

Informe PEC1

Inés Elena Fernández Benet

2024-11-06

Informe de Análisis de Datos de Metabolómica

Informe del proceso llevado para el desarrollo de la primera Prueba de Evaluación Continua (PEC1)

1. Tabla de contenidos

1. Tabla de Contenidos
2. Resumen Ejecutivo
3. Objetivos del estudio/la prueba
4. Materiales y Métodos
5. Resultados
 - Descarga de datos y configuración del entorno
 - Creación del contenedor SummarizedExperiment
 - Exploración de los datos
 - Análisis de Componentes Principales (PCA)
 - Análisis de Varianza (ANOVA)
6. Discusión, Limitaciones y Conclusiones
7. Referencias
8. Enlace al repositorio de PEC1

2. Resumen Ejecutivo

Este informe se ha desarrollado para dar respuesta a la Prueba de Evaluación Continua de la asignatura Análisis de Datos Ómicos del Máster de Bioinformática y Bioestadística de la Universitat Oberta de Catalunya.

En esta prueba, realizamos un estudio que analiza un conjunto de datos de metabolómica. Escogiendo una base de datos sobre caquexia humana, estudiamos los datos e identificamos patrones y diferencias entre las dos condiciones que encontramos en la muestra (cachexic y control). A partir de los datos, se creó un contenedor SummarizedExperiment para organizar los datos. Las técnicas utilizadas incluyeron estadísticas descriptivas, un Análisis de Componentes Principales (PCA) y un análisis ANOVA.

Los resultados revelan diferencias significativas en ciertos metabolitos y patrones de variación entre condiciones.

3. Objetivos del estudio

El objetivo principal de este estudio/prueba es explorar un conjunto de datos de metabolómica, encontrar sus datos y metadatos, y realizar un estudio analítico del mismo.

Específicamente, los objetivos de la prueba son: - Crear un contenedor estructurado para organizar los datos de metabolitos. - Realizar una exploración inicial de las características del dataset. - Realizar estudios estadísticos (ANOVA, PCA). - Reposicionar los datos y el código en un repositorio de GitHub.

4. Materiales y métodos

Los datos utilizados en este análisis fueron descargados del repositorio del profesor de la asignatura Alex Sanchez Pla, <https://github.com/nutrimetabolomics/metaboData>.

El dataset contiene información sobre 63 metabolitos en muestras clasificadas en dos grupos: cachexia y control. Cada muestra representa una observación independiente, y el dataset incluye un identificador único (Patient.ID) y una variable categórica (Muscle.loss) que indica la condición experimental.

Herramientas utilizadas: - R y Rstudio: Lenguaje de programación y software por el que programar los análisis - PAquetes de R: hemos utilizado BioConductor, POMA, ggplot2, entre otros más básicos típicos de programación en R - GitHub: repositorio donde he descargado el conjunto de datos de estudio, y donde he repositado todos mis documentos para evaluación

5. Resultados

5.1. Descarga de los datos y configuración del entorno

Tras leer las especificaciones del enunicado de la PEC1, accedí al repositorio del profesor donde recoge varias de los conjuntos de datos propuestos para hacer la PEC. Escogí el conjunto de datos de caquexia, debido a que es una condición que me resultó interesante y que aprendí sobre ella hace tan sólo unos meses.

El entorno de desarrollo del programa para realizar el código comenzó por conectar el entorno de RStudio con mi cuenta de GitHub. Una vez conectada, puedo realizar commit, push y pull directamente desde el software donde realizo la programación, y que se verá reflejado en la “nube”, en mi repositorio dentro de GitHub dedicada a esta PEC: <https://github.com/ifdzbenet/Fernandez-Benet-Ines-PEC1>.

5.2. Creación del contenedor SummarizedExperiment

Gracias a uno de los materiales proporcionados en la asignatura de Análisis de Datos Ómicos, conocí la existencia del paquete POMA, que hace el proceso de trasladar los datos de un dataset a un objeto SummarizedExperiment muy fácil, usando la función PomaSummarizedExperiment(). Esta función simplemente necesita un argument target, que contiene la información descriptiva de la muestra (patient.id y muscle.loss, este último la condición de tener caquexia o ser control); el otro argumento que necesita es features, que es la parte del assay, es decir, la parte de los metabolitos del conjunto de datos.

