PEC1

Eduard Rodríguez Pérez

2024-11-04

Contenidos

Introducción

Este informe tratará de analizar en poca profundidad un conjunto de datos de metabolómica de un estudio en humanos.

Se parte del *dataset* "cachexia" disponible en: https://rest.xialab.ca/api/download/metaboanalyst/human_cachexia.csv Dicho dataset proviene del estudio de Eisner et al. (2010) *Learning to predict cancer-associated skeletal muscle wasting from 1h-nmr profiles of urinary metabolites Metabolomics* 7:25-34

En el mencionado estudio se comparan las concentraciones de distintas molèculas presentes en la orina de 47 pacientes de cáncer de colon o pulmón, asi como de 30 individuos sanos, con el objetivo de predecir la cachexia o pérdida de masa muscular debida al cáncer. Con dichos datos, los investigadores efectúan un análisis de aquellos marcadores con mayor capacidad de predicción de pérdida muscular mediante la creación de modelos predictivos. Para llegar a estos modelos usan métodos estadísticos estándar así como herramientas de inteligencia artificial tipo *machine learning*.

Los datos de metabolómica cuantitativos se obtuvieron a partir de espectros de resonancia magnética nuclear de muestras de orina, usando métodos en paralelo para corroborar la extrapolación cuantitativa a partir de las concentraciones d creatinina i de varios aminoácidos.

Adicionalmente, se realizaron métodos estadísticos para la normalización de los datos teniendo en cuenta que las muestras no se habían recogido en condiciones semejantes i por tanto podía haber dilución de la orina en función de la ingesta de agua horas antes del muestreo. En ese sentido, se usó tanto una normalización respecto a la concentración de creatinina, respecto al área total de picos de cada muestra (asumiendo que el área bajo el espectro de RMN es una función linear de la concentración de los metabolitos detectables), así como una normalización respecto al cociente de probabilidad que estima el factor de dilución más probable a partir de las amplitudes del espectro a partir de un espectro de referencia.

Finalmente, se realizó una clasificación de los individuos bajo estudio en función de si presentaban cachexia o no, a través del análisis de imágenes de tomografia axial computerizada de la superfície de tejido muscular en el corte a la altura de la tercera vértebra lumbar. El porcentaje de diferencia en masa muscular se usó como la variable para dividir en sendos grupos.

Objetivos del Estudio

- Describir el conjunto de datos.
- Organizar los datos en un objeto `SummarizedExperiment` y/o `ExpressionSet` según convenga.
- Explorar el *dataset* para obtener una visión general relativa al contenido trabajado en la asignatura hasta el momento.

Entre los objetivos del trabajo se incluye la creación del repositorio de `github` RODRIGUEZ-Perez-Eduard-PEC1 que no forma parte del informe en sí pero que se referencia en el mismo, en el apartado de Resultados.

Materiales y métodos

Preprocesado de los datos

El *dataset* bajo estudio se ha procesado mediante distintos paquetes de lenguaje R usando el software RStudio. El objetivo es preparar los datos presentes en el dataset `human_chachexia.csv`, si bien estos ya parecen presentar un preprocesado medio (normalización) según el estudio, cosa que tratamos de confirmar más adelante en este informe. ### Implementación de Bioconductor: SummarizedExperiment y ExpressionSet Con tal de implementar las herramientas bioinformáticas presentadas en la asignatura, en este trabajo se hace uso de los objetos SummarizedExperiment y ExpressionSet del paquete Bioconductor para análisis de datos ómicos.

Si bien se aplica SummarizedExperiment a forma de ejercicio práctico y como requerimiento del presente informe, éste tipo de objeto suele implementarse en experimentos con secuencias genéticas, donde las filas representan genes, transcritos o exones, entre otros (Figura 1).

Mientras que nuestro *dataset* se corresponde más con un objeto del tipo `ExpressionSet` para experimentos tipo array, de expresión génica y datos ómicos de similar estructura. En este objeto, las filas se corresponden a los genes u otros carácteres a medir, como metabolitos en nuestro caso, mientras que las columnas se corresponden a las muestras con su respectivo fenotipo (ver Figura 2).

Para ambos casos, tendremos que trasponer la tabla inicial `human_cachexia.csv`antes de transformarla a sendos objetos.

What is a `SummarizedExperiment`?

