Universitat Oberta de Catalunya

uoc.edu

# Estudio metabolómico mediante Resonancia Magnética Nuclear de biomarcadores en suero de pacientes con cáncer colorrectal

**Ana del Mar Salmerón López**
Máster interuniversitario UOC/UB en Bioinformática y Bioestadística
Trabajo Fin de Máster
Área 2, subárea 13: Resonancia Magnética Nuclear en metabolómica

**Consultores: Ignacio Fernández de las Nieves y Ana Cristina Ralha de Abreu**
**Profesor responsable de la asignatura: Marc Maceira Duch**

Fecha de entrega: 08/06/2021

**FICHA DEL TRABAJO FINAL**

| | |
|---|---|
| **Título del trabajo:** | *Estudio metabolómico mediante Resonancia Magnética Nuclear de biomarcadores en suero de pacientes con cáncer colorrectal* |
| **Nombre del autor:** | *Ana del Mar Salmerón López* |
| **Nombre del consultor/a:** | *Ignacio Fernández de las Nieves y*<br>*Ana Cristina Ralha de Abreu* |
| **Nombre del PRA:** | *Marc Maceira Duch* |
| **Fecha de entrega:** | 08/06/2021 |
| **Titulación:** | *Máster en bioinformática y bioestadística* |
| **Área del Trabajo Final:** | *Trabajo Fin de Máster* |
| **Idioma del trabajo:** | Castellano |
| **Número de créditos:** | 15 créditos ECTS |
| **Palabras clave** | *Cáncer colorrectal, Metabolómica, Machine Learning* |

**Resumen del Trabajo:**

El uso de la metabolómica como herramienta de detección temprana de cáncer colorrectal ha despertado un alto interés en el área clínica como alternativa diagnóstica a métodos tan invasivos como la colonoscopia. Este *Trabajo de Fin de Máster* (TFM) engloba (1) la realización de una breve revisión bibliográfica de algunas de las principales investigaciones dedicadas a este fin empleando *Resonancia Magnética Nuclear de Protón* (RMN de $^{1}$H) como técnica analítica, y (2) el uso de dicha herramienta combinada con métodos de análisis multivariante de datos de naturaleza supervisada y no supervisada en el análisis de 90 muestras de suero de pacientes con cáncer colorrectal para la obtención de posibles biomarcadores y su correlación con esta enfermedad. Se han verificado cambios en los perfiles metabólicos de un total de 12 metabolitos discriminantes entre pacientes enfermos y control a través de un modelo *Análisis Discriminante de Mínimos Cuadrados Parciales Ortogonal* (OPLS-DA) y 16 correspondientes a un modelo *Random Forest* (RF), posiblemente relacionados con disturbios en la glucólisis, el metabolismo del piruvato y de la alanina, del aspartato y del glutamato. Se concluye que la RMN acoplada a técnicas multivariantes es una poderosa herramienta de predicción y obtención de biomarcadores para esta enfermedad.

**Abstract:**

The use of metabolomics as a tool for detecting early colorectal cancer has developed great interest in the clinical field as a diagnostic alternative for invasive methods as colonoscopies. This dissertation contains (1) a bibliographic review where some of the main investigations about this topic applying *Proton Nuclear Magnetic Resonance* ($^1$H NMR) as the main analytical technique were analyzed, and (2) the use of this platform combined with multivariate unsupervised and supervised methods towards obtaining and correlating potential biomarkers of 90 samples of human colorectal serum. In addition, changes in the metabolomic profile of a total of 12 and 16 discriminant metabolites were reported from an *Orthogonal Partial Least Squares Discriminant Analysis* (OPLS-DA) model and a *Random Forest* (RF) representation, respectively, possibly related to glycolysis, and the metabolism of pyruvate, alanine, aspartate and glutamate. It was concluded that NMR in combination with multivariate techniques is such a powerful technique for biomarkers prediction and identification on this disease.

# Índice

**1. Resumen** ..... **2**
**2. Introducción** ..... **2**
2.1. Contexto y justificación del Trabajo ..... 2
2.1.1. Descripción general ..... 2
2.1.2. Justificación del TFM ..... 2
2.2. Objetivos del Trabajo ..... 4
2.2.1. Objetivos generales ..... 4
2.2.2. Objetivos específicos ..... 4
2.3. Enfoque y método seguido ..... 4
2.4. Planificación del Trabajo ..... 5
2.4.1. Tareas ..... 5
2.4.2. Calendario ..... 6
2.4.3. Hitos ..... 7
2.4.4. Análisis de riesgos ..... 8
2.5. Breve sumario de contribuciones y productos obtenidos ..... 8
2.6. Breve descripción de los otros capítulos de la memoria ..... 8
**3. Estado del arte** ..... **9**
3.1. Cáncer de colon ..... 9
3.1.1. La realidad del cáncer de colon ..... 9
3.1.2. Metabolómica para el diagnóstico de cáncer de colon ..... 10
3.2. Análisis de datos metabolómicos analizados mediante RMN de $^{1}H$ ..... 13
3.2.1. Procesado de los espectros ..... 13
3.2.2. Normalización ..... 13
3.2.3. Centrado y escalado ..... 14
3.2.4. Análisis estadístico ..... 15
3.2.5. Interpretación biológica ..... 17
3.3. Metabolómica en el estudio del cáncer de colon: Revisión bibliográfica ..... 17
**4. Metodología** ..... **22**
4.1. Revisión bibliográfica ..... 22
4.2. Selección del *dataset* ..... 23
4.3. Procesado de los datos ..... 23
4.4. Análisis estadístico ..... 24
4.5. Interpretación biológica ..... 26
**5. Resultados** ..... **27**
5.1. Aplicación de métodos lineales ..... 27
5.2. Aplicación de métodos supervisados no lineales ..... 32
5.3. Interpretación biológica ..... 39

**6. Discusión ..... 40**
**7. Conclusiones ..... 44**
7.1. Conclusiones de este trabajo ..... 44
7.2. Líneas de futuro ..... 44
7.3. Seguimiento de la planificación ..... 45
**8. Glosario ..... 45**
**9. Bibliografía ..... 47**
**Anexo A ..... 50**
**Anexo B ..... 63**

# Lista de figuras

**Figura 1.** (**a**) Espectrómetro *Bruker Avance* III 600 equipado con (**b**) automuestreador termostatizado *SampleCase* de 24 posiciones y (**c**) un manipulador de muestras automático *SampleJet*. Imágenes extraídas de la **referencia 14**……..11

**Figura 2.** Avance de las investigaciones centradas en metabolómica y RMN desde el año 2001 hasta la actualidad. Obtenida de *ISI Web of Knowledge* mediante la búsqueda de las palabras "*NMR*" y "*metabolomics*"…….……12

**Figura 3.** Diagrama de bloques según el área de investigación, obtenido de *ISI Web Of Knowledge* introduciendo las palabras clave "NMR" y "metabolomics"…….12

**Figura 4.** Proceso de análisis de un espectro de RMN hasta su transformación en *buckets*. Adaptado de la **referencia 19**……...……16

**Figura 5.** Avance de las investigaciones centradas en metabolómica y RMN desde el año 2001 hasta la actualidad. Obtenida de *ISI Web of Knowledge* mediante la búsqueda de las palabras "*nmr*" or "*nuclear magnetic resonance*", "*metabolomics*" or "*metabonomics*", or "*metabolite*" or "*metabolic*", y "*colorectal cancer*" or "*colon cancer*" or "*colorectal cancer*" or "*colon cancer*" or "*colorectum cancer*"…….……18

**Figura 6.** Diagrama de bloques según el área de investigación, obtenido de *ISI Web Of Knowledge* introduciendo las palabras clave "*nmr*" or "*nuclear magnetic resonance*", "*metabolomics*" or "*metabonomics*", or "*metabolite*" or "*metabolic*", y "*colorectal cancer*" or "*colon cancer*" or "*colorectal cancer*" or "*colon cancer*" or "*colorectum cancer*"……..……..18

**Figura 7**. Red de citaciones obtenida mediante el software *CitNetExplorer*……19

**Figura 8.** Mapa de calor obtenido mediante el software *VosViewer* de las palabras clave más empleadas en los resúmenes (*abstracts*) de las publicaciones……...20

**Figura 9.** Fragmento de la *Bucket table* obtenida para el análisis estadístico……..……..21

**Figura 10.** Gráfico PCA (PC1/PC2) de (**a**) *scores* y de (**b**) *loadings* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Unit Variance*). Es posible observar una discriminación evidente de las muestras en el gráfico de *scores* de acuerdo con la variable *Sucia/Limpia*……...……..27

**Figura 11.** Señales en espectros de RMN de muestras de tipo *Limpia*, en azul, y *Sucia*, en verde…….28

**Figura 12**. Gráfico PCA (PC1/PC2) de (**a**) *scores* y de (**b**) *loadings* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Unit Variance*). Las muestras quedan agrupadas en su gran mayoría en el centro de la gráfica, sin observarse una discriminación clara……..……28

**Figura 13.** Gráfico PLS-DA de *scores* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Unit Variance*). Puede observarse una discriminación entre las muestras de tipo *Cáncer* y *Control*. Los parámetros de calidad del modelo fueron $R^2X = 0.361$, $R^2Y = 0.918$, $Q^2Y = 0.713$, y RMSEE = 0.137……..……29

**Figura 14.** Gráfico OPLS-DA de *scores* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Pareto*). Puede observarse una discriminación entre las muestras de tipo *Cáncer* y *Control*. Los parámetros de calidad del modelo fueron $R^2X = 0.729$, $R^2Y = 0.809$, $Q^2Y = 0.687$, y RMSEE = 0.211……..30

**Figura 15.** Gráfica de contribuciones generada a partir del modelo OPLS-DA. En color verde quedan señalados los *buckets* más relevantes para las muestras de tipo *Control*, mientras que en color rojo quedan plasmados los correspondientes a las muestras de tipo *Cáncer*……..31

**Figura 16. (a**) Gráfico curva ROC/AUC para distintos modelos obtenidos mediante *Random Forest* (modelo escalado a *Pareto*) en función de las variables seleccionadas, y (**b**) Gráfico de precisión de la predicción en función de las variables seleccionadas. El modelo con valores más adecuados en ambas representaciones es el correspondiente a 100 variables, que presenta un valor AUC de 0. 969 en un *Intervalo de Confianza* al 95% de 0.878-0.999 y una precisión de predicción del 91.5%. ……34

**Figura 17.** Gráfico representando la matriz de confusión de las muestras de 100 *features* según el modelo *Random Forest* escalado a *Pareto* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal. Debido a que el algoritmo emplea un método de submuestreo balanceado, el límite de clasificación se encuentra localizado en el centro de la gráfica (x= 0.5, línea). Se observa una buena discriminación de las muestras en función de los grupos *Cáncer* y *Control*, destacando la presencia de cinco muestras incorrectamente clasificadas señalizadas mediante el color rojo (muestras *Cáncer* incorrectas) y el color azul (muestras *Control* incorrectas)……..……34

**Figura 18.** (**a**) Gráfico curva ROC/AUC para distintos modelos obtenidos mediante *Support Vector Machine* (modelo escalado a *Pareto*) en función de las *features* seleccionadas, y (**b**) Gráfico de precisión de la predicción en función de las variables seleccionadas. El modelo considerado como el más adecuado en ambas representaciones es el correspondiente a 100 variables, que presenta un valor AUC de 0.899 en un *Intervalo de Confianza* al 95% de 0.759-0.968 y una precisión de predicción del 82.8%.....35

**Figura 19.** Gráfico representando la matriz de confusión de las muestras de 100 *features* según el modelo *Support Vector Machine* escalado a *Pareto* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal. Debido a que el algoritmo emplea un método de submuestreo balanceado, el límite de clasificación se encuentra localizado en el centro de la gráfica (x= 0.5, línea). Se observa una discriminación adecuada de las muestras en función de los grupos *Cáncer* y *Control*, destacando la presencia de nueve muestras incorrectamente clasificadas señalizadas mediante el color rojo (muestras *Cáncer* incorrectas) y el color azul (muestras *Control* incorrectas)……..…….36

**Figura 20.** Gráfica de la frecuencia de selección de cada *bucket* generado a partir del modelo RF. A la derecha se observa una escala que identifica la probabilidad para cada grupo de selección de biomarcadores, siendo el color rojo la mayor probabilidad y el azul la mínima…….……37

**Figura 21.** (**a**) Regiones espectrales correspondientes a muestras de tipo *Cáncer* (tres primeras), y de tipo *Control* (tres últimas), en las cuales se observa claramente un cambio en el desplazamiento de la señal del piruvato acompañado de la presencia de otras señales (± 0.02 ppm), y (**b**) diagrama de caja del *bucket* en $\delta_H$ 2.38 ppm, en el que resalta la presencia en su mayoría en el grupo *Cáncer*……...……..39

# Lista de tablas

**Tabla 1.** Calendario de tareas propuestas……7

**Tabla 2.** Principales hitos del proyecto……..…**7**

**Tabla 3.** Tipos de escalado más empleados en análisis metabolómicos mediante RMN. Adaptada de la **referencia 18**……..…..14

**Tabla 4.** *Buckets* discriminantes (variables), su valor VIP de contribución y metabolito al que pertenecen de acuerdo a la presencia o no de cáncer en función del modelo OPLS-DA. Los *buckets* numéricos representan el centro de la región espectral (ppm) ± 0.02 ppm. [a], [b]……....32

**Tabla 5.** *Buckets* discriminantes (variables), su rango de contribución y metabolito al que pertenecen de acuerdo a la presencia o no de cáncer en función del modelo RF. Los *buckets* numéricos representan el centro de la región espectral (ppm) ± 0.02 ppm. [a], [b]……37

# 1. Resumen

El cáncer colorrectal es el tipo de cáncer más frecuente en España, y el tercero con mayor incidencia mundial. Actualmente existen métodos de diagnóstico de esta enfermedad muy invasivos tales como la colonoscopia, por lo que múltiples investigaciones se encuentran estudiando los perfiles metabólicos de muestras tales como el suero de pacientes con cáncer de colon, con el objetivo de determinar los posibles biomarcadores involucrados en este proceso. Para ello, se emplean múltiples técnicas analíticas tales como la *Resonancia Magnética Nuclear de Protón* (RMN de $^{1}$H), en combinación con técnicas de análisis multivariante supervisadas y no supervisadas. De esta forma, el presente *Trabajo de Fin de Máster* (TFM) engloba (1) la realización de una breve revisión bibliográfica de algunas de las principales investigaciones dedicadas a este fin empleando RMN de $^{1}$H como técnica analítica, y (2) el uso de dicha herramienta combinada con métodos de análisis multivariante de datos de naturaleza supervisada y no supervisada en el análisis de 90 muestras de suero de pacientes con cáncer colorrectal para la obtención de posibles biomarcadores y su correlación con esta enfermedad.

Para ello, se han empleado modelos de *Análisis Discriminante de Mínimos Cuadrados* (PLS-DA), *Análisis Discriminante de Mínimos Cuadrados Parciales Ortogonal* (OPLS-DA), *Random Forest* (RF) y *Support Vector Machine* (SVM). Se obtuvieron un total de 12 metabolitos discriminantes para el modelo OPLS-DA y 16 para el modelo RF, posiblemente relacionados con la glucólisis, y los metabolismos del piruvato y de la alanina, del aspartato y del glutamato. Se ha concluido que la RMN acoplada a técnicas multivariantes es una poderosa herramienta de predicción y obtención de biomarcadores asociados con la aparición y/o desarrollo de esta enfermedad.

# 2. Introducción

## 2.1. Contexto y justificación del Trabajo

### 2.1.1. Descripción general

Este TFM se centra en el estudio mediante RMN de $^{1}$H de un conjunto de muestras metabolómicas de suero de pacientes con cáncer colorrectal y de un grupo control, con el objetivo de determinar posibles biomarcadores de esta enfermedad mediante modelos de discriminación supervisados y no supervisados de las muestras.

Para ello, se llevará a cabo una breve revisión bibliográfica de esta temática mediante la cual se reconozcan las técnicas multivariantes más empleadas en metabolómica, para posteriormente aplicarlas en un conjunto de datos derivados de esta enfermedad empleando algunas de estas metodologías, para finalmente realizar una interpretación biológica de los resultados.

### 2.1.2. Justificación del TFM

El cáncer colorrectal es aquél que tiene su origen en el colon o en el recto, y comienza su desarrollo cuando las células comienza a crecer de manera descontrolada, modificando su forma, tamaño y otras características. Este tipo de cáncer es predominante en personas de edad superior a los 50 años, afectando a los dos sexos prácticamente por igual. Hoy en día, gran parte de los factores de riesgo de esta enfermedad están relacionados con hábitos de vida poco saludables, como el sobrepeso, la obesidad, el tabaquismo, el sedentarismo, y el consumo excesivo de alcohol. Por otro lado, algunos factores de riesgo que no dependen del individuo son aquellos que engloban

enfermedades y condiciones predisponentes (como la presencia de pólipos en el colon y/o recto, y enfermedades inflamatorias como la enfermedad de Crohn), el hecho de haber padecido un cáncer colorrectal previamente, y presentar factores genéticos o familiares.

En 2018, representó el tercer tipo de cáncer con mayor incidencia a nivel mundial, después del cáncer de pulmón y de mama, y en España se trata del tumor diagnosticado más frecuente, reportando 44.937 nuevos casos en 2019 según el informe de *la Sociedad Española de Oncología Médica* (SEOM). En cuanto a la mortalidad, la *Asociación Española Contra el Cáncer* (AECC) reportó que este es el segundo cáncer con mayor mortalidad, generando un total de 15.923 defunciones al año.

Así, uno de los mayores retos actuales para la biomedicina en este campo, es el de tratar de hallar biomarcadores novedosos que puedan ayudar en el diagnóstico temprano y en el consecuente tratamiento de este tipo de enfermedades.

Para ello, el estudio de distintas ciencias ómicas tales como la transcriptómica, la genómica, la proteómica, y la metabolómica, ha ayudado considerablemente en la identificación y cuantificación de dichos biomarcadores. Concretamente, la metabolómica es actualmente uno de los campos más estudiados con dicho objetivo, ya que, aunque se trata de la ciencia ómica más joven, está demostrando tener múltiples ventajas sobre las demás, tales como: (1) la obtención de un número inferior de metabolitos, que simplifica la complejidad final de los datos ómicos, (2) los metabolitos obtenidos consiguen reflejar de una forma más adecuada el nivel funcional de una célula, (3) la identificación y cuantificación de estos metabolitos finales es más correcta, ya que a lo largo de los flujos metabólicos se ven influenciados por el estrés ambiental, pudiendo variar los resultados, y (4) la concentración final de metabolitos puede variar, aunque la concentración de los flujos metabólicos a lo largo de una reacción bioquímica no varíe demasiado.

Los estudios metabolómicos se apoyan en distintas plataformas analíticas de alta resolución para la obtención de las medidas correspondientes a los metabolitos. En concreto, una de las técnicas más empleadas, que ofrece grandes ventajas, es la Resonancia Magnética Nuclear (RMN), la cual se presenta como una plataforma robusta y versátil que permite la medición de un gran número de metabolitos de forma fiable y repetitiva, sin necesidad de separación o derivatización, y presenta una alta sensibilidad gracias al uso de criosondas. Con esta plataforma, se obtienen las medidas de los metabolitos en forma de *dataset*, que generalmente suele ser dividido en zonas de 0.04 ppm (denominados *buckets* o *bins*) para seguidamente ser normalizado y escalado, haciendo que las medidas sean comparables entre sí.

Con el objetivo de determinar posibles biomarcadores de enfermedades y/o condiciones que puedan servir como marcadores-diagnóstico, en metabolómica suele emplearse el método no supervisado multivariante *Análisis de Componentes Principales* (PCA) sobre el set de datos. Comúnmente, además son aplicados modelos supervisados tales como PLS-DA y OPLS-DA. Sin embargo, en estos últimos años se ha observado una tendencia creciente a emplear modelos no lineales de *Machine Learning* tales como SVM, *Artificial Neural Networks* (ANN) y RF, los cuales agrupan las distintas muestras contenidas en el *dataset* de acuerdo con su similitud.

Estos modelos son posteriormente validados empleando técnicas de validación cruzada o de *bootstrapping*, y la selección de los metabolitos más relevantes para cada modelo es llevada a cabo en función del método escogido en cada caso, siendo el más común los valores *Variable Importance In Projection* (VIP) para los métodos PLS-DA y OPLS-DA. Finalmente, los estudios llevan a cabo la interpretación biológica de los resultados, con el objetivo de arrojar conclusiones relevantes asociadas a los perfiles metabolómicos identificados.

## 2.2. Objetivos del Trabajo

### 2.2.1. Objetivos generales

Este TFM persigue los siguientes objetivos generales:

(1) Objetivo 1: Llevar a cabo una revisión bibliográfica clara y concisa en aras de obtener información útil sobre modelos de predicción y búsqueda de biomarcadores sobre metabolómica mediante RMN de suero de pacientes con cáncer colorrectal.

(2) Objetivo 2: Preparar y analizar el conjunto de datos seleccionado en este TFM mediante técnicas de análisis multivariante y de *Machine Learning* para obtener posibles biomarcadores de cáncer colorrectal empleando distintos modelos de clasificación.

(3) Objetivo 3: Distinguir los mejores modelos de predicción e interpretar biológicamente los resultados.

### 2.2.2. Objetivos específicos

Los puntos generales serán abordados por los siguientes puntos específicos:

(1) Objetivo 1: Llevar a cabo una revisión bibliográfica clara y concisa en aras de obtener información útil sobre modelos de predicción y búsqueda de biomarcadores sobre metabolómica mediante RMN de suero de pacientes con cáncer colorrectal.
   a) Búsqueda exhaustiva de información sobre el análisis de muestras de este tipo.
   b) Breve revisión bibliográfica empleando bases de datos y literatura científica.

(2) Objetivo 2: Preparar y analizar el conjunto de datos seleccionado en este TFM mediante técnicas de análisis multivariante y de *Machine Learning* para obtener posibles biomarcadores de cáncer colorrectal empleando distintos modelos de clasificación.
   a) Preparación del set de datos tras el análisis de RMN en una *bucket table.*
   b) Comprobación de la estructura del *dataset*, recopilación de los datos más relevantes, y si es necesario, transformación de los datos.
   c) Aplicación de métodos no supervisados con el objetivo de obtener discriminación entre los grupos presentes en las variables.
   d) Aplicación de modelos supervisados lineales para obtener modelos de clasificación de las muestras y determinar aquellos metabolitos más relevantes para los modelos.

(3) Objetivo 3: Distinguir los mejores modelos de predicción e interpretar biológicamente los resultados.
   a) Obtención de los parámetros que determinan la calidad de los modelos anteriormente generados, indicando aquellos con los mejores resultados.
   b) Interpretación biológica de los metabolitos, explorando las rutas metabólicas posiblemente implicadas en el desarrollo de la enfermedad.

## 2.3. Enfoque y método seguido

El enfoque planteado para lograr los objetivos expuestos en los puntos del **apartado 2.2.** tiene su comienzo en una búsqueda exhaustiva de bibliografía empleando una serie de palabras clave relacionadas con el asunto principal de este trabajo en la base de datos *WebOfScience*, para elaborar posteriormente una breve revisión bibliográfica. Las publicaciones científicas

seleccionadas determinan los métodos más empleados actualmente para la búsqueda de potenciales biomarcadores empleando modelos de predicción. Con este sistema puede generarse una base realista para el flujo de trabajo a seguir para lograr los objetivos, certificando el cumplimiento de las estrategias más apropiadas para ello.

Seguidamente, se analiza y se prepara el *dataset*, facilitado por el grupo de investigación de metabolómica aplicada mediante RMN de la Universidad de Almería (*NMRMBC*). Este incluye datos espectrales de suero de pacientes con cáncer colorrectal y de suero de personas sanas que formaron parte de un grupo control.

Empleando distintos programas populares en el ámbito de la metabolómica, tales como *Amix*, *RStudio*, *SIMCA* y la herramienta web *MetaboAnalyst*, se lleva a cabo el análisis multivariante y la aplicación de algunos métodos supervisados y no supervisados observados en la revisión bibliográfica realizada, buscando potenciales biomarcadores por medio de modelos de clasificación. A continuación, se evalúan y se comparan dichas metodologías.

Finalmente, se interpretan biológicamente los resultados obtenidos y la bibliografía consultada.

## 2.4. Planificación del Trabajo

### 2.4.1. Tareas

Los objetivos planteados son desglosados en tareas, las cuales estarán marcadas mediante una duración determinada y serán las siguientes:

(1) Objetivo 1: Llevar a cabo una breve revisión bibliográfica con el objetivo de obtener información útil sobre modelos de predicción y búsqueda de biomarcadores sobre metabolómica mediante RMN de suero de pacientes con cáncer colorrectal.

a) Búsqueda exhaustiva de información sobre el análisis de muestras de este tipo.
   - ❖ Tarea 1: Elaboración de este documento (2 semanas y 1 día).
   - ❖ Tarea 2: Búsqueda bibliográfica de información sobre el tema en cuestión de este TFM (1 semana).

b) Breve revisión bibliográfica empleando bases de datos y literatura científica.
   - ❖ Tarea 3: Llevar a cabo una breve revisión bibliográfica (2 semanas).
   - ❖ Tarea 4: Elaboración de una introducción sintetizada del tema abordado (2 semanas).
   - ❖ Tarea 5: Selección de los métodos supervisados y no supervisados más empleados en los artículos seleccionados (1 semana).

(2) Objetivo 2: Preparar y analizar el conjunto de datos seleccionado en este TFM mediante técnicas de análisis multivariante y de *Machine Learning* para obtener posibles biomarcadores de cáncer colorrectal empleando distintos modelos de clasificación.

a) Preparación del set de datos tras el análisis de RMN en una *bucket table.*
   - ❖ Tarea 6: Obtención de los datos espectrales, que serán proporcionados por el grupo de metabolómica aplicada mediante RMN (*NMRMBC*) de la Universidad de Almería (6 días).
   - ❖ Tarea 7: Organización de los datos en una *bucket table* empleando el software *Amix* implementando un *bucketing* de 0.04 ppm y la normalización de los mismos.

Seguidamente, exportación y definición de los grupos de estudio de los datos en formato *.csv* mediante el software *Excel* (6 días).

b) Comprobación de la estructura del *dataset*, recopilación de los datos más relevantes, y si es necesario, transformación de los datos.
   - ❖ Tarea 8: Implementación de los datos en *R* mediante la interfaz *RStudio*, llevando a cabo un estudio del formato (1 semana).
   - ❖ Tarea 9: Preprocesamiento de los datos (1 semana).

c) Aplicación del modelo no supervisado PCA para contemplar una posible discriminación de las muestras entre los grupos de las variables.
   - ❖ Tarea 10: Implementación de distintos tipos de escalados empleados en metabolómica, tales como *Unit Variance, Pareto, Range Scale, y Vast Scale*, comentando los resultados obtenidos en sus agrupaciones mediante sus representaciones en PCA (1 semana).

d) Aplicación de modelos supervisados para obtener modelos de clasificación de las muestras, determinando aquellos metabolitos más relevantes para cada metodología.
   - ❖ Tarea 11: Escoger los paquetes y/o programas más adecuados para implementar cada metodología mediante las tendencias observadas en la revisión bibliográfica, el repositorio *CRAN* y *Bioconductor* (4 semanas).
   - ❖ Tarea 12: Implementar los algoritmos y las técnicas empleadas para hallar las variables más relevantes en cada modelo (4 semanas).

(3) Objetivo 3: Distinguir los mejores modelos de predicción e interpretar biológicamente los resultados.

a) Obtención de los parámetros que determinan la calidad de los modelos anteriormente generados, indicando aquellos con los mejores resultados.
   - ❖ Tarea 15: Generación de los parámetros necesarios para cada modelo de forma sincrónica a la obtención de los mismos (1 semana).

b) Interpretación biológica de los metabolitos, explorando las rutas metabólicas posiblemente implicadas.
   - ❖ Tarea 16: Empleando la web *MetaboAnalyst*, exploración de las rutas implicadas en los metabolitos de interés (1 semana).
   - ❖ Tarea 17: Reforzar las conclusiones obtenidas mediante la bibliografía consultada (1 semana).

### 2.4.2. Calendario

A continuación, se incluye en la **Tabla 1** un calendario de Tareas que incluye su duración junto con las fechas concretas, además de un diagrama de Gantt generado con el programa *ganttproject* en la **Figura A1 del anexo A** en el cual son organizadas de acuerdo al marco temporal, haciendo este esquema más visual.

