Inteligencia Artificial 19(57), 22-25
doi: 10.4114/ia.v19i57.1144

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Resumen de tesis: Aplicación de técnicas de minería de datos e inteligencia artificial a datos de espectrometría de masas para el descubrimiento de conocimiento

H. López-Fernández
Departamento de Informática, Escuela Superior de Ingeniería Informática, Universidad de Vigo. Campus Universitario As Lagoas S/N, 32004, Ourense.
hlfernandez@uvigo.es

Resumen La espectrometría de masas empleando desorción/ionización láser asistida por matriz con detector de tiempo de vuelo (MALDI-TOF) 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), aprendizaje automático (ML) 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.

Keywords: mass spectrometry, artificial intelligence, data mining, knowledge discovery.
Palabras clave: espectrometría de masas, inteligencia artificial, minería de datos, descubrimiento de conocimiento.

1 Introduction

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, 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. 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 limpia, eliminando los picos pertenecientes al ruido y dejando los picos reales.

La espectrometría de masas empleando desorción/ionización láser asistida por matriz con detector de tiempo de vuelo (MALDI-TOF, M*atrix Assi0sted Laser Desorption Ionization coupled to Time of Flight Analyzers*) ha

ISSN: 1988-3064(on-line)

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: la adquisición de datos, el preprocesado y el 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 [2-4], empleando para ello listas de picos.

Esta tesis explora el flujo de análisis de datos MALDI-TOF presentado, incluyendo una etapa adicional de control de calidad inmediatamente después del preprocesado, con el fin de detectar espectros de baja calidad o anómalos que puedan sesgar o dificultar los análisis posteriores. En este contexto, el objetivo principal de este trabajo es la aplicación de técnicas de minería de datos y AI al análisis de datos de espectrometría de masas para el descubrimiento de nuevo conocimiento.

2 Evolución de la investigación y contribuciones

En 2012, el trabajo doctoral comenzó en dos de las líneas de trabajo definidas. Por una parte, el desarrollo de *MLibrary* [5], una base de datos con un motor de búsqueda diseñados para asistir al usuario en la detección e identificación de anabolizantes androgénicos esteroideos (AAS, A*ndrogenic 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.

Por otra parte, el estudio y comparación de los métodos disponibles de preprocesado de datos en crudo de MALDI-TOF. En este momento, 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 [6].

En 2013, se propuso un novedoso proceso para la aplicación de biclustering a datos de MALDI-TOF [7]. 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.

En 2013, se decidió unificar todos los componentes desarrollados anteriormente en una única plataforma, dando lugar al desarrollo de *Mass-Up* [8], 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*) visualización basada en el análisis de componentes principales (PCA, *Principal Component Analysis*) y (*v*) clasificación de grandes conjuntos de muestras.

3 Estructura del trabajo

El trabajo realizado se ha organizado en torno a tres contribuciones principales, las cuales han sido publicadas en revistas internacionales de impacto indexadas en el *Journal Citation Reports* (JCR).

En la primera de estas contribuciones se presenta *MLibrary* [5], 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.

En la segunda contribución, se 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 [6]. Existen distintos métodos para llevar a cabo las principales tareas del preprocesamiento como la 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 trabajo doctoral concluye con *Mass-Up* [8], 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*, desarrollada empleando el framework AIBench [9], 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.

4 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. Además, se desarrolló un algoritmo de emparejamiento de picos llamado *Forward*, 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. 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.

Además, se puso a disposición de la comunidad científica el software *Mass-Up* (http://sing.ei.uvigo.es/mass-up/), 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 utilidad está siendo refrendada por el aumento del número de estudios que hacen uso de este software [10-12] 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 esta plataforma ha sido actualizada continuamente 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 (*ii*) incluir nuevos tipos de análisis.

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.

Agradecimientos

Me gustaría agradecer la ayuda y apoyo de mis directores, D. Glez-Peña y M. Reboiro-Jato, sin quienes este trabajo doctoral no hubiese sido posible, así como el apoyo y ánimo de F. Fdez-Riverola, líder del grupo de Sistemas Informáticos de Nueva Generación (SING), durante estos años. Quiero agradecer también a la Universidad de Vigo y a la Xunta de Galicia las becas predoctorales que he disfrutado y que me han permitido realizar esta tesis doctoral.

