E.T.S. de Ingeniería Industrial,
Informática y de Telecomunicación

Optimización del análisis de datos proteómicos procedentes del espectrómetro de masas Triple TOF 5600

Máster en Ingeniería Biomédica

Trabajo Fin de Máster

Xabier Martínez de Morentin Iribarren
Director; Javier Rodríguez Falces
Co-director; Joaquín Fernández Irigoyen
Pamplona, 29 Febrero 2016

upna Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

"Una de las cosas más fascinantes de los programadores es que no puedes saber si están trabajando o no sólo con mirarlos. A menudo están sentados aparentemente tomando café, chismorreando o mirando a las nubes. Sin embargo, es posible que estén poniendo en orden todas las ideas individuales y sin relación que pululan por su mente"
-- Charles M. Strauss

Resumen

Gracias a la infraestructura presente en la Unidad de Proteómica de Navarrabiomed en la que hay presentes máquinas como el espectrómetro de masas Triple TOF 5600 acoplado a un cromatógrafo líquido de alta resolución (HPLC), se han podido llevar a cabo los flujos de trabajo explicados en esta memoria. Los experimentos de proteómica realizados en esta unidad persiguen la identificación y cuantificación de proteínas entre dos o más condiciones experimentales y caracterizar aquellas que sean diferenciales en base a su cuantificación relativa. Esto permite por ejemplo cuantificar las diferencias entre sujetos sanos y sujetos con afecciones o ver las diferencias de expresión entre células estimuladas con un tratamiento y sin tratar.

Además veremos cómo no todos los programas diseñados para el análisis de los datos obtenidos por estas técnicas se comportan igual y decidiremos cuales son más eficientes en nuestro caso concreto y para el espectrómetro de masas Triple TOF 5600. De igual modo propondremos flujos de análisis bioinformáticos para extraer la información contenida en los experimentos de tipo diferencial, ya sea con R o Python.

La metodología desarrollada se utiliza hoy en día de manera rutinaria para analizar los datos de la Unidad de Proteómica y ha sido empleada con éxito en varias publicaciones científicas como son las presentes en el anexo 1.

Palabras clave: Proteómica, Cromatografía líquida de alta resolución, Espectrometría de masas, Triple TOF 5600, iTRAQ, Label-free, MaxQuant, Progenesis, R, Python.

Índice

1.1 ¿Qué son y de dónde vienen las proteínas?

El proceso por el cual se sintetizan las proteínas en bastante complejo y en él intervienen multitud de procesos.

Los nucleótidos son las unidades más básicas que componen los ácidos nucleicos, el ácido desoxirribonucleico (DNA) y el ácido ribonucleico (RNA). Estos están formados por una base nitrogenada (purinas y pirimidinas), una pentosa (desoxiribosa en DNA y ribosa en RNA) y un grupo fosfato, unidos mediante enlaces covalentes [1] como podemos ver en la figura 1.

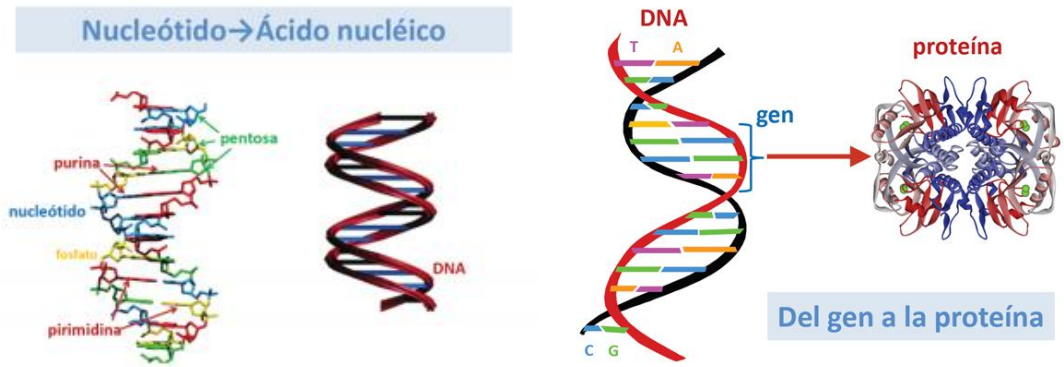


Figura 1: Estructura de DNA

Los nucleótidos que conforman el DNA se disponen en dos largas cadenas con forma helicoidal (doble hélice). Esta estructura se debe a la capacidad de las bases nitrogenadas de formar enlaces entre sí, pares de bases (adenina-timina (A-T) y citosina-guanina (C-G)). Las pentosas y los grupos fosfato se disponen verticalmente a los laterales de la cadena mientras que estos pares de bases serían las uniones horizontales de la estructura.

Este DNA se encuentra tanto en el núcleo celular como en la mitocondria. El DNA contiene la información para la síntesis de todas las proteínas y es el mismo en todas las células de un individuo. Esto se debe a la capacidad del DNA para replicarse a si mismo duplicando las cadenas de nucleótidos.

En la molécula de DNA se encuentran unas nuevas estructuras denominadas genes. Estos genes están formados por secuencias de nucleótidos de diferentes longitudes. En los genes se distinguen los exones, partes codificantes que darán lugar a las proteínas y los intrones, regiones encargadas de la transcripción primaria de RNA estos pueden verse en la figura 2.

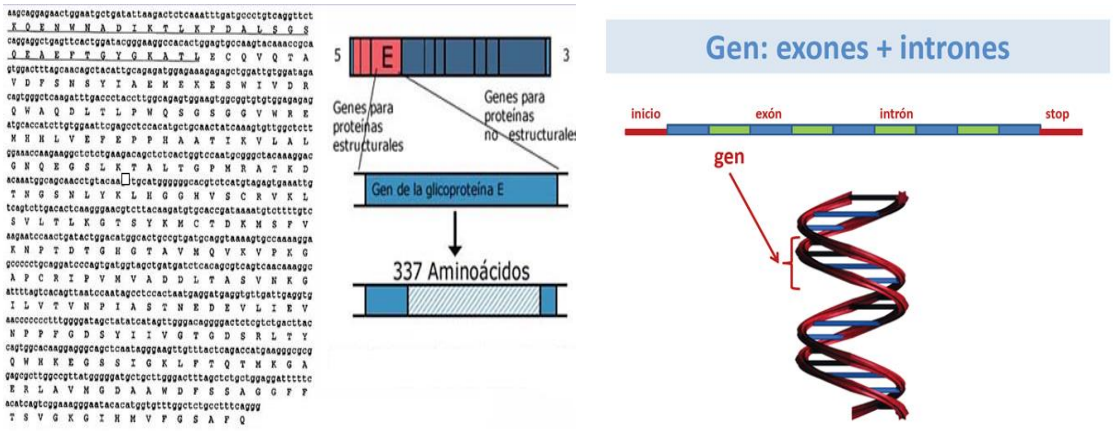


Figura 2: Secuencia de nucleótidos [2] y de aminoácidos del gen de la glicoproteína E

La información genética está contenida en códigos formados por los cuatro nucleótidos citados (A, T, C Y G) como podemos ver en el panel izquierda de la figura 2 donde cada letra. A los conjuntos de tres bases formados por estos nucleótidos se les denominan codones y las diferentes combinaciones de estos nucleótidos en cada codón determinan la secuencia aminoacídica de la proteína, estos pueden verse con letras mayúsculas en el panel de la izquierda de la figura 2.

Dicho esto es necesaria una comunicación para poder regular la expresión génica de los genes de modo que se sintetice una proteína o un conjunto de proteínas si es requerido. Para ello existen la transcripción y la traducción.

En la transcripción la información del gen que se desea expresar es transferida a un RNA mensajero (mRNA), de modo que esta información llega al citoplasma para su producción. En este proceso se forma el mRNA mediante complementariedad de nucleótidos, ya que por cada nucleótido del DNA, le corresponde el complementario en la secuencia de mRNA. De esta manera la secuencia de mRNA tiene toda la información necesaria para la síntesis de la proteína. En este proceso se eliminan los intrones, quedando únicamente los exones, con la información codificante. A este proceso se le llama *"splicing"*.

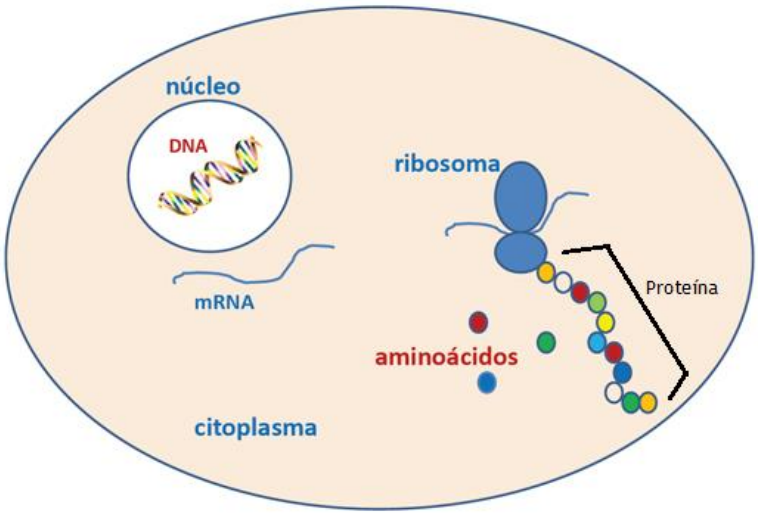


Figura 3: Proceso de traducción

La traducción es el proceso de lectura de la secuencia de mRNA llevada a cabo en los ribosomas celulares como podemos ver representado en la figura 3. Estos ribosomas son capaces de leer y traducir las secuencias de bases o codones del mRNA a la cadena de aminoácidos que conforman una proteína.

Las proteínas son moléculas complejas formadas por cadenas de aminoácidos, que forman parte de las células y los tejidos, ya que son componentes principales en las rutas metabólicas de las células. El conjunto de todas las proteínas y sus modificaciones producidas por un tejido, se denomina proteoma.

En organismos eucariotas existen veinte aminoácidos que constituyen las proteínas. Estas cadenas de aminoácidos adquieren una disposición con forma tridimensional en el espacio que les permite realizar correctamente su función.

1.2. ¿Qué es la proteómica?

La proteómica es el estudio y caracterización del conjunto de proteínas expresadas por un genoma (proteoma). Las técnicas de proteómica abordan el estudio de este conjunto de proteínas. La proteómica permite identificar, cuantificar, categorizar y clasificar las proteínas con respecto a su función y a las interacciones que establecen entre ellas. De este modo, se pueden caracterizar las redes funcionales que establecen las proteínas y su dinámica durante procesos fisiológicos y patológicos. A menudo se realizan estudios a gran escala *"shotgun"* (analizando cientos o miles de proteínas en un único experimento) o estudios dirigidos a analizar proteínas concretas.

Es por esta razón que las proteínas pueden ser motivo de dianas terapéuticas, con fines como biomarcadores ya sean para la monitorización de enfermedades o para anticiparse al diagnóstico de estas enfermedades (método de detección precoz) o bien para la generación de fármacos para el tratamiento de enfermedades [3].

En los estudios proteómicos de "*shotgun"*, la investigación se dirige a la identificación y cuantificación del proteoma de una especie. Esta descripción del proteoma nos permite tener una imagen de las proteínas expresadas en un momento dado y bajo unas determinadas condiciones (tiempo y ambiente), así como la abundancia relativa de estas proteínas si el estudio estaba dirigido al análisis diferencial del proteoma en diferentes condiciones biológicas. De esta manera seremos capaces de identificar posibles proteínas cuya presencia, ausencia o alteración se corresponde con unos determinados estadíos de interés. Este es el motivo por el cual la proteómica nos permite encontrar biomarcadores, ya que podemos encontrar proteínas que nos permitan identificar la presencia-ausencia de una enfermedad así como su evolución.

La proteómica es una ciencia cuya consolidación es bastante reciente (mediados de los 90). Para ello fueron importantes ciertos avances tecnológicos como la aplicación de la espectrometría de masas para el análisis de moléculas biológicas, el creciente aumento de información curada en las bases de datos de proteínas y genes además de la aplicación de potentes métodos de fraccionamiento y separación de péptidos y proteínas como la cromatografía liquida de alta resolución (HPLC).

No obstante, existe una complejidad inherente a la proteómica con respecto a su predecesora, la genómica. Mientras que el genoma de un individuo es relativamente homogéneo y estático, el proteoma es dinámico, difiere de una célula a otra y de un instante de tiempo a otro. Esto lleva a que un mismo genoma pueda dar lugar a diferentes proteomas. Para mayor complejidad, existe otro factor adicional, como son las modificaciones que pueden sufrir las estructuras o las secuencias básicas de la proteína. Estas modificaciones son debidas mayoritariamente al *"splicing"* alternativo de los mRNA y a las modificaciones postraduccionales (fosforilaciones, metilaciones, acetilaciones, oxidaciones, etc), que normalmente modifican la actividad, función o localización de las proteínas, en presencia de diferentes contextos fisiológicos. Por tanto, el número de proteínas es exponencialmente mayor al número de genes existentes en un genoma.

1.3. Estado actual y futuro de la proteómica

En la actualidad las áreas de estudio de la proteómica se pueden catalogar como:

- Identificación de proteínas y caracterización de sus modificaciones postraduccionales.
- Proteómica de "expresión diferencial".
- Estudio de las interacciones proteína - proteína.
- Integración "ómica" de la proteómica con el resto de ómicas (genómica, transcriptómica, RNAsec, epigenómica, etc).

Por otra parte, las líneas a futuro [4] de la proteómica tienen como objeto cinco puntos de interés listados a continuación:

1. Accesibilidad

   La proteómica es actualmente una tecnología costosa por ello es necesario el desarrollo (ya en proceso) de softwares gratuitos de alta calidad. También sería necesario reducir el precio de los equipos. De esta forma se podrían analizar proteomas a bajo coste con tecnologías robustas y de amplia implantación como lo son los microarrays.

2. Sensibilidad

   Los estudios realizados a nivel de un organismo o tejido a menudo enmascaran información biológica. Para ello es necesario mejorar las tecnologías para que sean capaces de extraer y procesar la información con una resolución a nivel de células independientes.

3. Resolución

   Actualmente la espectrometría de masas es capaz de detectar el proteoma entero de especies de baja complejidad genómica como la levadura, no obstante cuando se trata de proteomas más complejos como el humano, no se

es capaz de detectar todo el proteoma. Por ello es necesario mejorar la preparación de muestras para poder cubrir las proteínas poco abundantes o las proteínas de membrana, así como la creación de flujos de enriquecimiento de modificaciones sin estudiar o la mejora en la interpretación de los resultados biológicos. Actualmente el proyecto *"Human Protein Atlas"* ha conseguido cubrir el 80% de las proteínas codificadas en el genoma.

4. Caracterización de isoformas

   Los flujos analíticos deben de realizarse permitiendo la integración de múltiples ómicas. El mayor reto en este campo es la de obtener la información de las proteínas a nivel de isoformas. Un gen puede dar lugar a múltiples proteínas y por tanto la relación proteína-gen debe tener esto en cuenta.

5. Integración de datos multi-ómicos

   Actualmente existe una baja correlación en los datos obtenidos mediante experimentos de proteómica y transcriptómica. El objetivo en este campo es la de reducir las discrepancias generadas al realizar las aproximaciones de manera independientemente para así poder comprender los principios de la co-regulación (gen-proteína) y sus alteraciones que tienen lugar durante la progresión de enfermedades.

Capítulo 2. Cromatografía líquida de alta resolución y Espectrometría de masas

2.1. Cromatografía liquida de alta resolución (HPLC)

La cromatografía liquida de alta resolución (HPLC) es una técnica empleada para separar los componentes que están presentes en una mezcla basándose en diferentes tipos de interacciones entre la sustancia analizada y la columna empleada en la cromatografía [5].

La mezcla introducida al HPLC será en nuestro caso un conjunto de miles de péptidos procedentes de una muestra biológica o un conjunto de muestras biológicas.

Los péptidos son moléculas formadas por la unión de varios aminoácidos mediante enlaces peptídicos. Un conjunto de péptidos forman una proteína. En este caso los péptidos utilizados serán aquellos obtenidos al cortar las proteínas empleando una proteasa como puede ser la tripsina.

En esta tecnología la mezcla pasa por la columna cromatográfica a través de la fase estacionaria. Esta columna es normalmente un cilindro con pequeñas partículas con ciertas características químicas, en nuestro caso emplearemos cadenas alquil C18 $(CH_2)_{18}$ como material de la columna.

Tenemos una fase móvil liquida que mediante su bombeo a alta presión hará que la mezcla pase a través de la columna. La muestra se introduce toda junta en el sistema y los péptidos serán separados debido a las diferentes fuerzas hidrofóbicas establecidas entre los péptidos y la columna.

El grado de retención o tiempo necesario para que se produzca el paso de la mezcla presente en la fase móvil a través de la fase estacionaria, dependerá de la naturaleza propia de la mezcla, de la fase móvil y de la fase estacionaria. En la figura 4 podemos ver las diferentes partes por las que pasará la fase móvil y la mezcla de péptidos a analizar.

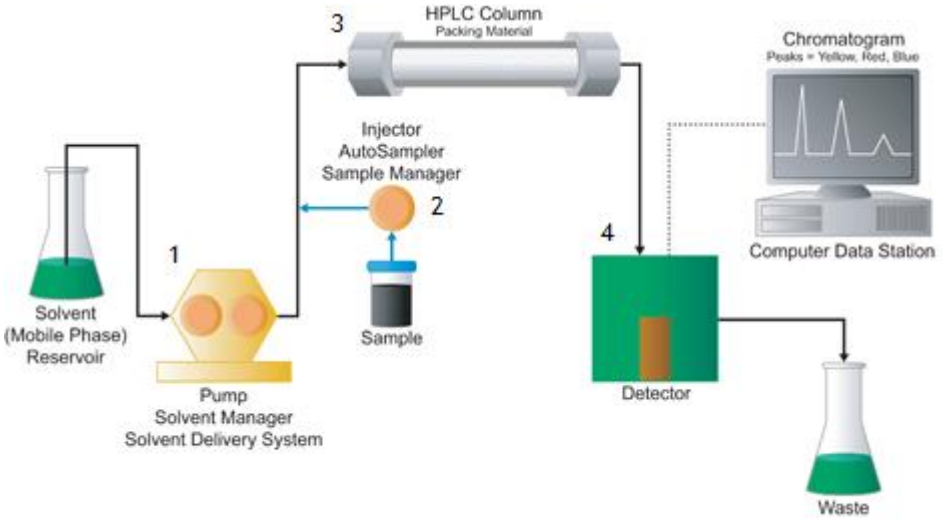


Figura 4: Componentes sistema HPLC. 1. Sistema de bombeo (*Pump*), 2. Sistema inyector de muestra (*AutoSampler*), 3. Columna, 4. Detector

En nuestro caso emplearemos un solvente orgánico, como el acetonitrilo a porcentajes de concentración variables como fase móvil. Se emplea esta fase móvil para experimentos de detección masiva de péptidos ya que en la literatura queda demostrado que esta es la que mejor interacciona con las características químicas de los péptidos [6].

El cambio en la concentración de acetonitrilo en la fase móvil provocará que los péptidos se liberen de la columna cromatográfica. Según la complejidad de la mezcla de péptidos a analizar, se empleará una rampa donde la pendiente de concentración de acetonitrilo aumentará lentamente y a lo largo del tiempo.

De esta manera los péptidos eluyen por orden de hidrofobicidad, de modo que los primeros péptidos en salir serán los hidrofílicos y después los hidrofóbicos en función de su coeficiente de reparto hidrocarburo-agua.

Esta cromatografía consiste en una fase estacionaria apolar y una móvil de polaridad moderada. El tiempo que tardan en liberarse los péptidos es mayor para aquellas moléculas de naturaleza apolar y menor para los polares. Esta cromatografía se basa en el principio de las interacciones hidrofóbicas las cuales resultan en fuerzas de repulsión entre el disolvente polar, un compuesto apolar y una fase estacionaria apolar. Cuanto menor sea el área o zona de la mezcla de los péptidos a ser analizados con el disolvente, mayor será la fuerza conductora que una a los péptidos con la fase estacionaria. El efecto hidrofóbico disminuye al introducir más disolvente apolar a la fase móvil, lo cual modifica el coeficiente de partición (cociente entre las concentraciones qué forman la fase móvil) de manera que la mezcla peptídica eluye a través de la columna. El tiempo de retención aumenta con el área de superficie hidrofóbica y normalmente esto es en ocasiones inversamente proporcional al tamaño del compuesto aunque no tiene por qué ser así.

El pH puede provocar cambios en la hidrofobicidad de los péptidos, por ello se utilizan tampones para mantener la mezcla a un pH ácido de modo que se facilite la ionización de los péptidos antes de ser introducidos en el espectrómetro de masas.
Además de la separación de mezclas complejas de péptidos, esta tecnología se puede emplear para multitud de usos como:

- Fármacos: Antibióticos, analgésicos, etc.
- Productos de alimentación: Edulcorantes artificiales, antioxidantes, aditivos, etc.
- Productos de la industria química: Aromáticos, colorantes, etc.
- Contaminantes: Pesticidas, herbicidas, etc.
- Química forense: Drogas, venenos, alcoholes en sangre, narcóticos, etc.

2.2. ¿Qué es la espectrometría de masas?

La espectrometría de masas se introdujo en el ámbito del análisis de biomoléculas al final de la década de los años 70 del siglo pasado. No fue hasta inicios de los años 90 cuando se crearon dos nuevos métodos de ionización, el electrospray (ESI) y la desorción por láser asistida por matriz (MALDI). Estos constituyeron los pilares básicos de la espectrometría de masas en la proteómica contemporánea.

La espectrometría se basa en la medición de la masa/carga de los iones presentes en la mezcla introducida, para ello se calienta un haz de los analítos (péptidos cargados) hasta vaporizarlos y se ionizan los diferentes átomos que los forman [7].

Esta tecnología tiene multitud de usos además del ya citado como:

- Industria: análisis de semiconductores, biosensores.
- Farmacéuticas: fármacos productos de síntesis química.
- Análisis forenses, perfumes, contaminación medioambiental, etc.

2.2.1 Fases y componentes en los espectrómetros de masas, Triple TOF 5600

El proceso de funcionamiento del espectrómetro de masas Triple TOF 5600 (su nombre técnico es Q-Q-TOF) se resume principalmente en cuatro pasos estos pueden verse relacionados con sus diferentes analizadores en el figura 5:

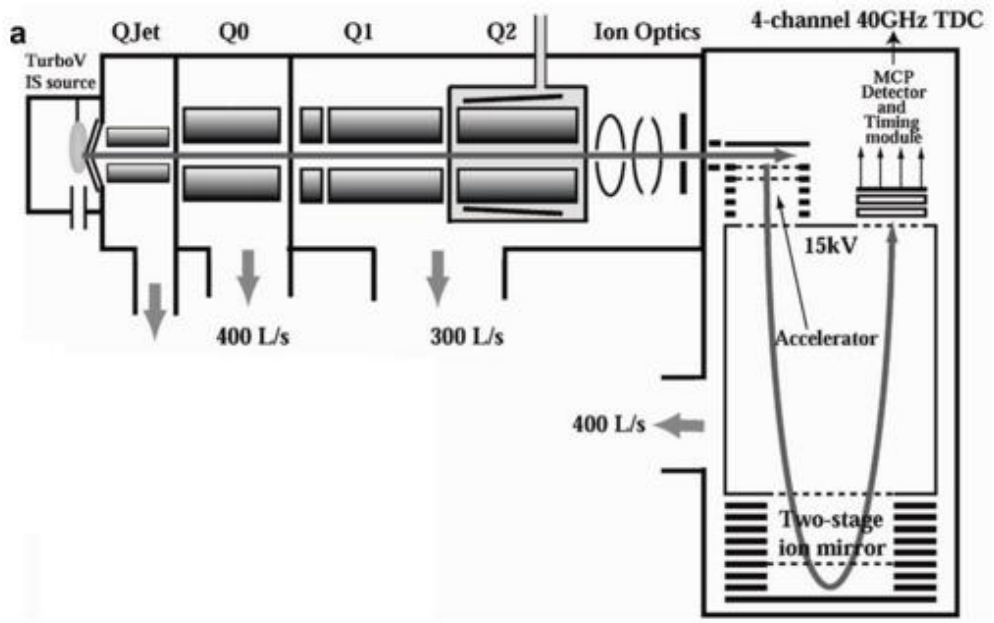


Figura 5: Esquema espectrómetro de masas (de izquierda a derecha, Fuente ionización (*IS source*), Q0 (cuadrupolo que dirige los iones cargados provenientes de la fuente de ionización), Q1 (cuadrupolo que filtra y selecciona las masas de los precursores a analizar en un instante de tiempo), Q2 (cuadrupolo denominado celda de colisión donde se fragmentan los péptidos), analizador TOF tiempo de vuelo y Detector en Triple TOF 5600)

1. Ionización de la muestra, Electrospray.

Existen diferentes formas de ionización de la muestra empleando bombardeos de electrones (-e), a continuación explicaré la técnica empleada en nuestro caso seguido de su uso.

El ESI (electrospray) es una técnica de ionización a presión atmosférica en la que, mediante la aplicación de nitrógeno líquido y en presencia de un campo eléctrico, se produce la nebulización de una solución del analito a su salida de un tubo capilar. En este proceso los péptidos presentes en fase líquida pueden ser desorbidos a fase gaseosa de una forma tan suave que se pueden conservar incluso complejos moleculares derivados de interacciones no covalentes existentes en la solución. En este estado se introducen los péptidos al espectrómetro de masas.

La ionización por ESI de péptidos y proteínas genera iones multicargados siendo la extensión de la carga dependiente del pH de la solución y del número de puntos básicos o ácidos de cada molécula. Este aumento de la carga produce la explosión de Coulomb, donde la tensión de superficie es lo suficientemente grande como para romper los enlaces, liberando el analito (péptido aislado del medio) [8].

2. Aceleración de iones por un campo cuadrupolar, analizador cuadrupolo.

A continuación explicaremos el funcionamiento del cuadrupolo (Q) [9], que es uno de los tipos de analizadores empleados en nuestro espectrómetro de masas por los que pasarán los péptidos a ser analizados. Consta de tres de ellos como podemos ver en la figura 5. Los cuadrupolos son un caso particular de sistemas multipolares y están

presentes en casi todos los tipos de espectrómetros de masa modernos como filtros y guías de banda ancha de iones.

Un cuadrupolo consiste en cuatro barras paralelas de sección hiperbólica en la cara interna, generalmente de entre 15-20 cm de largo y 0.5 cm de radio, separadas entre sí unos 2 cm, a las que se aplica un potencial combinado de corriente continua y de radiofrecuencia que crean en su interior un campo denominado cuadrupolar.

Los péptidos generados en la fuente deben atravesar longitudinalmente el recinto limitado por estas barras para incidir en el detector. Estos péptidos, que entran en el analizador (Q1, figura 5) a una energía de unos pocos electronvoltios, son sometidos al efecto del campo cuadrupolar que los hace oscilar y los desvía en función de su valor m/z (masa por unidad de carga) de forma que para una combinación de potenciales solo una estrecha ventana de péptidos llega a incidir en el detector. Los analizadores de cuadrupolo actúan por tanto como filtros de iones y los espectros de masas en estos sistemas se obtienen mediante un barrido del potencial aplicado a las barras. En consecuencia, en cada instante solo una pequeña fracción del total de péptidos es monitorizada mientras que el resto se desecha.

Durante el barrido, solo los péptidos con una determinada m/z llegan al detector, no obstante si dos proteínas de igual masa o muy parecida tienen unas propiedades químicas similares, estas eluyen a través del HPLC en el mismo instante de tiempo y el filtro selectivo de masas (cuadrupolo Q1) permitirá el paso de ambas, llegando estas al detector al mismo tiempo. Este efecto sucede en un número de ocasiones muy reducido.

Es importante establecer que el espectrómetro de masas funciona realizando dos ciclos a lo largo de un tiempo establecido y se comportan de la siguiente manera:

- En el primer ciclo denominado MS1, se adquieren las masas e intensidades asociadas a los péptidos (denominados precursores) identificados y cuantificados cuando el primer cuadrupolo (Q1) está actuando como filtro selectivo de masas, el segundo cuadrupolo (Q2) como conductor de péptidos y llegan directamente al detector tras pasar por el analizador TOF (tiempo de vuelo) [8].

- En el segundo ciclo denominado MS2 se adquieren las masas e intensidades de cada uno de los fragmentos generados a partir de los péptidos identificados en el ciclo MS1. Para ello el primer cuadrupolo actuará como filtro selectivo de masas, el segundo cuadrupolo actuará como celda de colisión y por último los péptidos pasarán a través del analizador TOF y llegarán al detector. Tendremos tantos ciclos MS2 como péptidos identificados en el ciclo MS1 se desee fragmentar.

En el ciclo MS1 se generará la lista de picos, donde se almacena la información de masa e intensidad de cada uno de los péptidos identificados a lo largo de todos los ciclos de MS1 que llegan al detector.
En la figura 6 podemos ver los péptidos identificados en un instante de tiempo en un ciclo MS1 y en un instante de tiempo posterior los péptidos identificados en un segundo ciclo MS1.