**Tabla 1.** Calendario de tareas propuestas.

| Tareas | Fechas inicio | Fechas final |
|---|---|---|
| ***PEC 0.*** **Definición de los contenidos del trabajo** | 17/02/2021 | 01/03/2021 |
| ***PEC 1.*** **Plan de trabajo** | 02/03/2021 | 16/03/2021 |
| ***PEC 2.*** **Desarrollo del trabajo – Fase 1** | 17/03/2021 | 19/04/2021 |
| **Búsqueda bibliográfica** | 17/03/2021 | 23/03/2021 |
| **Revisión bibliográfica** | 24/03/2021 | 06/04/2021 |
| **Elaboración de breve introducción al tema** | 24/03/2021 | 06/04/2021 |
| **Selección de métodos** | 31/03/2021 | 06/04/2021 |
| **Obtención del *dataset*** | 07/04/2021 | 12/04/2021 |
| **Organización de los datos en una *bucket table*** | 07/04/2021 | 12/04/2021 |
| **Implementación de los datos en R** | 13/04/2021 | 19/04/2021 |
| **Preprocesamiento de los datos** | 13/04/2021 | 19/04/2021 |
| ***PEC 3.*** **Desarrollo del trabajo – Fase 2** | 20/04/2021 | 17/05/2021 |
| **PCA con distintos escalados del *dataset*** | 20/04/2021 | 26/04/2021 |
| **Selección de programas para métodos supervisados** | 27/04/2021 | 10/05/2021 |
| **Aplicación de métodos supervisados** | 27/04/2021 | 10/05/2021 |
| **Interpretación biológica de los resultados** | 11/05/2021 | 17/05/2021 |
| **Refuerzo de conclusiones mediante bibliografía** | 11/05/2021 | 17/05/2021 |
| ***PEC 4.*** **Cierre de la memoria** | 18/05/2021 | 08/06/2021 |
| ***PEC 5 a.*** **Elaboración de la presentación** | 09/06/2021 | 13/06/2021 |
| ***PEC 5 b.*** **Defensa pública** | 16/06/2021 | 23/06/2021 |

### 2.4.3. Hitos

La **Tabla 2** muestra los hitos de este proyecto, concretados en el Plan Docente de la asignatura.

**Tabla 2.** Principales hitos del proyecto.

| Hitos | Fechas |
|---|---|
| ***PEC 0. Definición de los contenidos del trabajo*** | 1 de Marzo de 2021 |
| ***PEC 1. Plan de trabajo*** | 16 de Marzo de 2021 |
| ***PEC 2. Desarrollo del trabajo – Fase 1*** | 19 de Abril de 2021 |
| ***PEC 3. Desarrollo del trabajo – Fase 2*** | 17 de Mayo de 2021 |
| ***PEC 4. Cierre de la memoria*** | 08 de Junio de 2021 |
| ***PEC 5a. Elaboración de la presentación*** | 13 de Junio de 2021 |
| ***PEC 5b. Defensa pública*** | 23 de Junio de 2021 |

### 2.4.4. Análisis de riesgos

A continuación, se resumen los riesgos asociados a este Trabajo de Fin de Máster, los cuales pueden llegar a afectar al desarrollo del Plan de Trabajo. En sí, se dividen en:

- Riesgos referentes al *factor tiempo*, ya que puede que algunas de las tareas planificadas requieran más tiempo del programado.
- Riesgos referentes al *alcance del proyecto*, ya que se debe determinar qué pasos seguir y qué modelos implementar para completar el estudio dentro del marco temporal planteado.
- Riesgos referentes a la *búsqueda bibliográfica*, ya que se pretende generar una base adecuada para llevar a cabo el proyecto, por lo que la selección de los artículos debe ser estudiada en profundidad.
- Riegos referentes a la *búsqueda de paquetes y programas* adecuados para el tratamiento del *dataset*, con el objetivo de seleccionar aquellos que tengan las funciones necesarias para el análisis que se llevará a cabo.
- Riesgos referentes al *procesamiento y tratamiento de los datos*, ya que cualquier tipo de error puede suponer la pérdida de información relevante.
- Riesgos referentes a la *privacidad de los individuos participantes* en el estudio, que quedan solventados ya que se siguen los preceptos determinados por la *declaración de Helsinki (1964/1975/2000).*
- Riesgos referentes a *no obtener modelos válidos*, en el caso de disponer de un tamaño muestral (n) muy bajo.
- Riesgos referentes a la *identificación de biomarcadores por RMN*, ya que, si se encuentran en bajas concentraciones, puede suponer el no lograr dilucidarlos.

## 2.5. Breve sumario de contribuciones y productos obtenidos

Tras finalizar el presente TFM, se habrán obtenido los siguientes documentos:

- *Plan de Trabajo* (PEC 0) en el cual se presentarán los objetivos perseguidos y el flujo de trabajo para alcanzarlos.
- *Informe de la parte trabajada en R*, en el que se incluirán todos los métodos aplicados al *dataset* y los resultados obtenidos con dicho lenguaje.
- *Memoria*, que describirá el proyecto, el proceso de análisis, y los resultados y conclusiones.
- *Presentación virtual*, en la cual se expondrán todos los apartados de la memoria, haciendo una mención especial de aquellos puntos más relevantes.
- *Autoevaluación del proyecto* realizado.
- *Artículo de revisión bibliográfica*, que será desarrollado a partir del apartado de revisión de este Trabajo de Fin de Máster.
- *Artículo de análisis metabolómico* del conjunto de datos analizado en este trabajo, que será posteriormente completado con otras variables y conclusiones.

## 2.6. Breve descripción de los otros capítulos de la memoria

El resto de los capítulos contenidos en la presente memoria vienen resumidos en los siguientes apartados:

- *Introducción* en la cual se detalla el contexto del proyecto, aportando una visión más detallada del problema que se desea mitigar y de las alternativas actuales más efectivas para tratar de solucionarlo.
- *Metodología*, capítulo en el que se incluyen los métodos empleados para realiza el estudio.

- *Resultados* obtenidos mediante los algoritmos implementados.
- *Discusión* de los resultados anteriormente obtenidos.
- *Conclusiones* e interpretación biológica de los resultados.
- *Glosario* que detalle aquellos acrónimos y/o anglicismos empleados.
- *Bibliografía* empleada para el desarrollo del proyecto.
- *Anexo* en el cual se incluyen tablas y figuras.

# 3. Estado del arte

## 3.1. Cáncer de colon

### 3.1.1. La realidad del cáncer de colon

El cáncer colorrectal es aquel tumor maligno localizado en el colon y/o en el recto y que constituyen la parte final del tracto digestivo. Su desarrollo se inicia cuando las células comienzan a crecer descontroladamente, modificando su forma, tamaño y otras características. Así, se genera una serie de crecimientos en el revestimiento interno del colon o del recto, conocidos como pólipos. Estos, pueden derivar en cáncer a lo largo del tiempo, y pueden ser de dos clases, *pólipos adenomatosos* (adenomas), que pueden llegar a desembocar en cáncer, y *pólipos inflamatorios* y *pólipos hiperplásicos*, que son más frecuentes, y en general no son precancerosos [1, 2].

En 2018, el cáncer colorrectal representó el tercero cáncer con mayor incidencia a nivel mundial, después del cáncer de pulmón y de mama según el proyecto GLOBOCAN [3], y en España se trata del tumor diagnosticado más frecuente, con 44.937 nuevos casos en 2019 según el informe de la SEOM [4].

En cuanto a los datos de mortalidad, según la AECC en 2018, se trata del segundo cáncer con mayor mortalidad, generando un total de 15.923 defunciones al año [5]. Sin embargo, si se consigue llevar a cabo un diagnóstico temprano del cáncer de colon, la tasa relativa de supervivencia a 5 años ha demostrado ser muy alta, del 91% de acuerdo con la información obtenida de las personas diagnosticadas entre los años 2010 y 2016, por lo que se recomienda llevar a cabo cribados masivos en la población que supere los 50 años, circunstancia que constituye uno de los principales factores de riesgo de padecer esta enfermedad [6].

Hoy en día, otras de las causas que pueden derivar en el desarrollo de estos tumores están relacionadas con hábitos de vida poco saludables, como el sobrepeso, la obesidad, el tabaquismo, el consumo de carnes procesadas, el sedentarismo, y el consumo excesivo de alcohol. Por otro lado, algunos factores de riesgo no dependientes del individuo se pueden clasificar en: (1) enfermedades y condiciones predisponentes, como la presencia de pólipos en el colon y/o recto, y enfermedades inflamatorias tales como la enfermedad de Crohn y la colitis ulcerosa, (2) haber padecido anteriormente un cáncer colorrectal, que aumenta el riesgo de uno posterior y (3) presentar factores genéticos, como el síndrome de *Lynch* y la *poliposis adenomatosa familiar* (PAF), o factores familiares, ya que la incidencia ha demostrado ser mayor en aquellas personas con parientes que han presentado cáncer colorrectal [7].

### 3.1.2. Metabolómica para el diagnóstico de cáncer de colon

Actualmente, la detección de pólipos cancerosos de esta clase es llevada a cabo mediante análisis visuales de la estructura del colon y del recto, tales como la *colonoscopia* y la *sigmoidoscopia*, siendo la primera de ellas la técnica más empleada y con mayor sensibilidad de detección de este tipo de patologías.

Sin embargo, esta metodología presenta una clara desventaja, y es que resulta altamente invasiva para los pacientes que se someten a ella, además de requerir una preparación previa que implica un aumento de ciertos riesgos tales como la perforación del intestino y la propia ansiedad que puede llegar a generar en los individuos [8].

Por esta razón, técnicas no invasivas como las implicadas en las distintas ciencias ómicas como la genómica, la proteómica, la transcriptómica, y la metabolómica, se encuentran en pleno auge en el campo de la biotecnología contribuyendo de forma fundamental al entendimiento y a la predicción de cuestiones biológicas básicas. En primer lugar, la genómica es la ciencia que estudia el conjunto de genes perteneciente a un genoma determinado; la transcriptómica es aquella que analiza el conjunto de ARN derivado de una célula, tejido u órgano; la proteómica, por su lado, estudia las proteínas y modificaciones postranscripcionales que las regulan; y la metabolómica es la disciplina que identifica y cuantifica el metaboloma de un sistema biológico, es decir, estudia los metabolitos de bajo peso molecular. Esta última ciencia permite obtener una visión general del estado final de un organismo ya que ofrece información sobre la actividad celular.[9]

A medida que los avances tecnológicos progresan y a la información que nos aporta cada una, las ciencias ómicas están haciéndose cada vez más notables en el ámbito sanitario, permitiendo el desarrollo de diagnósticos personalizados cada vez más tempranos a los pacientes, e incluso, previniendo el desarrollo de ciertas enfermedades mediante la formación de equipos multidisciplinares que ayudan a la interpretación de la alta cantidad de datos generados mediante estos métodos [10].

La metabolómica suele ser considerada un "análisis complementario" al resto de ómicas, aunque en los últimos años se ha apostado por su desarrollo en solitario debido a sus ventajas, haciendo de ella una técnica viable en investigaciones dedicadas a la búsqueda de posibles biomarcadores de, en este caso, cáncer colorrectal, y en el desarrollo de modelos de clasificación. Algunas de estas ventajas se proporcionan a continuación:

(1) La obtención de un número de metabolitos sustancialmente inferior a lo obtenido mediante otras ciencias ómicas, lo que simplifica el tratamiento de datos;
(2) Los metabolitos obtenidos consiguen reflejar de forma fiel el nivel funcional de una célula;
(3) La identificación y cuantificación de estos metabolitos es fiable y reproducible, y permite cuantificar de forma precisa los niveles de concentración que por ejemplo a lo largo de los flujos metabólicos pueden verse influenciado por el estrés ambiental o a lo largo de una reacción bioquímica [11].

Los estudios metabolómicos se apoyan en distintas plataformas analíticas de alta resolución para la obtención de las medidas correspondientes que permitan llegar al conjunto de metabolitos involucrados. En concreto, una de las técnicas más empleadas y que a su vez ofrece grandes ventajas, es la RMN, que se presenta como una plataforma robusta y versátil que permite la medición de un gran número de metabolitos de forma fiable y repetitiva, presentando una alta sensibilidad gracias al uso de criosondas. Aunque existen múltiples herramientas analíticas para este propósito como la *espectrometría de masas* (MS) acoplada a métodos de separación como la

*cromatografía de gases* (GC) y de *líquidos* (LC), la RMN está demostrando suplir muchas de las desventajas que estas otras técnicas presentan: (1) se tratan de métodos destructivos de la muestra, (2) requieren de un paso previo de separación, (3) al método GC-MS suele acompañarle un paso de derivatización, (4) en el caso de LC-MS se requiere de algún tipo de analizador, ya que no puede llevarse a cabo la identificación de los metabolitos directamente con alguna biblioteca debido a la formación de aductos [12, 13], (5) dependen de la ionización de los analitos, (6) presentan efecto matriz por lo que las condiciones no solo de separación sino también de ionización dependen fuertemente de la matriz en el que se deseen cuantificar los metabolitos objeto de estudio, (7) se necesitan de patrones externos para cuantificar concentración y recuperación.

Así, la espectroscopía de RMN se presenta como una técnica poderosa para la identificación de metabolitos concretos incluso en mezclas complejas al no presentar efecto matriz. No requiere de pasos de separación o derivatización, además de realizar un análisis no destructivo y no invasivo de la muestra, proporcionando información cuantitativa y estructural sobre la misma simultáneamente. En la **Figura 1** se muestra el espectrómetro de RMN 600 MHz dotado de criosonda cuádruple empleado para el análisis del conjunto de datos empleado en este Trabajo de Fin de Máster.


**a)** **b)** **c)**

**Figura 1.** (**a**) Espectrómetro *Bruker Avance* III 600 equipado con automuestreador termostatizado (**b**) *SampleCase* de 24 posiciones y (**c**) *SampleJet* de hasta 480 posiciones. Imágenes extraídas de la **referencia 14**.

Haciendo uso de esta técnica, en metabolómica suele ser necesaria la eliminación de la señal del agua de las muestras, ya que normalmente presentan una gran diferencia de concentración de agua entre ellas y con respecto a sus propios metabolitos. Para este propósito generalmente se emplean experimentos tales como la presaturación de la señal del disolvente empleando un pulso de onda continua, y 1D-NOESY PRESAT, que junto con el módulo de presaturación, introduce una secuencia de triple pulso de 90º que consigue eliminar dicha señal efectivamente y sin causar grandes distorsiones en señales adyacentes.

También puede llegar a ser necesaria la eliminación de señales en función del peso molecular, para lo que son empleados los llamados *filtros de difusión*, [15] que emplean una combinación de pulsos de radiofrecuencia y de gradiente de campo magnético que consigue atenuar las señales procedentes de moléculas de menor tamaño que suelen ser las provenientes del disolvente empleado, aunque tienen como inconveniente que el resto de metabolitos también sufren atenuación en sus señales en mayor o menor medida en función de su tamaño. También se aplican

los denominados *filtros de relajación* $T_2$, como el de *Carr Purcell Meiboon Gill* (CPMG), que elimina las señales con tiempos de relajación transversal ($T_2$) pequeños, que suelen estar asociados a sistemas de grandes tiempos de correlación generalmente presentes en macromoléculas o proteínas [13].

En la **Figura 2**, obtenida de la base de datos *Web Of Science* [16] en la *Web Of Science Core Collection* mediante la búsqueda de las palabras clave "*NMR*" y "*metabolomics*", se muestra la evolución de los 5801 resultados obtenidos desde el año 2001 hasta la actualidad, eliminando los documentos de tipo *Early Access*. El desarrollo y mejora de las técnicas analíticas en estos años han permitido aumentar la sensibilidad de la RMN hasta en un factor cinco, haciendo que la metabolómica sea un campo de investigación cada vez más estudiado y aplicado, como refleja la tendencia ascendente de la gráfica en estos últimos años.

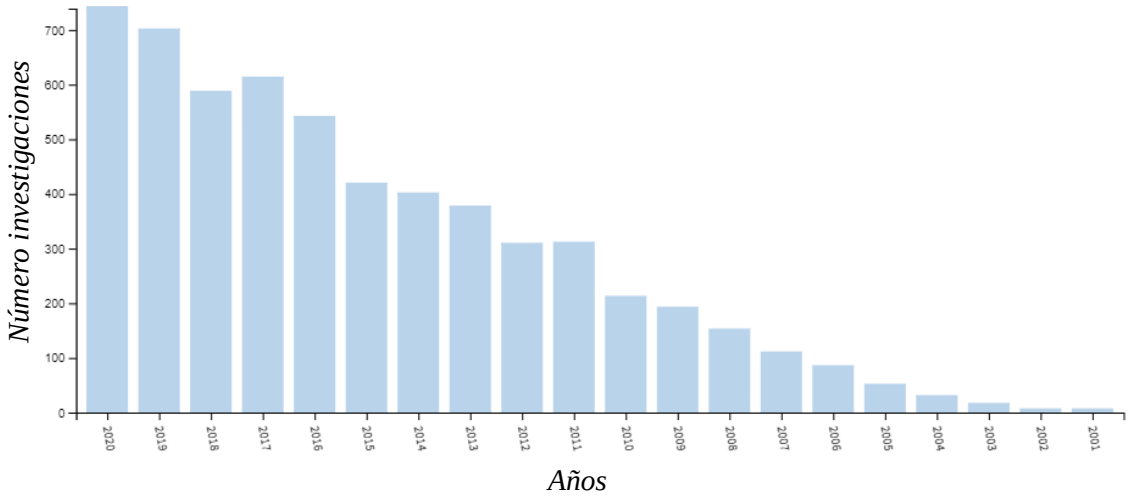



**Figura 2.** Avance de las investigaciones centradas en metabolómica y RMN desde el año 2001 hasta la actualidad. Obtenida de *ISI Web of Knowledge* mediante la búsqueda de las palabras "*NMR*" y "*metabolomics*".

Además, puede verse reflejada la aceptación recibida por los campos de la bioquímica, de la química analítica y de la salud en general mediante un análisis de los temas en los cuales se agrupan dichas investigaciones (**Figura 3**).

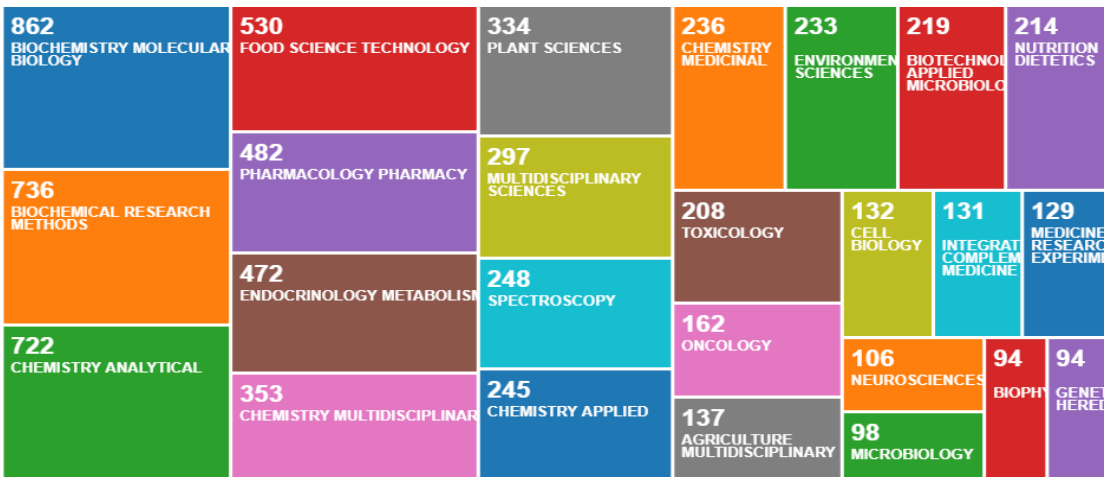



**Figura 3.** Diagrama de bloques según el área de investigación, obtenido de *ISI Web Of Knowledge* introduciendo las palabras clave "NMR" y "metabolomics".

## 3.2. Análisis de datos metabolómicos analizados mediante RMN de $^{1}H$

Los estudios metabolómicos, como el resto de las ciencias ómicas, generan una gran cantidad de datos y precisan números elevados de muestras, por lo que es de especial interés la reducción de su dimensión, con el propósito de generar una interpretación más adecuada y correcta.

Para ello, se emplea la *quimiometría*, que es una técnica que combina procedimientos matemáticos y estadísticos que permiten extraer la información más relevante de los datos experimentales obtenidos, mejorando el proceso de interpretación de grandes conjuntos de datos y aportando calidad a los resultados. Hoy en día, en química, se emplea sobre todo para el procesado de señales, diseños experimentales, reducción de variables, exploración de datos, análisis multivariantes y reconocimiento de patrones [17].

El proceso de análisis de datos metabolómicos analizados mediante RMN sigue una serie de pasos, recogidos en los siguientes apartados:

(1) *Procesado de los espectros*
(2) *Normalización*
(3) *Centrado y escalado*
(4) *Análisis estadístico*
(5) *Interpretación biológica.*

### 3.2.1. Procesado de los espectros

Este paso engloba la transformación de los datos espectrales en su versión óptima para el posterior análisis estadístico, comprobando ausencias de datos posiblemente debidas a metabolitos por debajo del límite de detección, ajuste de línea base, referenciado del espectro de forma que el patrón interno se encuentre localizado en el mismo desplazamiento químico en todos los espectros, multiplicación del espectro por funciones que suavicen o acentúen la resolución espectral, aplicación de algoritmos que minimicen la fluctuación en desplazamiento químico como consecuencia de variaciones en la temperatura, supresión de regiones espectrales defectuosas o en donde existen desplazamientos de señales, habitualmente provenientes de grupos ácidos, intercambio químico, etc.

### 3.2.2. Normalización

En el análisis mediante RMN son adquiridos volúmenes idénticos de muestra con el propósito de hacer todas las muestras comparables entre sí. Sin embargo, en el caso de muestras correspondientes a biofluídos, existen múltiples variables externas que pueden llegar a afectar a la concentración de los metabolitos, tales como el estado de hidratación de cada individuo, o incluso posibles inexactitudes experimentales o errores técnicos.

Con el objetivo de obtener volúmenes y concentraciones comparables, se aplica un paso de *normalización*, que logra corregir estos factores de dilución o concentración entre muestras. En metabolómica se emplean una serie de métodos, aunque generalmente se aplica la *normalización a la intensidad del área total* del espectro, mediante la cual se lleva a cabo la división de los valores de cada región espectral en la que se divide el espectro o *bucket* (ver más adelante) entre la suma de todos ellos, de forma que la suma de todos los resultados a de proporcionar un valor igual a la unidad.

### 3.2.3. Centrado y escalado

Generalmente, en los estudios metabolómicos en los cuales el objetivo es identificar nuevos biomarcadores, suelen emplearse técnicas de análisis multivariante que extraen información de los datos mediante su proyección en la dirección de la máxima varianza. Estos análisis de datos se centran en el perfil espectral, y cualquier información de variación biológica puede verse solapada, por lo que el *centrado mediante la media* de los datos es un paso bastante común, ya que permite compensar este problema, enfocándose en la variación biológica y en las posibles diferencias y similitudes entre las muestras.

Sin embargo, aquellos metabolitos que sean más abundantes en las muestras mostrarán valores más altos en la tabla de datos, por lo que terminarán contribuyendo en mayor medida al modelo que se genere posteriormente. Para evitar este sesgo se emplean métodos de *escalado*, tales como los mostrados en la **Tabla 3**.

**Tabla 3.** Tipos de escalado más empleados en análisis metabolómicos mediante RMN. Adaptada de la **referencia 18**.

| Escalado | Fórmula | Suposición | Ventajas | Desventajas |
|---|---|---|---|---|
| ***Unit-Variance (Autoscaling)*** | $\widetilde{x_{ij}} = \frac{x_{ij} - \overline{x_i}}{s_i}$ | Compara los metabolitos en función de sus correlaciones | Todos los metabolitos resultan igual de importantes | Aumento de los errores de medida |
| ***Pareto scaling*** | $\widetilde{x_{ij}} = \frac{x_{ij} - \overline{x_i}}{\sqrt{s_i}}$ | Reduce la importancia relativa de valores altos, pero deja la estructura de los datos prácticamente intacta | Los valores obtenidos son más cercanos a los originales que con *autoscaling* | Sensible a grandes cambios |
| ***Range scaling*** | $\widetilde{x_{ij}} = \frac{x_{ij} - \overline{x_i}}{(x_{imax} - x_{imin})}$ | Compara los metabolitos en función de su rango de respuesta biológico | Todos los metabolitos resultan igual de importantes, y el escalado está más relacionado con la biología. | Aumento de los errores de medida, y se vuelve más sensible a los *outliers* |

| | | | | |
|---|---|---|---|---|
| ***Vast scaling*** | $\widetilde{x_{ij}} = \frac{x_{ij} - \bar{x_i}}{s_i} \cdot \frac{\bar{x_i}}{s_i}$ | Se centra en los metabolitos con pequeñas fluctuaciones | Persigue la robustez, y puede emplear conocimiento previo sobre los grupos | No es recomendado para una gran variación inducida sin estructura de los datos |
| ***Level scaling*** | $\widetilde{x_{ij}} = \frac{x_{ij} - \bar{x_i}}{\bar{x_i}}$ | Se centra en la respuesta relativa | Adecuado para identificación | Aumento de los errores de media |

### 3.2.4. Análisis estadístico

Los análisis metabolómicos, en términos globales, pueden dividirse en función de si se tiene algún tipo de conocimiento previo sobre los metabolitos de interés, o si no se posee información sobre los mismos. Los primeros se centran en el seguimiento de dichos compuestos seleccionados anteriormente en función de las rutas metabólicas conocidas, o de si se tratan de biomarcadores asociados sólidamente con la condición de estudio. Así, estos metabolitos deben ser apropiadamente asignados y cuantificados en las muestras.

Por otro lado, aquellos análisis metabolómicos en los cuales no se posee ningún tipo de metabolito objetivo, se centran en el estudio del perfil espectral como un todo, y por tanto considera todos los analitos presentes en la muestra. Para ello, en primer lugar, pueden emplearse dos técnicas: (a) el método *fingerprinting*, con el cual se obtiene una evaluación rápida del total de los metabolitos presentes en los espectros mediante su transformación en matrices de datos empleando el método *bucketing* (o *binning*), con el que se toman pequeñas porciones (*buckets*) de los espectros de una anchura de entre 0.02-0.04 ppm, posteriormente empleados para llevar a cabo los análisis estadísticos pertinentes y llevar a cabo clasificaciones, y/o (b) el *profiling*, que consiste en estudiar el conjunto del espectro empleando algoritmos específicos de alineamiento de los picos, y que es empleado para determinar las concentraciones de todos los metabolitos cuantificables en las muestras biológicas, aportando información útil desde el punto de vista bioquímico. La **Figura 4** ilustra gráficamente todo este proceso.

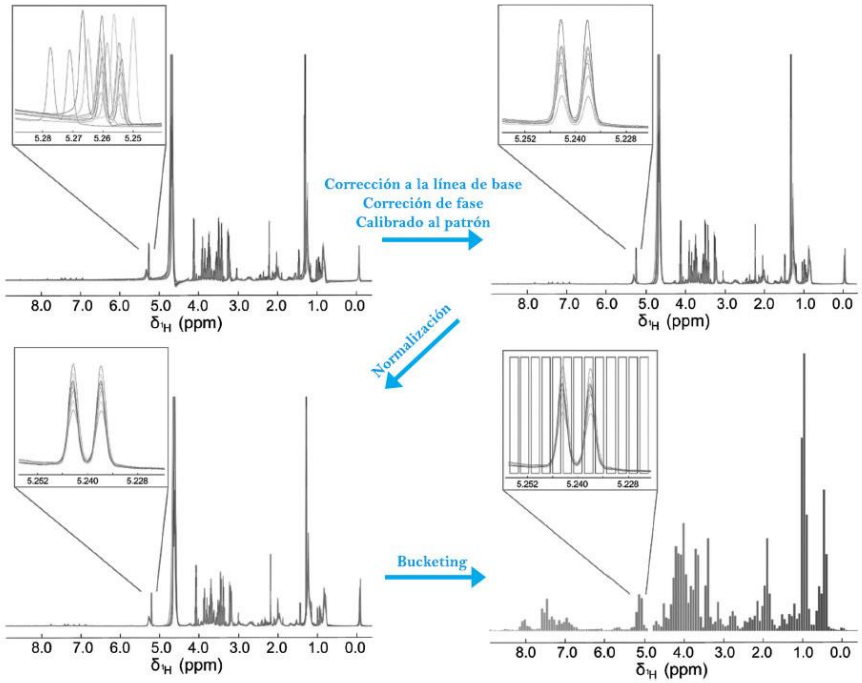



**Figura 4.** Proceso de análisis de un espectro de RMN hasta su transformación en *buckets*. Adaptado de la **referencia 19.**

Una vez obtenidos los datos de interés en la denominada *bucket table*, y en la cual las filas corresponden a las muestras del estudio y las columnas a las variables, se lleva a cabo el análisis estadístico multivariante de los datos. Para ello hay que tener en mente el objetivo de estudio, que puede ser (a) visualización de las diferencias generales entre las muestras, como tendencias o correlaciones, (b) detección de diferencias significativas entre grupos, (c) resaltar las zonas espectrales que contribuyen en mayoría a estas diferencias y (d) la construcción de un modelo predictivo para la correcta clasificación de nuevas muestras.