Referencias

[1] Eidhammer I, Flikka K, Martens L, Mikalsen S-O. Computational Methods for Mass Spectrometry Proteomics. 1st edition. Wiley-Interscience; 2008. doi: 10.1002/9780470724309.

[2] Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. OMICS J Integr Biol. 2013; 17:595–610. doi: 10.1089/omi.2013.0017.

[3] McDonald RA, Skipp P, Bennell J, Potts C, Thomas L, O'Connor CD. Mining whole-sample mass spectrometry proteomics data for biomarkers - An overview. Expert Syst Appl. 2009; 36:5333–5340. doi:10.1016/j.eswa.2008.06.133.

[4] Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A, Le Q-T. Sample classification from protein mass spectrometry, by "peak probability contrasts." Bioinformatics. 2004; 20:3034–3044. doi: 10.1093/bioinformatics/bth357.

[5] Galesio M, López-Fdez H, Reboiro-Jato M, Gómez-Meire S, Glez-Peña D, Fdez-Riverola F, Lodeiro C, Diniz ME, Capelo JL. Speeding up the screening of steroids in urine: Development of a user-friendly library. Steroids. 2013; 78:1226–1232. doi:10.1016/j.steroids.2013.08.014.

[6] Fernández HL, Jato MR, Peña DG, Riverola FF. A comprehensive analysis about the influence of low-level preprocessing techniques on mass spectrometry data for sample classification. Int J Data Min Bioinforma. 2014; 10:455. doi:10.1504/IJDMB.2014.064897.

[7] López-Fernández H, Reboiro-Jato M, Madeira SC, López-Cortés R, Nunes-Miranda JD, Santos HM, Fdez-Riverola F, Glez-Peña D. A Workflow for the Application of Biclustering to Mass Spectrometry Data. In 7th International Conference on Practical Applications of Computational Biology & Bioinformatics. Edited by Mohamad MS, Nanni L, Rocha MP, Fdez-Riverola F. Springer International Publishing; 2013:145–153. [Advances in Intelligent Systems and Computing, vol. 222]. doi: 10.1007/978-3-319-00578-2_19

[8] López-Fernández H, Santos HM, Capelo JL, Fdez-Riverola F, Glez-Peña D, Reboiro-Jato M: Mass-Up. an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery. BMC Bioinformatics. 2015; 16:318. doi: 10.1186/s12859-015-0752-4.

[9] Fdez-Riverola F, Glez-Peña D, Lõpez-Fernández H, Reboiro-Jato M, Méndez JR. A JAVA application framework for scientific software development. Softw - Pract Exp. 2012; 42:1015–1036. doi: 10.1002/spe.1108.

[10] Fernández-Costa C, Reboiro-Jato M, Fdez-Riverola F, Ruiz-Romero C, Blanco FJ, Capelo-Martínez J-L. Sequential depletion coupled to C18 sequential extraction as a rapid tool for human serum multiple profiling. Talanta. 2014; 125:189–195. doi: 10.1016/j.talanta.2014.02.050

[11] Araújo JE, Santos T, Jorge S, Pereira TM, Reboiro-Jato M, Pavón R, Magriço R, Teixeira-Costa F, Ramos A, Santos HM. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry-based profiling as a step forward in the characterization of peritoneal dialysis effluent. Anal Methods. 2015; 7:7467–7473. doi: 10.1039/C5AY00620A.

[12] López-Cortés R, Formigo J, Reboiro-Jato M, Fdez-Riverola F, Blanco FJ, Lodeiro C, Oliveira E, Capelo JL, Santos HM. A methodological approach based on gold-nanoparticles followed by matrix assisted laser desorption ionization time of flight mass spectrometry for the analysis of urine profiling of Knee Osteoarthritis. Talanta. 2016; 150: 638–645. doi:10.1016/j.talanta.2015.06.043.