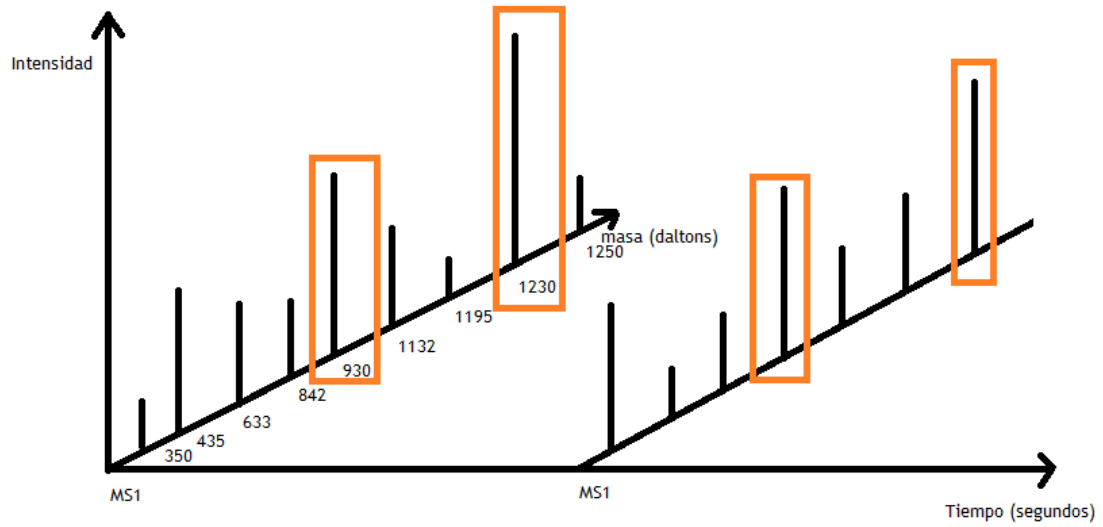


Figura 6: Ejemplo de masas identificadas en dos ciclos de MS1 diferentes para dos instantes de tiempo. Recuadradas en naranja se encuentran la masa que pasarán al segundo cuadrupolo (Q2) para un instante de tiempo si empleásemos Top 2 (los dos péptidos más intensos para realizar el ciclo MS2).

Como ya hemos comentado en el segundo ciclo MS2 el primer cuadrupolo (Q1) actúa como filtro y permitirá el paso de un péptido de una masa concreta. No obstante en cada ciclo MS1 se identifican múltiples péptidos de diferentes masas como podemos ver en la figura 6 y no todas pueden pasar al segundo cuadrupolo para su fragmentación ya que esto requeriría muchos ciclos MS2 (tantos como péptidos identificados en el ciclo MS1) y por tanto mucho tiempo de análisis por parte del espectrómetro de masas. Para decidir qué péptidos pasan al segundo cuadrupolo para su posterior fragmentación, el espectrómetro de masas detecta cuales son los N (valor definidos por nosotros y que sería de 2 en el caso del ejemplo de la figura 6) péptidos más intensos y estos serán los péptidos qué pasarán al segundo cuadrupolo y de los que se realizará el proceso de fragmentación.

Solo esta pequeña fracción de péptidos pasará al segundo cuadrupolo (Q2), que actúa como celda de colisión (ver figura 5). Aquí se fragmentarán y tras pasar por el analizador tiempo de vuelo (TOF) y llegar al detector se generarán los espectros de fragmentación. Los espectros de fragmentación contienen la información de masa e intensidad asociados a cada uno de los fragmentos generados a partir del péptido de la masa seleccionada en el primer cuadrupolo (Q1). Estos espectros de fragmentación tendrán asociado el precursor identificado y cuantificado en el ciclo MS1 al que pertenece dicho espectro de fragmentación como podemos ver en la figura 7.

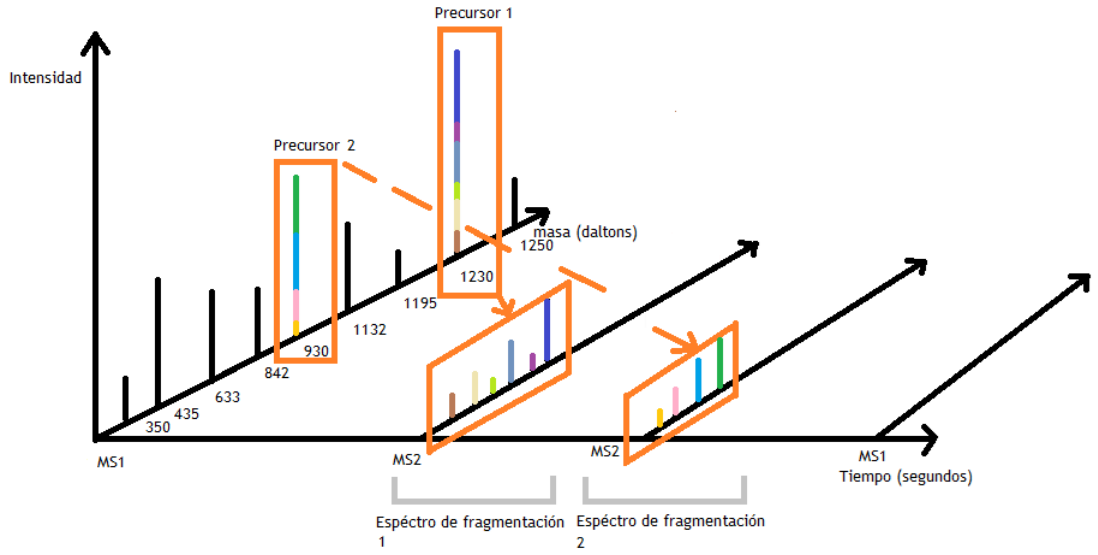


Figura 7: Precursores y su espectros de fragmentación asociados, en el caso de emplear un TOP2 (donde solo los dos péptidos más intensos de cada MS1 se fragmentan, empleando un ciclo MS2)

Este modo de funcionamiento repercute negativamente en el límite de detección de estos instrumentos, especialmente cuando se requiere la obtención de espectros completos en rangos de masa amplios [8].

Todo este proceso está regido por unos tiempos empleados los cuales se explicarán más a fondo en el capítulo 6 (*Parámetros en la adquisición de datos en el espectrómetro de masas*).

El principio físico que rige el primer analizador de tipo cuadrupolo queda definido de la siguiente manera, de forma análoga funcionara el segundo analizador de tipo cuadrupolo:

En los analizadores se ejerce una fuerza sobre cada péptido regida según la fórmula de la fuerza de Lorentz:

$$F = q\,(E + v \times B)$$

Donde E es el vector del campo eléctrico, B el vector del campo magnético, q será la carga del péptido, v el vector de velocidad y x el producto vectorial entre v y B.

Los péptidos sometidos a esta fuerza serán acelerados con una velocidad en base a su masa y carga que se rige según la fórmula:

$$v = [2eV/m]\ ½$$

Donde "V" es el potencial aplicado, "e" la carga y "m" la masa.

Con la acción de un campo magnético (H), los péptidos cargados se verán obligados a describir una trayectoria circular alrededor de este campo desarrollando una fuerza

centrífuga según la fórmula mv2/r, la cual será igual a la producida por el campo electromagnético. Se deduce pues que el radio descrito será igual a:

$$r = (2Vm/H2e)\ ½$$

Según la ecuación descrita la relación masa/carga (e) será:

$$m/e = H2.r2/2V$$

Es aquí donde podremos discernir entre masas de péptidos para su paso a la celda de colisión (cuadrupolo Q2). Esta es una de las propiedades más importantes de un espectrómetro de masas y de la cual depende la resolución del instrumento.

3. Tiempo de vuelo, analizador TOF.

Este es el tercer analizador presente en el Triple TOF 5600 (representado en la figuras 5 y 6).

Se basa en el hecho de que los iones acelerados por un campo eléctrico adquieren distintas velocidades según el valor de su relación m/z y por tanto tardan distinto tiempo en recorrer una determinada distancia. Considerando constantes la carga y la energía cinética de los iones formados en la fuente, la medida del tiempo de vuelo permite determinar de forma muy precisa la masa de cada uno de estos iones. A diferencia de los sistemas de cuadrupolo, que como se indicó anteriormente filtran en cada instante grupos de iones dentro de un pequeño rango de valores m/z desechando el resto de la población de iones, el analizador TOF separa y detecta en una escala de tiempo (tiempo de vuelo) el paquete completo de iones procedente de la fuente [8].

4. Detección de iones y producción de la correspondiente señal eléctrica.

El detector es el elemento final del flujo de análisis (figura 5). El tipo de detector presente se basa en el registro de la carga inducida o la corriente producida cuando un ion pasa cerca o golpea una superficie. Actualmente se utilizan electromultiplicadores (multiplicador de electrones) para poder amplificar la señal a una mínimamente procesable ya que el número de iones que llegan en un instante particular es muy pequeño.

El ordenador al que está conectado el espectrómetro de masas recoge las distintas señales y las reproduce en forma de espectros, formato de fácil interpretación.

2.3. Resultados obtenidos

Los resultados obtenidos se almacenan en archivos de extensión ".wiff". Estos archivos son específicos del fabricante y son distintos en función del espectrómetro empleado. Es un formato comprimido donde se encuentra la información de tiempo de retención-masa (mz) y de la intensidad de cada ion analizado (lista de picos y espectros de

fragmentación, *"peaklist"*) así como información adicional de la caracterización de estos iones.

Capítulo 3. Análisis de datos obtenidos por espectrometría de masas

3.1. Flujos de trabajo empleados en proteómica

Existe un gran abanico de técnicas empleadas para el análisis de los datos proteómicos por espectrometría de masas. Entre ellos citaremos y explicaremos los principios de dos de ellos; iTRAQ (*isobaric tag for relative and absolute quantification*) [10] y Label free (libre de marcaje) [11]. Estos flujos están actualmente implantados en la Unidad de Proteómica de Navarrabiomed. Ambas técnicas requieren de espectrómetros de masas de alta resolución como lo es el Triple TOF 5600.

Estas técnicas tienen como ventajas principales:

- Alto número de proteínas identificadas, hablamos en rangos de miles de proteínas.
- Es necesaria menor cantidad de muestra que en técnicas como los geles bidimensionales (*diferential in gel electrophoresis*).
- Es más reproducible que técnicas proteómicas como los geles bidimensionales.
- Pueden llegar a detectarse proteínas de membrana (aunque normalmente son fragmentos o degradaciones).
- Alta poder de resolución (capaz de distinguir dos picos cromatográficos de masas muy próximas).

Como desventajas principales serían:

- Alto precio de la tecnología.
- Alto precio de reactivos y especialmente de los marcadores isobáricos (iTRAQ).
- Límite de las muestras para realizar análisis mediante iTRAQ (8 muestras diferentes).

3.1.1. iTRAQ

Esta técnica nos permite no solo la identificación de las proteínas presentes sino además cuantificarlas para realizar análisis diferenciales [12]. Estos análisis nos permiten estudiar los cambios presentes en un conjunto de proteínas en diferentes condiciones experimentales. Esta técnica desarrollada inicialmente por Applied Biosystems, consiste en marcar las proteínas o los péptidos mediante un marcaje isobárico. Este marcaje se une a todos los residuos de argininas (Nter) y lysinas (Lys).

Con el objetivo de aumentar el número de proteínas identificadas en un experimento de proteómica, se realiza un proceso de digestión para obtener secuencias de péptidos de una proteína. Para lograr estos cortes de las secuencias proteicas, normalmente se emplea la proteasa tripsina, capaz de cortar después de las lysinas y las argininas. Podemos ver un ejemplo de este proceso en la figura 8.

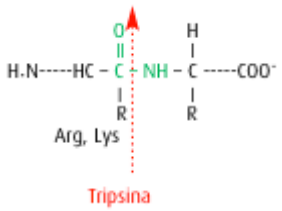


Figura 8: Digestión mediante tripsina

Este proceso de digestión se realiza de forma rutinaria no solo en la técnica de ITRAQ sino también en la técnica de label-free que explicaremos en el siguiente punto.

Al realizarse estos cortes que dejan péptidos de tamaños adecuados para la posterior identificación por parte del espectrómetro de masas, quedan descubiertos los radicales Nter y Lys, de modo que todos los péptidos quedarán marcados mediante el marcaje iTRAQ. Hay que tener en cuenta que si uno de estos radicales esta descubierto las probabilidades de quedar marcado por iTRAQ son casi del 100% según el fabricante.

La molécula de iTRAQ consta de tres regiones como podemos ver en la figura 9 ya sea para el caso de iTRAQ 4plex (cuatro marcajes diferentes) o iTRAQ 8plex (ocho marcajes diferentes).

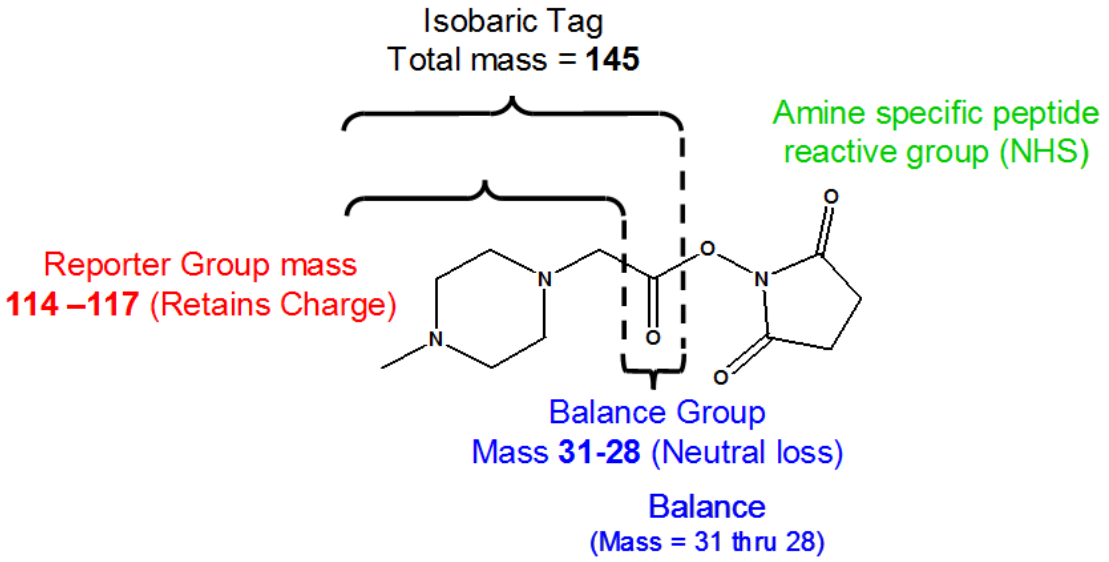


Figura 9: Estructura de iTRAQ; masa del grupo reportero, grupo balance y N-hidroxisuccimida

Como vemos en la figura, la molécula de iTRAQ siempre tiene una masa total de 145 daltons en el caso de iTRAQ 4 plex (305 daltons en el caso del iTRAQ 8plex). Esta masa se obtiene de dos regiones de esta molécula, la parte correspondiente al reportero y la parte correspondiente al balance.

La región referida al grupo reportero, corresponde a la masa variable según el reportero utilizado para marcar la muestra y puede ser 113, 114, 115, 116, 117, 118, 119 o 121. Esta parte nos sirve para identificar cada una de las muestras a analizar ya que una vez marcadas todas las muestras se mezclarán en un solo vial y de otra manera no podríamos saber de qué muestra proviene cada péptido. Este reportero

mantiene la carga y la eficacia en la ionización de los péptidos. Además no provoca alteraciones físico-químicas y mantiene el comportamiento cromatográfico de los péptidos marcados. Produce una señal intensa en el segundo analizador (de tipo cuadrupolo Q2, MS/MS).

La región referida al grupo balance, corresponde a la masa requerida para que la suma de cada reportero más el propio balance sumen 145 daltons (305 daltons en iTRAQ 8plex) (ver figuras 9 y 10). Esta región no aporta carga en MS/MS. Este grupo balance permitirá que todas las muestras de una misma condición (marcadas con el mismo reportero) tengan el mismo incremento en su masa, manteniendo constante la suma de ambas regiones (reportero + balance = 145 daltons en el caso de iTRAQ 4plex).

Por último está la región activa de unión a péptido gracias a un grupo N-hidroxisuccimida, capaz de reaccionar con el grupo amino N-terminal de los péptidos y el grupo amino de la cadena lateral de las lisinas.

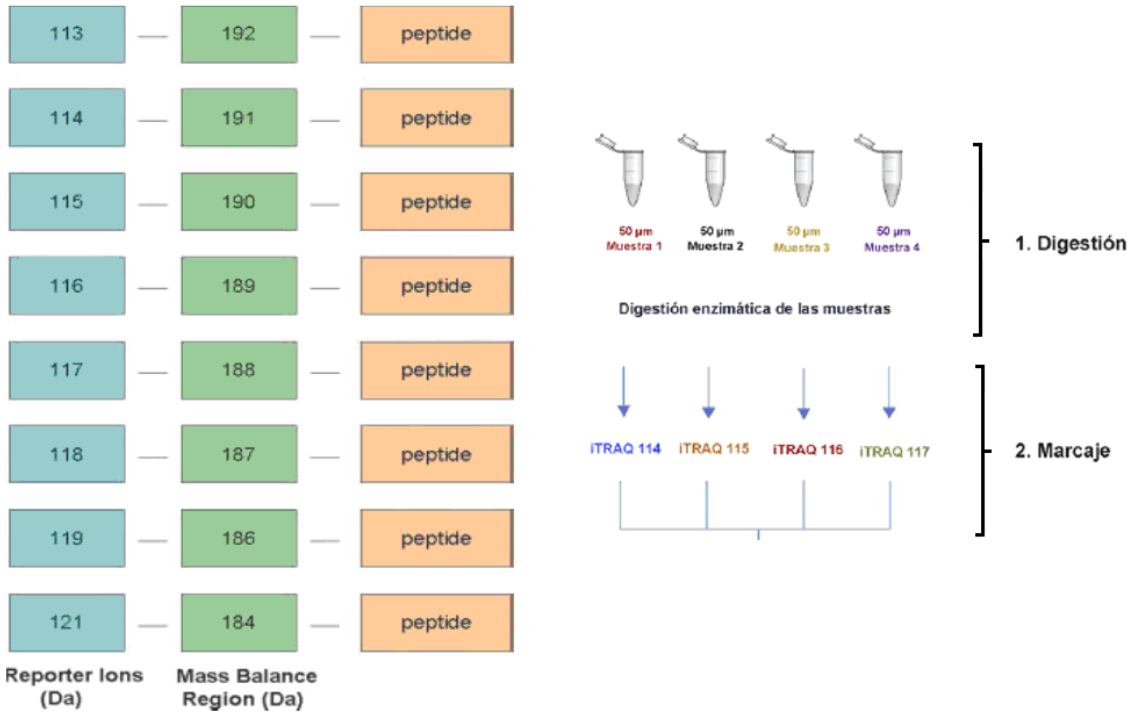


Figura 10: Proceso de digestión seguido del proceso de marcaje de cada una de las muestras por independiente. La suma de la región reportero y la balance siempre suman 305 daltons (iTRAQ 8plex)

Una vez digeridos, marcados y posteriormente mezclados en un único vial, las muestras no pueden ser analizadas directamente por espectrometría de masas dada su complejidad. Por lo tanto será necesario realizar un paso previo de fraccionamiento, permitiendo incrementar de esta manera el número de proteínas identificadas y a la vez favorecer la detección de péptidos presentes en menor cantidad.

Una vez realizada esta técnica, las fracciones resultantes se introducen en el espectrómetro de masas. En el primer ciclo (MS1), no podremos distinguir péptidos idénticos marcados con diferentes regiones reportero ya que en este paso únicamente

tendremos la información de todos los péptidos vistos en un instante de tiempo en función a su masa y tiempo de retención. Como ya hemos dicho todos los marcajes tienen una masa total igual y por tanto al marcar un mismo péptido la masa vista en el primer cuadrupolo será la misma venga el péptido de una muestra A o una muestra B. Por este motivo no podemos realizar la cuantificación de los péptidos mediante un ciclo MS1.

Será necesario liberar la región del grupo reportero para que podamos saber de qué muestra proviene un péptido. Para liberar las regiones del grupo reportero, se emplea el segundo ciclo MS2. Como ya hemos explicado previamente en este ciclo el segundo cuadrupolo actúa en modo de celda de colisión fragmentando los péptidos. En el caso del iTRAQ, en este proceso de fragmentación se liberarán las regiones reportero.

Una vez realizada la fragmentación tendremos una masa de la región reportero diferente en función de la muestra de la que proviene el péptido y por tanto podremos identificar a que muestra pertenece. La región del grupo balance seguirá unida al péptido razón por la que en los análisis posteriores hay que tener en cuenta la modificación de masa sufrida por los péptidos marcados con iTRAQ.

La cuantificación de un péptido se hará en función de la intensidad de este reportero como podemos ver en la parte derecha de la figura 11 y 12. Esto hace que la relación señal ruido sea mejor, este hecho lo vemos presente en los datos obtenidos.

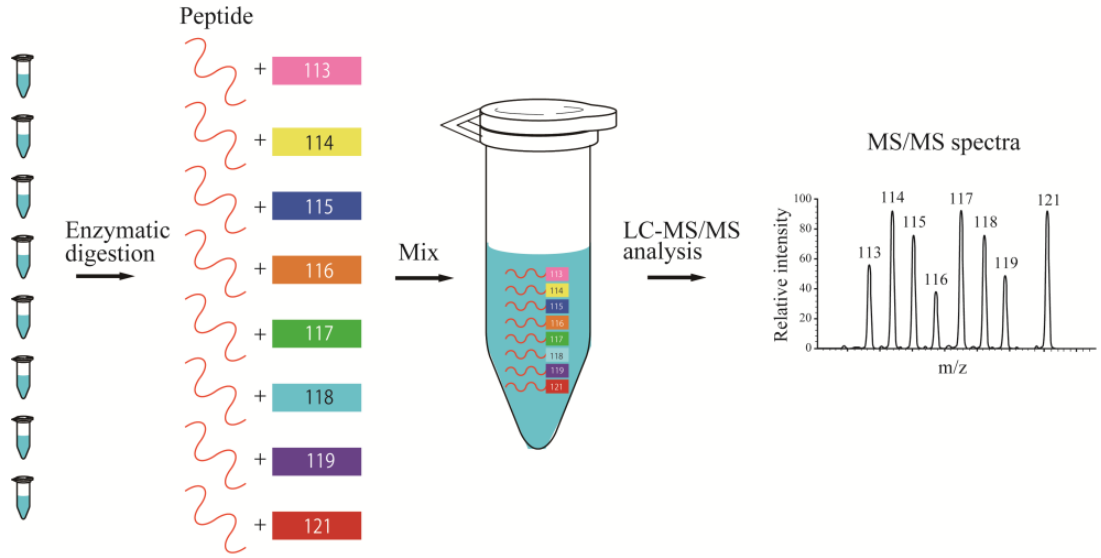


Figura 11: Flujo de trabajo iTRAQ e intensidad de cada reportero asociada a un péptido [10]

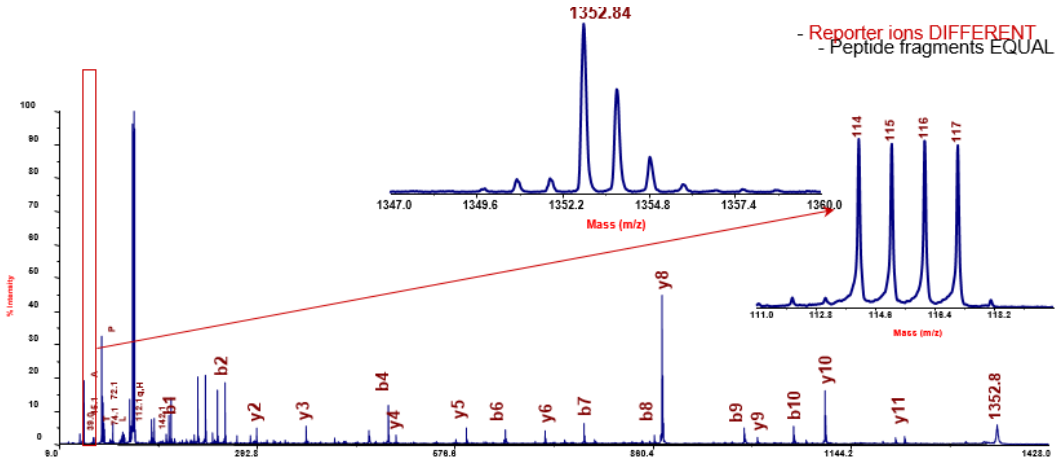


Figura 12: Intensidad de los reporteros iTRAQ 4plex (MS2) para una masa en un tiempo de retención concreto obtenido de un ciclo MS

Existen ciertas limitaciones inherentes a esta técnica, entre ellas la limitación en el número de muestras que pueden ser marcadas para un mismo experimento. Existen únicamente ocho marcajes distintos. Cada muestra se marca con uno de los marcajes pudiendo tener únicamente ocho muestras en el experimento.

Existen algunas vías para salvar la limitación del número de muestras a analizar (ocho), como la inclusión de estándares internos en todas las muestras de nuestro experimento, el cual nos permita analizar de manera global todas las muestras presentes en estudios de mayor envergadura.

3.1.2. Label free

A diferencia de otras técnicas como la citada previamente, esta aproximación no utiliza ningún tipo de marcaje. Por este motivo aunque la preparación de la muestra sea similar a la realizada en iTRAQ, al no llevar marcajes la cuantificación en el espectrómetro de masas se hará de manera diferente [11].

En este tipo de técnicas se introducen las muestras de manera independiente en el espectrómetro de masas, sin necesidad de mezclarlas, por lo que se conoce la muestra de la que provienen los péptidos identificados.

Tiene una ventaja frente a iTRAQ que es la ausencia de limitaciones en el número de muestras que se pueden analizar en un mismo experimento. No obstante las cuantificaciones no son en principio tan robustas como por ejemplo en iTRAQ, ya que las cuantificaciones se realizan en el primer ciclo MS (cuando los cuadrupolo Q1 y Q2 actúan únicamente como filtros selectivos de masas) y por tanto hay una necesidad férrea de ajustar muy bien los alineamientos de los cromatogramas en las diferentes muestras, un desajuste significativo de estos puede provocar errores en las cuantificaciones y dar lugar a falsos positivos. De igual modo, una cromatografía (HPLC) que introduzca grandes desviaciones en los tiempos de retención puede provocar perdida de la información al no encontrarse en los distintos cromatogramas de las diferentes muestras un mismo péptido a cuantificar, dentro de unos límites tolerables de tiempo de retención.

Como es lógico un péptido no eluye exclusivamente de modo fugaz en un instante de tiempo, sino que eluye a lo largo de un intervalo de tiempo gracias al HPLC acoplado al espectrómetro de masas. Esto permite al espectrómetro de masas generar un conjunto de picos cromatográficos para un mismo péptido obtenidos en diferentes ciclos MS1 consecutivos (XIC, ver figura 13). Estos picos tendrán asociados un valor de intensidad y se integrarán en un único valor de intensidad para dicho péptido empleando diferentes aproximaciones a la hora de analizar mediante la bioinformática los datos obtenidos del espectrómetro de masas como: el valor de intensidad máximo encontrado, el valor del área bajo la curva (curva roja presente en la figura 13) o el valor del área media de la curva centrada en el punto de máxima intensidad. Estos serán los valores de cuantificación a partir de los cuales se realizarán los análisis estadísticos (matriz de cuantificación).

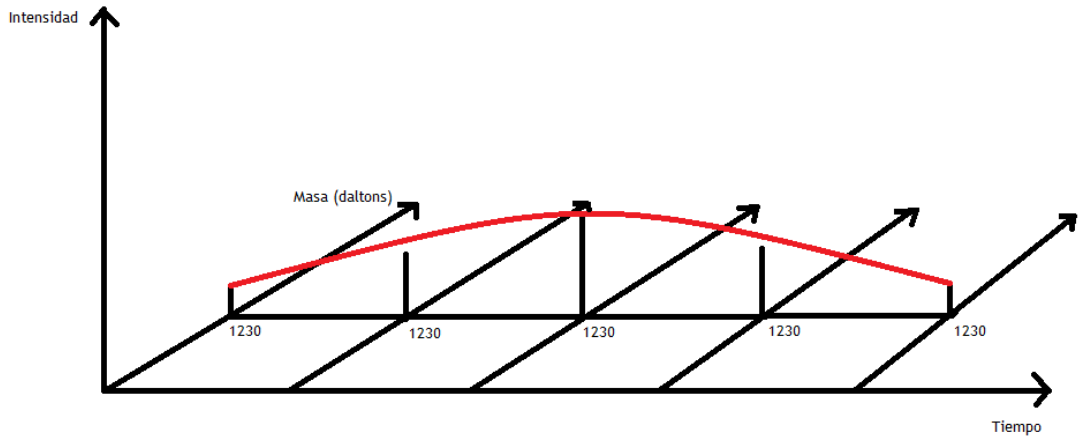


Figura 13: XIC, elución de un péptido de un masa concreta (1230) a lo largo del tiempo y sus niveles de intensidad en los diferentes ciclos de MS1 (precursores).

A menudo en este tipo de técnica se requiere seleccionar una muestra como control frente al resto para poder cuantificar el éxito en el alineamiento de cada una de las muestras. Existen diferentes aproximaciones para decidir qué control emplear, como la selección de la muestra con mayor número de iones o la selección de la muestra con mayor semejanza al resto.

Figura 14: Imagen tomada de Progenesis, proceso de alineamiento cuya muestra control para realizar el alineamiento será la muestra 362 (en rojo)

Para realizar el proceso de alineamiento se emplean vectores que unen los péptidos presentes en una muestra control (considerando la muestra control, la usada para realizar el alineamiento), con los péptidos presentes en cada una de las demás muestras. Según la proporción de vectores obtenidos en relación al número de péptidos presentes en las muestras se asigna una puntuación o *"Score"*. Este proceso puede verse en la figura 14.

3.2. Limitaciones

Explicadas las técnicas de análisis empleadas en el laboratorio y la tecnología disponible, el siguiente paso es realizar los análisis a partir de los datos adquiridos por el espectrómetro de masas.

En el momento de la realización de este proyecto, se empleaba el software comercial de la casa Sciex, proveedor del espectrómetro de masas Triple TOF 5600. Este software llamado Protein Pilot (versión 4.5) permite cargar los archivos de extensión ".wiff" ya citados y realizar el proceso de identificación generando en este proceso archivos de extensión ".group".

Además la estadística es realizada empleando una plantilla en Excel considerablemente pesada y lenta. Se encuentra además protegida mediante contraseña de modo que no es posible modificar el flujo de trabajo. Por último, este excel solo es capaz de realizar de manera directa el análisis diferencial semi-cuantitativo si se ha empleado la técnica iTRAQ.

Capítulo 4. Objetivos

Dadas estas limitaciones el objetivo de este proyecto es proponer y caracterizar opciones para realizar los análisis de iTRAQ y label-free de manera robusta y poder implementarlos en la rutina diaria de la Unidad de Proteómica de Navarrabiomed.