Estas técnicas de análisis multivariante se dividen en (a) *métodos no supervisados*, empleados para resumir, explorar y descubrir posibles agrupaciones de los datos sin conocer las agrupaciones de los mismos, siendo algunas de estas técnicas PCA, *Independent Component Analysis* (ICA), *k-means* (KM) y *Partition Around Medoids* (PAM), y (b) *métodos supervisados*, que emplean datos conocidos de las muestras con el objetivo de generar modelos que estudien los efectos de interés, y lograr clasificar nuevas muestras, como por ejemplo las técnicas PLS-DA, OPLS-DA, *k-Nearest Neighbours* (k-NN) y ANN.

Una vez aplicados, deben ser correctamente validados para evitar riesgos tales como el sobreajuste mediante técnicas como la *validación cruzada* o el *bootstrapping*. Además, existen métodos tales como las curvas *Receiver Operating Characteristic* (ROC), con las cuales se controla la proporción de falsos positivos generados en el modelo.

El proceso de *profiling*, es decir, de identificación de los metabolitos, en espectros unidimensionales de RMN se lleva a cabo mediante la asignación directa empleando multiplicidades y desplazamientos químicos y con la ayuda de bases de datos tales como la *Human Metabolome Database* (HMDB) y múltiples herramientas disponibles, tales como el *software Chenomx*, y algunos paquetes disponibles para *R* tales como BATMAN o ASICS [17, 18]. Por otro lado, además es necesaria la confirmación de las asignaciones realizadas empleando diferentes *espectros bidimensionales* de tipo homonuclear tales como $^{1}$H,$^{1}$H-COSY o $^{1}$H,$^{1}$H-

TOCSY, y de tipo heteronuclear tales como $^{1}H,^{13}C$-HMQC, $^{1}H,^{13}C$-HSQC, $^{1}H,^{13}C$-HMBC, $^{1}H,^{15}N$-HMQC/HMBC y $^{1}H,^{31}P$-HMQC/HMBC, en los cuales se obtiene información aún más detallada de la estructura de los metabolitos de interés.

Tras aplicar los distintos métodos multivariantes y llevada a cabo la asignación de metabolitos, se aplican diferentes *test de hipótesis* con el objetivo de identificar metabolitos regulados diferencialmente a los que se denomina biomarcadores de condición de estudio. Para ello, en el caso de datos que siguen una distribución normal, son empleados el *t-test* (o *t de Student*), y en el caso de tener más de dos condiciones a comparar, el método de *Análisis de la Varianza* (ANOVA). Además, el método de *Benjamini y Hochberg* es empleado con el propósito de controlar la proporción de falsos positivos, y el *test* de *Bonferroni* para controlar la proporción de error general [20, 21].

### 3.2.5. Interpretación biológica

El objetivo final del proceso de análisis metabolómico es la correcta interpretación de los resultados obtenidos en el análisis estadístico mediante el reconocimiento de las rutas metabólicas y comportamientos de los diferentes metabolitos y biomarcadores identificados. La metabolómica mediante RMN ha expandido por completo la comprensión del metabolismo celular y fisiológico, ayudando a la identificación de múltiples asociaciones bioquímicas inesperadas en distintas condiciones y enfermedades.

Existen múltiples bases de datos, tales como KEG, *Pathway Database*, y MSEA, que recogen diferentes rutas metabólicas y sus metabolitos involucrados. Además, también se puede hacer uso de herramientas online que ayudan al análisis y comprensión de los datos tales como *MetaboAnalyst*, que es una herramienta de análisis de datos provenientes de RMN y MS que examina los metabolitos presentes en la matriz biológica, proporcionando las posibles rutas metabólicas implicadas y, por lo tanto, ayudando en el análisis de la importancia biológica de los resultados.

Además, para obtener un resultado del análisis completo con un mayor fundamento y conocimiento de la temática, es aconsejable consultar publicaciones anteriores que hayan podido aportar información relevante sobre el tema en cuestión [19].

## 3.3. Metabolómica en el estudio del cáncer de colon: Revisión bibliográfica

Con el objetivo de obtener una base sólida para el estudio que se llevará a cabo en este TFM, en esta sección se realizará una breve revisión bibliográfica sobre las principales investigaciones dedicadas al estudio de los cambios metabólicos producidos en muestras de suero de cáncer colorrectal fundamentalmente involucrando a la RMN como plataforma analítica en el estudio metabolómico, con el objetivo de determinar posibles biomarcadores de esta enfermedad.

Así, refinando la búsqueda realizada en el **apartado 3.1.2.** en *Web Of Science Core Collection* mediante las palabras clave "*nmr*" or "*nuclear magnetic resonance*", "*metabolomics*" or "*metabonomics*", or "*metabolite*" or "*metabolic*", y "*colorectal cancer*" or "*colon cancer*" or "*colorectal cancer*" or "*colon cancer*" or "*colorectum cancer*", se obtuvo un total de 376 publicaciones, eliminando aquellos documentos de tipo *Early Access*, que fueron organizadas en función de su año de publicación desde el 2001 hasta la actualidad (mayo de 2021) en la **Figura 5**, en la cual vuelve a verse reflejada esta tendencia a emplear RMN en el campo de la metabolómica.

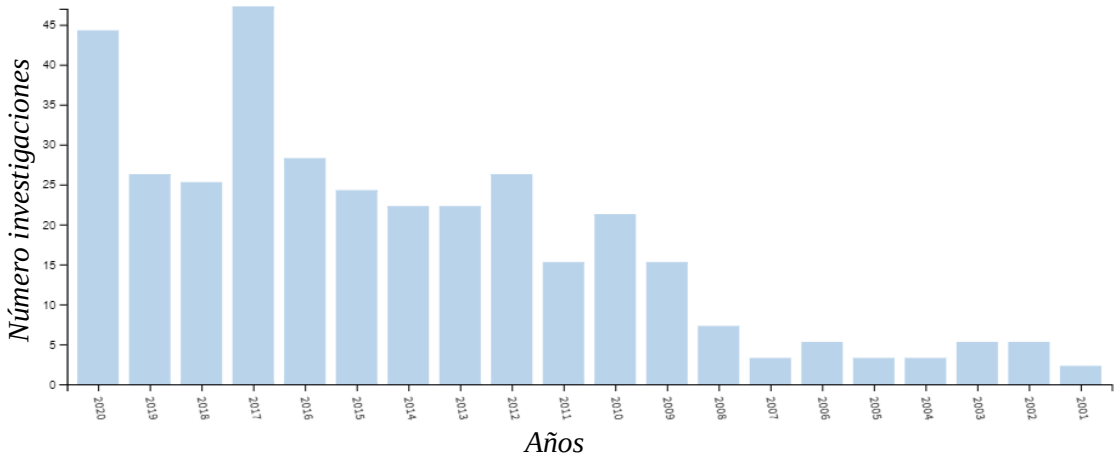



**Figura 5.** Avance de las investigaciones centradas en metabolómica y RMN desde el año 2001 hasta la actualidad. Obtenida de *ISI Web of Knowledge* mediante la búsqueda de las palabras "*nmr*" or "*nuclear magnetic resonance*", "*metabolomics*" or "*metabonomics*", or "*metabolite*" or "*metabolic*", y "*colorectal cancer*" or "*colon cancer*" or "*colorectal cancer*" or "*colon cancer*" or "*colorectum cancer*".

Además, al organizar los resultados en la misma página web en función del campo de aplicación, como puede observarse en la **Figura 6**, se aprecia fácilmente cómo ha sido de gran relevancia, concretamente en el campo de la investigación sanitaria, y más concretamente, en tercer lugar con 55 contribuciones, en el de la oncología.


**Figura 6.** Diagrama de bloques según el área de investigación, obtenido de *ISI Web Of Knowledge* introduciendo las palabras clave "*nmr*" or "*nuclear magnetic resonance*", "*metabolomics*" or "*metabonomics*", or "*metabolite*" or "*metabolic*", y "*colorectal cancer*" or "*colon cancer*" or "*colorectal cancer*" or "*colon cancer*" or "*colorectum cancer*".

Se llevó a cabo un análisis de estas palabras clave empleando el software *CitNetExplorer* [22], con el cual se obtuvo una red de citaciones en base a su relevancia, organizadas en el eje vertical por año de publicación. Esta red puede observarse en la **Figura 7**, en la cual destaca la formación de dos agrupaciones centrales, en función de los temas de los artículos contenidos en cada una. En la principal, posicionada a la izquierda de la imagen, se centraron todas las publicaciones asociadas con el análisis metabolómico del cáncer de colon u otros tipos de cáncer, destacando la

convergencia de todas ellas en una única publicación de 1956. En general, se observó una gran cantidad de estudios centrados en el análisis metabolómico de tejido tumoral y de fluidos tales como suero u orina.

Aunque todas estas publicaciones tuvieron en común una misma temática, mostraron una gran variedad de objetivos, ya que en función de las variables de estudio, en general se enfocaban en la obtención de modelos de clasificación y en la búsqueda de biomarcadores de (1) cáncer colorrectal, empleando muestras de esta enfermedad, muestras de tipo control y muestras positivas en pólipos, (2) cáncer de colon en etapa de metástasis, empleando muestras correspondientes a este grupo, y muestras positivas en cáncer en otros momentos de la enfermedad, y (3) de cáncer de colon en una etapa temprana, empleando muestras en ese estadío y muestras correspondientes a otras etapas del cáncer.

Como se ha mencionado anteriormente, en la **Figura 7** se han resaltado en color rosa aquellas publicaciones que se centran en el análisis metabolómico de fluidos biológicos tales como suero u orina procedentes de pacientes de cáncer colorrectal, con un total de 20 de 100.

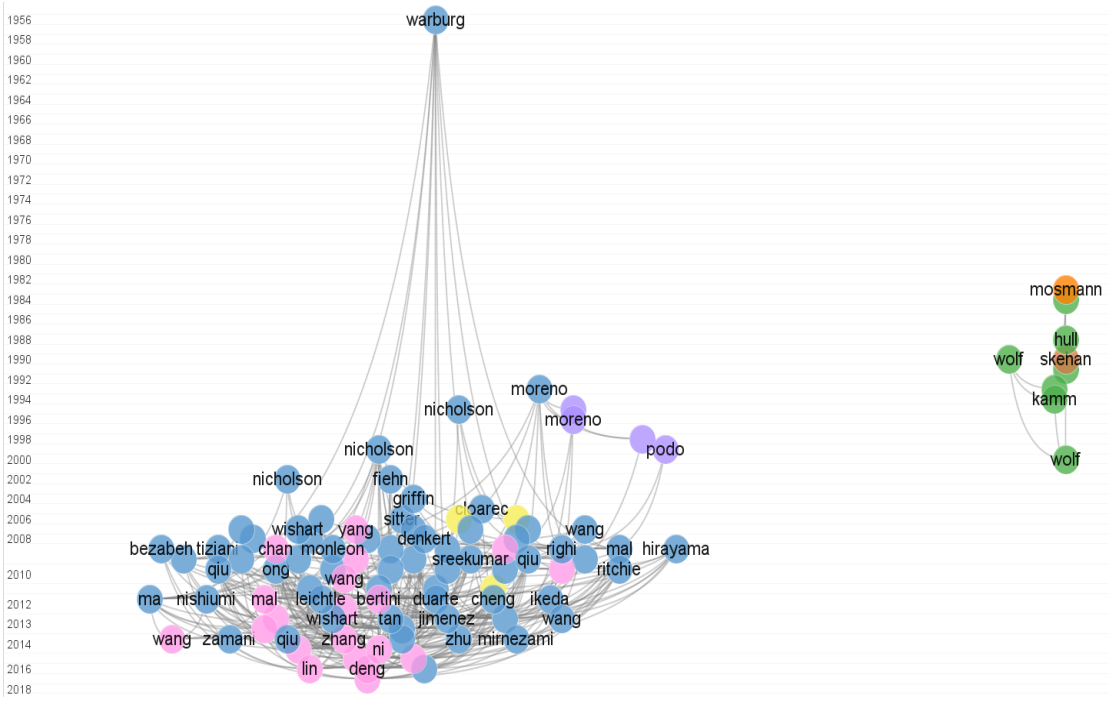



**Figura 7**. Red de citaciones obtenida mediante el software *CitNetExplorer*.

Además, empleando el software *VosViewer* [23], se llevó a cabo un análisis de las palabras clave más empleadas en los diferentes resúmenes (*abstracts*) de estas publicaciones, como se refleja en la **Figura 8**. Se han marcado en naranja las palabras procedentes de los estudios relacionados con *Análisis Multivariante* y *Machine Learning*, destacando los términos *PCA*, *OPLS-DA, Biomarker discovery* y *AUC*.

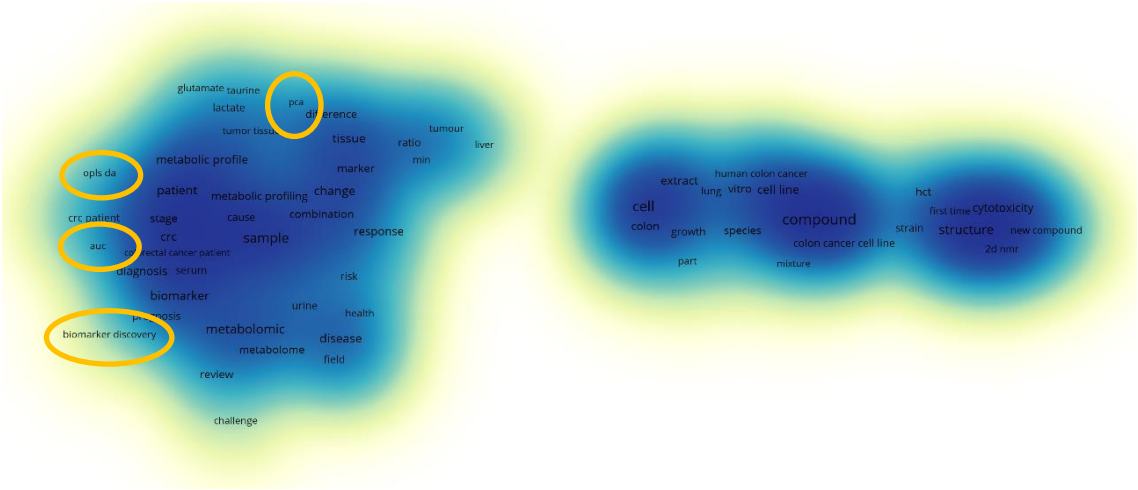



**Figura 8.** Mapa de calor obtenido mediante el software *VosViewer* de las palabras clave más empleadas en los resúmenes (*abstracts*) de las publicaciones.

Seguidamente, se llevó a cabo una breve revisión bibliográfica más centrada en el contenido de las publicaciones relacionadas con el tema de estudio en este TFM, es decir, en el análisis metabolómico, mayoritariamente mediante RMN, de muestras provenientes de suero de cáncer colorrectal. Todos estos artículos se han resumido en la **Tabla A1 del anexo A** destacando lo más importante de cada uno de ellos.

El estudio de Ludwig *et al*. (2009) [24] empleó RMN de $^{1}$H para el análisis metabolómico de 38 muestras de suero positivas en cáncer de colon, 16 de tipo control y 8 positivas en adenoma. Se aplicaron PCA y PLS-DA para la obtención de modelos de clasificación, en los cuales se obtuvo un resultado positivo. Además, se concluyó que las muestras cancerosas presentaban mayores niveles de lactato y de piruvato que el resto. En este trabajo solventan el hecho de emplear correlaciones de tipo $^{1}$H,$^{1}$H-TOCSY de larga duración con la sustitución de la transformada de Fourier por la transformada de *Hadamard*, que se encuentra dentro de la plétora de experimentos denominados rápidos o del inglés *fast NMR*.

Qiu *et al.* (2009) [25] llevó a cabo un análisis metabolómico de suero proveniente de 64 pacientes positivos en cáncer colorrectal y de 65 personas de un grupo control empleando LC-MS y GC-MS-TOF. Se generaron modelos mediante PCA y OPLS-DA que mostraron una alta discriminación entre ambos grupos, y se determinó que el grupo de cáncer, al igual que en el estudio de Ludwig *et al*., presentó un aumento del piruvato y de lactato, acompañado por una disminución de triptófano, tirosina, uridina y oleamida, siendo este último el metabolito más consumido, y cuya disminución se evidenció de forma más significativa en las muestras cancerosas.

Seguidamente, en 2014, Cross *et al*. [26] trataron de obtener nuevos biomarcadores de cáncer colorrectal empleando UPLC-MS y GC-MS, analizando un total de 254 muestras de suero de pacientes con cáncer de colon y 254 de tipo control mediante pruebas estadísticas. Para ello fue empleada la *prueba exacta de Fisher* para las variables categóricas, y el *test* de *Wilcoxon* para las continuas. Seguidamente fue empleada la *corrección de Bonferroni* para la detección de metabolitos significantes, y se comprobó el poder de predicción del set de datos mediante un modelo RF. Aunque obtuvieron algunas posibles asociaciones con metabolitos concretos, ninguno fue estadísticamente significante a excepción del glicoquenodesoxicolato de sodio, que fue positivamente asociado con el grupo de mujeres con cáncer colorrectal.

En ese mismo año, Zamani *et al.* [27] llevaron a cabo un estudio metabolómico de 33 muestras de suero correspondientes a un grupo positivo en cáncer colorrectal y 33 muestras de tipo control empleando RMN de $^{1}H$ con el objetivo de obtener un modelo de predicción y posibles biomarcadores empleando los modelos PCA y PLS-DA. Como resultado, se obtuvo una discriminación positiva entre ambos grupos, y se identificó en el grupo de cáncer una disminución de los niveles de piridoxina, orotidina, s-adenosilhomocisteína, piridoxamina, ácido glicocólico, *β*-leucina, 5-metilcitidina, ácido taurocólico, ácido 3-hidroxibutírico, 7-acetocolesterol, ácido 3-hidroxiisovalerico, l-fucosa, colesterol y L-palmitoilcarnitina, además de un aumento de glicina. Además, destacó la proporción de ácido litocólico (LCA)/ácido desoxicólico (DCA) como posible biomarcador de cáncer de colon.

Dos años después, en 2016, el grupo de Deng *et al.* [28] llevó a cabo un estudio metabolómico de suero de cáncer colorrectal empleando LC-MS y RMN de $^{1}H$ con objeto de examinar el poder de estas dos técnicas para la identificación de biomarcadores. Para ello, emplearon un grupo positivo en cáncer colorrectal de 28 sujetos, un total de 44 individuos con pólipos y otro grupo de 55 controles. Generaron un algoritmo con el cual se examinaron todas las variables y a cada iteración se eliminó una en función de su precisión de predicción (*Backward Variable Elimination*) junto con la *validación cruzada Monte Carlo* (MCCV) en combinación con PLS-DA. Como resultado principal se obtuvo que en el grupo correspondiente a cáncer de colon se observaron mayores niveles de glucosa, menores de adenosina, y alteraciones en los niveles de piruvato y de glutamina. Por otro lado, en el grupo positivo en pólipos se encontró un descenso de orotato y un aumento de adenosina. Se observaron alteraciones en los niveles de aminoácidos, fumarato, citrato, oxaloacetato, ácido linoléico y de lípidos tanto para el grupo de cáncer como para el de pólipos en comparación al control.

En la investigación llevada a cabo en 2019 por Gu *et al.* [8] se realizó el análisis metabolómico de 40 muestras provenientes de suero de cáncer de colon, 32 muestras positivas en pólipos y 38 de tipo control empleando RMN de $^{1}H$ con el propósito de generar modelos de discriminación e identificar posibles biomarcadores. Con este objetivo se emplearon los métodos PCA, PLS-DA, OPLS-DA, RF y SVM, consiguiendo identificar un total de 23 metabolitos, y determinando que la proporción acetato/glicerol podría ser un biomarcador de presencia de pólipos, y que lactato/citrato podría serlo de cáncer colorrectal.

En general se ha comprobado que los metabolitos más relevantes de cada estudio están ampliamente relacionados con el metabolismo de carbohidratos, en concreto con la glucólisis, en la cual se observa un consumo de glucosa y liberación de lactato. También se observaron en general alteraciones en el metabolismo de aminoácidos (adenosina, tirosina, alanina, histidina, triptófano, etc.), relacionado con la actividad cancerígena, y alteraciones en algunos metabolitos tales como el citrato, fumarato y oxaloacetato, posiblemente derivados del ciclo del ácido cítrico. En el estudio de Deng *et al*. además, se hace referencia a la biosíntesis de ácidos biliares primarios, el metabolismo de la vitamina B6 y la síntesis y degradación de cuerpos cetónicos, entre otros.

Estos estudios se centraron en el uso de PCA como técnica no supervisada de discriminación entre grupos y de detección de *outliers*, y en gran mayoría emplearon técnicas multivariantes lineales supervisadas como PLS-DA y OPLS-DA, a excepción de los estudios de Cross *et al*. [26], y de Gu *et al.* [8] en los que se emplearon técnicas no lineales para el desarrollo de modelos de clasificación, RF y SVM respectivamente. Con el objetivo de seleccionar las variables que contribuyeron más a los modelos, dos de estos estudios (Qiu *et al*.[25] y Gu *et al*.[8]) emplearon el método VIP en combinación a los métodos lineales empleados, que determina que aquellas

variables con un valor VIP superior a 1 son estadísticamente significantes para el modelo y causantes de la discriminación en mayor medida. Por otro lado, para el asesoramiento de aquellos metabolitos más discriminantes en los procedimientos no lineales de estas investigaciones, en primer lugar se llevó a cabo el análisis de los metabolitos más diferenciales, bien mediante distintos tests estadísticos, como en el estudio de Cross *et al.*, o bien empleando los resultados de VIP obtenidos mediante los métodos lineales, para posteriormente examinar la frecuencia de selección de las variables por parte del algoritmo en función de múltiples modelos generados según el *Área Bajo la Curva* (AUC) ROC, como puede observarse en el estudio de Gu *et al*.[8].

Con el objetivo de validar los modelos lineales PLS-DA y OPLS-DA, varios de los estudios coinciden en emplear el test paramétrico de validación cruzada CV-ANOVA, mientras que la precisión de los modelos no lineales fue evaluada empleando curvas ROC (AUC) con las cuales se logra comprobar la proporción de falsos positivos derivados de matrices de confusión, en combinación con sus respectivos *intervalos de confianza* (IC). Por último, la mayoría de los estudios realizaron análisis de rutas metabolómicas empleando la corrección de *Holm-Bonferroni*, con el objetivo de determinar aquellas más enriquecidas.

Por otro lado, aunque la mayoría de estos estudios emplearon RMN de $^{1}$H como herramienta analítica para la identificación de biomarcadores, no todos se apoyaron únicamente en ella. Sin embargo, debido a la baja proporción de estudios metabolómicos de suero de cáncer colorrectal en la que emplean RMN de $^{1}$H como única plataforma analítica, estos fueron incluidos en la presente revisión. Así, el equipo de Deng *et al.*, además de la RMN empleó LC-MS, mientras que el estudio de Qiu *et al*. utilizó LC-MS y GC-MS-TOF únicamente, y el de Cross *et al*.[26] hizo uso de UPLC-MS y GC-MS. Es de gran importancia remarcar estas diferencias ya que, en función del equipo empleado, pueden detectarse distintos tipos de metabolitos en función de la *sensibilidad* y *especificidad* de cada plataforma.

Otro factor de gran importancia es el referente al procesamiento de los datos, que mostró una gran variación entre todos estos artículos en función del tipo de normalización, escalado y/o transformación empleados. De esta manera, en función de los métodos empleados se priorizan distintos tipos de señales, tal y como puede verse resumido en la **Tabla 3**, lo cual puede llevar a múltiples conclusiones. En general estos artículos presentaron tamaños muestrales bastante reducidos a excepción del de Cross *et al*.[26], por lo que convendría llevar a cabo nuevas investigaciones aumentando dicho tamaño para generar resultados más robustos.

Un aspecto a remarcar es el de que en ninguno de estos estudios se llevó a cabo la caracterización de los ácidos grasos obtenidos, por lo que una posible línea de investigación de alto interés podría ser la búsqueda de correlación entre perfiles metabolómicos empleando RMN con perfiles de ácidos grasos mediante la técnica *cromatografía de gases con detector de ionización de llama* (GC-FID).

# 4. Metodología

## 4.1. Revisión bibliográfica

Llevando a cabo diferentes búsquedas en la base de datos *Web Of Science* [16] mediante palabras clave, y en combinación con los programas *CitNetExplorer* [22] y *VosViewer* [23], se llevó a cabo el análisis bibliográfico de un pequeño conjunto de artículos centrados en la búsqueda de biomarcadores de cáncer de colon empleando metabolómica mediante RMN de $^{1}$H. Se recopiló la información obtenida para obtener una base sólida con objeto de la elaboración de este TFM.

## 4.2. Selección del *dataset*

Para este trabajo, se seleccionó un conjunto de datos cedido por el grupo de investigación *NMRMBC* [29], conteniendo los datos metabolómicos espectrales de 90 muestras correspondientes a un grupo positivo en cáncer de colon (n = 64) y un grupo control (n = 26) analizados mediante RMN de $^{1}$H. Estos espectros fueron automáticamente ajustados en fase, línea base y alineados a la señal de referencia, en este caso correspondiente al patrón interno *ácido 2,2,3,3-$d_4$-(trimetilsilil)propanoico* (TSP). Los espectros obtenidos emplearon la secuencia de pulsos optimizada para la supresión de señales procedentes de macromoléculas, CPMG con supresión de agua, tal y como se ha comentado en el **apartado 3** del presente Trabajo Fin de Máster.

## 4.3. Procesado de los datos

Empleando el *software Amix* [30] (v. 3.9.12, Bruker BioSpin GmbH, Rheinstetten, Alemania) se llevó a cabo el *bucketing* de los espectros en intervalos de 0.04 ppm, normalizando la intensidad de los picos individuales en relación a la intensidad del área total y posteriormente exportando este archivo en formato *.csv*. La región espectral de $\delta_H$ 4.67 a 5.12 ppm fue excluída, ya que corresponde con la región que incluye pequeñas distorsiones espectrales procedentes de la eliminación de la señal de agua.

La *bucket table* anteriormente obtenida fue analizada en primer lugar mediante el lenguaje de programación *R* empleado el *software RStudio* [31] (v. 1.4.1106, PBC, Boston, MA). En primer lugar, se cargaron los paquetes necesarios para llevar a cabo la lectura del *dataset* (*readxl*, *rJava*, *xlsxjars* y *xlsx*), y se definieron los nombres de las columnas correctamente, como puede verse en el resultado obtenido en la **Figura 9**. El código en *R* empleado en esta parte del análisis se encuentra recogido en el **anexo B**.

| Names <chr> | Sucia/Limpia <chr> | Control/Cancer <chr> | 8.58 <dbl> | 8.54 <dbl> | 8.5 <dbl> |
|---|---|---|---|---|---|
| 20201103_SUE_10_L_ | Limpia | Cancer | -1.010e-05 | 2.912e-05 | -3.980e-06 |
| 20201103_SUE_11_L_ | Limpia | Cancer | -7.990e-06 | 1.063e-05 | 7.780e-06 |
| 20201103_SUE_12_L_ | Limpia | Cancer | -1.389e-05 | -2.590e-06 | 6.630e-06 |
| 20201103_SUE_13_L_ | Limpia | Cancer | -1.249e-05 | 1.460e-06 | 1.203e-05 |
| 20201103_SUE_14_S_ | Sucia | Cancer | 1.310e-06 | 7.340e-06 | 9.230e-06 |
| 20201103_SUE_16_S_ | Sucia | Cancer | -6.840e-06 | -2.140e-06 | -1.470e-06 |
| 20201103_SUE_17_L_ | Limpia | Cancer | -1.062e-05 | -3.970e-06 | 8.240e-06 |
| 20201103_SUE_20_S_ | Sucia | Cancer | 4.650e-06 | 4.140e-06 | -5.210e-06 |
| 20201103_SUE_21_L_ | Limpia | Cancer | -3.270e-06 | -2.130e-05 | -1.012e-05 |
| 20201103_SUE_22_S_ | Sucia | Cancer | 3.870e-06 | 5.900e-07 | 3.410e-06 |

**Figura 9.** Fragmento de la *Bucket table* obtenida para el análisis estadístico.

Se guardaron como factores los datos correspondientes a las variables del *dataset*, *Control/Cancer*, y *Limpia/Sucia*, además de definir un objeto con únicamente los valores numéricos del *dataset* (*cancernum1*). La variable *Control/Cancer* correspondió al grupo positivo en cáncer de colon y al grupo de tipo control, y la variable *Sucia/Limpia* clasificó las muestras en función del tipo de muestreo de las mismas: las limpias fueron recopiladas con un producto de una casa comercial, mientras que las sucias fueron tratadas con un producto del propio laboratorio.

Seguidamente, se comprobó la existencia de valores de tipo *Not Available* (NA) y se eliminaron aquellas variables iguales a 0 al corresponderse con señales debidas al ruido del espectro. No fue aplicado ningún escalado en este primer paso de procesado ya que conformó uno de los apartados del análisis estadístico en este caso.

## 4.4. Análisis estadístico

Tras el primer paso de procesado del set de datos, se continuó el análisis llevando a cabo la evaluación de las agrupaciones o *clústers* presentes de forma natural en el set de datos. Esta valoración es posible gracias a los métodos de análisis exploratorio no supervisados, los cuales tratan de encontrar posibles patrones presentes entre los datos suministrados en ausencia de información sobre los grupos.

