Informe PEC1

Fernando Santiago Hontoria

2024-11-06

Contents

Introducción

La PEC 1 de la asignatura de Análisis de Datos Ómicos pretende que se planifique y ejecute una versión simplificada del proceso de análisis de datos ómicos. Estos datos se han podido escoger de entre una gran variedad, empleándose finalmente un dataset que contiene mediciones de distintos metabolitos en pacientes con caquexia e indiivudos control.

La caquexia es un síndrome complejo caracterizado por una pérdida extrema de peso corporal, masa muscular y tejido adiposo, asociado comúnmente con enfermedades crónicas graves, como el cáncer, la insuficiencia cardíaca, la enfermedad pulmonar obstructiva crónica (EPOC) y la insuficiencia renal. A diferencia de la pérdida de peso común, la caquexia implica una desnutrición profunda y se relaciona con alteraciones metabólicas, inflamación y un desequilibrio energético que el cuerpo no puede compensar, lo que lleva a una pérdida de masa muscular y grasa, teniendo un gran impacto en la calidad de vida y supervivencia de los individuos afectados.

El estudio de metabolitos en pacientes con caquexia y en grupos de control es esencial porque los metabolitos reflejan cambios bioquímicos en tiempo real que ocurren debido a esta condición. Al analizar el perfil metabólico, se pueden identificar alteraciones específicas, proporcionando una visión profunda sobre cómo la

caquexia afecta el metabolismo celular y sistémico. Los metabolitos específicos pueden servir como biomarcadores de diagnóstico para identificar pacientes en etapas tempranas de caquexia o para predecir la progresión de la enfermedad, lo que finalmente puede contribuir a mejorar la calidad de vida y aumentar la supervivencia de los pacientes.

Objetivos

Sabiendo la relevancia y motivación biológica detrás de este estudio, se tiene por objetivo de este trabajo explorar el archivo de datos sobre los metabolitos con pacientes con y sin caquexia, y comprobar si existen diferencias o si se pueden obtener conclusiones del mismo. Siguiendo esta línea, se pueden plantear las siguientes cuestiones:

1. ¿Son los datos recogidos comprensibles, y se tiene suficiente información para obtener conclsuiones?
2. ¿Existen diferencias significativas entre los grupos del ensayo?
3. ¿Podrían obtenerse marcadores predictivos a partir de las conclusiones del estudio de los datos?

Materiales y métodos

Materiales

Los datos para este trabajo se han encontrado a través del repositorio "metaboData" de Álex Sanchez Pla (https://github.com/nutrimetabolomics/metaboData/). Dentro de la sección de Datasets se encuentra una carpeta que contiene el archivo csv y la información del mismo, donde se encuentra el link de descarga y el registro de un sanity check sobre el archivo.

Este dataset de cachexia (caquexia en inglés) contiene dos grupos entre sus muestras, los pacientes y los contros, sus datos no están pareados y no se detectaron datos que falten. Se han tomado datos de 63 metabolitos distintos, además de incluirse el ID de cada paciente empleado en el ensayo.

Procedimiento

Descarga de datos

La descarga de datos se realizó directamente a partir del link de descarga del archivo (https://rest.xialab.ca/api/download/metaboanalyst/human_cachexia.csv). A partir de este, se señaol que era un archivo .csv y se guardó como dataframe directamente en una variable denominada "datos".

Creación del contenedor

En primer lugar deben analizarse los datos de los que se parten, modificar las categorias necesarias para homogeneizar el formato y entender cómo está contruido el dataset. Luego, se extraen las filas y columnas que contengan los nombres clave, en este caso los ids de los pacientes y los tipos de metabolito, que se asignarán a los nombres de las filas y las columnas respectivamente. Los datos numéricos que contienen las medidas de distintos metabolitos en los pacientes, además de la columna de los grupos, se agrupan en una matriz, formato que permitirá crear el contenedor.