- El primer objetivo será testar flujos de trabajo capaces de trabajar con los datos obtenidos del espectrómetro de masas Triple TOF 5600. Este punto engloba el estudio del flujo planteado por Laurent Gato, el software MaxQuant desarrollado por Jürgen Cox y el software Progenesis desarrollado por Nonlinear Dynamics (Waters).

- El segundo objetivo será la realización de flujos de análisis en R o Python capaces de procesar los datos obtenidos por los flujos de trabajo candidatos y dar una salida robusta para su interpretación.

- Por último, tras analizar los resultados obtenidos (R o Python) mediante el uso de estos flujos de trabajo, deberemos ser capaces de decidir qué flujos nos dan mejores resultados para cada aproximación proteómica (iTRAQ y label free) y por tanto serán implementados para su uso por el personal científico de la Unidad de Proteómica de Navarrabiomed.

En un primer proceso de búsqueda encontramos tres candidatos capaces de trabajar con los datos de nuestro equipo. Estas tres opciones son: el workflow planteado por Lurent Gato en forma de scripts de R [13], el software libre Maxquant [14] y el software comercial Progenesis [15].

5.1. Softwares empleados para el análisis

5.1.1. Lurent Gatto R scripts

Esta fue la primera opción como método analítico. Una serie de paquetes desarrollados en R por Laurent Gato, (actualmente director del área de proteómica computacional de la universidad de Cambridge) y su grupo. Estos paquetes tal como MSnbase o RforProteomics [16], acompañados de otros como isobar no desarrollado por él, permiten realizar los procesos de identificación así como la realización del análisis diferencial posterior ya sea mediante label-free (MSnbase) o mediante iTRAQ (isobar) [17].

Este flujo de trabajo requiere de una conversión de los formatos comprimidos de la casa comercial (“.wiff”) a los formatos estándar actuales como son (“.mzml” o “.mzXml”). Estos son los formatos capaces de ser cargados mediante las funciones presentes en los paquetes citados.

Para la realización de esta conversión es necesario emplear 2 conversores. Primero el proporcionado por la casa comercial Sciex, capaz de convertir de formato “.wiff” a formato “.mzml”. Para esta tarea emplearemos un archivo batch para automatizar la conversión ya que son numerosos los archivos generados en un experimento de proteómica.

El código del archivo batch quedaría así:

```
@ECHO OFF
cd C:\prueba mzml con r\MSconvert
FOR %%I IN (*.wiff) DO (
  SET "ext=%%~xI"
  SETLOCAL EnableDelayedExpansion
  start /i "" "C:\Program Files (x86)\AB SCIEX\MS Data Converter\AB_SCIEX_MS_Converter.exe" WIFF "%%I" -proteinpilot MZML "%%~nI.mzml"
  TIMEOUT 10
  ENDLOCAL
)
```

Una vez generados los archivos en formato “.mzml” es necesaria una segunda conversión ya que de no hacerlo así la ausencia de determinadas cabeceras en los archivos xml generarán problemas posteriores. Este es un problema intrínsecamente relacionado a los archivos generados por Sciex, pues con la casa comercial Thermo no sucede. Esto puede ser debido a que en la actualidad muchos de los softwares se

desarrollan para este fabricante y posteriormente se extienden a otras casas comerciales.
En esta segunda fase de conversión empleamos el conversor por excelencia proporcionado por Proteowizard y de libre distribución. Podemos ver una captura de este conversor en la figura 15.

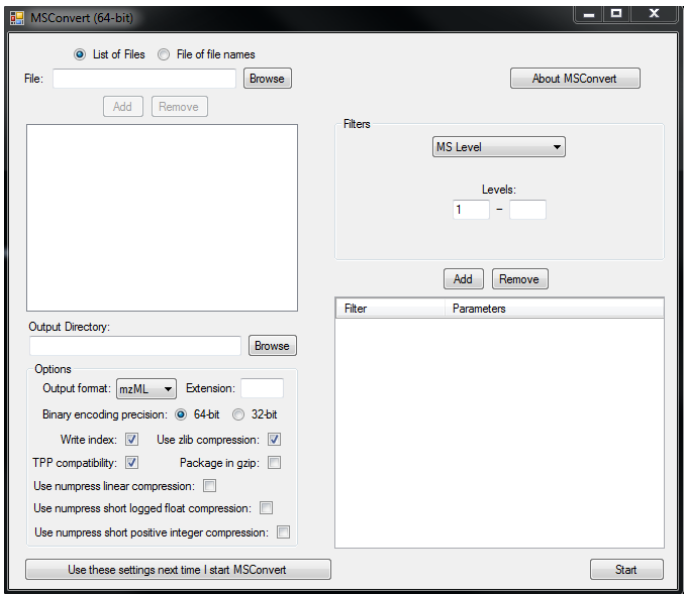


Figura 15: Captura MsConvert de Proteowizard

En este punto generamos los archivos ".mzxml", para cargarlos en R empleando la función "readMSData".

En este proceso de doble conversión pueden modificarse señales de espectros generando así datos de poca fiabilidad sin tener control sobre ello.

Para adquirir los datos de identificación de los iones presentes en los archivos brutos ".wiff" ahora convertidos a ".mzxml" utilizaremos motores de búsqueda de libre distribución como Rtandem (Xtandem) o MS-GF+ y bases de datos anotadas (.fasta).

Utilizaremos nuevamente un archivo batch o secuencias en DOS para la transformación de los archivos de identificación mediante MSFG+ a formato estándar (.mzid):

```
java -Xmx3500M -jar MSGFPlus.jar -s C:\autoAnalisisR\Fraccion16.mzxml -d C:\autoAnalisisR\HUMAN.fasta -o C:\autoAnalisisR\Fraccion16.mzid -t 20ppm -ti 1,2 -thread 10 -tda 1 -m 1 -inst 2 -e 1 -protocol 2 -ntt 1 -minLength 6 -maxLength 40 -minCharge 2 -maxCharge 6 -n 1 -addFeatures 0 -mod Mods.txt
```

Repetimos el proceso para los datos obtenidos mediante el motor de búsqueda Xtandem (este puede realizarse desde R utilizando Rtandem):

```
java -jar mzidentml-lib.jar Tandem2mzid TandemOutput.xml TandemOutput.mzid -outputFragmentation false -decoyRegex Rev_ -databaseFileFormatID MS:1001348 -massSpecFileFormatID MS:1001062 -idsStartAtZero false -compress false
```

La visualización del flujo de generación de todos estos archivos pueden observarse en la figura 16.

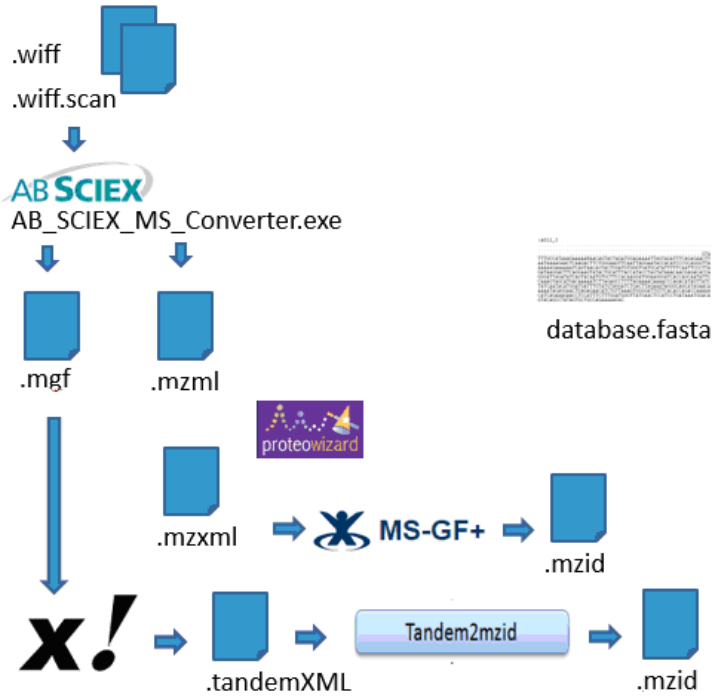


Figura 16: Flujo conversión e inclusión de identificaciones

Los resultados obtenidos no son satisfactorios en el número de identificaciones empleando estas vías si las comparamos con las obtenidas en Protein Pilot empleando su motor de búsqueda (Paragon) [18].

En la figura 17 podemos ver el proceso de identificación de péptidos y proteínas. Para ello se empleará una base de datos (.fasta) de la especie a estudiar, donde se encuentra anotada la información de cada proteína (secuencias de aminoácidos, taxonomía, gen, etc.). Estas anotaciones pueden contener información curada a mano o generada de manera automática. De manera continua se depuran estas bases de datos y por ello es importante mantener actualizadas las bases de datos empleadas a las que se enfrentan los espectros de fragmentación generados en los ciclos MS2.

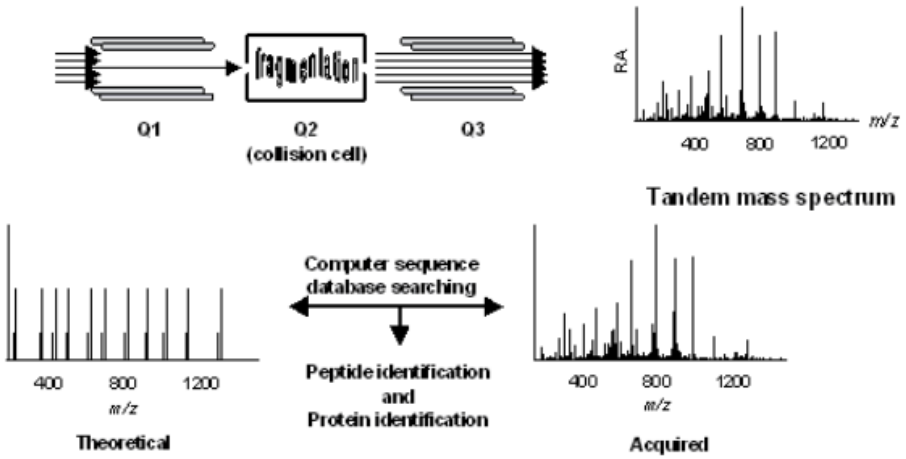


Figura 17: Proceso de identificación de péptidos y proteínas. Búsqueda de coincidencias entre los espectros de fragmentación teóricos (base de datos, .fasta) y los espectros de fragmentación reales generados en cada ciclo MS2 obtenidos en el espectrómetro de masas.

De este modo decidimos no utilizar esta vía de trabajo y buscar otra alternativa. Esto no significa que esta vía no sea una buena práctica, solo que con los datos obtenidos mediante el espectrómetro de masas Triple TOF 5600 de Sciex, no es viable emplear este flujo de trabajo actualmente, si no se disponen de otros motores de búsqueda más efectivos y se mejora en la conversión de datos.

5.1.2. MaxQuant

En este caso estamos ante un software desarrollado por Jürgen Cox investigador del instituto de bioquímica Max Planck [14].

Este software de distribución libre permite realizar tanto el proceso de identificación de proteínas gracias al motor de búsqueda integrado Andrómeda [19] (actualmente posee una versión independiente a MaxQuant) como el análisis diferencial semi-cuantitaivo. Es capaz de desarrollar esta tarea tanto para datos provenientes de la técnica Label-free como de iTRAQ.

No es necesario convertir de manera externa los datos procedentes de los archivos crudos ".wiff". Así mismo es capaz de procesar datos de casas comerciales como Thermo y Bruker además de Sciex, actualmente pese a que dispone de la opción de carga de archivos en formato estándar ".mzXml", esta opción no funciona al menos para los archivos generados a partir de datos ".wiff" de Sciex.

Los resultados obtenidos mediante el uso de MaxQuant son satisfactorios en cuanto al número de identificaciones y su cuantificación en el caso de utilizar datos generados mediante la técnica iTRAQ aunque son mejorables como veremos a lo largo del manuscrito. Por otra parte no son lo suficientemente buenos al emplear Label-free ya que se tiene un número de identificaciones inferior y hay una ausencia de valores de cuantificación en numerosas muestras para diferentes proteínas, situación que no se da al emplear otras alternativas como también veremos a lo largo de la memoria. Esto se debe únicamente a que de nuevo MaxQuant fue en origen generado para funcionar y procesar datos de la casa comercial Thermo (.raw) y por tanto está optimizado para este formato de datos.

Llegados a este punto ya tenemos un candidato para procesar los datos generados mediante iTRAQ. Posteriormente los datos serán procesados empleando scripts de desarrollo propio en R y Python.

Más información sobre el manejo y las características así como de los parámetros utilizados en MaxQuant en el Anexo 2.

En cuanto al procesado estadístico de los datos empleando los lenguajes citados serán explicados a fondo en el capítulo 7 de esta memoria.

5.1.3. Progenesis

Progenesis QI, es un software desarrollado por Nonlinear Dynamics, una compañía de Waters. Es un software de pago que requiere de una licencia para su uso. Como veremos en el capítulo *"5.2. Comparativa de softwares"* MaxQuant no dio muy buenos resultados en los análisis mediante la técnica Label-Free, no obstante Progenesis sí nos los ha dado.

Fundamentalmente se debe a que Progenesis nos permite insertar los datos de identificación obtenidos empleando el motor de búsqueda de Protein Pilot (Paragon), el cual arrojaba los mejores resultados. Para cuantificar este hecho realizaremos un breve estudio comparativo de los resultados obtenidos mediante ambos softwares en análisis mediante la técnica label-free.

El análisis semi-cuantitativo es realizado mediante este software y al igual que en el caso de MaxQuant realizaremos los análisis estadísticos mediante scripts desarrollados en R y Python que se explicarán en el capítulo 7.

Más información sobre el manejo y las características así como de los parámetros utilizados en Progénesis en el Anexo 3.

5.2. Comparativa de softwares

Para unas mismas muestras biológicas, realizaremos un breve experimento para comparar los resultados obtenidos mediante la técnica iTRAQ, empleando MaxQuant (Andromeda) y Protein Pilot (Paragon). Podemos ver el flujo de trabajo en la figura 18.

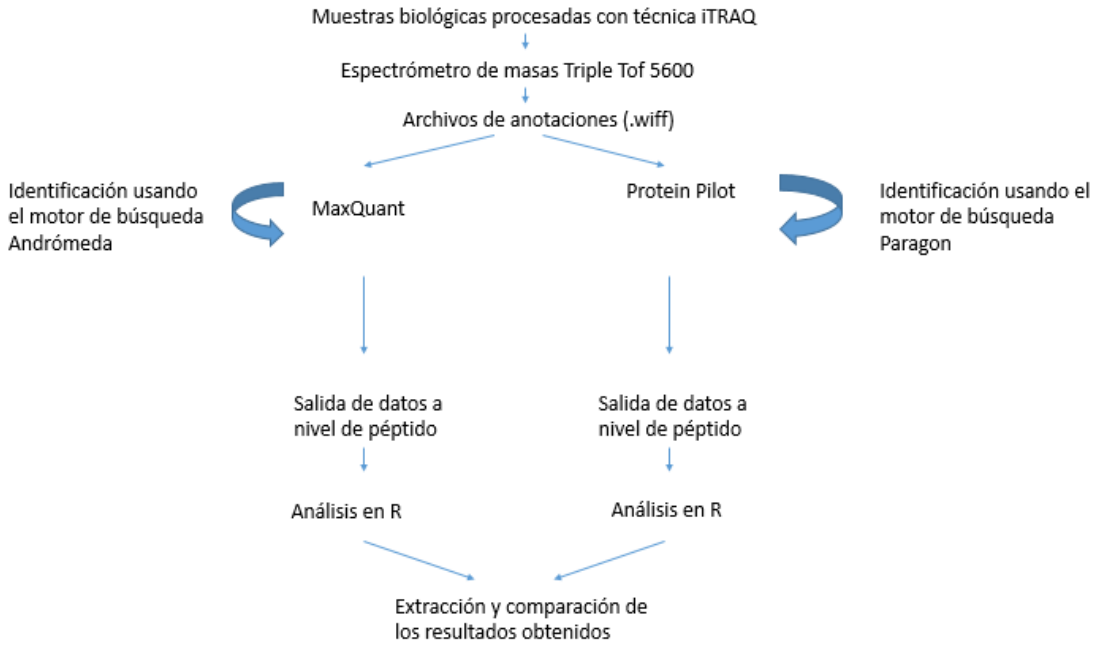


Figura 18: Flujo de trabajo para comparar los resultados obtenidos

5.2.1. Solapamiento en las identificaciones

En primer lugar realizaremos los análisis empleando ambos flujos y obtendremos los siguientes resultados, arrojados por los motores de búsqueda que utilizaremos en cada flujo de trabajo.

Paragon + Protein Pilot:

- Protein Pilot versión 5.0 -> 2885 Proteínas Cuantificadas
- Protein Pilot *("Protein groups")* -> 4872 Proteínas Cuantificadas

Andromeda + MaxQuant:

- MaxQuant-> 2026 Proteínas Cuantificadas
- MaxQuant(*"Protein groups"*)-> 4284 Proteínas Cuantificadas

Si realizamos un diagrama de Venn con el objeto de visualizar los solapamientos de los nombres de las proteínas identificadas en ambos flujos, obtenemos un 45% de solapamiento como se muestra en la figura 19 adjunta a continuación.

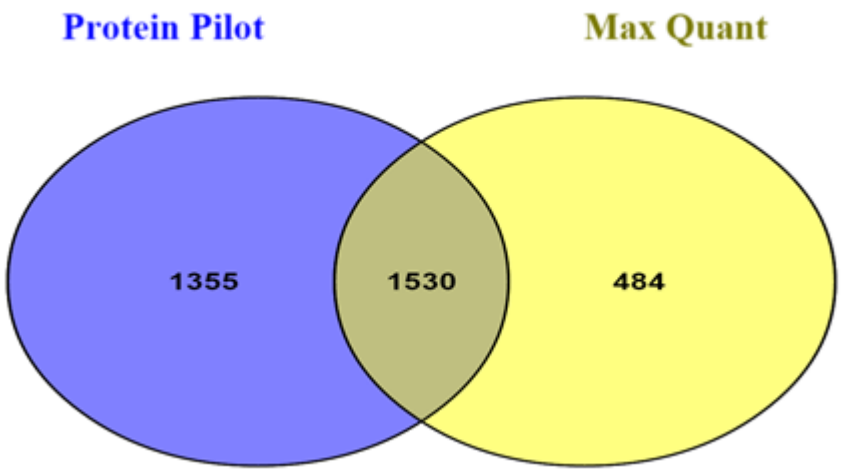


Figura 19: Diagrama Venn, solapamientos en códigos asociados a proteínas

Aparentemente, el solapamiento no es demasiado alto teniendo en cuenta que estamos procesando los mismos datos, no obstante esto puede deberse a la generación de los *"protein groups"*.

Los *"protein groups"* son la agrupación de aquellas proteínas que pueden explicarse por compartir péptidos comunes. Esta agrupación esta ordenada según el número de péptidos comunes o únicos utilizados para confirmar la presencia de la proteína. De esta manera la primera proteína de un *"protein groups"* será la que consideremos que hemos identificado ya que su presencia se basa en un mayor número de péptidos utilizados en su identificación.

Para comprobar que esta diferencia se debe a la generación de los *"protein groups"* y por tanto a la diferencia en las listas de péptidos identificadas en cada programa, emplearemos R para separar cada uno de los códigos asociados a cada proteína y de

este modo comprobar todos los códigos asociados a las posibles proteínas en cada uno de los *"protein groups"*.

Realizando este proceso, obtenemos el siguiente diagrama de Venn (figura 20):

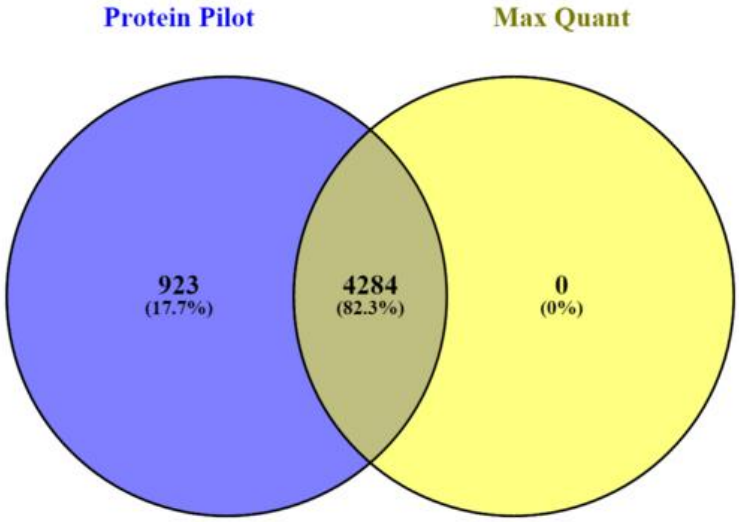


Figura 20: Diagrama Venn, solapamientos de los *"protein groups"*. En azul todas las proteínas presentes en los *"protein groups"* obtenidos mediante Protein Pilot, en amarillo los obtenidos mediante MaxQuant

Comprobamos que las no coincidencias se deben a la identificación por parte de Protein Pilot de un mayor número de péptidos asociados a proteínas de modo que las no coincidencias se deben al ordenamiento de los *"protein groups"*, dando lugar a proteínas ganadoras diferentes, ya que este incremento en la identificación de péptidos hará más plausibles otras proteínas.

Además tenemos un mayor número de proteínas identificadas adicionales, en Protein Pilot.

5.2.2. Solapamientos en identificaciones al 0.05 y 0.01 de p-valor

Si realizamos el procesado estadístico como se explicara a fondo en el capítulo 7, obtendremos aquellas proteínas que rechazan la hipótesis nula de no cambio entre las condiciones experimentales y por tanto cambian de manera significativa y además tienen una tasa media de cambio de al menos un 30%.

Para el cálculo de los niveles de significancia emplearemos un test Anova de un factor, de modo que nos quedemos con las proteínas que más cambian entre sus condiciones experimentales. Obtenemos los diagramas de Venn presentes en la figura 21.

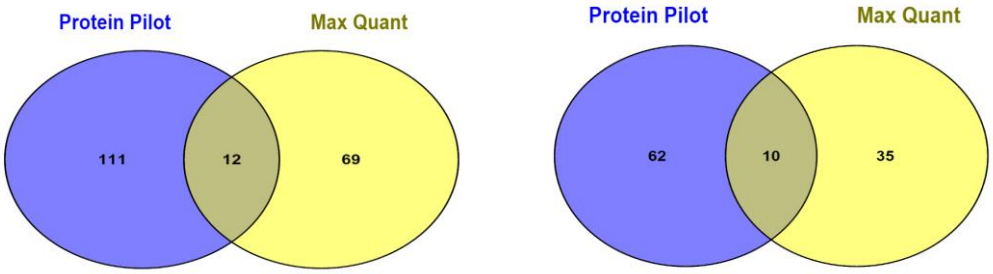


Figura 21: A izquierda p-valor < 0.05 a derecha p-valor <0.01

A la vista de los resultados, únicamente un 8.5% (12 proteínas) y un 9.5% (10 proteínas) de los resultados obtenidos respectivamente solapan al emplear ambos softwares (ver figura 21).

Podríamos pensar que puede deberse a diferencias en el ordenamiento de los *"protein groups".* No obstante si observamos algunos de los valores asociados a las proteínas, como LogFC1 (tasa media de cambio en escala log2, entre las dos primeras condiciones experimentales), LogFC2 (tasa media de cambio en escala log2 entre la primera y la tercera condición experimental), y Anova (p-valor, obtenido mediante el análisis de varianzas explicado en el capítulo 7) vemos diferencias apreciables en la cuantificación debido a la gran disparidad de péptidos utilizados en cada análisis así como en los valores P obtenidos (tabla 1).

Código de proteínas (Uniprot)	LogFC1	LogFC2	Anova
Protein Pilot -> P31949	0.897516	0.676249	0.000907
MaxQuant -> P31949	0.918813194	0.624465019	0.005177
Protein Pilot -> P09429	0.975228	0.71571	0.000166
MaxQuant -> P09429	0.952706512	0.702830896	0.00954
Protein Pilot -> P05161	1.033059	1.888534	2.32E-08
MaxQuant -> P05161	0.978726211	2.365954022	1.25E-07
Protein Pilot -> Q96D15	0.967888	1.357889	2.21E-05
MaxQuant -> Q96D15	1.02448312	1.475413789	0.000253
Protein Pilot -> O14879	1.051885	1.521224	1.94E-06
MaxQuant -> O14879	1.039159829	1.875808582	0.000137
Protein Pilot -> P20591	1.05483	1.498419	5.85E-05
MaxQuant -> P20591	1.042465671	2.08667313	7.73E-08
Protein Pilot -> P80723	1.072549	1.694116	1.82E-05
MaxQuant -> P80723	1.069252603	1.809195434	0.000419

Tabla 1: Valores de cuantificación y estadístico asociado a algunas de las proteínas obtenidas en los datos obtenidos en Protein Pilot y en MaxQuant. En verde los niveles de sub-expresión y en rojo los niveles de sobre-expresión.

Desechamos la opción de los *"protein groups"* como causa única de diferencia en los resultados ya que como vemos en la figura 22 hay presencia de proteínas significativas en un caso y no en otro, siendo estas proteínas identificadas en ambos casos:

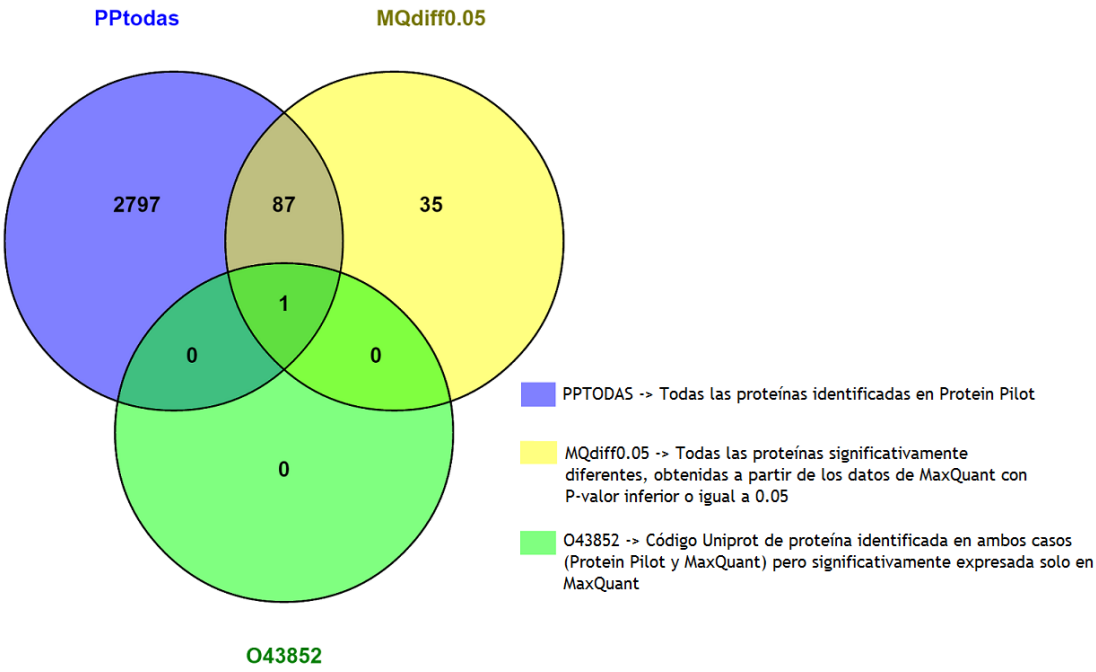
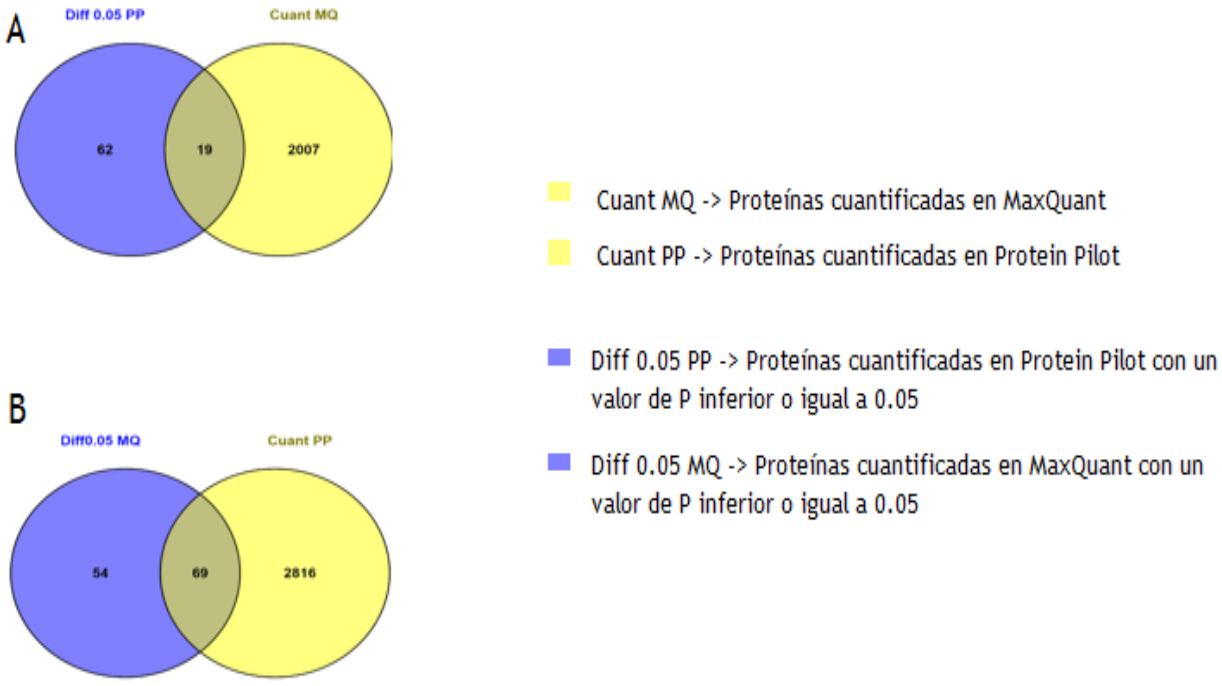


Figura 22: Código Unipot 043852, proteína identificada en ambos casos pero significativa únicamente en MaxQuant

Si profundizamos más en los datos y extraemos únicamente las proteínas significativamente expresadas en cada caso que además están identificadas en el caso alternativo y las comparamos entre si obtenemos:



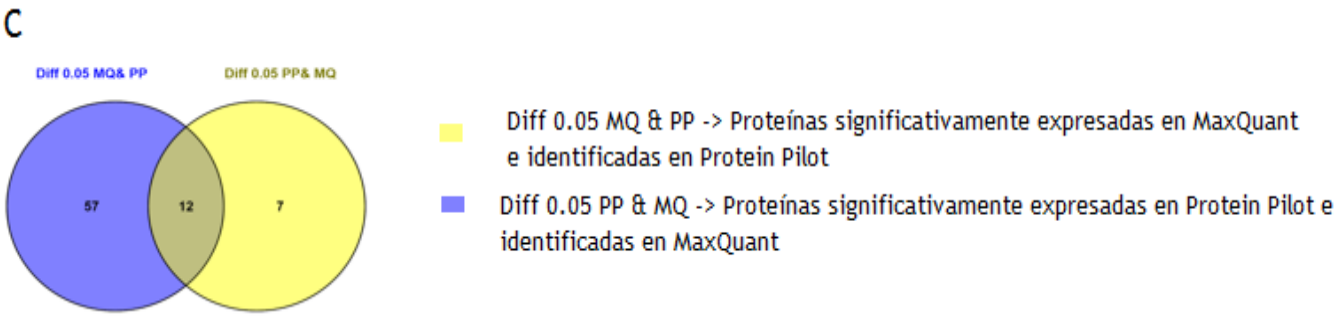


Figura 23: Proteínas significativas e identificadas enlos casos “A”, “B” Y “C”

En el caso “A”, podemos ver las proteínas cuantificas en MaxQuant (Cuant MQ) frente a las significativamente expresadas en Protein Pilot con P-valor inferior o igual a 0.05 (Diff 0.05 PP). Diecinueve proteínas se encontrarán en ambas listas.

