Application of data mining and artificial intelligence techniques to mass spectrometry data for knowledge discovery

Hugo López Fernández

January 2016
UNIVERSITY OF VIGO

Department of Informatics

APPLICATION OF DATA MINING AND ARTIFICIAL INTELLIGENCE TECHNIQUES TO MASS SPECTROMETRY DATA FOR KNOWLEDGE DISCOVERY

THESIS FOR OBTAINING THE ACADEMIC DEGREE OF DOCTOR BY THE UNIVERSITY OF VIGO WITH INTERNATIONAL MENTION

Hugo López Fernández

Ourense, January 2016
To my family.
Acknowledgements

I would like to offer my most sincere thanks to all those who, directly or indirectly, contributed to this doctoral work, enabling it to come to fruition.

Firstly, I would like to thank my PhD supervisors Prof. Daniel Gonzalez Peña and Prof. Miguel Reboiro Jato for all their support and invaluable guidance throughout this work.

A very special thanks goes to Prof. Florentino Fernández Riverola for his support since he recruited me for the scientific research.

I would also like to thank BIOSCOPE group, particularly Prof. José Luis Capelo Martinez, for their remarkable contribution to this doctoral work.

A special thanks to my family, for all the encouragement given throughout my academic life. Thanks to my parents, without you all this would not have been possible. Thanks to my wife Luisa and our beloved Dogs, Tomy and Kata, for supporting me these years and giving me the necessary motivation.

Last, but definitely not least, I would like to thank Sabela for her support, understanding and encouragement.
I would like to thank University of Vigo and Xunta de Galicia for the predoctoral research grants.

I would like to thank the funded provided by (i) INOU-14-08 project from the Provincial Council of Ourense, (ii) TIN2009-14057-C03-02 project from the Spanish Ministry of Science and Innovation, the Plan E from the Spanish Government and the European Union from the ERDF, (iii) FP7/REGPOT-2012-2013.1 project from the European Union Seventh Framework Programme under grant agreement n° 316265, BIOCAPS, (iv) 15V1013 Contract-Programme from the University of Vigo, (v) 08VIB6 project research on Translational Bioinformatics from University of Vigo, (vi) Integrated action AIB2010PT-00353 from the Spanish Ministry of Science and Innovation, (vii) Agrupamento INBIOMED (2012/273) from DXPCTSUG-FEDER unha maneira de facer Europa, and (viii) ProteoMass Scientific Association (Caparica, Portugal).

Hugo López Fernández
Resumen

La espectrometría de masas empleando desorción/ionización láser asistida por matriz con detector de tiempo de vuelo (MALDI-TOF, matrix assisted laser desorption ionization coupled to time of flight analyzers) ha ganado popularidad durante la última década debido a su rapidez, sensibilidad y robustez para detectar péptidos y proteínas. Esta técnica de proteómica de alto rendimiento permite analizar rápidamente grandes conjuntos de muestras en una única tanda. En este escenario, las herramientas computacionales y los métodos bioinformáticos juegan un papel clave en el análisis de datos de MALDI-TOF, puesto que son capaces de manejar las grandes cantidades de datos en crudo generados para extraer nuevo conocimiento y conclusiones útiles.

El flujo típico de análisis de datos de MALDI-TOF tiene tres etapas principales: la adquisición de datos, el preprocesado y el análisis. Aunque el uso más popular de esta tecnología es la de identificar proteínas a través de sus péptidos, también se pueden llevar acabo otros análisis que hacen uso de inteligencia artificial (AI, Artificial Intelligence), aprendizaje automático (ML, Machine Learning) y métodos estadísticos, a fin de realizar identificación de biomarcadores, diagnóstico automático o descubrimiento de conocimiento.

En este trabajo de investigación se explora en profundidad este flujo de análisis y se proponen nuevas soluciones basadas en la aplicación de AI, ML y métodos estadísticos. Además, se ha desarrollado una plataforma software que da soporte al flujo completo de análisis de datos de MALDI-TOF y facilita el trabajo de los investigadores del ámbito de la proteómica que no poseen un alto grado de conocimiento de bioinformática.

Ámbito y motivación

La espectrometría de masas (MS, Mass Spectrometry) es una técnica utilizada para medir la relación masa-carga (m/z), a menudo llamada simplemente masa, de los componentes de una muestra [1]. Los instrumentos empleados se llaman espectrómetros de masas y constan de tres partes principales: la fuente de ionización, el analizador de masa y el detector. Mediante esta técnica, es posible medir de una manera rápida y precisa los tamaños y las abundancias relativas de las proteínas presentes en una mezcla.
biológica/química compleja. De manera general, los componentes de la muestra se pasan a través de estos tres componentes generando un espectro de masa (Figura 1.1). En primer lugar, los componentes de la muestra se ionizan en la fuente de ionización, de manera que sea posible medir sus masas. Hay dos tipos principales de fuentes [2]: las basadas en ionización por electrospray (ESI, Electrospray Ionization) o las basadas en ionización láser asistida por matriz (MALDI, Matrix Assisted Laser Desorption Ionization), que es la fuente de ionización predominante para espectrometría de masas simple [1]. A continuación, los componentes de la muestra se separan en el analizador de masa de acuerdo a la m/z de los iones. Hay cuatro tipos básicos de analizadores de masa utilizados actualmente en proteómica [3]: (i) trampa de iones (ion trap), (ii) tiempo de vuelo (TOF, time-of-flight), (iii) cuadrupolar (quadrupole) y (iv) resonancia ciclotrónica por transformada de Fourier (FT-MS, Fourier transform ion cyclotron resonance). Mientras que ESI se suele emplear con trampas de iones y analizadores triple cuadrupolo, MALDI suele emplearse con analizadores TOF y son los instrumentos más sencillos para el análisis de péptidos y proteínas. Tras la separación, los componentes chocan con el detector y sus valores de m/z con calculados por un ordenador conectado al espectrómetro que construye el espectro de masa. Un espectro de masa es una representación donde las masas medidas se sitúan en el eje horizontal y la intensidad de la señal de cada masa se sitúa en el eje vertical.

![Figura 1.1 Adquisición de datos utilizando un espectrómetro de masas.](image)

Este proceso genera datos en crudo, esto es, grandes conjuntos de espectros donde cada uno de ellos contiene cientos de mediciones de señales de m/z con sus respectivas sus intensidades. Los datos en crudo se caracterizan porque contienen señales que provienen tanto de los péptidos y proteínas presentes en la muestra como señales derivadas de diversas formas de ruido. Por este motivo, es necesario preprocesar dichos datos en crudo y convertirlos en una lista de picos (masas) limpia, eliminando los picos pertenecientes al ruido y dejando los picos reales. Las tareas más importantes de la etapa de preprocesado son la corrección de la línea base, el suavizado, la detección de picos, el emparejamiento de picos, la normalización de intensidades y el calibrado.
La espectrometría de masas empleando desorción/ionización láser asistida por matriz con detector de tiempo de vuelo (MALDI-TOF, *matrix assisted laser desorption ionization coupled to time of flight analyzers*) ha ganado popularidad durante la última década debido a su rapidez, sensibilidad y robustez para detectar péptidos y proteínas. Esta técnica de proteómica de alto rendimiento permite analizar rápidamente grandes conjuntos de muestras en una única tanda. En este escenario, los métodos bioinformáticos y las herramientas computacionales juegan un papel clave en el análisis de datos de MALDI-TOF, ya que son capaces de manejar las grandes cantidades de datos en crudo generados para extraer nuevo conocimiento y conclusiones útiles.

El flujo de trabajo típico en el análisis de datos de MALDI-TOF tiene tres etapas principales: (*i*) la adquisición de datos (Figura 1.1), (*ii*) el preprocesado y el (*iii*) análisis. En cuanto a la etapa de análisis, el uso más popular de esta tecnología es la de identificar proteínas a través de sus péptidos, un proceso conocido como *peptide-mass fingerprinting* (PMF). En este tipo de análisis, los espectros son preprocesados a fin de obtener una lista de masas experimentales de péptidos, la cual se empleará para buscar las proteínas asociadas en una base de datos. Sin embargo, también se pueden llevar a cabo análisis que hacen uso de inteligencia artificial (AI, *Artificial Intelligence*), aprendizaje automático (ML, *Machine Learning*) y métodos estadísticos, con el fin de realizar la identificación de biomarcadores, diagnóstico automático y descubrimiento de conocimiento [4-6], empleando para ello listas de picos.

Tal y como se explicó anteriormente, cada espectro preprocesado o lista de picos contiene un número finito y discreto de picos. Un análisis de búsqueda de biomarcadores se puede llevar a cabo empleando métodos estadísticos especialmente adaptados, los cuales conducen a la identificación de los picos que están asociados con los factores de interés [7].

El diagnóstico automático o clasificación dado un conjunto de muestras previamente clasificadas es un problema de ML supervisado [8]. Por ejemplo, dada una muestra de suero de un paciente sin clasificar, la cual puede venir de una o más réplicas (espectros), el objetivo de la clasificación sería el de asignarlo a un determinado grupo de diagnóstico (por ejemplo, sano o enfermo). En este caso, el modelo de clasificación se construye a partir de un conjunto de muestras clasificadas empleando la intensidad o la presencia/ausencia de los diferentes picos (m/z) como características [4]. Es importante señalar que cuando se emplean las intensidades de los picos, el conjunto de datos debe
ser normalizado para hacer las intensidades comparables. Los datos utilizados para construir este modelo constituyen el conjunto de entrenamiento. Una vez entrenado, el modelo se emplea para predecir la clase de muestras sin clasificar. Técnicas habituales de ML supervisado son [4], entre otras: (i) clasificadores Bayesinos, como el algoritmo Naïve Bayes, basados en el teorema de Bayes; (ii) aprendizaje basado en reglas, basados en la creación de reglas legibles para los humanos que puedan explicar por qué ciertas muestras pertenecen a una determinada clase; (iii) árboles de decisión, basados en estructuras en forma de árbol que estructuran el conocimiento para discriminar entre muestras y predecir así su clase; (iv) selvas aleatorias, basadas en el uso de varios árboles de decisión para predecir la clase de una muestra; (v) máquinas de soporte vectorial (SVMs, support vector machines), como el algoritmo Sequential Minimal Optimization (SMO), basadas en el concepto de separabilidad lineal entre clases; y (vi) redes de neuronas artificiales (ANNs, artificial neural networks), basadas en la simulación del funcionamiento del cerebro humano para construir un modelo capaz de predecir la clase de cada muestra. Cuando se aplican técnicas de aprendizaje basado en reglas o árboles de decisión, también es posible considerar los picos empleados para separar las clases como biomarcadores para las clases definidas. A pesar del hecho de que estos algoritmos toman listas de picos como entrada, estas pueden contener todavía picos irrelevantes o redundantes que pueden reducir la precisión de los clasificadores. Para aliviar estos síntomas, es posible aplicar selección de características para generar un conjunto de datos más limpio sobre el que aplicar los algoritmos de clasificación. Los métodos de selección de características también pueden ayudar al descubrimiento de potenciales biomarcadores.

En contraste con el ML supervisado, en la clasificación no supervisada o agrupamiento (clustering) las muestras de las que se dispone no tienen una clasificación asociada y el objetivo de estas consiste en agrupar las muestras con perfiles de picos similares. Los principales enfoques son [8]: (i) particional (por ejemplo, el algoritmo K-means), (ii) jerárquico y (iii) modelos mixtos. Estos enfoques se caracterizan por el hecho de que realizan un agrupamiento en una dimensión, empleando los atributos de las muestras. Un subtipo de agrupamiento conocido como agrupamiento doble (biclustering), lleva a cabo un agrupamiento en dos dimensiones, es decir, los grupos se crean en base las muestras y a los atributos de las muestras. Las técnicas no supervisadas llevan a la creación de nuevas hipótesis (por ejemplo, los grupos propuestos) que deben ser explorados y evaluados a posteriori.
Esta tesis explora el flujo de análisis de datos MALDI-TOF presentado, incluyendo una etapa adicional de control de calidad entre las fases de preprocesado y de análisis (Figura 1.2).

Objetivos y metodología

El objetivo principal de esta tesis fue la aplicación de técnicas de minería de datos y AI al análisis de datos de espectrometría de masas para el descubrimiento de nuevo conocimiento. De cara a alcanzar este objetivo principal y considerando el escenario presentado anteriormente, el trabajo de investigación se estructuró en torno a los siguientes objetivos:

O1. Estudiar los métodos de preprocesado de datos de MALDI-TOF disponibles.
O2. Analizar cómo la minería de datos y la AI pueden ayudar a mejorar técnicas existentes de preprocesado de datos de MALDI-TOF.
O3. Estudiar el problema de la aplicación de minería de datos y AI a datos de MALDI-TOF preprocesados.
O4. Analizar cómo la minería de datos y la AI pueden ayudar a: (i) descubrir potenciales biomarcadores, (ii) generar nuevas hipótesis y (iii) clasificar o diferenciar muestras.

O5. Estudiar y analizar cómo la minería de datos y la AI pueden ayudar a visualizar y representar datos de MALDI-TOF.

O6. Desarrollar una plataforma integrada de software que soporte el flujo de análisis de datos de MALDI-TOF presentado en la Figura 1.2.

O7. Facilitar el trabajo de los investigadores en proteómica que no tienen conocimientos avanzados de bioinformática, proveyéndoles de un software que les permita manejar y analizar datos de MALDI-TOF de un modo sencillo e intuitivo.

En relación a las metodologías utilizadas durante la realización del trabajo de investigación, todas las líneas de estudio siguen una estructura similar con las siguientes fases principales: (i) estudio pormenorizado del problema en cuestión y estado del arte, (ii) definición de objetivos específicos, (iii) desarrollo de los componentes de software necesarios, empleando un proceso iterativo, y (iv) validación y prueba del progreso realizado.

**Evolución de la investigación**

El trabajo doctoral aquí presentado tiene su origen en dos aplicaciones software presentadas en el año 2011 [9]: *Decision Peptide-Driven*, una aplicación para cuantificar proteínas de un modo rápido y preciso, y *Bacterial Identification*, una aplicación para la búsqueda de biomarcadores y diagnóstico de tuberculosis. Estas aplicaciones, desarrolladas utilizando el framework AIBench [10] en colaboración con el grupo BIOSCOPE (*Universidade NOVA de Lisboa*, Caparica, Portugal), fueron importantes, ya que permitieron una primera toma de contacto con datos de MALDI-TOF y establecieron un punto de partida para futuros desarrollos y colaboraciones.

La colaboración con el grupo BIOSCOPE llevó a la participación en algunos estudios sobre expresión diferencial (*profiling*) [11-13], con el objetivo de obtener perfiles diferenciales de las muestras que puedan ser utilizados para diferenciar entre grupos de éstas (por ejemplo, sanos frente a enfermos). Estos estudios también fueron importantes porque proveyeron un conjunto de datos de doce muestras que fue empleado para alcanzar los objetivos de esta tesis.
En 2012, el trabajo doctoral comenzó en dos de las líneas de trabajo definidas. Por una parte, el desarrollo de MLibrary [14], una base de datos con un motor de búsqueda diseñados para asistir al usuario en la detección e identificación de anabolizantes androgénicos esteroideos (AAS, androgenic anabolic steroids) y sus metabolitos mediante espectrometría de masas MALDI-TOF. Esta aplicación permite a los investigadores manejar repositorios de biomarcadores que pueden ser utilizados para detectar e identificar la presencia de AAS en muestras MALDI-TOF. El desarrollo de MLibrary fue muy importante para este trabajo doctoral, ya que permitió alcanzar un mejor conocimiento de los datos de MALDI-TOF preprocesados y de otros problemas comunes, como los formatos de almacenamiento o el emparejamiento de picos, entre otros.

Por otra parte, el estudio y comparación de los métodos disponibles de preprocesado de datos en crudo de MALDI-TOF. Mientras que MLibrary se situaría al final del flujo de análisis de datos explorado, este estudio se centra en el paso situado inmediatamente después de la adquisición de datos. Los estudios y colaboraciones llevados a cabo hasta este momento se realizaron sobre listas de picos, es decir, datos preprocesados. Sin embargo, surgió la necesidad de manejar datos en crudo y se establecieron dos objetivos: (i) crear una plataforma para automatizar la carga y el preprocesado de los datos en crudo y (ii) utilizar dicha plataforma para evaluar distintos métodos de preprocesado. Después de comprender los distintos formatos empleados para el almacenamiento de los datos en crudo, se evaluó la influencia de distintos métodos de preprocesado en el rendimiento de una tarea de clasificación de muestras [15]. Con este estudio se alcanzaron dos objetivos importantes. En primer lugar, un mejor conocimiento del flujo completo de análisis de datos de MALDI-TOF. En segundo lugar, se desarrollaron e integraron en el framework AIBench distintos componentes software para cargar, visualizar y preprocesar datos.

En 2013, se propuso un novedoso proceso para la aplicación de biclustering a datos de MALDI-TOF [16]. Este estudio profundiza en el área de descubrimiento de información, ya que evalúa la viabilidad de la aplicación de biclustering para analizar datos de MALDI-TOF, comparando biclustering y agrupamiento jerárquico sobre dos conjuntos de datos reales. Los resultados fueron prometedores, ya que revelaron la habilidad de este tipo de técnicas para extraer información útil y generar nuevas hipótesis.

Hasta este momento se habían desarrollado varios componentes software específicos e independientes, los cuales dieron soporte a las necesidades de estudios
previos, tales como la carga de datos en crudo, el preprocesado, la aplicación de algoritmos de ML y diferentes técnicas de visualización, entre otras. A pesar del hecho de que son independientes, también están muy relacionadas y tienen una base común: están desarrollados en Java e integrados con el framework AIBench. En este punto, se unificaron todos esos componentes en una única plataforma, dando lugar al desarrollo de Mass-Up [17, 18], una aplicación multiplataforma de código libre para el descubrimiento de nuevo conocimiento sobre datos de espectrometría de masas MALDI-TOF. Mass-Up permite visualización de espectros, carga y preprocesado de datos en crudo y distintos tipos de análisis, incluyendo (i) búsqueda de biomarcadores, (ii) agrupamiento, (iii) biclustering, (iv) visitación basada en el análisis de componentes principales (PCA, Principal Component Analysis) y (v) clasificación de grandes conjuntos de muestras.

Finalmente, gracias a los trabajos presentados aquí, especialmente a Mass-Up, se obtuvo un contrato programa con la Universidad de Vigo titulado Desarrollo de técnicas y herramientas para tratamiento de datos ómicos. Este contrato programa, que contó con una financiación de 7500€, permitió profundizar en esta línea de trabajo, así como en otras nuevas. La experiencia adquirida en el análisis de datos de MALDI-TOF abre nuevas líneas de investigación en espectrometría de masas. En este sentido, se acaba de iniciar una colaboración en el área de bioimagen por espectrometría de masas empleando ablación laser con fuente de plasma de acoplamiento inductivo (LA-ICP MS, Laser Ablation and Inductively Coupled Plasma Mass Spectrometry).

El progreso la investigación y del trabajo doctoral presentados en esta sección se ilustran en la Figura 1.3, donde se resaltan los hitos más importantes.
Estructura del trabajo

Mientras que el capítulo actual motiva el interés que suscita el problema objeto de estudio, establece los objetivos fundamentales del trabajo a realizar y expone el avance en la investigación realizada, los Capítulos 2, 3 y 4 presentan los resultados más importantes derivados de este trabajo doctoral.

El Capítulo 2 presenta MLibrary, el proyecto que representa el inicio de la tesis. En este trabajo, se desarrollaron una base de datos con un motor de búsqueda para asistir al usuario en la detección e identificación de AAS y sus metabolitos mediante espectrometría de masas MALDI-TOF. La búsqueda de agentes anabólicos en la orina juega un papel muy importante en los laboratorios anti dopaje puesto que se trata de la droga más empleada en el mundo del deporte. MLibrary facilita el uso de la espectrometría de masas MALDI-TOF para realizar controles anti dopaje y reduce el tiempo necesario para la evaluación e interpretación de los resultados. En pocas palabras, la detección de AAS en las muestras se puede realizar comparando un espectro de masa contra la librería desarrollada, a fin de identificar los posibles positivos y comparando un espectro de masa/masa (MS/MS) producido después de la fragmentación de los posibles positivos contra un conjunto de espectros completo previamente establecido en MLibrary. La aplicabilidad de MLibrary se evalúa mediante el análisis de cinco muestras de orina.
marcadas, siendo la aplicación desarrollada capaz de identificar con éxito todos los componentes marcados. Además, el motor de búsqueda es, potencialmente, extensible para el análisis de otros componentes distintos a los AASs.

El Capítulo 3 describe el estudio sobre la influencia de los métodos de preprocesado en el descubrimiento de información, centrándose en estudiar el impacto en problemas de clasificación. Como se explicó anteriormente, existen distintos métodos para llevar a cabo las tareas de corrección de la línea base, suavizado, detección de picos, emparejamiento de picos, normalización de intensidades y calibrado. En este trabajo se lleva a cabo una comparación sistemática de diferentes paquetes software para llevar a cabo el preprocesado de datos de MALDI-TOF. Para garantizar la validez del estudio, se testean múltiples configuraciones de cada técnica de preprocesado, cuyas listas de picos resultantes se emplean para entrenar un conjunto de clasificadores. El rendimiento de estos clasificadores, medido empleando la precisión y el coeficiente kappa, proporciona información precisa para la comparación final. Los resultados mostraron el impacto real de cada técnica de preprocesado y de cada configuración en la clasificación, mostrando que MassSpecWavelet obtiene el mejor rendimiento y que las máquinas de soporte vectorial son uno de los clasificadores más precisos.

Finalmente, el Capítulo 4 presenta Mass-Up, una aplicación multiplataforma de código libre para el descubrimiento de nuevo conocimiento sobre datos de espectrometría de masas MALDI-TOF que cubre el flujo de análisis completo. Mass-Up permite a los investigadores cargar y visualizar tanto datos en crudo como datos preprocesados, preprocesar estos datos y realizar distintos tipos de análisis, tales como (i) búsqueda de biomarcadores, (ii) agrupamiento, (iii) biclustering, (iv) visualización basada en PCA y (v) clasificación de grandes conjuntos de muestras. Aunque existen varias librerías software y herramientas que pueden ser combinadas para llevar a cabo todas estas tareas, todavía existía la necesidad de soluciones que diesen un soporte completo y que incluyesen una interfaz gráfica amigable, evitando que los usuarios tuviesen que poseer conocimientos informáticos avanzados y de programación para poder analizar sus datos.

Los Capítulos 2, 3 y 4 que forman parte del presente documento de tesis, también pueden ser consultados en el formato original de las revistas donde fueron publicados (Apéndices A, B y C, respectivamente).
Conclusiones y trabajo futuro

El objetivo principal de esta tesis fue la aplicación de técnicas de minería de datos y AI para el descubrimiento de nuevo conocimiento con datos de MALDI-TOF.

En esta tesis, distintos métodos de preprocesado de datos MALDI-TOF fueron estudiados y comparados (O1). Además, se desarrolló un algoritmo de emparejamiento de picos llamado Forward (O2), el cual fue utilizado en casi todos los desarrollos y colaboraciones. El trabajo futuro en esta línea incluye la comparación de más librerías disponibles públicamente así como la inclusión de más conjuntos de datos.

Durante el curso de la investigación, la técnica de agrupamiento doble o biclustering se aplicó para en análisis de datos de MALDI-TOF, siendo capaz de extraer información útil y generar nuevas hipótesis (O3, O4). Su adecuación fue evaluada comparándola contra el agrupamiento jerárquico empleando dos conjuntos de datos reales. Aunque los resultados fueron prometedores, se debe continuar trabajando en esta línea en el futuro para profundizar y expandir este estudio.

Integrando todos los desarrollos e investigaciones llevadas a cabo para los estudios presentados en la sección Evolución de la investigación, se desarrolló Mass-Up (O5, O6, O7). Con Mass-Up, se pone a disposición de la comunidad científica una herramienta de código libre que da un soporte completo al flujo de análisis de datos MALDI-TOF incluyendo, además, una interfaz gráfica intuitiva que permite su empleo por parte de usuarios no expertos en bioinformática y programación. Su utilizad está siendo refrendada por el aumento del número de estudios que hacen uso de él [19-21] y por el hecho de que ha sido incluido en repositorios públicos de software de espectrometría de masas y en proyectos mayores, como, por ejemplo, MASSyPup(64), una distribución de Linux que incluye diferentes herramientas para el análisis de datos de espectrometría de masas.

En cuanto al trabajo futuro, esta tesis tiene dos líneas principales de continuación. Por una parte, continuar desarrollando y mejorando Mass-Up. Aunque Mass-Up ha sido actualizado continuadamente para solucionar fallos reportados por los usuarios, se han identificado algunas mejoras importantes: (i) soportar más formatos de almacenamiento de datos de MALDI-TOF, (ii) incluir más algoritmos de preprocesado y hacerlos más configurables e (iii) incluir nuevos tipos de análisis. Como se explicó previamente, Mass-Up integra diferentes componentes software que tienen una base común y, además, ha evolucionado y crecido incrementalmente con el paso de los años. Desde un punto de
vista de ingeniería de software, también se han identificado ciertas mejoras y se planea su implementación mediante una gran refactorización de código, aplicando todas las lecciones aprendidas desde el comienzo del proyecto.

Por otra parte, se acaba de iniciar una colaboración en el área de bioimagen por espectrometría de masas empleando ablación láser con fuente de plasma de acoplamiento inductivo (LA-ICP MS). Los objetivos de esta colaboración consisten en proporcionar una base analítica para emplear la técnica LA-ICP MS y en el desarrollo de una herramienta para automatizar el proceso.