Una vez que los datos están preparados, se integran en un contenedor único, el **SummarizedExperiment**, que estructura toda la información de forma compacta y organizada. Este contenedor incluye:

- Las mediciones de los metabolitos (la matriz de datos)
- Los identificadores de los pacientes como metadatos de filas
- Los nombres de los metabolitos como metadatos de columnas

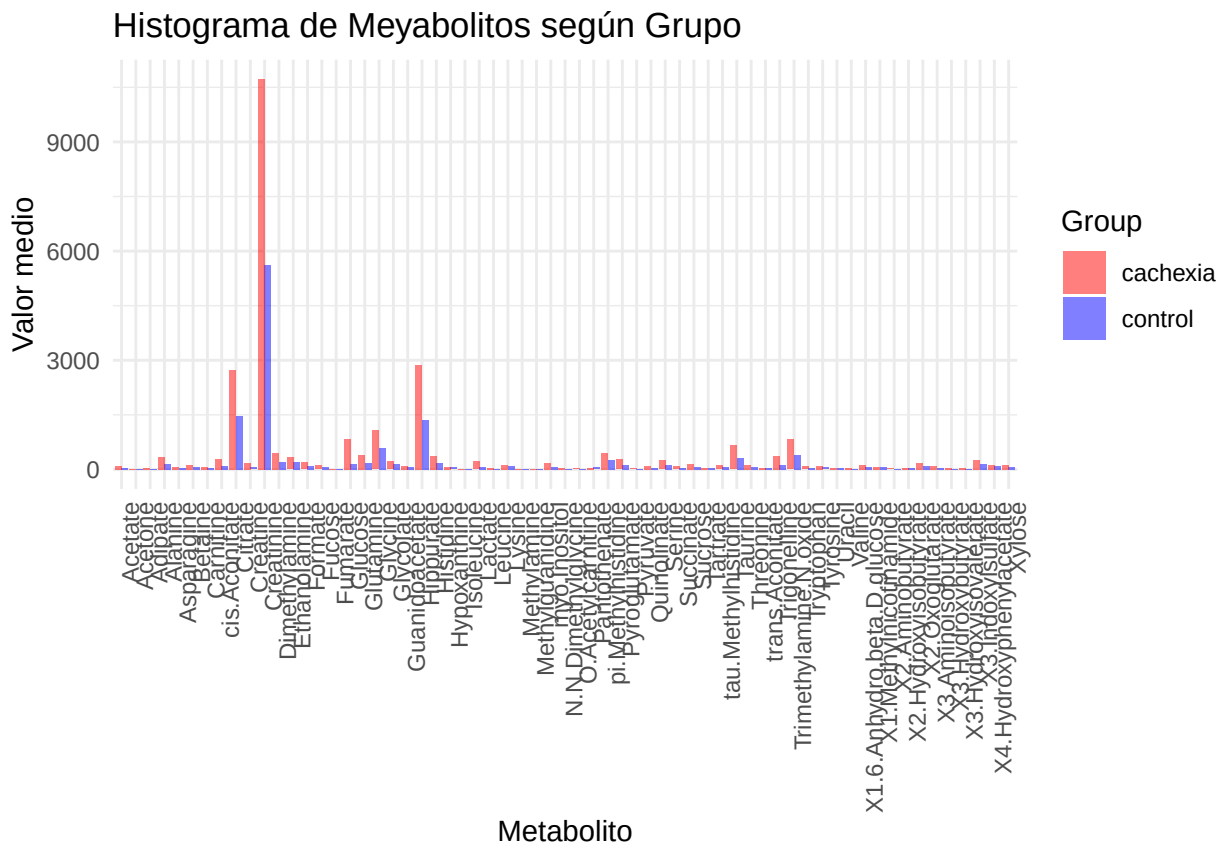
Para proporcionar un contexto completo sobre el dataset, se añaden detalles adicionales como una descripción general del estudio, la fuente de los datos, la fecha de creación, el autor, y notas específicas sobre las filas y columnas del objeto. Estos metadatos ayudan a comprender el contenido del dataset, su estructura y su uso potencial.

```
## class: SummarizedExperiment
## dim: 77 64
## metadata(8): descripcion fuente ... notas Tipos_metabolitos
## assays(1): counts
## rownames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## rowData names(1): ID_Paciente
## colnames(64): Muscle.loss X1.6.Anhydro.beta.D.glucose ...
##   pi.Methylhistidine tau.Methylhistidine
## colData names(1): Tipos_metabolitos
```