En el caso “B”, podemos ver las proteínas cuantificadas en Protein Pilot (Cuant PP) frente a las significativamente expresadas en MaxQuant con P-valor inferior o igual a 0.05 (Diff0.05 MQ). Sesenta y nueve proteínas se encuentran en ambas listas.

En el caso “C”, podemos ver las proteínas identificadas en MaxQuant y significativamente expresadas en Protein Pilot (Diff 0.05 MQ&PP) frente a proteínas identificadas en Protein Pilot y significativamente expresadas en MaxQuant (Diff 0.05 PP&MQ) . En este caso doce proteínas solaparan.

A la vista de los resultados obtenidos en el caso “C”, doce de las diecinueve proteínas significativas en Protein Pilot solapan mientras que solo doce de las sesenta y nueve proteínas identificadas en MaxQuant solapan. Ya que las proteínas significativamente expresadas en cada caso son inversamente proporcionales a las proteínas identificadas con cada software, en el caso de MaxQuant tendremos un mayor número de falsos positivos ya que es en este caso donde tenemos un menor número de identificaciones y un mayor número de proteínas diferenciales.

5.2.3. ¿Por qué estas diferencias?

A fin de entender estas relaciones en las proteínas diferenciales en cada análisis, profundizaremos nuevamente en los datos para poder explicar por qué suceden estas relaciones inversamente proporcionales entre identificaciones y significancias en Protein Pilot y MaxQuant.

Para ello estudiaremos las proteínas diferencialmente expresadas en cada uno de los grupos presentes en la figura 23 “C”.

El primer grupo contendrá las 57 proteínas identificadas en MaxQuant y significativamente expresadas en Protein Pilot (Diff 0.05 MQ&PP) que no solapan con el segundo grupo que cotendrá las 7 proteínas identificadas en Protein Pilot y significativamente expresadas en MaxQuant (Diff 0.05 PP&MQ) que no solapan con el primer grupo.

Con el objetivo de conocer por qué no solapan, veremos una a una si las proteínas presentes en un grupo quedan excluidas del otro grupo debido a cambios en las tasas medias obtenidas, por los P-valores asociados o por ambos motivos.

Si realizamos esta subdivisión para el primer grupo obtenemos lo presente en la figura 24.

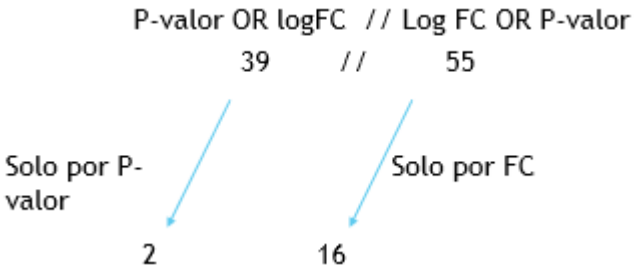


Figura 24: Proteína identificada en ambos casos pero significativamente expresadas únicamente en MaxQuant.

Como vemos en la figura 24, treinta y nueve proteínas quedan excluidas del segundo grupo por su P-valor (estas pueden tener además una tasa de cambio no significativa) dos de las cuales son exclusivamente por este factor. Si nos fijamos en su tasa media de cambio cincuenta y cinco quedarán excluidas de aparecer en el segundo grupo (estas pueden tener además un P-valor no significativo) dieciséis de las cuales son exclusivamente por este factor.

La explicación de la no significancia de estas proteínas en Protein Pilot es fundamentalmente debida a la diferencia en las tasas medias de cambio (FC) calculadas y en menor medida aunque también de manera importante al estadístico asociado P.

De manera más específica las 57 proteínas a estudiar se encuentran en MaxQuant de la siguiente manera:

- 35 Proteínas tienen FC cercanos al valor fijado 07 o 1.3
- 7 Proteínas tienen valores de P-valor cercano a 0.05
- 20 Proteínas serian diferencias claras según MQ

A la vista de los datos podríamos asumir que la discrepancia en los solapamientos no es tan grande como inicialmente parecía, si tenemos en cuenta la diferencia en los datos de partida utilizados en cada motor de búsqueda.

Si realizamos esta subdivisión para el segundo grupo obtenemos lo presente en la figura 25.

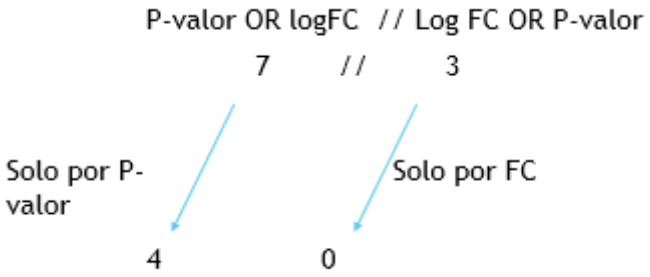


Figura 25: Proteína identificada en ambos casos pero significativamente expresadas únicamente en Protein Pilot.

En este caso las proteínas significativamente expresadas en Protein Pilot pero no significativamente expresadas en MaxQuant, quedarían rechazadas casi exclusivamente por un motivo estadístico.

5.2.4. Diferencias en los resultados

Además graficaremos las distribuciones de las desviaciones estandar de los valores en ambos experimentos para poder visualizar donde y porqué en el caso de Andromeda - MaxQuant existen estas disparidades y para poder visualizar si las proteínas significativas, realmente son proteínas de interés o por otra parte no lo son.

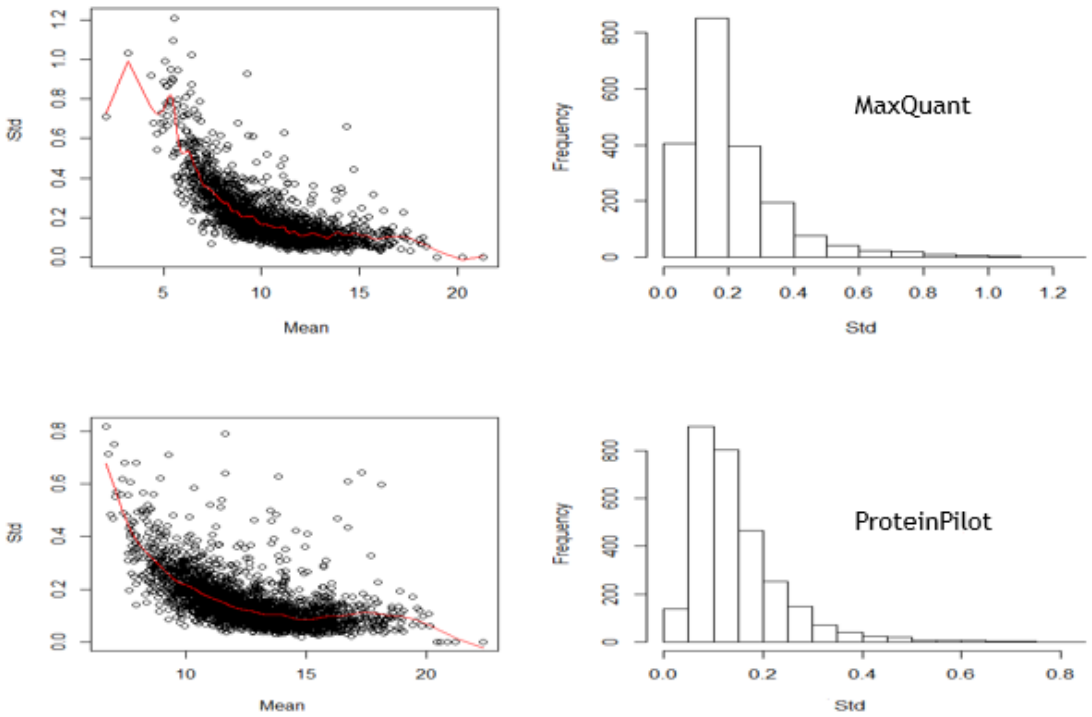


Figura 26: Desviaciones estándar y sus frecuencias (MaxQuant en la fila 1 y Protein Pilot en fila 2) frente a la media de la intensidad

Si observamos la figura 26, la distribución de las desviaciones asociadas en MaxQuant arroja un mayor número de desviaciones de un alto valor que la distribución arrojada por Protein Pilot. Podemos afirmar que los datos obtenidos en Protein Pilot son de mayor calidad, ya que las desviaciones en los valores de cuantificación son más homogéneos y solo las proteínas que realmente tienen cambio significativo se obtienen como tal.

Si realizamos un gráfico con el objetivo de visualizar donde se encuentran las proteínas significativas en MaxQuant para comprobar la existencia de proteínas significativamente diferentes en zonas de dudosa calidad, obtenemos la figura 27.

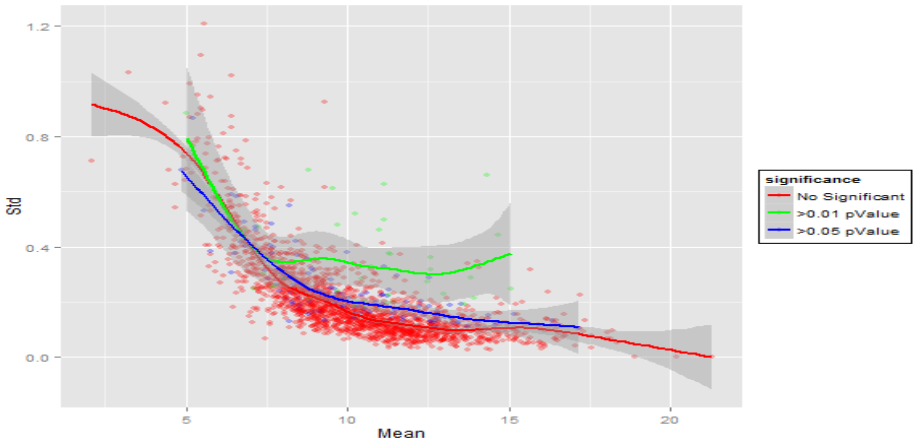


Figura 27: Desviaciones estándar y sus frecuencias coloreando las proteínas significativas frente a la media de la intensidad

Los puntos presentan una curva con una tendencia similar y se encuentran emplazados en mayor medida en la zona de intensidades media, lo cual es correcto porque es la zona de confort del espectrómetro de masas y donde actualmente mejor se realizan las calibraciones en masas.

Sin embargo tenemos un valor de significancia 99% con un valor de intensidad en escala logarítmica de 5. Este es un valor en escala lineal de 32 de intensidad. Este valor es considerado actualmente como ruido ya que los niveles de cuantificación son muy bajos teniendo en cuenta que barajamos rangos de 0 – 10^6 de nivel de intensidad.

A la vista de esta información es conveniente visualizar además las tasas de cambio frente a las intensidades de las proteínas. Como hemos visto, podemos tener proteínas de alta significancia que se encuentran en zonas de ruido y además tenemos un mayor número de proteínas significativas en MaxQuant pese a que el número de identificaciones es menor. Esto puede ser debido a diferencias provocadas en las tasas medias de cambio.

Si graficamos los datos obtenidos mediante los dos análisis obtenemos los resultados presentes en la figura 28.

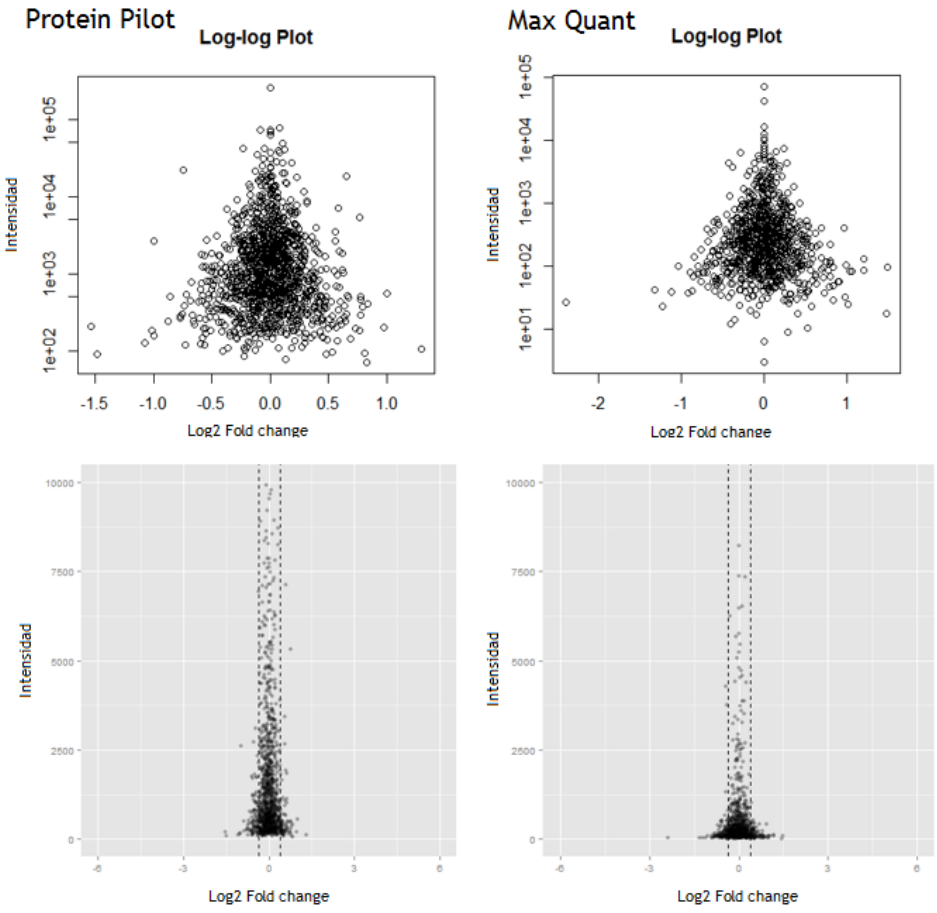


Figura 28: Tasas medias de cambio frente a las intensidades de las proteínas (Columna 1 Protein Pilot y columna 2 MaxQuant)

Observamos cómo pese al estar trabajando con las mismas muestras, los valores de intensidades obtenidas en MaxQuant a nivel de proteínas son muy inferiores a los obtenidos en Protein Pilot. Esto es debido al menor número de péptidos identificados mediante el motor de búsqueda de Andromeda (MaxQuant).

No solo vemos este efecto sino que además la anchura de las plumas, es decir, las tasas medias de cambio calculadas en MaxQuant son mayores en la parte inferior, la que llamaremos a partir de ahora zona de ruido.

Todo esto está relacionado directamente con los datos empleados para la realización del análisis. Para poder comprobar estos datos, se presentan los porcentajes de eficiencia del número de espectros asignados a péptidos mediante MaxQuant en 4 análisis diferentes (tabla 2).

Experimento	Espectros	No asignados	Porcentaje perdidos
1	120.000	10.000	8%
2	100.000	20.000	20%
3	75.000	25.000	33%
4	251.000	100.000	40%

Tabla 2: Espectros totales, espectros no asignados y el porcentaje de péptidos no asignados para cuatro análisis diferentes usando MaxQuant.

Podemos concluir que MaxQuant en comparación con Protein Pilot utiliza un menor número de espectros para la realización del análisis y además estos pueden no ser los mejores (muchos están en la zona de ruido).

Nos quedaría estudiar la diferencia en los valores obtenidos si empleamos Progenesis y R para procesar los datos de identificaciones obtenidas de Protein Pilot, en vez de analizar los obtenidos usando MaxQuant y R.

Para ello realizaremos otro experimento, esta vez mediante la técnica de label free y veremos cómo se replican los resultados obtenidos previamente, frente a los obtenidos mediante MaxQuant.

- Péptidos detectados y cuantificados:
  - Max Quant -> 19.896 péptidos.
  - Progenesis -> 21.245 péptidos.

- Proteínas:
  - Max Quant -> 2.122 proteínas detectadas y 397 proteínas cuantificadas.
  - Progenesis -> 2.792 proteínas detectadas y 1.517 proteínas cuantificadas.

Este experimento consistía en 4 grupos condicionales con 3 réplicas técnicas cada una, si realizamos un análisis por componentes principales obtenemos en cada caso las gráficos representados en la figura 29.

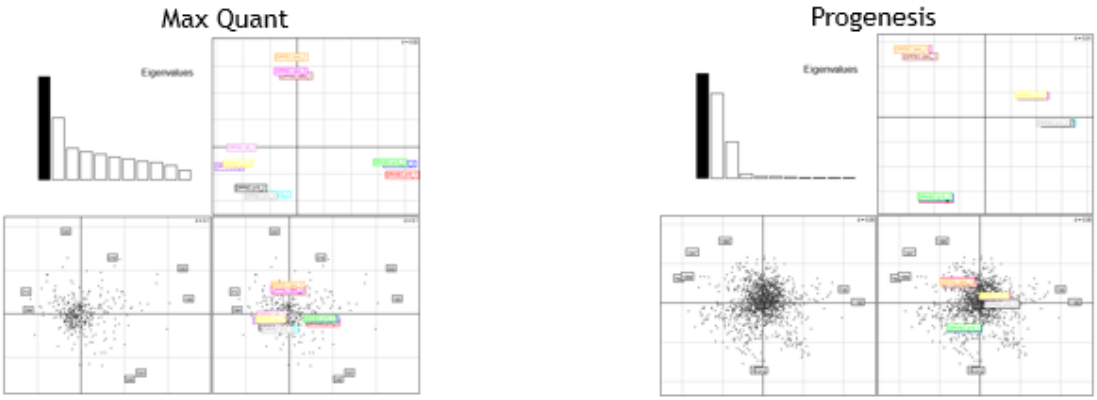


Figura 29: Análisis de coinercia para MaxQuant y Progenesis

Podemos observar como en Progenesis las réplicas técnicas se comportan como tal, mientras que en MaxQuant aunque cercanas entre si tienen un comportamiento mucho más variable. Esto confirma la variabilidad vista en la figura 29 y es el motivo por el que pueden surgir falsos positivos en la identificación de proteínas diferenciales. Es decir, el empleo del software Progenesis reduce considerablemente el número de falsos positivos, parámetro a tener muy en cuenta en el ámbito biomédico. Si empleamos estos falsos positivos para realizar análisis funcionales nos puede llevar a conclusiones equivocadas.

Si además realizamos un *"profile plot"* para ver los cambios en los niveles de expresión proteica a nivel global, nuevamente veremos un gran cambio (figura 30).

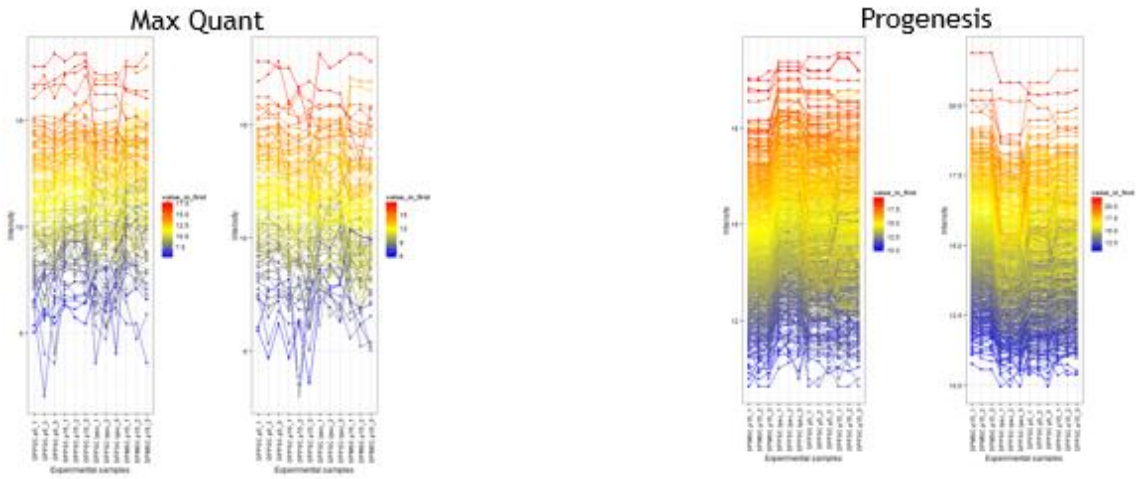


Figura 30: Profile plot para MaxQuant y Progénesis. Se observa como en MaxQuant los niveles de intensidades para las réplicas técnicas de la misma condición oscilan de manera notable.

Nuevamente las desviaciones estándar:

- En el caso de Progenesis:
  - Máxima: 1.8151356 std.
  - Mínima: 0.02607005 std.

- En el caso de Max Quant:
  - Máxima: 3.58537012153519 std.
  - Mínima: 0.17587579691086 std.

Como vemos en la figura 30 la reproducibilidad del nivel de cuantificación en las réplicas técnicas de una misma condición en MaxQuant es muy pobre si la comparamos con la obtenida en Progenesis. Si además tenemos un mayor número de proteínas identificadas empleando Protein Pilot (Paragon) que empleando MaxQuant (Andromeda) podemos concluir que el uso de Protein Pilot y posteriormente Progenesis para su procesado en vez de MaxQuant (Andromeda) es la mejor praxis.

Capítulo 6. Algunos parámetros en la adquisición de datos en el espectrómetro de masas

6.1. ¿Qué podemos modificar?

Según las necesidades analíticas, la metodología seguida y la complejidad de la muestra, es posible modificar varios parámetros que cambian significativamente las características de los datos adquiridos mediante el espectrómetro de masas. En la figura 31 quedan reflejados algunos aspectos que comentaremos a continuación.

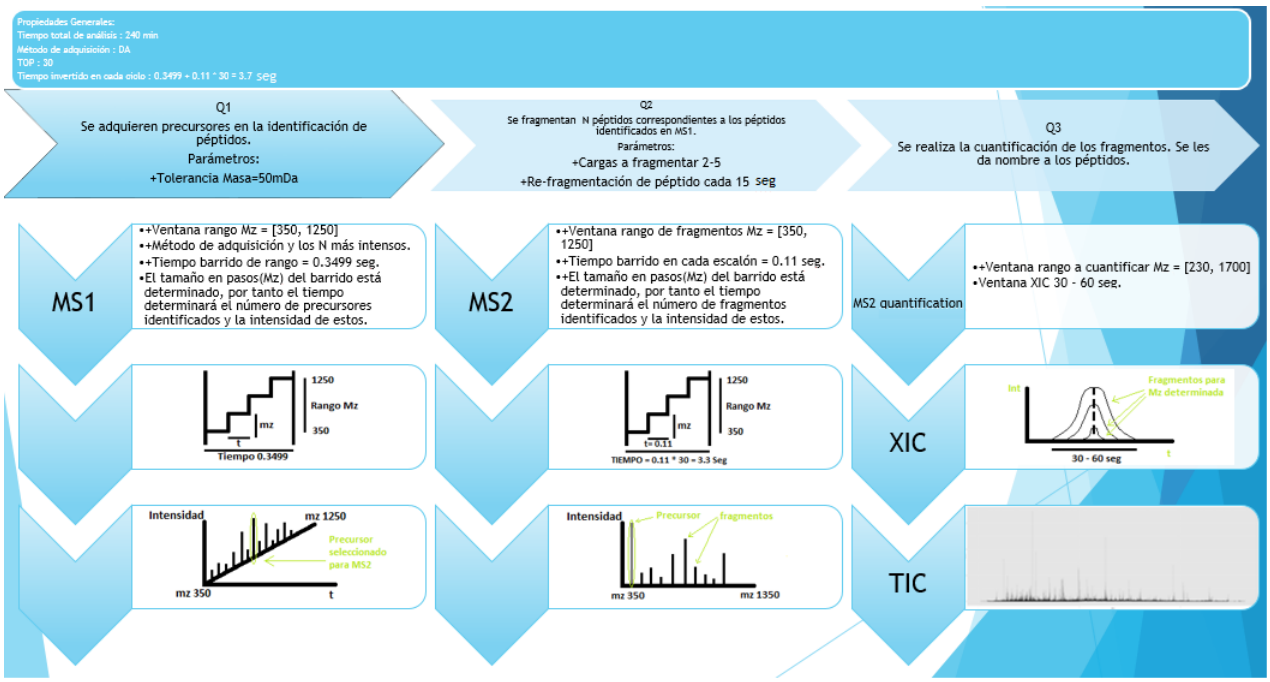


Figura 31: Sintesis del flujo y parametros que modifican la adquisición de datos

Principalmente las características de los datos adquiridos por el espectrómetro de masas vienen regidos por varios parámetros:

- ✓ El tiempo total de análisis (~1-4 horas).
- ✓ Método de adquisición empleado (DA *("dynamic accumulation"*), etc).
- ✓ Ventana de masas a analizar en el primer analizador (MS1) (350-1250 daltons).
- ✓ Número de péptidos más intensos seleccionados para la fragmentación de cada barrido MS1 (TOP30(los treinta más intensos), TOP50 (los cincuenta mas intensos), etc).
- ✓ Tiempo empleado para adquirir intensidades de masas en un instante de tiempo en el primer cuadrupolo (MS1) (0.3499 segundos).
- ✓ Ventana de masas a analizar en el segundo analizador (MS2) (350-1250 daltons).
- ✓ Tiempo asignado a cada escalón en el segundo cuadrupolo (MS2) (0.11 segundos).
- ✓ Tiempo de exclusión para un péptido (tiempo mínimo hasta la re-fragmentación de un mismo péptido).

En base a estos parámetros podremos ajustar la adquisición de datos con el fin de encontrar un equilibrio entre la identificación y cuantificación, paso clave en experimentos diferenciales realizados en el área biomédica.

Con el fin de observar la dependencia entre los resultados obtenidos y estos parámetros, realizaremos un pequeño estudio cambiando algunos de estos parámetros y al mismo tiempo compararemos estos resultados empleando nuevamente ambos softwares (MaxQuant y Progenesis) con el fin de corroborar las diferencias halladas previamente.

Realizaremos 2 análisis utilizando datos procedentes de un experimento iTRAQ con las siguientes paramétricas con cada uno de los softwares:

1. Primer método:

MaxQuant (Andromeda) y Protein Pilot (Paragon) -> TOP30 (fragmentaremos los 30 péptidos más intensos obtenidos en un instante de tiempo determinado MS1), 0.12 (tiempo asignado a cada escalón en el proceso de fragmentación, junto al tiempo total de ciclo, determinarán el número de escalones empleados en cada fragmentación) y DA ("*"dynamic accumulation"*, modificará el tiempo empleado en la fragmentación para obtener una intensidad suficiente para la obtención de un correcto espectro de fragmentación).

2. Segundo método:

MaxQuant (Andromeda) y Protein Pilot (Paragon) -> TOP 30 y 0.12.

6.2. Comparativa en los resultados obtenidos

Compararemos la disposición de los valores de la tasa de cambio y la intensidad promedio de las proteínas con muestras de la misma condición y para mayor homogeneidad emplearemos el control como condición.

Para el primer método obtenemos:

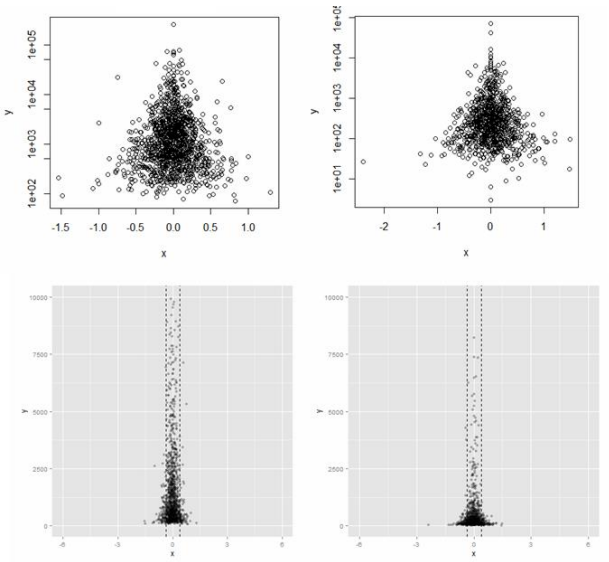


Figura 32: En eje "x" la tasa de cambio y en eje "y" la intensidad promedio de cada péptido (a izquierda Protein Pilot y a derecha MaxQuant). Las gráficas situadas en la parte superior y en la inferior representan la misma información pero para visualizar el efecto de ambos ejes por igual lo representamos con un zoom del eje Y diferente.