Como pudo observarse anteriormente en la revisión bibliográfica (**apartado 3.3**), el método no supervisado más empleado en metabolómica es el PCA, que permite el estudio de las agrupaciones presentes de forma natural en el set de datos revelando posibles tendencias en los datos de RMN de $^{1}$H, así como la existencia de otras variables no previstas que puedan estar afectando a la discriminación de los datos y atendiendo también a la posible presencia de valores atípicos. La aplicación del PCA supone la transformación lineal de los datos en componentes principales, logrando visualizarlos a lo largo de múltiples dimensiones en función del total de varianza explicada por cada una, y siendo la primera aquella que contiene la mayor parte de la información útil del conjunto de datos inicial.

Generalmente, este modelo es representado en función de las dos o tres primeras componentes con el objetivo de visualizar los vectores de *scores*, es decir, las muestras de cada componente, y los vectores de *loadings*, los cuales representan las variables del set de datos y sus relaciones.

Con el objetivo de obtener una mayor probabilidad de obtener agrupaciones no supervisadas, fueron empleados algunos de los escalados más comunes en metabolómica, generando una serie de funciones de acuerdo a la información detallada en la **Tabla 3** y en base a la **referencia 32.** Su orden de aplicación fue: (1) *Unit Variance*, (2) *Pareto*, (3) *Range scaling*, y (4) *Vast scaling*. Posteriormente se comprobó la validez de estas funciones representando mediante diagramas de cajas las variables del set de datos empleando la función de *R boxplot()*.

Así, fueron generados modelos PCA para cada uno de los escalados, y para ello fueron empleados dos paquetes provenientes del repositorio *CRAN*, que correspondieron con: (1) *FactoMineR* [33], del que se empleó la función *prcomp()* para el cálculo de los valores propios del PCA, es decir, sus *eigenvalues*, y (2) *factoextra* [34], con el que se llevaron a cabo las representaciones gráficas correspondientes. La función *fviz_pca_ind()* permite la visualización de la gráfica de *loadings*, y la función *fviz_pca_var()* la gráfica de *scores*. Empleando *fviz_contrib() se* lleva a cabo la representación de la contribución de las variables a la dimensión de interés, y utilizando *fviz_screeplot()* se logra representar la varianza explicada para cada dimensión del PCA.

Seguidamente, con el objetivo de optimizar la separación de los grupos *Control* y *Cáncer*, fueron empleados métodos de carácter supervisado. Estos modelos emplean la información disponible sobre los grupos del conjunto de datos responsables de las distintas agrupaciones permitiendo determinar las variables con mayor relevancia para cada uno. Existen múltiples tipos dentro de esta categoría de métodos, sin embargo, en este TFM, fueron seleccionados cuatro en total de acuerdo a las tendencias observadas en los estudios de la revisión bibliográfica.

Por un lado, fue aplicado el algoritmo lineal PLS-DA, que consiste en una regresión por mínimos cuadrados parciales que permite llevar a cabo una separación entre las clases, y seguidamente fue empleado el método OPLS-DA, que se trata de una modificación ortogonal del primero que permite maximizar la varianza entre grupos en una única dimensión. En ambos casos fueron evaluados los distintos escalados anteriormente indicados, incluyendo en este TFM aquellos que aportaron los mejores resultados.

Para la implementación de estos modelos en R, fue empleada la función *opls()* del paquete *ropls* [35] descargado del repositorio *Bioconductor*, que genera el modelo PLS-DA por defecto, y que mediante la opción *orthoI = NA* construye el modelo OPLS-DA aplicando *validación cruzada* de 7 *interacciones* mediante el algoritmo NIPALS para su validación. Esta función genera automáticamente los parámetros de calidad del modelo $R^2$ (porcentaje de variación), que debe aportar un valor cercano a 1 para ser considerado adecuado [36], $Q^2$ (porcentaje de variación de acuerdo a la validación cruzada), cuyo valor debe ser superior a 0.5 para y *Root Mean Square Error of the Estimation* (RMSEE). Por otra parte, empleando el *software SIMCA* (v. 14.0, Umetrics, Suecia), uno de los más populares en esta área, se generó además el valor asociado a *Root Mean Square Error of Cross Validation* (RMSEcv) de cada grupo, se comprobó visualmente la normalidad de los modelos, la existencia de *outliers*, y se determinó el valor CV-ANOVA asociado a cada uno. Seguidamente, se generaron los gráficos de *loadings* correspondientes, permitiendo observar qué metabolitos aumentan o disminuyen para cada grupo.

Una vez seleccionado el modelo lineal más relevante, se procedió a la identificación de las regiones espectrales o *buckets* discriminantes y que tras su identificación estructural se asignan como pertenecientes a metabolitos concretos que denominamos biomarcadores. Para ello, se empleó el método VIP, muy común para este tipo de modelos en estudios metabolómicos. Mediante el programa *SIMCA* [37] *(v. 17.0.0.24543, Sartorius, Goettingen, Alemania)*, se llevó a cabo la selección manual de los VIP superiores a 1, es decir, los más importantes, además de su representación en una gráfica de contribuciones y de su obtención en forma de tabla. Por último, las variables fueron identificadas de acuerdo a la lista de asignaciones facilitada por el grupo de investigación [38].

Por otro lado, fueron implementados los dos algoritmos de naturaleza no lineal recogidos en la revisión bibliográfica. El primero fue *Random Forest*, que consiste en múltiples árboles de decisión que actúan como un ensamblaje, dando cada árbol individual una predicción de clase, de la cual se selecciona la más común. Se trata de un algoritmo de alta versatilidad, de fácil implementación, robustez y sin tendencia al sobreajuste. El otro algoritmo empleado consistió en *Support Vector Machine*, que se apoya en la generación de un hiperplano de máxima separación para lograr la partición de los datos en grupos de clases similares. En el caso de que los datos no sean linealmente separables, es empleado un parámetro llamado *kernel*, que fuerza la separación de los datos. Para generar dichos modelos, fue empleada la herramienta online *MetaboAnalyst* [39] *(MetaboAnalyst v.5.0, Xia Lab, Quebec, Canada)*, que se trata de una plataforma que ofrece múltiples herramientas de interés para la metabolómica, tales como análisis de tipo univariante, multivariante, búsqueda de biomarcadores y análisis de rutas biológicas, basada en su mayoría en funciones de R. Actualmente, múltiples estudios de esta clase emplean dicha herramienta gracias a su facilidad de uso, y, de hecho, existe un paquete disponible para *R* del mismo desarrollador llamado *MetaboAnalystR* [40], aunque en el caso de este TFM fue empleada la herramienta *online*.

Las muestras de tipo Sucia fueron descartadas del estudio estadístico, y fue seleccionada la opción *Biomarker Analysis*, que proporcionó varios tipos de estudios: (1) análisis univariante mediante curvas ROC, (2) análisis exploratorio multivariante mediante curvas ROC, y (3) Evaluación del modelo en base a curvas ROC.

La opción preferente fue el análisis exploratorio multivariante, que permite seleccionar tanto el método de clasificación, como el de la obtención de las variables más relevantes para cada modelo. En primer lugar, fue aplicado el método RF mediante distintos escalados, seleccionando aquél con mejores resultados para este trabajo, y utilizado como método de selección de variables

la opción *RF built-in*, que discrimina en función de los pesos de las mismas en el modelo RF. Los resultados obtenidos correspondieron a gráficas de curvas ROC, obtenidas mediante el método MCCV empleando un submuestreo balanceado. En cada MCCV, dos tercios de las muestras fueron empleadas para la evaluación de la importancia de las variables o *features*, las cuales son seleccionadas en la construcción de los modelos de clasificación. Dichos modelos son seguidamente validados empleando el tercio restante con el objetivo de evitar un posible sobreajuste u *overfitting*.

A continuación, fue aplicado el método SVM de *kernel* lineal mediante distintos tipos de escalados, seleccionando aquél con mejores resultados para este trabajo, e indicando en este caso la opción *SVM built-in* para la selección de los *buckets* más relevantes para el modelo, que lleva a cabo esta discriminación en función de los pesos de las variables en el modelo SVM. Se repitió el mismo procedimiento que con el análisis anterior.

En función de los valores del AUC ROC, el IC, y la precisión de la predicción, uno de estos modelos fue seleccionado como el más óptimo. La métrica AUC aporta la posibilidad de que un modelo clasifique una muestra positiva como negativa, siendo el valor AUC= 1 el correspondiente a un clasificador perfecto, mientras que un valor AUC= 0.7 se trata del valor mínimo que debe adoptar un modelo para ser clínicamente útil [41]. Los biomarcadores asociados fueron identificados mediante la tabla de asignaciones proporcionada por el grupo de investigación [38], y seguidamente fueron evaluados mediante la opción de análisis univariante con el objetivo de determinar aquellos más estadísticamente significantes para el modelo mediante sus valores AUC, que indican la utilidad como biomarcadores de cada metabolito, acompañados de sus IC al 95% correspondientes, calculados empleando un método de 500 *bootstraps*. Un valor AUC entre 0.9-1.0 determina un biomarcador notable, entre 0.8-0.9 resulta muy adecuado, entre 0.7-0.8 origina uno adecuado, entre 0.6-0.5 se trata de uno intermedio, y entre 0.5-0.6 es inadecuado. Por otro lado, un IC asociado al valor AUC de biomarcadores identificados de entre 0.9-0.7 resulta un rango muy adecuado, en el cual el valor AUC asociado se encontrará el 95% de las veces [42].

## 4.5. Interpretación biológica

Con el objetivo de llevar a cabo un análisis de las rutas biológicas más relevantes asociadas a los biomarcadores obtenidos en ambos modelos, se empleó la opción *Pathway Analysis* del programa *MetaboAnalyst*, que proporciona información sobre las rutas metabólicas en donde los metabolitos introducidos se encuentran involucrados, empleando bases de datos tales como KEGG y/o *Small Molecule Pathway Database* (SMPDB). Para ello fue empleada la librería de rutas metabólicas relacionadas con *Homo Sapiens*. Mediante esta herramienta pudieron obtenerse también valores de impacto de cada ruta, y valores asociados a *False Discovery Rate* (FDR), valores *p*, y valores *p* de Holm (ajustados mediante el método de *Holm-Bonferroni*). Las rutas son consideradas enriquecidas significantemente si el valor *p* de *Holm*< 0.05, FDR<0.05, el valor de *impacto*>0, y el *número de metabolitos* de cada ruta>1.

Finalmente, empleando la bibliografía consultada, se interpretaron biológicamente dichos resultados, contrastando los resultados obtenidos.

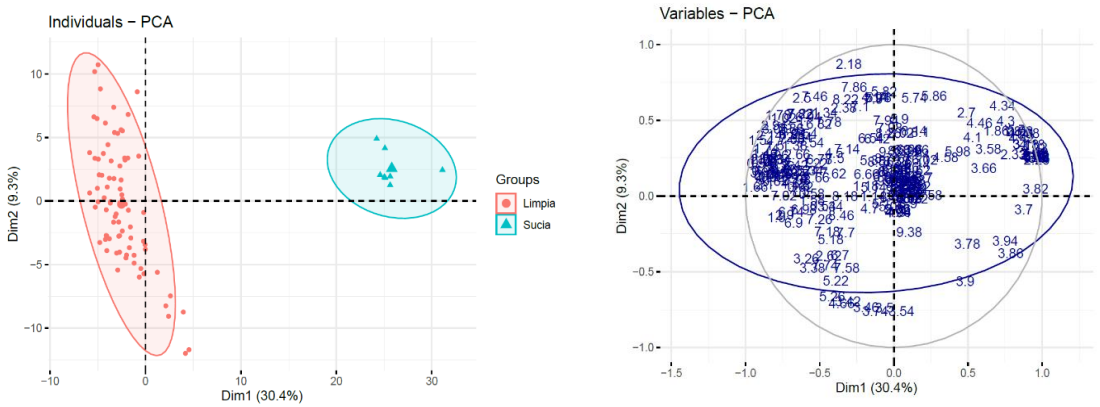
# 5. Resultados

## 5.1. Aplicación de métodos lineales

Este TFM se centra en el estudio del perfil metabólico de muestras de suero de cáncer colorrectal respecto al obtenido de muestras control. Para ello fueron empleadas técnicas de análisis multivariante de datos tanto lineales como no lineales, las cuales permiten la correlación de los biomarcadores identificados por RMN con el diagnóstico de cáncer en las muestras.

Tras la adquisición de los espectros se generó una tabla de *bucketizado*, y seguidamente se llevó a cabo su tratamiento, de acuerdo al procedimiento descrito en la sección de metodología. Las variables incluidas en este trabajo recopilan información sobre el tratamiento de las muestras y la presencia o no de cáncer, sin embargo, se excluyeron otras variables tales como el estado del cáncer, la edad del paciente o su sexo con el objetivo de no vulnerar la confidencialidad de los individuos participantes.

De esta forma, se realizó en primer lugar un PCA mediante escalado *Unit Variance* para obtener una visión general de la varianza metabólica del conjunto de datos de RMN. En la **Figura 10a**, queda representado el gráfico de *scores* correspondiente a dicho modelo, el cual considera las dos primeras componentes principales (PC1 y PC2), revelando un 30.4% y un 9.3% de la varianza total, respectivamente. Puede observarse que existe una discriminación evidente de las muestras de tipo *Limpia* y *Sucia*, que puede verse evidenciada de acuerdo a la representación de tipo *loadings* de la **Figura 10b**, en la cual se observa un aumento de las regiones espectrales centradas en $\delta_H$ 4.34, 2.70, 4.46 y 4.30 ppm, entre otros.




**a)** **b)**

**Figura 10.** Gráfico PCA (PC1/PC2) de (**a**) *scores* y de (**b**) *loadings* obtenido a partir de espectros de RMN de $^1$H de muestras de suero de cáncer colorrectal (modelo escalado con *Unit Variance*). Es posible observar una discriminación evidente de las muestras en el gráfico de *scores* de acuerdo con la variable *Sucia/Limpia.*

Al consultar los espectros de RMN asociados a muestras de tipo *Limpia* y de tipo *Sucia*, se comprobó que mostraban señales diferentes de sus metabolitos, que quedan plasmadas en la

**Figura 11**. Estas regiones espectrales marcadas en azul presentes en las muestras de tipo *Sucia* son los causantes del *clustering* observado.

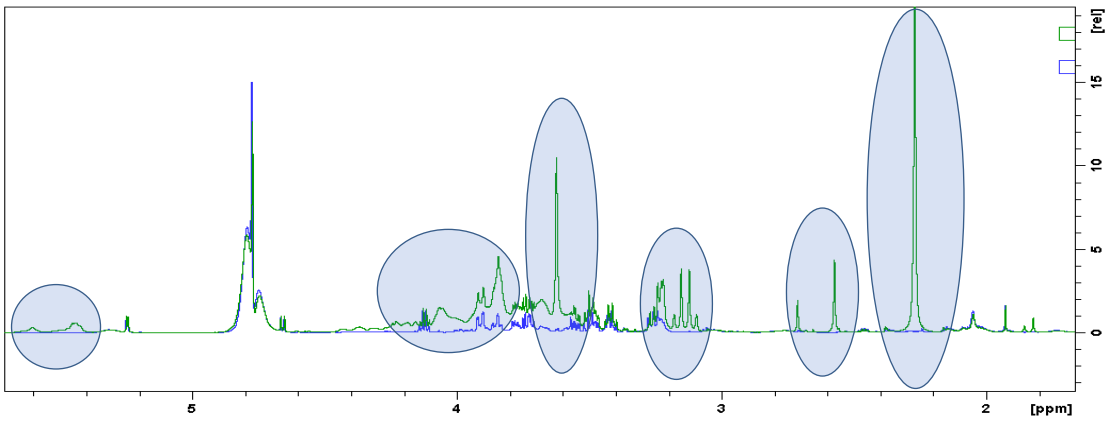
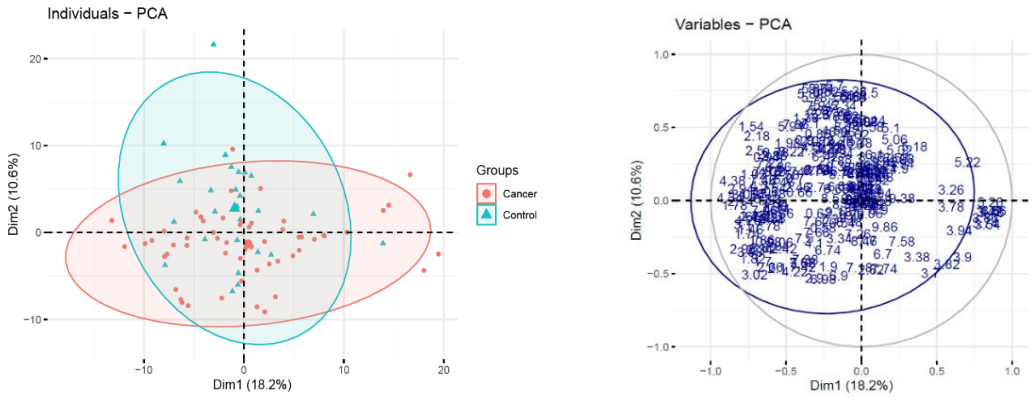



**Figura 11.** Señales en espectros de RMN de muestras de tipo *Limpia*, en azul, y *Sucia*, en verde.

Contrastando las evidencias espectrales observadas con la información obtenida y disponible sobre las muestras, se procedió a la eliminación de las muestras de tipo *Sucia* del conjunto de datos, con el objetivo de que los resultados obtenidos no quedasen sesgados.

Seguidamente, se procedió a una nueva representación de *scores* de PCA (**Figura 12**), esta vez teniendo en cuenta únicamente las muestras de tipo *Limpia* de acuerdo con la variable *Cancer/Control*, y empleando de nuevo el escalado de tipo *Unit Variance*. Como puede apreciarse, estas dos componentes principales revelan un 18.2% y un 10.6% de la varianza total, y es posible observar que la mayoría de las muestras se agrupan en una única región central en donde ambos clústeres (azul y rojo) solapan con la excepción de algunos valores atípicos (*outliers*).




**a)** **b)**

**Figura 12**. Gráfico PCA (PC1/PC2) de (**a**) *scores* y de (**b**) *loadings* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Unit Variance*). Las muestras quedan agrupadas en su gran mayoría en el centro de la gráfica, sin observarse una discriminación clara.

Además, fueron también obtenidos los gráficos correspondientes a la contribución de las variables a la dimensión 1 (es decir, a la componente PC1) y al porcentaje de varianza explicada por cada componente, que fueron incluidas en la **Figura A2 del anexo A.**

Como puede comprobarse, no se obtuvo una agrupación adecuada de las muestras de *Cáncer* y *Control* mediante este método no supervisado empleando el escalado de tipo *Unit Variance*, por lo que se procedió a examinar los resultados de aplicar otros como parte de un paso de posible mejora de los resultados. Para ello se empleó el escalado *Pareto*, cuyo resultado se muestra en la **Figura A3a del anexo A** junto a su gráfico de *loadings* (**Figura A3b del anexo A**). Este nuevo modelo explica la varianza total en un 53.2%, en el caso de PC1, y un 19.3% en el caso de PC2, y de nuevo muestra una disposición de muestras en su parte central y una serie de *outliers*. Sin embargo, estos resultados no supusieron una mejora con respecto a los resultados de la **Figura 12**, ya que no dieron ningún tipo de agrupación discriminante.

A continuación, se comprobaron los resultados arrojados por el escalado de tipo *Range Scaling*, de nuevo mediante los correspondientes gráficos de *scores* y de *loadings* (**Figura A4 del anexo A**). En esta ocasión, la varianza total explicada por PC1 y PC2 fue de un 58.9% y de un 19.4%, respectivamente, aunque, de nuevo, modelo no permitió distinguir entre agrupaciones.

Por último, se evaluó el escalado de tipo *Vast Scaling*, cuyas representaciones de *scores* y de *loadings* quedan recogidas en la **Figura A5 del anexo A**. El PC1 explicó un 44.7% de la varianza total, mientras que el PC2 un 22.2%. Sin embargo, una vez más, no se observó ningún tipo de agrupación discriminante clara. En ninguno de los escalados empleados se observó una mínima discriminación entre los grupos *Cáncer* y *Control*, aunque esto se trata de un hecho frecuente en el campo de la metabolómica, sobre todo en el área de las muestras clínicas, las cuales presentan una alta variabilidad intrínseca debida a las múltiples variables biológicas que les afectan, tales como la dieta, la presencia de enfermedades, o la edad. Así, en estos casos, los modelos no supervisados no suelen revelar agrupaciones altamente diferenciadas [43].

Por esta razón, se generaron modelos de análisis supervisado con el objetivo de forzar esta discriminación y posteriormente lograr obtener posibles biomarcadores de cáncer colorrectal. Se aplicaron, por un lado, dos de los métodos lineales más empleados en metabolómica actualmente, siendo en primer lugar el modelo PLS-DA escalado a *Unit Variance*, escalado junto al cual se obtuvieron los parámetros más adecuados, cuyo gráfico de *scores* queda recogida en la **Figura 13**.

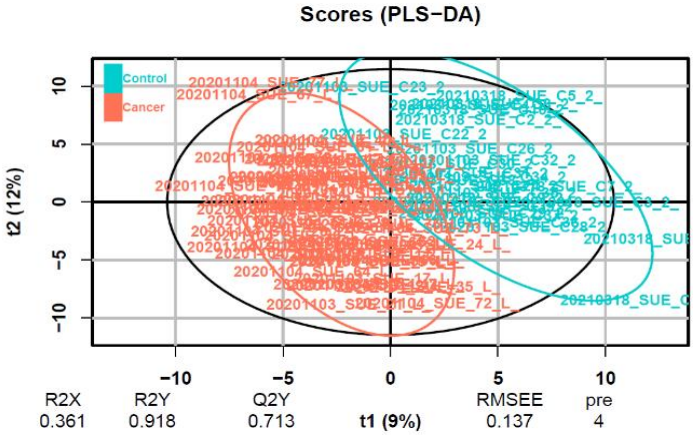



**Figura 13.** Gráfico PLS-DA de *scores* obtenido a partir de espectros de RMN de $^{1}H$ de muestras de suero de cáncer colorrectal (modelo escalado con *Unit Variance*). Puede observarse una discriminación entre las muestras de tipo *Cáncer* y *Control*. Los parámetros de calidad del modelo fueron $R^2X = 0.361$, $R^2Y = 0.918$, $Q^2Y = 0.713$, y RMSEE = 0.137.

Los resultados obtenidos del modelo fueron adecuados en general, con un valor $R^2$ muy cercano a 1 en el caso de su componente Y, y el valor asociado a $Q^2$ superior a 0.5, lo que indica una buena capacidad predictiva. El valor de RMSEE fue de 0.137, que en el caso del área de la metabolómica es un buen valor, y el de RMSEcv, observado en la **Figura A6 del anexo A** fue inferior a 0.5, por lo que se determinó que el modelo es capaz de predecir los datos en buena fiabilidad.

Seguidamente, se comprobó la normalidad de los residuos del modelo mediante la gráfica de la **Figura A7 del anexo A**. Se observó que este modelo PLS-DA sí presenta una distribución normal de sus residuos y que además su valor CV-ANOVA fue muy inferior a 0.05 ($1.66\times10^{-18}$), por lo que se trata por tanto de un modelo muy válido.

Además, el gráfico de *loadings* correspondiente a este modelo (**Figura A8 del anexo A**), permitió deducir un aumento de intensidad de las regiones espectrales centradas a $\delta_H$ 8.46, 4.1, 1.94, 2.38 ppm, entre otros, y en cambio, el grupo *Control* presenta un aumento de intensidad, por ejemplo, en las regiones espectrales a $\delta_H$ 2.58, 2.54, 6.78, y 3.22 ppm.

A continuación, los parámetros obtenidos mediante el modelo PLS-DA se trataron de mejorar empleando un segundo modelo lineal, OPLS-DA, que en general suele aportar valores más adecuados en metabolómica mediante la aplicación del escalado de tipo *Pareto*, junto al cual se obtuvieron los parámetros más adecuados. La **Figura 14** recoge el gráfico de *scores* en donde las muestras con diagnóstico positivo quedan en el clúster de la izquierda marcado en color rojo, mientras que las muestras control quedan a la derecha del eje PC2 marcadas en azul.

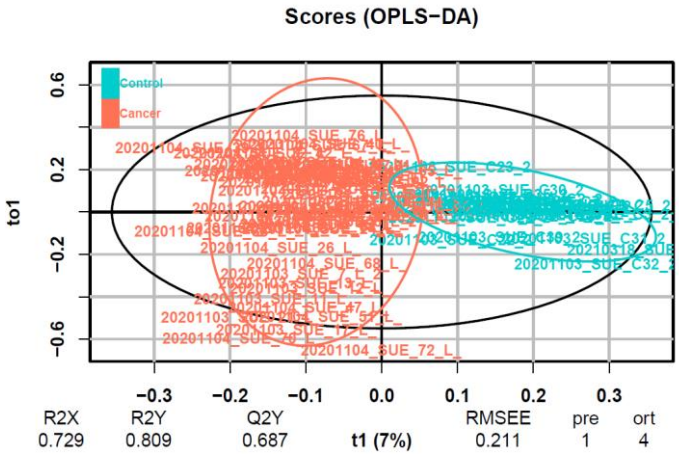



**Figura 14.** Gráfico OPLS-DA de *scores* obtenido a partir de espectros de RMN de $^1$H de muestras de suero de cáncer colorrectal (modelo escalado con *Pareto*). Puede observarse una discriminación entre las muestras de tipo *Cáncer* y *Control*. Los parámetros de calidad del modelo fueron $R^2X = 0.729$, $R^2Y = 0.809$, $Q^2Y = 0.687$, y RMSEE = 0.211.

De nuevo, se obtuvieron unos parámetros de regresión lineal ($R^2X$ y $R^2Y$) y capacidad predictiva ($Q^2Y$) adecuados, en donde el valor $R^2$ obtenido es del orden de 0.8 y el valor asociado a $Q^2$ resulta ser superior a 0.5. El valor de RMSEE de 0.211 aumenta con respecto al obtenido en el anterior modelo, pero sigue siendo aceptable, en relación con estudios similares de metabolómica basados en RMN. Por otro lado, el valor asociado a RMSEcv, mostrado en la **Figura A9 del anexo A**, aunque fue inferior a 0.5 volvió a aumentar, y al comprobar la normalidad del modelo en la **Figura A10 del anexo A**, se observó una leve disminución con

respecto a la del modelo PLS-DA. El *gráfico S* asociado a este modelo también queda plasmado en la **Figura A11 del anexo A**, en el cual queda reflejada la contribución de las variables al modelo OPLS-DA.

El parámetro CV-ANOVA también experimentó un aumento ($2.56\times10^{-14}$), por lo que, aunque este modelo obtenido mediante el método OPLS-DA es un modelo muy válido, sigue siendo algo inferior al que proporcionó el PLS-DA.

Seguidamente, se procedió a la determinación de posibles biomarcadores de cáncer de colon. Para ello se generaron gráficas de contribución para ambos métodos mediante la selección de aquellos metabolitos con menor desviación estándar y con un valor VIP superior a 1. Los metabolitos poseedores de regiones espectrales con VIP mayores a 1 son considerados relevantes para el modelo y son los que contribuyen a la discriminación entre muestras.

Estos *buckets* más relevantes asociados al modelo PLS-DA se encuentran recogidos en la **Figura A12 del anexo A**, mientras que los correspondientes al modelo OPLS-DA se encuentran representados en la **Figura 15**. Este último modelo fue el seleccionado como el óptimo para la obtención de biomarcadores ya que, aunque sus parámetros fueron algo inferiores a los obtenidos con el primero, al consultar los resultados obtenidos para ambos se observó cómo el modelo PLS-DA arrojaba múltiples señales correspondientes a ruido del espectro, en gran parte debido al escalado empleado para el mismo, *Unit Variance*, que infunde la misma importancia para todas las señales presentes, sin discriminar entre metabolitos, errores de medida y/o señales residuales.

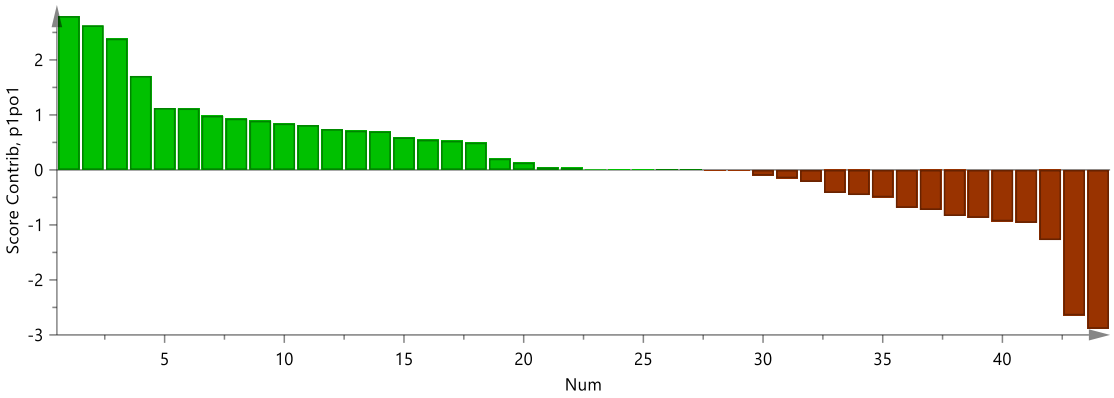



**Figura 15.** Gráfica de contribuciones generada a partir del modelo OPLS-DA. En color verde quedan señalados los *buckets* más relevantes para las muestras de tipo *Control*, mientras que en color rojo quedan plasmados los correspondientes a las muestras de tipo *Cáncer*.