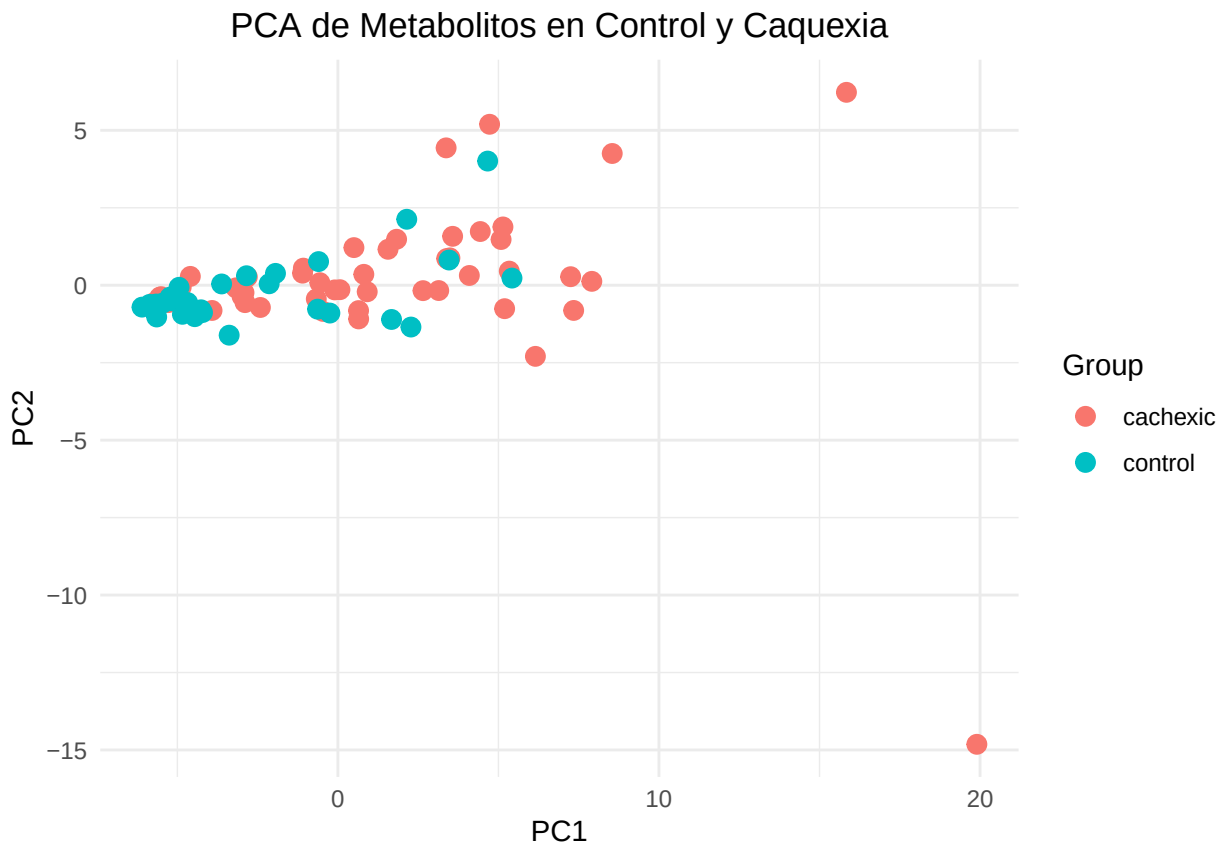
Exploración de datos

La exploración de datos del dataset escogido se ha dividido en tres apartados. Primero, la creación de un histograma, que ha permitido una comparativa visual de las diferencias en los metabolitos en cada grupo. También para el análisis estadístico-visual, se ha realizado un análisis de componentes principales, para que se vea cómo se agrupan los datos en los dos grupos. Ya por último, se ha realziado un análisis estadístico para determinar si existen diferencias significativas en los niveles de los metabolitos entre el grupo control y el grupo de pacientes, lo que daría una idea de que metabolitos varían más, y su medición pueda emplearse en futuras pruebas.

Creación de Histograma Para la creación del histograma, primero se realizaron las medias del grupo de las muestras control (30 muestras) y del de las muestras de pacientes con caquexia (47 muestras). Así, se obtiene un único valor para cada metabolito en cada grupo, lo que permite una comparación visual. Esta comparación se realzia mediante histograma, que permite un solapamiento de las columnas que ayuda a distinguir posibles diferencias.