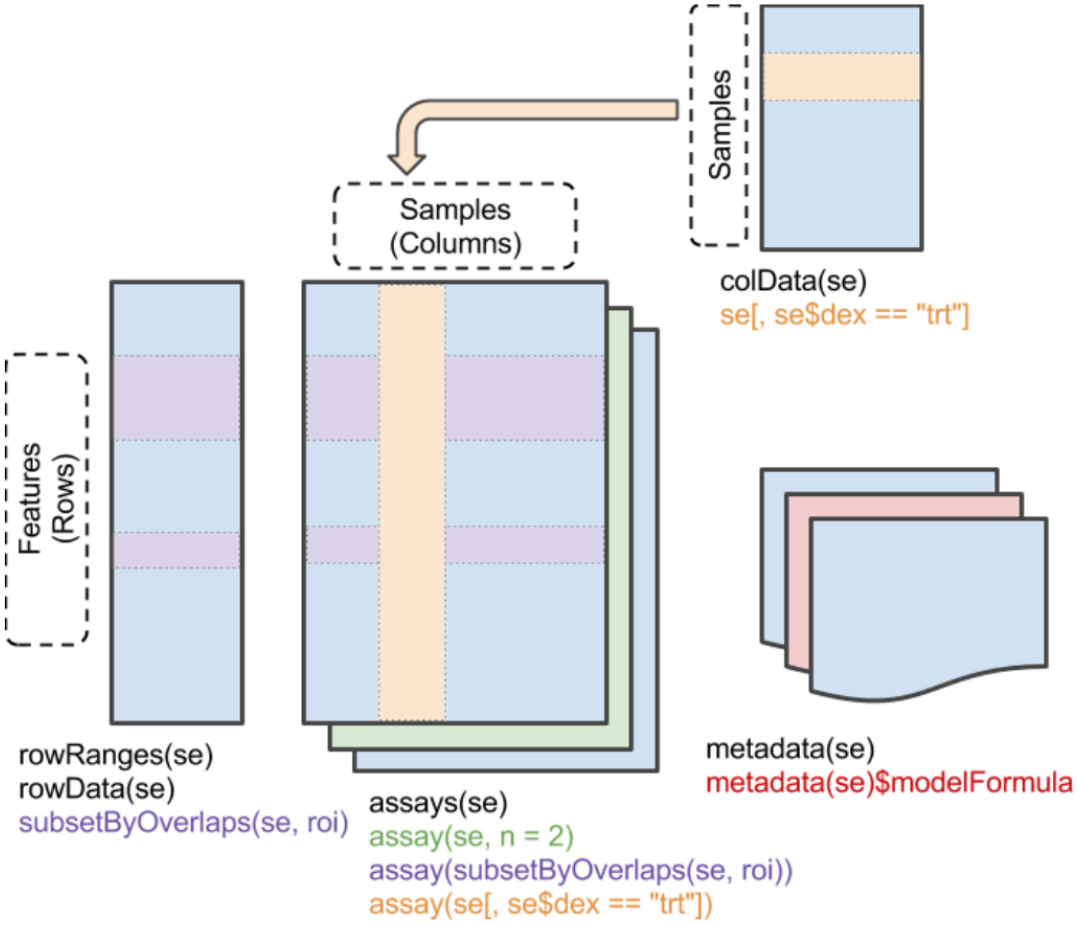


Figure 1: Fuente:*https://montilab.github.io/BS831/articles/docs/ExpressionSet.html#the-summarized experiment-object*