Para el segundo método obtenemos:

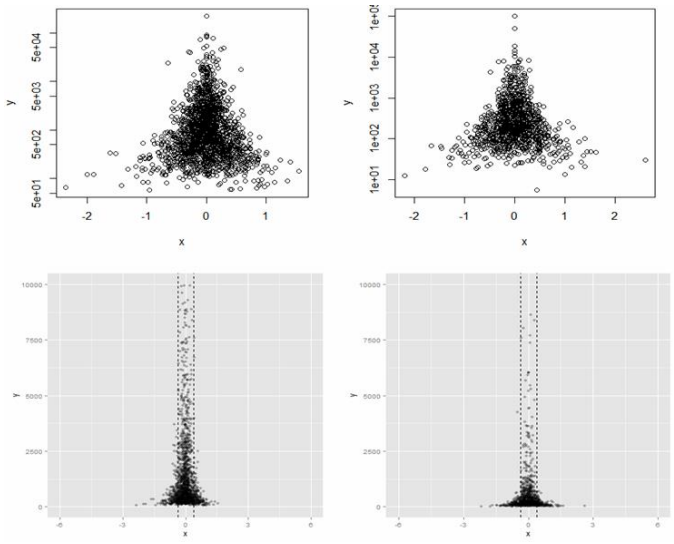


Figura 33: En eje "x" la tasa de cambio y en eje "y" la intensidad promedio de cada péptido (a izquierda Protein Pilot y a derecha MaxQuant) Las gráficas situadas en la parte superior y en la inferior representan la misma información pero para visualizar el efecto de ambos ejes por igual lo representamos con un zoom del eje Y diferente.

Si comparamos los resultados obtenidos entre MaxQuant y Protein Pilot para cualquiera de los métodos empleados obtenemos la información contenida en la figura 34.

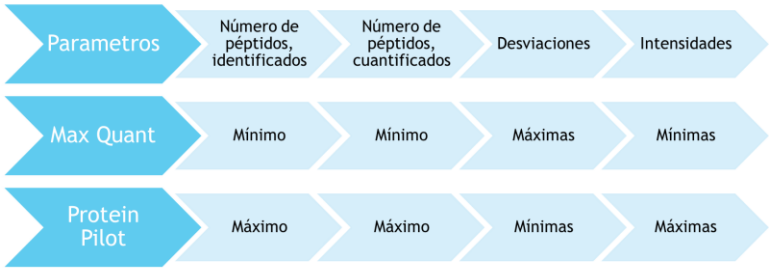


Figura 34: Resumen de resultados obtenidos entre MaxQuant y Protein Pilot

De igual manera si comparamos los resultados obtenidos de cada método para cada software por independiente obtenemos:

- MaxQuant:

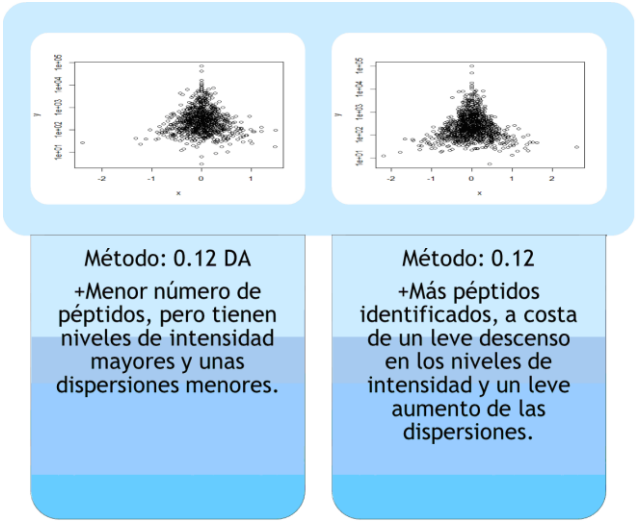


Figura 35: Comparativa métodos de adquisición

A nivel de proteína obtendremos:

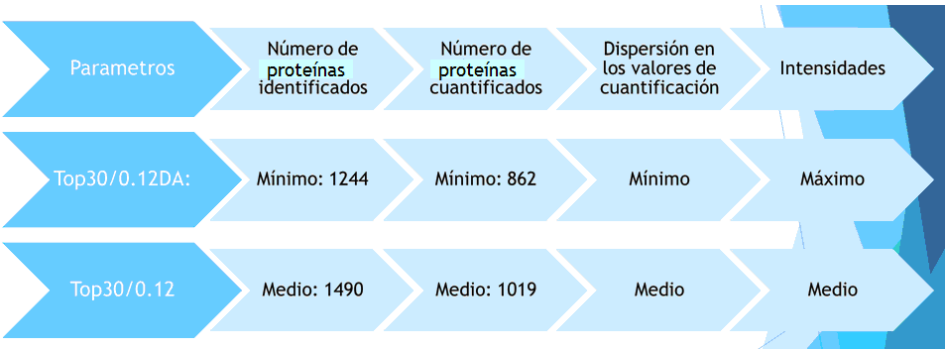


Figura 36: Resultados obtenidos a nivel de proteína

- Protein Pilot:

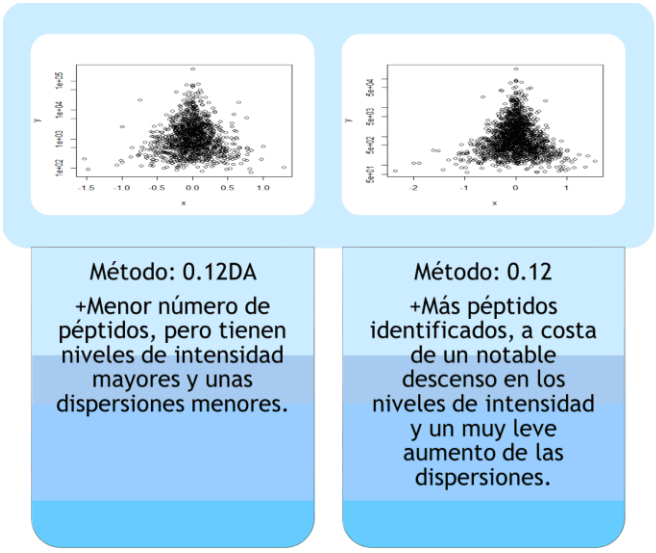


Figura 37: Comparativa métodos de adquisición

A nivel de proteína obtendremos:

Figura 38: Resultados obtenidos a nivel de proteína

A la vista de los resultados podemos apreciar de las figuras 36 y 38 la importancia en la adquisición de los datos y la definición de unos valores equilibrados para la identificación y la cuantificación de los mismos.

Observamos como el método DA *("dynamic accumulation")* citado previamente disminuye notablemente el número de identificaciones en MaxQuant y Protein Pilot ya que al aumentar los tiempos asignados para la fragmentación de los péptidos se generan menos ciclos de MS1 y por tanto un menor número de péptidos diferentes serán fragmentados y analizados. Por otra parte los espectros de fragmentación tienen uno valores de intensidad mayores lo cual hace que las desviaciones entre las dos muestras control sean menores.

Capítulo 7. Análisis estadísticos y visualización de datos

7.1. Flujo desarrollado en R

En este primer flujo de trabajo el objetivo era desarrollar código capaz de procesar los datos de salida de MaxQuant ya sean obtenidos mediante la técnica iTRAQ o Label-free [10] [11].

Para ello se elige el lenguaje de programación R en el cual se encuentran muchos de los paquetes desarrollados con anterioridad para realizar análisis genómicos como por ejemplo el paquete estadístico de limma [20].

Alrededor de este código se generan otros de utilidad para desarrollar ciertas tareas relacionadas con los análisis, entre ellos scripts de conversión o de análisis funcionales empleando herramientas como String, Reactome o Kegg, incluidos en este código y que serán explicadas más adelante.

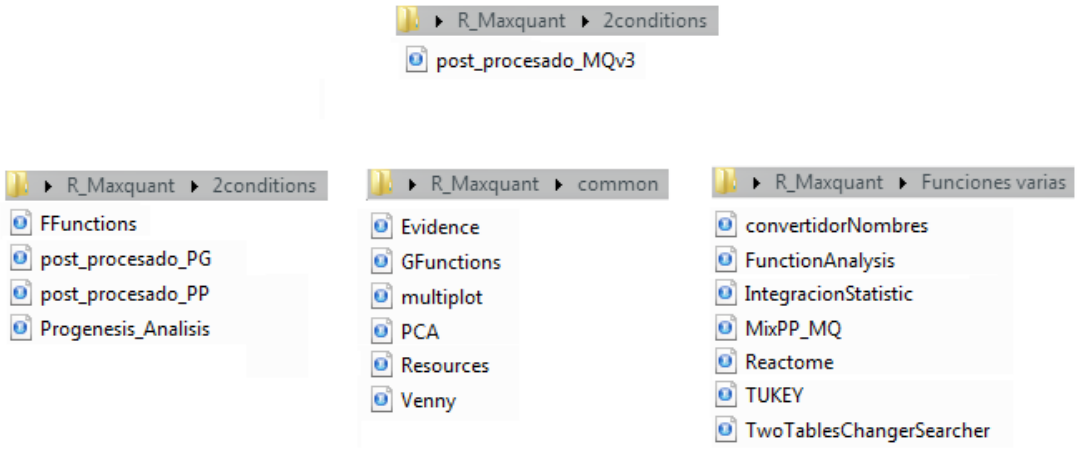


Figura 39: Scripts de R para realizar análisis

Los datos de partida obtenidos mediante Maxquant son archivos que contienen las matrices de cuantificación y además información asociada a ellos en un formato de texto tabulado por separaciones (.txt). Estas tablas tendrán un número de columnas dependiente del número de muestras del experimento así como del tipo de experimento realizado (iTRAQ o Label-free).

El modo de generación de estas tablas es explicado en el Anexo 2.

Antes de comenzar con el análisis utilizaremos la tabla generada por MaxQuant (evidence.txt) donde tenemos información de algunos parámetros de calidad sobre el experimento.

Generaremos varios gráficos para poder visualizar el estado de los datos y así detectar posibles fallos para así poder reanalizar los datos con distinta paramétrica, estos pueden verse en la figura 40.

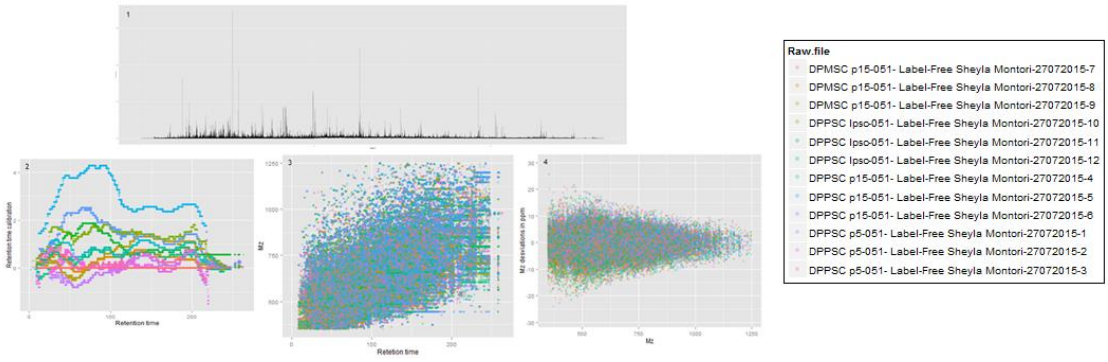


Figura 40: 1. Cromatograma de los datos identificados y cuantificados
2. Calibraciones realizadas a los datos obtenidos en cada uno de los *runs*
3. Masas en relación a los tiempos de retención para cada uno de los *runs*
4. Desviaciones en masa (ppm, partes por millón) en relación a la masa para cada *runs*

En la figura 40.1 tenemos un cromatograma, en este se disponen la suma de las intensidades de todos los iones (eje y) para un instante de tiempo de retención dado (eje x) a lo largo de todo el tiempo de análisis.

También tenemos el gráfico 40.2 en el que se muestra la corrección de masas en partes por millón (eje y) realizada para cada instante del tiempo de retención (eje x) en cada una de las muestras.

Además el gráfico 40.3 muestra los mapas de péptidos presentes en cada muestra para cada masa (eje y) y el tiempo de retención (eje x).

Por último tenemos el gráfico 40.4 en el que se muestran las masas de los iones de cada una de las muestras (eje x) frente a la desviación en masa asociado en partes por millón (eje y).

A continuación iniciaremos algunas variables antes de realizar la carga del fichero, entre ellas:

```
experimento -> definiremos aquí el tipo de experimento (‘itraq’ o ‘labelfree’)
distribución -> definiremos el número de muestras por condición experimental
canalesNames -> definiremos los nombres de las columnas de las muestras
nombres -> definiremos los nombres a reemplazar por los nombres originales de las muestras
devnombres -> definiremos los nombres de las columnas correspondientes a los valores Z de cada
una de las muestras
```

Definidas estas cinco variables cargaremos el archivo de datos a nivel de péptido (peptide.txt) si se trata de un análisis mediante label-free o a nivel de espectro (PSMs, msms.txt) si se trata de un experimento de iTRAQ. Esto se realiza de esta manera para maximizar el control de los datos, ya que en el caso de iTRAQ se cuantifica a nivel de MS/MS y en label-free a nivel de MS. Este será el punto desde el cual partiremos para filtrar y generar nuestro propio set de proteínas.

Figura 41: Breve esquema del flujo de análisis realizado en R o Python para los datos obtenidos mediante el uso de iTRAQ

Realizaremos una primera fase de filtrado donde eliminaremos aquellas proteínas identificadas en la base de datos *"decoy"* y en la base de datos *"contaminants"*.

Llegados a ese punto conviene explicar qué son estas bases de datos. En proteómica se almacena la información en bases de datos cuya extensión es ".fasta", son archivos de texto. En el caso de los experimentos realizados en MaxQuant se emplean dos bases de datos de búsqueda una con la información del proteoma de la especie de interés y de la cual forman parte las muestras a analizar y una segunda anotada con queratinas y posibles proteínas indeseadas consideradas contaminantes presentes en los experimentos. Por último, los datos se enfrentan a la base de datos del proteoma deseado aplicándole a esta una transformación tal que darle la vuelta o aleatorizarla, de modo que las secuencias de proteínas presentes en la base de datos sean ficticias. Este proceso se realiza con el objetivo de crear una manera de filtrar las listas de proteínas identificadas, denominado FDR (*False Discovery rate*), de modo que cuando el número de secuencias encontradas pertenecientes a esta base de datos decoy sea superior al 1% del tamaño de la lista total (valor modificable), se filtrarán todas las identificaciones que siguen a esta. De esta manera se mantiene controlada la identificación errónea de proteínas. Asumiremos en nuestro caso una tasa de FDR del 1% a nivel de PSMs, péptido y proteína.

Una vez realizada esta operación realizaremos una agregación de los valores de intensidad propios de cada muestra para cada proteína, sumando los niveles propios de cada uno de los péptidos pertenecientes a dicha proteína. De esta manera constituiremos los niveles de intensidad propios de cada proteína. Este proceso se realizará previamente para pasar de PSMs a péptidos si estamos en el caso de iTRAQ.

A lo largo de este proceso podemos realizar 2 filtrados adicionales, uno a nivel de péptido y uno posterior a nivel de proteína. Como hemos visto en el estudio de los softwares, MaxQuant tiene una tendencia a acumular valores de cuantificación en la zona de ruido a nivel de proteína. Según el rango de intensidades propio del experimento, filtraremos por lo general a nivel de péptido por un umbral dependiente del experimento. De esta manera filtraremos los datos de muy bajo nivel de intensidad. Como hemos visto en capítulos anteriores estos péptidos tendrán valores de intensidad cuantificados de manera muy variable.

Por convenio internacional se ha establecido la necesidad de tener al menos 2 péptidos entre los cuales puede haber o no uno de ellos como péptido único para la identificación fiable de una proteína. Las proteínas en base a un único péptido visualizado se eliminan.

Cabe citar que en la generación de los *"protein groups"*, de donde obtenemos la proteína ganadora para dicho *"protein group"*, pueden contribuir péptidos comunes, que explican la presencia de múltiples *"protein groups"* o péptidos únicos que son aquellos, que únicamente explican la presencia de un *"protein group"*.

Incluiremos un proceso adicional de filtrado en el que estableceremos que la ausencia de valores de cuantificación en más de una muestra para una misma condición supondrá el filtrado de la proteína. De este modo permitimos cierta flexibilidad en las cuantificaciones y así continuamos trabajando con aquellas proteínas donde un valor de cuantificación por condición no había sido posible de conseguir. Los valores ausentes serán completados por la media del resto de valores de cuantificación para dicha muestra. Esto puede hacer variar las desviaciones estándar de manera positiva, sin embargo supondrá una menor variación del valor medio de la condición el cual nos interesa no sufra una deformación al incluir estos valores de sustitución. Ver figura 42.

Gene names	Intensity Control	Intensity Control	Intensity Control	Intensity Initial	Intensity Initial	Intensity Initial
APOB	153.0950779	138.1414826	152.0955043	82.08381285	0	106.7865155
CPSF7	43.25671038	63.2633048	64.38179548	0	0	30.57555484
ERC1	21019.06138	32066.82584	31934.75513	22633.83227	26456.49903	20465.62603

Figura 42: En verde proteínas que pasarán el filtro y en rojo proteínas que no pasaran el filtro de ausencia de valores de cuantificaión. En el caso de APOB, reemplazaremos el valor faltante por la media de las otras dos muestras de la condición *Initial*.

Llegados a este punto existen al menos dos métodos de normalización de interés para este tipo de experimentos. Uno que permita comparar los niveles de cuantificación para las proteínas dadas entre las diferentes condiciones experimentales y un segundo método que permita comparar los niveles de cuantificación de las proteínas dentro de una misma muestra.

Según el interés del experimento emplearemos uno de ellos o ambos. Por defecto mostraremos solo aquel de interés, no obstante en Python (flujo desarrollado a posteriori) incluiremos por defecto ambos valores de normalización, esto será explicado a continuación.

Para realizar la normalización entre condiciones experimentales emplearemos la normalización por quantiles como vemos en la figura 43, técnica basada en el ordenamiento por columnas en orden creciente de los valores asociados y su posterior cálculo de las medias. Una vez calculadas las medias, se procede a sustituir los valores originales de la matriz por los valores obtenidos en el cálculo de las medias. De este modo los rangos así como la media asociada a cada columna (muestra del experimento) será la misma y podrán ser comparadas entre sí.

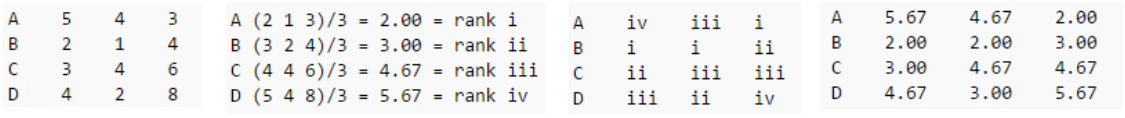


Figura 43: Proceso de normalización por quantiles explicado en cuatro pasos

Para realizar la normalización entre proteínas de una misma muestra emplearemos la normalización iBAQ [21], técnica basada en el cálculo de los péptidos teóricos en base a la secuencia de una proteína digerida mediante una proteasa. En nuestro caso procederemos al cálculo de todos los péptidos teóricos obtenidos a partir de la digestión de las secuencias de las proteínas empleando tripsina. Una vez calculado este número se dividen los valores de cuantificación obtenidos por él, de modo que todas las proteínas quedan relativizadas según su tamaño y de este modo pueden ser comparados sus niveles de cuantificación.

Una vez normalizados los datos procedemos a realizar una transformación de estos a escala logarítmica con el objetivo de trabajar en una escala normal y para poder graficar los datos de manera adecuada.

Para realizar el cálculo de los estadísticos asociados a los datos de cuantificación emplearemos el paquete de R limma, éste y otros paquetes empleados pueden ser encontrados en la web de Bioconductor (https://www.bioconductor.org/). Éste fue diseñado originalmente para realizar análisis de microarrays (experimentos donde se analiza la expresión de los genes que componen un genoma).

Para ello emplearemos fundamentalmente dos funciones comprendidas en este paquete, “lmFit” nos permitirá realizar el ajuste lineal de las muestras agrupadas previamente según un modelo o diseño definido y “eBayes” la cual calculará los estadísticos moderados t y F así como las tasas de cambio asociadas a cada par de condiciones experimentales frente a la condición que definiremos como control.

Limma emplea un método empírico de Bayes para reducir las varianzas de las muestras para cada *“feature”* o proteína (fila) frente a un valor común y para aumentar los grados de libertad para las varianzas individuales.

Empleando el cálculo de los test estadísticos moderados t, se realizarán los contrastes de hipótesis por pares frente a la condición control para cada proteína (fila), de modo que se calculen de manera independiente si los contrastes se rechazan y por tanto hay cambio significativo o por el contrario no lo hay. De igual manera para cada fila se realizará el test estadísticos moderados F este test global se calcula a partir de los test t calculados previamente para cada proteína o fila. Este proceso es análogo al realizado

entre los test estadísticos t y los test estadísticos F convencionales aplicados en los análisis de varianza anova salvo por que los cuadrados medios residuales y los grados de libertad han sido moderados entre las proteínas (filas).

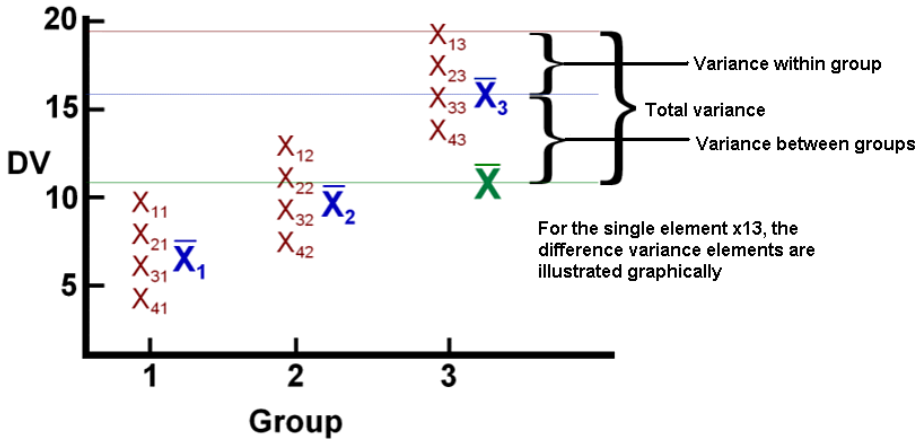


Figura 44: Parámetros empleados para el análisis de varianza

Los análisis de varianza (anova) parten de los conceptos de regresión lineal. Un análisis de varianza permite determinar si diferentes condiciones experimentales muestran diferencias significativas o por el contrario sus medias no cambian. El análisis de varianza permite superar las limitaciones de hacer contrastes de hipótesis bilaterales (entre dos condiciones) que son un mal método para determinar si un conjunto de variables mayor que 2 difieren entre sí.

El funcionamiento del análisis de varianzas simple (1 factor) como es el empleado en este tipo de análisis, se puede entender como una comparación de la medida de la variación entre diferentes grupos condicionales *"Variance between groups"* con una medida de la variación dentro de cada grupo condicional *"Variance within group"*. Si la varianza entre grupos es significativamente mayor que la varianza dentro de cada grupo, concluiremos que las medias asociadas a diferentes condiciones experimentales (factor de estudio) son distintas. Si por el contrario la varianza entre grupos no es significativamente mayor que la varianza dentro de un grupo condicional no se rechazará la hipótesis nula de que las medias asociadas a diferentes niveles del factor coinciden y por tanto no habrá diferencia significativa. Esta explicación puede verse gráficamente en la figura 44.

En los análisis donde únicamente tenemos dos grupos diferenciales tendremos el test moderado t, y en los análisis de múltiples comparaciones (más de dos grupos condicionales) tendremos el test moderado F. Además se calculan otros estadísticos como el B que no serán utilizados ya que se requiere de prefijar el número de proteínas que cambian en un experimento, dato desconocido en este tipo de análisis.

Actualmente se sabe que las distribuciones de los datos ómicos (proteómica, genómica, transcriptómica, etc.) no se ajustan correctamente a un modelo lineal simple o robusto los cuales son procesados en limma. No obstante, se siguen empleando estas aproximaciones en la actualidad.

Además se pueden computar los test estadísticos de Benjamini o Bonferroni con el objeto de ajustar los valores P, ya que se realizan múltiples análisis en un experimento, lo cual implica el aumento del error tipo 1 asociado a la realización del cálculo de cada test y proporcional al valor umbral seleccionado (ejemplo: 99% de nivel de confidencia, error asociado del 1%). Dada la naturaleza y la cantidad de los datos en proteómica rara vez será posible emplear estos valores ajustados, ya que nos quedaríamos sin información significativa.

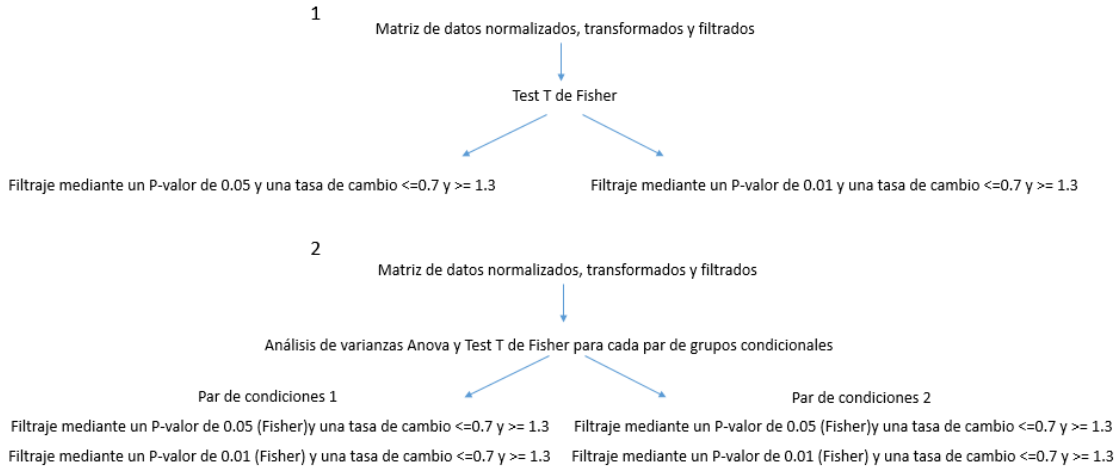


Figura 45: 1. Generación de tablas en el caso de 2 condiciones experimentales, 2. Generación de tablas en el caso de más de dos condiciones experimentales

En este flujo generaremos dos listas de proteínas significativas como vemos en la figura 45, una primera con un nivel de significancia del 99% (p-valor < 0.01) y una segunda con un nivel de significancia del 95% (p-valor < 0.05). Además se generará una tercera lista con aquellas proteínas cuyos cambios a nivel de tasa media de cambio son grandes pero según el test estadístico no son cambios estadísticamente significativos (también puede ser de interés estudiar estas proteínas desde un punto de vista biológico).

Las proteínas deben cumplir un umbral definido de cambio para poder estar en estas listas de proteínas significativas, por regla general consideraremos un cambio de la expresión de un 30% suficiente para considerar que la proteína cambia entre condiciones experimentales. Si enfrentamos la tasa media de cambio (*fold change*), frente al estadístico asociado calculado (P-valor), nos permite generar los gráficos denominados volcano (ver figura 46).

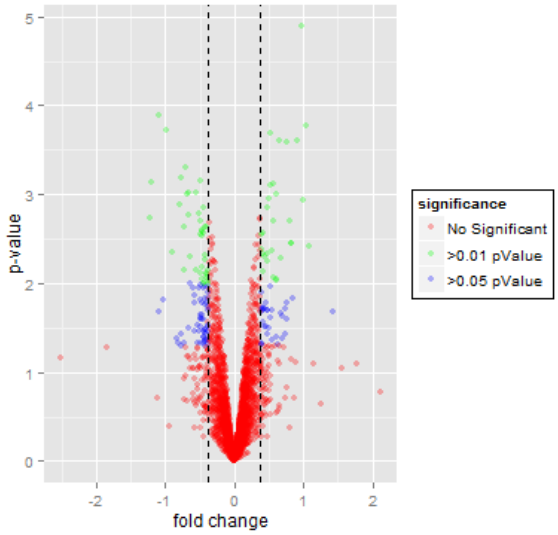


Figura 46: A izquierda sobre-expresión en control y a derecha sobre- expresión en otra condición experimental, los ejes verticales a trazos marcan los valores 0.7 y 1.3 en escala logarítmica. En el eje y utilizamos -log10 de los valores de los estadísticos obtenidos por limma [20]

Generaremos una tabla con los valores Z, estos valores de desviaciones normalizados y estandarizados frente a la media de los niveles de cuantificación presentes en cada una de las muestras para una proteína dada, serán empleados para pintar los Heatmaps (ver figura 47), donde se manera rápida y visual detectamos las proteínas que cambian en las diferentes condiciones experimentales.

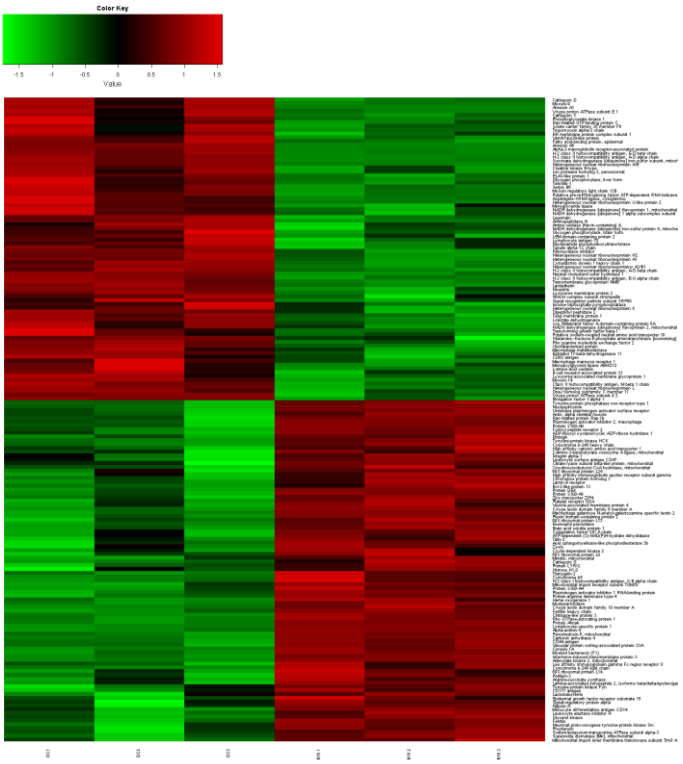


Figura 47: Heatmap clasterizado (eje "x", muestras analizadas y eje "y" nombres de proteínas) [22]

Emplearemos la técnica de análisis por componentes principales así como de análisis de coinercia para observar la correlación de las muestras de nuestro experimento

generaremos también otro tipo de gráficas (dendogramas, heatmaps, volcanos, etc) realizadas a lo largo del proceso.