Referencias


## Contents

1 INTRODUCTION .................................................................................................................. 1

1.1 ABSTRACT ......................................................................................................................... 1

1.2 SCOPE AND MOTIVATION ................................................................................................. 1

1.3 OBJECTIVES AND METHODOLOGY .................................................................................. 5

1.4 RESEARCH PROGRESS ...................................................................................................... 6

1.5 STRUCTURE OF THE WORK* ............................................................................................ 8

1.6 CONCLUSION AND FUTURE WORK ................................................................................ 10

1.7 REFERENCES ..................................................................................................................... 11

2 SPEEDING UP THE SCREENING OF STEROIDS IN URINE: DEVELOPMENT OF A USER-FRIENDLY LIBRARY ............................................................................................................. 13

2.1 ABSTRACT ......................................................................................................................... 13

2.2 INTRODUCTION ................................................................................................................ 14

2.3 MATERIALS AND METHODS ........................................................................................... 15

2.3.1 Chemicals ...................................................................................................................... 15

2.3.2 Apparatus ....................................................................................................................... 16

2.3.3 Sample preparation .......................................................................................................... 16

2.3.4 Experimental design ....................................................................................................... 17

2.3.5 MALDI-TOF–MS and MALDI-TOF–TOF–MS analysis ................................................... 18

2.3.6 MLibrary software ........................................................................................................... 19

2.4 RESULTS AND DISCUSSION ............................................................................................ 19

2.4.1 MS mode ......................................................................................................................... 20

2.4.2 MS/MS mode .................................................................................................................. 21

2.5 CASE STUDY ..................................................................................................................... 23

2.6 CONCLUSIONS .................................................................................................................. 25

2.7 SUPPLEMENTARY MATERIAL .......................................................................................... 26

2.8 APPENDIX A ..................................................................................................................... 32

2.9 REFERENCES ..................................................................................................................... 39
1. Introduction

3 A COMPREHENSIVE ANALYSIS ABOUT THE INFLUENCE OF LOW-LEVEL PREPROCESSING TECHNIQUES ON MASS SPECTROMETRY DATA FOR SAMPLE CLASSIFICATION

3.1 Abstract .................................................................................................................................................................43

3.2 Introduction .............................................................................................................................................................43

3.3 Related Work on Mass Spectrometry ..........................................................................................................................44

3.4 Available Preprocessing Methods for MS Data ...........................................................................................................46

3.4.1 Baseline correction ..................................................................................................................................................46

3.4.2 Smoothing ............................................................................................................................................................47

3.4.3 Peak detection ........................................................................................................................................................47

3.4.4 Peak alignment .......................................................................................................................................................48

3.5 Experimental Setup ....................................................................................................................................................50

3.5.1 Mass Spectrometry datasets ................................................................................................................................50

3.5.2 Machine learning algorithms ................................................................................................................................51

3.5.3 Preprocessing configuration ..................................................................................................................................52

3.6 Results and Discussion ................................................................................................................................................55

3.7 Conclusions ..............................................................................................................................................................61

3.8 References ...............................................................................................................................................................61

4 MASS-UP: AN ALL-IN-ONE OPEN SOFTWARE APPLICATION FOR MALDI-TOF MASS SPECTROMETRY KNOWLEDGE DISCOVERY

4.1 Abstract .................................................................................................................................................................65

4.2 Background .............................................................................................................................................................65

4.3 Implementation ..........................................................................................................................................................67

4.3.1 Mass-Up workflow ................................................................................................................................................68

4.3.2 Third-party libraries ..............................................................................................................................................68

4.4 Results and Discussion ..............................................................................................................................................69

4.4.1 Sample Datasets ...................................................................................................................................................70

4.4.2 Preprocessing .......................................................................................................................................................70

4.4.3 Quality Control ....................................................................................................................................................73

4.4.4 Biomarker discovery ..............................................................................................................................................75

4.4.5 Principal component analysis ................................................................................................................................77
4.4.6  Cluster analysis ................................................................. 78
4.4.7  Bicluster analysis .............................................................. 80
4.4.8  Classification analysis ....................................................... 81
4.4.9  Performance notes ............................................................ 82
4.5  CONCLUSIONS ................................................................. 83
4.6  AVAILABILITY AND REQUIREMENTS .................................. 83
4.7  REFERENCES ................................................................. 84

A.  STEROIDS 78/12, PP. 1226-1232; DOI:10.1016/J.STEROIDS.2013.08.14 ................. 89
Introduction

1.1 Abstract

Mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analyzers (MALDI-TOF MS) has become popular during the last decade due to its high speed, sensitivity and robustness for detecting proteins and peptides. This allows quickly analyzing large sets of samples are in one single batch and doing high-throughput proteomics. In this scenario, bioinformatics methods and computational tools play a key role in MALDI-TOF data analysis, as they are able handle the large amounts of raw data generated in order to extract new knowledge and useful conclusions.

A typical MALDI-TOF MS data analysis workflow has three main stages: data acquisition, preprocessing and analysis. Although the most popular use of this technology is to identify proteins through their peptides, analyses that make use of artificial intelligence (AI), machine learning (ML), and statistical methods can be also carried out in order to perform biomarker discovery, automatic diagnosis, and knowledge discovery.

In this research work, this workflow is deeply explored and new solutions based on the application of AI, ML, and statistical methods are proposed. In addition, an integrated software platform that supports the full MALDI-TOF MS data analysis workflow that facilitate the work of proteomics researchers without advanced bioinformatics skills has been developed and released to the scientific community.

1.2 Scope and Motivation

Mass spectrometry (MS) is a technique used for measuring the mass-to-charge ratio (m/z), often simply called mass, of the components in a sample [1]. The instruments employed are called mass spectrometers and consists of three main parts: the ionization source, the
mass analyzer, and the detector. This technique provides rapid and precise measurements of the sizes and relative abundances of the proteins present in a complex biological/chemical mixture. In short, the components of the sample are passed through these three components generating a mass spectrum (Figure 1.1). First, the components of the sample are ionized in the ionization source, in order to be able to measure their masses. There are two main types of sources [2]: based on electrospray ionization (ESI) or based on matrix assisted laser desorption ionization (MALDI), which is the dominating ionization source for (single) mass spectrometry [1]. Then, in the mass analyzer, the components of the sample are separated according to the m/z of the ions. There are four basic types of mass analyzers currently used in proteomics research [3]: (i) ion trap, (ii) time-of-flight (TOF), (iii) quadrupole, and (iv) Fourier transform ion cyclotron resonance (FT-MS). While ESI are usually coupled to ion traps and triple quadrupole analyzers, MALDI are usually coupled to TOF analyzers and are the simplest instruments for protein and peptide analysis. After separation, the components hit the detector and their m/z values can be calculated by a connected computer that constructs a mass spectrum. A mass spectrum is a representation where the measured masses are placed along the horizontal axis and the intensity of the signal for each component (m/z) along the vertical axis.

This process generates raw data: large spectra sets where each individual spectrum contains thousands of measurements involving m/z signals and intensity (i.e., {m/z, intensity} pairs). Since raw data contains signals coming from the peptides or proteins and signals derived from several forms of noise, it is necessary to preprocess this raw data and convert it into a clean peak list by removing only noisy peaks and not real peaks. The most important preprocessing tasks are baseline correction, smoothing, peak detection, peak matching, intensity normalization and calibration.

Mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analyzers (MALDI-TOF MS, hereinafter called MALDI) has become popular during the last decade due to its high speed, sensitivity and robustness for detecting
proteins and peptides. This allows quickly analyzing large sets of samples are in one single batch and doing high-throughput proteomics. In this scenario, bioinformatics methods and computational tools play a key role in MALDI data analysis, as they are able to preprocess raw data registered in different formats, compare them, and apply complex algorithms in order to finally extract new knowledge and useful conclusions.

A typical MALDI data analysis workflow has three main stages: (i) data acquisition (Figure 1.1), (ii) preprocessing, and (iii) analysis. Regarding the analysis stage, the most popular use of MALDI is to identify proteins through their peptides, a process known as peptide-mass fingerprinting (PMF). In this cases, the mass spectrum must be preprocessed in order to obtain a list of peptide experimental masses, which can be searched against a database to identify proteins. Nevertheless, analyses that make use of artificial intelligence (AI), machine learning (ML), and statistical methods are also carried out in order to perform biomarker discovery, automatic diagnosis and knowledge discovery [4–6] taking peak lists as input.

As explained before, each preprocessed spectrum or peak list contains a finite number of peaks. A biomarker analysis can be done by using adapted statistical methods that led to identify which of those peaks are associated with factors of interest [7].

Automatic diagnosis or classification given a set of previously classified samples is a supervised ML problem [8]. For example, given an unlabeled serum sample from an individual, which can come from one or more replicates (i.e., spectra), the purpose of classification could be to assign it to a specific diagnostic group (e.g., healthy or diseased). In this case, a classification model is built from a set of labeled samples using the intensity or the presence/absence of the different peaks (m/z) as features [4]. It is important to note that when intensity values are used, the dataset must be normalized in order to make intensities comparable. Data used to build this model is called training data. The model is then used to predict the class of unlabeled samples. Common types of ML supervised techniques are [4], among others: (i) Bayesian classifiers, such as Naïve Bayes, which are based on Bayes theorem; (ii) rule-based learners, which are based on the creation of human-readable rules that could explain why certain samples belong to a class; (iii) decision trees, which are based on tree-like structures that structure the knowledge to discriminate between samples and predict their class; (iv) random forests, which use several decision trees to predict the class of each sample; (v) support vector machines (SVMs), such as Sequential Minimal Optimization (SMO), which are based in the concept of linear separability between classes; and (vi) artificial neural networks (ANNs), which
simulate brain’s operation in order to build the model and predict the class of each sample. In algorithms such as rule-based learners or decision trees, it is also possible to consider the specific peaks used to separate the classes as biomarkers for the defined classes. Despite the fact that this algorithms take peak lists as input, they can still contain noisy, irrelevant or redundant peaks, which can reduce the accuracy of the classifiers. To mitigate this symptoms, feature selection can be applied prior to the classification algorithms, generating a cleaner dataset on which apply them. Feature selection methods can also be used to discover potential biomarkers.

In contrast with supervised machine learning, in unsupervised classification or clustering, samples do not have associated class labels and they consist in grouping together samples with similar peak profiles. The main clustering approaches are [8]: (i) partition clustering (e.g., K-means algorithm), (ii) hierarchical clustering, and (iii) mixture models. This techniques are characterized by the fact that they perform a one dimensional clustering using samples’ attributes. A sub-type of clustering called biclustering or co-clustering performs a two dimensional clustering, that is, clusters are modeled with both samples and samples’ attributes. These unsupervised techniques lead to the creation of new hypotheses (e.g., proposed groups) that must be further explored and evaluated.

This thesis work will explore the MALDI data analysis workflow presented in this section, including an additional quality control step between preprocessing and analysis (Figure 1.2).
1.3 Objectives and Methodology

The main objective of this thesis was the application of data mining and AI techniques to the analysis of mass spectrometry data as a way of knowledge discovery. In order to attain this main objective and considering the scenario presented in the previous section, the research work was structured around the following topics:

- Study the available preprocessing methods for MALDI data.
- Analyze how data mining and AI can help to improve existing preprocessing methods for MALDI data.
- Study the problem of applying data mining and AI techniques to preprocessed MALDI data.
- Analyze how data mining and AI can help to: (i) discover potential biomarkers in MALDI data, (ii) generate knowledge in form of new hypotheses, and (iii) classify or differentiate samples.
- Study and analyze how data mining and AI can help to visualize and represent MALDI data.
1. Introduction

- Develop an integrated software platform that supports the full MALDI data analysis workflow depicted in Figure 1.2.
- Facilitate the work of proteomics researchers without advanced bioinformatics skills by providing them a software to easily handle and analyze MALDI data.

Regarding the methodologies used during the work carried out, all lines of study follow a similar structure with the following main phases: (i) in-depth study of the problem and state of the art, (ii) definition of specific objectives, (iii) development of software components needed, using an iterative process, and (iv) validation and test of the progress made.

1.4 Research Progress

The doctoral work here presented had its genesis in two software applications presented in 2011 [9]: Decision Peptide-Driven, an application for rapid and accurate protein quantification, and Bacterial Identification, an application for tuberculosis biomarker search and diagnosis. These applications, developed using the AIBench framework [10] in collaboration with the BIOSCOPE group (Universidade NOVA de Lisboa, Caparica, Portugal), were very important since they allowed a first contact with MALDI data and established a starting point for future developments and collaborations.

The collaboration with the BIOSCOPE group lead to the participation in a few profiling studies [11–13], with the objective of obtaining sample profiles that can be used to distinguish between different groups of samples (e.g., healthy vs. diseased). These studies were also important since they provided a dataset of twelve samples which were used to address thesis objectives.

In 2012, the doctoral work for this thesis started with two of the defined lines of work. By one hand, the development of MLibrary [14], a novel database search engine designed to assist the user in the detection and identification of androgenic anabolic steroids (AAS) and its metabolites by MALDI-TOF MS. This application allows researchers to manage repositories of biomarkers that can be then used to detect and identify the presence of AAS in MALDI samples. The development of MLibrary was very important for this thesis work since it allowed gaining a better understanding of MALDI preprocessed data and common problems, such as data formats or peak matching, among others.
On the other hand, the study and comparison of the available preprocessing methods for raw MALDI data. While MLibrary stands at the end of the explored workflow, this study is focused immediately after the MALDI data acquisition. Studies and collaborations carried out until this moment worked with peak lists (i.e., preprocessed data). However, the need to deal with raw data arisen and the two objectives were established: create a platform to automate the loading and preprocessing of raw data and use this platform to evaluate different preprocessing methods. After understanding the different formats used to store raw data, the influence of the available preprocessing methods on the performance of sample classification was evaluated [15]. With this study, two important objectives were achieved. First, a better understanding of the MALDI workflow, especially regarding how raw data is stored and preprocessed to obtain peak lists. Second, software components to load, visualize, and preprocess MALDI data were created and integrated into the AIBench framework.

In 2013, a novel workflow to the application of biclustering to MALDI-TOF mass spectrometry data [16] was proposed. This study deepens in the knowledge discovery area, since it evaluates the adequacy of applying biclustering to analyze MALDI data by comparing between biclustering and hierarchical clustering over two real datasets. Results were promising since they revealed the ability of these techniques to extract useful information and generate new hypotheses.

At this moment, there has been developed specific, independent software components that allowed meeting the needs arisen in previous studies such as raw data loading, preprocessing, application of ML algorithms, and different data visualization techniques, among others. Despite the fact that they are independent, they are also closely related and have a common base: they are developed in Java and integrated with the AIBench framework. In this scenario, all these components were put together in order to develop Mass-Up [17, 18], an open software multiplatform application for MALDI-TOF MS knowledge discovery that allows spectra visualization, raw data preprocessing and different kinds of analyses, including (i) biomarker discovery, (ii) clustering, (iii) biclustering, (iv) three-dimensional PCA visualization, and (v) classification of large sets of spectra data.

Finally, a contract program with University of Vigo has been obtained thank to the works presented here, especially Mass-Up. This contract, entitled Development of techniques and tools for omics data treatment, has a funding of 7500€ and allowed the
continuation of this line of research, as well as new collaborations. The experience acquired in MALDI data analysis opens new research lines in computational mass spectrometry. In this sense, a collaboration in the area of laser ablation and inductively coupled plasma mass spectrometry (LA-ICP MS) bioimaging has just started.

The research progress and the doctoral work done presented in this section are illustrated by Figure 1.3, where the most significant milestones are highlighted.

![Figure 1.3 Research progress timeline.](image)

1.5 **Structure of the Work**

While this chapter motivates the interest of the problem, establishes the main objectives of the thesis and discusses the progress of the research work carried out, Chapters 2 to 4 present the most significant results of this doctoral work.

Chapter 2 presents MLibrary, the project that represents the thesis start. In this work, a database search engine to assist users in the detection and identification of androgenic anabolic steroids (AAS) and its metabolites by MALDI was developed. The urinary screening for anabolic agents plays a major role in anti-doping laboratories as they represent the most abused drug class in sports. MLibrary eases the use of MALDI techniques for doping control and reduces the time needed for the evaluation and
interpretation of the results. In short, the detection of the AAS in the samples can be done by searching the MS spectra against the library developed in order to identify possible positives and by comparison of the tandem mass spectrometry (MS/MS) spectra produced after fragmentation of the possible positives with a complete set of spectra that have previously been assigned to the software. The applicability of the MLiBrary tool is assessed through the analysis of spiked urine samples, being able to successfully identify all spiked compounds. Moreover, the search engine can, potentially, be further expanded to other compounds in addition to AASs.

Chapter 3 describes the study about the influence of MALDI preprocessing techniques on knowledge discovery, focused on studying their impact in classification problems. As stated previously, different preprocessing methods have been successfully developed for baseline correction, smoothing, normalization, peak detection and peak matching. In this work, a systematic comparison of different software packages aiding in the compulsory preprocessing of MALDI data is performed. In order to guarantee the validity of the study, multiple configurations of each preprocessing technique that are subsequently used to train a set of classifiers are tested. Classifiers’ performance (i.e., kappa and accuracy) provided accurate information for the final comparison. Results showed the real impact of each preprocessing technique and configuration on classification, evidencing that MassSpecWavelet provides the best performance and Support vector machines (SVM) are one of the most accurate classifiers.

Finally, Chapter 4 presents Mass-Up, an open software multiplatform application for MALDI knowledge discovery that covers the full data analysis workflow. Mass-Up allows researchers to manage and visualize raw data or peak lists, to preprocess data, and different types of analyses such as (i) biomarker discovery, (ii) clustering, (iii) biclustering, (iv) three-dimensional PCA visualization and (v) classification of large sets of spectra data. Although there exists many software libraries and tools that can be combined for these kind of tasks and analysis, there was still a need for all-in-one solutions with graphical user-friendly interfaces and avoiding the need of programming skills.

Chapters 2 to 4 can be also read in their original journal format as they were annexed to this document (Appendix A to C).
1. Introduction

1.6 Conclusion and Future Work

The main objective of this thesis was the application of data mining and AI techniques to perform knowledge discovery with MALDI data.

In this thesis, different MALDI preprocessing methods were studied and compared (O1). In addition, an algorithm for peak matching called *Forward* was developed (O2) and used in almost all the developments and collaborations. Future work in this line includes the comparison of more publicly available libraries and working with more datasets.

During the course of this work, biclustering technique was applied to analyze MALDI data, being able to extract useful information and generate new hypotheses (O3, O4). Its adequacy was assessed by comparison against hierarchical clustering over two real datasets. Although results were promising, future work must deepen in this area and expand this study.

Integrating all specific developments and researches carried out for the studies presented in section 1.4, Mass-Up was developed (O5, O6, O7). With Mass-Up, an open-source tool with a friendly graphical user interface designed to allow proteomics researchers analyze MALDI data without the need to be bioinformatics experts is provided to the scientific community. Its usefulness is being demonstrated by the increasing number of studies that use it [19–21] and by the fact that it has been included in public mass spectrometry software repositories and projects, such as, for instance, MASSyPup(64), the Mass Spectrometry Live Linux, a Puppy Linux based Live distribution that includes several tools focused on the analysis of mass spectrometry data.

Regarding future work, this thesis work have two main lines of continuation. By one hand, continue developing and improving Mass-Up. Although Mass-Up has been continuously updated to fix bugs reported by users, some significant improvements have been identified: (i) support for more MALDI data formats, (ii) include more preprocessing algorithms and make them more flexible, and (iii) include new types of analyses. As explained previously, Mass-Up is a platform that integrates different pieces of software with a common basis and it has been continuously growing and evolving during these years. From a software engineering point of view, some improvements has also been identified and are planned to be carried out through a major code refactorization with all lessons learned from project start.
On the other hand, a collaboration with the GEPAM group *(University of Campinas, Sao Paulo, Brazil)* in order to perform bioimaging from elemental distribution using LA-ICP MS has been just started. The objectives of this collaboration are to provide the analytical basis for handling LA-ICP MS imaging technique and to develop a software tool to automate the process.

### 1.7 References


1. Introduction


Speeding up the screening of steroids in urine: Development of a user-friendly library

2.1 Abstract

This work presents a novel database search engine – MLibrary – designed to assist the user in the detection and identification of androgenic anabolic steroids (AAS) and its metabolites by matrix assisted laser desorption/ionization (MALDI) and mass spectrometry-based strategies. The detection of the AAS in the samples was accomplished by searching (i) the mass spectrometric (MS) spectra against the library developed to identify possible positives and (ii) by comparison of the tandem mass spectrometric (MS/MS) spectra produced after fragmentation of the possible positives with a complete set of spectra that have previously been assigned to the software. The urinary screening for anabolic agents plays a major role in anti-doping laboratories as they represent the most abused drug class in sports. With the help of the MLibrary software application, the use of MALDI techniques for doping control is simplified and the time for evaluation and interpretation of the results is reduced. To do so, the search engine takes as input several MALDI-TOF–MS and MALDI-TOF–MS/MS spectra. It aids the researcher in an automatic mode by identifying possible positives in a single MS analysis and then confirming their presence in tandem MS analysis by comparing the experimental tandem mass spectrometric data with the database. Furthermore, the search engine can, potentially, be further expanded to other compounds in addition to AASs. The applicability of the MLibrary tool is shown through the analysis of spiked urine samples.
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

2.2 Introduction

The use of androgenic anabolic steroids (AAS) and hormones to enhance athletic performance has important health and social implications. Their use was first introduced in sports as agents supporting the athlete recuperation after extreme stress and fatigue, but rapidly became the main agents used in doping abuse [1].

Nowadays, this class of drugs is a major group included in the prohibited list of the world anti-doping agency (WADA) as well as of major sports authorities [2-5]. In the WADA statistic report for 2011, the AAS represented 59.4% of all adverse analytical findings reported by WADA accredited laboratories [6]. Although this data may not reflect the real doping abuse statistical status, because of the well-known problems in the detectability of clandestinely designed AAS, micro dosages of endogenous AAS and “modern” doping agents (e.g., peptide hormones) [1, 7-9].

The use of AAS to increase muscle mass and strength is not a behaviour strictly related to elite athletes, as their use is increasing amongst amateur athletes as well as outside sports as an expression of an improved life style [10, 11]. The illicit AAS use is an increasing trend in western societies and the emergent AAS dependence is a matter of growing public health concern [12].

Quickly following the development of mass spectrometry (MS) detectors, its use coupled to gas chromatography (GC) has become the standard technique for AAS control. Currently, most methods for routine detection of these compounds and their metabolites, comprising both screening and confirmatory analysis, are still based in GC–MS techniques [13-17]. More recently, due to the increasing complexity of doping analyses and in order to enhance the detection of this group, liquid chromatography coupled with MS/MS is gaining ground within anti-doping laboratories [18-20]. In particular for the detection of thermo-labile and polar steroids, such as Trenbolone. It avoids the derivatisation step required by GC–MS and provides good sensitivity for the determination of these compounds [21, 22]. Moreover, the low throughput provided by LC–MS techniques are to some extent being surpassed by the introduction of uHPLC systems coupled with LC-columns containing solid core particles that allows high speed and high efficiency separations. Unfortunately, for the majority of the AAS, due to their poor ionisation efficiencies, their determination by LC–MS may lead to losses in sensitivity.
The combination of these two factors, the long separation times of gas chromatographic techniques and the increasing workloads within anti-doping laboratories, expose an urgent need for an analytical technique allowing simplicity, speed and high throughput for the screening of the huge number of banned compounds, particularly the AAS.

Recently, the use of matrix-assisted laser desorption/ionization (MALDI) for the analysis of small molecules, has grown as a potential technique, which is reflected by the increasing number of studies reported in literature [23-27]. Moreover, it appears extremely promising for high-throughput, which is a major demand for future anti-doping methods.

In light of the latest technological improvements of this analytical technique we have recently study the applicability of a wide variety of commercial MALDI matrices for the rapid screening of AAS [28]. The matrix 2-(4-hydroxyphenylazo)-benzoic acid (HABA) was found to be the most robust for the analysis of anabolic steroids after a derivatisation step with the reagent Girard T hydrazine. The Girard T hydrazone derivative produced after derivatisation is a quaternary ammonium ion that originates a strong [M]+ ion signal in the MALDI mass spectrum, as a result it increases the intensity of the steroid signal. In the aforementioned work it was demonstrated that positive identification of the characteristic peaks for all the compounds studied is possible for a sample concentration of 10 ng/mL in the MALDI sample plate. The sensitivity achieved with the HABA matrix after derivatisation was similar to that achieved by GC/MS – around 4–10 ng/mL in the single ion monitoring mode.

In the present work it is presented a step forward in simplifying AAS control through the use of and easy sample treatment and friendly software. The software is freely source code available, and it can be run as a multiple platform. As a proof-of-concept, the rapid screening of AAS in urine is reported using a sample treatment previously published by our team [28].

2.3 Materials and methods

2.3.1 Chemicals

Standards of 17-α-methyltestosterone and nandrolone were purchased from Riedel-de Haen (Seelze, Germany). The standards 17α-trenbolone, 2α-methyl-5β-androstane-3α-ol-17-one, mesterolone, methandienone, calusterone, fluoxymesterone, ethisterone and
mibolerone were kindly provided by the Portuguese National Anti-doping Laboratory and the Italian National Anti-doping Laboratory. A solution of $\beta$-glucuronidase from Escherichia coli K12 with a specific activity approximately of 140 U/mg at 37 °C and pH 7 with nitrophenyl-$\beta$-d-glucuronidase as substrate (1 mL contains at least 140 U) was purchased from Roche Diagnostic (Mannheim, Germany). Sodium hydrogen phosphate, sodium phosphate dibasic, tert-butylmethyl ether, methanol (MeOH), acetonitrile (ACN) and the derivatisation reagent, Girard T (GT) hydrazine, used for sample and matrix preparation were purchased from Sigma (Steinheim, Germany); glacial acetic acid (>99.5%), matrices αCHCA and HABA were purchased from Fluka (Buchs, Switzerland); trifluoroacetic acid (TFA, 99%) was from Riedel-de Haën.

Urine samples used in this work were obtained from healthy volunteers from the research team. Volunteer’s age ranged between 22–30 years, including both male and female. The research ethical committee from the Science Faculty of Ourense approved the study protocol and all the volunteers gave their consent.

2.3.2 Apparatus
A model UNIVAPO 100H vacuum concentrator centrifuge (UniEquip, Martinsried, Germany) with a model Unijet II refrigerated aspirator vacuum pump (UniEquip) was used for (i) sample drying and (ii) sample pre-concentration. A Spectrafuge-mini minicentrifuge (Labnet, Madrid, Spain) and a Sky Line minicentrifuge-vortex (ELMI, Riga, Latvia) were used throughout the sample treatment, when necessary. A Simplicity 185 system (Millipore, Milan, Italy) was used to obtain Milli-Q ultrapure water throughout all the experiments. The derivatisation procedure was performed in a 1.5 mL microtube flat cap from Delta Lab (Barcelona, Spain). Separation of the steroid Girard T (GT) hydrazones from the unreacted Girard T reagent was carried out in a 2 mL empty reversible solid-phase extraction (SPE) cartridge from Supelco (Belefonte, PA, USA) packed with a preparative C18 resin (125 Å, 55–105 μm; Waters, Barcelona, Spain).

2.3.3 Sample preparation

2.3.3.1 Standard solutions
Individual stock standard solutions of each compound (500 mg/L) were prepared by weighing 0.0125 g of analyte in a 25 mL volumetric flask and making it to volume with methanol. These standard solutions were stored in the dark at −20 °C. Working standard
solutions were prepared by dilution of the stock standard solutions in the appropriate volume of methanol.

2.3.3.2 Urine hydrolysis procedure

Urine samples (2 mL) were hydrolysed with 50 μL of the commercial solution of β-glucuronidase, after the addition of 0.750 mL of phosphate buffer (0.8 M, pH 7). The hydrolysis was performed at 55 °C during 60 min.

2.3.3.3 Liquid–liquid extraction of target analytes

After cooling to room temperature, 0.5 mL of carbonate buffer (pH 9) was added to alkalize the hydrolyzed solution. Liquid–liquid extraction was carried out by agitation with 5 mL of tert-butyl methyl ether for around 10 min; after centrifugation, the organic phase layer was transferred to a new vessel.

2.3.3.4 Derivatisation procedure

The procedure for derivatisation with Girard T hydrazine was performed based on the protocol described by Wheeler [29], as follows: The collected organic phase was dried under a gentle nitrogen stream at 40 °C. After the addition of 500 μL of a methanolic solution with 10% glacial acetic acid and 4 mg of Girard T hydrazine, the vial was closed and the derivatisation reaction was then performed at 60 °C during 30 min. After cooling, the solution was evaporated to dryness in a vacuum concentrator centrifuge and then reconstituted with 1 mL of methanol/water (10:90, v/v).

2.3.3.5 SPE clean-up

After derivatisation, the steroid GT hydrazones were separated from un-reacted GT hydrazine reagent by SPE in a C18 cartridge, according to the protocols described by Khan et al. [24] and Griffiths et al. [25]. Briefly, before use, the cartridges were conditioned with 5 mL of methanol plus 10 mL of MilliQ-water without allowing the cartridges to dry out. After loading the sample, the cartridge was washed with 2 mL of methanol/water (10:90, v/v) in order to remove impurities from the cartridge and, finally, the steroid GT hydrazones were eluted from the cartridge with 1 mL of methanol.

2.3.4 Experimental design

To assess the use of the MLibrary software for the analysis of AAS present in human urine, five urine samples were spiked with different AAS at different concentration levels.
The five urine samples were designated as Urine 1, Urine 2, Urine 3, Urine 4 and Urine 5. Urine 1 was spiked with 17α-methyltestosterone (100 ng/mL). Urine 2 was spiked with calusterone (250 ng/mL). Urine 3 was spiked with nandrolone (10 ng/mL). Urine 4 was spiked with fluoxymesterone (200 ng/mL) and ethisterone (150 ng/mL). Urine 5 was spiked with 17α-trenbolone (25 ng/mL) and mesterolone (300 ng/mL).