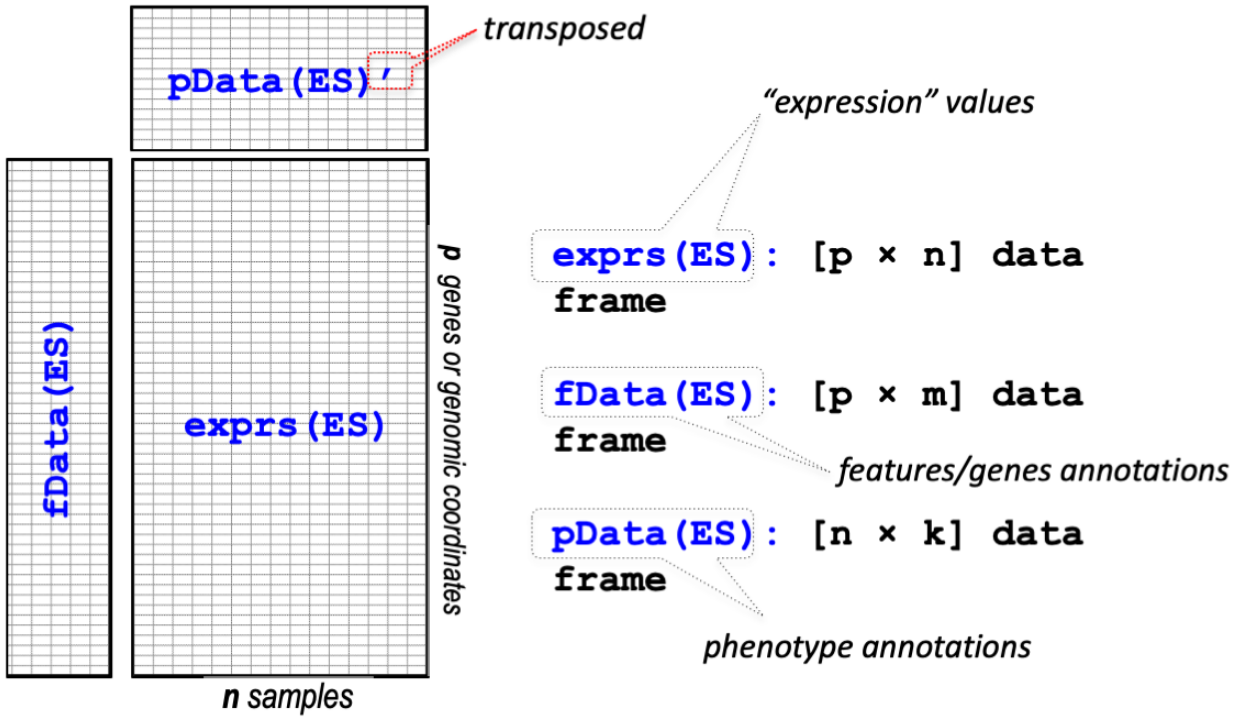


Figure 2: Fuente:*https://montilab.github.io/BS831/articles/docs/ExpressionSet.html#the-expressionset-object*

Exploración de los datos

Resultados

Preprocesado de los datos

Importamos los datos *cachexia* de nuestra carpeta de trabajo:

```
#Tratamos los datos para obtener una tabla traversa.
tabla_datos <- read.csv("human_cachexia.csv", row.names = 1)
tabla_datos$Muscle.loss <- as.factor(tabla_datos$Muscle.loss)
datos_tr <- t(tabla_datos)
```

Implementación de SummarizedExperimetn

Instalamos los paquetes necesarios para ejecturar `SummarizedExperiment`.

```
# Instalamos los paquetes necesarios
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("SummarizedExperiment")
## Warning: package(s) not installed when version(s) same as or greater than current; use
##   `force = TRUE` to re-install: 'SummarizedExperiment'



library(SummarizedExperiment)

datos_se <- SummarizedExperiment(assays = list(counts=datos_tr))
```

Vemos qué tal se muestra el objeto:

```
# Visualizamos el formato SummarizedExperiment resultante
datos_se
## class: SummarizedExperiment
## dim: 64 77
## metadata(0):
## assays(1): counts
## rownames(64): Muscle.loss X1.6.Anhydro.beta.D.glucose ...
##   pi.Methylhistidine tau.Methylhistidine
## rowData names(0):
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(0):
```