Los *buckets* más relevantes para este modelo quedan recopilados en la **Tabla 4** conjuntamente a sus valores VIP, distinguiendo entre aquellos que aumentaron para cada grupo y mostrando el metabolito al que pertenecen.

**Tabla 4.** *Buckets* discriminantes (variables), su valor VIP de contribución y metabolito al que pertenecen de acuerdo a la presencia o no de cáncer en función del modelo OPLS-DA. Los *buckets* numéricos representan el centro de la región espectral (ppm) ± 0.02 ppm. [a], [b]

| Muestras de tipo Control | | | Muestras de tipo Cáncer | | |
|---|---|---|---|---|---|
| **Asignación** | **Loading** | **VIP** | **Asignación** | **Loading** | **VIP** |
| Colina | 3.22 | 2.80 | Lactato | 1.34, 4.1, 4.14 | 2.89, 2.65, 1.28 |
| Etanol | 1.18, 3.66 | 2.63, 0.98 | Piruvato | 2.38 | 0.87 |
| Glutamina | 2.46 | 1.71 | Acetato | 1.94 | 0.50 |
| Leucina | 0.94, 0.98 | 1.13, 0.20 | | | |
| Isoleucina | 0.94, 1.02 | 1.13, 0.93 | | | |
| FA | 0.94, 1.62, 0.86, 1.26, 1.3, 2.02, 1.38, 0.9. | 1.13, 0.84, 0.81, 0.71, 0.70, 0.59, 0.55, 0.13 | | | |
| Valina | 1.02, 1.06 | 0.93, 0.89 | | | |
| Alanina | 1.5 | 0.74 | | | |
| UFA | 5.34 | 0.50 | | | |

[a] Los *buckets* presentes en la **Figura 15** no incluidos en esta tabla no fueron asignados debido a la presencia de ruido, señales solapadas o se corresponden con señales de metabolitos no identificados. Aquellas señales identificadas más de una vez con distintos metabolitos corresponden a zonas espectrales con señales solapadas. [b] 3-HB: *3-Hidroxibutirato*, FA: Ácido graso, UFA: Ácido grado insaturado y PUFA: *Ácido graso poliinsaturado*.

Como puede observarse, aquellos biomarcadores con mayor relevancia, es decir, con valores VIP superiores a 1, fueron la colina, el etanol, la glutamina, la leucina, la isoleucina, los ácidos grasos, el lactato, el piruvato y el acetato. En el caso de la leucina, la isoleucina y los ácidos grasos se observó que algunas de sus señales presentaron valores inferiores a 1, sin embargo, esta disminución puede estar debida a múltiples factores, tales como el solapamiento con otras señales, por lo que continúan siendo relevantes.

## 5.2. Aplicación de métodos supervisados no lineales

Una vez se obtuvieron los metabolitos discriminantes para el modelo lineal OPLS-DA, se procedió con la aplicación de métodos no lineales supervisados con objeto de corroborar los biomarcadores encontrados, e incluso de vislumbrar nuevos metabolitos relevantes. Para ello, tal y como se ha especificado en la sección de **metodología**, se emplearon los algoritmos *Random Forest* y *Support Vector Machine* acompañados del escalado que aportase mejores resultados para cada uno, que coincidió con *Pareto* en ambos casos. De esta forma, además, se aseguró la obtención del mínimo número posible de señales de tipo ruido. Estos dos métodos fueron evaluados en el estudio de Gu *et al.* mediante respectivos análisis multivariantes ROC con el

propósito de obtener una mejor comprensión de los metabolitos del set de datos a partir de los VIPS obtenidos en el análisis lineal, pudiendo de esta forma obtener biomarcadores más robustos. Este tipo de gráficos permiten comparar *sensibilidad* y *especificidad* entre un amplio rango de valores para obtener la predicción de un resultado dicotómico. En este caso, la sensibilidad se refiere al porcentaje de individuos con cáncer colorrectal y resultados clasificados correctamente, es decir, a los verdaderos positivos, mientras que la *especificidad* viene dada por el porcentaje de individuos sin cáncer colorrectal y resultados negativos, es decir, verdaderos negativos.

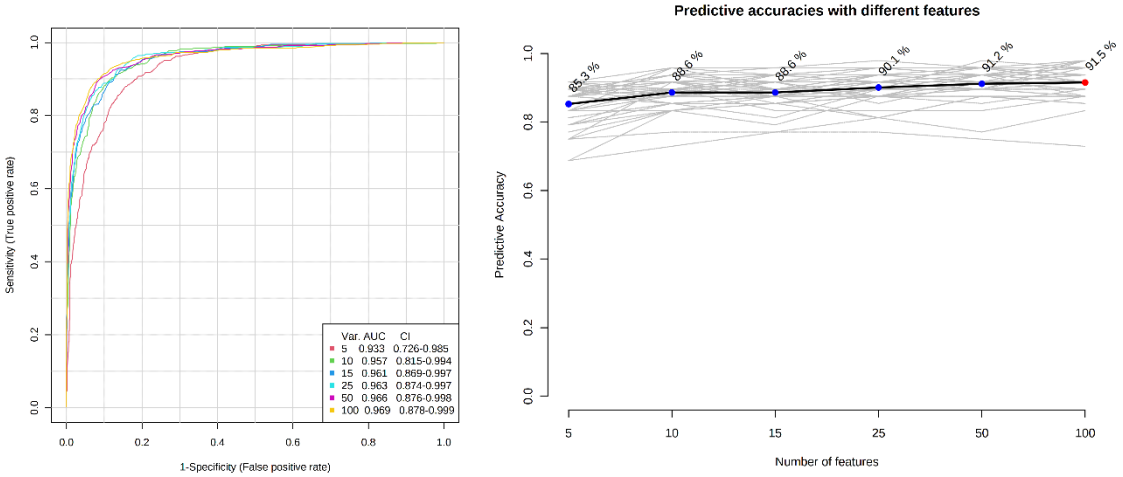
De esta manera fue aplicado el método *Random forest* escalado a *Pareto*. En el caso de este TFM, se llevó a cabo este análisis con el *dataset* original tras la eliminación de las muestras de tipo *Limpia* y *Sucia*, con el objetivo de no suprimir posibles variables que para este método podrían resultar más relevantes que con el anterior, y de esta manera no reducir posible información útil. Así, fueron generados una serie de modelos de predicción en función del número de variables escogidos para cada nodo de RF, que correspondieron a 5, 10, 15, 25, 50 y 100. Fueron representados el gráfico de curvas ROC multivariante para determinar la proporción de falsos positivos, es decir, la relación *sensibilidad/especificidad*, junto a un gráfico de precisión de las predicciones en función del número de variables seleccionadas, ambos recogidos en la **Figura 16a**.

Como puede observarse, todos los modelos generados mostraron unos valores AUC e IC muy adecuados, destacando el valor AUC proveniente del modelo de 100 *features*, igual a 0.969 en un IC estrecho de 0.878-0.999, indicando una relación *sensibilidad/especificidad* superior al resto (representado con una curva de color amarillo).

De esta forma, podría concluirse que todos estos modelos con distinto número de variables seleccionadas son capaces de distinguir en alta proporción las muestras de tipo *Cáncer* de las de tipo *Control*.

Por otro lado, en la **Figura 16b** se representa el número de variables incluidas en dichos modelos en función de la precisión de la predicción de cada uno en el conjunto de test, siendo aquél con un valor de predicción superior una vez más el correspondiente a 100 variables, con un 91.5% de precisión de predicción en el conjunto de entrenamiento. Se observa una mejora en la precisión del modelo en función de las variables seleccionadas, que comprende desde un 85.3% para el modelo de 5 variables, hasta un 91.5% para el correspondiente a 100.

En conclusión, el modelo RF que fue considerado el más adecuado en este caso correspondió al generado mediante 100 *features*, con un valor de precisión del 91.5% y un error global de 0.085.




**a)** **b)**

**Figura 16. (a**) Gráfico curva ROC/AUC para distintos modelos obtenidos mediante *Random Forest* (modelo escalado a *Pareto*) en función de las variables seleccionadas, y (**b**) Gráfico de precisión de la predicción en función de las variables seleccionadas. El modelo con valores más adecuados en ambas representaciones es el correspondiente a 100 variables, que presenta un valor AUC de 0. 969 en un *Intervalo de Confianza* al 95% de 0.878-0.999 y una precisión de predicción del 91.5%.

Así, el modelo RF obtenido mediante la selección de 100 variables logró predecir las clases de las muestras, como puede verse representado en el gráfico correspondiente a su matriz de confusión, **Figura 17,** en el que son plasmadas las muestras en función de la probabilidad de la predicción de las clases, es decir, del valor medio de la validación cruzada. Se observa una clara discriminación de las muestras según el grupo, *Cáncer*, en color blanco, o *Control*, en color negro, destacando 5 muestras incorrectamente clasificadas, 3 de cáncer marcadas en rojo y dos de control marcadas en azul.

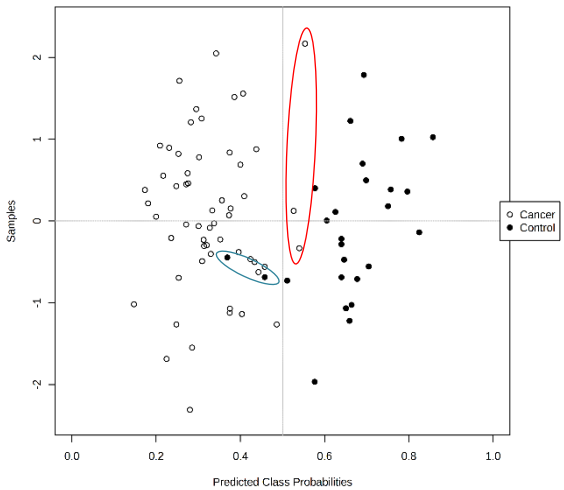
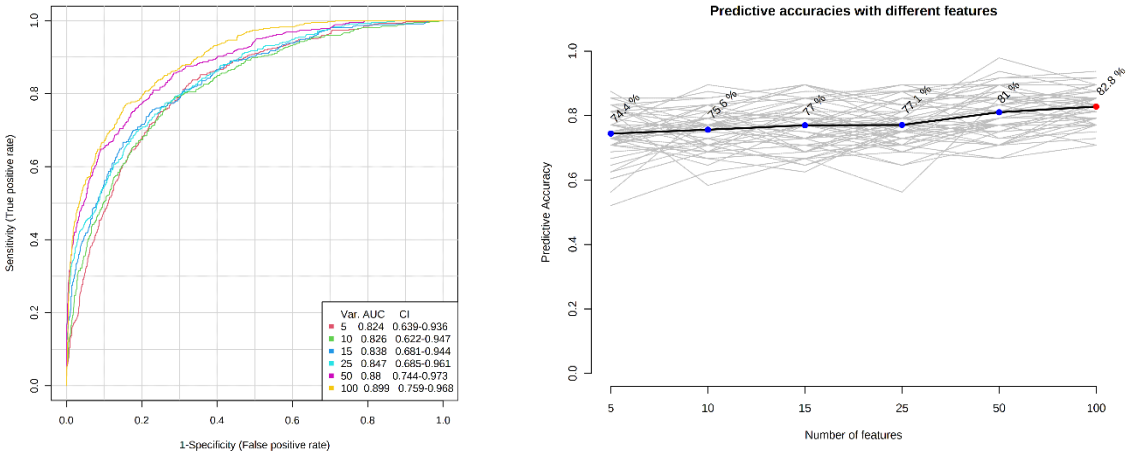



**Figura 17.** Gráfico representando la matriz de confusión de las muestras de 100 *features* según el modelo *Random Forest* escalado a *Pareto* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal. Debido a que el algoritmo emplea un método de submuestreo balanceado, el límite de clasificación se encuentra localizado en el centro de la gráfica (x= 0.5, línea). Se observa una buena discriminación de las muestras en función de los

grupos *Cáncer* y *Control*, destacando la presencia de cinco muestras incorrectamente clasificadas señalizadas mediante el color rojo (muestras *Cáncer* incorrectas) y el color azul (muestras *Control* incorrectas).

Seguidamente se aplicó el método *Support Vector Machine* con *kernel* lineal empleando el escalado *Pareto*. Se repitió el mismo procedimiento que el empleado para *Random Forest*, obteniendo la representación de las curvas ROC multivariantes asociadas a distintos modelos SVM de 5, 10, 15, 25, 50 y 100 variables de la **Figura 18a**, de los cuales, aquél con los valores óptimos fue el correspondiente a 100 variables, representado mediante la curva de color amarillo. Este modelo aporta una relación *especificidad/sensibilidad* superior al resto de modelos, presentando un valor AUC de 0.899 en un intervalo de confianza del 95% desde 0.759 a 0.968, que indica una presencia inferior de *falsos positivos* y de *falsos negativos*. El resto de ellos presentaron valores de AUC también elevados, siendo el inferior el correspondiente a 5 variables con un valor AUC de 0.824 (curva de color rojo), por lo que teniendo en cuenta que el valor AUC de 0.7 es el mínimo considerado para que un modelo sea considerado útil clínicamente, sigue siendo un buen resultado. De esta forma, el predictor correspondiente a 100 *features* fue seleccionado como el óptimo, presentando un valor AUC de 0.899 en un IC de 0.759 a 0.968. representado mediante una curva de color turquesa.

En la **Figura 18b**, por otro lado, es representado el número de variables incluidas en los distintos modelos en función de la precisión de la predicción de cada uno en el conjunto de entrenamiento, siendo aquél con un valor de predicción superior el de 100 variables con un 82.8% de precisión y un error del 0.172. En este caso, se observa una mejora considerable de la precisión del modelo en función de las variables seleccionadas, que comprende desde un 74.4% para el modelo de 5 variables, hasta el valor de precisión correspondiente a 100, detallado anteriormente.




**a)** **b)**

**Figura 18.** (**a**) Gráfico curva ROC/AUC para distintos modelos obtenidos mediante *Support Vector Machine* (modelo escalado a *Pareto*) en función de las *features* seleccionadas, y (**b**) Gráfico de precisión de la predicción en función de las variables seleccionadas. El modelo considerado como el más adecuado en ambas representaciones es el correspondiente a 100 variables, que presenta un valor AUC de 0.899 en un *Intervalo de Confianza* al 95% de 0.759-0.968 y una precisión de predicción del 82.8%.

Los parámetros correspondientes al modelo de predicción SVM obtenido mediante 100 variables seleccionadas quedan plasmados en la **Figura 19** mediante el gráfico de la matriz de confusión, en el que son plasmadas las muestras en función de la probabilidad de predicción de las clases, es decir del valor medio de la validación cruzada. Se observa una clara discriminación de las muestras según el grupo, *Cáncer*, en color blanco, o *Control*, en color negro, destacando 9 muestras incorrectamente clasificadas y que se indican en círculos rojo y azul.

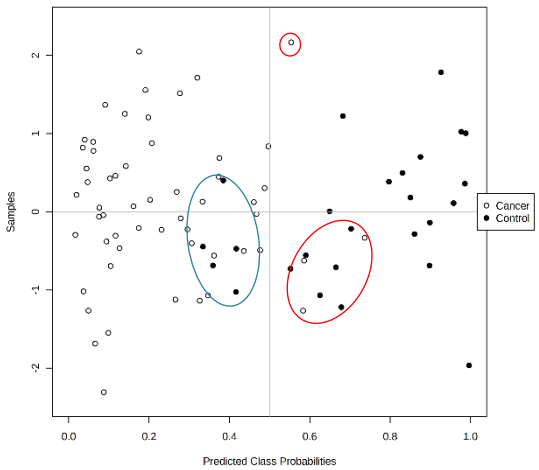



**Figura 19.** Gráfico representando la matriz de confusión de las muestras de 100 *features* según el modelo *Support Vector Machine* escalado a *Pareto* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal. Debido a que el algoritmo emplea un método de submuestreo balanceado, el límite de clasificación se encuentra localizado en el centro de la gráfica (x= 0.5, línea). Se observa una discriminación adecuada de las muestras en función de los grupos *Cáncer* y *Control*, destacando la presencia de nueve muestras incorrectamente clasificadas señalizadas mediante el color rojo (muestras *Cáncer* incorrectas) y el color azul (muestras *Control* incorrectas).

Como pudo comprobarse, ambos modelos demostraron ofrecer una óptima discriminación entre los grupos *Cáncer* y *Control* de acuerdo a los parámetros estudiados, ofreciendo valores de AUC superiores a 0.7 y precisiones elevadas, es decir, demostrando ser clínicamente útiles. Sin embargo, el modelo RF obtuvo un valor AUC superior al generado mediante SVM (RF: AUC= 0.966; SVM: AUC= 0.847), además de presentar un IC al 95% de confianza mucho más estrecho, dando así la posibilidad de obtener clasificaciones más precisas. Asimismo, presentó una precisión de las predicciones bastante más elevada que aquella generada con SVM (RF: Precisión= 89,8% ; SVM: Precisión= 77.1%), acompañada de un valor del error global del modelo inferior, valores que pueden verse reflejados en las **Figuras 17** y **19**.

De esta manera, se procedió a determinar aquellos metabolitos discriminantes o biomarcadores más relevantes para este modelo RF. Para ello se optó por su clasificación en función de la frecuencia de selección por parte del algoritmo, marcando un umbral de 0.5 [8]. Estos metabolitos quedan representados en la **Figura 20**, mientras que los correspondientes al modelo SVM también fueron incluidos en la **Figura A13 del anexo A**.

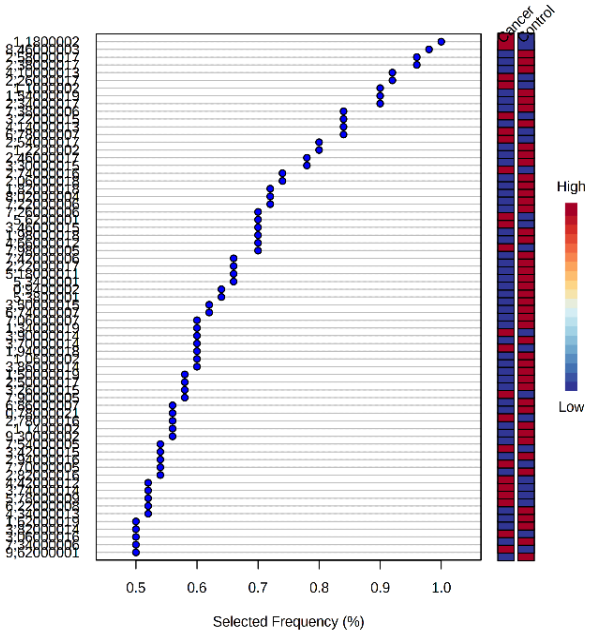



**Figura 20.** Gráfica de la frecuencia de selección de cada *bucket* generado a partir del modelo RF. A la derecha se observa una escala que identifica la probabilidad para cada grupo de selección de biomarcadores, siendo el color rojo la mayor probabilidad y el azul la mínima.

De esta forma, a partir de los resultados y consultando la tabla de asignaciones de las señales de RMN proporcionada, se generó la **Tabla 5** en la cual se resumen los biomarcadores obtenidos para cada grupo empleando este modelo RF. Además, también fueron incluidos los valores AUC correspondientes a cada metabolito, calculados mediante un análisis univariante mediante 500 *bootstraps.* Todos estos valores AUC estuvieron en un rango comprendido desde 0.90 hasta 0.60 (exceptuando una de las señales del lactato, $\delta_H$ 1.34 ppm, que presentó un valor AUC de 0.58), que indicó que la utilidad de estos biomarcadores fue adecuada, sobre todo en el caso de las variables con valores AUC en el rango de 0.9 a 0.8, tales como el formiato, el 3-hidroxibutirato, el citrato (en el cual una de sus señales, $\delta_H$ 2.58 ppm, estuvo comprendida en dicho intervalo, mientras que la otra, 2.54, fue de 0.73, valor también muy adecuado), y el lactato (en el cual una de sus señales, $\delta_H$ 4.10 ppm, estuvo incluida dentro de este intervalo, mientras que su señal a 4.14 presentó un valor AUC adecuado de 0.74, y su señal $\delta_H$ 1.34 ppm, sin embargo, fue la señal con un valor AUC más bajo).

**Tabla 5.** *Buckets* discriminantes (variables), sus valores AUC y metabolito al que pertenecen de acuerdo a la presencia o no de cáncer en función del modelo RF. Los *buckets* numéricos representan el centro de la región espectral (ppm) ± 0.02 ppm. [a], [b]

| Muestras de tipo Control | | | Muestras de tipo Cáncer | | |
|---|---|---|---|---|---|
| **Asignación** | **Loading** | **AUC** | **Asignación** | **Loading** | **AUC** |
| Formiato | 8.46 | 0.85 | 3-HB | 1.18 | 0.86 |
| Citrato | 2.58, 2.54 | 0.84, 0.73 | Lactato | 4.10, 4.14, 1.34 | 0.81, 0.74, 0.58 |

| | | | | | |
|---|---|---|---|---|---|
| Piruvato | 2.38 | 0.76 | Fenilalanina | 7.38, 7.42, 7.34 | 0.79, 0.70, 0.64 |
| Valina | 2.22, 1.06 | 0.65, 0.72 | | | |
| FA | 2.34, 2.06, 0.94, 1.62 | 0.72, 0.71, 0.71, 0.70 | | | |
| Colina | 3.22 | 0.79 | | | |
| Glutamina | 2.46, 2.50 | 0.79, 0.65 | | | |
| PUFA | 2.74, 2.78, | 0.64, | | | |
| Leucina | 0.94 | 0.71 | | | |
| Isoleucina | 0.94, 1.94 | 0.71, 0.63 | | | |
| UFA | 5.34, 0.94, 5.38 | 0.71, 0.71, 0.68 | | | |
| Alanina | 1.50 | 0.69 | | | |
| Isobutirato | 1.14 | 0.63 | | | |

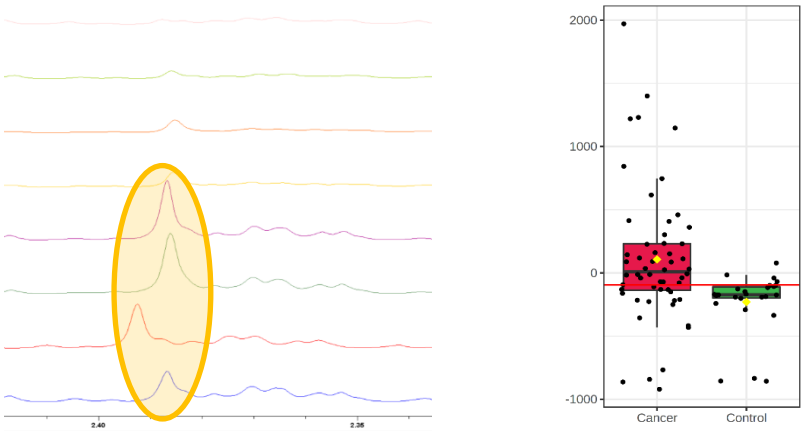
[a] Los *buckets* presentes en la **Figura 20** no incluidos en esta tabla no fueron asignados debido a la presencia de ruido, señales solapadas o se corresponden con señales de metabolitos no identificados. Aquellas señales identificadas más de una vez con distintos metabolitos corresponden a zonas espectrales con señales solapadas. [b] 3-HB: *3-Hidroxibutirato*, FA: Ácido graso, UFA: Ácido grado insaturado y PUFA: *Ácido graso poliinsaturado*.

Fueron generadas las gráficas de curvas ROC para las señales anteriormente detalladas con valores AUC de entre 0.8 a 0.9 con el objetivo de observar los valores asociados a los biomarcadores más destacados y que aportan un mínimo de falsos positivos, junto al IC al 95% correspondiente a cada uno, calculado mediante 500 *bootstraps*, que pueden ser observados en la **Figura A14 del anexo A**. En todos los casos los IC fueron relativamente estrechos, desde un 0.7 a un 0.9, que indican una mayor precisión de la estimación. Además, los diagramas de caja (o *boxplots*) de cada una de estas variables fue incluido en la **Figura A15 del anexo A**.

Además, las estructuras químicas de cada uno de los metabolitos identificados como posibles biomarcadores de cáncer colorrectal mediante ambos métodos se encuentran representadas en la **Figura A16 del anexo A**.

Finalmente, se compararon los resultados de biomarcadores obtenidos tanto por el modelo lineal OPLS-DA como por el modelo no lineal RF. En general, se observó que los biomarcadores de ambos modelos coincidieron en su mayoría, destacando colina, glutamina, leucina, ácidos grasos, valina, alanina y ácidos grasos insaturados para el grupo de tipo *Control*, y lactato para el grupo de tipo *Cáncer*. Por otro lado, uno de los biomarcadores, el piruvato (señal en $\delta_H$ 2.38 ppm) fue clasificado como relevante para el grupo de *Cáncer* en el modelo OPLS-DA, y para el grupo *Control* en el modelo RF. Este detalle puede verse explicado en base a que la región $\delta_H$ 2.38 ± 0.02 ppm incluye otras señales además de la del piruvato, que pueden solapar con la de interés. Además, la naturaleza ácida del piruvato hace que esta señal pueda aparecer en distintos

desplazamientos, por lo que esto puede haber contribuido también en parte a estos resultados. Para comprobar visualmente el grupo en el cual la señal del piruvato es más intensa, se consultaron una serie de espectros de tipo *Cáncer* y otros de tipo *Control* y se identificó dicha señal, que puede verse resaltada de color naranja en el caso de las muestras de tipo *Cáncer* en la **Figura 21a**, grupo en el cual la señal pareció mostrar mayor intensidad. Además, se realizó un diagrama de caja del *bucket* $\delta_H$ 2.38 ppm, en el cual, aun presentando una gran dispersión, se confirmó su presencia mayoritaria en el grupo *Cáncer*, observado en la **Figura 21b**. Por lo tanto, se concluye que el piruvato probablemente predomina en el grupo *Cáncer*, tal y como el modelo OPLS-DA intuye, y que el *bucket* $\delta_H$ 2.38 ppm del modelo RF posiblemente esté debido a otro metabolito no identificado, por lo que se desestimó de este modelo.




**a)** **b)**

**Figura 21.** (**a**) Regiones espectrales correspondientes a muestras de tipo *Cáncer* (tres primeras), y de tipo *Control* (tres últimas), en las cuales se observa claramente un cambio en el desplazamiento de la señal del piruvato acompañado de la presencia de otras señales (± 0.02 ppm), y (**b**) diagrama de caja del *bucket* en $\delta_H$ 2.38 ppm, en el que resalta la presencia en su mayoría en el grupo *Cáncer*.

Por otro lado, en ambos modelos se obtuvieron también metabolitos no comunes entre sí. En el caso del modelo OPLS-DA destacó la presencia de etanol e isoleucina en las muestras de tipo *Control*, mientras que para las de tipo *Cáncer* fue el acetato. El modelo RF resaltó un aumento de formiato, citrato, ácidos grasos poliinsaturados, isoleucina e isobutirato en el grupo *Control*, mientras que en el grupo de *Cáncer* los metabolitos no comunes a ambos modelos más destacados fueron el 3-hidroxibutirato y la fenilalanina.

## 5.3. Interpretación biológica

Una vez obtenidos los metabolitos más relevantes para los modelos OPLS-DA y RF, se procedió a la interpretación biológica de los mismos empleando la herramienta especificada en el **apartado 4** de metodología y consultando de los resultados de la revisión bibliográfica del **apartado 3.3.**

Así, en primer lugar, se obtuvieron para el modelo OPLS-DA como más relevantes las rutas metabólicas observadas en la **Figura A17 del anexo A**. Fueron incluidas únicamente aquellas con un impacto superior a 1 y con más de un metabolito implicado en ellas (coincidencia superior a 1). Aquellas rutas metabólicas con valores *p* ajustados de Holm y FDR inferiores a 0.05

definieron las rutas mayormente enriquecidas en este análisis, que correspondieron con el metabolismo del piruvato, la glicólisis/gluconeogénesis, y el metabolismo de la alanina, del aspartato y del glutamato. Los valores de impacto, *p* valores, valores *p* de *Holm*, y FDR se encuentran descritos en la **Tabla A2 del anexo A**.