2.3.5 MALDI-TOF–MS and MALDI-TOF–TOF–MS analysis

The mass spectrometric analyses were performed on Applied Biosystems 4700 Proteomics Analyzer with TOF/TOF™ Optics system (Applied Biosystems, Foster City, CA, USA) equipped with a diode pumped Nd:YAG laser with 200 Hz repetition rate. The instrument was operated for detection in positive ion reflectron mode. The MS spectrum for each sample was based on the average of 1000 laser shots; for the MS/MS up to 4000 shots were accumulated. MS/MS mode was operated with 1 kV collision energy; air was used as the collision gas such that nominally single collision conditions were achieved. For MS analysis, laser desorbed ions were accelerated from the source at 20 kV. For MS/MS analysis, ions were accelerated from the source at 8.0 kV. Both modes employed delayed ion extraction for improved ion focusing. The MS/MS data was acquired using the instrument default calibration, whilst the MS was acquired using the peaks from the HABA matrix for internal calibration. Prior to MALDI analysis, the sample was mixed with an equal volume of the MALDI matrix solution and homogenised in a vortex instrument. The HABA MALDI matrix used in this work was prepared according to the developed method, 0.52 mg in 1 mL of a solution of ACN/H2O/MeOH (40/40/20, v/v/v) [28]. The matrix αCHCA was prepared by dissolving 10 mg in 1 mL of a solution ACN/H2O/TFA (50/49.9/0.1, v/v/v). An aliquot of the sample/matrix solution (0.5 μL) was hand-spotted onto the MALDI sample plate and the sample was allowed to dry. Figure 2.1 presents a scheme of the sample treatment followed in this study.
2.4. Results and discussion

The MLibrary software is explained in detail in the following sections. First we describe the use of MLibrary to detect the presence of possible AASs in the target samples (MS mode) and then its use to confirm the identity of the compound (MS/MS mode).

2.3.6 MLibrary software

To assess the use of the MLibrary software for the analysis of AAS present in human urine, five urine samples were spiked with different AAS at different concentration levels. The five urine samples were designated as Urine 1, Urine 2, Urine 3, Urine 4 and Urine 5. Urine 1 was spiked with 17α-methyltestosterone (100 ng/mL). Urine 2 was spiked with calusterone (250 ng/mL). Urine 3 was spiked with nandrolone (10 ng/mL). Urine 4 was spiked with fluoxymesterone (200 ng/mL) and ethisterone (150 ng/mL). Urine 5 was spiked with 17α-trenbolone (25 ng/mL) and mesterolone (300 ng/mL).
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

2.4.1 MS mode

2.4.1.1 Construction of MS database

The MS database in the MLibrary software contains the characteristic mass values for each target AAS. Additionally, the mass values of AAS glucuronides, which are the main excretion metabolites of AAS in the human body, were also introduced in the MLibrary database. In our previous work, the analysis of AAS by MALDI techniques was performed after derivatisation with Girard T hydrazine and therefore only this specific modification was introduced in the repository. Nevertheless, the introduction of specific modifications to the database is easily performed by typing the name and the mass variation in the modifications correspondent line within the XML file.

2.4.1.2 Detecting the presence of AASs

After the SPE clean-up procedure, as explained in the experimental section, the collected sample solution was mixed with the MALDI matrix HABA and spotted onto the MALDI sample plate. The MALDI ion source is a soft ionisation technique and therefore the MALDI-TOF–MS analysis measure, primarily, singly charged ions that correspond to the molecular ions of the sample solution species as well as to the MALDI matrix characteristic ions. The mass spectrum obtained after the MALDI-TOF–MS analysis is exported as a list of peaks to a CSV file. Only the centroid mass and relative intensity of each peak is used by the MLibrary software (see Figure 2.SM.1 of the Supplementary Material section).

The AASs detection process in MLibrary starts with the MS data loading. Through the “Load MS Data” operation (see Figure 2SM.2), the user can input the CSV file and filter peaks by their intensity, avoiding the load of peaks with lower intensities. Alternatively, the spectrum can be previously processed using the respective MS data software and only then transferred to a CSV file. It is important to stress that the MLibrary software permits the loading of multiple CSV files corresponding to several spectra.

After loading the MS data, the user can perform a compound search, in order to identify which AAS are present in the loaded data. The “MS Analysis” operation (see Figure 2.SM.3) determines if a specific AAS is present in the MS spectrum. This operation contemplates several parameters that provide flexibility to the software. It takes into account if the compound is in its “conjugated” or “free” form and also allows the
user to select the derivatisation agent employed. It is important to stress that the “conjugated” parameter only takes into account the glucuronide conjugates, which are the main excretion metabolites of AAS in the human body. The search is performed taking into account the error between the database and the experimental mass values. This error value can be selected as percentage, parts per million (ppm) or absolute mass units (amu) (Figure 2.SM.3). Then, the search retrieves the mass values that matched between the experimental data and the database values, showing both experimental and theoretic values as well as the name of the compound and the experimental peak intensity (see Figure 2.2). MLibrary also provides an additional operation named “MS Full Analysis”, which performs a search within all the MS databases present in the MLibrary repository.

Figure 2.2 MLibrary MS spectrum analysis display window. Data identification: blue line: experimental spectrum mass peaks; green line: database reference masses; red line: experimental mass that matched with the reference mass.

Since the AAS mass values are close to the matrix characteristic ion mass ones, it is possible to perform an internal calibration for each MALDI plate spot using the matrix peaks as reference. Consequently, the experimental mass values obtained are very accurate and precise.

2.4.2 MS/MS mode

In the first step of the analysis, the MLibrary software detected the ions that matched with the theoretic mass values of AAS within the MS database. In the second step, the ions detected by the MLibrary are selected for fragmentation in a second round of MS analysis, in which MS/MS spectra are acquire. Each AAS compound presents a characteristic
MS/MS fragmentation pattern that will be used by the MLibrary as a signature of that compound. Likewise the MS mode, the mass spectrum obtained after the MALDI-TOF–TOF–MS analysis is exported as a list of peak to a CSV file.

2.4.2.1 **Construction of MS/MS spectral library**

The MS databases in the MLibrary contain the characteristic fragmentation ions for each associated compound. The construction of the database can be easily performed using standard solutions of AAS. To ensure reliable spectral data, the software allows the input of several replicates corresponding to each standard. With these data, the MLibrary software generates a list of common fragmentation mass values for all replicates, along with their average relative intensities. In addition, when comparing with loaded data, the software allows the user to choose the discriminate power of the generated list, meaning that the user can decide to include within this list, the mass values that are present, for instance, at least in 95% percent of the spectral mass data inputted for each compound.

The construction of the mass spectral library in the MLibrary software is an ongoing process that is simple to achieve and updated by any qualified user, with the advantage that it can be easily adapted to numerous experimental conditions and compounds.

2.4.2.2 **Confirming the presence of AASs**

This process is similar to the AAS detection process except that, in this case, the user can perform two additional analysis operations. As before, the process starts with the data loading, in this case, through the “Load MS/MS Data” operation (see Figure 2.SM.4). This operation also provides the “Peak Intensity” parameter for peak filtering.

After loading the sample CSV file, the user can use the “MS/MS Library Analysis” operation (see Figure 2SM.5) to compare it with the characteristic MS/MS spectra of the AAS compounds stored in the MLibrary databases. At this stage the user has to select the database, the modifications performed in the sample treatment procedure, the mass value of the precursor molecular mass and the mass tolerance permitted for each mass peak. Additionally, the user has to select the mass tolerance within the database mass spectra and its discriminate power. This operation retrieves a list of all compounds within the MS/MS library that fit the search criteria, ranked by similarity to the inputted file. The similarity is attained regarding the number of mass values that matched and the
relative intensity of all peaks. To ensure reliable results, the search score concerns only to a limited number of mass values, defined previously by the user. For instance, the user may limit the search to the 10 most intense peaks within the inputted file and the library data (see Figure 2.SM.6). This tool is essential to avoid misinterpretation of the results due to the fact that different spectra of the same sample, generally present distinct overall number of mass peaks.

Another important tool within the MLibrary software is the “MS/MS Std. Match Analysis” operation. With this tool the user can compare two experimental spectra, which is extremely important if we are working at different conditions than the one recorded in the MLibrary MS/MS database. By adding a standard solution of a specific compound to the MALDI analysis and comparing the two spectra using the MLibrary (see Figure 2.SM.7), the user can confirm the identity of this compound. The results appear in the same way that showed in Figure 2.SM.6.

A third MS/MS Analysis tool is available in the MLibrary software. The “MS/MS Marker Analysis” operation allows the user to locate concrete biomarkers into the loaded MS/MS data (see Fig. 2.SM.8). This feature is particularly important for the analysis of isobaric compounds having very similar MS/MS fragments or compounds having poor fragmentation pattern. For isobaric compounds the MS/MS Library analysis does not allow the differentiation between the two species, since it will retrieve very close results. For this reason, the identification of exclusive fragments corresponding to each compound is essential for the interpretation of the spectra and it will become essential for their identification. Additionally, the user can use the “MS/MS Full Marker Analysis” in order to search for all the biomarkers stored in the MLibrary repository.

2.5 Case study

Table 2.1 presents the results obtained with the data acquired from five urine samples spiked with different AAS at different concentration levels as described in the sample preparation section. All spiked compounds were detected by single MS as possible positives and their identity was confirmed by MS/MS. In the MS mode, besides the spiked compounds, the MLibrary retrieved other AAS as possible positives. This result occurs due to the fact that some synthetic AAS possesses the same molecular mass than some endogenous AAS, which are present in urine at low concentration values. For instance,
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

4-androsten-3,17-dione is a minor AAS metabolite that possesses the same mass than boldenone.

<table>
<thead>
<tr>
<th>AAS compounds spiked into urine</th>
<th>MLLibrary data analysis</th>
<th>AAS identified by MS mode</th>
<th>AAS confirmed by MS/MS mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urine 1 17α-methyltestosterone</td>
<td>Boldenone</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17α-methyltestosterone</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mibolerone</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mesterolone</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Urine 2 Calusterone</td>
<td>Calusterone</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bolasterone</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Urine 3 Nadrolone</td>
<td>Nadrolone</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Boldenone</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mesterolone</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Urine 4 Ethisterone Fluoxymesterone</td>
<td>Mesterolone</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ethisterone</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fluoxymesterone</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Urine 5 Trenbolone Mesterolone</td>
<td>Trenbolone</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Boldenone</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mesterolone</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.SM.9 presents the results retrieved by the MLibrary for the sample Urine 1. As it may be seen in Figure 2.SM.9A, in the MS mode, the software identified three peaks corresponding to four possible positives: boldenone, 17α-methyltestosterone, mibolerone, which has the same molecular mass that 17α-methyltestosterone, and mesterolone. As mentioned above, following the detection of the possible presence of specific AASs in the urine sample, a MALDI-TOF–TOF–MS analysis was performed to the ions retrieved as possible positives. The MLibrary software was used to confirm the identity of the compounds. Figures 2.SM.9B, 2.SM.9C, 2.SM.9D and 2.SM.9E show the results obtained for the four compounds. The presence of both boldenone and mesterolone is easily proved false by MLibrary MS/MS data analysis, showing less than 10% of positive match to the reference database spectra within the 10 most intense peaks.

Regarding 17α-methyltestosterone and mibolerone, the MLibrary MS/MS data analysis clearly confirms the presence of 17α-methyltestosterone. MLibrary software retrieves a higher percentage of mass values matching with 17α-methyltestosterone reference data within the 10 more intense peaks; 70% against 50% for mibolerone.
Furthermore, the peak intensity of all matched mass ions present in the MS/MS spectrum is similar to the MLibrary reference data for 17α-methyltestosterone.

Figures 2.AA.1–4 of the Appendix A present the results retrieved by MLibrary for the samples Urine 2, Urine 3, Urine 4 and Urine 5, respectively.

In sample Urine 2, the presence of the isobaric compounds calusterone and bolasterone are detected by the MS mode (see Figure 2.AA.1 of the Appendix A). Although both calusterone and bolasterone present similar fragmentation patterns, the intensity of the peaks are different. For this reason, despite the similarity of the matched peaks between the sample compound and the reference data for calusterone and bolasterone, the MLibrary software easily confirms the identity of calusterone. It retrieves a higher percentage of mass values matching with calusterone within the 10 more intense peaks; 80% against 40%.

In sample Urine 3, the presence of nandrolone was confirmed using the MS/MS Marker Analysis tool (see Figure 2.AA.2 of the Appendix A). As it was mentioned above for poor fragmentation pattern compounds, such as nandrolone, the detection of specific markers, previously identified by the user, is of major importance.

Figure 2.AA.3 of the Appendix A shows the results retrieved by MLibrary that allowed the identification of both fluoxymesterolone and ethisterone in sample Urine 4. The identification of ethisterone was performed in the same manner than for nandrolone by using the MS/MS Marker Analysis Tool. The presence of mesterolone was also proved false, showing 0% of matching within the most intense peaks.

In sample Urine 5, the MLibrary identified the presence of 17α-trenbolone and mesterolone (see Figure 2.AA.4 of the Appendix A). The identification of 17α-trenbolone was performed using the tool MS/MS Standard Mach Analysis. The MS/MS identification of 17α-trenbolone was performed, in a first approach, using the Library analysis tool, however only 60% of matching was achieved. The MS/MS Standard Mach Analysis tool allowed the unequivocal confirmation of this compound by comparing it with a standard solution analysed in the same experimental conditions.

2.6 Conclusions

We have developed friendly software to help in an automated mode to detect and identify the presence of AAS in urine samples by MALDI-TOF–TOF–MS. MLibrary software allows the user to perform robust and accurate screening and confirmation for
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

AAS in both MS and MS/MS mode. As a proof-of-concept, when applied to five urine samples spiked with AAS, the software was able to successfully identify all spiked compounds. The software presented here is a very versatile tool that can be simply adapted to any future modifications carried out on the sample treatment procedure, as well as, be easily applied to other compounds. The MLibrary software saves times and it is a simple tool to work with. Additionally, MLibrary software has a wizard easy to follow for its installation.

2.7 Supplementary Material

This section includes supplementary material figures.

Figure 2.SM.1 CSV file showing centroid mass and relative intensity of each peak. This is an example of the final data list retrieved from the MALDI analysis.

Figure 2.SM.2 MLibrary interface for MS spectrum data load
2.7. Supplementary Material

Figure 2.SM.3 MLibrary MS spectrum analysis search window. Input information: Database, compound state, modification and mass tolerance.

Figure 2.SM.4 MLibrary interface for MS/MS spectrum data load.
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

Figure 2.SM.5 MLibrary MS/MS spectrum analysis search window. Input information: Database, compound mass or compound name, compound state, modification, mass tolerance concerning the experimental and reference mass values, mass tolerance within the database spectra and discriminate power.

Figure 2.SM.6 MS/MS Library Analysis result display window. Data identification: blue line: experimental spectrum mass peaks; green line: database reference masses; red line: experimental mass that matched with the reference mass.
2.7. Supplementary Material

Figure 2.SM.7 MS/MS Std. Match Analysis display window.

Figure 2.SM.8 MS/MS Marker Analysis display window.
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

Figure 2.SM.9A MLibrary sample Urine 1 retrieved result (a) MLibrary MS spectrum analysis display window. Data identification: blue line: experimental spectrum mass peaks; green line: database reference masses; red line: experimental mass that matched with the reference mass.

Figure 2.SM.9B MLibrary MS/MS spectrum analysis display window for boldenone.
Figure 2.SM.9C MLibrary MS/MS spectrum analysis display window for 17α-methyltestosterone.

Figure 2.SM.9D MLibrary MS/MS spectrum analysis display window for mibolerone.
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

Figure 2.SM.9E MLibrary MS/MS spectrum analysis display window for mesterolone.

2.8 Appendix A

This section includes supplementary figures showing the results retrieved by MLibrary for the samples Urine 2, Urine 3, Urine 4 and Urine 5.

Figure 2.AA.1A Sample Urine 2: MLibrary MS spectrum analysis display window. Data identification: 
*blue line*: experimental spectrum mass peaks; *green line*: database reference masses; *red line*: experimental mass that matched with the reference mass.
2.8. Appendix A

Figure 2.AA.1B Sample Urine 2: MLibrary MS/MS spectrum analysis display window for calusterone.

Figure 2.AA.1C Sample Urine 2: MLibrary MS/MS spectrum analysis display window for bolasterone.
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

Figure 2.AA.2A Sample Urine 3: MLibrary MS spectrum analysis display window. Data identification: 
*blue line*: experimental spectrum mass peaks; *green line*: database reference masses; *red line*: experimental mass that matched with the reference mass.

Figure 2.AA.2B Sample Urine 3: MLibrary MS/MS spectrum analysis display window for nandrolone.
Figure 2.AA.2C Sample Urine 3: MLlibrary MS/MS spectrum analysis display window for boldenone.

Figure 2.AA.2D Sample Urine 3: MLibrary MS/MS spectrum analysis display window for mesterolone.
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

Figure 2.AA.3A Sample Urine 4: MLibrary MS spectrum analysis display window. Data identification: 
*blue line*: experimental spectrum mass peaks; *green line*: database reference masses; *red line*: experimental mass that matched with the reference mass.

Figure 2.AA.3B Sample Urine 4: MLibrary MS/MS spectrum analysis display window for mesterolone.
Figure 2.AA.3C Sample Urine 4: MLibrary MS/MS spectrum analysis display window for ethisterone.

Figure 2.AA.3D Sample Urine 4: MLibrary MS/MS spectrum analysis display window for fluoxymesterone.
2. Speeding up the screening of steroids in urine: Development of a user-friendly library

Figure 2.AA.4A Sample Urine 5: MLibrary MS spectrum analysis display window. Data identification: 
*blue line*: experimental spectrum mass peaks; *green line*: database reference masses; *red line*: experimental mass that matched with the reference mass.

Figure 2.AA.4B Sample Urine 5: MLibrary MS/MS spectrum analysis display window for trenbolone.
2.9 References


2. Speeding up the screening of steroids in urine: Development of a user-friendly library


References


A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification

3.1 Abstract
Matrix assisted laser desorption ionization time-of-flight (MALDI-TOF) is one of the high-throughput mass spectrometry technologies able to produce data requiring an extensive preprocessing before subsequent analyses. In this context, several low-level preprocessing techniques have been successfully developed for different tasks, including baseline correction, smoothing, normalization, peak detection and peak alignment. In this work, we present a systematic comparison of different software packages aiding in the compulsory preprocessing of MALDI-TOF data. In order to guarantee the validity of our study, we test multiple configurations of each preprocessing technique that are subsequently used to train a set of classifiers whose performance (kappa and accuracy) provide us accurate information for the final comparison. Results from experiments show the real impact of preprocessing techniques on classification, evidencing that MassSpecWavelet provides the best performance and Support vector machines (SVM) are one of the most accurate classifiers.

3.2 Introduction
In the last years, high-throughput mass spectrometry (MS) based proteomic data analysis has been an active research area. MS technology allows researchers to measure the mixture of peptides or proteins present in biological samples, such as urine, serum or tissues. These measurements can be further used for discovering condition related patterns (biomarker discovery) and subsequently, for classify samples automatically.
A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification

A typical mass spectrometry experiment always produces a large volume of raw data. Tens to hundred MS spectra are generated, each one containing thousands of measurements (i.e. \{m/z, intensity\} pairs). In such a situation, dimensionality reduction is a crucial task that must be carried out before subsequent classification or biomarker discovery [1]. This is done in a step globally called preprocessing, an extensive low-level procedure able to clean raw data and detect true signals in the noisy spectra [2]. The whole preprocessing procedure comprises several tasks such as baseline correction, smoothing, normalization, peak detection and peak alignment. Since the application of inadequate or incorrect preprocessing methods can result in biased dataset, also hindering the achievement of meaningful biological conclusions [3], preprocessing is a critical task in rigorous MS data analysis. For this reason, several algorithms have been proposed to address each preprocessing task. These algorithms differ from each other in their principles, implementations and performance [4], and therefore, their output can affect subsequent analyses.

Related with this situation, several studies have been carried out in order to compare different preprocessing algorithms and alternatives. In these works, comparisons are performed in terms of false discovery rate (FDR), sensitivity and reproducibility [4-6]. Unlike these studies and from another perspective, the goal of this work is to systematically study how different preprocessing algorithms and configurations affect classifiers accuracy when dealing with MS data.

After motivating the work, the rest of the paper is structured as follows: section 2 gives a general overview about mass spectrometry domain. Section 3 identifies and briefly describes the analyzed preprocessing steps. While Section 4 introduces the experimental setup giving details for guaranteeing reproducibility, Section 5 presents and discusses obtained results. Finally, Section 6 concludes and outlines future work.

### 3.3 Related Work on Mass Spectrometry

In proteomics, mass spectrometry stands for a commonly used technique for measuring the mass-to-charge ratio (m/z) of the components in a sample [7]. MS provides rapid and precise measurements about the sizes and relative abundances of the proteins present in a complex biological/chemical mixture [8]. Mass spectrometric measurements are carried out using mass spectrometers, comprising an ionization source, a mass analyzer and a detector.
In a typical analysis, the sample is passed through these three components generating a mass spectrum. First, the components of the sample are ionized in the ionization source in order to be able to measure their masses. There are two main types of sources [9]: based on electrospray ionization (ESI) or based on matrix assisted laser desorption ionization (MALDI), which represents the dominating ionization source for (single) mass spectrometry [7]. Then, in the mass analyzer, the components of the sample are separated according to the mass-to-charge ratio (m/z) of the ions. After separation, the components hit the detector and their m/z values can be calculated. There are four basic types of mass analyzers currently used in proteomics research [10]: (i) ion trap, (ii) time-of-flight (TOF), (iii) quadrupole and (iv) Fourier transform ion cyclotron (FT-MS). While ESI are usually coupled to ion traps and triple quadrupole analyzers, MALDI are usually coupled to TOF analyzers and are the simplest instruments for protein and peptide analysis.

A mass spectrometer (e.g. MALDI-TOF) operates connected to a computer, which constructs a mass spectrum after analyzing a sample in the instrument (Figure 3.1). A mass spectrum is a representation where the measured masses are placed along the horizontal axis and the intensity of the signal for each component (m/z) along the vertical axis. As a result of its simplicity, excellent mass accuracy, high resolution and sensitivity [7, 10], MALDI-TOF is widely used to identify proteins thought their peptides, a process known as peptide-mass fingerprinting. In this process, the mass spectrum must be preprocessed in order to obtain a list of peptide experimental masses, which can be searched against a database to identify proteins.

![Figure 3.1 Mass spectrometer.](image)

Mass spectrometry is generally applied both to biomarker discovery [11-14] and classification [15-18]. In this context, and as pointed out in [3], it is important to note that a typical dataset arising in an application of mass spectrometry contains from tens to hundreds spectra. Each spectrum has between 10,000 and 100,000 intensity measurements representing an unknown number of protein peaks. Such huge volume of
3. A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification

raw data requires extensive low-level preprocessing in order to clean the data and to detect the true signals in the noisy spectra [2].

3.4 Available Preprocessing Methods for MS Data

As stated before, preprocessing of MS data is a critical stage that transforms raw data into a suitable input for further analysis, such as machine learning or biomarker discovery. Inadequate or incorrect preprocessing methods can result in biased dataset and hinder to reach meaningful biological conclusions [3]. In such a situation, preprocessing is necessary since raw data contains signals coming from the real peptides/proteins, as well as signals derived from several forms of noise (e.g. chemical, electronic factors, etc). The specific goals of this phase are [8] (i) to remove noisy peaks without discarding any of the true peaks, and (ii) to determine the m/z and intensity values with the best accuracy. Since there is no standard mass spectrometry data preprocessing pipeline, some authors proposed different guidelines to establish a design/data analysis protocol (DAP) [2, 19]. In this work, we use the following preprocessing workflow: (i) baseline correction, (ii) smoothing, (iii) peak detection and (iv) peak alignment. Next subsections introduce the most popular preprocessing steps and techniques in detail.

3.4.1 Baseline correction

Baseline is a specific form of noise mainly motivated by chemical perturbations. In [7] baseline is defined as an offset of the intensities of masses that often shows a dependency on the m/z value such that it is highest at low /z values, presenting an exponential decay towards higher masses. For MALDI-TOF mass spectrometry, baseline is a monotonically decreasing bias resulting from matrix clusters formed during ionization [20, 21]. The most common baseline correction methods are monotone minimum, linear interpolation, loess, moving average of minima and continuous wavelet transform, being available as free software. The Bioconductor PROcess package [22] implements both loess and linear interpolation. Moreover, the Cromwell package [23] implements monotone minimum methods. Figure 3.2 shows an example of a raw spectrum with baseline noise (a) and the corresponding baseline corrected spectrum (b).
3.4. Available Preprocessing Methods for MS Data

3.4.2 Smoothing

Usually, spectra are jagged, making it difficult to detect true peaks amongst the noise. Therefore, a smoothing algorithm is usually applied in order to soften the data. Again, several methods have been successfully proposed in the literature for correcting this issue. The simplest technique consists of using a sliding window, where a new value is calculated for the point in the middle, based on the values of the points in the window. Commonly used filters are moving average, Savitzky-Golay, Gaussian and the Kaiser window. However, MS community appears to be converging on the use of wavelets for denoising [3]. In this context, Combes et al. [23] presented a denoising method based on the undecimated discrete wavelet transform (UDWT) that is available in the Cromwell package.

3.4.3 Peak detection

Peak detection can be defined as the process of selecting true (i.e. peptide/protein-related) peaks from a given spectrum. As in previous subsections, there are many peak detection algorithms, most of them applied after baseline correction and smoothing. However, Du et al. proposed a method that performs peak detection without explicit smoothing and baseline correction [24]. This method is based on the continuous wavelet transform (CWT) and is publicly available in the MassSpecWavelet package, which is also in the Bioconductor project.

Peak detection algorithms use one or more of the following criteria in order to identify true peaks [4]:

![Figure 3.2 Example of baseline correction.](image)
3. A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification

- **Signal to Noise Ratio (SNR).** SNR is a measure of the signal relative to the background noise. Peaks are selected if their SNR is larger than a given threshold.
- **Detection/intensity threshold.** This criterion is used to remove small peaks in flat regions. The use of SNR alone in such regions may select noisy points as peaks, since they may have a high SNR.
- **Slopes of peaks.** Within this criterion the shape of peaks is used to filter out false occurrences. Any potential peak is eliminated if both left and right slopes are less than a pre-established threshold. The limit is defined as half of the local noise level [25].
- **Local Maximum.** Following this criterion a peak is selected if it is a local maximum of N neighboring points.
- **Shape ratio.** Within this criterion, a peak is selected if its shape ratio [4] exceeds a certain threshold.
- **Model-based criterion.** These methods use a model function in order to fit peaks.
- **Peak width.** A peak is detected as true if its peak width is contained in a given range.

For example, in the Cromwell package peaks are selected if they are local maxima and larger than a given SNR. However, in the PROcess package in addition to the SNR, intensity threshold and shape ratio criterion are also used.

### 3.4.4 Peak alignment

This process, also referred as peak matching, consists on determining which peaks correspond to the same peptide/protein in different samples. Apart from the existence of other standard algorithms, in this work we analyzed our own peak alignment method based on a moving window successfully used in previous works [26, 27]. As a result of the execution of the algorithm, all the aligned peaks have the same mass values in all spectra. These mass values correspond to the virtual centroid mass. Figure 3.3 shows the pseudocode algorithm of the proposed method.
3.4. Available Preprocessing Methods for MS Data

Figure 3.3 Pseudocode of the proposed peak alignment algorithm.