Implementación de ExpressionSet

```
BiocManager::install("Biobase")
## Bioconductor version 3.20 (BiocManager 1.30.25), R 4.4.1 (2024-06-14 ucrt)
## Warning: package(s) not installed when version(s) same as or greater than current; use
##   `force = TRUE` to re-install: 'Biobase'
## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.4.1/library
##   packages:
##     boot, foreign, MASS, Matrix, nlme, survival
## Old packages: 'curl'
datos_es<-ExpressionSet(datos_tr)
class(datos_es)
```

```
## [1] "ExpressionSet"
## attr(,"package")
## [1] "Biobase"
datos_es
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 64 features, 77 samples
##   element names: exprs
## protocolData: none
## phenoData: none
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation:
```

```
# Añadimos anotaciones de los datos a myAnnotDF
columnDesc <- data.frame(labelDescription=rownames(datos_tr))
datos_tr <- as.data.frame(datos_tr)
myAnnotDF <- new("AnnotatedDataFrame", data=tabla_datos, varMetadata= columnDesc)
show(myAnnotDF)
## An object of class 'AnnotatedDataFrame'
##   rowNames: PIF_178 PIF_087 ... NETL_003_V2 (77 total)
##   varLabels: Muscle.loss X1.6.Anhydro.beta.D.glucose ...
##     tau.Methylhistidine (64 total)
##   varMetadata: labelDescription
```

```
# Incluimos las anotaciones al ExpressionSet "datos_es"
phenoData(datos_es) <- myAnnotDF
show(datos_es)
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 64 features, 77 samples
##   element names: exprs
## protocolData: none
## phenoData
##   rowNames: PIF_178 PIF_087 ... NETL_003_V2 (77 total)
##   varLabels: Muscle.loss X1.6.Anhydro.beta.D.glucose ...
##     tau.Methylhistidine (64 total)
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation:
```

Exploración de los datos

Realizamos un breve análisis estadístico de los datos.

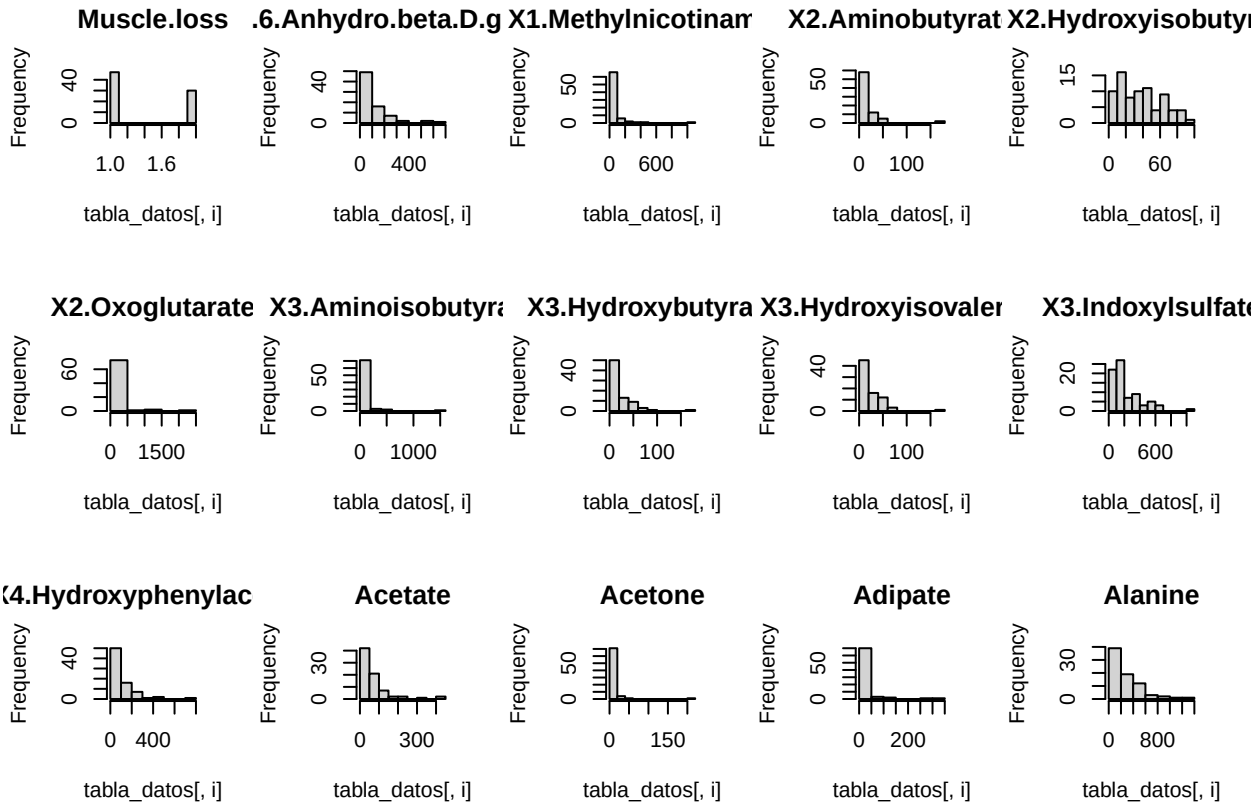
Tenemos 77 muestras, 47 de cachexia y 30 controles, así como 64 marcadores distintos en orina.