En función de los resultados obtenidos mediante el modelo OPLS-DA, se concluyó que aquellos biomarcadores posiblemente implicados en el metabolismo del piruvato y en la glucólisis, es decir, el lactato, piruvato y acetato, aumentaron en las muestras de suero de cáncer colorrectal, tal y como se mostró en los estudios de Ludwig *et al*.[24], Qiu *et al.* [25] y Gu *et al*.[8], en los cuales se reportó un aumento de los niveles de lactato y de piruvato en este tipo de muestras. El hecho de que en este presente trabajo se reporte también el acetato como metabolito implicado, consigue reforzar la teoría de que las rutas metabólicas correspondientes sean estas, ya que este compuesto forma parte del mecanismo en ambos procesos. En general, los metabolitos que mostraron una clara disminución de sus niveles, coincidieron con los resultados obtenidos en el tercer estudio anteriormente mencionado, correspondiente a Gu *et al*.[8], en el cual también se reportó una disminución de leucina, valina, ácidos grasos, ácidos grasos insaturados, alanina y glutamina, siendo estos dos últimos compuestos dos de los metabolitos implicados (junto al piruvato) en el metabolismo de la alanina, del aspartato y del glutamato. Además, también se reportó una disminución en los niveles de etanol y de leucina, compuesto que llama la atención ya que, en el mencionado estudio, se indica un aumento del mismo en muestras de cáncer. También cabe mencionar el estudio de Deng *et al*. [28], en el cual también se describieron alteraciones en los niveles de piruvato y glutamina en muestras de cáncer de colon.

Seguidamente, para el modelo RF se obtuvieron las rutas metabólicas observadas en la **Figura A18 del anexo A**. Sin embargo, aunque dos de las rutas, el metabolismo de la alanina, aspartato y glutamato, y la ruta del glioxilato y dicarboxilato, mostraron un impacto superior a 0, una coincidencia superior a 1, y mostraron *p* valores y FDR inferiores a 0.05, al llevar a cabo una corrección múltiple según el método de *Holm-Bonferroni*, se obtuvieron valores de *p* de *Holm* superiores a 0.05 por lo que se concluyó en que ninguna de las rutas obtenidas mediante dicho método fueron estadísticamente significantes. Los valores de impacto, *p* valores, valores *p* de *Holm*, y FDR se encuentran descritos en la **Tabla A3 del anexo A**.

# 6. Discusión

En primer lugar, en este Trabajo de Fin de Máster, se llevó a cabo una breve revisión bibliográfica recogida en la **Tabla A1** sobre algunas investigaciones relacionadas con el estudio de los metabolitos presentes en el suero de cáncer colorrectal. En dicha revisión, se observó que en general, los estudios se centraron en el papel de los metabolitos envueltos en la glucólisis, ya que un aumento en la actividad en esta ruta puede suponer un incremento de tumores malignos, conocido como *efecto Warburg*. Este proceso envuelve una acumulación anormal de piruvato, glucosa y de lactato (metabolitos intermedio y final de la glucólisis), que a su vez, puede estar relacionada con una alta demanda de aminoácidos por parte de los tejidos tumorales, provocando alteraciones en el metabolismo asociado a estos compuestos, y una consecuente disminución de sus niveles en muestras cancerígenas [24, 25]. Además, otros metabolitos relacionados con esta ruta tales como el citrato y el succinato en bajas cantidades también fueron indicados como parte de este *efecto Warburg* [8]. Algunos de los metabolismos de aminoácidos reportados en estos artículos son el de la arginina, glutamina y prolina, el de la alanina, aspartato y glutamato, y el del piruvato [28, 8].

Por otro lado, fueron identificados también altos niveles de cuerpos cetónicos tales como acetato, acetoacetato y 3-hidroxibutirato en muestras de cáncer, siendo en concreto este último el más identificado [24, 25, 27], junto a una disminución de los niveles de ácidos grasos, tanto UFA como PUFA [8] que hacen prever que el metabolismo de glicerolípidos y de ácidos grasos podrían estar implicados en estos procesos cancerígenos. Además, otros procesos mencionados entre otros fueron el metabolismo de la cianoamina, de la timina, del metano, del glutatión, y de la fucosa y manosa en el estudio de Zamani *et al*.[27], el metabolismo de ácidos biliares y el del tocoferol en el artículo de Cross *et al*. [26], y el de la colina en el respectivo a Gu *et al*.[8].

A excepción del estudio de Deng *et al*.[28], todos aplicaron métodos no supervisados de tipo PCA y seguidamente llevaron a cabo análisis supervisados para mejorar la discriminación entre grupos y lograr obtener potenciales biomarcadores de cáncer colorrectal. En los casos de Ludwig [24], Zamani [27], Deng [28] y Gu [8], fueron aplicados modelos PLS-DA, y determinaron los biomarcadores asociados mediante distintos métodos, destacando los valores VIP más relevantes. Los estudios de Qiu [25] y Gu [8] emplearon OPLS-DA, y coincidieron en la metodología de selección de los biomarcadores, empleando ambos los valores VIP. El estudio de Cross *et al.* [26] aplicó un modelo RF y escogió los metabolitos más relevantes según el test de *Bonferroni*. Por último, el estudio de Gu *et al*. [8], empleó un clasificador RF, seleccionando los potenciales biomarcadores en función de su frecuencia de ser escogidos por el algoritmo. Seguidamente, emplearon un modelo SVM con el objetivo de validar los resultados obtenidos.

En general, estos estudios demostraron un gran poder de predicción, clasificación y selección de biomarcadores empleando RMN de $^{1}$H para el análisis de metabolitos, sin embargo, aunque la tendencia actual sigue aumentando, no se logró obtener un amplio número de artículos que empleasen dicha técnica como herramienta analítica principal. Por esta razón, algunos de los estudios incluidos en la revisión emplearon otras técnicas de análisis. En el caso de Deng *et al*. [28] además de la RMN emplearon LC-MS, por otro lado, el estudio de Qiu *et al*. utilizó LC-MS y GC-MS-TOF, y en el de Cross *et al*. [26] se hizo uso de UPLC-MS y GC-MS. Es de gran importancia remarcar estas diferencias ya que, en función del equipo empleado, pueden detectarse distintos tipos de metabolitos en función de la *sensibilidad* y *especificidad* de cada plataforma.

Otro factor de gran importancia es el referente a los datos. Por un lado, se dispuso de tamaños muestrales bastante reducidos en la mayoría de estudios, a excepción del de Cross *et al*. [26], aspecto que pudo condicionar muchos de los resultados estadísticos obtenidos. Además, se observó una falta de consenso entre los estudios a la hora de afrontar el análisis estadístico de los datos, comenzando por la alta variación en los tipos de escalados, normalizados y transformaciones empleadas. De esta manera, algunos de los metabolitos obtenidos en estos estudios varían entre unos y otros, ya que en función del procesado de los datos se llegan a priorizar distintos tipos de señales, tal y como puede verse resumido en la **Tabla 3**.

Los variables seleccionadas en estos estudios correspondieron con muestras de cáncer y muestras control, exceptuando los estudios de Deng *et al*. [28] y de Gu *et al*. [8], en los cuales se consideraron también muestras de pólipos. De esta forma, también podría resultar interesante el análisis de muestras de cáncer colorrectal en función de su estadío (si presenta metástasis o se encuentra al principio de la enfermedad), empleando otros tipos de muestras, tales como la orina o tejido tumoral, o teniendo en cuenta la presencia de otras variables tales como la edad, la presencia de otras enfermedades o el estado físico del individuo. Además, tampoco se observó la caracterización de los ácidos grasos identificados en los artículos, por lo que podría ser una futura línea de investigación a estudiar empleando técnicas analíticas tales como GC-FID.

En segundo lugar, en este TFM se llevó a cabo un estudio en el cual fueron aplicados métodos de análisis multivariante en datos de RMN procedentes de muestras de suero de cáncer colorrectal con el propósito de obtener sus correspondientes perfiles metabólicos. Para ello, fueron tomadas en cuenta aquellas técnicas estadísticas observadas en el apartado de revisión bibliográfica (**apartado 3.3**), realizando de esta forma un recorrido por algunas de las metodologías más aplicadas actualmente. De esta manera, fueron empleados los métodos PCA, PLS-DA, OPLS-DA, SVM y RF, llevando a cabo una separación entre métodos lineales y no lineales para una mayor organización. Para su implementación, fueron utilizados además algunos de los programas más comunes en el campo de la metabolómica, correspondientes a *Amix*, *RStudio*, *SIMCA*, y la herramienta web *MetaboAnalyst*.

En general no fue observada ningún tipo de discriminación entre grupos empleando el modelo PCA no supervisado, aun aplicando diferentes tipos de escalado con el objetivo de priorizar distintos tipos de señales. Sin embargo, sí fueron obtenidos resultados satisfactorios empleando métodos supervisados. Los modelos lineales PLS-DA (escalado a *unit variance*) y OPLS-DA (escalado a *Pareto*) obtuvieron valores de regresión, capacidad predictiva y de validez del modelo muy adecuados (PLS-DA: $R^2X$= 0.361, $R^2Y$= 0.918, $Q^2Y$= 0.713, RMSEE= 0.137, CV-ANOVA= $1.66\times10^{-18}$; OPLS-DA: $R^2X$= 0.7, $R^2Y$= 0.809, $Q^2Y$= 0.687, RMSEE= 0.211, CV-ANOVA= $2.56\times10^{-14}$). Por otro lado, los modelos no lineales de clasificación SVM (de *kernel* lineal) y RF generaron valores AUC y de precisión de predicción altamente aceptables (RF: AUC= 0.969, IC 95%= 0.878-0.999, Precisión= 91.5%; SVM: AUC= 0.899, IC= 0.759-0.968, Precisión= 82.8%). Seguidamente se procedió a la identificación de biomarcadores, para lo cual fue seleccionado en primer lugar el modelo OPLS-DA en función de sus parámetros estudiados y del tipo de escalado empleado (*Pareto*) con el objetivo de evitar una alta proporción de señales de tipo ruido. En segundo lugar, fue seleccionado el modelo RF de 100 *features* en función de los parámetros de robustez anteriormente comentados.

Así, fue aplicado el método de los VIP más relevantes en el modelo OPLS-DA con el objetivo de identificar aquellos metabolitos más contribuyentes para dicho modelo. Fueron obtenidos un total de 12 metabolitos con un VIP superior a 1, entre los que se encontró una disminución de los aminoácidos glutamina, leucina, isoleucina, valina y alanina, también reportados en estudios tales como el de Qiu *et al*. [25], en el cual se menciona su posible relación con el metabolismo de aminoácidos como resultado de una alta demanda por parte de los tejidos cancerígenos; la presencia de una menor proporción de ácidos grasos (FA y UFA), y el aumento del cuerpo cetónico acetato, hecho reportado en el estudio de Gu *et al*. [8] en relación al metabolismo de glicerolípidos como ruta implicada en procesos cancerígenos; el aumento del nutriente esencial colina, que según este último estudio mencionado podría provenir del metabolismo de lípidos de membrana debido a una rápida proliferación celular [8]; un aumento de lactato y de piruvato, que junto al incremento de acetato anteriormente reportado podrían tener relación con el metabolismo del piruvato, ruta identificada en el estudio de Gu *et al*. [8] en relación con las muestras de pólipos, y que en combinación con una disminución de etanol, también identificada en el presente TFM, podrían implicar una relación con la glucólisis [24, 25]. Como se ha comentado anteriormente, el aumento de la actividad glucolítica, en concreto un aumento de piruvato y de lactato, constituyen una clara firma denominada *efecto Warburg*, característico del comportamiento cancerígeno.

Al llevar a cabo el análisis de las rutas metabólicas implicadas en estos metabolitos, es decir, el *Pathway Analysis*, se reportaron la glucólisis/gluconeogénesis, el metabolismo del piruvato, y el metabolismo de la alanina, del aspartato y del glutamato como rutas mayormente enriquecidas, coincidiendo con gran parte de los resultados de los artículos de la revisión. Además, fueron

identificadas también otras de las rutas mencionadas en estos estudios, asociadas a los aminoácidos y a los ácidos grasos, sin embargo, al llevar a cabo la corrección de *Holm-Bonferroni*, no constituyeron resultados estadísticamente significantes.

Por otro lado, los metabolitos más estadísticamente significantes para el modelo RF (escalado a *Pareto*) de 100 *features* fueron identificados en función de la frecuencia de ser seleccionados por el algoritmo, cuyo valor mínimo considerado fue 0.5. De esta forma, se lograron vislumbrar un total de 16 metabolitos, en mayoría comunes a aquellos obtenidos mediante el análisis de VIP anterior con el modelo OPLS-DA, entre los que se encontró la disminución de los aminoácidos valina, glutamina, leucina, isoleucina, alanina y fenilalanina, la disminución de ácidos grasos (FA, UFA, a los cuales añade también PUFA en este caso), la disminución del nutriente colina, y el aumento del lactato característico del *efecto Warburg*. También fue identificado el *bucket* correspondiente a la señal $\delta_H$ 2.38 ppm, identificado como piruvato en el modelo anterior. Sin embargo, en este caso fue identificado como grupo *Control*, por lo que se dedujo que el algoritmo RF seleccionó este metabolito en función de otras señales presentes en el rango ±0.02 ppm, por lo que no correspondió a la señal del piruvato. Además, un nuevo metabolito identificado en este modelo correspondió al citrato, que presentó una disminución de su concentración para las muestras cancerígenas, y que de acuerdo con las conclusiones comentadas en el estudio de Gu *et al*. [8] y con las rutas resultantes del modelo anterior, podría estar relacionado con la ya identificada glucólisis. También fue reconocido una disminución de isobutirato y un aumento de 3-hidroxibutirato, que de acuerdo a los estudios de Ludwig [24], Qiu [25] y Zamani *et al*. [27], este último compuesto podría estar influenciado por el metabolismo de glicerolípidos y de ácidos grasos, también posiblemente relacionados con la disminución de ácidos grasos. Fue también vislumbrada la disminución de dos metabolitos anteriormente no reportados correspondientes al formiato y a la fenilalanina.

Seguidamente, se llevó a cabo el análisis de las rutas metabólicas posiblemente relacionados con estos biomarcadores, sin embargo, ninguna de ellas fue reportada como estadísticamente significante de acuerdo a la corrección de *Holm-Bonferroni.* Algunas de estas rutas correspondieron al metabolismo de la alanina, aspartato y glutamato, y al metabolismo del glioxilato y dicarboxilato que, aunque presentaron unos valores *p* de 0.0013 y de 0.0020, al observar sus valores *p* de *Holm*, estos fueron superiores a 0.05. Estos resultados podrían deberse a no haber llevado a cabo un filtrado previo al modelo RF de los metabolitos más diferenciales, sin embargo, fue un paso que optó por evitarse debido al alto número de señales disponibles para un mismo metabolito, por lo que, si se llevase a cabo el filtrado de los mismos desde un primer momento, podría perderse gran parte de la información. Por otro lado, una gran parte de los *buckets* más relevantes correspondieron a señales de solapamientos, que conllevó una alta dificultad añadida. Además, al aplicar el método de supresión de agua, una zona concreta del espectro ($\delta_H$ 4.67 a 5.12 ppm) es eliminada, incluyendo posibles señales de interés que pudieran contribuir a los resultados del proceso estadístico.

En conclusión, este estudio siguió un procedimiento bien fundamentado en base al apartado de revisión bibliográfica, obteniendo diversos resultados satisfactorios a lo largo de los métodos empleados y por lo tanto demostrando la posibilidad de obtener biomarcadores de relevancia para el cáncer de colon, que podría suponer un gran avance en el diagnóstico médico de esta enfermedad.

# 7. Conclusiones

## 7.1. Conclusiones de este trabajo

(1) La RMN es una técnica novedosa que está demostrando una gran aplicabilidad en el área de la metabolómica clínica. Por este motivo actualmente no existe un gran número de contribuciones en el estudio del cáncer colorrectal.

(2) No existe un procedimiento estadístico estándar entre estudios para la identificación de biomarcadores relevantes en metabolómica de RMN, que conlleva a múltiples conclusiones y metodologías.

(3) En general, los estudios han mostrado resultados satisfactorios en el estudio de metabolitos involucrados en el cáncer colorrectal, por lo que es un área que debe seguir en desarrollo. Los metabolitos identificados podrían estar relacionados con la glucólisis, el metabolismo de aminoácidos y el metabolismo de glicerolípidos y de ácidos grasos, ente otras.

(4) Aunque *R* es un poderoso entorno de trabajo para la ejecución de análisis estadísticos, los paquetes disponibles para el estudio de datos metabolómicos resultan poco intuitivos, por lo que podría suponer una dificultad añadida para el análisis a personas que están iniciándose en el lenguaje. Otros programas tales como *SIMCA* y *MetaboAnalyst* aportan múltiples opciones de análisis interesantes, que en múltiples ocasiones pueden completar las necesidades del estudio.

(5) Los resultados obtenidos mediante el método no supervisado PCA no obtuvieron ningún tipo de discriminación, posiblemente debido al alto número de variables desconocidas envueltas en muestras clínicas.

(6) La preparación de los datos supone un paso crucial para lograr una correcta interpretación de los resultados.

(7) Los resultados obtenidos empleando técnicas supervisadas mejoraron considerablemente. Se lograron vislumbrar múltiples biomarcadores de cáncer colorrectal tanto para el modelo OPLS-DA como para el modelo RF, por lo que los objetivos de este trabajo fueron logrados.

(8) La interpretación biológica de estos resultados coincidió con parte de los estudios revisados, por lo que podría suponer una contribución de gran importancia.

(9) Se mostraron algunas limitaciones, tales como un tamaño muestral reducido, la presencia de pocas variables de estudio, y la corta experiencia empleando *R*.

## 7.2. Líneas de futuro

(1) Un aspecto que surgió en la elaboración de este TFM fue el estudio de métodos no lineales tales como RF y SVM, los cuales fueron desarrollados sin un paso previo de filtrado de variables, por lo que sería de gran interés combinar métodos estadísticos tales como LASSO, VIP más relevantes o el coeficiente de *Gini*.

(2) Uso de otras técnicas supervisadas de *Deep Learning*, o la implementación de otros tipos de *kernel* en el algoritmo SVM.

(3) Implementación de librerías de *R* tales como ASICS o BATMAN para la asignación de las señales de los espectros de RMN de $^{1}$H.

(4) Este conjunto de datos fue desarrollado con una base de *bucketing* en regiones de 0.04 ppm, pero sería interesante desarrollar un *workflow* a partir de otros tipos de *bucketizados*, tales como el variable.

(5) Algunas variables no pudieron ser identificadas debido a solapamientos, por lo que podría ser un problema a solventar empleando espectros bidimensionales de $^{1}$H-RMN con los cuales confirmar las relaciones entre señales.
(6) Uso de nuevas variables tales como el estadío de la enfermedad, la dieta del individuo, y la edad, entre otras.
(7) Exploración de posibles relaciones entre los biomarcadores identificados examinando sus ratios en función de las variables de interés.
(8) Como se ha mencionado anteriormente, no fue observada la caracterización de los ácidos grasos en ninguno de los estudios de la bibliografía. Por lo tanto, sería de alto interés emplear técnicas como GC-FID y posteriormente realizar análisis estadísticos de los resultados.

### 7.3. Seguimiento de la planificación

Se lograron cumplir todos los objetivos marcados al inicio del semestre, sin embargo tuvieron que introducirse algunas modificaciones en la planificación del estudio y en consecuencia, en el marco temporal.

El primer objetivo, referente al apartado de revisión bibliográfica, se cumplió sin ningún tipo de complicación y siguiendo todos los pasos previstos. Seguidamente, el segundo objetivo se llevó a cabo según lo previsto, sin embargo al llegar a la aplicación de modelos supervisados se previó emplear paquetes de *R* para su implementación, que tuvieron que ser reemplazados por el uso de los programas *SIMCA* y *MetaboAnalyst* para lograr el objetivo final. Esto supuso un retraso en la programación de 1 semana, por lo que el objetivo 3 tuvo que ser pospuesto. Este último paso fue llevado a cabo en la primera semana correspondiente a la PEC 4, conjuntamente a la redacción del trabajo, lo cual no supuso una gran carga adicional.

# 8. Glosario

**RMN de $^{1}$H**: Acrónimo de *Resonancia Magnética Nuclear de Protón*. Técnica analítica de determinación estructural de compuestos de alta sensibilidad, especificidad y robustez.

***Bucket***: Regiones de entre 0.04 a 0.02 ppm de ancho en las cuales se divide un espectro de RMN para poder llevar a cabo posteriores análisis estadísticos.

**Metabolómica**: Se trata de la ciencia ómica que estudia el metabolismo de un sistema biológico.

**Metabolitos**: Aquellos compuestos derivados de la metabolómica.

**PCA**: *Análisis de Componentes Principales* (del inglés *Principal Component Analysis*). Se trata de un método no supervisado multivariante que permite tanto la reducción de las dimensiones de un conjunto de datos, como la observación de las agrupaciones existentes entre las variables del mismo.

**PLS-DA**: *Análisis Discriminante de Mínimos Cuadrados Parciales* (del inglés *Partial Least Squares Discriminant Analysis*). Se trata de una técnica de clasificación supervisada multivariante que permite tanto la reducción de las dimensiones de un conjunto de datos, como la discriminación entre las variables del mismo encontrando una regresión lineal mediante la proyección de las variables de predicción y las observables en un nuevo espacio.

**OPLS-DA**: *Análisis Discriminante de Mínimos Cuadrados Parciales Ortogonal* (del inglés *Orthogonal Partial Least Squares Discriminant Analysis*). Es la versión ortogonal de PLS-DA,

se trata de una técnica de clasificación supervisada multivariante que permite la discriminación entre las variables del mismo encontrando una regresión lineal a lo largo de una única componente ortogonal.

**RF:** *Random Forest* (o *Bosque Aleatorio*). Se trata de un método supervisado de *Machine Learning* de clasificación o regresión basado en múltiples *árboles de decisión* que emplea *bagging*, es decir, que cada árbol entrene distintas porciones de los datos de entrenamiento, obteniendo predicciones.

**SVM**: *Support Vector Machine* (o *Máquinas de Vectores de Soporte*). Son métodos supervisados de *Machine Learning* de clasificación o regresión basados en un concepto de jiperplano que permita la agrupación de las variables, empleando un parámetro de *kernel* (lineal, *gaussiano*, entre otros).

**VIP**: *Variable Importance in Projection* (o *Importancia de las Variables para la Proyección*). Se trata de un método comúnmente empleado con PLS-DA y OPLS-DA para la identificación de posibles biomarcadores mediante la proyección de un conjunto grande de variables en uno más reducido, en función de su importancia para el modelo.

**Curva ROC**: Gráfica de la curva *Receiver Operating Characteristic*,. Permite la evaluación de la robustez de un modelo mediante el diagnóstico de la sensibilidad y la especifidad del mismo.

**AUC**: *Área Bajo la Curva* (o *Area Under the Curve*). Se trata de la proporción del área que queda bajo la curva ROC, que aporta la relación sensibilidad/especifidad del modelo.

**Biomarcadores**: Compuestos relacionados en un proceso biológico, generalmente empleados en el diagnóstico clínico de enfermedades.

**Validación cruzada:** o *cross-validation*. Se trata de una metodología de valoración de los resultados de un test en función de las proporciones de entrenamiento y test del *dataset*.

***Bootstrapping*:** Método de remuestreo comúnmente empleado para la construcción de *Intervalos de Confianza* o para contraste de hipótesis.

**IC**: *Intervalo de Confianza*. Rango en el cual se estima que se encontrará cierto valor respecto a una población con un determinado nivel de confianza.

**MCCV:** *Validación Cruzada de Monte Carlo* (del inglés *Monte-Carlo Cross Validation*). Es una técnica de validación cruzada en la cual se selecciona aleatoriamente una fracción del conjunto de datos (sin reemplazo) para formar la fracción de entrenamiento, y el resto de datos se emplea como conjunto de *test*. Este procedimiento es repetido múltiples veces, generando múltiples conjuntos de entrenamiento y *test* independientes.

**Corrección *Holm-Bonferroni***: Se trata de un método de corrección de comparaciones múltiples en el cual se lleva a cabo un *t-test* común, organizando los datos en orden ascendente de *p* valor.

**RMSEE:** *Error cuadrático medio de la estimación* de un modelo lineal (del inglés *Root Mean Square Error of Estimation*).

# 9. Bibliografía

(1) American Cancer Society. *Acerca del cáncer colorrectal.* **2018** [Online] https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/acerca/que-es-cancer-de-colon-o-recto.html%22%20/l%20%22:~:text=El%20c%C3%A1ncer%20de%20colon%20y,a%20otras%20partes%20del%20cuerpo (Página accedida a 2 de abril de 2021).

(2) Asociación Española Contra el Cáncer. *Cáncer de colon.* **2020** [Online] https://www.aecc.es/es/todo-sobre-cancer/tipos-cancer/cancer-colon (Página accedida a 2 de abril de 2021).

(3) IARC Global Cancer Observatory. **2018** [Online] https://gco.iarc.fr/ (Página accedida a 2 de abril de 2021).

(4) Sociedad Española de Oncología Médica. *Las cifras del cáncer en España.* **2020** [Online] https://seom.org/seomcms/images/stories/recursos/Cifras_del_cancer_2020.pdf (Página accedida a 2 de abril de 2021).

(5) Asociación Española Contra el Cáncer. *Incidencia y mortalidad de cáncer colorrectal en España en la población entre 50 y 69 años.* **2018** [Online] https://www.aecc.es/sites/default/files/content-file/Informe-incidencia-colon.pdf (Página accedida a 7 de abril de 2021).

(6) American Cancer Society. *Tasas de supervivencia por etapas para el cáncer colorrectal*. **2021** [Online] https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/deteccion-diagnostico-clasificacion-por-etapas/tasas-de-supervivencia.html (Página accedida a 7 de abril de 2021).

(7) Sociedad Española de Oncología Médica. *Cáncer de colon y recto.* **2021** [Online] https://seom.org/info-sobre-el-cancer/colon-recto?showall=1 https://www.cdc.gov/cancer/colorectal/basic_info/risk_factors.htm (Página accedida a 7 de abril de 2021).

(8) Gu, J., Xiao Y., Su D., Liang X., Hu, X., Xie, Y., Lin, D., Li H. *Metabolomics Analysis in Serum from Patients with Colorectal Polyp and Colorectal Cancer by $^{1}$H-NMR Spectrometry*. Dis. Markers **2019**, ID: 3491852, doi: 10.1155/2019/3491852

(9) Gutiérrez-Aguilar, R., Frigolet-Vázquez, M.E. *Ciencias Ómicas, ¿cómo ayudan a las ciencias de la salud?* Revista digital universitaria **2017**, 18(7).

(10) Bernal-Ruiz, M. L. *La era de las ciencias ómicas.* Colegio Oficial de Farmacéuticos de Aragón **2015**, Zaragoza.

(11) Yanes, O. *Metabolómica: la ciencia Ómica más multidisciplinaria*. Sociedad Española de Bioquímica y Biología Molecular **2020**.

(12) Abreu, A. C., Fernández, I. *Nuclear magnetic resonance to study bacterial biofilms structure, formation and resilience (chapter 2).* Recent Trends in Biofilm Science and Technology. Simões M. (ed.), Elsevier **2020**, 23-70.

(13) Castejón-Ferrer, D. *Avances en el Estudio de Matrices Alimentarias mediante RMN metabolómica*. Universidad Complutense de Madrid, Facultad de Ciencias Químicas, Tesis doctoral **2015**.

(14) Fernández de las Nieves, I. *Advanced NMR Methods and Metal-Based Catalysts*. **2019**. [Online] https://www.nmrmbc.com/gallery/ (Página accedida a 9 de abril de 2021).

(15) Esturau, N., Espinosa, J. F. *Optimization of Diffusion-Filtered NMR Experiments for Selective Suppression of Residual Nondeuterated Solvent and Water Signals from $^{1}$H NMR Spectra of Organic Compounds.* J. Org. Chem. **2006**, 71, 4103-4110.