The window size of the algorithm showed in Figure 3.3 may be specified in three different ways: (i) absolute, (ii) relative and (iii) PPM (points per million). In this context, a PPM window size of 500 means that at a given m/z value m, the window is m ± (m • 500 / 106). When working with PPM, the higher mass value the wider window size, based on the fact that mass measurements at higher masses tend to be more distant.

With the goal of clarifying the computation of a peak alignment, Figure 3.4 shows an example that illustrates the alignment process of three sample spectra. Considering a window of 500 PPM, the three peaks with masses 999.7, 1000.1 and 1000.3 are aligned (i.e. they represent the same peak), so that after alignment the three spectra contain the same peak. This new peak is the centroid of the three peaks being initially aligned, and it is computed as their mean.
3. A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification

Figure 3.4 Peak alignment process with a window of 150 PPM.

3.5 Experimental Setup

With the goal of conducting a straightforward and reproducible experiment for evaluating the influence of different techniques supporting the previous commented preprocessing pipeline, this section introduces in detail important questions regarding our experimental setup: (i) dataset preparation procedure, (ii) machine learning techniques and (iii) global preprocessing configuration and classification workflow.

3.5.1 Mass Spectrometry datasets

A typical MS dataset contains samples divided into two or more classes (e.g. healthy and diseased). Usually, for each sample, there are several biological replicates and, for each one, there are several technical replicates (i.e. the result of applying the same experimental procedure several times to the same biological replicate). Finally, each technical replication is spotted several times into the mass spectrometer.

For this work we have used a large dataset consisting of fourteen different Spanish wines with five bottles of each wine [26]. Each bottle (i.e. sample) was spotted five times so that the wine dataset contains a total of 350 \((14\times5\times5)\) spectra.

In order to further analyze the effect of commonly used preprocessing techniques in classification problems, the MS dataset must be stored in a proper format, such as WEKA’s own ARFF [28]. For this reason, the set of attributes consists of the masses of all spectra together with the corresponding class attribute. Then, each spectrum is added
as a new data line (i.e. instance) with ones where peaks are present and zeros otherwise. Figure 3.5 shows an example of an MS dataset exported as ARFF file.

```
@relation WineMassData
@attribute sample {A#1, A#2, ..., B#1, B#2, ...} numeric
@attribute mass1 numeric
@attribute mass2 numeric
...@
@attribute massN numeric
@attribute class {A, B, ...} numeric
@data
A#1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, ..., 0, A
A#2, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, ..., 0, A
...B#1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, ..., 0, B
B#2, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, ..., 0, B
```

Figure 3.5 MS dataset in ARFF format.

### 3.5.2 Machine learning algorithms

It is generally accepted by the scientific community that there is no a superior classification algorithm/technique for being applied to all domains. Taking into account this limitation, in this work we have selected several algorithms covering different learning approaches such as decision trees, instance-based learning, support vector machines or probabilistic models.

As instance-based learning approach we have chosen the K-nearest neighbors (KNN) algorithm [29], implemented by WEKA in the class IBk (weka.classifiers.IBk). In this classification algorithm, no internal model is built, and the training stage just consists of storing training samples. To classify a new case, its K-nearest neighbors (i.e. the K samples most similar to it in terms of the specific problem domain) are selected. The most frequent class among these K neighbors is the class predicted for the new case.

As probabilistic models, we have chosen Bayesian networks [30] and Naïve Bayes [31] classifiers available at BayesNet and NaiveBayes classes belonging the weka.classifiers.bayes package. Naïve Bayes is a simple classifier that assigns a case x to the most probable class given x, relying on the Bayes’ theorem and the variable independence assumption.

As decision tree (DT) learning algorithms, we have selected C4.5 [32] and Random Forest (RF) [33]. RF is an ensemble classifier that consists of many DTs, where
the output class is computed as a combination from the output classes of each individual tree. Briefly, a decision tree is built by recursively partitioning the training data with the aim of maximizing the class homogeneity of the resulting subsets. The selected variable will be the one that ensures the maximal reduction of class heterogeneity measured in different ways (e.g. in the case of C4.5 this measure is the entropy). Both J48 (the implementation of C4.5 algorithm) and RandomForest are available in the weka.classifiers.trees package.

Finally, we have also opt for several classifiers based on Support Vector Machines (SVMs) [34], which have been applied extensively to mass spectra [1]. On the one hand, a SVM trained with Sequential Minimal Optimization (SMO) [35], is available in the class SMO (weka.classifiers.functions). On the other hand, two SVMs: a radial basis function (SVM-RBF), and a linear function (SVM-L). These SVMs are available through the WEKA’s class LibSVM, which acts as wrapper for the libsvm tools [36].

3.5.3 Preprocessing configuration

As it was previously commented, the selected dataset consists of 350 MALDI-TOF raw spectra stored in mzML format [37], the PSI (Proteomics Standards Initiative) standard for MS data. In our experiments spectra are read using jmzML [38], an open-source Java library for mzML.

Each single spectrum is treated by using a combination of the preprocessing steps described in Section 3. In our experimental workflow (Figure 3.6), peak detection (PD) is always performed, whereas baseline correction (B) and smoothing (S) are optional. As a result of this operation, a list of representative peaks for each spectrum is obtained.
3.5. Experimental Setup

Figure 3.6 Experimental workflow followed in the present study.

In order to execute the preprocessing tasks, we have used PROcess and MassSpecWavelet (MSW) packages partly because (i) they are commonly used mass spectrometry preprocessing software and (ii) the availability of integration facilities using RJava [39] to invoke them. Table 3.1 shows the preprocessing steps available in PROcess and MSW together with the set of parameters under study based on the experiments of Yang et al. [4]. In PROcess, smoothing and peak detection should be executed together, resulting in a single step.

Table 3.1 Preprocessing configuration.

<table>
<thead>
<tr>
<th>Step</th>
<th>Package</th>
<th>Parameters</th>
<th>Values</th>
<th>Combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>PROcess</td>
<td>method</td>
<td>approx, loess</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bandwidth</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>S+PD</td>
<td>PROcess</td>
<td>SNR</td>
<td>{1, 3, 5}</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>peak neighbor</td>
<td>{0.3, 0.003, 0.0003}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>peak ratio</td>
<td>{0.001, 0.01, 0.1, 0.5}</td>
<td></td>
</tr>
<tr>
<td>PD</td>
<td>MSW</td>
<td>SNR</td>
<td>{1, 3, 5}</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>peak scale range</td>
<td>{2, 4, 6}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>peak amplitude</td>
<td>{1, 3, 5}</td>
<td></td>
</tr>
</tbody>
</table>

The first preprocessing block depicted in Figure 6 using the preprocessing configurations showed in Table 3.1 leads to 189 different setups: 36 (PROcess S+PD) + 27 (MSW PD) + 2 × 36 (PROcess B | PROcess S+PD) + 2 × 27 (PROcess B | MSW PD).
In the second preprocessing block showed in the experimental workflow of Figure 6, peak alignment is performed by using the algorithm described in Subsection 3.4. Also in this step, a sample (i.e. a bottle, in the wine dataset) is summarized into a single spectrum (called sample fingerprint) instead of the 5 replicates per spectrum. Consequently, after peak alignment, the wine dataset is represented by 70 spectra, corresponding to 14 classes with 5 samples per class, in which each spectrum represents one sample. The process of generating a sample fingerprint is parameterized by the intra-sample percentage of presence (POP), which stands for the minimum number of replicates in where a peak must be present in order to be added to the fingerprint. The higher POP, the less number of peaks selected. For instance, considering that there are 5 replicates for each sample, POP varies from 20% (which can be seen as the union of the 5 spectra), to 100% (which can be understood as the intersection of the 5 spectra).

Peak alignment is performed inside a loop, where a new version of the dataset is created at each iteration. In this loop, the intra-sample POP varies from 1 to the maximum replicates in a sample (5 in our dataset). Thus, after applying the same preprocessing pipeline to the original dataset, there are several additional variations due to the intra-sample POP used to create the sample fingerprints. As a consequence, there is a total of 945 preprocessing configurations (189×5).

In order to systematically study the influence of the preprocessing steps (algorithms and their parameter configurations), the set of classifiers presented in Subsection 4.2 was applied to each preprocessed dataset resulting from the previous experimental workflow. All the classifiers were executed with default parameters except for IBk, where K was configured considering 10 neighbors.

With the goal of guaranteeing the validity of the results, we conducted an ad hoc 10-fold cross-validation experiment. All the runs were executed on the AIBench platform [40] using a WEKA plugin. In order to precisely measure the accuracy of each classifier we have used both (i) the percentage of correct classifications (accuracy) and (ii) the Cohen’s kappa statistic [41]. The kappa index compensates for classifications that may be due to chance and it is considered a standard statistically-robust measure useful to assess the accuracy in multiclass problems [42]. Kappa compares the real class with the predicted one, giving values ranging from 0 (random classification) to 1 (perfect classification) and being computed as showed in Expression (1).
3.6 Results and Discussion

As previously mentioned, we selected kappa as performance measure in order to carry out subsequent analyses from the outcome of the statistical tests. For each one of the 945 possible configurations, we collected in a single table those values corresponding to baseline correction, peak detection, POP, type of classifier, fold and kappa.

First, we carried out a multi-factor ANOVA [43] for analyzing the full table of results. This test allows us to see: (i) for each parameter, if there are significant differences among its possible values and (ii) if there are interactions between parameters. The goal of this test is two-fold: detecting parameters that significantly impact on the classification and identifying possible interactions among them. The results of the test are showed in Table 3.2.

Table 3.2 Results of a multi-factor ANOVA.

<table>
<thead>
<tr>
<th>Factors</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>classifier</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>POP</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>Peakdetection</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>Baselinecorrection</td>
<td>0.2075</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interactions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>classifier:POP</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>classifier:Peakdetection</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>POP:Peakdetection</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>classifier:Baselinecorrection</td>
<td>0.9999</td>
</tr>
<tr>
<td>POP:Baselinecorrection</td>
<td>0.9999</td>
</tr>
<tr>
<td>Peakdetection:Baselinecorrection</td>
<td>0.9734</td>
</tr>
<tr>
<td>classifier:POP:Peakdetection</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>classifier:POP:Baselinecorrection</td>
<td>1</td>
</tr>
<tr>
<td>classifier:Peakdetection:Baselinecorrection</td>
<td>1</td>
</tr>
<tr>
<td>POP:Peakdetection:Baselinecorrection</td>
<td>1</td>
</tr>
</tbody>
</table>

Results from Table 3.2 reveal that the classifier, POP and peak detection factors significantly affect the classification performance, while baseline correction parameter
3. A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification

does not. This means that there are no significant differences among the baseline correction methods employed (none and both loess and approx from PROcess). This conclusion was expected since in a visual inspection of the spectra, baseline artifact was not detected.

On the other hand, the test has identified differences among the different peak detection algorithms, the POP values and the classifiers. Moreover, the test has revealed that these three factors have an interaction. To precisely study this interaction, we carried out a Tukey's HSD (Honestly Significant Difference) post-hoc test.

The interaction between classifier and POP means that each classifier is not affected in the same way by the different values of POP (Figure 3.7). While most of classifiers perform worse with higher POP values, SVM-L and SVM-RBF does not seem to be affected by this factor.

![Figure 3.7 Boxplot of classifiers by POP.](image)

Tukey’s HSD test performs pair-wise comparisons for each possible combination of classifier, peak detection configuration and POP, which results in millions of comparisons. In order to reduce such huge size, we have studied each peak detection
package (MSW and PROcess) separately. This analysis allowed us to select the best peak detection configurations in each package, and then filter the Tukey’s HSD output to reduce its size.

Figure 3.8 shows a boxplot of PROcess peak detection parameters. First, they are grouped by ratio, then by SNR, and finally by area.w. A visual inspection of the boxplot suggests that SNR and ratio may not affect the performance of the classifiers, while area.w clearly affects it.

![Figure 3.8 Boxplot of PROcess parameters.](image)

Again, we carried out an ANOVA test, whose results are showed in Table 3.3. The test confirms that there are no statistically significant differences in SNR and ratio, something that does not occur with area.w.

<table>
<thead>
<tr>
<th>Factors</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>area.w</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>snr</td>
<td>0.9790</td>
</tr>
<tr>
<td>ratio</td>
<td>0.8750</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interactions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>area.w:snr</td>
<td>0.9480</td>
</tr>
</tbody>
</table>
To analyze in detail the differences in the parameter \textit{area.w}, we carried out a Tukey’s HSD test, whose results are showed in Table 3.4. A positive difference denotes that the mean of the first factor of the comparison is greater than the second, while a negative difference indicates that the mean of the second is greater. The value 0.3 of \textit{area.w} is better than 0.003 and 0.0003. This result is clearly visible in Figure 8.

<table>
<thead>
<tr>
<th>\textit{area.w} pairwise comparisons</th>
<th>difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{area.w} = 0.3 vs. \textit{area.w} = 0.003</td>
<td>0.1028</td>
<td>0</td>
</tr>
<tr>
<td>\textit{area.w} = 3.0E-4 vs. \textit{area.w} = 0.003</td>
<td>-0.0350</td>
<td>0</td>
</tr>
<tr>
<td>\textit{area.w} = 3.0E-4 vs. \textit{area.w} = 0.3</td>
<td>-0.1378</td>
<td>0</td>
</tr>
</tbody>
</table>

In order to analyze MSW package we proceeded in the same way as in the previous case. Therefore, Figure 3.9 shows a boxplot of MSW peak detection parameters. First, they are grouped by \textit{amp.Th}, then by \textit{SNR}, and finally by \textit{peakScaleRange}. A visual inspection of the plotbox suggests that \textit{peakScaleRange} may not affect to the performance of the classifiers, while both \textit{amp.Th} and \textit{SNR} affect it.
3.6. Results and Discussion

The results of the corresponding ANOVA test are presented in Table 3.5. In this case, here we detected statistically significant differences for the three factors, so that we should study the interaction among them.

Table 3.5. Results of MSW ANOVA.

<table>
<thead>
<tr>
<th>Factors</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ampTh</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>snr</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>peakScaleRange</td>
<td>0.0366</td>
</tr>
<tr>
<td>Interactions</td>
<td></td>
</tr>
<tr>
<td>ampTh:snr</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>ampTh:peakScaleRange</td>
<td>4.76E-05</td>
</tr>
<tr>
<td>snr:peakScaleRange</td>
<td>0.0117</td>
</tr>
<tr>
<td>ampTh:snr:peakScaleRange</td>
<td>1.57E-10</td>
</tr>
</tbody>
</table>

To study the interaction among amp.Th, SNR and peakScaleRange, we have applied a Tukey’s HSD. Tukey’s HSD test agrees with ANOVA and confirms that parameters strongly affect each other, since there is not a clear best combination of parameters. However, these results suggest that a value 0.1 of the parameter amp.Th performs clearly worse. The best configurations (i.e. those that win in more times) are summarized in Table 3.6.

Table 3.6. MassSpecWavelet best parameter combinations.

<table>
<thead>
<tr>
<th>SNR</th>
<th>amp.Th</th>
<th>peakScaleRange</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.01</td>
<td>2</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0001</td>
<td>2</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0001</td>
<td>4</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0001</td>
<td>6</td>
</tr>
<tr>
<td>5.0</td>
<td>0.0001</td>
<td>2</td>
</tr>
</tbody>
</table>

At this time we are able to study the global interaction among classifier, POP and peak detection. By analyzing their behaviour, we expect to be able to answer the following questions: (i) what is the best peak detection algorithm (MSW vs. PROcess), (ii) what are the best POP values and (iii) what are the most suitable classifiers for MS data. In order to gain knowledge about this issues, we carried out a Tukey’s HSD test and filtered the results to the best configurations of PROcess (area.w = 0.3) and MSW (Table 6). Moreover, we also exclude those rows with POP values of 20 and 40, since these values are not usually acceptable.

The Tukey’s HSD filtered test table compares 408 different configurations. By analyzing the comparisons, we observed a group of 19 configurations that won in the 42%
3. A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification

of times, tying in the remaining cases. In these configurations, SVM-L or SVM-RBF and MSW are always present, with a variety of parameters as concluded previously. Moreover, SVM-RBF and SVM-L are present in the top 27 configurations. Figure 3.10 shows a boxplot of the overall kappa and accuracy of each classifier supporting this conclusion.

![Figure 3.10 Classifiers boxplot comparison.](image1)

Regarding the peak detection algorithm used, MSW is present in the best 71 configurations with 0% of losses and a win percentage that ranges from 22% to 42%. This fact indicates that MSW clearly outperforms PROcess. Figure 3.11 shows a boxplot of the overall kappa and accuracy measures obtained by each package.

![Figure 3.11 PROcess vs. MSW boxplot comparison.](image2)

It is also remarkable the fact that the best configurations involving PROcess also include SVM-L as classifier, supporting the conclusion that this type of classifier performs better in this case. Finally, results do not reveal significant differences among the values
of POP. Although Fig. 7 shows that most of classifiers perform worse with higher values of POP, SVMs seem not to be affected.

### 3.7 Conclusions

High-throughput mass spectrometry data analysis requires an important stage of preprocessing, since it could bias subsequent analysis such as classification or biomarker discovery. In this work, we have studied the effect of such preprocessing on MS sample classification, comparing two standard packages: MassSpecWavelet and PROcess.

First, we found that baseline correction has no effect in classification. However, this result cannot be extrapolated to other situations since the spectra in our dataset does not exhibit baseline.

Regarding peak detection, results show that MassSpecWavelet outperforms PROcess. This conclusion agrees with previous studies, where MassSpecWavelet also exhibits a better performance in terms of FDR and sensitivity [4, 6].

Although this study is mainly focused on the comparison of preprocessing techniques, we have also found that support vector machines achieved the best performance in our experiments.

In order to complement the present study, further work includes the comparison of more publicly available libraries, such as Cromwell [24] or LIMPIC [44] working with more MS datasets. However, within this study we have also established a valuable framework for extensively testing MS preprocessing techniques following a reproducible procedure.

### 3.8 References


3. A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification


3.8. References


[38] Cote RG, Reisinger F, Martens L. jmnzML, an open-source Java API for mzML, the PSI standard for MS data. Proteomics. 2010; 10(7):1332–1335.


3. A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification


Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

4.1 Abstract
Mass spectrometry is one of the most important techniques in the field of proteomics. MALDI-TOF mass spectrometry has become popular during the last decade due to its high speed and sensitivity for detecting proteins and peptides. MALDI-TOF-MS can be also used in combination with Machine Learning techniques and statistical methods for knowledge discovery. Although there are many software libraries and tools that can be combined for these kind of analysis, there is still a need for all-in-one solutions with graphical user-friendly interfaces and avoiding the need of programming skills.

Mass-Up, an open software multiplatform application for MALDI-TOF-MS knowledge discovery is herein presented. Mass-Up software allows data preprocessing, as well as subsequent analysis including (i) biomarker discovery, (ii) clustering, (iii) biclustering, (iv) three-dimensional PCA visualization and (v) classification of large sets of spectra data.

Mass-Up brings knowledge discovery within reach of MALDI-TOF-MS researchers. Mass-Up is distributed under license GPLv3 and it is open and free to all users at http://sing.ei.uvigo.es/mass-up.

4.2 Background
Mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analysers, MALDI-TOF-MS, referred to herein as MALDI, has become popular during the last decade due to its high speed and sensitivity for detecting proteins and
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

peptides. Large sets of samples are analysed quickly in one single batch. The aforementioned reasons have led to the use MALDI for the classification of large sets of samples from different sources and/or characteristics [1]. In this sense, computational tools play a key role in MALDI experiments, as they are able to preprocess raw data registered in different formats, compare them, and apply complex algorithms in order to finally extract new knowledge and useful conclusions.

Raw data generated by MALDI is usually composed of large spectra sets. Each single spectrum contains thousands of measurements entailing mass to charge (m/z) signals and intensity (i.e. \{m/z, intensity\} pairs). These spectra are usually stored using open xml-based formats such as mzXML [2], mzML [3] and PeakML [4]. In addition, several open-source libraries to handle these data formats have been developed in the last years, among which the following are noteworthy: mzMatch [4], jmzML [5], jmzReader [6], the ProteomeCommons.org IO Framework [7] and different R packages [8, 9].

The spectra generated by MALDI apparatus usually contain a high level of noisy signals, making data preprocessing a crucial task that must be carried out before subsequent analysis [10]. This preprocessing is an extensive low-level procedure able to clean raw data and identify true signals in the noisy spectra [11]. Preprocessing comprises several tasks, such as baseline correction, smoothing, normalization, peak detection and peak matching. The use of inadequate or incorrect preprocessing methods can result in a biased dataset, hindering the achievement of meaningful biological conclusions [12]. Therefore, preprocessing is a critical stage in rigorous MALDI data analysis. To accomplish the aforementioned tasks, different algorithms and tools have been developed. Most of them are publicly available as R packages [8, 13, 14], Matlab packages [15], Java libraries [16, 17] or standalone applications [18–20].

Although MALDI is commonly used to identify and characterize molecules, such as peptides or proteins, it can be also used in combination with Machine Learning (ML) techniques and statistical methods [1] to perform biomarker discovery [21, 22], automatic sample classification [23–26], and sample clustering [27, 28]. However, there are no tools devoted to performing these analyses, thus forcing researchers to use more general tools such as R, SPSS, Weka [29] or RapidMiner [30] to carry out them. This makes it necessary to include an intermediate adaptation step to convert the preprocessed MALDI data into the input format required by each tool.
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

In order to make the development of MS proteomics applications easier, some frameworks such as OpenMS [31] and ProteoWizard [32], in C++, and MsInspect [16] in Java have been published. An example of a tool developed using such frameworks is TOPP (The OpenMS Proteomics Pipeline) [33], which is based on the OpenMS framework.

In spite of the existence of such a great variety of tools and techniques for both the preprocessing and data analysis of MALDI based proteomic datasets, there is still a lack of specific tools that cover the whole process of MALDI data analysis, allowing the users to manage raw datasets, preprocess them and perform several analyses in a row, and allow the user to apply different ML and statistical techniques to analyze MALDI data. Moreover, most of the tools are intended to be used by a user with a bioinformatic profile, requiring programming skills.

This paper presents Mass-Up, an extensible open-source platform for MALDI data processing and analysis with ML and statistical techniques that has arisen from our previous experience working with MALDI data [34–36]. Mass-Up is an AIBench [35] based desktop application specifically created to perform complete analyses of MALDI data, allowing the users to: (i) import raw data from different formats (mzML, mzXML, csv); (ii) preprocess raw data; and (iii) perform different type of analyses, including supervised (e.g. biomarker discovery, predictor building, etc.) as well as unsupervised (e.g. clustering, biclustering, etc.) techniques.

The Mass-Up design is focused on two main objectives: coverage of the whole process of data analysis and simplicity of use. The first objective is accomplished in the way Mass-Up covers the whole process of MALDI data analysis, from data preprocessing to different types of analysis. The second is achieved through a design that allows Mass-Up to be used in a straightforward manner by non-informatician users. In addition, Mass-Up is multiplatform, open source and designed using a pluggable architecture which makes it easier for programmers to develop and include new algorithms and analysis tools.

4.3 Implementation

Mass-Up is a computer application for managing, preprocessing and analyzing MALDI data. Mass-Up is implemented in Java and it was constructed using the AIBench framework, which has been demonstrated to be suitable for developing proteomics
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

applications [36], as it is the base framework of previously developed MS applications [37, 38]. Currently, Mass-Up has distributions for Windows and Linux operative systems.

This section briefly describes the Mass-Up workflow and the main algorithms and third-party libraries employed in each Mass-Up task.

4.3.1 Mass-Up workflow

Mass-Up includes a serie of operations that can be classified into (i) input/output operations, (ii) preprocessing operations, and (iii) analysis operations. Figure 4.1 depicts the Mass-Up main workflow, where the most important operations are represented, along with the input files and data types managed by the application.

![Mass-Up workflow](image)

Figure 4.1 Mass-Up main workflow operations and datatypes. Different colors have been used to identify input/output operations (green), preprocessing operations (orange), analysis operations (blue), and datatypes (red).

4.3.2 Third-party libraries

With the main goal of covering the whole process of MALDI data analysis, Mass-Up integrates several open source third-party libraries in order to accomplish different tasks, such as reading different MS data formats, preprocessing spectra, applying ML techniques, or visualizing data, among others. Table S1 of the Additional File 1 shows a
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

general overview of the Mass-Up, including the algorithms and libraries used by each operation. All of these libraries has been transparently integrated into Mass-Up so that final users does not have to install them manually, since they are built-in in each Mass-Up distribution.

Mass-Up uses jmzReader 1.2.0 [6] in order to read the mzXML and mzML MS data formats. To visualize MS spectra and to display quality control charts, Mass-Up uses JFreeChart 1.0.13, an open source Java library.

Mass-Up integrates two R packages for raw MALDI data preprocessing: MALDIquant [8] and MassSpecWavelet [13]. In addition, custom implementation of a fast peak matching algorithm based on a forward sliding window, named Forward, is also incorporated. Similarly to the alignment algorithm proposed by Kazmi et al. [39], this algorithm iterates the peaks from minimum to maximum m/z, adding them to the last cluster created if their m/z is within a distance from the average m/z of the cluster or creating a new cluster if not. This clustering algorithm does not allows clusters with two peaks from the same spectrum. In such case, only the peak that minimizes the average m/z of the cluster is kept.

Mass-Up makes use of Weka [29], a collection of ML algorithms for data mining tasks implemented in Java. These algorithms are used for classification and for principal component analysis (PCA). Three-dimensional PCAs are rendered by using Jzy3d [40], an open source Java library which can easily draw three dimensional scientific data. Clustering is executed using a custom implementation of an agglomerative hierarchical clustering algorithm and is rendered using an adapted version of JTreeview [41]. Biclustering is performed with Bimax [42], a powerful algorithm capable of generating all optimal biclusters, and BiBit [43], a novel approach for the extraction of biclusters from binary datasets that can obtain similar results to Bimax by using significantly less computation time and reducing the total number of generated biclusters. The aforementioned software, as well as a biclusters viewer, is integrated through the adaptation available in BiMS [44].

4.4 Results and discussion

Mass-Up is a flexible tool that includes several operations whose application depends on the analysis objectives. Therefore, there is no single way to use Mass-Up, and researchers
must determine which analyses apply in their studies. In this section, several practical applications of the Mass-Up operations are presented, in order to demonstrate its usefulness and applicability.

### 4.4.1 Sample Datasets

Two datasets from previous studies were selected to illustrate the Mass-Up functionality. A brief description of the main characteristics of both datasets is given in this section.

#### 4.4.1.1 Cancer dataset

R. López-Cortés *et al.* [45] propose the use of gold-nanoparticles to separate the proteins and peptides in human serum as a way to improve MALDI-based sample profiling. The protocol described in this work divides each sample into two sub-samples: pellet and supernatant. The MALDI spectra of both sub-samples are grouped by their corresponding conditions using three-dimensional PCA. The dataset is composed of sera from 5 patients with lymphoma, sera from 5 patients with myeloma, and sera from 2 healthy donors. As the classifications using pellet or supernatant are similar, only the latter sub-samples are used in the present work.

#### 4.4.1.2 Wine dataset

Nunes *et al.* [46] propose a fast MALDI-based methodology to identify different types of wines. The authors carry out a preliminary study with 5 wines of different denominations of origin, in order to identify the most appropriate MALDI matrix. The study of the matrices found that CHCA is the most suitable for the purpose of classification. Each wine was spotted five times (i.e. 25 samples in total). Those 25 samples corresponding to the use of CHCA matrix are used as proof of concept.