Análisis de componentes principales El **análisis de componentes principales (ACP)** es una técnica estadística que permite visualizar y explorar patrones en los datos, lo que facilita la interpretación de datos complejos. Para realziar el análisis, se seleccionan los datos numéricos de los metabolitos y se normalizan, lo que permite que cada variable tenga la misma importancia en el análisis, eliminando el efecto de diferentes escalas o unidades entre variables. Entonces, se calculan los componentes principales, que representan las direcciones en las que los datos tienen mayor variabilidad. Los resultados del ACP se almacenan en un nuevo conjunto de datos, donde cada fila corresponde a una muestra y cada columna a un componente principal. Estos resultados se exponene mediante un gráfico bidemensional, siendo los ejes las dos muestras que más variabilidad presentan. Si los grupos de estudio (control y caquexia) se separan claramente en el gráfico, esto sugiere que hay diferencias significativas en los perfiles metabólicos entre ellos, lo cual puede indicar biomarcadores o cambios metabólicos específicos asociados con la caquexia.



Estudio de diferencias significativas Finalmente, tras los análisis visuales, se realiza un análisis estadísitco **t de student** que compara las medias del metabolito entre ambos grupos. La prueba devuelve un valor p que indica si hay diferencias significativas en ese metabolito entre los grupos. Dado que se realizan múltiples pruebas (una para cada metabolito), se ajustan los valores p obtenidos para reducir el riesgo de detectar diferencias significativas por azar (error de Tipo I). Después del ajuste, se seleccionan solo los metabolitos cuyo valor p ajustado es menor a un umbral (usualmente 0.05), indicando diferencias estadísticamente significativas entre los grupos.

Así, se obtienen 45 metabolitos que presentan diferencias significativas entre los grupos caquexia y control. Estos metabolitos son X1.6.Anhydro.beta.D.glucose, X2.Aminobutyrate, X2.Hydroxyisobutyrate, X3.Hydroxybutyrate, X3.Hydroxyisovalerate, X3.Indoxylsulfate, Acetate, Adipate, Alanine, Asparagine, Betaine, Carnitine, Citrate, Creatine, Creatinine, Dimethylamine, Ethanolamine, Formate, Fucose, Fumarate, Glucose, Glutamine, Glycine, Hippurate, Histidine, Leucine, Methylamine, N.N.Dimethylglycine, O.Acetylcarnitine, Pyroglutamate, Pyruvate, Quinolinate, Serine, Succinate, Taurine, Threonine, Trigonelline, Trimethylamine.N.oxide, Tryptophan, Tyrosine, Valine, cis.Aconitate, myo.Inositol, trans.Aconitate, tau.Methylhistidine.

Reposición de datos en Github

Todos los estudios realizados sobre los datos y los resultados obtenidos se suben a un repositorio en Github. Este repositorio, denominado Santiago-Hontoria-Fernando-PEC1, se ha creado específicamente en la web de Github para este fin. Tras crearse, se suben todos los archivos requeridos, además de un archivo README.md donde se explican los contenidos del repositorio. Así, se puede hacer público los avances realizados y añadirse un contexto a todo el proceso para terceras personas que se encuentren con el repositorio.

Resultados

Al haberse obtenido 45 metabolitos con diferencias significativas entre grupos, debería realizarse futuros experimentos centrándose en estos metabolitos. Esto supondría un avance en el conocimiento sobre la caquexia, y acercarse a la obtención de biomarcadores que faciliten el diagnóstico y la pronosis de los pacientes.

Enlace al repositorio Github

A través del siguiente enlace, se puede visitar el repositorio donde se tienen todos los archivos empleados para el análisis y aquellos obtenidos a partir del mismo. En dicho repositorio se puede encontrar:

- el informe, bajo el nombre de "Informe PEC1.pdf"
- el objeto contenedor con los datos y los metadatos en formato binario (.Rda), bajo el nombre de "Caquexia_SExperiment.Rda"
- el código R para la exploración de los datos, bajo el nombre de "Exploración de datos PEC1.R"
- los datos en formato texto, bajo el nombre de "human_cachexia.csv"
- los metadatos acerca del dataset en un archivo markdown, bajo el nombre de "metadatos_dataset.md"
- el archivo R markdown empleado para la creación del informe y con el código empleado, bajo el nombre de "Informe PEC1.Rmd"

https://github.com/FernandoSantiagoHontoria/Santiago-Hontoria-Fernando-PEC1/tree/2b27194acfd1190740b92220b9a6159411586b54