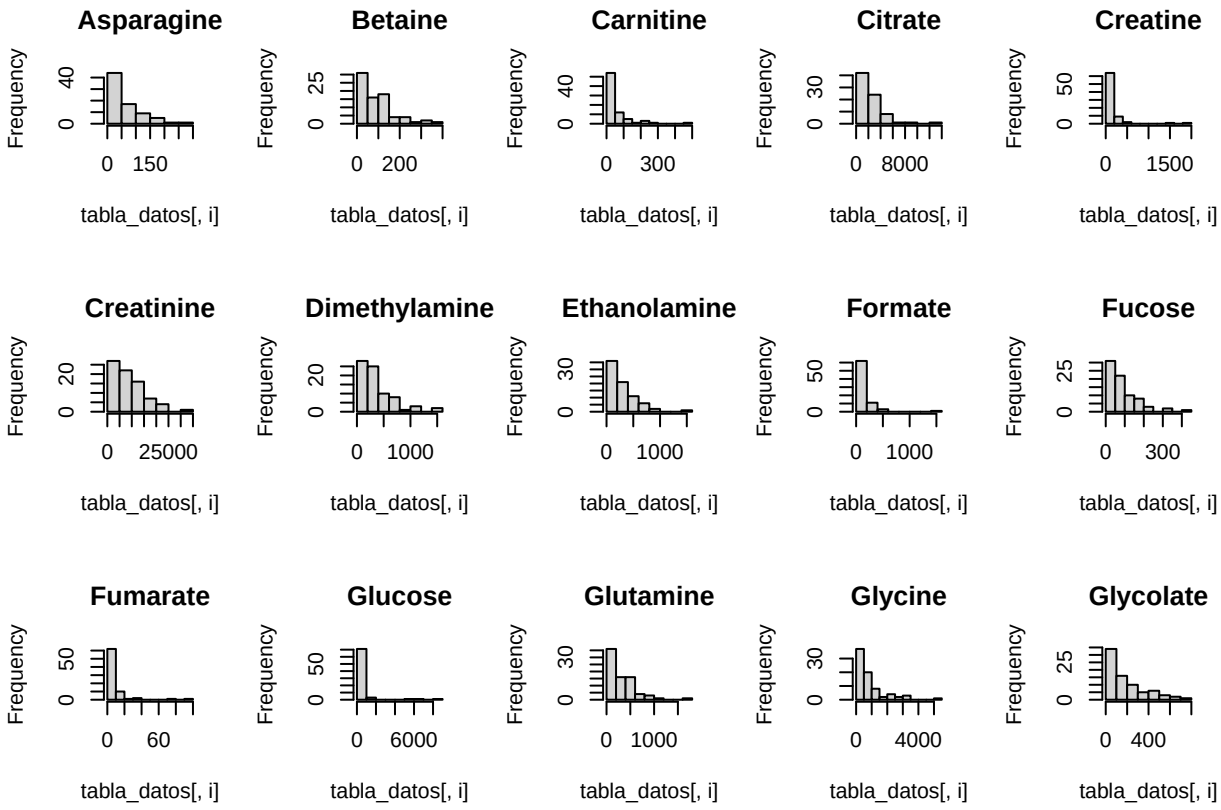
Los dos grupos (cachexia y control) han de esperar tener diferencias significativas en algunos de los marcadores. Sin embargo, es posible que existan relaciones entre los marcadores que provoquen que de forma individual no sean lo suficientemente predictivos.