### 4.4.2 Preprocessing

As previously stated, the preprocessing of MS data is a critical stage that converts raw data into a suitable input for further analysis. Inadequate or incorrect preprocessing methods can result in biased datasets, hindering the achievement of meaningful biological conclusions [12]. Preprocessing is essential since raw data contains both m/z values belonging to analytes, as well as m/z values derived from several forms of noise (e.g. chemical, electronic factors, etc.). The main objectives of preprocessing are [47] to remove noise without discarding any of the m/z values of interest, and to determine the
m/z and intensity values with the best accuracy. The most common preprocessing tasks (shown in orange in Figure 4.1) are smoothing, baseline correction, normalization, peak detection, and peak matching.

Usually, spectra are jagged, making it difficult to detect the m/z values of interest from the noise [47]. Thus, smoothing algorithms are usually applied to soften the spectra. The simplest techniques are based on the use of a sliding window, where the intensity of each m/z value is adjusted based on the intensity of the neighbor m/z values. Commonly used filters are moving average, Savitzky-Golay, Gaussian and the Kaiser window. Mass-Up provides two smoothing methods: moving average window and Savitzky-Golay, both from the MALDIquant library [8].

Baseline is a specific form of noise mainly driven by chemical perturbations, defined as an offset of the intensities of peaks that often show a dependency on the m/z value such that it is highest at low m/z values, presenting an exponential decay towards higher masses [47]. The most common baseline correction methods are monotone minimum, linear interpolation, LOESS, moving average of minima and continuous wavelet transform, all of which are available as free software in different packages such as Cromwell [15] (Matlab), PROCess [14] (R), MALDIquant [8] (R) or SpecAlign [19] (Java). Mass-Up allows the user to make use of all the baseline correction methods provided by MALDIquant (i.e. Top Hat, SNIP, Convex Hull, and Median).

A major constraint of MALDI is that the intensity of the m/z values is relative and can vary among spots of the same sample. For this reason, normalization is typically used, making the intensities of different spectra comparable. The most common normalization methods are Total Ion Current (TIC), Probabilistic Quotient Normalization (PQN), Z-score, Linear, Mean or Median. Mass-Up allows the user to perform normalization using TIC, PQN or Median, all provided by the MALDIquant library [8].

The m/z detection can be defined as the process of selecting values of interest (i.e. related with target analytes) from a given spectrum, and it is normally applied after baseline correction and smoothing. Most of the peak detection methods are based on setting a threshold value in order to discard low intensity m/z values. The threshold can be absolute (e.g. minimum intensity) or relative (e.g. signal-to-noise ratio, SNR). However, Du et al. [13] proposed a method that performs m/z detection without explicit smoothing and baseline correction. This method is based on the continuous wavelet
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

transform (CWT) and is publicly available in the MassSpecWavelet package. Mass-Up includes two m/z selection methods: the CWT-based method implemented in MassSpecWavelet [13], and a SNR-based method provided by MALDIquant [8], which uses a sliding window.

Finally, m/z matching is needed in order to make different spectra comparable. Without this matching procedure, the same molecule or metabolite (e.g. a certain peptide) can have different m/z values across replicates or samples. The objective of m/z matching methods is to find a common set of m/z locations in several spectra, so that all spectra will have the same m/z values for the same biological entities. In Mass-Up there are two fundamental types of m/z matching: intra-sample and inter-sample. The intra-sample matching is applied to the spectra obtained for the replicates of the same sample, while the inter-sample matching is applied to match m/z values across different samples, making them comparable and suitable for the subsequent analysis stage. Peak matching algorithms, are classified into two main groups: sequential algorithms based on a sliding window (e.g. the Forward algorithm, available in Mass-Up) and clustering based approaches (e.g. the MALDIquant algorithm [8], also available in Mass-Up).

The Mass-Up workflow also incorporates an additional filtering step that is very closely related to the matching process. This step is performed after the intra-sample matching and before the inter-sample matching, and allows the creation of a consensus spectrum for a sample, which summarizes the replicates of a sample in one single spectrum. In this step, the Percentage of Presence (POP) parameter allows the user to set the number of replicates where an m/z value must be present in order to be considered a valid consensus m/z value.

Finally, it is important to note that, while smoothing, baseline correction, normalization, and m/z detection are applied individually to each single spectrum in the Preprocess data operation, the m/z matching is applied to several spectra at the same time and is carried out by using the Match Peaks operation.

The new data generated by the Preprocess data and Match Peaks operations can be exported as comma-separated value files, allowing users to load them later with Mass-Up or to analyze them with other software packages. Mass-Up documentation includes information about exporting data and examples describing how it can be loaded in other languages such as R.
4.4.3 Quality Control

When working with MALDI, low quality spectra may occasionally be generated. For example, spectra showing a low number of m/z values in comparison with other spectra, or containing many unique m/z values not present in their sibling replicates. These spectra may lead to failure when carrying out an analysis, or to incorrect conclusions. To prevent such a scenario, a quality control (QC) step was included, which may be performed between the preprocessing and the analysis tasks. The QC can be done at two levels: replicates, a low level QC analysis focused on the replicates of each sample; and samples, a high level QC analysis with additional information from the intra-sample m/z matching process.

At the replicates level, the user can check basic information about each individual spectrum (i.e. peak count, m/z range, intensity ranges, etc.) and compare all spectra in the dataset. Figure 4.2A shows a replicate QC analysis applied to the samples from conditions A, B, C, D, and E of the Wine dataset previously described. As can be noted from the boxplot, there are two outliers (red circles) and one extreme outlier (red triangle) in the Masses count chart. Specifically, the QC analysis has marked the E-CHCA.3-4, A-CHCA.1-4 and E-CHCA.2-2 samples as outliers due to the number of m/z values of their spectra. Therefore, before continuing with further analysis, it is recommended to carefully revise these samples and even to repeat their analysis.
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

Figure 4.2 Details of the quality control analysis views for (A) replicates, and (B) samples. Box plot charts are used to summarize the more detailed information presented in the tables.

At the samples level, the user can check the performance of the intra-sample peak matching process, by comparing the percentages of presence (POP) counts (globally and by conditions) and the POPs of each sample. As previously stated, the data table is more detailed and contains additional information from the intra-sample peak matching process, specifically: (i) POPXX columns, where XX is a percentage of the number of spectra, which show the number of peaks with a POP value exactly equal to XX; (ii) Align. Masses column, which shows the number of masses that have been matched across the spectra in the sample; (iii) Split >= XX columns, which show the percentage of masses that have a POP value higher or equal to XX; and (iv) Count >= XX columns, which show the number of masses that have a POP value higher or equal to XX (these are the columns used as categories in the charts).

Figure 4.2B shows a samples QC applied to the same samples as in the previous example. In this case, the box plot corresponds to the global POP count and shows that
there are two outliers for the category “Count >= 60” and one outlier for the category “Count >= 100”. Again, the outliers are highlighted in bold in the table.

4.4.4 Biomarker discovery

One of the main purposes of the MS analyses is the biomarker discovery [21, 22, 48]. A biomarker is a peptide, protein or other element of a sample that can identify and differentiate certain conditions such as phenotypes, strains, diseases or infections.

When identifying new biomarkers, it is necessary to distinguish between two types of data sets that can be analyzed: (i) those cases where there are a known and well defined number of conditions (e.g. healthy vs. diseased, different stages of a disease, etc.), and (ii) those cases where there are no conditions or where they are not clearly defined. In accordance with this differentiation, Mass-Up provides two types of biomarker discovery analysis: (i) the inter-label analysis, for the former type of data, and (ii) the intra-label analysis, for the latter.

In the inter-label analysis, the user can perform the appropriate statistic tests to identify those peaks that can be potential biomarkers to differentiate the conditions. Four different tests of independence were included in Mass-Up following the recommendations given by McDonald [49], where tests are chosen depending on the number of samples and conditions of the dataset, as shown in Table 4.1. Taking into account that the number of samples in MALDI experiments is generally below 1000, the Fisher’s exact test and the randomization test are the tests more commonly applied. As each test is performed independently for each m/z value, the Benjamini-Hochberg FDR correction is applied to take into account the number of m/z values analyzed and reduce the number of false positives.

<table>
<thead>
<tr>
<th></th>
<th>&lt;= 1000 samples</th>
<th>&gt; 1000 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 conditions</td>
<td>Fisher’s exact test</td>
<td>Yates’ chi-square test</td>
</tr>
<tr>
<td>&gt;2 conditions</td>
<td>Randomization test</td>
<td>Chi-square test</td>
</tr>
</tbody>
</table>

By using the inter-label analysis in the Wine dataset (shown in Figure 4.3A), we can analyze all the samples of conditions A, B, C, D, and E. In this case, the randomization test is applied in order to identify statistically relevant m/z values, as the number of samples is lower than 1000 (5 samples for each of the 5 wines for a total of 25 samples) and the number of conditions is higher than 2 (5 wine denominations). The first three columns contain the m/z value, the p-value, and the q-value respectively; while the other
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

Columns show in which samples the m/z values are present. As can be seen, the peaks with a q-value < 0.05 are clear candidates to be biomarkers as they differentiate certain conditions from others.

Figure 4.3 (A) Inter-label biomarker discovery view. Depending on the number of samples and conditions, Mass-Up automatically selects the appropriate statistical test to apply. (B) Intra-label biomarker discovery view. Filters are configured to select only the m/z values present in the MA samples and absent in the other samples.

In the intra-label analysis, the user can identify those m/z values that are representative of one or more samples, in a more exploratory fashion. In this scenario, it is possible to identify the biomarkers of a specific sample or discover groups of samples.
with a similar profile that may, therefore, be related. This analysis is particularly useful, for example, when working with different strains of the same bacteria and the user wants to identify those peaks that are unique for a certain strain.

By using the intra-label analysis in the Cancer dataset (shown Figure 4.3B), we can analyze the samples of the condition Myeloma and configure the analysis to identify those peaks present in the "MA" sample (i.e. Myeloma A) and not present in the rest of the samples. The identification of these peaks may be useful, for example, to explain the abnormal behaviour of a sample when compared to other samples from the same condition. Specifically, the Intra-label Biomarker Discovery view shows how we are looking for specific peaks of the sample MA (i.e. Myeloma A), that is, peaks that are in this sample but not in the others.

4.4.5 Principal component analysis

PCA is a mathematical procedure that uses orthogonal transformation to convert a set of observations (i.e. samples) of possibly correlated variables (i.e. m/z values) into a set of values of linearly uncorrelated variables called principal components (PC), whose dimensionality is expected to be lower than the dimensionality of the original data set.

Once the PC are calculated, they can be used to represent the samples in a 3-dimensional space. By assigning a different color to each condition’s samples, users can visually identify if there is a separation between conditions. If such were the case, then the conditions would be distinguishable. The PCA view also includes additional information about the PCA, such as the eigenvectors and their corresponding eigenvalues and retained variances, for a better results interpretation.

As previously stated, López-Cortés et al. [45] demonstrate that the spectra of supernatant sub-samples of the Cancer dataset can be grouped by their corresponding conditions using PCA. Figure 4.4A shows the result of applying PCA to this set of samples in Mass-Up. As it can be clearly seen, the three conditions are separable in the 3-dimensional space.
Figure 4.4 (A) Principal component analysis view presenting three different clusters, one for each condition. (B) Detail of the hierarchical clustering visualization using JTreeView. The upper dendrogram automatically colors the tree branches that only include samples from the same condition, while the side dendrogram groups the more similar m/z values. (C) Class-biclusters of the Cancer dataset extracted with the Mass-Up Biclustering Viewer. Purple rectangles denote the existence of biclusters associated with one condition.

4.4.6 Cluster analysis

Cluster analysis allows finding groups of similar spectra among all the samples being studied. In the case of unlabeled data, it allows discovering hidden or previously unknown subgroups of samples. In the case of labeled data, it allows the user to check if the different conditions present in a dataset are separable by means of the m/z values of each sample.
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

Mass-Up incorporates a hierarchical clustering algorithm for the construction of a hierarchy of sample groups (named clusters). The algorithm included is agglomerative and follows a bottom-up approach, meaning that it is constructed iteratively, starting with each sample in its own cluster, and merging the closest pair of clusters on each step. In order to decide which clusters should be merged, a measure of dissimilarity between clusters is required. In our case, this is achieved by using a distance metric, which measures the distance between two samples, and a linkage criterion, which specifies the dissimilarity of clusters. Mass-Up includes the Euclidean and Hamming distances as distance metrics, and the complete, single and average functions as linkage criteria. The results of a hierarchical clustering are usually presented in a dendrogram.

An important aspect when performing a cluster analysis in Mass-Up is that the user can decide whether to use intensities (i.e. a m/z value is represented by the value of its peak intensity) or not (i.e. a m/z value is represented by its peak presence or absence). The Euclidean distance is the most suitable when using intensities while the Hamming distance is the most appropriate when using presence/absence of peaks.

In each cluster analysis, two hierarchical clusterings are constructed: one for the samples and one for the m/z values. For the visualization of the results, Mass-Up incorporates an adapted version of JTreeView, a software for the visualization and analysis of gene expression data. We have adapted it to MS, so that in our specific case the rows represent peaks instead of genes, while columns still represent samples. This representation also includes a heat map, which is combined with two dendrograms that represent the aforementioned hierarchical clusterings. The individual values contained in the heat map matrix are displayed as colors and they can represent (i) the intensity level of the corresponding peak (red if the peak has an intensity value of 1; green if the peak has an intensity of 0; and intermediate colors for intensities between 0 and 1), or (ii) the presence or absence of the peak (red if the peak is present and green if the peak is not present). It is important to note that to achieve a correct representation using intensities, the m/z values must be scaled between 0 and 1 during the raw data preprocessing.

Figure 4.4B shows the results of applying hierarchical clustering to the Cancer dataset used as proof-of-concept. As the dendrogram illustrates, the three conditions are well separated since the samples of each condition can be grouped together.

Finally, it is worth noting that the cluster analysis can be used with a list of previously selected peaks. This way, the cluster analysis will be focused on analyzing
only these peaks. This list can be obtained by exporting the biomarkers identified in the inter-label analysis. In such a situation, this feature is useful to qualitatively verify if a list of potential biomarkers is enough to separate or differentiate between the conditions of study.

4.4.7 Bicluster analysis

Although biclustering techniques have been successfully used with gene expression data for over a decade, it is only very recently that those techniques have been applied to MS data [50]. Biclustering is a data mining technique that allows simultaneous clustering of the rows and columns of a matrix. It has been successfully applied to analyze microarray data due to their ability to discover co-expressed genes under certain samples [51]. In contrast to traditional clustering techniques, where each gene in a given cluster is defined under all the samples, biclustering algorithms propose groups of genes that show similar activity patterns under a subset of the experimental samples.

In previous studies, we have proposed a novel workflow for the application of biclustering to MALDI data. In addition, the adequacy of applying biclustering to analyze such data by comparing biclustering and hierarchical clustering over two real datasets has also been evaluated [44]. Biclustering has shown the ability to discover groups of samples that are similar but only in a subset of m/z values, which represent a new kind of hidden hypothesis that are difficult to be discovered by classic clustering algorithms, such as hierarchical clustering, which are based on a global comparison of samples including all m/z values.

The biclustering algorithms selected in the study and included in Mass-Up (i.e. Bimax and BiBit) use a binary dataset as input where 1 represents a peak presence, and 0 represents a peak absence. These algorithms will look for groups (i.e. biclusters) of 1’s, that we call presence patterns. Nevertheless, in certain cases, it can be desirable to extract other type of patterns, such as absence patterns (i.e. biclusters of 0’s) or simple presence/absence patterns (i.e. biclusters of 1’s and 0’s in one direction). López-Fernández et al. [44] further discuss how to prepare an input MALDI dataset into a suitable form to look for these three types of patterns.

Mass-Up provides an operation to apply this technique to both labeled and unlabeled samples. The user has to select the biclustering algorithm to use, the type of pattern and the biclustering mode (i.e. whether rows of the biclustering binary matrix are...
peaks or samples). In addition, the user can also establish the minimum dimensions of the output biclusters. If the input data is labeled, the user can also indicate whether the output of the biclustering must be filtered in order to only retrieve those biclusters where most of the samples belong to the same condition or label, known as class-biclusters. After performing a biclustering analysis, results can be inspected in the biclustering viewer, an intuitive view that shows a list of the generated biclusters as well as a heat map. If a bicluster is selected, it will be highlighted in the heat map, which is automatically rearranged in order to show the bicluster in the upper left corner.

In order to demonstrate the usefulness of this module, we considered the Cancer dataset used in previous sections, and applied biclustering by means of the BiBit algorithm in the hope of finding presence class-biclusters. Figure 4.4C shows one presence class-bicluster for each class, where each column represents one m/z value and each row represents a sample. As shown, each class bicluster includes a group of m/z values with the same pattern of presence in the samples of one condition, and a variable pattern of presence in the rest of the samples. When using a presence class-bicluster, only presence is taken into account to create the class-bicluster, whereas when using a presence/absence class-bicluster, the absence is also taken into account.

### 4.4.8 Classification analysis

Sample classification is the ability to predict the label of a sample given a training set of labelled samples, therefore, the capacity of producing a diagnosis machine [10, 24, 26]. Through the “Classification Analysis” operation, the user can determine which classifier performs best for the data under analysis. This operation provides an interface adapted from the Weka software that allows the user to select and to configure a classifier, and to evaluate its performance by means of a cross-validation scheme. The output log of the evaluation process summarizes the performance of the classifier using different statistical measurements, such as accuracy, kappa, precision, recall, etc. In addition, you can make a ROC analysis per condition.

Classification analyses are performed in the classification view (shown in Figure 4.5), which was adapted from the Weka software. Through this view, the user can select a classifier and a validation scheme (i.e. cross-validation or percentage split) to perform an evaluation. As shown, the results report includes several global and per-class statistics, as well as the resulting confusion matrix. Using these operations, users can assess whether
the data being analyzed is suitable for classification, as well as determine which classification algorithm is best.

Figure 4.5 Classification analysis view presenting the result of executing a Bayes Net classifier using a 10-fold cross validation scheme. The resulting confusion matrix is presented along with several statistical measurements. ROC curve corresponding to condition C of the Wine dataset is also showed.

4.4.9 Performance notes

Although the performance is very dependent on the number of samples and the computer being used, some tests has been carried out in order to provide some performance guidelines. We have created a test dataset of 490 samples based on the Wine dataset, and then, we have executed the most common workflow of Mass-Up under an Intel Core i5 M520 with 8GB of RAM and Kubuntu 13.10 as OS. It is important to note that the size of this test dataset clearly exceeds the common size of a dataset in a MALDI-TOF MS experiment, which usually are no longer than 200 samples.

It took about 90 seconds to load 490 raw samples and about 200 seconds to fully preprocess them. Once the data is preprocessed and prior to perform any analysis, we must apply the Match Peaks operation, which could be executed in less than 30 seconds using the MALDIquant algorithm and in less than 3 seconds using the Forward algorithm. Most of the analyses (quality control, PCA, classification and intra-label analyses) could be executed in less than 5 seconds, while clustering, biclustering and inter-label analysis took more time. On one hand, clustering analysis took less than 20 seconds and the
biclustering execution time depends on the algorithm selected (less than 20 seconds for Bibit and about 15 minutes for Bimax). On the other hand, inter-label biomarker discovery based on 10000 randomizations took about 8 minutes.

4.5 Conclusions

In this paper we have presented Mass-Up, a new software for the analysis of MALDI data. This is an application that covers the whole process of MALDI data analysis, from data preprocessing to complex data analyses.

Mass-Up incorporates the most common analyses, aside from protein identification and focusing in biomarker discovery, such as statistical tests-based biomarker discovery, clustering, PCA, and classification. In addition, other less common analyses such as quality control and biclustering are also included. Therefore, Mass-Up provides users with a wide range of tools to analyze and explore their MALDI data.

Unlike other MS tools, Mass-Up provides a friendly graphical user interface designed to avoid the need for a bioinformatics expert to use it. The tutorial and examples included in Mass-Up tool and in the project homepage will guide users through the different operations included, making it use suitable for any user.

Finally, Mass-Up is open to further extension, such as including new operations or improving the available ones.

4.6 Availability and requirements

The Mass-Up software is freely available from the project homepage on http://sing.ei.uvigo.es/mass-up. Additionally, source code can be downloaded from https://sourceforge.net/projects/mass-up/.

Project name: Mass-Up.

Project homepage: http://sing.ei.uvigo.es/mass-up

Operating system: Platform independent, packaged for Windows and Linux.

Programming language: Java version 7.

Other requirements: Mass-Up has no other requirements since distributions are self-contained.
Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

License: Version 3 of the GNU General Public License (GPLv3).

4.7 References


Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery


Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry
to knowledge discovery


Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery


Appendix

Steroids 78/12, pp. 1226-1232;
doi:10.1016/j.steroids.2013.08.14
Speeding up the screening of steroids in urine: Development of a user-friendly library

M. Galesio a, H. López-Fdez a, C. Reboiro-Jato a, S. Gómez-Meire a, D. Glez-Peña a, F. Fdez-Riverola a, C. Lodeiro b, c, M.E. Diniz a, b, J.L. Capelo a, c

a REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Montijo de Caparica, Portugal
b BIOCOP Group, REQUIMTE/QTOF, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Montijo de Caparica, Portugal
c INIC Group, Informatics Department, Higher Technical School of Computer Engineering, University of Vigo, Ourense Campus, E-32004 Spain

ARTICLE INFO

Article history:
Received 3 June 2013
Received in revised form 14 August 2013
Accepted 23 August 2013
Available online 11 September 2013

Keywords:
Androgenic anabolic steroids
MALDI-TOF-MS/MS
MLibrary software
Bioinformatics
Anti-doping

ABSTRACT

This work presents a novel database search engine – MLibrary – designed to assist the user in the detection and identification of androgenic anabolic steroids (AAS) and its metabolites by matrix assisted laser desorption/ionization (MALDI) and mass spectrometry-based strategies. The detection of the AAS in the samples was accomplished by searching (i) the mass spectrometric (MS) spectra against the library developed to identify possible positives and (ii) by comparison of the tandem mass spectrometric (MS/MS) spectra produced after fragmentation of the possible positives with a complete set of spectra that have previously been assigned to the software. The urinary screening for anabolic agents plays a major role in anti-doping laboratories as they represent the most abused drug class in sports. With the help of the MLibrary software application, the use of MALDI techniques for dope control is simplified and the time for evaluation and interpretation of the results is reduced. To do so, the search engine takes as input several MALDI-TOF-MS and MALDI-TOF-MS/MS spectra. It aids the researcher in an automatic mode by identifying possible positives in a single MS analysis and then confirming their presence in tandem MS analysis by comparing the experimental tandem mass spectrometric data with the database. Furthermore, the search engine can, potentially, be further expanded to other compounds in addition to AASs. The applicability of the MLibrary tool is shown through the analysis of spiked urine samples.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

The use of androgenic anabolic steroids (AAS) and hormones to enhance athletic performance has important health and social implications. Their use was first introduced in sports as agents supporting the athlete recuperation after extreme stress and fatigue, but rapidly became the main agents used in doping abuse [1].

Nowadays, this class of drugs is a major group included in the prohibited list of the world anti-doping agency (WADA) as well as of major sports authorities [2–5]. In the WADA statistic report for 2011, the AAS represented 59.4% of all adverse analytical findings reported by WADA accredited laboratories [6]. Although this data may not reflect the real doping abuse statistical status, because of the well-known problems in the detectability of clandestinely designed AAS, micro dosages of endogenous AAS and “modern” doping agents (e.g., peptide hormones) [1,7–9].

The use of AAS to increase muscle mass and strength is not a behaviour strictly related to elite athletes, as their use is increasing amongst amateur athletes as well as outside sports as an expression of an improved life style [10,11]. The illicit AAS use is an increasing trend in western societies and the emergent AAS dependence is a matter of growing public health concern [12].

Quickly following the development of mass spectrometry (MS) detectors, its use coupled to gas chromatography (GC) has become the standard technique for AAS control. Currently, most methods for routine detection of these compounds and their metabolites, comprising both screening and confirmatory analysis, are still based in GC–MS techniques [13–17]. More recently, due to the increasing complexity of doping analyses and in order to enhance the detection of this group, liquid chromatography coupled with MS/MS is gaining ground within anti-doping laboratories [18–20]. In particular for the detection of thermo-labile and polar steroids, such as Trenbolone. It avoids the derivatisation step required by GC–MS and provides good sensitivity for the determination of these compounds [21,22]. Moreover, the low throughput provided by LC–MS techniques are to some extent being surpassed by the introduction of uHPLC systems coupled with LC-columns containing solid core particles that allows high speed and high efficiency
separations. Unfortunately, for the majority of the AAS, due to their poor ionisation efficiencies, their determination by LC–MS may lead to losses in sensitivity.

The combination of these two factors, the long separation times of gas chromatographic techniques and the increasing workloads within anti-doping laboratories, expose an urgent need for an analytical technique allowing simplicity, speed and high throughput for the screening of the huge number of banned compounds, particularly the AAS.

Recently, the use of matrix-assisted laser desorption/ionization (MALDI) for the analysis of small molecules, has grown as a potential technique, which is reflected by the increasing number of studies reported in literature [23–27]. Moreover, it appears extremely promising for high-throughput, which is a major demand for future anti-doping methods.

In light of the latest technological improvements of this analytical technique we have recently study the applicability of a wide variety of commercial MALDI matrices for the rapid screening of AAS [28]. The matrix 2-(4-hydroxyphenylazobenzene acid (HABA) was found to be the most robust for the analysis of anabolic steroids after a derivatisation step with the reagent Girard T hydrazine. The Girard T hydrazine derivative produced after derivatisation is a quaternary ammonium ion that originates a strong [M+H]+ ion signal in the MALDI mass spectrum, as a result it increases the intensity of the steroid signal. In the aforementioned work it was demonstrated that positive identification of the characteristic peaks for all the compounds studied is possible for a sample concentration of 10 ng/mL in the MALDI sample plate. The sensitivity achieved with the HABA matrix after derivatisation was similar to that achieved by GC/MS − around 4–10 ng/mL in the single ion monitoring mode.

In the present work it is presented a step forward in simplifying AAS control through the use of easy sample treatment and friendly software. The software is freely source code available, and it can be run as a multiple platform. As a proof-of-concept, the rapid screening of AAS in urine is reported using a sample treatment previously published by our team [28].

2. Material and methods

2.1. Chemicals

Standards of 17α-methyltestosterone and nandrolone were purchased from Riedel-de Haën (Seelze, Germany). The standards 17α-trenbolone, 2α-methyl-5α-androstane-3α-ol-17-one, mestanolone, methandienone, calusterone, fluoroxymesterone, ethisterone and mibolerone were kindly provided by the Portuguese National Anti-doping Laboratory and the Italian National Anti-doping Laboratory. A solution of β-glucuronidase from Escherichia coli K12 with a specific activity approximately of 140 U/mg at 37 °C and pH 7 with nitrophenyl-β-D-glucuronidase as substrate (1 mL contains at least 140 U) was purchased from Roche Diagnostic (Mannheim, Germany). Sodium hydrogen phosphate, sodium phosphate dibasic, tert-butylmethyl ether, methanol (MeOH), acetoniitrite (ACN) and the derivatisation reagent, Girard T (GT) hydrazine, used for sample and matrix preparation were purchased from Sigma (Steinheim, Germany); glacial acetic acid (>99.5%), matrix; zetGICA and HABA were purchased from Huka (Buchs, Switzerland); trifluoroacetic acid (TFA, 99%) was from Riedel-de Haën.

Urine samples used in this work were obtained from healthy volunteers from the research team. Volunteer’s age ranged between 22–30 years, including both male and female. The research ethical committee from the Science Faculty of Ourense approved the study protocol and all the volunteers gave their consent.

