controles de calidad, lo que es fundamental para asegurar la validez de los análisis.

La variable **Class**, esta tabla muestra que hay 43 muestras correspondientes a cáncer gástrico, 40 muestras benignas y 40 sanas, lo que garantiza una representación balanceada entre las clases de interés para el análisis. Además, se incluyen 17 muestras de control de calidad (QC).

- **Visualizar la distribución de un metabolito en concreto.** Se realizó un gráfico tipo boxplot para visualizar la distribución del metabolito M1 en función de las clases (Class). Este gráfico me permite identificar posibles diferencias entre las clases en relación con los niveles del metabolito M1 y proporciona información sobre la dispersión de los datos dentro de cada clase.

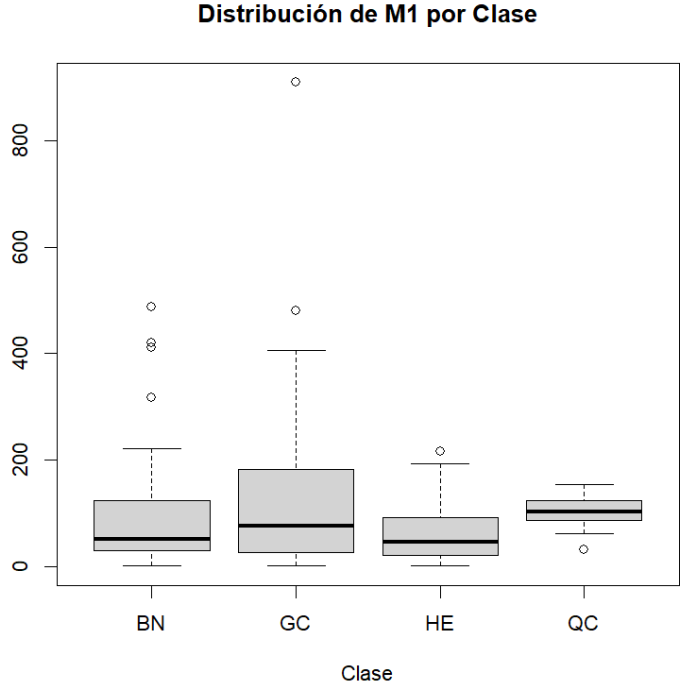

Figura 1. Gráfico boxplot que muestra la distribución del metabolito 1 en función de las clases.

El gráfico de boxplot muestra que el metabolito M1 tiene una mayor concentración en la clase GC en comparación con las otras clases. En BN, HE, y QC, la mediana es más baja y la dispersión es menor, lo que indica que las concentraciones de M1 son más homogéneas en esas clases. Se observan algunos outliers en varias clases, particularmente en GC, lo que podría ser indicativo de diferencias biológicas o técnicas en esas muestras.

- **Análisis de valores faltantes en el conjunto de datos.** Para ello, se verificó la calidad de los datos, se calculó el porcentaje de datos

  faltantes por cada metabolito. Este análisis nos muestra si había una cantidad significativa de datos faltantes. El análisis de los valores faltantes es crucial para evaluar la calidad del dataset antes de realizar análisis posteriores. A continuación, algunos ejemplos representativos de los valores faltantes para ciertos metabolitos:

  Metabolitos con alto porcentaje de valores faltantes:

  El metabolito **M9** presenta el mayor porcentaje de datos faltantes con **19.29%**. Este es un valor elevado y podría influir significativamente en los resultados si no se maneja adecuadamente.
  El metabolito **M1** también muestra un porcentaje de valores faltantes relativamente alto con **11.43%**, lo cual puede sugerir problemas en la detección o cuantificación de este metabolito.

  Metabolitos con pocos o ningún valor faltante:

  En el extremo opuesto, algunos metabolitos, como **M2** y **M8**, tienen porcentajes bajos de valores faltantes (**0.71%** y **0.00%**, respectivamente), lo que indica que los datos para estos metabolitos están casi completos y son fiables para el análisis.

5. Discusión, limitaciones y conclusiones

Discusión

Este análisis sobre los metabolitos de pacientes con cáncer gástrico y controles sanos mediante espectroscopía de resonancia magnética nuclear (NMR) ha revelado importantes patrones que podrían ser indicativos de diferencias metabólicas significativas entre los grupos de muestra. El enfoque empleado para la limpieza y organización de los datos, incluyendo la identificación de valores faltantes y la transposición de matrices, ha permitido realizar un análisis más robusto y confiable de los datos disponibles. Sin embargo, los resultados obtenidos muestran que algunos metabolitos presentan altos porcentajes de valores faltantes, lo que podría influir en la robustez de los análisis posteriores, en especial cuando se consideran metabolitos clave como el M9.

Además, la distribución observada en el análisis de los primeros nueve metabolitos destaca una amplia variabilidad entre las muestras, lo que sugiere diferencias individuales significativas. Esta variabilidad debe ser considerada cuidadosamente al interpretar los resultados, especialmente en estudios con un tamaño de muestra limitado. La alta concentración de ciertos metabolitos en muestras específicas podría estar relacionada con características individuales o clínicas, lo que abriría nuevas áreas de investigación.

Limitaciones

Una de las principales limitaciones de este estudio es la presencia de valores faltantes en algunos metabolitos, como el caso de M9, que presentó hasta un 19.29% de datos faltantes. El tratamiento de estos valores, ya sea mediante imputación o exclusión, podría afectar la interpretación de los resultados. Otra limitación importante es el número relativamente pequeño de muestras disponibles, lo que podría limitar la generalización de los hallazgos a una población más amplia.

Asimismo, la naturaleza transversal del dataset limita la capacidad de establecer relaciones causales entre los niveles de metabolitos y el estado clínico de los pacientes. En futuros estudios, sería útil contar

con un diseño longitudinal que permita realizar un seguimiento de las alteraciones metabólicas a lo largo del tiempo y en diferentes etapas del desarrollo del cáncer.

Conclusiones

Este estudio preliminar ha proporcionado una base sólida para la comprensión de las diferencias metabólicas entre pacientes con cáncer gástrico y controles sanos. A pesar de las limitaciones mencionadas, los resultados destacan la importancia del análisis de datos metabólicos en la identificación de posibles biomarcadores para el diagnóstico y pronóstico del cáncer gástrico. En futuras investigaciones, se recomienda abordar las limitaciones relacionadas con los valores faltantes y considerar un tamaño de muestra más amplio para mejorar la precisión y generalización de los resultados.

El uso de enfoques computacionales avanzados, como el procesamiento de datos en R, ha sido clave para este análisis y demuestra el potencial de las herramientas bioinformáticas en estudios de metabolómica. Sin embargo, es crucial continuar optimizando los métodos de análisis y aumentar la calidad de los datos para obtener conclusiones más confiables y aplicables en un contexto clínico.

6. Repositorio de los datos en GitHub

Antes de nada, me he creado una cuenta en GitHub ya que no tenía. Una vez creada desde la página oficial he creado el repositorio y he incluido todos los entregables que menciona la PEC1. La URL del repositorio es la siguiente: https://github.com/irenecastellanos/Castellanos-Gonzalez-Irene-PEC1.git