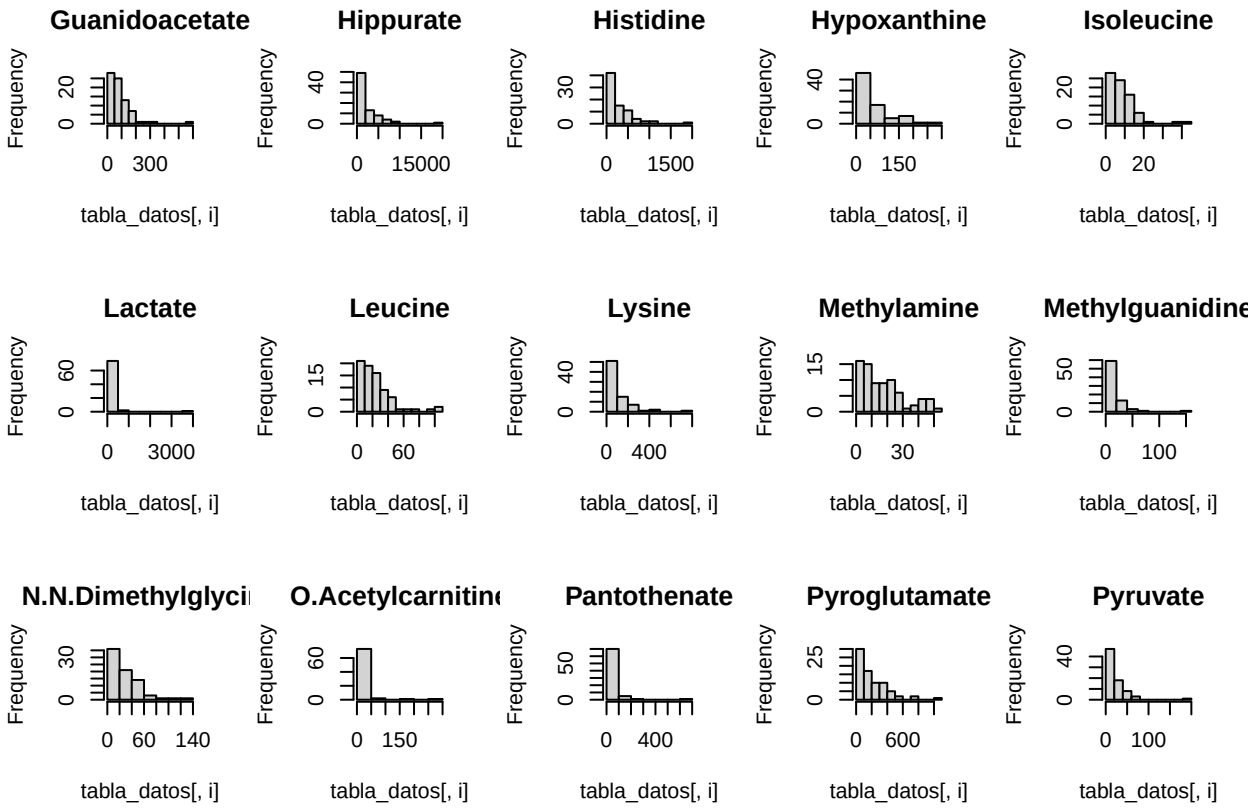
Realizamos un estudio de cada una de las variables, para ver de forma preliminar si hay una distribución en dos grupos (resultados al final del documento).