2.2. Apparatus

A model UNIVAPO 100H vacuum concentrator centrifuge (Uni-Equip, Martinsried, Germany) with a model Unijet II refrigerated aspirator vacuum pump (Unilab) was used for (i) sample drying and (ii) sample pre-concentration. A Spectralfuge-mini minicentrifuge (Labnet, Madrid, Spain) and a Sky Line minicentrifuge-vortex (ELMI, Riga, Latvia) were used throughout the sample treatment, when necessary. A Simplicity 185 system (Millipore, Milan, Italy) was used to obtain Milli-Q ultrapure water throughout all the experiments. The derivatisation procedure was performed in a 1.5 mL microtube flat cap from Delta Lab (Barcelona, Spain). Separation of the steroid Girard T (GT) hydrazones from the unreacted Girard T reagent was carried out in a 2 mL empty reversible solid-phase extraction (SPE) cartridge from Supelco (Bellefonte, PA, USA) packed with a preparative C18 resin (125 Å, 55–105 μm; Waters, Barcelona, Spain).

2.3. Sample preparation

2.3.1. Standard solutions

Individual stock standard solutions of each compound (500 mg/mL) were prepared by weighing 0.0125 g of analyte in a 25 mL volumetric flask and making it to volume with methanol. These standard solutions were stored in the dark at −20 °C. Working standard solutions were prepared by dilution of the stock standard solutions in the appropriate volume of methanol.

2.3.2. Urine hydrolysis procedure

Urine samples (2 mL) were hydrolysed with 50 μL of the commercial solution of β-glucuronidase, after the addition of 0.750 mL of phosphate buffer (0.8 M, pH 7). The hydrolysis was performed at 55 °C during 60 min.

2.3.3. Liquid–liquid extraction of target analytes

After cooling to room temperature, 0.5 mL of carbonate buffer (pH 9) was added to alkalize the hydrolyzed solution. Liquid–liquid extraction was carried out by agitation with 5 mL of tert-butyl methyl ether for around 10 min; after centrifugation, the organic phase layer was transferred to a new vessel.

2.3.4. Derivatisation procedure

The procedure for derivatisation with Girard T hydrazine was performed based on the protocol described by Wheeler [29], as follows: The collected organic phase was dried under a gentle nitrogen stream at 40 °C. After the addition of 500 μL of a methanolic solution with 10% glacial acetic acid and 4 mg of Girard T hydrazine, the vial was closed and the derivatisation reaction was then performed at 60 °C during 30 min. After cooling, the solution was evaporated to dryness in a vacuum concentrator centrifuge and then reconstituted with 1 mL of methanol/water (10:90, v/v).

2.3.5. SPE clean-up

After derivatisation, the steroid GT hydrazones were separated from un-reacted GT hydrazine reagent by SPE in a C18 cartridge, according to the protocols described by Khan et al. [24] and Griffiths et al. [25]. Briefly, before use, the cartridges were conditioned with 5 mL of methanol plus 10 mL of MilliQ-water without allowing the cartridges to dry out. After loading the sample, the cartridge was washed with 2 mL of methanol/water (10:90, v/v) in order to remove impurities from the cartridge and, finally, the steroid GT hydrazones were eluted from the cartridge with 1 mL of methanol.
2.4. Experimental design

To assess the use of the MLibrary software for the analysis of AAS present in human urine, five urine samples were spiked with different AAS at different concentration levels. The five urine samples were designated as Urine 1, Urine 2, Urine 3, Urine 4 and Urine 5. Urine 1 was spiked with 17α-methyltestosterone (100 ng/mL), Urine 2 was spiked with calusterone (250 ng/mL), Urine 3 was spiked with nandrolone (10 ng/mL), Urine 4 was spiked with fluoromesterone (200 ng/mL) and ethisterone (150 ng/mL). Urine 5 was spiked with 17α-trenbolone (25 ng/mL) and mesterolone (300 ng/mL).

2.5. MALDI-TOF-MS and MALDI-TOF-TOF-MS analysis

The mass spectrometric analyses were performed on Applied Biosystems 4700 Proteome Analyzer with TOF/TOF™ Optics system (Applied Biosystems, Foster City, CA, USA) equipped with a and a diode pumped Nd:YAG laser with 200 Hz repetition rate. The instrument was operated for detection in positive ion reflection mode. The MS spectrum for each sample was based on the average of 1000 laser shots; for the MS/MS up to 4000 shots were accumulated. MS/MS mode was operated with 1 kV collision energy; air was used as the collision gas such that nominally single collision conditions were achieved. For MS analysis, laser desorbed ions were accelerated from the source at 20 kV. For MS/MS analysis, ions were accelerated from the source at 80.0 kV. Both modes employed delayed ion extraction for improved ion focusing. The MS/MS data was acquired using the instrument default calibration. All the MS data was acquired on the basis of the HABA matrix used for internal calibration. Prior to MALDI analysis, the sample was mixed with an equal volume of the MALDI matrix solution and homogenised in a vortex instrument. The HABA MALDI matrix used in this work was prepared according to the developed method, 0.52 mg in 1 mL of a solution of ACN/H2O/MeOH (40/40/20, v/v/v) [28]. The matrix CHCA was prepared by dissolving 10 mg in 1 mL of a solution ACN/H2O/TFA (50/49.9/0.1, v/v/v). An aliquot of the sample/matrix solution (0.5 μL) was hand-spotted onto the MALDI sample plate and the sample was allowed to dry. Fig. 1 presents a scheme of the sample treatment followed in this study.

2.6. MLibrary software

Current versions of the software and their supporting user manuals are freely available for download and use, without restriction, via Internet at http://sins.ei.uvigo.es/MLibrary. The program was developed based on previous work on a tool for accurate protein quantification [30]. MLibrary operates on comma-separated-values (CSV) files with centroid mass and relative intensity data extracted from the instrument software (Data ExplorerTM Software, version 4.5). This data can be analysed and compared with the compound data stored in the MLibrary repository, which contains the characteristic mass values of the molecular ion and the fragmentation ions for each target AAS in separated databases. For each database created and stored in the MLibrary repository, the user may include standard modification mass variations corresponding to specific derivatisation reagents and the consequent MS/MS spectra for each AAS presenting that particular modification. The MLibrary repository comprises also MS/MS compound markers. The MLibrary repository is stored in a single standard XML file, which can be easily modified with any plain text editor. More information about how to edit this file can be found in the software web page.

The installation wizard is available from the MLibrary web site as an executable file that depends on the final user operating system: Windows, Linux or MAC. By executing the setup file, the installation wizard will be automatically launched. The user has to simply follow the instructions on the screen to successfully complete the installation.

3. Results and discussion

The MLibrary software is explained in detail in the following sections. First we describe the use of MLibrary to detect the presence of possible AASs in the target samples (MS mode) and then its use to confirm the identity of the compound (MS/MS mode).

3.1. MS mode

3.1.1. Construction of MS database

The MS database in the MLibrary software contains the characteristic mass values for each target AAS. Additionally, the mass values of AAS glucuronides, which are the main excretion metabolites of AAS in the human body, were also introduced in the MLibrary database. In our previous work, the analysis of AAS by MALDI techniques was performed after derivatisation with Girard T hydrazide and therefore only this specific modification was introduced in the repository. Nevertheless, the introduction of specific modifications to the database is easily performed by typing the name and the mass variation in the modifications corresponding line within the XML file.

3.1.2. Detecting the presence of AASs

After the SPE clean-up procedure, as explained in the experimental section, the collected sample solution was mixed with the MALDI matrix HABA and spotted onto the MALDI sample plate. The MALDI ion source is a soft ionisation technique and therefore the MALDI-TOF-MS analysis measure, primarily, singly charged ions that correspond to the molecular ions of the sample solution species as well as to the MALDI matrix characteristic ions. The mass spectrum obtained after the MALDI-TOF-MS analysis is exported as a list of peaks to a CSV file. Only the centroid mass and relative intensity of each peak is used by the MLibrary software (see Fig. 1 Supplementary material Fig. 1SM).

The AAS detection process in MLibrary starts with the MS data loading. Through the “Load MS Data” operation (see Fig. 2SM), the user can input the CSV file and filter peaks by their intensity, avoiding the load of peaks with lower intensities. Alternatively, the spectrum can be previously processed using the respective MS data software and only then transferred to a CSV file. It is important to stress that the MLibrary software permits the loading of multiple CSV files corresponding to several spectra.

After loading the MS data, the user can perform a compound search, in order to identify which AAS are present in the loaded data. The “MS Analysis” operation (see Fig. 4) determines if a specific AAS is present in the MS spectrum. This operation contemplates several parameters that provide flexibility to the software. It takes into account if the compound is in its “conjugated” or “free” form and also allows the user to select the derivatisation agent employed. It is important to stress that the “conjugated” parameter only takes into account the glucuronide conjugates, which are the main excretion metabolites of AAS in the human body. The search is performed taking into account the error between the database and the experimental mass values. This error value can be selected as percentage, parts per million (ppm) or absolute mass units (amu) (see Fig. 3SM). Then, the search retrieves the mass values that matched between the experimental data and the database values, showing both experimental and theoretical values as well as the name of the compound and the experimental peak intensity (see Fig. 2). MLibrary also provides an additional operation named “MS Full Analysis”, which performs a
search within all the MS databases present in the MLibrary repository. Since the AAS mass values are close to the matrix characteristic ion mass ones, it is possible to perform an internal calibration for each MALDI plate spot using the matrix peaks as reference. Consequently, the experimental mass values obtained are very accurate and precise.

3.2. MS/MS mode

In the first step of the analysis, the MLibrary software detected the ions that matched with the theoretic mass values of AAS within the MS database. In the second step, the ions detected by the MLibrary are selected for fragmentation in a second round of MS analysis, in which MS/MS spectra are acquire. Each AAS compound

Fig. 1. Schematic diagram of the MALDI-MS/MS strategy to analyse AAS using MLibrary software.
presents a characteristic MS/MS fragmentation pattern that will be used by the MiLibrary as a signature of that compound. Likewise the MS mode, the mass spectrum obtained after the MALDI-TOF-TOF-MS analysis is exported as a list of peak to a CSV file.

3.2.1. Construction of MS/MS spectral library
The MS databases in the MiLibrary contain the characteristic fragmentation ions for each associated compound. The construction of the database can be easily performed using standard solutions of AAS. To ensure reliable spectral data, the software allows the input of several replicates corresponding to each standard. With these data, the MiLibrary software generates a list of common fragmentation mass values for all replicates, along with their average relative intensities. In addition, when comparing with loaded data, the software allows the user to choose the discriminate power of the generated list, meaning that the user can decide to include within this list, the mass values that are present, for instance, at least in 95% percent of the spectral mass data inputted for each compound.

The construction of the mass spectral library in the MiLibrary software is an ongoing process that is simple to achieve and updated by any qualified user, with the advantage that it can be easily adapted to numerous experimental conditions and compounds.

3.2.2. Confirming the presence of AAS
This process is similar to the AAS detection process except that, in this case, the user can perform two additional analysis operations. As before, the process starts with the data loading. In this case, through the "Load MS/MS Data" operation (see Fig. 4SM). This operation also provides the "Peak Intensity" parameter for peak filtering.

After loading the sample CSV file, the user can use the "MS/MS Library Analysis" operation (see Fig. 5SM) to compare it with the characterstic MS/MS spectra of the AAS compounds stored in the MiLibrary databases. At this stage the user has to select the database, the modifications performed in the sample treatment procedure, the mass value of the precursor molecular mass and the mass tolerance permitted for each mass peak. Additionally, the user has to select the mass tolerance within the database mass spectra and its discriminate power. This operation retrieves a list of all compounds within the MS/MS Library that fit the search criteria, ranked by similarity to the inputted file. The similarity is attained regarding the number of mass values that matched and the relative intensity of all peaks. To ensure reliable results, the search score concerns only to a limited number of mass values, defined previously by the user. For instance, the user may limit the search to the 10 most intense peaks within the inputted file and the library data (see Fig. 6SM). This tool is essential to avoid misinterpretation of the results due to the fact that different spectra of the same sample, generally present distinct overall number of mass peaks.

Another important tool within the MiLibrary software is the "MS/MS Std. Match Analysis" operation. With this tool the user can compare two experimental spectra, which is extremely important if we are working at different conditions than the one recorded in the MiLibrary MS/MS database. By adding a standard solution of a specific compound to the MALDI analysis and comparing the two spectra using the MiLibrary (see Fig. 7SM), the user can confirm the identity of this compound. The results appear in the same way as shown in Fig. 9SM.

A third MS/MS Analysis tool is available in the MiLibrary software. The "MS/MS Marker Analysis" operation allows the user to locate concrete biomarkers into the loaded MS/MS data (see Fig. 8SM). This feature is particularly important for the analysis of isobaric compounds having very similar MS/MS fragments or compounds having poor fragmentation pattern. For isobaric compounds the MS/MS Library analysis does not allow the differentiation between the two species, since it will retrieve very close results. For this reason, the identification of exclusive fragments corresponding to each compound is essential for the interpretation of the spectra and it will become essential for their identification. Additionally, the user can use the "MS/MS Full Marker Analysis" in order to search for all the biomarkers stored in the MiLibrary repository.

4. Case study
Table 1 presents the results obtained with the data acquired from five urine samples spiked with different AAS at different concentration levels as described in the sample preparation section. All spiked compounds were detected by single MS as possible positives and their identity was confirmed by MS/MS. In the MS mode, besides the spiked compounds, the MiLibrary retrieved other AAS as possible positives. This result occurs due to the fact that some synthetic AAS possesses the same molecular mass than some endogenous AAS, which are present in urine at low concentration values. For instance, 4-androsten-3,17-dione is a minor AAS metabolite that possesses the same mass than boldenone.

Fig. 9SM presents the results retrieved by the MiLibrary for the sample Urine 1. As it may be seen in Fig. 9SM, in the MS mode, the software identified three peaks corresponding to four possible positives: boldenone, 17α-methyltestosterone, milborenone, which has the same molecular mass that 17α-methyltestosterone, and mesterolone. As mentioned above, following the detection of the possible presence of specific AAS in the urine sample, a MALDI-TOF-TOF-MS analysis was performed to the ions retrieved as possible positives. The MiLibrary software was used to confirm the identity of the compounds. Figs. 9BSM, 9CSM, 9DSM and 9EMS shows the results obtained for the four compounds. The presence of both boldenone and mesterolone is easily proved false by MiLibrary MS/MS data analysis, showing less than 10% of positive match to the reference database spectra within the 10 most intense peaks.

Regarding 17α-methyltestosterone and milborenone, the MiLibrary MS/MS data analysis clearly confirms the presence of 17α-methyltestosterone. MiLibrary software retrieves a higher percentage of mass values matching with 17α-methyltestosterone reference data within the 10 more intense peaks; 70% against 50% for milborenone. Furthermore, the peak intensity of all matched mass ions present in the MS/MS spectrum is similar to the MiLibrary reference data for 17α-methyltestosterone.

Figs. 1–4 of the Appendix A present the results retrieved by MiLibrary for the samples Urine 2, Urine 3, Urine 4 and Urine 5, respectively.

In sample Urine 2, the presence of the isobaric compounds calusterone and bolasterone are detected by the MS mode (see Fig. 1 of the Appendix A). Although both calusterone and bolasterone present similar fragmentation patterns, the intensity of the peaks are different. For this reason, despite the similarity of the matched peaks between the sample compound and the reference data for calusterone and bolasterone, the MiLibrary software easily confirms the identity of calusterone. It retrieves a higher percentage of mass values matching with calusterone within the 10 more intense peaks: 80% against 40%.

In sample Urine 3, the presence of nandrolone was confirmed using the MS/MS Marker Analysis tool (see Fig. 2 of the Appendix A). As it was mentioned above for poor fragmentation pattern compounds, such as nandrolone, the detection of specific markers, previously identified by the user, is of major importance.

Fig. 3 of the Appendix A shows the results retrieved by MiLibrary that allowed the identification of both fluoromesterone and ethisterone in sample Urine 4. The identification of ethisterone
Table 1
Urine samples data analysis by the MLibrary software.

<table>
<thead>
<tr>
<th>Urine</th>
<th>AAS compounds spiked into urine</th>
<th>MLibrary data analysis</th>
<th>AAS identified by MS mode</th>
<th>AAS confirmed by MS/MS mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urine 1</td>
<td>17α-methyltestosterone</td>
<td>Boldenone</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>17α-methyltestosterone</td>
<td>Boldenone</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Mibolerone</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Urine 2</td>
<td>Calusterone</td>
<td>Calusterone</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Bolasterone</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Urine 3</td>
<td>Nandrolone</td>
<td>Nandrolone</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Boldenone</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Urine 4</td>
<td>Ethisterone</td>
<td>Mibolerone</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Fluoxymesterone</td>
<td>Ethisterone</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Fluoxymesterone</td>
<td>Fluoxymesterone</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Urine 5</td>
<td>Trenbolone</td>
<td>Trenbolone</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Mibolerone</td>
<td>Boldenone</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Mibolerone</td>
<td>Mibolerone</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

was performed in the same manner than for nandrolone by using the MS/MS Marker Analysis Tool. The presence of mibolerone was also proved false, showing 0% of matching within the most intense peaks.

In sample Urine 5, the MLibrary identified the presence of 17α-trenbolone and mibolerone. The identification of 17α-trenbolone was performed using the tool MS/MS Standard Mach Analysis. The MS/MS identification of 17α-trenbolone was performed in a first approach, using the Library analysis tool, however only 60% of matching was achieved. The MS/MS Standard Mach Analysis tool allowed the unequivocal confirmation of this compound by comparing it with a standard solution analysed in the same experimental conditions.

5. Conclusions

We have developed friendly software to help in an automated mode to detect and identify the presence of AAS in urine samples by MALDI-TOF–TOF–MS. MLibrary software allows the user to perform robust and accurate screening and confirmation for AAS in both MS and MS/MS mode. As a proof-of-concept, when applied to five urine samples spiked with AAS, the software was able to successfully identify all spiked compounds. The software presented here is a very versatile tool that can be simply adapted to any future modifications carried out on the sample treatment procedure, as well as, be easily applied to other compounds. The MLibrary software saves time and it is a simple tool to work with. Additionally, MLibrary software has a wizard easy to follow for its installation.

Acknowledgements

M. Galesio acknowledges the Fundação para a Ciência e a Tecnologia (FCT, Portugal) for their post-doctoral grant SFRH/BPD/77094/2011. The Laboratório Antidoping di Roma (National Antidoping Laboratory from Italy) is gratefulfully acknowledged for providing us with steroids standards. This work was partially funded by the project Research on Translational Bioinformatics (08V186) from University of Vigo. Xunta de Galicia (Spain) is also acknowledged for financial support under project 09CSA043323P-2009. C.L. and JLC thanks to Scientific Association PROTEOMASS (Portugal) for financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.steroids.2013.08.014.

References

[18] Mazzarino M, de la Torre X, Botri F. A screening method for the simultaneous detection of glucuronoid metabolites, diuretics, stimulants, anti-oestrogens, beta-


Appendix

International Journal of Data Mining and Bioinformatics 10/4, pp. 455-473; doi: 10.1504/IJDMB.2014.064897
A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification

Hugo López-Fernández*, Miguel Reboiro-Jato, Daniel Glez-Peña and Florentino Fernández-Riverola

Department Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
Email: hlfernandez@uvigo.es
Email: mjrjato@uvigo.es
Email: dgpena@uvigo.es
Email: riverola@uvigo.es
*Corresponding author

Abstract: Matrix-Assisted Laser Desorption Ionisation Time-of-Flight (MALDI-TOF) is one of the high-throughput mass spectrometry technologies able to produce data requiring an extensive preprocessing before subsequent analyses. In this context, several low-level preprocessing techniques have been successfully developed for different tasks, including baseline correction, smoothing, normalisation, peak detection and peak alignment. In this work, we present a systematic comparison of different software packages aiding in the compulsory preprocessing of MALDI-TOF data. In order to guarantee the validity of our study, we test multiple configurations of each preprocessing technique that are subsequently used to train a set of classifiers whose performance (kappa and accuracy) provide us accurate information for the final comparison. Results from experiments show the real impact of preprocessing techniques on classification, evidencing that MassSpecWavelet provides the best performance and Support Vector Machines (SVM) are one of the most accurate classifiers.

Keywords: mass spectrometry; low-level preprocessing; sample classification; model comparison.

Reference to this paper should be made as follows: López-Fernández, H., Reboiro-Jato, M., Glez-Peña, D. and Fernández-Riverola, F. (2014) ‘A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification’, Int. J. Data Mining and Bioinformatics, Vol. 10, No. 4, pp.455–473.

Biographical notes: Hugo López-Fernández is a PhD fellow in the Computer Science Department at the University of Vigo. He is currently working in bioinformatics annotation tools using biomedical ontologies, as well as mass spectrometry data analysis.
1 Introduction

In the last years, high-throughput mass spectrometry (MS) based proteomic data analysis has been an active research area. MS technology allows researchers to measure the mixture of peptides or proteins present in biological samples, such as urine, serum or tissues. These measurements can be further used for discovering condition related patterns (biomarker discovery) and subsequently, for classify samples automatically.

A typical mass spectrometry experiment always produces a large volume of raw data. Tens to hundred MS spectra are generated, each one containing thousands of measurements (i.e. \( m/z \), intensity) pairs. In such a situation, dimensionality reduction is a crucial task that must be carried out before subsequent classification or biomarker discovery (Hilario et al., 2006). This is done in a step globally called preprocessing, an extensive low-level procedure able to clean raw data and detect true signals in the noisy spectra (Armananzas et al., 2011). The whole preprocessing procedure comprises several tasks such as baseline correction, smoothing, normalisation, peak detection and peak alignment. Since the application of inadequate or incorrect preprocessing methods can result in biased data set, also hindering the achievement of meaningful biological conclusions (Coombes et al., 2007), preprocessing is a critical task in rigorous MS data analysis. For this reason, several algorithms have been proposed to address each preprocessing task. These algorithms differ from each other in their principles, implementations and performance (Yang et al., 2009); therefore, their output can affect subsequent analyses.

Related with this situation, several studies have been carried out in order to compare different preprocessing algorithms and alternatives. In these works, comparisons are performed in terms of False Discovery Rate (FDR), sensitivity and reproducibility (Cruz-Marcelo et al., 2008; Yang et al., 2009; Zou et al., 2011). Unlike these studies and from another perspective, the goal of this work is to systematically study how different preprocessing algorithms and configurations affect classifiers accuracy when dealing with MS data.

After motivating the work, the rest of the paper is structured as follows: Section 2 gives a general overview about mass spectrometry domain. Section 3 identifies and briefly describes the analysed preprocessing steps. While Section 4 introduces the experimental set-up giving details for guaranteeing reproducibility, Section 5 presents and discusses obtained results. Finally, Section 6 concludes and outlines future work.
2 Related work on mass spectrometry

In proteomics, mass spectrometry stands for a commonly used technique for measuring the mass-to-charge ratio (m/z) of the components in a sample (Eidhammer et al., 2008). MS provides rapid and precise measurements about the sizes and relative abundances of the proteins present in a complex biological/chemical mixture (Shin and Markey, 2006). Mass spectrometric measurements are carried out using mass spectrometers, comprising an ionisation source, a mass analyser and a detector.

In a typical analysis, the sample is passed through these three components generating a mass spectrum. First, the components of the sample are ionised in the ionisation source in order to be able to measure their masses. The two main types of sources are (Colinge and Bennett, 2007): based on electrospray ionisation (ESI) or based on Matrix-Assisted Laser Desorption Ionisation (MALDI), which represents the dominating ionisation source for (single) mass spectrometry (Eidhammer et al., 2008). Then, in the mass analyser, the components of the sample are separated according to the mass-to-charge ratio (m/z) of the ions. After separation, the components hit the detector and their m/z values can be calculated. There are four basic types of mass analysers currently used in proteomics research ( Aebersold and Mann, 2003): (a) ion trap, (b) Time-of-Flight (TOF), (c) quadrupole and (d) Fourier transform ion cyclotron (FT-MS). While ESI are usually coupled to ion traps and triple quadrupole analysers, MALDI are usually coupled to TOF analysers and are the simplest instruments for protein and peptide analysis.

A mass spectrometer (e.g. MALDI-TOF) operates and connects to a computer, which constructs a mass spectrum after analysing a sample in the instrument (Figure 1). A mass spectrum is a representation where the measured masses are placed along the horizontal axis and the intensity of the signal for each component (m/z) along the vertical axis. As a result of its simplicity, excellent mass accuracy, high resolution and sensitivity (Aebersold and Mann, 2003; Eidhammer et al., 2008), Matrix-Assisted Laser Desorption Ionisation Time-of-Flight (MALDI-TOF) is widely used to identify proteins though their peptides, a process known as peptide-mass fingerprinting. In this process, the mass spectrum must be preprocessed in order to obtain a list of peptide experimental masses, which can be searched against a database to identify proteins.

Figure 1 Mass spectrometer

Mass spectrometry is generally applied both to biomarker discovery (Issaq et al., 2002; Prados et al., 2004; Diamandis, 2004; McDonald et al., 2009) and classification (Tibshirani et al., 2004; Alagaratnam et al., 2008; Pietrowska et al., 2009; Long et al.,
2011). In this context, as pointed out in the work of Coombes et al. (2007), it is important to note that a typical data set arising in an application of mass spectrometry contains from tens to hundreds spectra. Each spectrum has between 10,000 and 100,000 intensity measurements representing an unknown number of protein peaks. Such huge volume of raw data requires extensive low-level preprocessing in order to clean the data and to detect the true signals in the noisy spectra (Armananzas et al., 2011).

3 Available preprocessing methods for MS data

As stated before, preprocessing of MS data is a critical stage that transforms raw data into a suitable input for further analysis, such as machine learning or biomarker discovery. Inadequate or incorrect preprocessing methods can result in biased data set and hinder to reach meaningful biological conclusions (Coombes et al., 2007). In such a situation, preprocessing is necessary since raw data contain signals coming from the real peptides/proteins, as well as signals derived from several forms of noise (e.g. chemical, electronic factors, etc.). The specific goals of this phase are (Eidhammer et al., 2008) (a) to remove noisy peaks without discarding any of the true peaks and (b) to determine the m/z and intensity values with the best accuracy.

Since there is no standard mass spectrometry data preprocessing pipeline, some authors proposed different guidelines to establish a design/Data Analysis Protocol (DAP) (Barla et al., 2008; Armananzas et al., 2011). In this work, we use the following preprocessing workflow: (a) baseline correction, (b) smoothing, (c) peak detection and (d) peak alignment. Next subsections introduce the most popular preprocessing steps and techniques in detail.

3.1 Baseline correction

Baseline is a specific form of noise mainly motivated by chemical perturbations. Eidhammer et al. (2008) defined baseline as an offset of the intensities of masses that often shows a dependency on the m/z value such that it is highest at low m/z values, presenting an exponential decay towards higher masses. For MALDI-TOF mass spectrometry, baseline is a monotonically decreasing bias resulting from matrix clusters formed during ionisation (Shin et al., 2007; Sun and Markey, 2011). The most common baseline correction methods are monotone minimum, linear interpolation, loess, moving average of minima and Continuous Wavelet Transform (CWT), being available as free software. The Bioconductor PROcess package (Li et al., 2005) implements both loess and linear interpolation. Moreover, the Cromwell package (Coombes et al., 2005) implements monotone minimum methods. Figure 2 shows an example of a raw spectrum with (a) baseline noise and (b) the corresponding baseline corrected spectrum.

3.2 Smoothing

Usually, spectra are jagged, making it difficult to detect true peaks amongst the noise. Therefore, a smoothing algorithm is usually applied in order to soften the data. Again, several methods have been successfully proposed in the literature for correcting this issue. The simplest technique consists of using a sliding window, where a new value is calculated for the point in the middle, based on the values of the points in the window.
Commonly used filters are moving average, Savitzky-Golay, Gaussian and the Kaiser window. However, MS community appears to be converging on the use of wavelets for denoising (Coombes et al., 2007). In this context, Combes et al. (2005) presented a denoising method based on the Undecimated Discrete Wavelet Transform (UDWT) that is available in the Cromwell package.