5.3. Exploración de los datos

Se calcularon estadísticas descriptivas básicas para cada columna del conjunto. Estos resultados incluyen la media, mediana, desviación estándar, mínimo y máximo.

De esta información, la más relevante resultó ser que el conjunto contiene 47 observaciones de sujetos con caquexia, y 30 de sujetos con condición de control.

5.4. Análisis de Componentes Principales (PCA)

El PCA reveló patrones de agrupamiento entre las muestras. Los dos primeros componentes principales explicaron un porcentaje considerable de la varianza total, indicando que los datos de metabolitos pueden diferenciar en cierta medida las condiciones cachexia y control.

El gráfico PCA muestra una separación parcial entre las dos condiciones. Hay una tendencia de agrupación diferente para ambos grupos, siendo control la más uniforme, con excepciones, mientras que la condición de caquexia está más dispersa entre ambas componentes.

5.5. Análisis de Varianza (ANOVA)

Para evaluar diferencias entre las condiciones, se realizó un análisis ANOVA en cada metabolito. Los resultados mostraron que 40 de los 63 metabolitos tenían diferencias significativas ($p < 0.05$) entre cachexia y control, lo cual sugiere una variación relevante entre grupos.

6. Discusión, Limitaciones y Conclusiones

Discusión

Los resultados de este estudio sugieren que existen diferencias en ciertos metabolitos entre las condiciones caquexia y control. Estos hallazgos podrían ser útiles para estudios adicionales que busquen identificar biomarcadores de la condición caquexia, en comparación con condiciones de control. El PCA fue útil para visualizar los patrones generales de variación, aunque la separación entre grupos no fue absoluta, lo que indica que se requerirían más estudios para identificar asociaciones fuertes.

Limitaciones

La limitación principal para este estudio fue el nivel de conocimiento personal para desarrollarlo. Mientras el nivel es suficiente para realizar el estudio sugerido por la PEC, muchas cosas quedan en el tintero, y algunos estudios estadísticos más ajustados a las necesidades del estudio son necesarios.

Conclusiones

Este estudio proporciona una exploración preliminar y básico de un conjunto de datos de metabolómica, identificando varios metabolitos con diferencias significativas entre las condiciones caquexia y control.

7. Referencias

- Sanchez, A., Carmona, F. (2024). Casos y Ejemplos de Análisis Multivariante con R. Análisis de Datos Ómicos. Universitat Oberta de Catalunya. Acceso a través del enlace: <https://aspteaching.github.io/AMVCasos/>
- Sanchez, A. (2024). Introduction to microarray data exploration and analysis with basic R functions. Análisis de Datos Ómicos. Universitat Oberta de Catalunya. Acceso a través del enlace: https://aspteaching.github.io/Analisis_de_datos_omicos-Ejemplo_0-Microarrays/ExploreArrays.html#2_A_first_look_at_microarray_data_The_study
- Sanchez, A. (2024). Exploración multivariante de datos ómicos: Descriptivo, PCA y Clustering. Análisis de Datos Ómicos. Universitat Oberta de Catalunya.

- Sanchez, A. (2024). MetaboData datasets [Repositorio en GitHub]. GitHub. <https://github.com/nutrimetabolomics/metaboData/blob/main/Datasets/2024-Cachexia>
- Bioconductor (2023). POMA Workflow. Análisis de Datos Ómicos. Acceso a través del enlace: <https://web.archive.org/web/20240415023513/http://bioconductor.org/packages/release/bioc/vignettes/POMA/inst/doc/POMA-demo.html>
- Morgan, M., Obenchain, V., Hester, J., Pagès, H. (2023). SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest. Acceso a través del enlace: <https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html>
- Morgan, M., Obenchain, V., Hester, J., Pagès, H. (2023). SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest. Acceso a través del enlace: <https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html>

8. Enlace al repositorio de PEC1

El enlace a mi repositorio donde encontrarás todos los documentos y archivos pedidos es: <https://github.com/ifdzbenet/Fernandez-Benet-Ines-PEC1>