Todos las tablas de interés son almacenados en el interior de un archivo Excel, así como la imágenes de forma automática gracias al paquete "xlsx".

El código asociado a este flujo se encuentra en los archivos adjuntos a esta memoria.

7.2. Flujo desarrollado en Python

Este segundo flujo se desarrolló posteriormente con el objetivo de comprender y aprender el lenguaje de programación de Python orientado al análisis de datos, así como realizar un flujo único capaz de procesar de manera más exhaustiva los datos provenientes tanto de MaxQuant [14] como de Progenesis [15].

En este flujo trataremos de hacer la obtención de datos de manera automática ya que necesariamente debe obtener la información de dos fuentes diferentes.

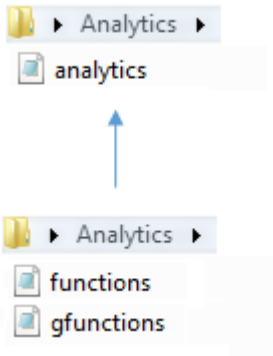


Figura 48: Scripts de Pyhton para realizar análisis

Definiremos varios parámetros antes de iniciar el experimento, estos serán dependientes del experimento:

```
experimentType = 'labelfree'     # 'labelfree' or 'itraq'
software = 'progenesis'     # 'maxquant' or 'progenesis'
distribution = np.array([3,2])
```

Para obtener las columnas con la información de cuantificación llamaremos a aquellas columnas que empiezan por *"Intensity "*, palabra clave presente en estas columnas. A partir de aquí agruparemos estos nombres en las diferentes condiciones experimentales en base a los valores proporcionado por los índices del array, *"distribution"*.

Emplearemos fundamentalmente tres paquetes de Python aplicados al análisis de datos, además de otros como son: Pandas.py, scipy.py y numpy.py.
El paquete de pandas nos dará la opción de poder trabajar en un entorno de tablas (*data frames*) al igual que R el cual es muy cómodo para realizar este tipo de análisis. Pese a la posibilidad de realizar los códigos en un entorno llano como Atom, sublime text 2, notepad ++ u otros editores de texto con asistencia a la programación,

utilizaremos el entorno gráfico de Spyder para Python 2.7. Este entorno nos permite mantener una apariencia similar a R y también bastante parecida a Matlab, además tiene integrada una herramienta desarrollada para Python llamada IPython, la cual nos permite visualizar los resultados de los códigos ejecutados. La interfaz de esta herramienta puede visualizarse en la figura 49.

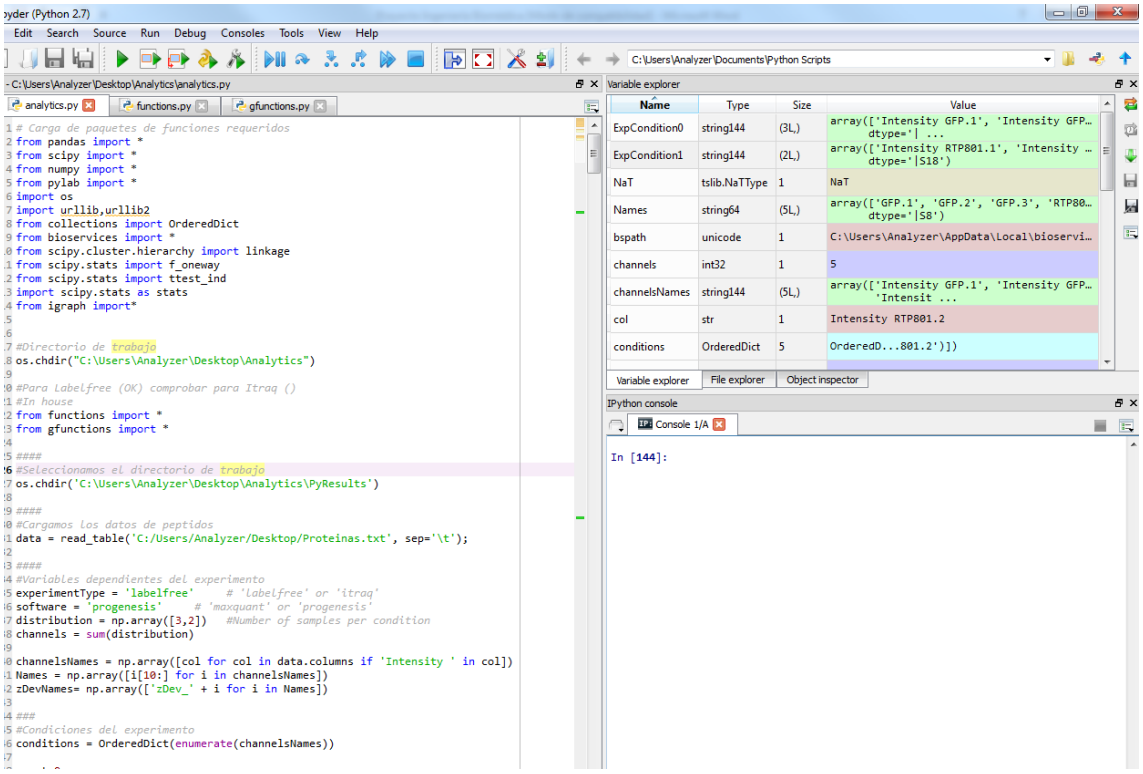


Figura 49: Spyder GUI

El flujo de trabajo será similar salvo por algunas diferencias, entre las cuales están el cálculo por independiente de todos los test estadísticos t para cada par de condiciones presentes en el experimento hasta un máximo de 5 grupos condicionales. De esta manera además de tener una significancia a nivel global calculando el análisis de varianzas Anova, tendremos el nivel de significancia independiente de cada par de condiciones experimentales asociado a una tasa de cambio media entre ellas.

Realizaremos una re-estructuración del conjunto de tablas mostradas. De esta nueva manera tendremos una tabla por cada par de condiciones experimentales cuyo contenido serán las proteínas significativamente expresadas en esta comparación. Mantendremos la tabla con los datos de cuantificación de todas las proteínas identificadas que satisfacen los filtros mencionados previamente y otra tabla con todas las proteínas identificadas en el análisis. Tendremos además una pestaña con todos los gráficos de calidad.

Incluiremos además una nueva tabla a modo de sumario donde quedarán recogidas las significancias de cada una de las proteínas (figura 50) del experimento en relación a cada una de las comparativas posibles. Esto puede resultar útil para realizar agrupaciones para posteriores análisis funcionales.

	Protein names	Proteins	Situation0	Situation1	Situation2
2	atidylinositol 4,5-bisphosphate phosphodiesterase	Q8N3E9	1	0	0
3	10 kDa heat shock protein, mitochondrial	P61604	-1	0	0
4	14-3-3 protein eta	Q04917	0	-1	0
5	-dienoyl-CoA reductase, mitochondrial (Fragmer	E5RFV2	0	1	1
6	2-oxoglutarate dehydrogenase, mitochondrial	E9PDF2	-1	0	1
7	26S protease regulatory subunit 10B	A0A087X2I1	0	-1	-1
8	6S proteasome non-ATPase regulatory subunit 2	Q13200	0	1	1
9	3-hydroxyacyl-CoA dehydrogenase type-2	Q99714	0	-1	-1
10	3-hydroxybutyrate dehydrogenase type 2	Q9BUT1	0	0	1
11	3-ketoacyl-CoA thiolase, mitochondrial	A0A0B4J2A4	0	-1	0
12	3-ketoacyl-CoA thiolase, peroxisomal	P09110	0	0	1
13	4-trimethylaminobutyraldehyde dehydrogenase	P49189	0	1	1
14	40S ribosomal protein S14	P62263	0	1	1
15	40S ribosomal protein S20	P60866	0	0	-1
16	40S ribosomal protein S25	P62851	-1	0	0
17	40S ribosomal protein S28	P62857	0	0	1
18	40S ribosomal protein S4, X isoform	P62701	0	0	1
19	40S ribosomal protein S8	Q5JR95	0	0	-1
20	hosphogluconate dehydrogenase, decarboxylati	P52209	0	1	1
21	6-phosphogluconolactonase	O95336	1	1	0
22	A-kinase anchor protein 12	Q02952	0	-1	0
23	ADP/ATP translocase 1	P12235	-1	-1	0
24	ADP/ATP translocase 3	P12236	-1	0	0
25	ARF GTPase-activating protein GIT1	A0A0C4DGN6	-1	-1	0
26	P synthase F(0) complex subunit B1, mitochondr	P24539	0	-1	0
27	ATP synthase subunit O, mitochondrial	P48047	0	-1	-1
28	ATP synthase subunit alpha, mitochondrial	P25705	0	-1	-1
29	ATP synthase subunit d, mitochondrial	O75947	-1	-1	0
30	ATP synthase subunit e, mitochondrial	P56385	-1	-1	0
31	ATP synthase subunit epsilon, mitochondrial	P56381	0	-1	0
32	P-dependent 6-phosphofructokinase, muscle typ	P08237	1	1	0
33	ATP-dependent RNA helicase A	Q08211	-1	-1	0
34	ATP-dependent RNA helicase DDX1	F1T0B3	0	0	-1
35	ATP-dependent RNA helicase DDX19A	F6QDS0	0	1	1
36	Abscission/NoCut checkpoint regulator	Q96K21	0	0	1
37	Acid ceramidase	E7EMM4	0	1	1
38	Actin, cytoplasmic 1	P60709	0	0	-1
39	Actin-related protein 2/3 complex subunit 5	O15511	0	1	0
40	Acylamino-acid-releasing enzyme (Fragment)	H7C393	0	1	1
41	Acylpyruvase FAHD1, mitochondrial	Q6P587	0	0	-1
42	Adenylyl cyclase-associated protein 1	Q01518	0	1	0
43	Adseverin	Q9Y6U3	0	1	0
44	toxin B1 aldehyde reductase member 2 (Fragme	H3BLU7	0	1	0
45	Alcohol dehydrogenase class-3	P11766	1	1	0

Figura 50: Sumario de expresión y significancia para cada par de condiciones experimentales (-1 = descenso de expresión significativa, 0 = Sin cambio significantivo, +1= aumento de expresión significativa)

Actualmente este flujo de trabajo más automatizado y detallado nos permite trabajar con ambos softwares (MaxQuant y Progenesis) de manera análoga y ha sustituido en mayor grado al flujo creado en R. Generaremos de igual modo gráficos con el fin de observar los datos así como para comprender la información presente.

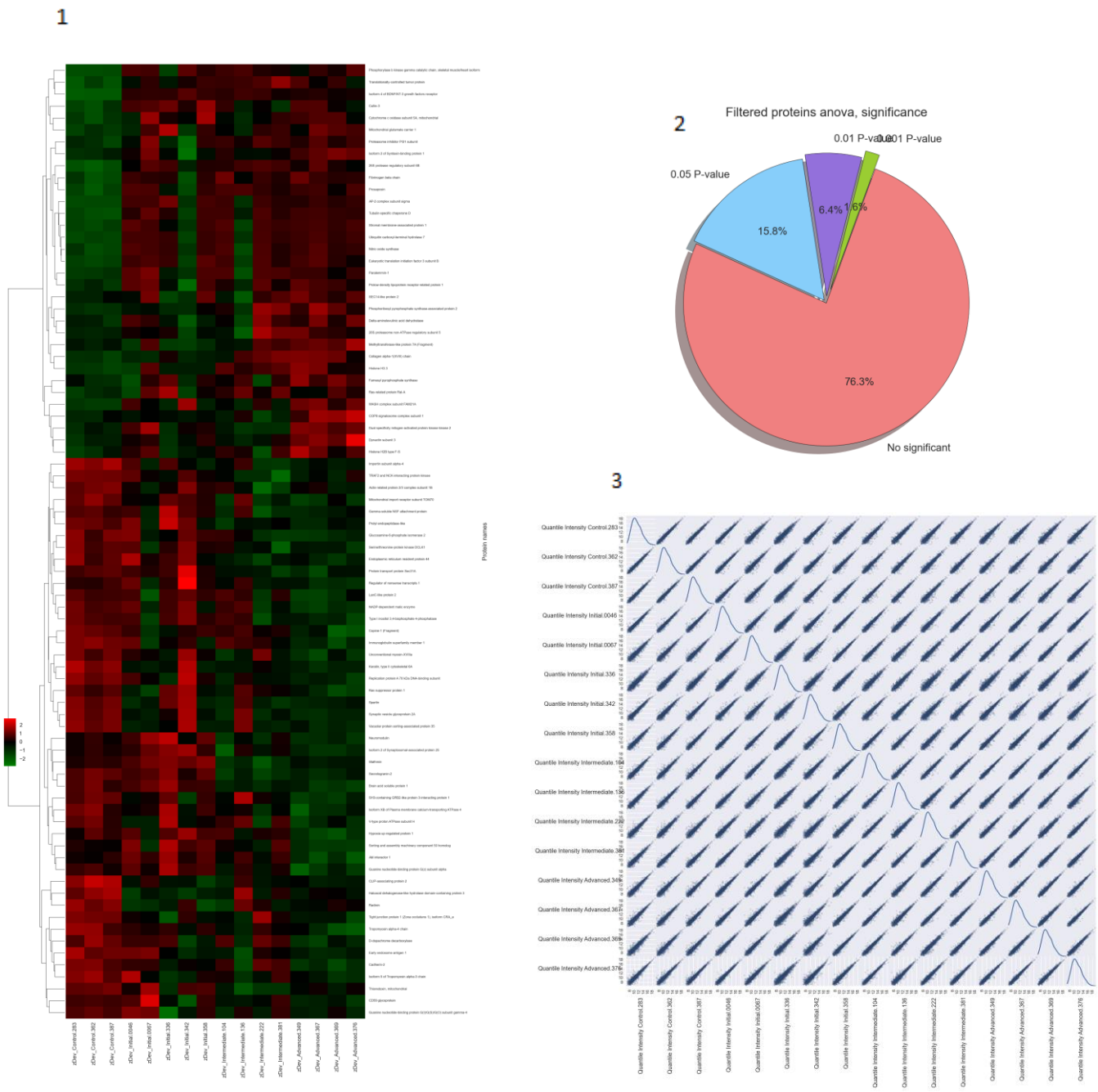


Figura 51: Gráficos generados en el flujo; 1. heatmap, 2. tarta de significancias, 3. Rectas de correlación junto con los gráficos de densidad heatmap de correlaciones Pearson

1. Heatmap: este gráfico muestra los cambios en los niveles de cuantificación asociados a cada proteína para cada una de las muestras con respecto al valor medio de los niveles de cuantificación de cada proteína. Para ello se utilizan los valores Z calculados. Se crean agrupaciones de proteínas según sus cambios de expresión. En el eje x están dispuestas las muestras empleadas en el análisis y en el eje y los nombres de las proteínas.

2. *"Pie"* o tarta: este gráfico nos sirve para visualizar los porcentajes de proteínas que cambian significativamente o por el contrario no cambian en el experimento de manera global. Emplearemos el análisis de varianzas (1-factor anova) para calcular los P-valores y subdividimos el gráfico en cuatro partes

según su no significancia, 95% de significancia, 99% de significancia y 99.9% de significancia.

3. Rectas de correlación: graficaremos las rectas de correlación entre cada par de muestras para observar su semejanza o disparidad en sus niveles de cuantificación obtenidos.

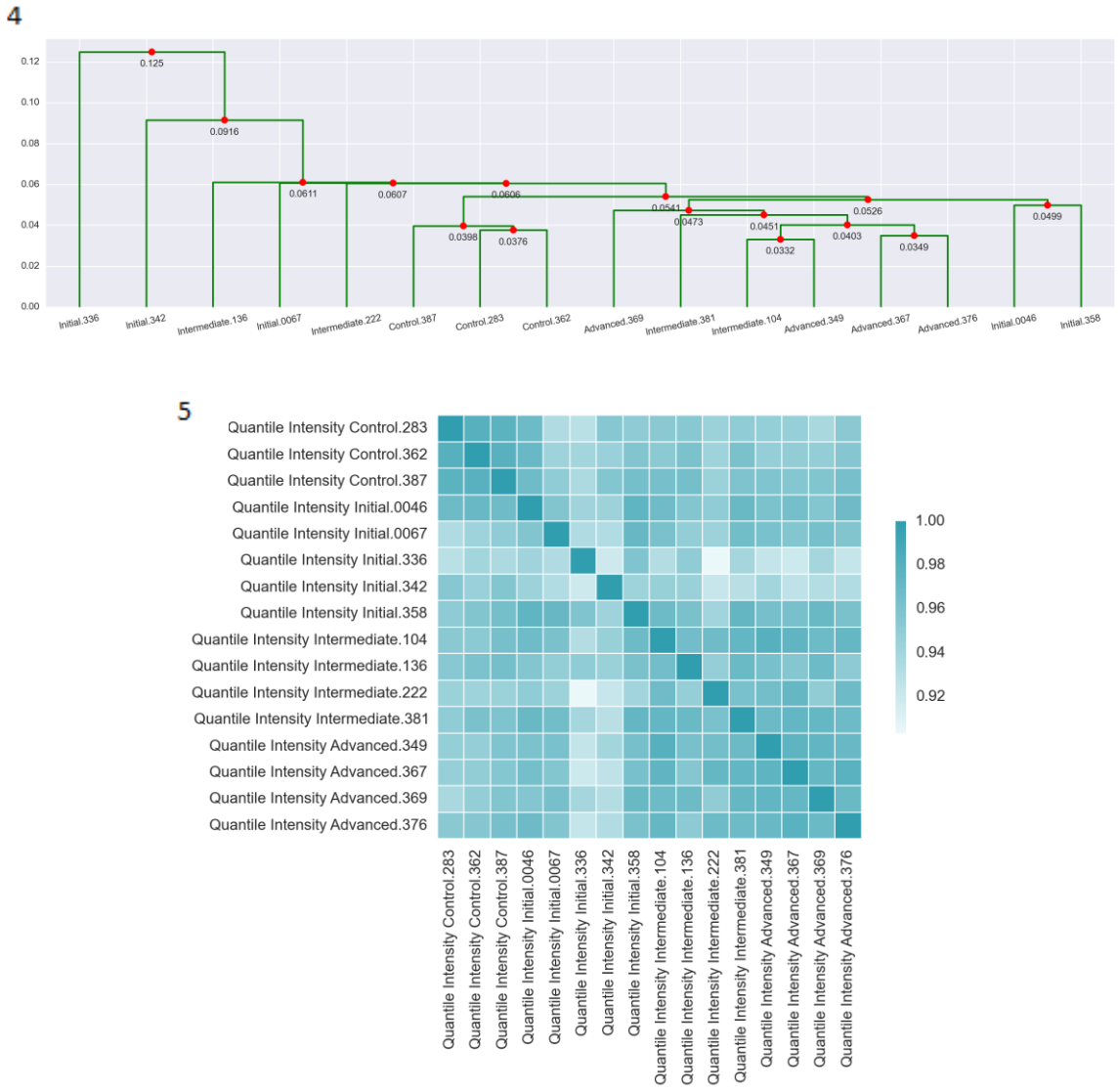


Figura 52: Gráficos generados en el flujo; 4. Dendograma de correlaciones de Pearson, 5. heatmap de correlaciones Pearson

4. Dendograma: forma alternativa para visualizar las correlaciones entre las muestras, adjuntaremos la diferencia en los niveles de correlación para cada nueva rama del dendograma.

5. Heatmap de correlaciones Pearson: en el graficaremos los coeficientes de correlación de Pearson para tener una idea de la similitud o diferencia a nivel global entre las muestras analizadas.

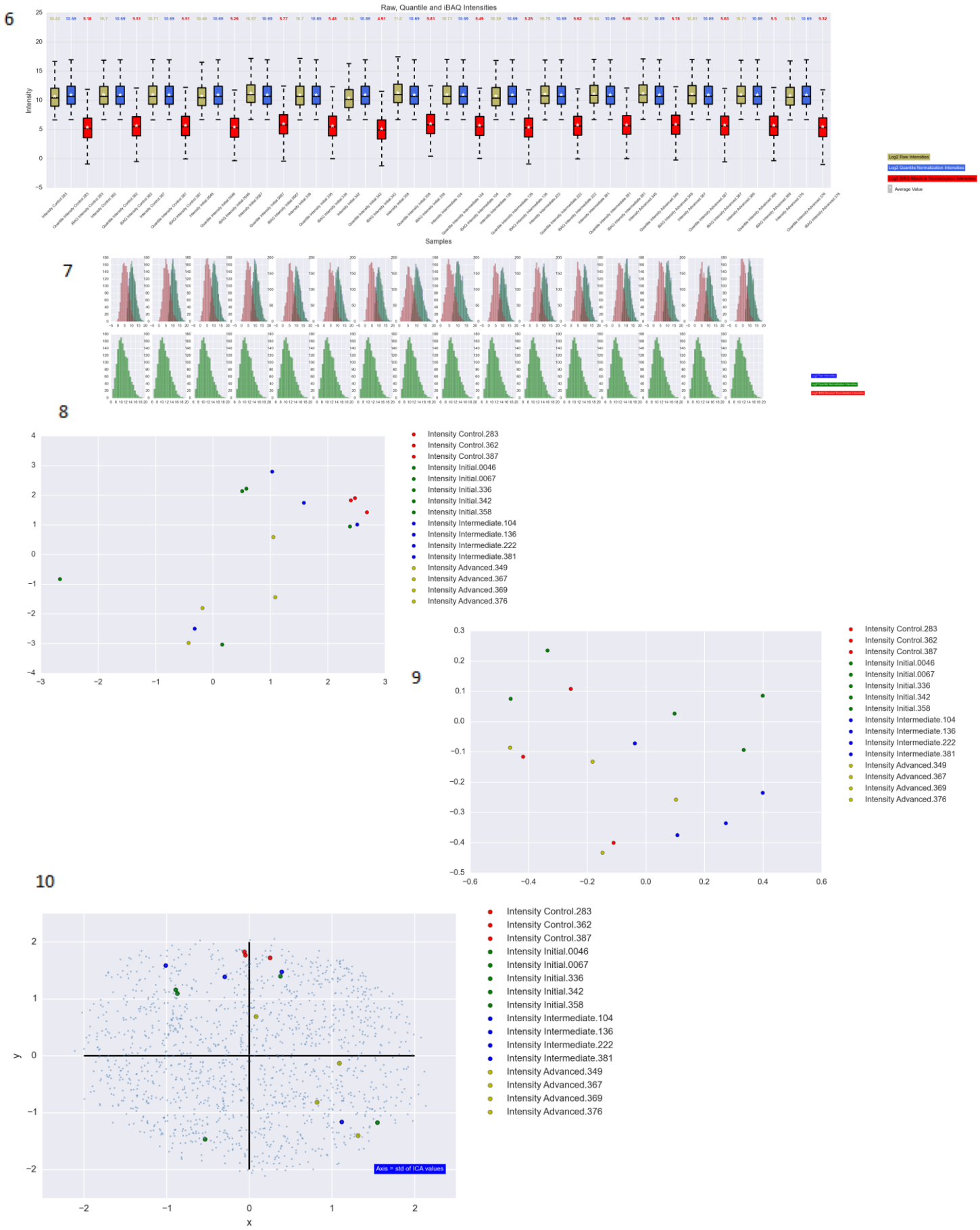


Figura 53: Gráficos generados en el flujo; 6. gráfico de cajas de los valores de intensidades brutas), intensidades normalizadas por quantiles y normalizadas por iBAQ, 7. gráfico análogo al 6 pero representado mediante gráficos de barras (en verde las intensidades normalizadas por quantiles), 8. análisis por componentes principales (PCA), 9. análisis de componentes principales empleando NMDS , 10. análisis de componentes independientes ICA.

6. Box-plot o Gráfico de cajas: lo emplearemos para visualizar la distribución de los niveles de intensidad de cada muestra para los valores crudos de intensidad, los normalizados mediante quantiles y los normalizados mediante iBAQ. En la parte superior mostraremos la media de cada grupo con dos dígitos de precisión.

7. Histogramas: Realizaremos múltiples histogramas para visualizar la frecuencia de los valores de intensidad para cada una de las muestras. Utilizaremos la misma escala de colores representados en la figura 6.

8. PCA: emplearemos la técnica de escalamiento multidimensional clásica (MDS, *"multidimensional scaling"*) para visualizar la correlación entre las muestras del experimento.

9. PCA: emplearemos la técnica de escalamiento multidimensional no métrico NMDS (*"Non-metric multidimensional scaling"*) para visualizar la correlación entre las muestras del experimento.

10. Análisis de componentes independientes ICA: nuevamente utilizaremos esta aproximación para visualizar la correlación entre las muestras analizadas.

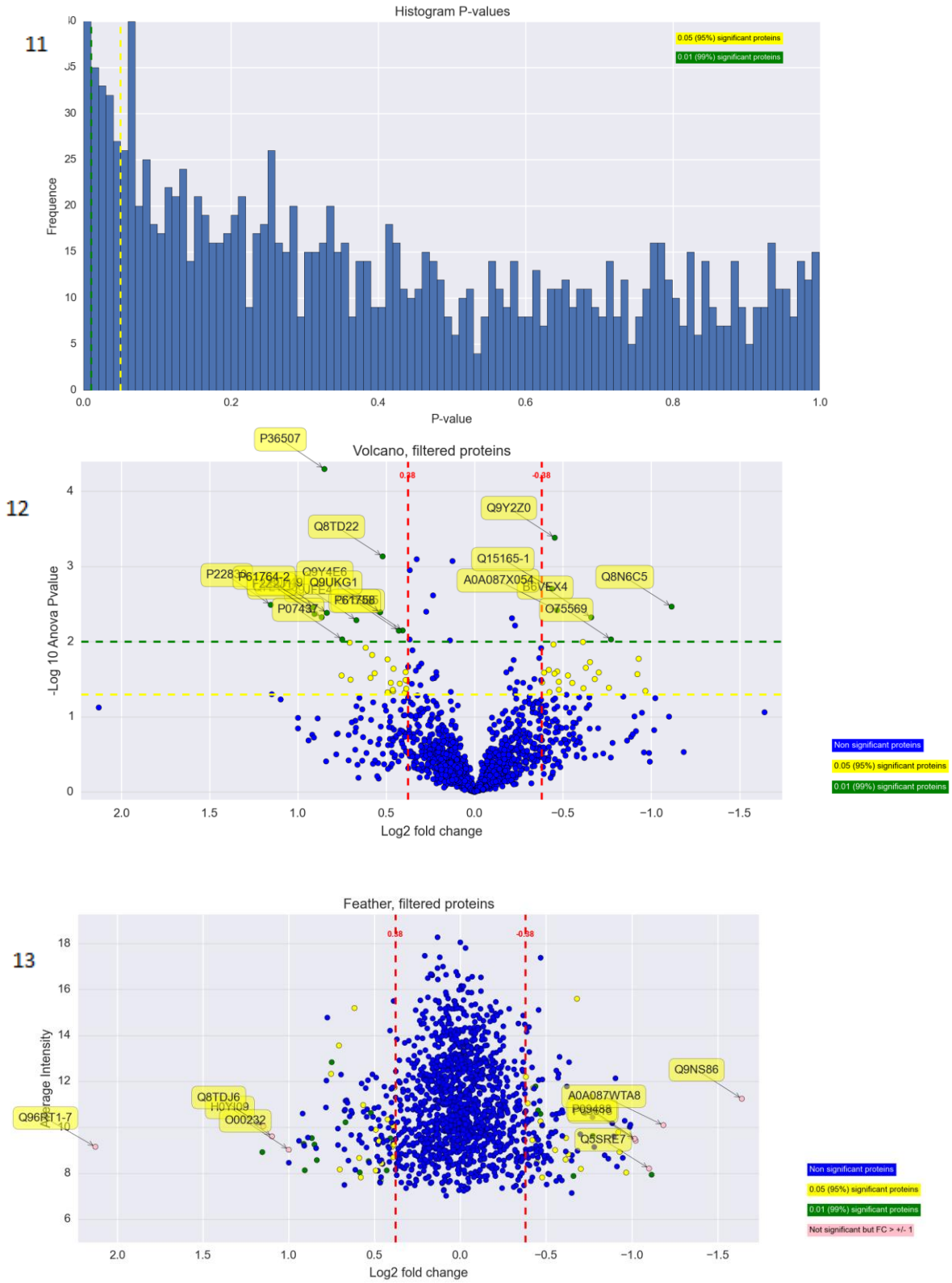


Figura 54: Gráficos generados en el flujo; 11. histograma de estadístico asociado a tasa de cambio (P-valor), 12. volcano, 13. pluma de intensidades medias frente a la tasa de cambio

11. Histograma de P-valores: utilizaremos este gráfico para echar un vistazo rápido a las diferencias o no presentes entre un par de condiciones experimentales. En el eje x los valores-P y en el eje y la frecuencia de estos valores. Añadimos dos rectas verticales en 0.01 de P-valor (verde) y 0.05 de P-valor (amarilla).

12. Volcano: en este gráfico mostraremos la estadística asociada a la comparativa de las proteínas para cada par de condiciones experimentales, en el eje y el logaritmo negativo en base 10 del P-valor y en el eje x la tasa media de cambio en escala logarítmica en base 2. Además colorearemos las proteínas según su significancia y mostraremos los códigos de acceso Uniprot de las proteínas con una significancia del 99% o mayor.