Figure 2  Example of baseline correction

![Baseline correction](image)

(a)  (b)

3.3 Peak detection

Peak detection can be defined as the process of selecting true (i.e. peptide/protein-related) peaks from a given spectrum. As in previous subsections, there are many peak detection algorithms, most of them applied after baseline correction and smoothing. However, Du et al. (2006) proposed a method that performs peak detection without explicit smoothing and baseline correction. This method is based on the CWT and is publicly available in the MassSpecWavelet package, which is also in the Bioconductor project.

Peak detection algorithms use one or more of the following criteria in order to identify true peaks (Yang et al., 2009):

- **Signal to Noise Ratio (SNR):** SNR is a measure of the signal relative to the background noise. Peaks are selected if their SNR is larger than a given threshold.

- **Detection intensity threshold:** This criterion is used to remove small peaks in flat regions. The use of SNR alone in such regions may select noisy points as peaks, since they may have a high SNR.

- **Slopes of peaks:** Within this criterion, the shape of peaks is used to filter out false occurrences. Any potential peak is eliminated if both left and right slopes are less than a pre-established threshold. The limit is defined as half of the local noise level (Coombes et al., 2003).

- **Local Maximum:** Following this criterion, a peak is selected if it is a local maximum of N neighbouring points.

- **Shape ratio:** Within this criterion, a peak is selected if its shape ratio (Yang et al., 2009) exceeds a certain threshold.

- **Model-based criterion:** These methods use a model function in order to fit peaks.
• **Peak width**: A peak is detected as true if its peak width is contained in a given range.

• **Ridge lines**: Used in the method based on the CWT (Du et al., 2006) for detecting true peaks.

For example, in the Cromwell package peaks are selected if they are local maxima and larger than a given SNR. However, in the PRO cess package in addition to the SNR, intensity threshold and shape ratio criterion are also used.

### 3.4 Peak alignment

This process, also referred as peak matching, consists on determining which peaks correspond to the same peptide/protein in different samples. Apart from the existence of other standard algorithms, in this work we analysed our own peak alignment method based on a moving window successfully used in previous works (Santos et al., 2010; Nunes-Miranda et al., 2012). As a result of the execution of the algorithm, all the aligned peaks have the same mass values in all spectra. These mass values correspond to the virtual centroid mass. Figure 3 shows the pseudocode algorithm of the proposed method.

![Figure 3](image)

**Pseudocode of the proposed peak alignment algorithm**

```plaintext
00 PROCEDURE peakAlignment(INPUT: peak_list; OUTPUT: alignments)
01 02 BEGIN
03 04 WHILE not_empty(peak_list) DO
05 06 BEGIN
07 08 group_peaks := []
09 10 start_peak := first(peak_list)
11 12 centroid_peak := start_peak
13 14 end_peak := start_peak
15 16 added_spectrum_list := []
17 18 group_peaks := group_peaks U start_peak
19 20 remove(start_peak, peak_list)
21 22 added_spectrum_list := added_spectrum_list U spectrum(start_peak)
23 24 finished := FALSE
25 26 WHILE not_finished AND not_empty(peak_list) DO
27 28 BEGIN
29 30 current_peak := next(end_peak)
31 32 tolerance := getTolerance(centroid_peak, ppm)
33 34 IF centroid_peak-start_peak < tolerance THEN
35 36 IF member(spectrum(current_peak, added_spectrum_list)) THEN
37 38 remove(current_peak, peak_list) /* ignore peak */
39 40 ELSE
41 42 group_peaks := group_peaks U current_peak
43 44 remove(current_peak, peak_list)
45 46 end_peak := current_peak
47 48 centroid_peak := avg(group_peaks)
49 50 added_spectrum_list := added_spectrum_list U spectrum(current_peak)
51 52 ELSE
53 54 finished := TRUE
55 56 END WHILE
57 58 alignments := alignments U new_group(start_peak, centroid_peak, end_peak, group_peaks)
59 60 END WHILE
61 62 END_peakAlignment
```

The window size of the algorithm showed in Figure 3 may be specified in three different ways: (a) absolute, (b) relative and (c) PPM (Points Per Million). In this context, a PPM window size of 500 means that at a given m/z value m, the window is m ± (m 500/10⁵). When working with PPM, the higher mass value the wider window size, based on the fact that mass measurements at higher masses tend to be more distant.
A comprehensive analysis about the influence of low-level preprocessing

With the goal of clarifying the computation of a peak alignment, Figure 4 shows an example that illustrates the alignment process of three sample spectra. Considering a window of 500 PPM, the three peaks with masses 999.7, 1000.1 and 1000.3 are aligned (i.e. they represent the same peak), so that after alignment the three spectra contain the same peak. This new peak is the centroid of the three peaks being initially aligned, and it is computed as their mean.

**Figure 4** Peak alignment process with a window of 150 PPM

4 Experimental set-up

With the goal of conducting a straightforward and reproducible experiment for evaluating the influence of different techniques supporting the previous commented preprocessing pipeline, this section introduces in detail important questions regarding our experimental set-up: (a) data set preparation procedure, (b) machine learning techniques and (c) global preprocessing configuration and classification workflow.

4.1 Mass spectrometry data sets

A typical MS data set contains samples divided into two or more classes (e.g. healthy and diseased). Usually, for each sample, there are several biological replicates and, for each one, there are several technical replicates (i.e. the result of applying the same experimental procedure several times to the same biological replicate). Finally, each technical replication is spotted several times into the mass spectrometer.

For this work, we have used a large data set consisting of 14 different Spanish wines with five bottles of each wine (Nunes-Miranda et al., 2012). Each bottle (i.e. sample) was spotted five times so that the wine data set contains a total of 350 (14 × 5 × 5) spectra.

In order to further analyse the effect of commonly used pre-processing techniques in classification problems, the MS data set must be stored in a proper format, such as
WEKA’s own ARFF (Hall et al., 2009). For this reason, the set of attributes consists of the masses of all spectra together with the corresponding class attribute. Then, each spectrum is added as a new data line (i.e. instance) with ones where peaks are present and zeros otherwise. Figure 5 shows an example of an MS data set exported as ARFF file.

**Figure 5** MS data set in ARFF format

```
@relation wineMassData

@attribute sample {A#1, A#2, ..., B#1, B#2, ...}
@attribute mass1 numeric
@attribute mass2 numeric
... 

@attribute massN numeric
@attribute class {A, B, ...}
@data
A#1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1; 0, 0, 0; ... 0, A
A#2, 0, 0, 1; 0, 0, 0, 0, 1, 1; 0, 0, 0; ... 0, A

B#1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, ... 0, B
B#2, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, ... 0, B
```

### 4.2 Machine learning algorithms

It is generally accepted by the scientific community that there is no a superior classification algorithm/technique for being applied to all domains. Taking into account this limitation, in this work we have selected several algorithms covering different learning approaches such as decision trees, instance-based learning, Support Vector Machines (SVMs) or probabilistic models.

As instance-based learning approach we have chosen the K-Nearest Neighbours (KNN) algorithm (Aha et al., 1991), implemented by WEKA in the class IBk (weka.classifiers.IBk). In this classification algorithm, no internal model is built, and the training stage just consists of storing training samples. To classify a new case, its K-nearest neighbours (i.e. the K samples most similar to it in terms of the specific problem domain) are selected. The most frequent class among these K neighbours is the class predicted for the new case.

As probabilistic models, we have chosen Bayesian networks (Pearl, 1985) and Naïve Bayes (John and Langle, 1995) classifiers available at BayesNet and NaiveBayes classes belonging the weka.classifiers.bayes package. Naïve Bayes is a simple classifier that assigns a case \( x \) to the most probable class given \( x \), relying on the Bayes’ theorem and the variable independence assumption.

As decision tree (DT) learning algorithms, we have selected C4.5 (Quinlan, 1993) and Random Forest (RF) (Breiman, 2001). RF is an ensemble classifier that consists of many DTs, where the output class is computed as a combination from the output classes of each individual tree. Briefly, a decision tree is built by recursively partitioning the training data with the aim of maximising the class homogeneity of the resulting subsets. The selected variable will be the one that ensures the maximal reduction of class heterogeneity measured in different ways (e.g. in the case of C4.5 this measure is the entropy). Both J48 (the implementation of C4.5 algorithm) and RandomForest are available in the weka.classifiers.trees package.
Finally, we have also opted for several classifiers based on SVMs (Vapnik, 1998), which have been applied extensively to mass spectra (Hilario et al., 2006). On the one hand, an SVM trained with Sequential Minimal Optimisation (SMO; Platt, 1999) is available in the class SMO (weka.classifiers.functions). On the other hand, two SVMs: a radial basis function (SVM-RBF) and a linear function (SVM-L). These SVMs are available through the WEKA’s class LibSVM, which acts as wrapper for the libsvm tools (Chang and Lin, 2001).

4.3 Preprocessing configuration

As it was previously commented, the selected dataset consists of 350 MALDI-TOF raw spectra stored in mzML format (Martens et al., 2011), the PSI (Proteomics Standards Initiative) standard for MS data. In our experiments, spectra are read using anmzML (Cote et al., 2010), an open-source Java library for mzML.

Each single spectrum is treated by using a combination of the preprocessing steps described in Section 3. In our experimental workflow (Figure 6), peak detection (PD) is always performed, whereas baseline correction (B) and smoothing (S) are optional. As a result of this operation, a list of representative peaks for each spectrum is obtained.

Figure 6 Experimental workflow followed in the present study

In order to execute the preprocessing tasks, we have used PROcess and MassSpecWavelet (MSW) packages partly because (a) they are commonly used mass spectrometry preprocessing software and (b) the availability of integration facilities using RJava (Urbanek, 2011) to invoke them. Table 1 shows the preprocessing steps available in PROcess and MSW together with the set of parameters under study based on the experiments of Yang et al. (2009). In PROcess, smoothing and peak detection should be executed together, resulting in a single step.
The first preprocessing block depicted in Figure 6 using the preprocessing configurations showed in Table 1 leads to 189 different set-ups: 36 (PROCess S + PD) + 27 (MSW PD) + 2 \times 36 (PROCess B | PROCess S+PD) + 2 \times 27 (PROCess B | MSW PD).

Table 1  Preprocessing configuration

<table>
<thead>
<tr>
<th>Step</th>
<th>Package</th>
<th>Parameters</th>
<th>Values</th>
<th>Combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>PROCess</td>
<td>Method</td>
<td>approx, loess</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bandwidth</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SNR</td>
<td>{1, 3, 5}</td>
<td></td>
</tr>
<tr>
<td>S+PD</td>
<td>PROCess</td>
<td>Peak neighbour</td>
<td>{0.3, 0.003, 0.0003}</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peak ratio</td>
<td>{0.001, 0.01, 0.1, 0.5}</td>
<td></td>
</tr>
<tr>
<td>PD</td>
<td>MSW</td>
<td>SNR</td>
<td>{1, 3, 5}</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peak scale range</td>
<td>{2, 4, 6}</td>
<td></td>
</tr>
</tbody>
</table>

In the second preprocessing block showed in the experimental workflow of Figure 6, peak alignment is performed by using the algorithm described in Section 3.4. Also in this step, a sample (i.e. a bottle, in the wine data set) is summarised into a single spectrum (called sample fingerprint) instead of the five replicates per spectrum. Consequently, after peak alignment, the wine data set is represented by 70 spectra, corresponding to 14 classes with five samples per class, in which each spectrum represents one sample. The process of generating a sample fingerprint is parameterised by the intra-sample percentage of presence (POP), which stands for the minimum number of replicates, where a peak must be present in order to be added to the fingerprint. The higher POP, the less number of peaks selected. For instance, considering that there are five replicates for each sample, POP varies from 20% (which can be seen as the union of the five spectra) to 100% (which can be understood as the intersection of the five spectra).

Peak alignment is performed inside a loop, where a new version of the data set is created at each iteration. In this loop, the intra-sample POP varies from 1 to the maximum replicates in a sample (five in our data set). Thus, after applying the same preprocessing pipeline to the original data set, there are several additional variations due to the intra-sample POP used to create the sample fingerprints. As a consequence, there is a total of 945 preprocessing configurations (189 \times 5).

In order to systematically study the influence of the preprocessing steps (algorithms and their parameter configurations), the set of classifiers presented in Section 4.2 was applied to each preprocessed data set resulting from the previous experimental workflow. All the classifiers were executed with default parameters except for IBk, where $K$ was configured considering ten neighbours.

With the goal of guaranteeing the validity of the results, we conducted an ad hoc tenfold cross-validation experiment. All the runs were executed on the AIBench platform (Glez-Peña et al., 2010) using a WEKA plugin. In order to precisely measure the accuracy of each classifier, we have used both (a) the percentage of correct classifications (accuracy) and (b) the Cohen’s kappa statistic (Cohen, 1960). The kappa index compensates for classifications that may be due to chance and it is considered a standard statistically robust measure useful to assess the accuracy in multi-class problems.
A comprehensive analysis about the influence of low-level preprocessing (Ben-David, 2008). Kappa compares the real class with the predicted one, giving values ranging from 0 (random classification) to 1 (perfect classification) and being computed as showed in expression (1).

$$Kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$  \hspace{1cm} (1)

where \(Pr(a)\) represents the observed agreement and \(Pr(e)\) stands for the expected agreement due to chance.

All the runs were executed using Ubuntu 2.6.32 64 bit-server over an 32-core AMD OpteronTM 6180 with 64 GB GB of RAM.

5 Results and discussion

As previously mentioned, we selected kappa as performance measure in order to carry out subsequent analyses from the outcome of the statistical tests. For each one of the 945 possible configurations, we collected in a single table those values which corresponding to baseline correction, peak detection, POP, type of classifier, fold and kappa.

First, we carried out a multi-factor ANOVA (McDonald, 2009) for analysing the full table of results. This test allows us to see: (a) for each parameter, if there are significant differences among its possible values and (b) if there are interactions between parameters. The goal of this test is twofold: detecting parameters that significantly impact on the classification and identifying possible interactions among them. The results of the test are showed in Table 2.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Results of a multi-factor ANOVA</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Factors</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>classifier</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>POP</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>Peakdetection</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>Baselinecorrection</td>
<td>0.2075</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interactions</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>classifier:POP</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>classifier:Peakdetection</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>POP:Peakdetection</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>classifier:Baselinecorrection</td>
<td>0.9999</td>
</tr>
<tr>
<td>POP:Baselinecorrection</td>
<td>0.9999</td>
</tr>
<tr>
<td>Peakdetection:Baselinecorrection</td>
<td>0.9734</td>
</tr>
<tr>
<td>classifier:POP:Peakdetection</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>classifier:POP:Baselinecorrection</td>
<td>1</td>
</tr>
<tr>
<td>classifier:Peakdetection:Baselinecorrection</td>
<td>1</td>
</tr>
<tr>
<td>POP:Peakdetection:Baselinecorrection</td>
<td>1</td>
</tr>
<tr>
<td>classifier:POP:Peakdetection:Baselinecorrection</td>
<td>1</td>
</tr>
</tbody>
</table>
Results presented in Table 2 reveal that the classifier, POP and peak detection factors significantly affect the classification performance, while baseline correction parameter does not. This means that there are no significant differences among the baseline correction methods employed (none and both loess and approx from PROcess). This conclusion was expected since in a visual inspection of the spectra, baseline artefact was not detected.

On the other hand, the test has identified differences among the different peak detection algorithms, the POP values and the classifiers. Moreover, the test has revealed that these three factors have an interaction. To precisely study this interaction, we carried out a Tukey’s HSD (Honestly Significant Difference) post-hoc test.

The interaction between classifier and POP means that each classifier is not affected in the same way by the different values of POP (Figure 7). While most of classifiers perform worse with higher POP values, SVM-L and SVM-RBF does not seem to be affected by this factor.

Figure 7  Boxplot of classifiers by POP (see online version for colours)

Tukey’s HSD test performs pairwise comparisons for each possible combination of classifier, peak detection configuration and POP, which results in millions of comparisons. In order to reduce such huge size, we have studied each peak detection package (MSW and PROcess) separately. This analysis allowed us to select the best peak detection configurations in each package, and then filter the Tukey’s HSD output to reduce its size.
Figure 8 shows a boxplot of PROcess peak detection parameters. First, they are grouped by ratio, then by SNR and finally by area.w. A visual inspection of the boxplot suggests that SNR and ratio may not affect the performance of the classifiers, while area.w clearly affects it.

**Figure 8**  Boxplot of PROcess parameters (see online version for colours)

Again, we carried out an ANOVA test, whose results are showed in Table 3. The test confirms that there are no statistically significant differences in SNR and ratio, something that does not occur with area.w.

**Table 3**  Results of PROcess ANOVA

<table>
<thead>
<tr>
<th>Factors</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>area.w</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>snr</td>
<td>0.9790</td>
</tr>
<tr>
<td>ratio</td>
<td>0.8750</td>
</tr>
<tr>
<td>Interactions</td>
<td></td>
</tr>
<tr>
<td>area.w:snr</td>
<td>0.9480</td>
</tr>
<tr>
<td>area.w:ratio</td>
<td>0.7610</td>
</tr>
<tr>
<td>snr:ratio</td>
<td>0.9920</td>
</tr>
<tr>
<td>area.w:Baselinecorrection</td>
<td>0.9980</td>
</tr>
</tbody>
</table>
To analyse in detail the differences in the parameter \( \text{area.w} \), we carried out a Tukey’s HSD test, whose results are showed in Table 4. A positive difference denotes that the mean of the first factor of the comparison is greater than the second, while a negative difference indicates that the mean of the second is greater. The value 0.3 of \( \text{area.w} \) is better than 0.003 and 0.0003. This result is clearly visible in Figure 8.

<table>
<thead>
<tr>
<th>( \text{area.w} ) pairwise comparisons</th>
<th>difference</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{area.w} = 0.3 ) vs. ( \text{area.w} = 0.003 )</td>
<td>0.1028</td>
<td>0</td>
</tr>
<tr>
<td>( \text{area.w} = 3.0\text{E-4} ) vs. ( \text{area.w} = 0.003 )</td>
<td>-0.0350</td>
<td>0</td>
</tr>
<tr>
<td>( \text{area.w} = 3.0\text{E-4} ) vs. ( \text{area.w} = 0.3 )</td>
<td>-0.1378</td>
<td>0</td>
</tr>
</tbody>
</table>

In order to analyse MSW package, we proceeded in the same way as in the previous case. Therefore, Figure 9 shows a boxplot of MSW peak detection parameters. First, they are grouped by \( \text{amp.Th} \), then by \( \text{SNR} \) and finally by \( \text{peakScaleRange} \). A visual inspection of the plotbox suggests that \( \text{peakScaleRange} \) may not affect to the performance of the classifiers, while both \( \text{amp.Th} \) and \( \text{SNR} \) affect it.

The results of the corresponding ANOVA test are presented in Table 5. In this case, here we detected statistically significant differences for the three factors, so that we should study the interaction among them.

To study the interaction among \( \text{amp.Th} \), \( \text{SNR} \) and \( \text{peakScaleRange} \), we have applied a Tukey’s HSD. Tukey’s HSD test agrees with ANOVA and confirms that parameters
strongly affect each other, since there is not a clear best combination of parameters. However, these results suggest that a value 0.1 of the parameter ampTh performs clearly worse. The best configurations (i.e. those that win in more times) are summarised in Table 6.

Table 5  Results of MSW ANOVA

<table>
<thead>
<tr>
<th>Factors</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ampTh</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>snr</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>peakScaleRange</td>
<td>0.0366</td>
</tr>
<tr>
<td>Interactions</td>
<td></td>
</tr>
<tr>
<td>ampTh:snr</td>
<td>&lt;2.0E-16</td>
</tr>
<tr>
<td>ampTh:peakScaleRange</td>
<td>4.76E-05</td>
</tr>
<tr>
<td>snr:peakScaleRange</td>
<td>0.0117</td>
</tr>
<tr>
<td>ampTh:snr:peakScaleRange</td>
<td>1.57E-10</td>
</tr>
</tbody>
</table>

Table 6  MassSpecWavelet best parameter combinations

<table>
<thead>
<tr>
<th>SNR</th>
<th>amp.Th</th>
<th>peakScaleRange</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.01</td>
<td>2</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0001</td>
<td>2</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0001</td>
<td>4</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0001</td>
<td>6</td>
</tr>
<tr>
<td>5.0</td>
<td>0.0001</td>
<td>2</td>
</tr>
</tbody>
</table>

At this time, we are able to study the global interaction among classifier, POP and peak detection. By analysing their behaviour, we expect to be able to answer the following questions: (a) what is the best peak detection algorithm (MSW vs. PROcess), (b) what are the best POP values and (c) what are the most suitable classifiers for MS data. In order to gain knowledge about this issues, we carried out a Tukey’s HSD test and filtered the results to the best configurations of PROcess (\(area.w = 0.3\)) and MSW (Table 6). Moreover, we also exclude those rows with POP values of 20 and 40, since these values are not usually acceptable.

The Tukey’s HSD filtered test table compares 408 different configurations. By analysing the comparisons, we observed a group of 19 configurations that won in the 42% of times, tying in the remaining cases. In these configurations, SVM-L or SVM-RBF and MSW are always present, with a variety of parameters as concluded previously. Moreover, SVM-RBF and SVM-L are present in the top 27 configurations. Figure 10 shows a boxplot of the overall kappa and accuracy of each classifier supporting this conclusion.

Regarding the peak detection algorithm used, MSW is present in the best 71 configurations with 0% of losses and a win percentage that ranges from 22% to 42%. This fact indicates that MSW clearly outperforms PROcess. Figure 11 shows a boxplot of the overall kappa and accuracy measures obtained by each package.
It is also remarkable the fact that the best configurations involving PROcess also include SVM-L as classifier, supporting the conclusion that this type of classifier performs better in this case. Finally, results do not reveal significant differences among the values of POP. Although Figure 7 shows that most of classifiers perform worse with higher values of POP, SVMs seem not to be affected.

**Figure 10** Classifiers boxplot comparison

**Figure 11** PROcess vs. MSW boxplot comparison

6 Conclusions

High-throughput mass spectrometry data analysis requires an important stage of preprocessing, since it could bias subsequent analysis such as classification or biomarker discovery. In this work, we have studied the effect of such preprocessing on MS sample classification, comparing two standard packages: MassSpecWavelet and PROcess.

First, we found that baseline correction has no effect in classification. However, this result cannot be extrapolated to other situations since the spectra in our data set do not exhibit baseline.
Regarding peak detection, results show that MassSpecWavelet outperforms PROCeSS. This conclusion agrees with previous studies, where MassSpecWavelet also exhibits a better performance in terms of FDR and sensitivity (Cruz-Marcelo et al., 2008; Yang et al., 2009).

Although this study is mainly focused on the comparison of preprocessing techniques, we have also found that SVMs achieved the best performance in our experiments.

In order to complement the present study, further work includes the comparison of more publicly available libraries, such as Cromwell (Du et al., 2006) or LIMPIC (Mantini et al., 2007) working with more MS data sets. However, within this study, we have also established a valuable framework for extensively testing MS preprocessing techniques following a reproducible procedure.

Acknowledgements

This work is partially funded by the (a) TIN2009-14057-C03-02 project from the Spanish Ministry of Science and Innovation, the Plan E from the Spanish Government and the European Union from the ERDF, (b) the integrated action AIB2010PT-00353 from the Spanish Ministry of Science and Innovation, and (c) Agrupamento INBIOMED (2012/273) from DXPCTSUG-FEDER unha maneira de facer Europa. H. López-Fernández was supported by a pre-doctoral fellowship from the University of Vigo.

References


A comprehensive analysis about the influence of low-level preprocessing


Appendix

Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery

H. López-Fernández1,2, H. M. Santos3, J. L. Capelo1, F. Fdez-Riverola1,2, D. Glez-Peña1,2 and M. Reboiro-Jato1,2*

Abstract
Background: Mass spectrometry is one of the most important techniques in the field of proteomics. MALDI-TOF mass spectrometry has become popular during the last decade due to its high speed and sensitivity for detecting proteins and peptides. MALDI-TOF-MS can be also used in combination with Machine Learning techniques and statistical methods for knowledge discovery. Although there are many software libraries and tools that can be combined for this kind of analysis, there is still a need for all-in-one solutions with graphical user-friendly interfaces and avoiding the need of programming skills.

Results: Mass-Up, an open software multipurpose application for MALDI-TOF-MS knowledge discovery is herein presented. Mass-Up software allows data preprocessing, as well as subsequent analysis including (i) biomarker discovery, (ii) clustering, (iii) biclustering, (iv) three-dimensional PCA visualization and (v) classification of large sets of spectra data.

Conclusions: Mass-Up brings knowledge discovery within reach of MALDI-TOF-MS researchers. Mass-Up is distributed under license GPLv3 and it is open and free to all users at http://sing.eiu vigoes/mass-up.

Keywords: Mass spectrometry, MALDI-TOF-MS, Knowledge discovery, Machine learning, Biomarker discovery

Background
Mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analysers, MALDI-TOF-MS, referred to herein as MALDI, has become popular during the last decade due to its high speed and sensitivity for detecting proteins and peptides. Large sets of samples are analysed quickly in one single batch. The aforementioned reasons have led to the use of MALDI for the classification of large sets of samples from different sources and/or characteristics [1]. In this sense, computational tools play a key role in MALDI experiments, as they are able to preprocess raw data registered in different formats, compare them, and apply complex algorithms in order to finally extract new knowledge and useful conclusions.

Raw data generated by MALDI is usually composed of large spectra sets. Each single spectrum contains thousands of measurements entailing mass-to-charge ratio \( m/z \) signals and intensity (i.e. \( m/z \), intensity) pairs. These spectra are usually stored using open xml-based formats such as mzXML [2], mzML [3] and PeakML [4]. In addition, several open-source libraries to handle these data formats have been developed in the last years, among which the following are noteworthy: mzMatch [4], jmzML [5], jmzReader [6], the ProteomeCommons.org IO Framework [7] and different R packages [8, 9].

The spectra generated by MALDI apparatus usually contain a high level of noisy signals, making data preprocessing a crucial task that must be carried out before subsequent analysis [10]. This preprocessing is an extensive low-level procedure able to clean raw data and identify true signals in the noisy spectra [11]. Preprocessing comprises several tasks, such as baseline correction, smoothing, normalization, peak detection and peak matching. The use of inadequate or incorrect preprocessing methods can result in a biased dataset, hindering the achievement of meaningful biological conclusions [12]. Therefore, preprocessing is a critical stage in rigorous MALDI data analysis. To accomplish the aforementioned tasks, different...
algorithms and tools have been developed. Most of them are publicly available as R packages [8, 13, 14], Matlab packages [15], Java libraries [16, 17] or standalone applications [18–20].

Although MALDI is commonly used to identify and characterize molecules, such as peptides or proteins, it can be also be used in combination with Machine Learning (ML) techniques and statistical methods [1] to perform biomarker discovery [21, 22], automatic sample classification [23–26], and sample clustering [27, 28]. However, there are no tools devoted to performing these analyses, thus forcing researchers to use more general tools such as R, SPSS, Weka [29] or RapidMiner [30] to carry out them. This makes it necessary to include an intermediate adaptation step to convert the preprocessed MALDI data into the input format required by each tool.

In order to make the development of mass spectrometry (MS) proteomics applications easier, some frameworks such as OpenMS [31] and ProteoWizard [32], in C++, and MsInspect [16] in Java have been developed. An example of a tool developed using such frameworks is TOPP (The OpenMS Proteomics Pipeline) [33], which is based on the OpenMS framework.