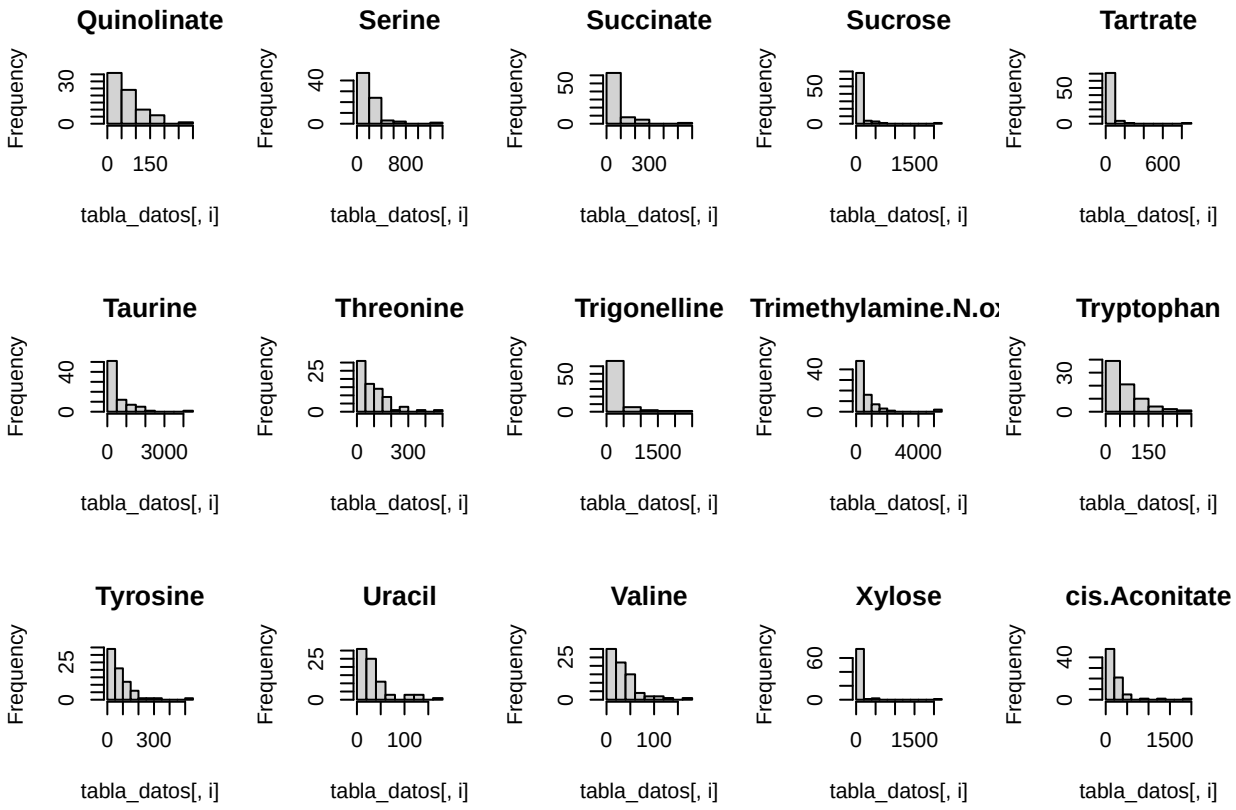
```
# Creamos histogramas individuales de cada factor
tabla_datos$Muscle.loss <- as.numeric(tabla_datos$Muscle.loss)
opt <- par(mfrow=c(3,5))
```

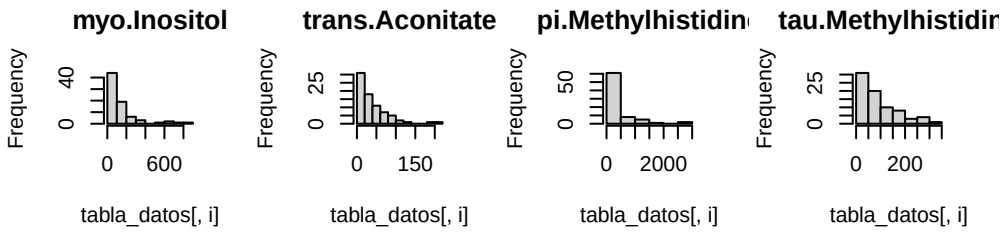
```r
for (i in 1:ncol(tabla_datos))
  hist(tabla_datos[,i], main = names(tabla_datos)[i])
par(opt)
```




Asparagine
Betaine
Carnitine
Citrate
Creatine
Creatinine
Dimethylamine
Ethanolamine
Formate
Fucose
Fumarate
Glucose
Glutamine
Glycine
Glycolate
Frequency
tabla_datos[, i]


Guanidoacetate
Hippurate
Histidine
Hypoxanthine
Isoleucine
Lactate
Leucine
Lysine
Methylamine
Methylguanidine
N.N.Dimethylglyci
O.Acetylcarnitine
Pantothenate
Pyroglutamate
Pyruvate
Frequency
tabla_datos[, i]


Quinolinate
Serine
Succinate
Sucrose
Tartrate
Taurine
Threonine
Trigonelline
Trimethylamine.N.o
Tryptophan
Tyrosine
Uracil
Valine
Xylose
cis.Aconitate
Frequency
tabla_datos[, i]



Creación del repositorio de github

```
# Creamos la versión en texto de los datos.
write.table(tabla_datos, "human_cachexia.txt",
            row.names = TRUE,
            quote = TRUE,
            col.names = TRUE)
```

- Enlace a github: *https://github.com/EduardRP/RODRIGUEZ-Perez-Eduard-PEC1.git*

Discusión y limitaciones y conclusiones del estudio

Este estudio ha resultado ser muy limitado debido a la falta de práctica con los objetos SummarizedExperiment y ExpressionSet. De haberle dedicado más tiempo, habría sido interesante realizar un estudio adecuado de las variables entre ambos grupos, si bien no he sabido analizarlos o bien he tenido problemas al intentarlo debido al excesivo número de variables.