13. Pluma: análogamente al Volcano esta vez enfrentaremos el eje x a la intensidad promedio de cada proteína (eje y), con el fin de ver la significancia relacionada con el nivel de intensidad medio asociada a una proteína. Añadiremos un nuevo color para aquellas proteínas que no siendo significativamente diferentes poseen una tasa media de cambio de 1 en escala logarítmica.

Todas las tablas de interés son almacenadas en el interior de un archivo Excel, así como las imágenes de forma automática.

7.3. Herramientas para el análisis funcional de los datos obtenidos

Emplearemos 5 herramientas llamadas desde R para realizar los análisis funcionales. Consideramos análisis funcionales aquellos que permiten interpretar la salida de datos desde un punto de vista biológico, como por ejemplo ver en que región celular se encuentran las proteínas identificadas, que función desempeñan, en que ruta metabólica participan, etc. Entre ellas se encuentran:

- String

Base de datos que contiene información acerca de las posibles interacciones físicas o funcionales entre un grupo de proteínas [23].

Haremos uso del servicio web (http://string-db.org/), utilizando la información presente en su base de datos con un corte de 0.7 en el nivel de fiabilidad de las anotaciones empleadas. Mediante los comandos presentes en post_procesado_MQv3.R, podremos colorear según su nivel de expresión (*fold change*) las proteínas presentes en la red de interacción a nivel funcional. Esto añade más información a la ya presente en la versión web de la herramienta. Accederemos a este servicio utilizando el paquete de R "STRINGdb". Esta herramienta como otras se desarrolló para genómica y por tanto funcionan con listas de nombres de genes que será lo que utilizaremos.

Para ello inicializamos una clase de referencia STRINGdb, determinamos la versión, la taxonomía referente a la especie de la cual provienen los datos a analizar (NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy)) y el umbral para las interacciones a ser consideradas.

```
string_db <- STRINGdb$new( version="9_05", species=9606,
                           score_threshold=700, input_directory="" )
```

Definimos los valores a ser utilizados así como la selección de color para la condición, tasa de cambio para cada par de condiciones experimentales.

```
genes<- string_db$map(significantPvalue1, "GeneNames", removeUnmappedRows = TRUE )

genes$color<-as.factor(ifelse(genes$FoldChange1 > 1.3,
                              "#FF0000",
                              ifelse(genes$FoldChange1 <0.77,
                                     "#00FF00",
                                     "#FFFFFF")
))
```

Obtenemos redes de interacción como la presente:

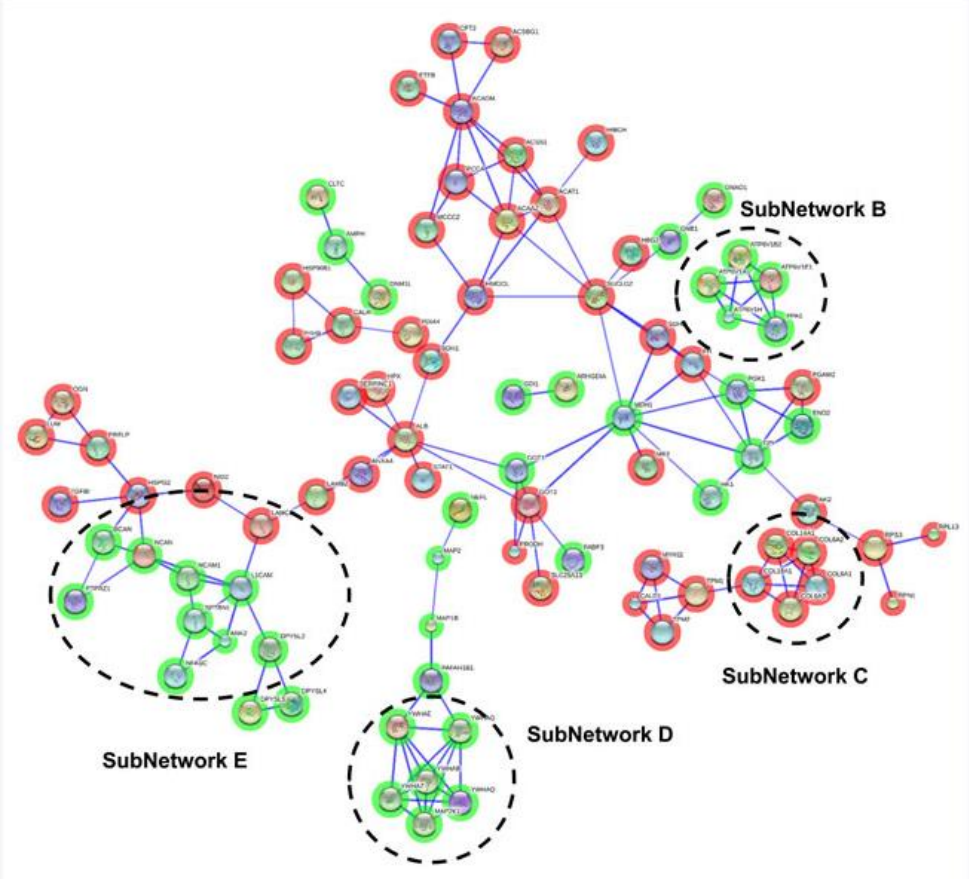


Figura 55: Red proporcionada por String [24], en verde se representan las proteínas que disminuyen su expresión y en rojo proteínas que aumentan de expresión en el sistema biológico analizado.

- David

Base de datos que contiene información funcional de genes. Permite identificar los procesos biológicos enriquecidos en una lista de genes, obtener los Go asociados a esos genes o encuentrar las vías en las que se hallan estos genes (Kegg). Permite también la conversión entre códigos asociados a genes entre otras utilidades. [25][26]

Podemos emplear esta herramienta desde R, y generar a partir de los datos obtenidos diferentes gráficos. Para ello emplearemos el paquete de R "FGNet".

Para mejorar el número de proteínas encontradas en David empleando este paquete, realizaremos primero una conversión de los códigos asociados a los genes, valores que utilizaremos para realizar la búsqueda ya que con los códigos de Uniprot (obtenidos a partir de las bases de datos ".fasta") se pierde un alto porcentaje de genes.

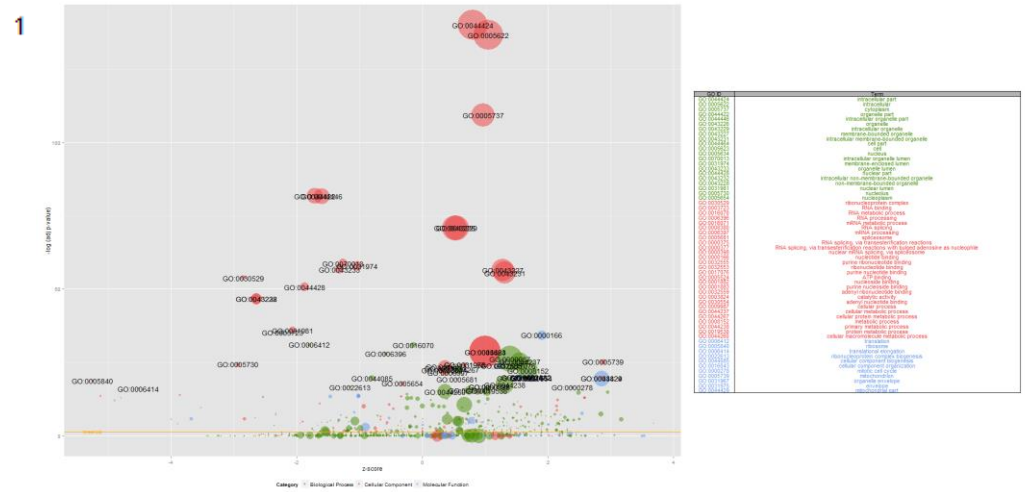
Utilizaremos para ello la función "convertidor" contenida en convertidorNombres.R, la cual gracias al uso de la base de datos biomart (www.biomart.org) nos servirá para realizar la conversión a código entrez, este se considera el estándar como referencia.

```
library('biomaRt')

convertidor<-function(x){
  listaDB<-listMarts(mart = NULL, host="www.biomart.org", path="/biomart/martservice",
                     port=80, includeHosts = FALSE, archive=FALSE, ssl.verifypeer=TRUE, verbose = FALSE)
  ensembl <- useMart('ensembl')
  lista<-listDatasets(ensembl)
  ensembl <- useMart('ensembl', dataset=x)
  listAtributes<-listAttributes(ensembl)
  Filters<-listFilters(ensembl)
  annotation <- getBM(attributes=c("entrezgene","uniprot_genename"), filters="uniprot_genename", Genes, mart=ensembl)

  return(annotation)
}
```

Una vez transformados utilizaremos la siguiente función para obtener los datos asociados (gen ontology) a estos genes a nivel de componente celular (CC), funcionamiento molecular (MF) y proceso biológico (BP). Estas anotaciones nos indican en que procesos o vías se encuentras los genes obtenidos en relación a las proteínas identificadas por espectrometría de masas.

```
feaResults_David <- fea_david(EGenes,email="martinezdemorentin.59596@e.unavarra.es", geneIdType="ENTREZ_GENE_ID")
```

Tras algunos ajustes y la selección de los términos de interés graficaremos los resultados obtenidos en relación a la estadística asociada a la obtención de estas anotaciones. Utilizaremos el estadístico corregido de Benjamin Hochberg (FDR), para el filtrado de las anotaciones obtenidas.



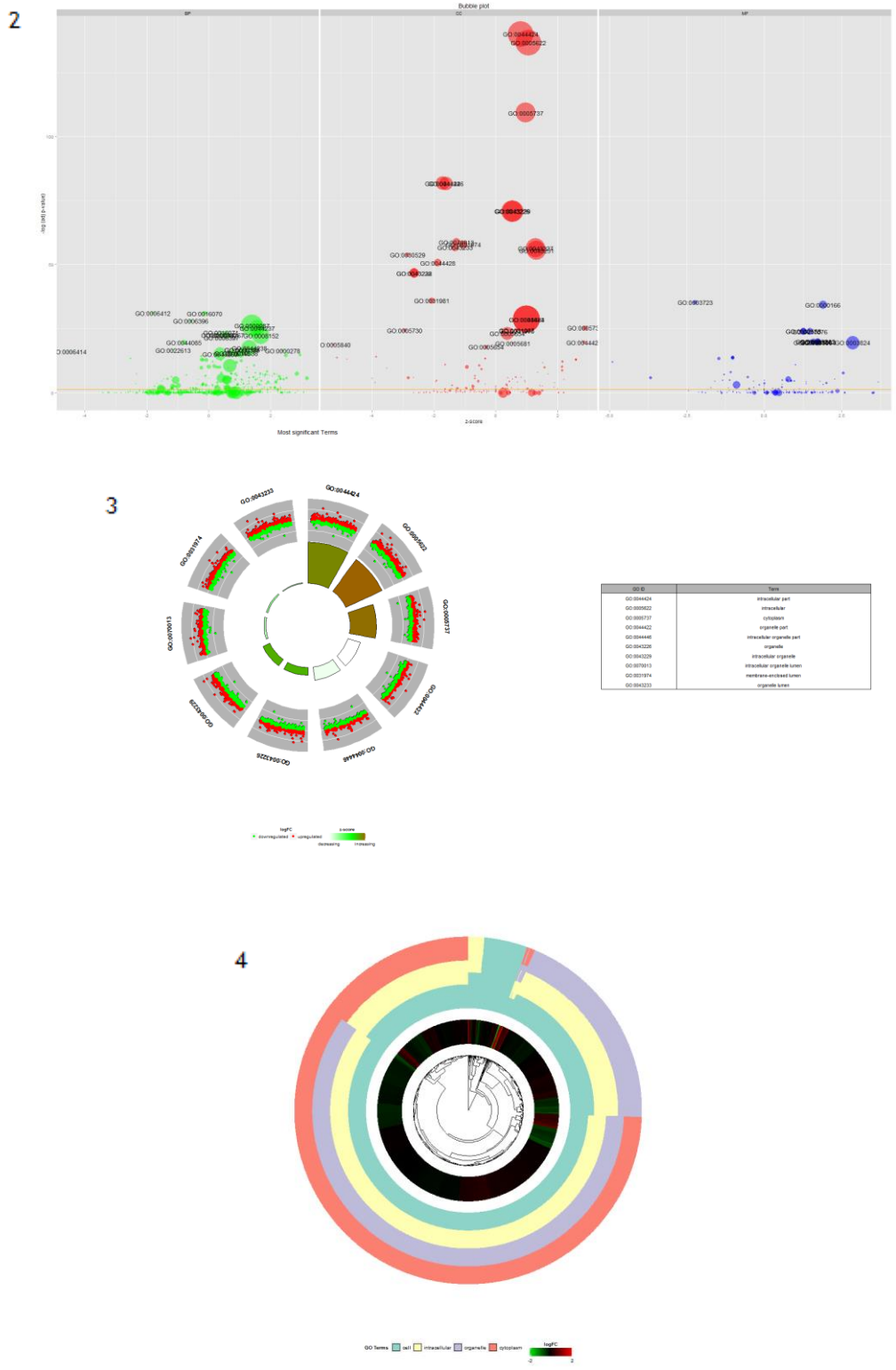


Figura 56: Gráficos obtenidos de Gen Ontology (David).

1. Gráfico con los términos obtenidos (eje y estadístico, eje x los valores Z) pintamos en diferentes colores la procedencia de los términos y los anotamos a la izquierda (verde los Go asociados a procesos biológicos, rojo los Go asociados a componentes celulares y azul los Go asociados a funciones moleculares)
2. Muestra la misma información que el gráfico 1 pero esta vez de manera separada (proceso biológico, componente celular y función molecular)
3. Gráfico donde representamos la funcionalidad obtenida (Go) más significativa en función a la expresión que presentan las proteínas cuantificadas por proteómica para un par de condiciones experimentales.
4. Gráfico de círculos donde se representan la expresión obtenida en los datos de cuantificación por poteómica (circulo interior) y la presencia de estas proteínas en los términos seleccionados (círculos externos concéntricos)

- Kegg, GO y Reactome

En el archivo FunctionAnalysis.R tenemos el código para obtener las tablas de información de las bases de datos de Kegg [27][28], Go y Reactome [29][30]. En el caso de los dos primeros emplearemos bases de datos locales, descargables desde el repositorio de Bioconductor. En el caso de Reactome utilizaremos su servicio web.

Estas bases de datos nos dan información sobre rutas o pathways más específicos y en el caso específico de Reactome nos aporta mayor información sobre reacciones químicas que ocurren en los diferentes pathways.

Llamando a este servicio proporcionado por Kegg desde R empleando el paquete "KEGG.db", podemos ilustrar las proteínas presentes en nuestra lista obtenida coloreándolas según su nivel de expresión.

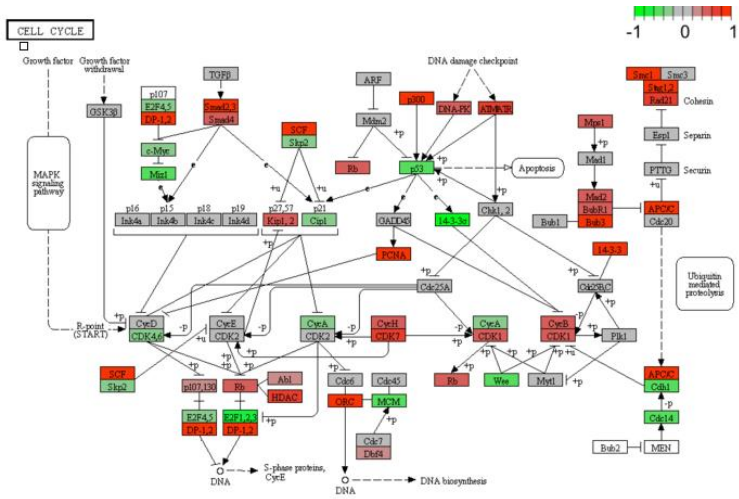


Figura 57: Pathway proporcionado por Kegg orientado a la comprensión del ciclo celular

I. Conclusiones

- ✓ El software MaxQuant es una elección válida para el análisis de datos si utilizamos la técnica de iTRAQ, no obstante el uso de Protein Pilot mejora los resultados obtenidos en el caso de usar un espectrómetro de masas Triple TOF 5600.

- ✓ El software Progenesis es una elección acertada para el análisis de datos si utilizamos la técnica de Label-free, mejorando notablemente los resultados obtenidos utilizando MaxQuant en el caso de usar un espectrómetro de masas Triple TOF 5600.

- ✓ Los flujos de análisis implementados son válidos y han sido utilizados en la realización de varias publicaciones científicas.

- ✓ La automatización en las salidas de datos permite la comprensión y el análisis biológico de los datos por parte de los investigadores que realizan experimentos de proteómica.

- ✓ A día de hoy, los flujos de análisis derivados de este proyecto se utilizan de manera rutinaria en la Unidad de Proteómica de Navarrabiomed.

- Implantación de nuevos flujos de análisis basados en otras tecnologías como SWATH [32].

- Una vez aprendidas las bases del lenguaje de programación Python así como las bases en la realización de análisis proteómicos por espectrometría de masas, se propone la realización de una herramienta online capaz de realizar estos análisis y representar de forma dinámica los datos resultantes de estos análisis. Para ello se propone el uso de HTML, CSS, JAVASCRIPT como lenguajes para la realización del fron-end y Python como lenguaje para la realización del back-end así como la utilización de bases de datos no relacionales como MONGODB.

III. Referencias Bibliográficas

[1] Guía metabólica. Del gen a la proteína (http://www.guiametabolica.org/noticia/gen-proteina-0)

[2] Carlos Yábar V, Ysabel Montoya P. "Síntesis *in vitro* de la proteína de la envoltura del virus peruano de la fiebre amarilla". Revista Peruana de Medicina Experimental y Salud Pública 2001, 18(2).

[3] Domon B, Aebersold R. Mass spectrometry and Protein analysis. Science. 2006 Apr 14; 321(5771):212-7.

[4] Breker M, Schuldiner M. The emergence of proteome-wide technologies: systematic analysis of proteins comes of age. Nat Rev Mol Cell Biol. 2014 Jul; 15(7):453-64. doi: 10.1038/nrm3821. Epub 2014 Jun 18.

[5] "How Does High Performance Liquid Chromatography Work?" Waters. Available from: http://www.waters.com/waters/es_ES/How-Does-High-Performance-Liquid-Chromatography-Work%3F/nav.htm?cid=10049055&locale=es_ES.

[6] Ji YB, Xu QS, Hu YZ, Heyden YV. Development, optimization and validation of fingerprint of Ginkgo biloba extracts by high-performance liquidchromatography. J Chromatogr A. 2005 Feb 25; 1066(1-2):97-104.

[7] Aebersold R, Mann Mass spectrometry-based proteomics. Nature. 2033 Mas 13; 422 (6928): 198-207.

[8] Joaquín Abián, Montserrat Carrascal y Marina Gay. Introducción a las técnicas espectrométricas para la caracterización de péptidos y proteínas en proteómica.

[9] Andrews GL, Simons BL, Young JB, Hawkridge AM, Muddiman DC. Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). Anal Chem. 2011 Jul 1; 83 (13):5442-6. doi: 10.1021/ac200812d.

[10] Unwin RD, Griffiths JR, Whetton AD. Simultaneous analysis of relative protein expression levels across multiple samples using iTRAQ isobaric tags with 2D nano LC-MS/MS. Nat Protoc. 2010 Sep; 5 (9): 1574-82. Doi: 10.1038/nprot.2010. 123. Epub 2010 Aug 20.

[11] Nahnses S, Bielow C, Reinert K, Kohlbacher O. Tools for label-free peptide quantification. Mol Cell Proteomics. 2013 Mar; 12(3):549-56. doi: 10.1074/mcp.R112.025163. Epub 2012 Dec 17.

[12] María Luz Valero, Virginia Rejas, Manuel Mateo Sánchez del Pino. "Luces y sombras de la cuantificación con iTRAQ". Proteómica 2010. 5, 96-97.

[13] Laurent Gatto and Sebastian Gibb. “MSnbase: labelled and label-free MS2 data pre-processing, visualisation and quantification.” October 13, 2015. Available from: https://www.bioconductor.org/packages/release/bioc/vignettes/MSnbase/inst/doc/MSnbase-demo.pdf.

[14] Cox, J. and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 2008, 26, pp 1367-72.

[15] Qi D, Brownridge P, Xia D, Mackay K, Gonzalez-Galarza FF, Kenyani J, Harman V, Beynon RJ, Jones AR. (2012) A software toolkit and interface for performing stable isotope labeling and top3 quantification using Progenesis LC-MS. Omics 16:489-495.

[16] S. Gibb, LM. Breckels, T. Lin Pedersen, VA. Petyuk, KS. Lilley and L. Gatto. A current perspective on using R and Bioconductor for proteomics data analysis. Computational Proteomics Unit and Cambridge Centre for Proteomics.

[17] Florian P. Breitwieser, Jacques Colinge. Isobar package for iTRAQ and TMT protein quantification. Bioconductor. October 13, 2015.

[18] Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA. The Paragon Algorith, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol Cell Proteomics. 2007 Sep; 6(9):1638-55. Epub 2007 May 27.

[19] Cox, J. Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda a peptide search engine integrated into MaxQuant environment. J Proteome Res. 2011 Apr 1; 10(4):1794-805. doi: 10.1021/pr101065j. Epub 2011 Feb 22.

[20] Gordon K. Smyth, Matthew Ritchie, Natalie Throne, James Wttenhall, Wei Shie and Yifang Hu Bioinformatics Division. Linear Model for Microarray and RNA-Seq Data.

[21] Mann K, Edsinger E. The Lottia gigantea shell matrix proteome: re-analysis including MaxQuant iBAQ quantitation and phosphoproteome analysis. Proteome Sci. 2014 May 18; 12:28. doi: 10.1186/1477-5956-12-28. eCollection 2014.

[22] Gato-Cañas M, Martinez de Morentin X, Blanco-Luquin I, Fernandez-Irigoyen J, Zudaire I, Liechtenstein T, Arasanz H, Lozano T, Casares N, Chaikuad A, Knapp S, Guerrero-Setas D, Escors D, Kochan G, Santamaría E. “A core of kinase-regulated interactomes defines the neoplastic MDSC lineage”. Oncotarget. 2015 Sep 29; 6(29):27160-75. doi: 10.18632/oncotarget.4746.

[23] Franceschini A, Szklarezyk D, Frankild S, Kuhn M, Siminivic M, Roth A, Lin J, Miguez P, Bork P, von Merging C and Jesen LJ. STRING v9.1: protein-protein interaction

networks, with increased coverage and interaction. Nucleic Acids Res. 2013; 41:D808-815.

[24] Zelaya MV, Pérez-Valderrama E, de Morentin XM, Tuñon T, Ferrer I, Luquin MR, Fernandez-Irigoyen J, Santamaría E. “Olfactory bulb proteome dynamics during the progression of sporadic Alzheimer's disease: identification of common and distinct olfactory targets across Alzheimer-related co-pathologies”. Oncotarget. 2015 Oct 28. doi: 10.18632/oncotarget.6254.

[25] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009; 4(1):44-57.

[26] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37(1):1-13.

[27] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016 Jan 4; 44(D1):D457-62. doi: 10.1093/nar/gkv1070. Epub 2015 Oct 17.

[28] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima. BlastKOALA nad GhostKOALA: KEGG Tool for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2015 Nov 14. pii: S0022-2836(15)00649-X. doi: 10.1016/j.jmb.2015.11.006. [Epub ahead of print].

[29] Joshi-Tope G, Gillespie M, Vastrik I, D’Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. *1:D428-32.*

[30] Matthews L, D’Eustachio P, Gillespie M, Croft D, de Bono B, Gopinath G, Jassal B, Lewis S, Schmidt E, Vastrik I, Wu G, Birney E, Stein L. An Introduction to the Reactome Knowledgebase of Human Biological Pathways and Processes. *Bioinformatics Primer, NCI/Nature Pathway Interaction Database.* doi:10.1038/pid.2007.3.

[31] Heaven MR, Funk AJ, Cobbs AL, Haffey WD, Norris JL, McCullumsmith RE, Greis KD. Systematic evaluation of data-independent acquisition for sensitive and reproducible proteomics-a prototype design for a single injection assay. J Mass Spectrom. 2016 Jan; 51(1): ii. doi: 10.1002/jms.3648.

[32] Gillet LC, Navarro P, Tate S et. al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012 Jun; 11(6):O111.016717. doi: 10.1074/mcp.O111.016717. Epub 2012 Jan 18.

IV. Anexos

Anexo 1: Publicaciones

Gato-Cañas M, Martinez de Morentin X, Blanco-Luquin I, Fernandez-Irigoyen J, Zudaire I, Liechtenstein T, Arasanz H, Lozano T, Casares N, Chaikuad A, Knapp S, Guerrero-Setas D, Escors D, Kochan G, Santamaría E. "A core of kinase-regulated interactomes defines the neoplastic MDSC lineage". Oncotarget. 2015 Sep 29; 6(29):27160-75. doi: 10.18632/oncotarget.4746.

Gato M, Blanco-Luquin I, Zudaire M, de Morentin XM, Perez-Valderrama E, Zabaleta A, Kochan G, Escors D, Fernandez-Irigoyen J, Santamaría E. "Drafting the proteome landscape of myeloid-derived suppressor cells". Proteomics. 2015 Sep 25. doi: 10.1002/pmic.201500229.

Zelaya MV, Pérez-Valderrama E, de Morentin XM, Tuñon T, Ferrer I, Luquin MR, Fernandez-Irigoyen J, Santamaría E. "Olfactory bulb proteome dynamics during the progression of sporadic Alzheimer's disease: identification of common and distinct olfactory targets across Alzheimer-related co-pathologies". Oncotarget. 2015 Oct 28. doi: 10.18632/oncotarget.6254

Anexo 2: **MaxQuant**

TUTORIAL BÁSICO DE ANÁLISIS EN MAXQUANT

1. Panel *"RAW FILES"*

En este panel seleccionamos la opción *"Load"*, para cargar nuestros archivos a analizar, estos deben ser de las siguientes extensiones: ".raw", ".mzxml", ".wiff" y ".p".

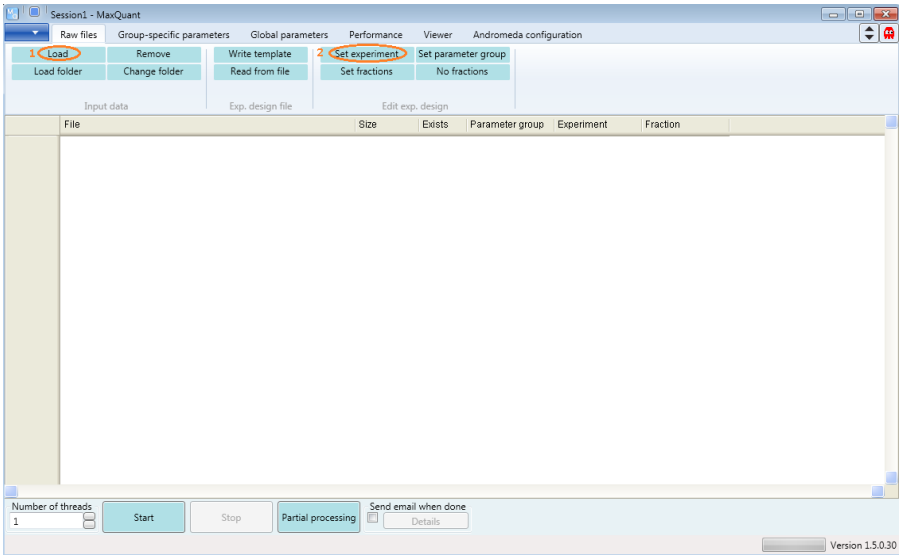


Figura 1: Panel *"Raw Files"*

Una vez seleccionados le asignamos un nombre a nuestro experimento, en la ventana *"Set experiment"*.

En la ventana azul del menú principal (ver figura 2), debajo del icono de la aplicación, tenemos a nuestra disposición una paleta desplegable donde podemos cargar parámetros para nuestro análisis "mqpar.xml".

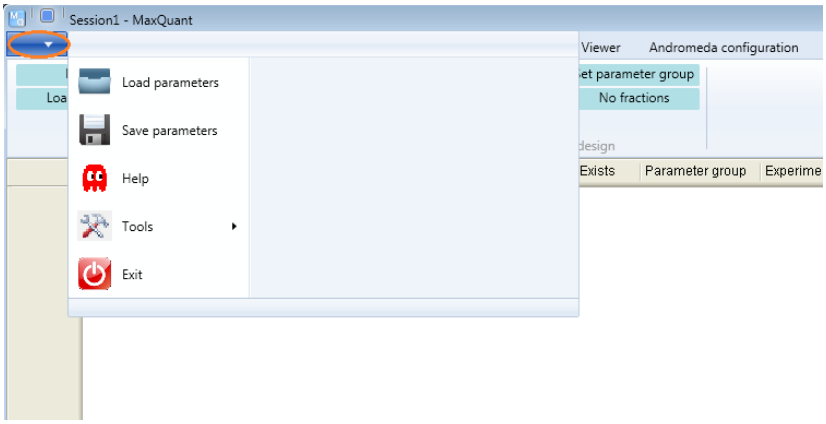


Figura 2: Menu de Inicio

Un detalle muy importante, es la celda en la parte izquierda inferior de la aplicación, donde podemos seleccionar el número de núcleos, que vamos a utilizar en nuestro análisis, en este caso está por defecto a 1 (ver figura 1).

2. Panel *"GROUP-SPECIFIC PARAMETERS"*

Aquí tenemos algunos de los parámetros que corresponden a la etapa MS o etapa de identificación.