In spite of the existence of such a great variety of tools and techniques for both the preprocessing and data analysis of MALDI based proteomic datasets, there is still a lack of specific tools that cover the whole process of MALDI data analysis, allowing the users to manage raw datasets, preprocess them and perform several analyses in a row, and allow the user to apply different ML and statistical techniques to analyze MALDI data. Moreover, most of the tools are intended to be used by a user with a bioinformatic profile, requiring programming skills.

This paper presents Mass-Up, an extensible open-source platform for MALDI data processing and analysis with ML and statistical techniques that has arisen from our previous experience working with MALDI data [34–36]. Mass-Up is an AIBench [35] based desktop application specifically created to perform complete analyses of MALDI data, allowing the users to: (i) import raw data from different formats (mzML, mzXML, csv); (ii) preprocess raw data; and (iii) perform different type of analyses, including supervised (e.g. biomarker discovery, predictor building, etc.) as well as unsupervised (e.g. clustering, biclustering, etc.) techniques.

The Mass-Up design is focused on two main objectives: coverage of the whole process of data analysis and simplicity of use. The first objective is accomplished in the way Mass-Up covers the whole process of MALDI data analysis, from data preprocessing to different types of analysis. The second is achieved through a design that allows Mass-Up to be used in a straightforward manner by non-informatician users. In addition, Mass-Up is multiplatform, open source and designed using a pluggable architecture which makes it easier for programmers to develop and include new algorithms and analysis tools.

Implementation

Mass-Up is a computer application for managing, preprocessing and analyzing MALDI data. Mass-Up is implemented in Java and it was constructed using the AIBench framework, which has been demonstrated to be suitable for developing proteomics applications [36], as it is the base framework of previously developed MS applications [37, 38]. Currently, Mass-Up has distributions for Windows and Linux operative systems.

This section briefly describes the Mass-Up workflow and the main algorithms and third-party libraries employed in each Mass-Up task.

Mass-Up workflow

Mass-Up includes a series of operations that can be classified into (i) input/output operations, (ii) preprocessing operations, and (iii) analysis operations. Figure 1 depicts the Mass-Up main workflow, where the most important operations are represented, along with the input files and data types managed by the application.

Third-party libraries

With the main goal of covering the whole process of MALDI data analysis, Mass-Up integrates several open source third-party libraries in order to accomplish different tasks, such as reading different MS data formats, preprocessing spectra, applying ML techniques, or visualizing data, among others. Additional file 1: Table S1 shows a general overview of the Mass-Up, including the algorithms and libraries used by each operation. All of these libraries have been transparently integrated into Mass-Up so that final users does not have to install them manually, since they are built-in in each Mass-Up distribution.

Mass-Up uses mzReader 1.2.0 [6] in order to read the mzXML and mzML MS data formats. To visualize MS spectra and to display quality control charts, Mass-Up uses JFreeChart 1.0.13, an open source Java library.

Mass-Up integrates two R packages for raw MALDI data preprocessing: MALDIquant [8] and MassSpecWavelet [13]. In addition, custom implementation of a fast peak matching algorithm based on a forward sliding window, named Forward, is also incorporated. Similarly to the alignment algorithm proposed by Kazmi et al. [39], this algorithm iterates the peaks from minimum to maximum m/z, adding them to the last cluster created if their m/z is within a distance from the average m/z of the cluster or creating a new cluster if not. This clustering algorithm does not allows clusters with two peaks from the same spectrum. In such case, only the peak that minimizes the average m/z of the cluster is kept.
Mass-Up makes use of Weka [29], a collection of ML algorithms for data mining tasks implemented in Java. These algorithms are used for classification and for principal component analysis (PCA). Three-dimensional PCAs are rendered by using Jzy3d [40], an open source Java library which can easily draw three dimensional scientific data. Clustering is executed using a custom implementation of an agglomerative hierarchical clustering algorithm and is rendered using an adapted version of JTreeView [41].

Biclustering is performed with Bimax [42], a powerful algorithm capable of generating all optimal biclusters, and BiBit [43], a novel approach for the extraction of biclusters from binary datasets that can obtain similar results to Bimax by using significantly less computation time and reducing the total number of generated biclusters. The aforementioned software, as well as a biclusters viewer, is integrated through the adaptation available in BiMS [44].

Results and discussion
Mass-Up is a flexible tool that includes several operations whose application depends on the analysis objectives. Therefore, there is no single way to use Mass-Up, and researchers must determine which analyses apply in their studies. In this section, several practical applications of the Mass-Up operations are presented, in order to demonstrate its usefulness and applicability.

Sample datasets
Two datasets from previous studies were selected to illustrate the Mass-Up functionality. A brief description of the main characteristics of both datasets is given in this section.

Cancer dataset
R. López-Cortés et al. [45] propose the use of gold nanoparticles to separate the proteins and peptides in human serum as a way to improve MALDI-based sample profiling. The protocol described in this work divides each sample into two sub-samples: pellet and supernatant. The MALDI spectra of both sub-samples are grouped by their corresponding conditions using three-dimensional PCA. The dataset is composed of sera from 5 patients with lymphoma, sera from 5 patients with myeloma, and sera from 2 healthy donors. As the classifications using pellet or supernatant are similar, only the latter sub-samples are used in the present work.

Wine dataset
Nunes et al. [46] propose a fast MALDI-based methodology to identify different types of wines. The authors carry out a preliminary study with 5 wines of different denominations of origin, in order to identify the most appropriate MALDI matrix. The study of the matrices found that CHCA is the most suitable for the purpose of classification. Each wine was spotted five times (i.e. 25 samples in total). Those 25 samples corresponding to the use of CHCA matrix are used as proof of concept.

Preprocessing
As previously stated, the preprocessing of MS data is a critical stage that converts raw data into a suitable input for further analysis. Inadequate or incorrect preprocessing...
methods can result in biased datasets, hindering the achievement of meaningful biological conclusions [12]. Preprocessing is essential since raw data contains both m/z values belonging to analytes, as well as m/z values derived from several forms of noise (e.g., chemical, electronic factors, etc.). The main objectives of preprocessing are [47] to remove noise without discarding any of the m/z values of interest, and to determine the m/z and intensity values with the best accuracy. The most common preprocessing tasks (shown in orange in Fig. 1) are smoothing, baseline correction, normalization, peak detection, and peak matching.

Usually, spectra are jagged, making it difficult to detect the m/z values of interest from the noise [47]. Thus, smoothing algorithms are usually applied to soften the spectra. The simplest techniques are based on the use of a sliding window, where the intensity of each m/z value is adjusted based on the intensity of the neighbor m/z values. Commonly used filters are moving average, Savitzky-Golay, Gaussian and the Kaiser window. Mass-Up provides two smoothing methods: moving average window and Savitzky-Golay, both from the MALDIquant library [8].

Baseline is a specific form of noise mainly driven by chemical perturbations, defined as an offset of the intensities of peaks that often show a dependency on the m/z value such that it is highest at low m/z values, presenting an exponential decay towards higher masses [47]. The most common baseline correction methods are monotone minimum, linear interpolation, LOESS, moving average of minima and continuous wavelet transform, all of which are available as free software in different packages such as Cromwell [15] (Matlab), PROCess [14] (R), MALDIquant [8] (R) or SpecAlign [19] (Java). Mass-Up allows the user to use the make up of all the baseline correction methods provided by MALDIquant (i.e. Top Hat, SNIP, Convex Hull, and Median).

A major constraint of MALDI is that the intensity of the m/z values is relative and can vary among species of the same sample. For this reason, normalization is typically used, making the intensities of different spectra comparable. The most common normalization methods are Total Ion Current (TIC), Probabilistic Quotient Normalization (PQN), Z-score, Linear, Mean or Median. Mass-Up allows the user to perform normalization using TIC, PQN or Median, all provided by the MALDIquant library [8].

The m/z detection can be defined as the process of selecting values of interest (i.e. related with target analytes) from a given spectrum, and it is normally applied after baseline correction and smoothing. Most of the peak detection methods are based on setting a threshold value in order to discard low intensity m/z values. The threshold can be absolute (e.g. minimum intensity) or relative (e.g. signal-to-noise ratio, SNR). However, Du et al. [13] proposed a method that performs m/z detection without explicit smoothing and baseline correction. This method is based on the continuous wavelet transform (CWT) and is publicly available in the MassSpecWavelet package. Mass-Up includes two m/z selection methods: the CWT-based method implemented in MassSpecWavelet [13], and a SNR-based method provided by MALDIquant [8], which uses a sliding window.

Finally, m/z matching is needed in order to make different spectra comparable. Without this matching procedure, the same molecule or metabolite (e.g. a certain peptide) can have different m/z values across replicates or samples. The objective of m/z matching methods is to find a common set of m/z locations in several spectra, so that all spectra will have the same m/z values for the same biological entities. In Mass-Up there are two fundamental types of m/z matching: intra-sample and inter-sample. The intra-sample matching is applied to the spectra obtained for the replicates of the same sample, while the inter-sample matching is applied to match m/z values across different samples, making them comparable and suitable for the subsequent analysis stage. Peak matching algorithms, are classified into two main groups: sequential algorithms based on a sliding window (e.g. the Forward algorithm, available in Mass-Up) and clustering based approaches (e.g. the MALDIquant algorithm [8], also available in Mass-Up).

The Mass-Up workflow also incorporates an additional filtering step that is very closely related to the matching process. This step is performed after the intra-sample matching and before the inter-sample matching, and allows the creation of a consensus spectrum for a sample, which summarizes the replicates of a sample in one single spectrum. In this step, the Percentage of Presence (POP) parameter allows the user to set the number of replicates where an m/z value must be present in order to be considered a valid consensus m/z value.

Finally, it is important to note that, while smoothing, baseline correction, normalization, and m/z detection are applied individually to each single spectrum in the Preprocess data operation, the m/z matching is applied to several spectra at the same time and is carried out by using the Match Peaks operation.

The new data generated by the Preprocess data and Match Peaks operations can be exported as comma-separated value files, allowing users to load them later with Mass-Up or to analyse them with other software packages. Mass-Up documentation includes information about exporting data and examples describing how it can be loaded in other languages such as R.

Quality control
When working with MALDI, low quality spectra may occasionally be generated. For example, spectra showing
to a low number of m/z values in comparison with other spectra, or containing many unique m/z values not present in their sibling replicates. These spectra may lead to failure when carrying out an analysis, or to incorrect conclusions. To prevent such a scenario, a quality control (QC) step was included, which may be performed between the preprocessing and the analysis tasks. The QC can be done at two levels: replicates, a low level QC analysis focused on the replicates of each sample; and samples, a high level QC analysis with additional information from the intra-sample m/z matching process.

At the replicates level, the user can check basic information about each individual spectrum (i.e. peak count, m/z range, intensity ranges, etc.) and compare all spectra in the dataset. Figure 2a shows a replicate QC analysis applied to the samples from conditions A, B, C, D, and E of the Wine dataset previously described. As can be noted from the boxplot, there are two outliers (red circles) and one extreme outlier (red triangle) in the Masses count chart. Specifically, the QC analysis has marked the E-CHCA.3-4, A-CHCA.1-4 and E-CHCA.2-2 samples as outliers due to the number of m/z values of their spectra. Therefore, before continuing with further analysis, it is recommended to carefully review these samples and even to repeat their analysis.

At the samples level, the user can check the performance of the intra-sample peak matching process, by comparing the percentages of presence (POP) counts (globally and by conditions) and the POP’s of each

---

**A Replicates Quality Control Analysis**

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Sample</th>
<th>Class</th>
<th>Mass count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum 1</td>
<td>E-CHCA.3-4</td>
<td>A</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 2</td>
<td>E-CHCA.1-2</td>
<td>A</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 3</td>
<td>E-CHCA.1-2</td>
<td>A</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 4</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 5</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 6</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 7</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 8</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 9</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 10</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 11</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 12</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 13</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 14</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 15</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
<tr>
<td>Spectrum 16</td>
<td>E-CHCA.1-3</td>
<td>B</td>
<td>32</td>
</tr>
</tbody>
</table>

**B Samples Quality Control Analysis**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mass</th>
<th>Masses</th>
<th>Masses</th>
<th>Masses</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-CHCA.1</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>A-CHCA.2</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>A-CHCA.3</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>A-CHCA.4</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>A-CHCA.5</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
</tbody>
</table>

**Fig. 2 Quality control view.** Details of the quality control analysis views for a replicates, and b samples. Box plot charts are used to summarize the more detailed information presented in the tables.
sample. As previously stated, the data table is more detailed and contains additional information from the intra-sample peak matching process, specifically: (i) POPXX columns, where XX is a percentage of the number of spectra, which show the number of peaks with a POP value exactly equal to XX; (ii) Align. Masses columns, which show the number of masses that have been matched across the spectra in the sample; (iii) Split > = XX columns, which show the percentage of masses that have a POP value higher or equal to XX; and (iv) Count >= XX columns, which show the number of masses that have a POP value higher or equal to XX (these are the columns used as categories in the charts).

Figure 2b shows a samples QC applied to the same samples as in the previous example. In this case, the box plot corresponds to the global POP count and shows that there are two outliers for the category "Count >= 60" and one outlier for the category "Count >= 100". Again, the outliers are highlighted in bold in the table.

**Biomarker discovery**

One of the main purposes of the MS analyses is the biomarker discovery [21, 22, 48]. A biomarker is a peptide, protein or other element of a sample that can identify and differentiate certain conditions such as phenotypes, strains, diseases or infections.

When identifying new biomarkers, it is necessary to distinguish between two types of data sets that can be analyzed: (i) those cases where there are a known and well defined number of conditions (e.g. healthy vs. diseased, different stages of a disease, etc.), and (ii) those cases where there are no conditions or where they are not clearly defined. In accordance with this differentiation, Mass-Up provides two types of biomarker discovery analysis: (i) the inter-label analysis, for the former type of data, and (ii) the intra-label analysis, for the latter.

In the inter-label analysis, the user can perform the appropriate statistic tests to identify those peaks that can be potential biomarkers to differentiate the conditions. Four different tests of independence were included in Mass-Up following the recommendations given by McDonald [49], where tests are chosen depending on the number of samples and conditions of the dataset, as shown in Table 1. Taking into account that the number of samples in MALDI experiments is generally below 1000, the Fisher’s exact test and the randomization test are the tests more commonly applied. As each test is performed independently for each m/z value, the Benjamini-Hochberg FDR correction is applied to take into account the number of m/z values analyzed and reduce the number of false positives.

By using the inter-label analysis in the Wine dataset (shown in Fig. 3a), we can analyze all the samples of conditions A, B, C, D, and E. In this case, the randomization test is applied in order to identify statistically relevant m/z values, as the number of samples is lower than 1000 (5 samples for each of the 5 wines for a total of 25 samples) and the number of conditions is lower than 2 (5 wine denominations). The first three columns contain the m/z value, the p-value, and the q-value respectively, while the other columns show in which samples the m/z values are present. As can be seen, the peaks with a q-value < 0.05 are clear candidates to be biomarkers as they differentiate certain conditions from others.

In the intra-label analysis, the user can identify those m/z values that are representative of one or more samples, in a more exploratory fashion. In this scenario, it is possible to identify the biomarkers of a specific sample or discover groups of samples with a similar profile that may, therefore, be related. This analysis is particularly useful, for example, when working with different strains of the same bacteria and the user wants to identify those peaks that are unique for a certain strain.

By using the intra-label analysis in the Cancer dataset (shown Fig. 3b), we can analyze the samples of the condition Myeloma and configure the analysis to identify those peaks present in the “MA” sample (i.e. Myeloma A) and not present in the rest of the samples. The identification of these peaks may be useful, for example, to explain the abnormal behaviour of a sample when compared to other samples from the same condition. Specifically, the Intra-label Biomarker Discovery view shows how we are looking for specific peaks of the sample MA (i.e. Myeloma A), that is, peaks that are in this sample but not in the others.

**Principal component analysis**

PCA is a mathematical procedure that uses orthogonal transformation to convert a set of observations (i.e. samples) of possibly correlated variables (i.e. m/z values) into a set of values of linearly uncorrelated variables called principal components (PC), whose dimensionality is expected to be lower than the dimensionality of the original data set.

Once the PC are calculated, they can be used to represent the samples in a 3-dimensional space. By assigning a different color to each condition’s samples, users can visually identify if there is a separation between conditions. If such were the case, then the conditions would be distinguishable. The PCA view also includes additional

---

**Table 1** Tests of independence applied depending on the number of samples and conditions

<table>
<thead>
<tr>
<th>Condition</th>
<th>1 &lt;= 1000 samples</th>
<th>&gt;1000 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 conditions</td>
<td>Fisher’s exact test</td>
<td>Yates’ chi-square test</td>
</tr>
<tr>
<td>&gt;2 conditions</td>
<td>Randomization test</td>
<td>Chi-square test</td>
</tr>
</tbody>
</table>
information about the PCA, such as the eigenvectors and their corresponding eigenvalues and retained variances, for a better results interpretation.

As previously stated, López-Cortés et al. [45] demonstrate that the spectra of supernatant sub-samples of the Cancer dataset can be grouped by their corresponding conditions using PCA. Figure 4a shows the result of applying PCA to this set of samples in Mass-Up. As it can be clearly seen, the three conditions are separable in the 3-dimensional space.

Cluster analysis
Cluster analysis allows finding groups of similar spectra among all the samples being studied. In the case of unlabeled data, it allows discovering hidden or previously unknown subgroups of samples. In the case of labeled data, it allows the user to check if the different conditions present in a dataset are separable by means of the m/z values of each sample.

Mass-Up incorporates a hierarchical clustering algorithm for the construction of a hierarchy of sample groups (named clusters). The algorithm included is agglomerative and follows a bottom-up approach, meaning that it is constructed iteratively, starting with each sample in its own cluster, and merging the closest pair of clusters on each step. In order to decide which clusters should be merged, a measure of dissimilarity between clusters is required. In our case, this is achieved by using a distance metric, which measures the distance between two samples, and a linkage criterion, which specifies the dissimilarity of clusters. Mass-Up includes the Euclidean and Hamming distances as distance metrics, and the
complete, single and average functions as linkage criteria. The results of a hierarchical clustering are usually presented in a dendrogram.

An important aspect when performing a cluster analysis in Mass-Up is that the user can decide whether to use intensities (i.e. a m/z value is represented by the value of its peak intensity) or not (i.e. a m/z value is represented by its peak presence or absence). The Euclidean distance is the most suitable when using intensities while the Hamming distance is the most appropriate when using presence/absence of peaks.

In each cluster analysis, two hierarchical clusterings are constructed: one for the samples and one for the m/z values. For the visualization of the results, Mass-Up incorporates an adapted version of JTreeView, a software for the visualization and analysis of gene expression data. We have adapted it to MS, so that in our specific case the rows represent peaks instead of genes, while columns still represent samples. This representation also includes a heat map, which is combined with two dendrograms that represent the aforementioned hierarchical clusterings. The individual values contained in the heat map matrix are displayed as colors and they can represent (i) the intensity level of the corresponding peak (red if the peak has an intensity value of 1; green if the peak has an intensity of 0; and intermediate colors for intensities between 0 and 1), or (ii) the presence or absence of the peak (red if the peak is present and green if the peak is not present). It is important to note that to achieve a correct representation using intensities, the m/z values must be scaled between 0 and 1 during the raw data preprocessing.

Figure 4b shows the results of applying hierarchical clustering to the Cancer dataset used as proof-of-concept. As the dendrogram illustrates, the three conditions are well separated since the samples of each condition can be grouped together.

Finally, it is worth noting that the cluster analysis can be used with a list of previously selected peaks. This
way, the cluster analysis will be focused on analyzing only these peaks. This list can be obtained by exporting the biomarkers identified in the inter-label analysis. In such a situation, this feature is useful to qualitatively verify if a list of potential biomarkers is enough to separate or differentiate between the conditions of study.

**Bicluster analysis**

Although biclustering techniques have been successfully used with gene expression data for over a decade, it is only very recently that those techniques have been applied to MS data [50]. Biclustering is a data mining technique that allows simultaneous clustering of the rows and columns of a matrix. It has been successfully applied to analyze microarray data due to their ability to discover co-expressed genes under certain samples [51]. In contrast to traditional clustering techniques, where each gene in a given cluster is defined under all the samples, biclustering algorithms propose groups of genes that show similar activity patterns under a subset of the experimental samples.

In previous studies, we have proposed a novel workflow for the application of biclustering to MALDI data. In addition, the adequacy of applying biclustering to analyze such data by comparing biclustering and hierarchical clustering over two real datasets has also been evaluated [44]. Biclustering has shown the ability to discover groups of samples that are similar but only in a subset of m/z values, which represent a new kind of hidden hypothesis that are difficult to be discovered by classic clustering algorithms, such as hierarchical clustering, which are based on a global comparison of samples including all m/z values.

The biclustering algorithms selected in the study and included in Mass-Up (i.e. Bimax and BiBit) use a binary dataset as input where 1 represents a peak presence, and 0 represents a peak absence. These algorithms will look for groups (i.e. biclusters) of 1’s, that we call presence patterns. Nevertheless, in certain cases, it can be desirable to extract other type of patterns, such as absence patterns (i.e. biclusters of 0’s) or simple presence/absence patterns (i.e. biclusters of 1’s and 0’s in one direction). López-Fernández et al. [44] further discuss how to prepare an input MALDI dataset into a suitable form to look for these three types of patterns.

Mass-Up provides an operation to apply this technique to both labeled and unlabeled samples. The user has to select the biclustering algorithm to use, the type of pattern and the biclustering mode (i.e. whether rows of the biclustering binary matrix are peaks or samples). In addition, the user can also establish the minimum dimensions of the output biclusters. If the input data is labeled, the user can also indicate whether the output of the biclustering must be filtered in order to only retrieve those biclusters where most of the samples belong to the same condition or label, known as class-biclusters. After performing a biclustering analysis, results can be inspected in the biclustering viewer, an intuitive view that shows a list of the generated biclusters as well as a heat map. If a bicluster is selected, it will be highlighted in the heat map, which is automatically rearranged in order to show the bicluster in the upper left corner.

In order to demonstrate the usefulness of this module, we considered the Cancer dataset used in previous sections, and applied biclustering by means of the BiBit algorithm in the hope of finding presence class-biclusters. Figure 4c shows one presence class-bicluster for each class, where each column represents one m/z value and each row represents a sample. As shown, each class bicluster includes a group of m/z values with the same pattern of presence in the samples of one condition, and a variable pattern of presence in the rest of the samples. When using a presence class-bicluster, only presence is taken into account to create the class-bicluster, whereas when using a presence/absence class-bicluster, the absence is also taken into account.

**Classification analysis**

Sample classification is the ability to predict the label of a sample given a training set of labelled samples, therefore, the capacity of producing a diagnosis machine [10, 24, 26]. Through the "Classification Analysis" operation, the user can determine which classifier performs best for the data under analysis. This operation provides an interface adapted from the Weka software that allows the user to select and to configure a classifier, and to evaluate its performance by means of a cross-validation scheme. The output log of the evaluation process summarizes the performance of the classifier using different statistical measurements, such as accuracy, kappa, precision, recall, etc. In addition, you can make a receiver operating characteristic (ROC) analysis per condition.

Classification analyses are performed in the classification view (shown in Fig. 5), which was adapted from the Weka software. Through this view, the user can select a classifier and a validation scheme (i.e. cross-validation or percentage split) to perform an evaluation. As shown, the results report includes several global and per-class statistics, as well as the resulting confusion matrix. Using these operations, users can assess whether the data being analyzed is suitable for classification, as well as determine which classification algorithm is best.

**Performance notes**

Although the performance is very dependent on the number of samples and the computer being used, some tests have been carried out in order to provide some performance guidelines. We have created a test dataset of 490 samples based on the Wine dataset, and then, we have executed the most common workflow of Mass-Up
under an Intel Core i5 M520 with 8GB of RAM and Kubuntu 13.10 as OS. It is important to note that the size of this test dataset clearly exceeds the common size of a dataset in a MALDI-TOF MS experiment, which usually are no longer than 200 samples.

It took about 90 s to load 490 raw samples and about 200 s to fully preprocess them. Once the data is preprocessed and prior to perform any analysis, we must apply the Match Peaks operation, which could be executed in less than 30 s using the MALDiquant algorithm and in less than 3 s using the Forward algorithm. Most of the analyses (quality control, PCA, classification and intra-label analyses) could be executed in less than 5 s, while clustering, biclustering and inter-label analysis took more time. On one hand, clustering analysis took less than 20 s and the biclustering execution time depends on the algorithm selected (less than 20 s for Bibit and about 15 min for Bimax). On the other hand, inter-label biomarker discovery based on 10000 randomizations took about 8 min.

Conclusions
In this paper we have presented Mass-Up, a new software for the analysis of MALDI data. This is an application that covers the whole process of MALDI data analysis, from data preprocessing to complex data analyses.

Mass-Up incorporates the most common analyses, aside from protein identification and focusing in biomarker discovery, such as statistical tests-based biomarker discovery, clustering, PCA, and classification. In addition, other less common analyses such as quality control and biclustering are also included. Therefore, Mass-Up provides users with a wide range of tools to analyze and explore their MALDI data.

Unlike other MS tools, Mass-Up provides a friendly graphical user interface designed to avoid the need for a bioinformatics expert to use it. The tutorial and examples included in Mass-Up tool and in the project homepage will guide users through the different operations included, making it use suitable for any user.

Finally, Mass-Up is open to further extension, such as including new operations or improving the available ones.

Availability and requirements
The Mass-Up software is freely available from the project homepage on http://sing.et.uvigo.es/mass-up. Additionally, source code can be downloaded from https://sourceforge.net/projects/mass-up/.

Project name: Mass-Up.
Project home page: http://sing.et.uvigo.es/mass-up
Operating system: Platform independent, packaged for Windows and Linux.
Programming language: Java version 7.
Other requirements: Mass-Up has no other requirements since distributions are self-contained.
License: Version 3 of the GNU General Public License (GPLv3).

Additional file

Additional file 1: Table S1. Detailed list of the source and version of the algorithms and libraries used in Mass-Up. (DOCX 15 kb)

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
JLC, HWS, FF, and MR conceived the idea of the software. FF, MR, HL, and DG designed the software architecture. MR, HL, and DG implemented the software. All authors wrote, read and approved the final manuscript.

Authors’ information
Not applicable.

Acknowledgements
This work was partially funded by the (l) IN2EU-14-08 project from the Provincial Council of Ourense, (ll) IN2009-14-057-C02-02 project from the Spanish Ministry of Science and Innovation, the Plan E from the Spanish Government and the European Union from the DEDD, (iii) FPI/RECPP-2012-2013.1 project from the European Union Seventh Framework Programme under grant agreement n° 316365, BIOCAPS, and (iv) DTH-DTO: Desarrollo de Tecnicas y Herramientas para el Tratamiento de Datos Oblicuos” Contract-Programme from the University of Vigo. H. López-Fernández was supported by pre-doctoral fellowships from the University of Vigo and Xunta de Galicia, H. M. Santos acknowledges the postdoctoral grant SRF/BPD/37997/2010 provided by Fundación para a Ciência e a Tecnologia, Ministério da Educação e Ciência (FCT-MEC, Portugal).

Author details
1Informatics Department, Universidad de Vigo, Campus Universitario A Lagoas s/n, 32004 Ourense, Spain. \Instituto de Investigación Biomédica de Vigo (IBV). Vigo, Pontevedra, Spain. 2BIOSCOPE Research Group, UCIBIO-REQUIME, Department of Chemistry, Faculty of Science and Technology, Universidade NOVA de Lisboa, Caparica, Setubal, Portugal.

Received: 6 May 2015 Accepted: 28 September 2015
Published online: 05 October 2015

References
5. Côe RG, Beiring R, Martens L, jMrML, an open-source Java API for mzML, the PSI standard for MS data. Proteomics. 2010;10:1332-5.


Page 11 of 12