(16) Base de datos *Web Of Science*. [Online] https://apps.webofknowledge.com/UA_GeneralSearch_input.do?product=UA&search_mo

de=GeneralSearch&SID=F1ArCdmHurumgsKc6yW&preferencesSaved= (Página accedida a 9 de abril de 2021).
(17) Lavine, B., Workman, J. *Chemometrics*. Anal. Chem **2008**, 80, 4519-4531.
(18) Van der Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., Van der Werd, M. J. *Centering, scaling, and transformations: improving the biological information content of metabolomics data.* BMC Genomics **2006**, 7, 142, doi: 10.1186/1471-2164-7-142
(19) Vignoli, A., Ghini, V., Meoni, G., Licari, C., Takis, P. G., Tenori, L., Turano, P., Luchinat C. *High-Troughput Metabolomics by 1D NMR.* Angew. Chem. Int. **2019**, 58, 968-994, doi:10.1002/anie.201804736.
(20) Zacharias, H. U., Altenbuchinger, M., Gronwald, W. *Statistical Analysis of NMR Fingerprints: Established Methods and Recent Advances.* Metabolites **2018**, 8, 47, doi:10.3390/metabo8030047.
(21) Karaman, I. *Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis (chapter 6).* Metabolomics: From Fundamental to Clinical Applications, Advances in Experimental Medicine and Biology, A. Sussulini (ed.) Springer **2017**, 145-160.
(22) Software CitnetExplorer ver. 1.0.0, **2020**; Centre for Science and Technology Studies, Leiden University, The Netherlands
(23) Software VosViewer ver. 1.6.15, **2020**; Centre for Science and Technology Studies, Leiden University, The Netherlands.
(24) Ludwig, C., Ward, D. G., Martin. A., Viant, M. R., Ismail, T., Johnson, P. J., Wakelam M. J. O., Günther, U. L. *Fast targeted multidimensional NMR metabolomics of colorectal cancer*. Magn. Reson. Chem. **2009**, 47, S68-S73, doi: 10.1002/mrc.2519.
(25) Qiu. Y., Cai, G., Su, M., Chen, T., Zheng, X., Xu, Y., Ni, Y., Zhao, A., Xu, L. X., Sanjun, C., Jia, W. *Serum Metabolite Profiling of Human Colorectal Cancer Using GC-TOFMS and UPLC-QTOFMS*. J. Proteome Res. **2009**, 8, 4844-4850, doi: 10.1021/pr9004162.
(26) Cross, A. J., Moore, S. C., Boca, S., Huang, W. Y., Xiong, X., Stolzenberg-Solomon, R., Sinha, R., Sampson, J. N. *A prospective study of serum metabolites and colorectal cancer risk.* Cancer **2014**, 120(19), 3049-3057, doi:10.1002/cncr.28799.
(27) Zamani, Z., Arjmand, M., Farideh, V., Hosseini, S. M. E., Fazeli, S. M., Iravani, A., Bayat, P., Oghalayee, A., Mehrabanfar, M., Hosseini, R. H., Tashakorpour, M., Tafazzoli, M., Sadeghi, S. *A Metabolic Study on Colon Cancer Using $^{1}H$ Nuclear Magnetic Resonance Spectroscopy*. Biochem. Res. Int. **2014**, ID: 348712, doi:10.1155/2014/348712
(28) Deng, L., Gu, H., Zhu, J., Nagana-Gowda, G. A., Djukovic, D., Chiorean, E. G., Raftery, D. *Combining NMR and LC/MS Using Backward Variable Elimination: Metabolomics Analysis of Colorectal Cancer, Polyps and Healthy Controls.* Anal. Chem. 2016, 88(16), 7975-7983, doi:10.1021/acs.analchem.6b00885.
(29) Advanced NMR Methods and Metal-based Catalysts. *NMRMBC Research Group.* **2021** [Online] https://www.nmrmbc.com/ (Página accedida a 17 de abril de 2021).
(30) Programa *Amix* v. 3.9.112, Bruker BioSpin GmbH, Rheinstetten, Alemania [Online] https://www.bruker.com/en/products-and-solutions/mr/nmr-software.html (Página accedida a 27 de marzo de 2021).
(31) Programa *RStudio* v. 1.4.1106, PBC, Boston, MA [Online] https://www.rstudio.com/products/rstudio/download/ (Página accedida a 21 de noviembre de 2020).
(32) Grace, S.C., Hudson, D.A. *Processing and Visualization of Metabolomics Data using R.* Metabolomics: Fundamentals and Applications, Dr Jeevan Prasain (Ed.), InTech, **2016**, 4, doi: 10.5772/65405.

(33) Paquete *FactoMineR* v. 2.4 [Online] https://CRAN.R-project.org/package=FactoMineR (Página accedida a 23 de abril de 2021).
(34) Paquete *factoextra* v.1.0.7 [Online] https://CRAN.R-project.org/package=factoextra (Página accedida a 23 de abril de 2021).
(35) Paquete *ropls* v.1.24.0 [Online] https://www.bioconductor.org/packages/release/bioc/html/ropls.html (Página accedida a 30 de abril de 2021).
(36) Worley, B., Powers, R. *Multivariate Analysis in Metabolomics*. Curr. Metabolomics, **2013**, 1: 92-107, doi: 10.2174/2213235X11301010092.
(37) Programa *SIMCA* v. 14.0, Umetrics, Suecia [Online] https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca?gclid=Cj0KCQjw2NyFBhDoARIsAMtHtZ4xDJGG997b_5pSZrCqIsU7T3wP63egUDGIkA5ntqqfecqw9TlKFb4aAj0LEALw_wcB (Página accedida a 3 de mayo de 2021).
(38) Tristán, A.I.; Salmerón, A.M.; Abreu, A.C., Fernández, I.; Prados, J.C. *Metabolomics applied in human serum to predict colon cancer using NMR spectroscopy*. [Artículo en curso]
(39) *Xia Lab*. MetaboAnalyst 5.0. **2021** [Online] https://www.metaboanalyst.ca/ (Página accedida a 17 de abril de 2021).
(40) Paquete *MetaboAnalystR* v.2.0 [Online] https://www.rdocumentation.org/packages/MetaboAnalystR/versions/2.0.0 (Página accedida a 4 de mayo de 2021).
(41) Alonso, A., Marsal, A., Julià, A. *Analytical methods in untargeted metabolomics: state of the art in 2015*. Front. Bioeng. Biotechnol. **2015**, 3(23), doi: 10.3389/fbioe.2015.00023.
(42) Xia, J., Broadhurst, D.I., Wilson, M., Wishart, D.S. *Translational biomarker Discovery in clinical metabolomics: an introductory tutorial.* J. Metabolomics, **2013**, 9: 280-299, doi: 10.1007/s11306-012-0482-9.
(43) Gertsman, I,. Barshop, B.A. *Promises and Pitfalls of Untargeted Metabolomics.* J Inherit Metab Dis. **2018**, 41(3): 355-366, doi: 10.1007/s10545-017-0130-7.

# Anexo A

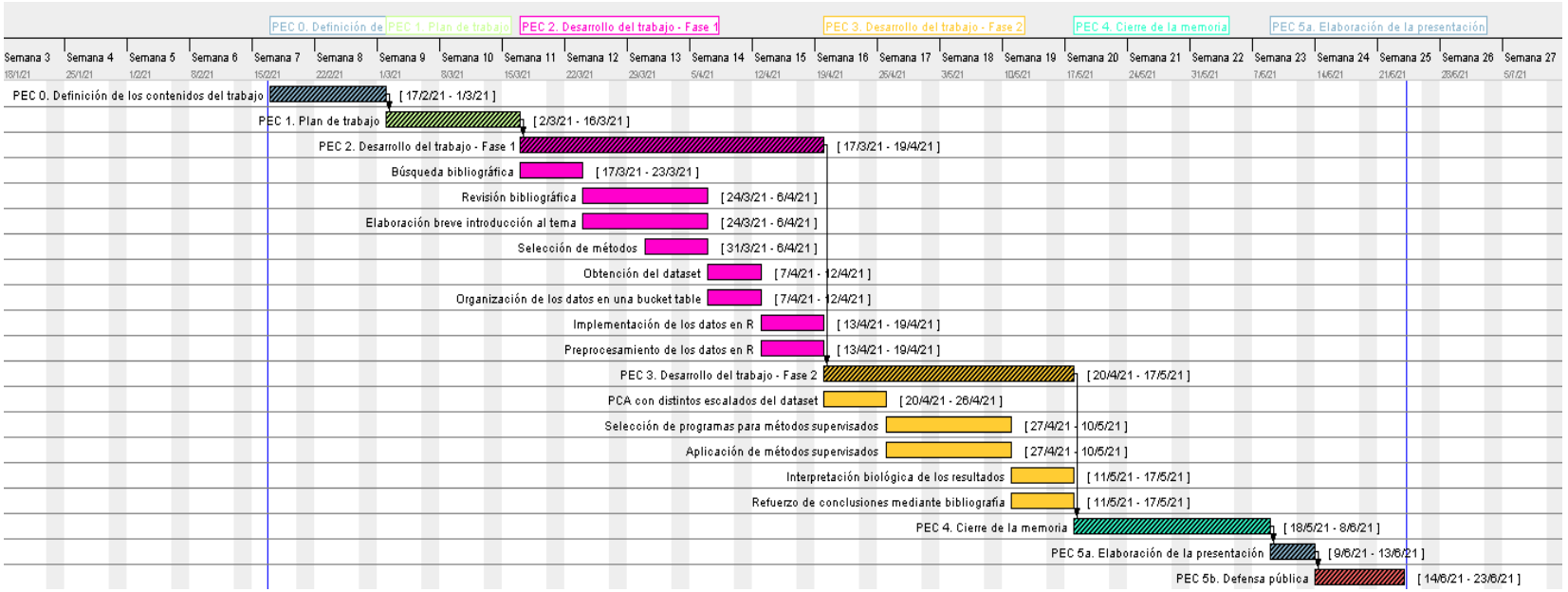
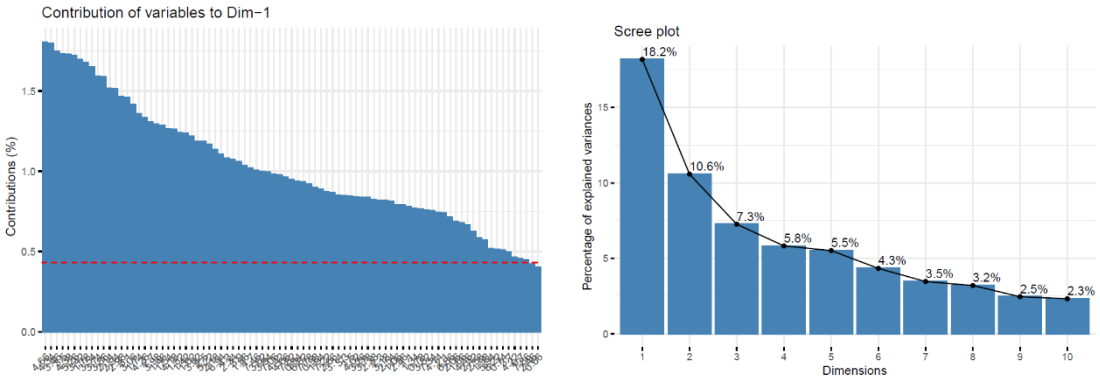
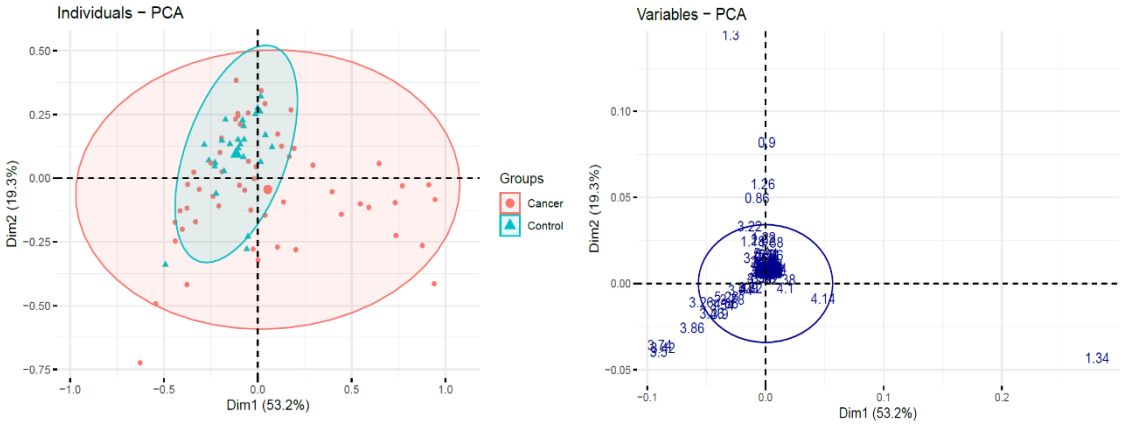



**Figura A1.** Diagrama de *Gantt* con la duración de las tareas propuestas.



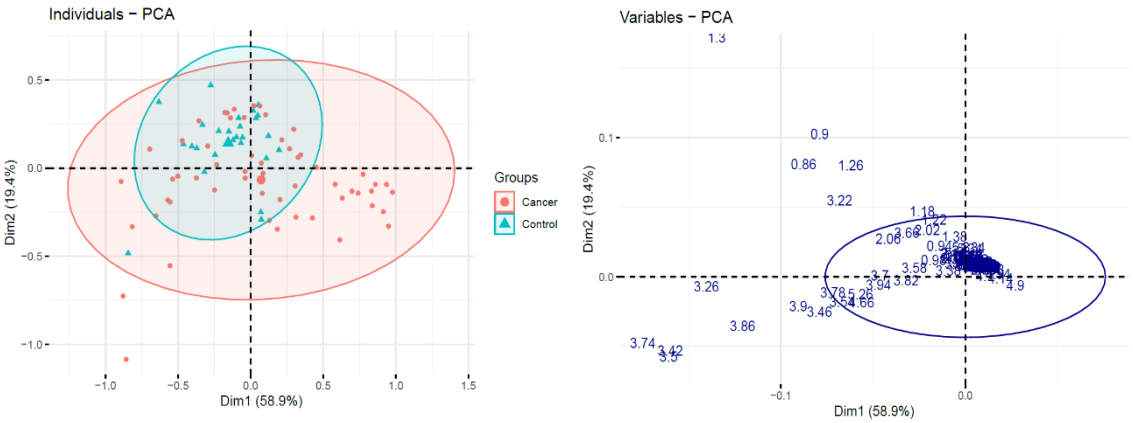

**a)** **b)**

**Figura A2.** Representación del PCA escalado a *Unit Variance* de (**a**) la contribución de las variables a la primera dimensión con un umbral de 0.47 y (**b**) porcentaje de varianza explicada por cada componente. La primera componente es la de mayor varianza explicada (18.2%), y un 58.4% de la varianza total es explicada hasta el PC8.



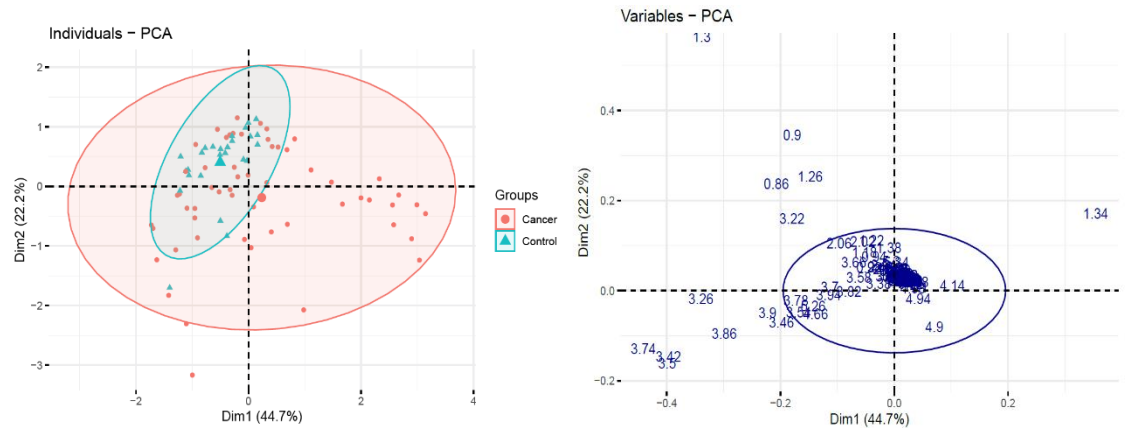

**a)** **b)**

**Figura A3.** Gráfico PCA (PC1/PC2) de (**a**) *scores* y de (**b**) *loadings* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Pareto*). Las muestras quedan agrupadas en mayoría en el centro de la gráfica, sin observarse una discriminación acentuada.




**a)** **b)**

**Figura A4.** Gráfico PCA (PC1/PC2) de (**a**) *scores* y de (**b**) *loadings* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Range Scaling*). Las muestras quedan agrupadas en mayoría en el centro de la gráfica, sin observarse una discriminación acentuada.




**a)** **b)**

**Figura A5.** Gráfico PCA (PC1/PC2) de (**a**) *scores* y de (**b**) *loadings* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Vast Scaling*). Las muestras quedan agrupadas en mayoría en el centro de la gráfica, sin observarse una discriminación acentuada.

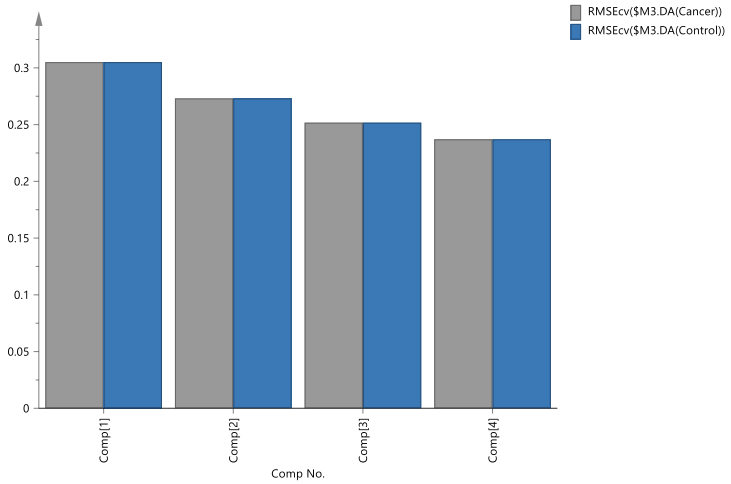



**Figura A6.** Representación del valor asociado a RMSEcv para cada componente del modelo PLS-DA escalado a *Unit Variance*. El valor de RMSEEcv asociado a la primera componente es de 0.31, que al ser inferior a 0.5 representa una capacidad predictiva precisa para el modelo. Las muestras de tipo *Cáncer* quedan representadas en color gris, y las de tipo *Control* en color azul.

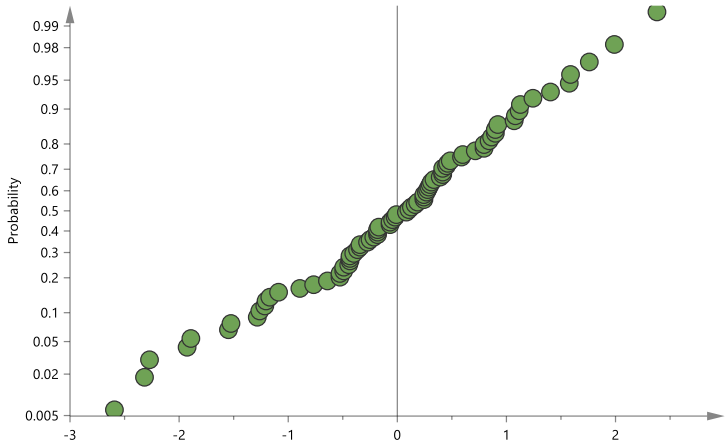



**Figura A7.** Gráfica de la normalidad de los residuos del modelo PLS-DA escalado a *Unit Variance*. Se observa un ajuste a la recta de normalidad muy adecuado, sin valores atípicos muy acentuados.

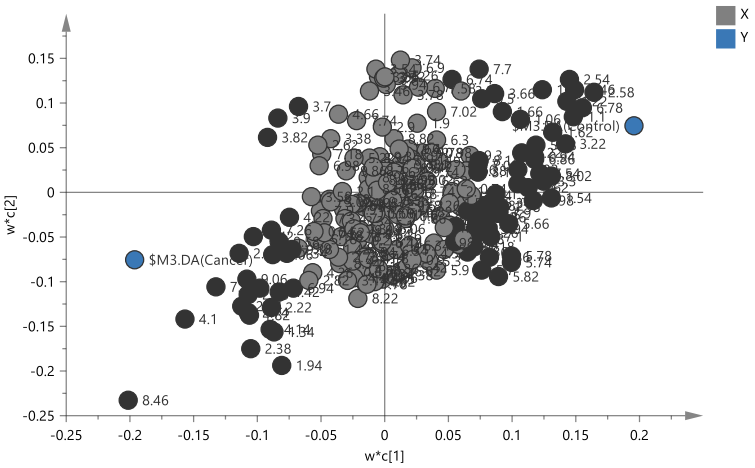



**Figura A8.** Gráfico PLS-DA de *loadings* obtenido a partir de espectros de RMN de $^{1}$H de muestras de suero de cáncer colorrectal (modelo escalado con *Unit Variance*). Pueden observarse aquellos metabolitos que aumentan según los grupos de la variable de estudio.

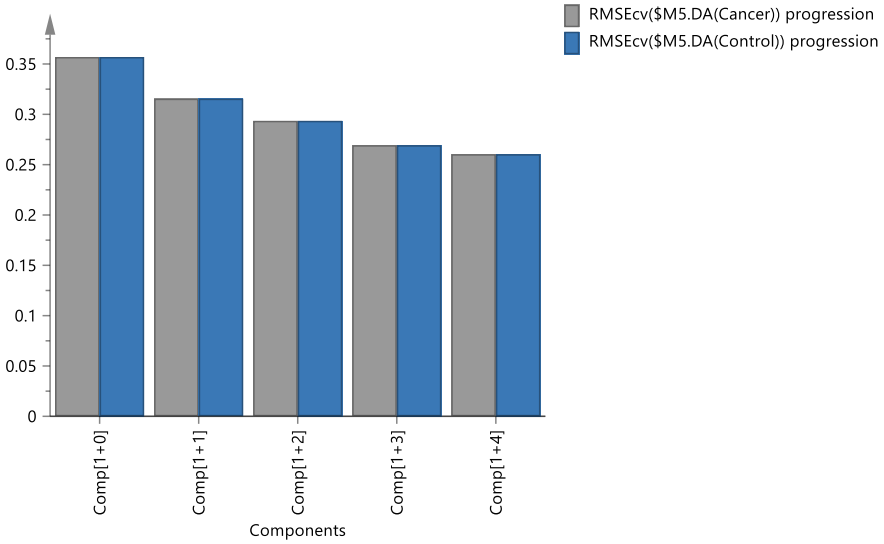



**Figura A9.** Representación del valor asociado a RMSEcv para cada componente del modelo OPLS-DA escalado a *Pareto*. El valor de RMSEEcv asociado a la primera componente es de 0.36, que al ser inferior a 0.5 representa una capacidad predictiva precisa para el modelo. Las muestras de tipo *Cáncer* quedan representadas en color gris, y las de tipo *Control* en color azul.

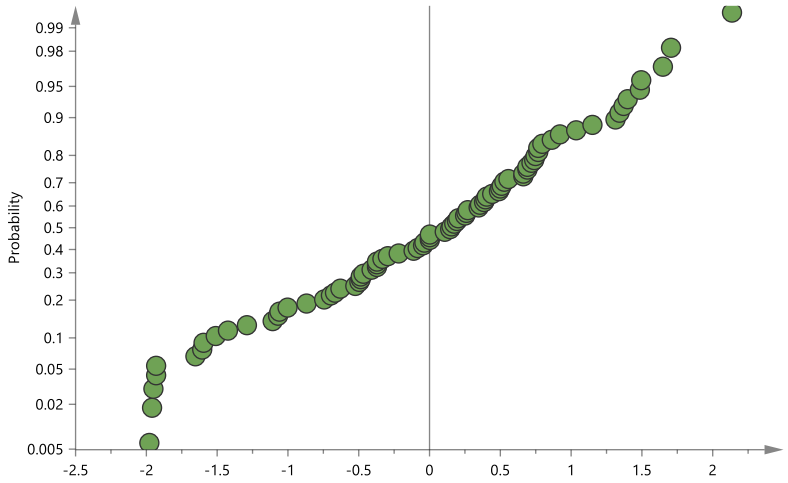



**Figura A10.** Gráfica de la normalidad de los residuos del modelo OPLS-DA escalado a *Pareto*. Se observa un ajuste a la recta de normalidad adecuado, esta vez observando una serie de *outliers*.

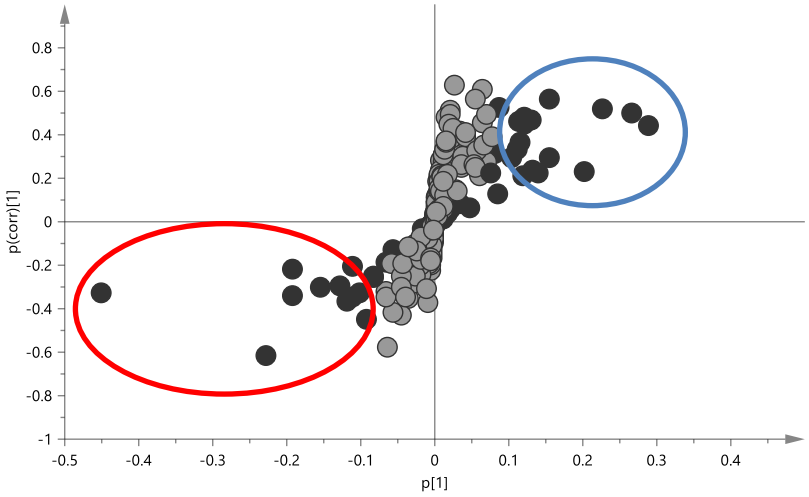



**Figura A11.** Gráfico S (*S-plot*) derivado del modelo OPLS-DA escalado con *Pareto*. Se muestra la contribución de las variables al modelo mediante la representación de la correlación de las variables frente a la magnitud de las mismas. Las variables que decrecieron para las muestras de tipo cáncer quedaron indicadas mediante un círculo azul, mientras que aquellas que aumentaron para este grupo, mediante un círculo rojo.

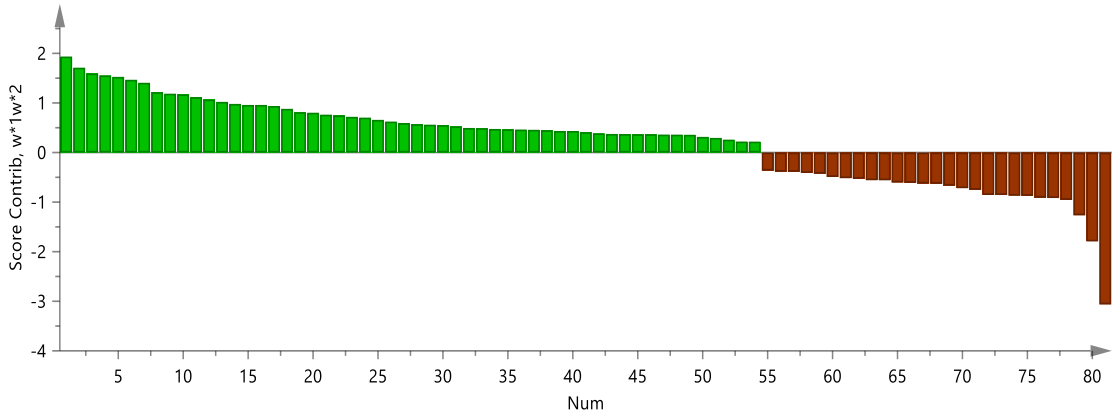



**Figura A12.** Gráfica de contribuciones generada a partir del modelo PLS-DA. En color verde quedan señalados los *buckets* más relevantes para las muestras de tipo *Control*, mientras que en color rojo quedan plasmados los correspondientes a las muestras de tipo *Cáncer*.

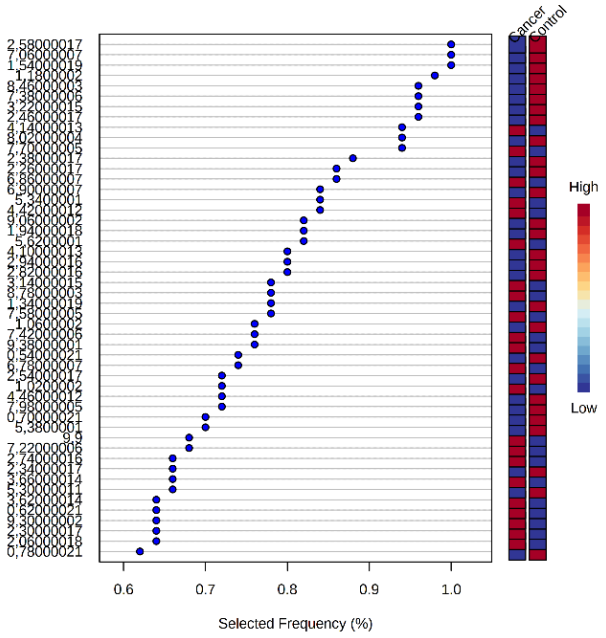



**Figura A13.** Gráfica de la frecuencia de selección de cada *bucket* generado a partir del modelo SVM de *kernel* lineal. A la derecha se observa una escala que identifica la probabilidad para cada grupo de selección de biomarcadores, siendo el color rojo la mayor probabilidad y el azul la mínima.