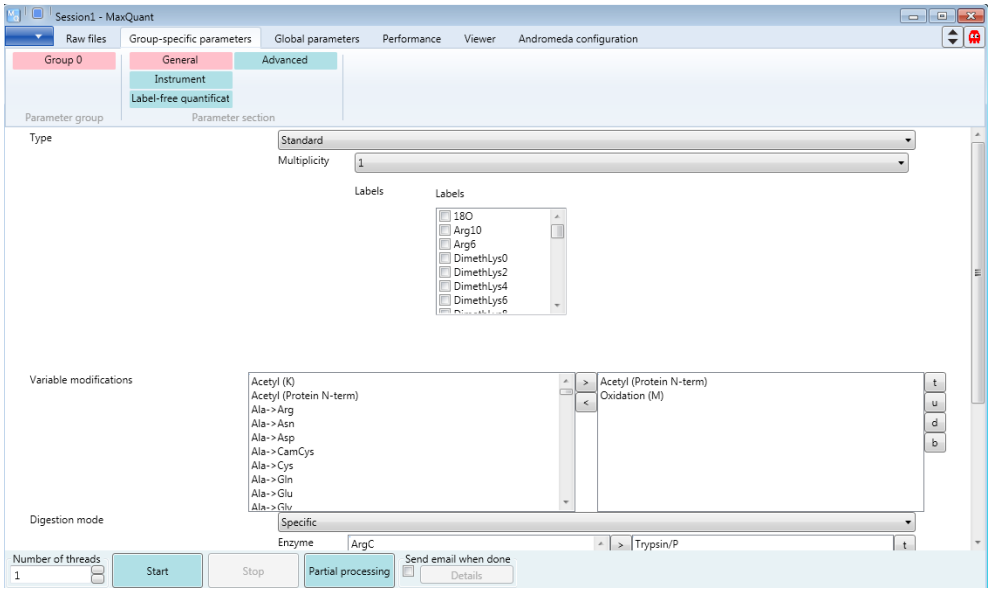


Figura 3: Ventana *"General"* del panel *"Group-specific parameters"*

En la venta seleccionada *"General"*, debemos elegir:

- El tipo de marcadores utilizados en el análisis, donde podemos elegir entre: *"Standard"* o *"Reporter Ion"*.

  - Si hemos elegido "Standard" podemos elegir entre 1, 2 o 3 niveles de *"Multiplicity"*, según nuestro tipo de experimento Silac, Label Free, si nos interesa hacer una comparativa entre diferentes *"labels"*.
  - Si hemos elegido *"Reporter Ion"*, debemos elegir los *"isobaric labels"*, propios de nuestro experimento.

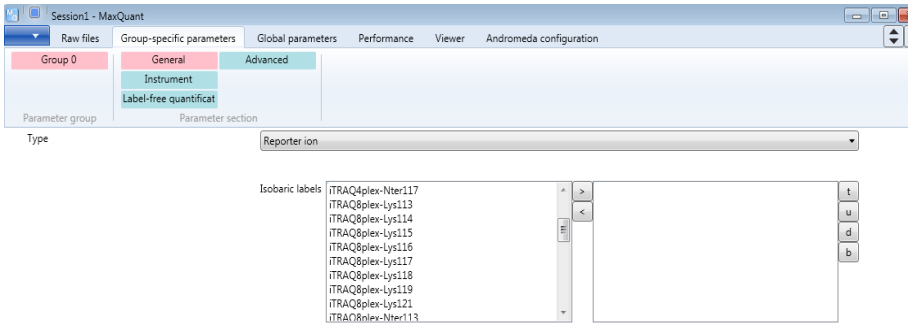


Figura 4: Selección de los reporteros iTRAQ o TMT

- Las modificaciones variables, se asignan aquí como podemos ver en la figura 5, no confundir con las modificaciones fijas que se asignan más adelante (ver figura 6).

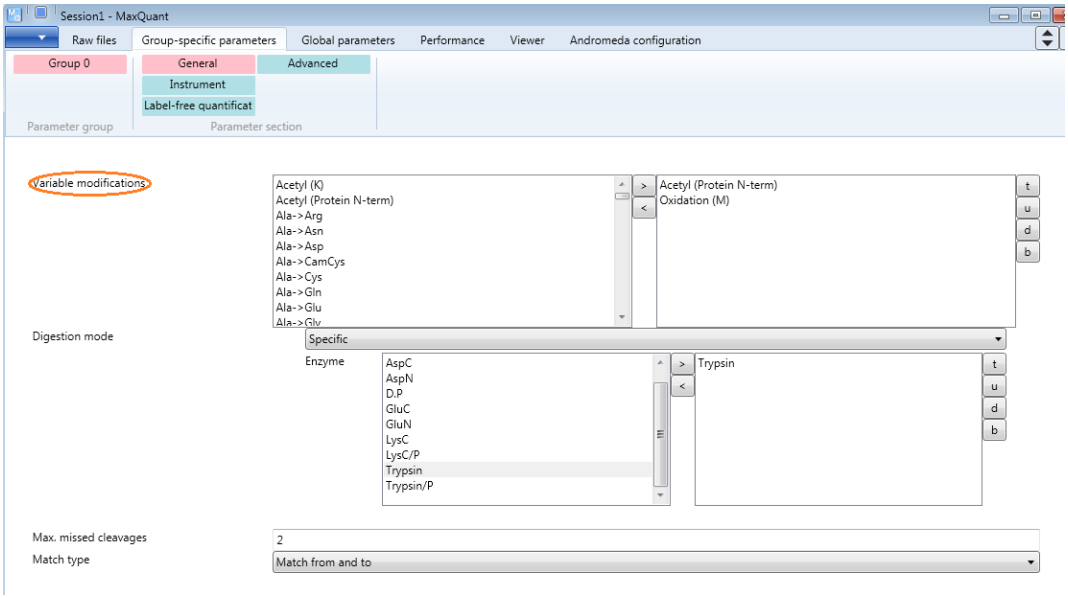


Figura 5: Selección de proteasa, modificaciones variables y missed cleavages permitidos

- Se selecciona la proteasa encargada de la digestión o corte de las proteínas en sus correspondientes péptidos.

No olvidar, que en la ventana *"Andromeda configuration"* se pueden crear nuevas Proteasas o Modificaciones para su uso, así como la modificación de las presentes si es necesario.

3. Panel *"GLOBAL PARAMETERS"*

Primero debemos elegir nuestro archivo ".fasta" (base de datos), para el proceso de identificación. Una opción para ello es: http://www.uniprot.org/.

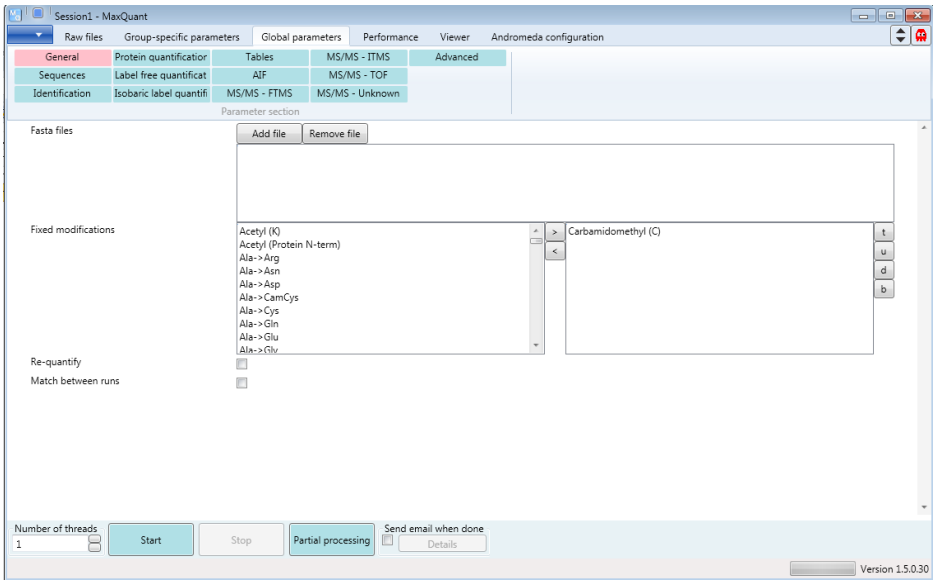


Figura 6: Selección de bases de datos ".fasta", modificaciones fijas y funciones Re-quantify y Match between runs.

Desde este enlace podemos descargar nuestra base de datos para su posterior importación en MaxQuant.

En segundo lugar elegiremos las modificaciones fijas que siempre sucederán ante la presencia de un aminoácido. (Aquí también podemos usar nuestras modificaciones creadas en Andrómeda).

La opción *"Re-quantify"*, está pensada para análisis en los que se han usado 2 o 3 niveles de multiplicidad, de *"labels"*, como por ejemplo en SILAC. No obstante mejora los resultados de la cuantificación en otros experimentos.

A continuación seleccionaremos los niveles de FDR así como el número mínimo de aminoácidos necesarios para la identificación de un péptido como vemos en la figura 7.

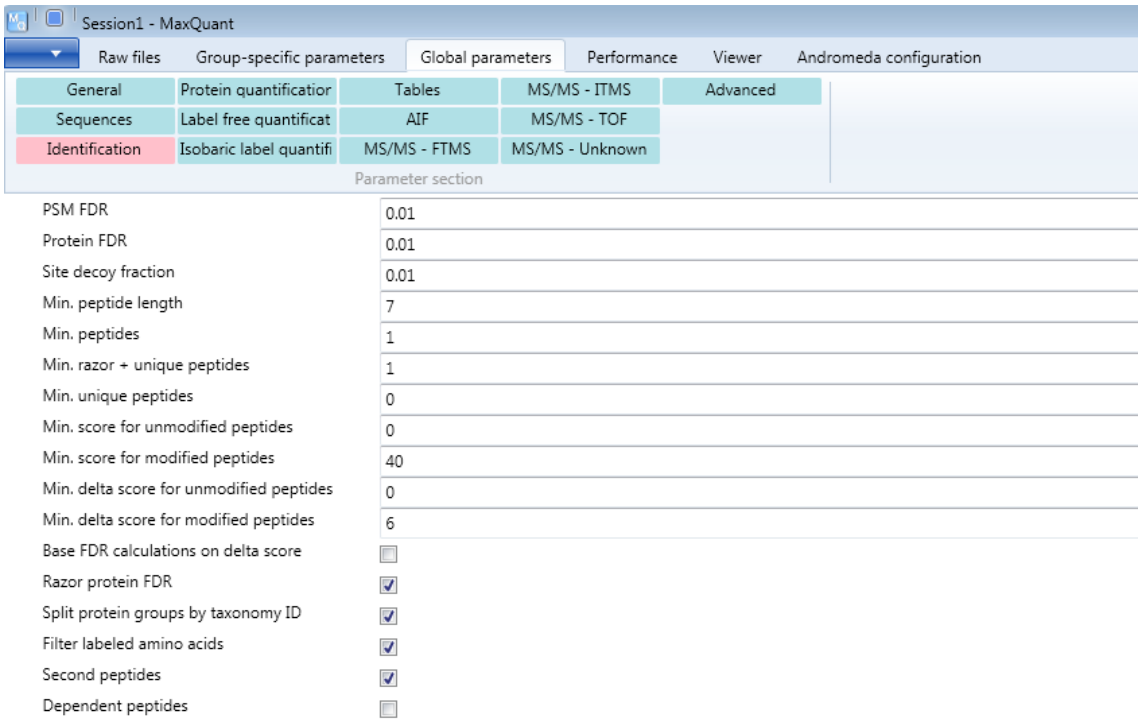


Figura 7: Panel de selección del porcentaje de falsos positivos permitidos asi como otros filtros de selección de información

Una vez seleccionadas estas opciones pasaremos a elegir los péptidos que serán empleados para realizar la integración a proteína (no utilizaremos los datos desarrollados de esta manera en nuestros análisis ya que realizaremos nuestra propia integración a proteínas).

En la pestaña *"Peptides for quantification"* se puede elegir entre usar todos los péptidos identificados, solo los únicos propios para una proteína o los únicos más los péptidos que principalmente identifican una proteína (razor).

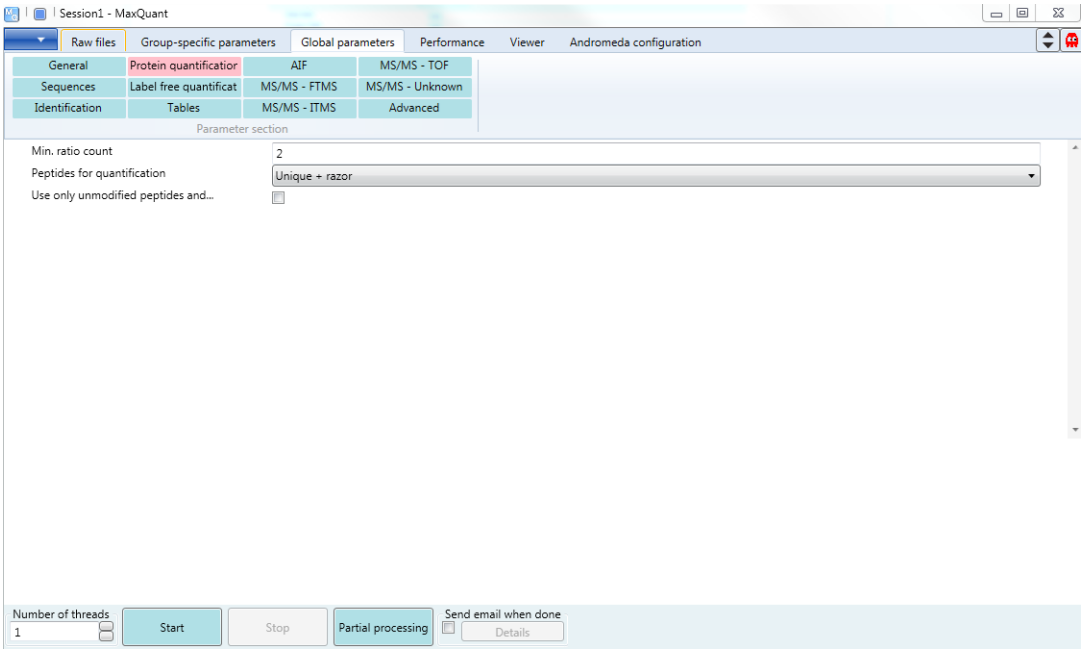


Figura 8: Selección de péptidos para realizar la cuantificación

Por último se selecciona la tolerancia de desviación de masa en partes por millon o daltons para la coincidencia de los picos obtenidos frente a la masa del péptido teórico.

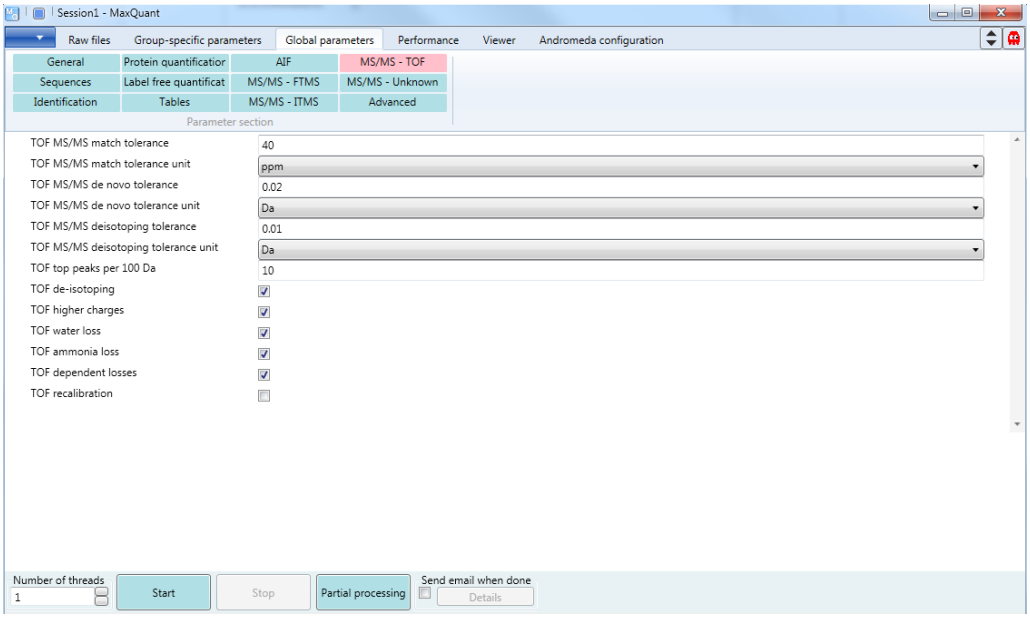


Figura 9: Selección de umbrales en desviaciones (ppm) para la coincidencia en masas de los péptidos identificados con las masas teóricas de los péptidos

Anexo 3: **Progenesis**

TUTORIAL BÁSICO DE ANÁLISIS EN PROGENESIS

1. Panel de *"Import Data"*

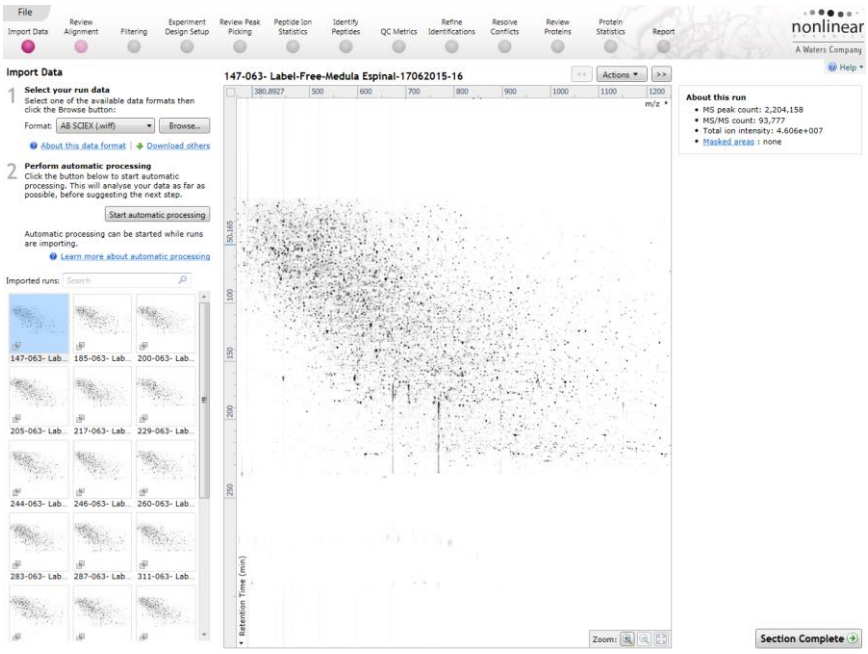


Figura 10: Panel de importe de archivos obtenidos de un espectrómetro de masas (".wiff" en el caso del espectrómetro de masas Triple TOF 5600)

En el panel presente en la figura 10 podemos ver como importar los archivos generados por el espectrómetro de masas. De cada uno de los archivos crudos ".wiff" tendremos un gráfico con los péptidos identificados (masa frente a tiempo de retención).

2. Panel de *"Alignment"*

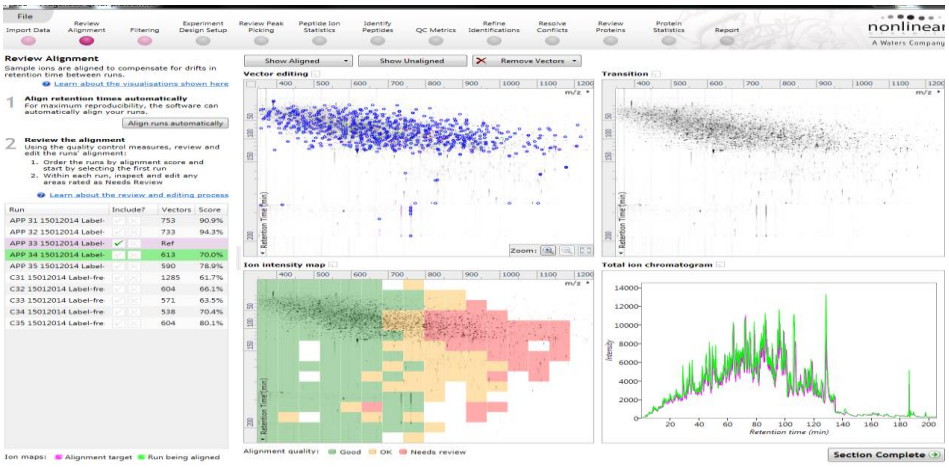


Figura 11: Panel para el ajuste y visualización de los alineamientos de los cromatogramas para cada run

En el panel de la figura 11, tenemos los porcentajes de coincidencia de péptidos (masa y tiempo) de cada uno de los archivos crudos si los comparamos con el archivo considerado como control para realizar el alineamiento. Podemos ver con colores (siendo rojo una mala alineación y verde una correcta) donde se han realizado las alineaciones de manera poco robusta. Estas zonas hay que valorarlas y puede ser motivo de exclusión del experimento de una de las muestras.

3. Panel de *"Filtering"*

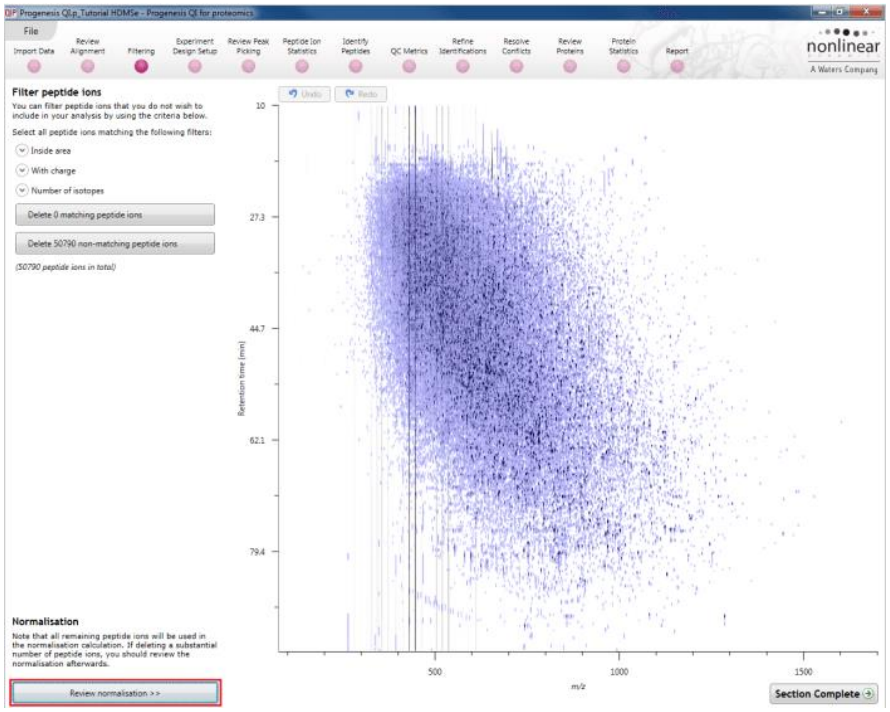


Figura 12: Filtrado a nivel de carga, masa y tiempo de retención

En el panel presente en la figura 12 se filtran de aquellos péptidos cuya carga es inferior a 2 (aquellos que tienen carga 1) o superior a 5 (cargas de 6 o mayores). De igual modo filtraremos las colas iniciales y finales hablando en términos de tiempos de retención donde veamos que las identificaciones no son las deseadas (contaminantes debidos a detergentes u otras causas presentes).

4. Panel de *"Experiment design"*

En este panel seleccionaremos que muestras pertenecen a cada condición experimental. Este al igual que otros pasos no tendrá mayor importancia para nosotros ya que analizaremos los datos posteriormente empleando R o Python.

5. Panel de *"Identify peptides"*

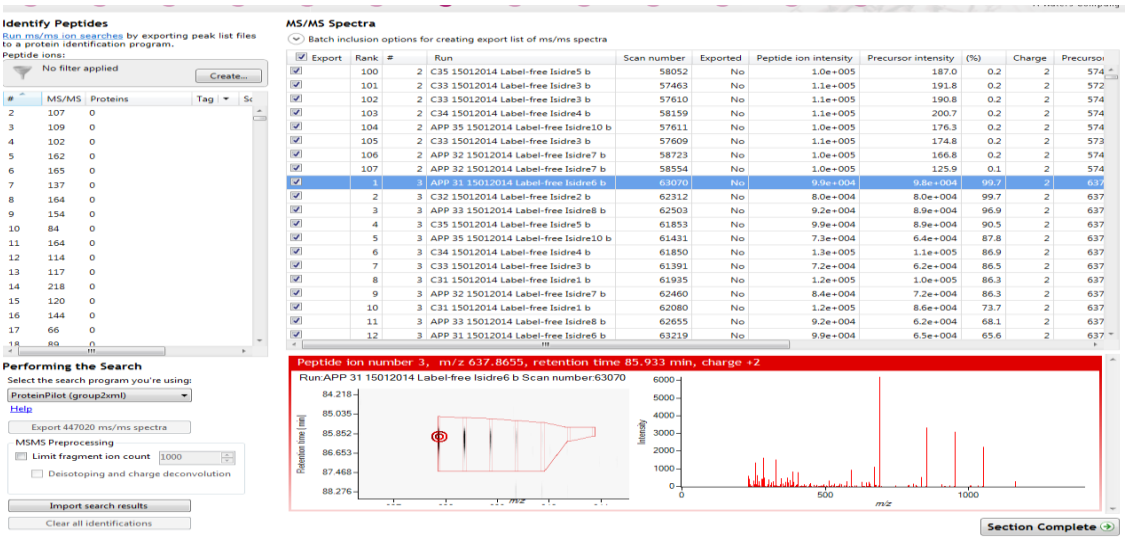


Figura 13: Panel para la importación de los datos de identificación generados mediante el uso de Protein Pilot (".group" en nuestro caso) u otros motores de búsqueda

Una vez generados los archivos ".group" con los datos de identificación, empleando la consola de comandos CMD para su realización en paralelo (manera más rápida y que permite la eliminación de ciertos archivos de identificación correspondientes a posibles muestras que se eliminen de análisis), se introducen en Progenesis y se añaden a los péptidos restantes en Progenesis (una vez estos han sido filtrados como explicamos en los paneles previos). En la configuración del motor de búsqueda se fija un valor de FDR del 1% (99% de confianza).

6. Panel de *"QC metrics"*

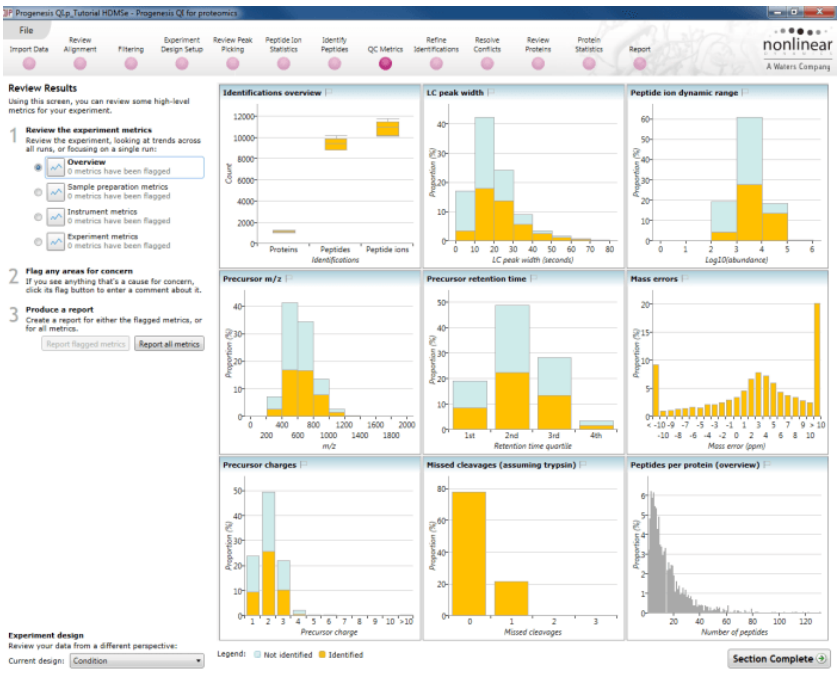


Figura 15: Gráficos de control de calidad donde se muestran las distribuciónes de las Intensidades entre otros datos de interes

Una vez en este panel, Progenesis presente algunos gráficos de calidad que nos pueden dar pistas sobre como continuar filtrando los datos o sobre si existe alguna irregularidad en los datos.

7. Panel de *"Refine identifications"*

Este panel consiste en una serie de filtros para filtrar los datos de entre ellos emplearemos los siguientes de esta manera:

- *"Score"*: Es el score a nivel de péptido dependiente del motor de búsqueda utilizado. En nuestro caso Protein Pilot (Paragon).
- *"Hits":* Eliminar los péptidos identificados en base a menos de 2 espectros.
- *"Mass error"* (ppm): Eliminaremos aquellos que no se encuentren entre +/-15 ppms.
- *"Sequence Length"*: Nos creemos las identificaciones de péptidos con 6 o más aminoácidos los péptidos con un menor número de aminoácidos quedan eliminados.
- *"Accession"*: Se eliminan aquellas identificaciones generadas por la base de datos decoy.

8. Panel de *"Review proteins"*

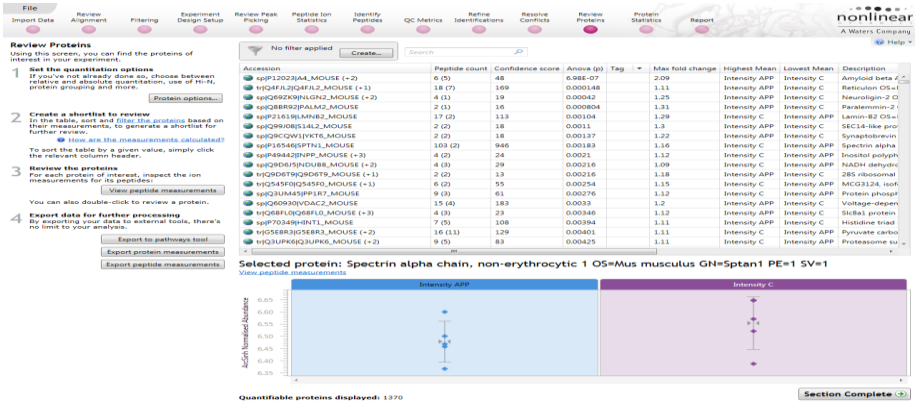


Figura 17: Panel de resultados a nivel de proteínas

En este panel exportaremos las tablas de péptidos y proteínas a partir de las cuales realizaremos el análisis en R o Python como se explica en la memoria.