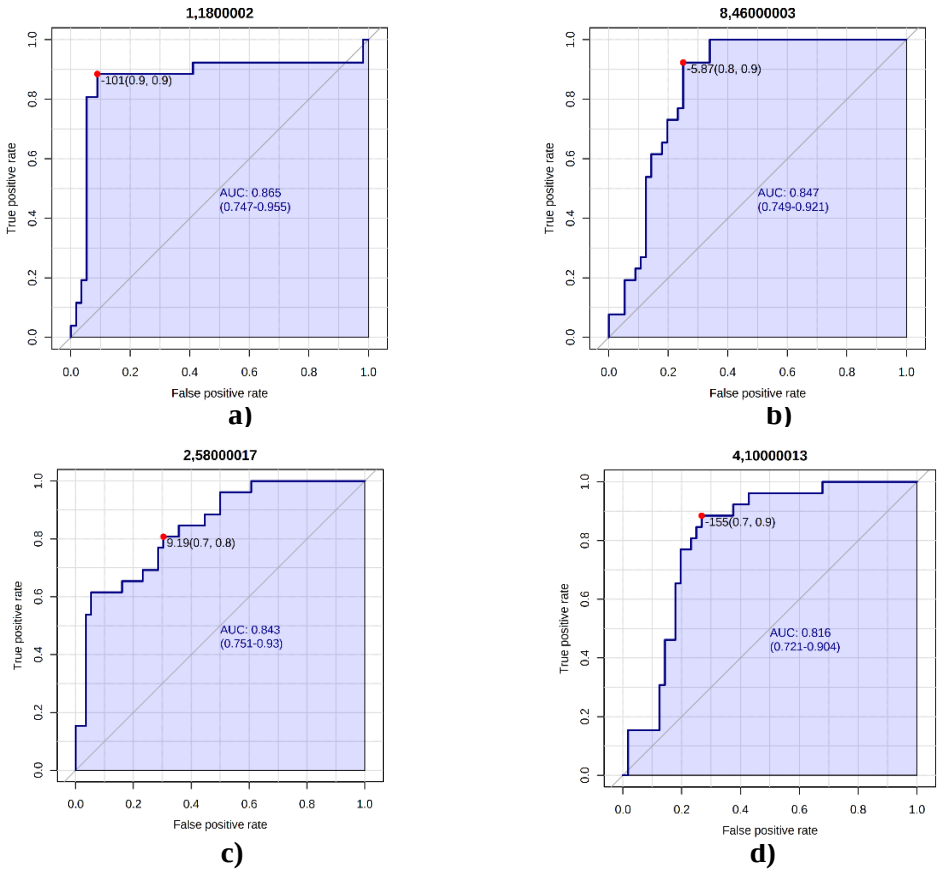



**Figura A14.** Gráficas de curvas ROC de los cuatro metabolitos con mayores niveles de AUC, en las cuales pueden observarse los Intervalos de Confianza asociados a cada uno. Como puede destacarse, todos estos intervalos al 95% comprendieron los valores 0.7-0.9. Las variables representadas son, por orden (**a**) $\delta_H$ 1.18 ppm, (**b**) $\delta_H$ 8.46 ppm, (**c**) $\delta_H$ 2.58 ppm, y (**d**) $\delta_H$ 4.10 ppm.

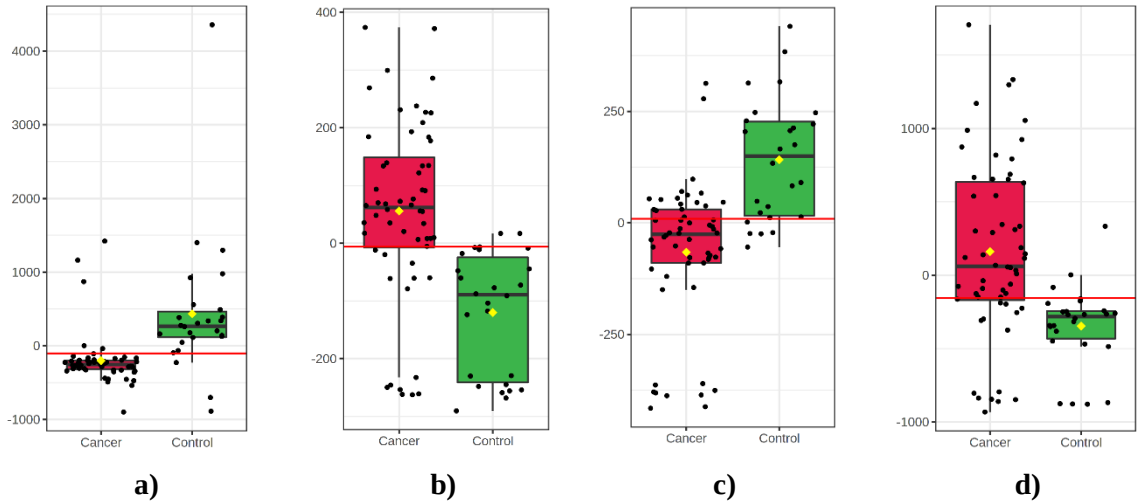



**Figura A15.** Diagramas de cajas (*boxplots*) de aquellos *buckets* con mayores valores de AUC en el modelo generado mediante RF. Las variables representadas son, por orden (**a**) $\delta_H$ 1.18 ppm, con valor AUC= 0.865, (**b**) $\delta_H$ 8.46 ppm, con valor AUC= 0.847, (**c**) $\delta_H$ 2.58 ppm, con valor AUC= 0.853, y (**d**) $\delta_H$ 4.10 ppm, con valor AUC= 0.816.

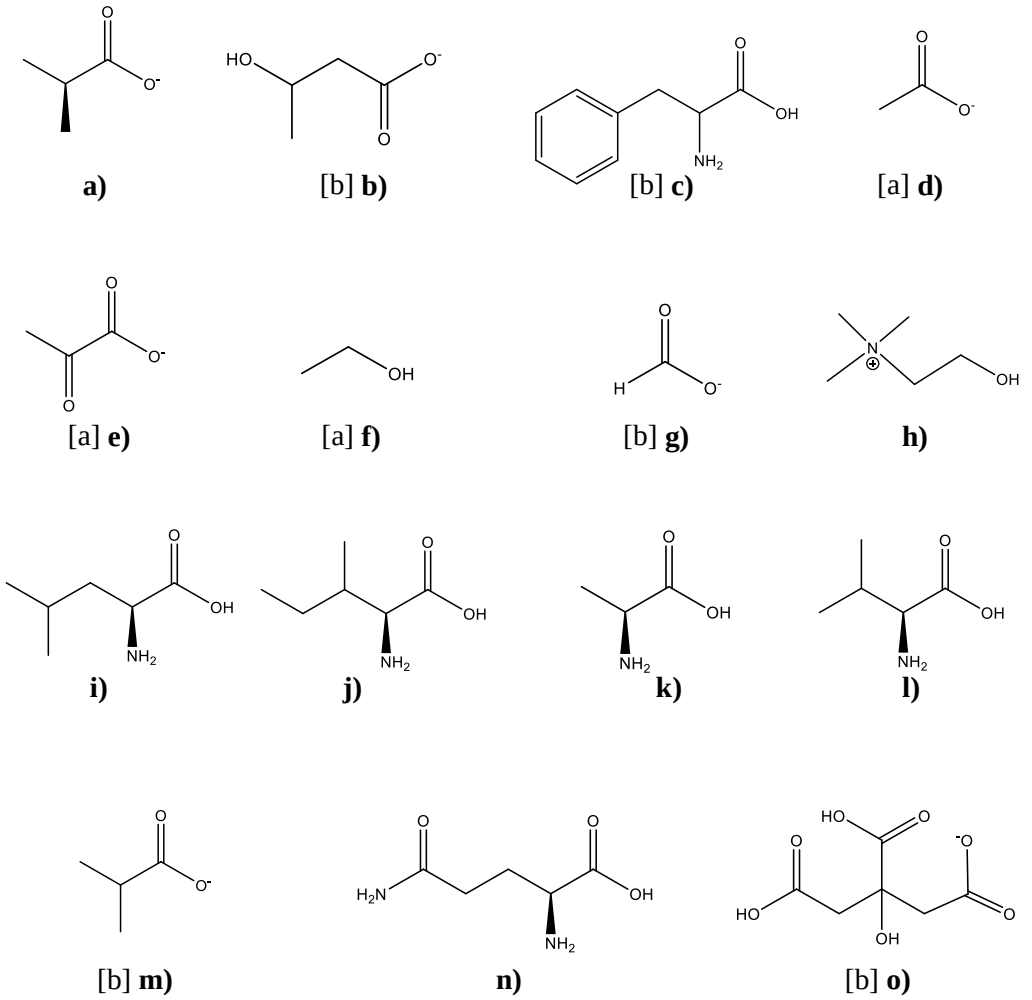



**Figura A16.** Estructuras químicas de los metabolitos identificados mediante los metodologías empleadas en este trabajo. Aquellos que aumentaron para el grupo *Cáncer* fueron (**a**) lactato, (**b**) 3-hidroxibutirato, (**c**) fenilalanina, (**d**) acetato, y (**e**) piruvato. Por otro lado, los que disminuyeron para este grupo fueron (**f**) etanol, (**g**) formiato, (**h**) colina, (**i**) leucina, (**j**) isoleucina, (**k**) alanina, (**l**) valina, (**m**) isobutirato, (**n**) glutamina, y (**o**) citrato. [a] Metabolito únicamente vislumbrado en el modelo OPLS-DA; [b] metabolito únicamente identificado en el modelo RF.

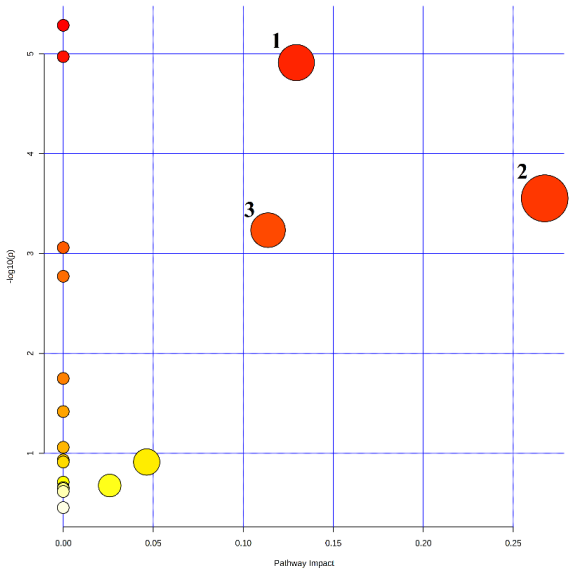



**Figura A17.** Rutas metabólicas significantemente afectadas por la presencia de Cáncer colorrectal en las muestras de suero (*Impacto>0, Coincidencia>1, p ajustada Holm* y *FDR < 0.05*). Para el análisis, fueron empleados los nombres de los biomarcadores obtenidos en el modelo OPLS-DA. 1, Glicólisis/ gluconeogénesis; 2, Metabolismo del piruvato; 3, Metabolismo de la alanina, del aspartato y del glutamato. Los valores de impacto, *p* valores, valores *p* de *Holm*, y FDR se encuentran descritos en la **Tabla A2 del anexo A**.

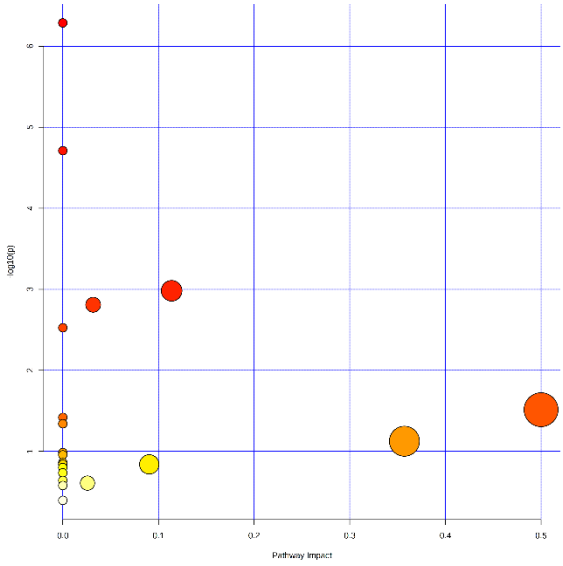



**Figura A18.** Rutas metabólicas afectadas por la presencia de Cáncer colorrectal en las muestras de suero. Ninguna de ellas cumplió los parámetros necesarios para ser considerada una ruta significantemente enriquecida (*Impacto>0, Coincidencia>1, p ajustada Holm* y *FDR < 0.05*). Los valores de impacto, *p* valores, valores *p* de *Holm*, y FDR se encuentran descritos en la **Tabla A3 del anexo A**.

**Tabla A1.** Revisión bibliográfica de aquellos artículos relacionados con el análisis metabolómico de muestras de suero de cáncer colorrectal empleando en su mayoría RMN de $^{1}H$.

| Estudio | Técnica analítica | Tamaño muestral | Modelos multivariantes | Objetivos | Resultados de metabolitos |
|---|---|---|---|---|---|
| ***Ludwig et al. (2009) [24]*** | RMN de $^{1}H$ | 38 crc[a], 8 con adenoma, y 19 controles | PCA, PLS-DA | Obtención de un modelo de clasificación e identificación de biomarcadores de crc frente a muestras control y muestras positivas en adenoma | Altos niveles de lactato y piruvato en las muestras de crc |
| ***Qiu et al. (2009) [25]*** | LC-MS y GC-TOFMS | 64 crc, y 65 controles | PCA, OPLS-DA | Obtención de un modelo de clasificación e identificación de biomarcadores de crc frente a muestras control | Altos niveles de lactato y piruvato en las muestras de crc. Disminución de triptófano, tirosina, uridina y oleamida |
| ***Cross et al. (2014) [26]*** | UPLC-MS, GC-MS | 254 crc, y 254 controles | PCA, RF | Identificación de biomarcadores de crc frente a muestras control | No se obtuvo relación entre los metabolitos y las muestras crc. Se observó una correlación positiva entre el glicoquenodesoxicolato de sodio y el grupo crc de mujeres |
| ***Zamani et al. (2014) [27]*** | RMN de $^{1}H$ | 33 crc, y 33 controles | PCA, PLS-DA | Obtención de un modelo de clasificación e identificación de biomarcadores de crc frente a muestras control | En grupo crc disminución de los niveles de piridoxina, orotidina, s-adenosilhomocisteína, piridoxamina, ácido glicocólico, β-leucina, 5-metilcitidina, ácido taurocólico, ácido 3-hidroxibutírico, 7-acetocolesterol, ácido 3-hidroxiisovalerico, l-fucosa, colesterol y L-palmitoilcarnitina, además de un aumento de glicina.<br>Proporción LCA/ DCA posible biomarcador de crc |

| | | | | | |
|---|---|---|---|---|---|
| ***Deng et al. (2016) [28]*** | LC-MS y RMN de $^{1}H$ | 28 crc, 44 con pólipos, y 55 controles | PLS-DA | Obtención de un modelo de clasificación e identificación de biomarcadores de crc y de pólipos frente a muestras control | En el grupo crc, se observaron mayores niveles de glucosa, menores de adenosina, y alteraciones en los niveles de piruvato, glutamina. Por otro lado, en el de pólipos se encontró un descenso de oroato y un aumento de adenosina.<br>Para ambos grupos se observaron alteraciones en niveles de aminoácidos, fumarato, citrato, oxaloacetato, ácido linoléico y lípidos |
| ***Gu et al. (2019) [8]*** | RMN de $^{1}H$ | 40 crc, 32 con pólipos, y 38 controles | PCA, PLS-DA, OPLS-DA, RF, SVM | Obtención de un modelo de clasificación e identificación de biomarcadores de crc y de pólipos frente a muestras control | La proporción de acetato/glicerol podría ser biomarcador de pólipos, y la de lactato/citrato de crc |

[a] Cáncer colorrectal.

**Tabla A2.** Rutas metabólicas significantemente afectadas por la presencia de Cáncer colorrectal en las muestras de suero (*Impacto>0, Coincidencia>1, p ajustada Holm* y *FDR < 0.05*). Para el análisis, fueron empleados los nombres de los biomarcadores obtenidos en el modelo OPLS-DA.

| Ruta metabólica | Coincidencia | *P* valor | *P* de *Holm* | FDR | Impacto |
|---|---|---|---|---|---|
| Glucólisis/ Gluconeogénesis | 4/26 | $1.22\times10^{-5}$ | 0.0010 | $3.4257\times10^{-4}$ | 0.1295 |
| Metabolismo del piruvato | 3/22 | $2.80\times10^{-4}$ | 0.0227 | 0.0058721 | 0.26749 |
| Metabolismo de alanina, aspartato y glutamato | 3/28 | $5.83\times10^{-4}$ | 0.0466 | 0.0097903 | 0.11378 |

**Tabla A3.** Rutas metabólicas significantemente afectadas por la presencia de Cáncer colorrectal en las muestras de suero (*Impacto>0, Coincidencia>1, p ajustada Holm* y *FDR < 0.05*). Para el análisis, fueron empleados los nombres de los biomarcadores obtenidos en el modelo RF.

| Ruta metabólica | Coincidencia | *P* valor | *P* de *Holm* | FDR | Impacto |
|---|---|---|---|---|---|
| Metabolismo de alanina, aspartato y glutamato | 3/28 | 0.001 | 0.0856 | 0.0292 | 0.11378 |
| Metabolismo del glioxilato y dicarboxilato | 3/32 | 0.002 | 0.1256 | 0.0326 | 0.03175 |

# Anexo B

**Código en R**

```
# Carga librerías correspondientes
library(readxl)
library(rJava)
library(xlsxjars)
library(xlsx)

# Lectura dataset
cancer<-read_xlsx("bucket_tablenew.xlsx")
# Nombres de columnas
colnames(cancer)<-cancer[1,]
cancer<-cancer[-1,]

# Conversión de las variables `Control/Cancer` y `Sucia/Limpia` en factores
cancer1<-as.data.frame(cancer)
cancer1$`Control/Cancer` <-as.factor(cancer1$`Control/Cancer`)
cancer1$`Sucia/Limpia`<-as.factor(cancer1$`Sucia/Limpia`)

# Comprobación de las dimensiones del dataset y
# de la posible presencia de valores faltantes
dim(cancer1)
table(is.na(cancer1))

# Creación del dataset numérico para el análisis
cancernum1<-cancer1[-c(1:3)]

# Nombres de las muestras como nombres de las columnas
rownames(cancernum1)<-cancer1$Names

# Eliminación de las variables iguales a 0, correspondientes a señales
# residuales del espectro
remcol1<-unlist(lapply(cancernum1, function(x) return (sum(x)==0)))
cancernum1<-cancernum1[,!remcol1]

# Nuevas dimensiones del dataset
dim(cancernum1)

# Creación función para escalado Unit Variance
UV_scale<-function(z){
  rowmean<-apply(z,1,mean)
  rowsd <- apply(z, 1, sd)
  rv <- sweep(z, 1, rowmean,"-")
  rv <- sweep(rv, 1, rowsd, "/")
  return(rv)
}

# Aplicación del escalado Unit Variance
```

```
cancernum0<-UV_scale(cancernum1)

# Comprobación de la aplicación del escalado a Unit Variance
boxplot(cancernum0, horizontal = F, names=F, main= "Escalado a Unit Variance")
boxplot(cancernum1, horizontal = F, names=F, main= "Sin escalado")

# Aplicación PCA Unit Variance
library(FactoMineR)
PCA_canc1<-prcomp(cancernum1, scale=TRUE)

# Representación PCA Unit Variance
library(factoextra)
library(ggplot2)
fviz_pca_ind(PCA_canc1, habillage=cancer1$`Sucia/Limpia`, addEllipses = TRUE, geom="point"
             ellipse.level=0.95, ggtheme= theme_minimal())

# Representación loadings Unit Variance
fviz_pca_var(PCA_canc1, addEllipses = TRUE, geom="text",  label = "all",
             col.var="darkblue", ellipse.level=0.95, ggtheme= theme_minimal())

# Eliminación de las variables de tipo `Sucia` Unit Variance
cancer11<-cancer1[!(cancer1$`Sucia/Limpia`=="Sucia"), ]
cancernum11<-cancer11[-c(1,2,3)]

# Eliminación de las variables iguales a 0
remcol11<-unlist(lapply(cancernum11, function(x) return (sum(x)==0)))
cancernum11<-cancernum11[,!remcol11]
rownames(cancernum11)<-cancer11$Names

# Aplicación PCA Unit Variance
PCA_cancl1<-prcomp(cancernum11, scale=TRUE)

# Representación PCA scores Unit Variance
fviz_pca_ind(PCA_cancl1, habillage=cancer11$`Control/Cancer`, addEllipses = TRUE,
             geom="point", ellipse.level=0.95, ggtheme= theme_minimal())

# Representación loadings Unit Variance
fviz_pca_var(PCA_cancl1, addEllipses = TRUE, geom="text",  label = "all",
             col.var="darkblue", ellipse.level=0.95, ggtheme= theme_minimal())

# Valores de loadings PCA Unit Variance
print("Seis primeros registros de Loadings")
head(PCA_cancl1$rotation)[,1:5]

# Valores de scores PCA Unit Variance
print("Seis primeros registros de Scores")
rownames(PCA_cancl1$x)<-cancer11$Names
head(PCA_cancl1$x)[,1:5]

# Representación de las variables que más contribuyen PCA Unit Variance
fviz_contrib(PCA_cancl1, choice="var", axes=1, top=85)

# Representación de la varianza explicada PCA Unit Variance
```

```
fviz_screeplot(PCA_cancl1, addlabels=TRUE)

# Creación función para escalado mediante Pareto
pareto_scale<-function(z){
  rowmean<-apply(z,1,mean)
  rowsd <- apply(z, 1, sd)
  rowsqrtsd <- sqrt(rowsd)
  rv <- sweep(z, 1, rowmean,"-")
  rv <- sweep(rv, 1, rowsqrtsd, "/")
  return(rv)
}

# Aplicación del escalado Pareto
cancernum2<-pareto_scale(cancernum1)

# Comprobación de la aplicación del escalado a Pareto
boxplot(cancernum2, horizontal = F, names=F, main= "Escalado a Pareto")
boxplot(cancernum1, horizontal = F, names=F, main= "Sin escalado")

# Eliminación muestras sucias
cancerl2<-cancer1[!(cancer1$`Sucia/Limpia`=="Sucia"), ]
cancernuml2<-cancerl2[-c(1,2,3)]

# Eliminación de las variables con 0
remcoll2<-unlist(lapply(cancernuml2, function(x) return (sum(x)==0)))
cancernuml2<-cancernuml2[,!remcoll2]
rownames(cancernuml2)<-cancerl2$Names

# Aplicación del escalado Pareto
cancernuml2<-pareto_scale(cancernuml2)

# Aplicación PCA Pareto
PCA_cancl2<-prcomp(cancernuml2, scale=FALSE)

# Representación PCA Pareto
fviz_pca_ind(PCA_cancl2, habillage=cancerl2$`Control/Cancer`, addEllipses = TRUE,
             geom="point", ellipse.level=0.95, ggtheme= theme_minimal())

# Representación loadings Pareto
fviz_pca_var(PCA_cancl2, addEllipses = TRUE, geom="text",  label = "all",
             col.var="darkblue", ellipse.level=0.95, ggtheme= theme_minimal())

# Valores de loadings PCA Pareto
print("Seis primeros registros de Loadings")
head(PCA_cancl2$rotation)[,1:5]

# Valores de scores PCA Pareto
print("Seis primeros registros de Scores")
rownames(PCA_cancl2$x)<-cancerl2$Names
head(PCA_cancl2$x)[,1:5]

# Representación de las variables que más contribuyen PCA Pareto
fviz_contrib(PCA_cancl2, choice="var", axes=1, top=15)
```

```r
# Representación de la varianza explicada PCA Pareto
fviz_screeplot(PCA_cancl2, addlabels=TRUE)

# Creación función para escalado mediante Range Scaling
range_scale<-function(z){
  rowmean<-apply(z,1,mean)
  maxran<-apply(z,1,max)
  minran<-apply(z,1,min)
  rv<-(z-rowmean)/(maxran-minran)
  return(rv)
}

# Aplicación del escalado Range Scaling
cancernum3<-range_scale(cancernum1)

# Comprobación de la aplicación del escalado a Range Scaling
boxplot(cancernum3, horizontal = F, names=F, main= "Escalado mediant Range Scaling")
boxplot(cancernum1, horizontal = F, names=F, main= "Sin escalado")

# Eliminación muestras sucias
cancerl3<-cancer1[!(cancer1$`Sucia/Limpia`=="Sucia"), ]
cancernuml3<-cancerl3[-c(1,2,3)]

# Eliminación de las variables con 0
remcoll3<-unlist(lapply(cancernuml3, function(x) return (sum(x)==0)))
cancernuml3<-cancernuml3[,!remcoll3]
rownames(cancernuml3)<-cancerl3$Names

# Aplicación del escalado Range Scaling
cancernuml3<-range_scale(cancernuml3)

# Aplicación PCA Range Scaling
PCA_cancl3<-prcomp(cancernuml3, scale=FALSE)

# Representación PCA Range Scaling
fviz_pca_ind(PCA_cancl3, habillage=cancerl3$`Control/Cancer`, addEllipses = TRUE,
             geom="point", ellipse.level=0.95, ggtheme= theme_minimal())

# Representación loadings
fviz_pca_var(PCA_cancl3, addEllipses = TRUE, geom="text",  label = "all",
             col.var="darkblue", ellipse.level=0.95, ggtheme= theme_minimal())

# Valores de loadings PCA Range Scaling
print("Seis primeros registros de Loadings")
head(PCA_cancl3$rotation)[,1:5]

# Valores de scores PCA Range Scaling
print("Seis primeros registros de Scores")
rownames(PCA_cancl3$x)<-cancerl3$Names
head(PCA_cancl3$x)[,1:5]

# Representación de las variables que más contribuyen PCA Range Scaling
fviz_contrib(PCA_cancl3, choice="var", axes=1, top=21)
```

```
# Representación de la varianza explicada PCA Range Scaling
fviz_screeplot(PCA_cancl3, addlabels=TRUE)

# Creación función para escalado mediante Vast Scaling
vast_scale<-function(z){
  rowmean<-apply(z,1,mean)
  rowdev<-apply(z,1,sd)
  rv<-((z-rowmean)/(rowdev))*((rowmean)/(rowdev))
  return(rv)
}

# Aplicación del escalado Vast Scaling
cancernum4<-vast_scale(cancernum1)

# Comprobación de la aplicación del escalado a Vast Scaling
boxplot(cancernum4, horizontal = F, names=F, main= "Escalado mediante Vast Scaling")
boxplot(cancernum1, horizontal = F, names=F, main= "Sin escalado")

# Eliminación muestras sucias
cancerl4<-cancer1[!(cancer1$`Sucia/Limpia`=="Sucia"), ]
cancernuml4<-cancerl4[-c(1,2,3)]

# Eliminación de las variables con 0
remcoll4<-unlist(lapply(cancernuml4, function(x) return (sum(x)==0)))
cancernuml4<-cancernuml4[,!remcoll4]
rownames(cancernuml4)<-cancerl4$Names

# Aplicación del escalado Vast Scaling
cancernuml4<-vast_scale(cancernuml4)

# Aplicación PCA Vast Scaling
PCA_cancl4<-prcomp(cancernuml4, scale=FALSE)

# Representación PCA Vast Scaling
fviz_pca_ind(PCA_cancl4, habillage=cancerl4$`Control/Cancer`, addEllipses = TRUE,
             geom="point", ellipse.level=0.95, ggtheme= theme_minimal())

# Representación loadings PCA Vast Scaling
fviz_pca_var(PCA_cancl4, addEllipses = TRUE, geom="text",  label = "all",
             col.var="darkblue", ellipse.level=0.95, ggtheme= theme_minimal())

# Valores de loadings PCA Vast Scaling
print("Seis primeros registros de Loadings")
head(PCA_cancl4$rotation)[,1:5]

# Valores de scores PCA Vast Scaling
print("Seis primeros registros de Scores")
rownames(PCA_cancl4$x)<-cancerl4$Names
head(PCA_cancl4$x)[,1:5]

# Representación de las variables que más contribuyen PCA Vast Scaling
fviz_contrib(PCA_cancl4, choice="var", axes=1, top=21)
```

```
# Representación de la varianza explicada
fviz_screeplot(PCA_cancl4, addlabels=TRUE)

# Preparación dataset muestras limpias
cancerp<-as.data.frame(cancer)
cancerp$`Control/Cancer` <-as.factor(cancerp$`Control/Cancer`)
cancerp$`Sucia/Limpia`<-as.factor(cancerp$`Sucia/Limpia`)
rownames(cancerp)<-cancerp$Names
cancerlp<-cancerp[!(cancerp$`Sucia/Limpia`=="Sucia"), ]
cancerlp<-cancerlp[-c(1,2)]

# Eliminación de las variables con 0
remcollp<-unlist(lapply(cancerlp[-1], function(x) return (sum(x)==0)))
cancerlp1<-cancerlp[-1][,!remcollp]

library(ropls)
# PLS-DA Unit Variance
plsdacanc<-opls(cancerlp1,cancerlp$`Control/Cancer`, scaleC="standard")

# Scores PLS-DA Unit Variance
plot(plsdacanc, typeVc="x-score", parPaletteVc=c("coral1", "cyan3"))

# OPLS-DA Pareto
oplsdacanc<-opls(cancerlp1,cancerlp$`Control/Cancer`, scaleC="pareto", predI=1, orthoI=NA)

# Scores OPLS-DA Pareto
plot(oplsdacanc, typeVc="x-score", parPaletteVc=c("coral1", "cyan3"))
```